A 360° CUSTOMER LIFETIME VALUE PREDICTION METHOD USING MACHINE LEARNING FOR MULTI CATEGORY E-COMMERCE COMPANIES

GÜLŞAH YILMAZ BENK

BOĞAZİÇİ UNIVERSITY

2023

A 360° CUSTOMER LIFETIME VALUE PREDICTION METHOD USING MACHINE LEARNING FOR MULTI CATEGORY E-COMMERCE COMPANIES

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Management Information Systems

by

Gülşah Yılmaz Benk

Boğaziçi University

2023

A 360° Customer Lifetime Value Prediction Method Using Machine Learning for Multi Category E-Commerce Companies

The thesis of Gülşah Yılmaz Benk

has been approved by:

Assoc. Prof Bertan Y. Badur _____
(Thesis Advisor)

Prof. Sona Mardikyan _____
(Thesis Co-Advisor)

Prof. Hande Türker _____

Assist. Prof. Nazım Taşkın _____

Assoc. Prof. Mehmet Nafiz Aydın _____
(External Member)

Assist. Prof. Nazım Ziya Perdahçı _____
(External Member)

January 2023

DECLARATION OF ORIGINALITY

I, Gülşah Yılmaz Benk, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution.
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature……

Date……

ABSTRACT

A 360° Customer Lifetime Value Prediction Method Using Machine Learning for Multi Category E-Commerce Companies

The Customer Lifetime Value (CLV) prediction methods that are used by e-commerce companies are mainly focusing on a specific group of customers with multiple transaction data and therefore remain unable to value the one-time purchasers. As a result, the management of these companies is incapable of valuing all customers and the overall company. Thus, each type of user needs a different way to predict CLV. In this thesis, we intend to develop a novel 360° holistic technique for the prediction and use of CLV models in the marketing management of multi-category e-commerce companies to enhance their strategic decision-making. The research compared the proposed framework which was constructed with several outputs (CLV, DPC, and TAS) with other ML models to evaluate the new variables created based on relationship marketing theory (RMT) and to demonstrate the best model which is more appropriate for multi-category e-commerce companies' usage. To make this result useful, we created customer clusters that enable management to separate end-users according to the three outputs. Finally, Shapley values obtained using explainable artificial intelligence (XAI) are then utilized to understand the DNN's findings. The results showed that using XAI shows which factors are more crucial to the results. Overall, the proposed model helps marketing management teams in planning their operations efficiently by differentially allocating their resources to specific types of customers based on their profitability which provides more strategic decision-making.

ÖZET

Çok Kategorili E-Ticaret Şirketleri için Makine Öğrenimi Kullanarak 360° Müşteri Yaşam Boyu Değer Tahmini

E-ticaret şirketleri tarafından kullanılan mevcut Müşteri Yaşam Boyu Değer (CLV) tahmin yöntemleri, yalnızca birden fazla işlem verisine sahip mevcut müşteriler olan belirli bir müşteri grubuna odaklanmakta ve bu nedenle tek seferlik satın alan yeni müşterilerin değerlerini hesaplayamamaktadır. Bu sebeple, bu şirketlerin yönetim kadroları, tüm müşterilerini ve şirketin değerlemesini yaparken yetersiz kalmaktadır. Bu soruna bir çözüm olarak, bu çalışmada, çok kategorili perakende e-ticaret şirketlerinin pazarlama yönetiminde CLV modellerinin öngörülmesi ve uygulanması için yeni bir 360° bütünsel yöntem geliştirmeyi amaçladık. Buna ek olarak gelecekte hangi müşterilerin en yüksek geliri getireceğini belirlemek için CLV'ye ek olarak iki yeni faktör (farklı ürün kategorisi (DPC) ve harcanan miktarın trendi (TAS)) ekleyerek mevcut kullanıcılar için CLV'yi tahmin etmek için kullanılan bu yöntemleri genişlettik. Bu iki yeni faktörü, pazarlama ilişki teorisini genişleterek oluşturduk. CLV, DPC ve TAS'ı birlikte tahmin ederken önerilen yapıyı test etmek için ilgili literatürde ilk kez CLV tahmini için çoklu tahminli bir derin sinir ağı (DNN) modelini kullandık. Bu sonucu gerçek hayatta uygulanabilir kılmak için, çok kategorili e-ticaret şirketlerinin son kullanıcıları üç değişkene göre gruplara ayırmasına olanak tanıyan müşteri kümeleri oluşturduk. Önerilen modelin çok kategorili e-ticaret firmalarının yöneticilerinin, kaynaklarını belirli müşteri türlerinin kârlılıklarına göre farklı şekilde dağıtarak operasyonlarını daha verimli bir şekilde planlamasına yardımcı sonuçlarına vardık.

CURRICULUM VITAE

NAME: Gülşah Yılmaz Benk

DEGREES AWARDED

PhD in Management Information Systems, 2023, Boğaziçi University

MA in Management Information Systems, 2011, Boğaziçi University

BA in Statistics and Computer Science, 2007, Baskent University

AREAS OF SPECIAL INTEREST

Customer Lifetime Value, Data Mining, Machine Learning, Deep Learning, Explainable Artificial Intelligence

PROFESSIONAL EXPERIENCE

Data Science Lead, Google 2013 – Present

Business Analyst, MSU Software and Consultancy, 2010 – 2013

Research Assistant, Department of Management Information Systems, Boğaziçi University, 2007 – 2011

PUBLICATIONS

International Journal Articles

Yılmaz Benk, G.; Badur, B.; Mardikyan, S. A New 360° Framework to Predict Customer Lifetime Value for Multi-Category E-Commerce Companies Using a Multi-Output Deep Neural Network and Explainable Artificial Intelligence. *Information* 2022, *13*, 373. https://doi.org/10.3390/info13080373

Yılmaz, G., B. Y. Badur, and S. Mardikyan. "Development of a Constraint Based Sequential Pattern Mining Tool." Computers and Software (2011): 191.

ACKNOWLEDGEMENTS

First and foremost, I am extremely grateful to my supervisors, Prof. Sona Mardikyan and Assoc. Prof. Bertan Badur for their invaluable advice, continuous support, and patience during my PhD study. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life.

I would like to send my thanks to the thesis committee –Prof. Hande Türker, Assoc. Prof. Mehmet Nafiz Aydın, and Assist. Prof. Nazım Taşkın and Assist. Prof. Nazım Ziya Perdahçı for caring to be a part of the committee, their constructive criticism, and valuable comments.

I am particularly grateful for the support I received from all my managers and colleagues at Google. I would especially like to express my gratitude to Mehmet Keteloğlu, Dennis Cardoso, James Hirschhorn, Rachel Meyers, Danny Bowen who are my lovely managers for their endless support during these nine years of time.

I am beyond grateful to my mother Ayşe Yılmaz, and my sisters Filiz, Funda and my brother Murat Yılmaz for their unconditional love and support throughout my entire life. I would not go this far without their unwavering supports. They are always there for me, which is something that cannot be provided by anyone but the family. Also, I'd like to thank my bellowed nieces Ece and Cemre and bellowed nephews Batuhan and Yağız for their unconditional support, encouragement and love.

I would like to offer my special thanks to my in-laws Hasan, Semihat, Mesut, Cebrail, Selin and Ahmet Benk also Prof. Dr. Dilek Benk Şilfeler, Murat, Mete and Zeynep Şilfeler from the bottom of my heart for their constant support and patience. I'd

especially like to thank Assoc. Prof. İbrahim Şilfeler who, although no longer with us, continues to inspire by his example.

I would like to express my sincere gratitude to my friends from Boğazici University, Dr. Aysun Bozanta Hakyemez, Dr. Tuğrul Cabir Hakyemez, Meltem Mutlutürk, Dr. Abide Coşkun and my childhood and lifelong friends Burçak Arıkan, Alper Tosyalı and Ceren Tosyalı, my best friends forever Çiğdem Canatan and İrem Erkaya for their unwavering support and belief in me.

Above all, it was my caring, loving and supporting husband, İsmail Benk, who firmly held my hands when I was lost many times searching for the light and courage. Your encouragement when the times got rough are much appreciated. This thesis would not be even possible without his loving and calming support. Without his tremendous understanding and encouragement in the past few years, it would be impossible for me to complete my study.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: LITERATURE REVIEW 9

2.1 Customer lifetime value (CLV) 12

2.2 Existing vs new users 20

2.3 Probability models 25

2.4 Machine learning models 27

2.5 This study 32

CHAPTER 3: PROBLEM DEFINITION AND MODELS 34

3.1 CLV prediction for new users 34

3.2 CLV prediction for existing users 36

3.3 CLV prediction models 44

CHAPTER 4: DATA ANALYSIS AND PRE-PROCESSING 57

4.1 Data 58

4.2 Data analysis 63

4.3 Pre-Processing 69

4.4 Performance measures 73

CHAPTER 5: NEW USER MODEL RESULTS 75

5.1 ML models 75

5.2 Hybrid model results for new users 78

5.3 Comparison of model results for new users 79

5.4 Customer segmentation for new users 81

CHAPTER 6: EXISTING USER MODEL RESULTS 83

6.1 Pareto/NBD models 83

6.2 ML models 84

6.3 Comparison of model results for existing users 87

6.4 Customer segmentation for existing users 95

CHAPTER 7: DISCUSSION AND CONCLUSION 97

7.1 Conclusion 107

7.2 Research implications 112

7.3 Managerial implications 114

7.4 Limitations 117

7.5 Future research 118

APPENDIX A: THE FEATURES EXTRACTED FROM THE DATASET 119

APPENDIX B: SHAP VALUES FOR DPC AND TAS 122

REFERENCES 124

LIST OF TABLES

Table 1. Hyperparameter Ranges and Optimal Hyperparameters for DNN Models 53

Table 2. Optimal Hyperparameters and Range of Hyperparameteres for Multi-Output DT 54

Table 3. Optimal Ranges and Hyperparameters for Multi-Output RF 55

Table 4. CLV Prediction with Decision Tree for New Users 76

Table 5. CLV Prediction with Random Forest for New Users 77

Table 6. CLV Prediction with Gradient Boosting for New Users 78

Table 7. CLV Prediction with XGBoost Regressor for New Users 78

Table 8. Logistic Regression Phase One of Hybrid Model 79

Table 9. ML and Hybrid Model Comparison for New Users 80

Table 10. Hit-ratios (Top 20%) for ML and Hybrid Model Outputs 81

Table 11. Cluster Results for New Users 82

Table 12. Pareto/NBD Model Results 83

Table 13. CLV Prediction with Decision Tree for Existing Users 85

Table 14. CLV Prediction with Random Forest for Existing Users 86

Table 15. CLV Prediction with Gradient Boosting for Existing Users 87

Table 16. CLV Prediction with XG Boost for Existing Users 87

Table 17. ML and Pareto/NBD Model Comparison for Existing Users 88

Table 18. Multi-Output and Single-Output Model Results Based on MAE, MSE and RMSE 89

Table 19. Hit-ratios (Top 20%) for Model Results 90

Table 20. Hit-ratios (Top 20%) for Multi-Output DNN Model, Malthouse and Blattberg's (2005) Study and Chian and Yang's (2018) Study 91

Table 21. Cluster Results 95

LIST OF FIGURES

Figure 1. 360° holistic CLV prediction framework 43

Figure 2. Multi-Output DNN structure 50

Figure 3. Model architecture for multi-output DNN 52

Figure 4. Product category distribution by count 66

Figure 5. Product category revenue distribution by percentage 66

Figure 6. Example transactions of two users one with positive trend one with negative trend 72

Figure 7. Trend segment by total amount spent (totalrevenue) 73

Figure 8. Random Forest feature importance 77

Figure 9. Random Forest feature importance for existing users 86

Figure 10. Decile charts for CLV, DPC, and TAS respectively (top row: multi-output model, bottom row: single-output model result). 92

Figure 11. Feature importance levels defined by average SHAP values for CLV (a) multi-output model CLV SHAP values, (b) single-output model CLV SHAP values. 93

ABBREVIATIONS

ANN: Artificial Neural Networks

BTYD: Buy Till You Die

CE: Customer Equity

CLV: Customer Lifetime Value

CRM: Customer Relationship Management

DNN: Deep Neural Network

DPC: Distinct Product Category

DT: Decision Tree

LF: Latent Factor

MAE: Mean Absolute Error

ML: Machine Learning

MSE: Mean Squared Error

NBD: Negative Binomial Distribution

RELU: Rectified Linear Unit

RF: Random Forest

RFM: Recency, Frequency, Monetary

RMSE: Root Mean Squared Error

RMT: Relationship Marketing Theory

SMOTE: Synthetic Minority Oversampling Technique

SVM: Support Vector Machine

TAS: Trend of Amount Spend

XAI: Explainable Artificial Intelligence

ZILN: Zero Inflated Log Normal

CHAPTER 1

INTRODUCTION

Online shopping has recently surpassed offline shopping in popularity for many reasons, including its greater efficiency, convenience, and affordability as well as its wider selection of products. Convenience drives the emergence of businesses that sell across numerous product categories and provide a range of goods at various price points. E-commerce companies have evolved into online malls due to their convenience and numerous advantages that online shopping has over in-store shopping (Schauer, 2022). As a consequence, there are now much more online shops and more people purchasing online. Competition heats up as more online retailers enter the market. As a result, to compete in the e-commerce market, businesses must improve their business planning (Jasek, Vrana, Sperkova, Smutny, & Kobulsky, 2018).

Having a clear and accurate customer value may facilitate more precise strategic decision making in company strategies and relationship marketing. For a business to stay ahead of the competition, it must make sure that all factors influencing its long-term performance are taken into account. One of the most important factors that organizations should use to get a competitive edge is customer lifetime value (CLV). CLV is one of the key metrics used to evaluate marketing initiatives and is a crucial part of customer relationship management (CRM) (Eid, Robin, & Wierdak, 2021). Earlier studies discovered that businesses may boost their profitability by determining which clients are the most valuable and by concentrating their marketing efforts to build strong, long-term connections with them. The CLV is a popular value that businesses use to assess the profitability and worth of a client. It displays the consumer's net present value based on

potential purchases (Kumar & Reinartz, 2016). The CLV helps firms rate clients according to how much they contribute to profits. This may be utilized to create and run programs at the customer level that will increase lifetime profits. One big challenge that modern marketers are facing is they frequently target average consumers. However, marketing to the average customer is no longer sufficient because of the complexities of business strategy and consumer purchasing habits (Kumar & Reinartz, 2016). The best advertisers, according to Jasek et al. (2018), prioritize lifetime value over temporary metrics such as clicks, installs, and even revenues. These marketers, according to Belsky (2017), are at the top of the advertising food chain. As a result, rather than treating all customers the same, businesses must treat each customer differently based on their CLV. Another reason for the growing interest in CLV is its enormous potential for boosting marketing management decision-making and even company strategies. CLV can be used by businesses to create a wide range of marketing strategies for a wide range of user groups (Kitchens, Dobolyi, Li, & Abbasi, 2018). Capon and Go (2016), for example, provide marketing strategy frameworks and elaborate on six marketing requirements for the two major client categories. To determine whether new customers should be acquired, retrieved, or excluded, a decision should be made about the desired customer qualities and investments in prospective customers who will bring good CLV and are aligned with the business' market strategy. Customer churn prevention is critical for both attracting new customers and retaining existing ones.

To choose an appropriate, comprehensive strategy to CLV adoption and using this metric for the decision making, each organization must conduct a thorough assessment of these unique circumstances for the two primary customer groups. As a result of numerous studies focusing on calculating the CLV for a company's existing clients,

retail e-commerce companies now have the opportunity to use the CLV strategy as a competitive advantage. However, according to Wang, Liu, and Miao (2019) there are significant research gaps surrounding the CLV model calculation, particularly for new customers (first-time buyers). Khajvand, Zolfaghar, Ashoori, and Alizadeh (2011) states that a significant portion of the progress can be attributed to the expansion of the Recency, Frequency, and Monetary (RFM) model for existing customers. A review of the literature revealed that, the ability of different machine learning algorithms to anticipate CLV for both new and existing users has not been empirically compared in any study (Kitchens et al., 2018).

Distinct data and models are needed for these two consumer groups: Companies must first forecast the future worth of their existing users whose transaction histories have been gathered. To comprehend the general worth of a corporation, second, they should forecast the future value for new users who have just made their initial buy (Kitchens et al., 2018).

In addition to forecasting CLV for various consumer groups for multi-category e-commerce businesses, there are additional aspects that require additional research when calculating CLV.

The CLV of the company's clients should be determined prior to developing a plan tailored to a specific customer. A company can use the CLV to fund ongoing improvements to infrastructure and advertising campaigns. The CLV framework can be used by businesses to figure out: consumers that will generate the most revenue in the future, what factors increase CLV, and how much money should be allocated to each communication channel (Jasek et al., 2018). The CLV indicates the value of a particular customer to e-commerce retailers. According to Kumar (2016), CLV is not the only

element that should be considered when determining which consumers provide the greatest revenue for a firm. To identify the really valued consumers, businesses must consider a few key factors that differentiate each of them. According to Kitchens et al. (2018), in anticipating consumer behavior, there are insufficient studies that adopt a full and holistic relationship approach, as opposed to a transactional one. Most of the prior studies have mainly focused on Recency, Frequency, Monetary (RFM) based CLV predictions which has a lot of criticism on because it does not cover the CLV for first time purchasers. Also, numerous research has analyzed relationship-oriented constructs related with consumer behaviors and CLV using explanatory models, but few have used predictive models (Kitchens et al., 2018). Using relationship marketing theory (RMT) as a foundation, one of the purposes of this research is to examine the additional constructs influencing customer lifetime value (CLV).

CLV forecasting and evaluation necessitate the application of specific models and approaches. CLV models, according to Kumar and Reinartz, (2016) and Zhang, Bradlow, and Small (2015), come in a variety of sizes and shapes. Still, there are big research gaps on this topic when it comes to multi-category e-commerce businesses. Therefore, to fill all of these research gaps, we present a revolutionary 360° framework for forecasting of CLV and segmenting consumers that enables online businesses which sells multiple categories of products to control every aspect of CLV and get a competitive edge. The suggested holistic framework calls for the integration of a 360° holistic CLV prediction system with the company's decision-making system that prioritizes the strategic decision making of the company. We divide our proposed framework into two parts and concentrated on the approaches for both new and existing clients individually.

The framework's first section focuses on new customers. In this research, a new user is defined as a customer who makes one purchase from an e-commerce website during the observation period, while an existing user is someone who has made many transactions from the same website.

Because it employs frequency and recency to distinguish clients, the conventional family of probabilistic models is inapplicable to new users who have same purchase recency and frequency. Due to this, as the initial component of our new 360° framework, we are concentrating on discovering the optimal Machine Learning (ML) algorithm to forecast the CLV for new users who have made one transaction. As a result, our questions for new users are:

Which of the CLV prediction methods is most suitable at predicting new users of multi-category e-commerce companies?

How to leverage the CLV to enhance the company's decision-making strategy?

The second component of the proposed framework focuses on existing users in order to close the research gaps mentioned earlier. In addition to the RFM based studies that dominated CLV prediction in marketing literature, machine learning models are also focused mainly on existing users' CLV prediction. However, most of these studies ignore the features that were used as the input features to the models. Without taking into account any additional inputs to the analytics process, these studies advance knowledge on model selection elements of customer analytics (Kitchens et al., 2018). The proposed framework not only focuses on finding the best model to predict CLV for existing users but also includes the addition of new relationship-oriented constructs to improve the prediction accuracy of CLV for existing users and improve the capabilities of companies especially on strategic decision making.

Using our new 360° framework, we separate the second part of the framework into four parts for the existing users to address the following four questions:

Which model is preferable to predict CLV for existing users of multi-category e-commerce companies?

How predicting CLV combined with novel relationship-oriented factors improve the ability to segment customers and improve the accuracy of CLV prediction?

Which model is the most effective to use when predicting the newly introduced constructs in addition to CLV?

Which segment of existing users will have the highest value in the future and how these segments impact strategic decision making?

To assist multi-category e-commerce businesses in addressing the aforementioned research questions problems and developing a thorough strategy that will provide them a competitive edge and provide strategic decision making, a new 360° framework for predicting CLV and segmenting consumers is developed. Based on the implementations, here is a summary of our study's contributions:

In order to give multi-category e-commerce businesses a comprehensive understanding of the value of their overall consumers and not just the consumers who have already made multiple purchases from the same company, two different user types—new user and existing user— are investigated together for the first time. Several models are examined for both new and existing users, and the best models to forecast CLV for each different group has been presented.

Multi-category e-commerce enterprises may concentrate on developing strategies that will bring them more valued consumers by adding the two new relationship-based constructs (Distinct Product Category (DPC) and Trend in Amount Spend (TAS)) built

on relationship marketing theory to the model for prediction in addition to CLV for the existing customers. The findings show that the newly suggested variables improve the prediction accuracy and they are valuable constructs. In addition, it showed that CLV should not be the only factor used to gauge a customer's value. It is crucial to forecast CLV, DPC, and TAS combined in order to identify the real high-value clients since doing so improves the results' precision.

The multi-output DNN model is proposed and evaluated to benefit from a more resilient and simpler to maintain model and better at capturing the interactions between the three outputs, CLV, DPC, and TAS, than building a single-output model for each output independently. We showed that the multi-output DNN model is superior than the single-output DNN model, as well as the multi-output Decision Tree and multi-output Random Forest models, on all of the assessment criteria and in the identification of high-value clients making it more suitable and dependable for multi-category e-commerce businesses.

The multi-output DNN model results are interpreted when predicting CLV for the first time using Explainable Artificial Intelligence (XAI), which increase user trust in DNN models that were previously thought of as "black-box" models, by explaining the feature significance details.

Rather than relying just on CLV, client segmentation is carried out based on the three outputs. Multicategory e-commerce enterprises may utilize segmentation based on the combined findings of the three outcomes as a valuable guide to help them concentrate on the correct clients and define their strategy appropriately.

Finally, the primary contribution of this study is the proposal for the holistic use of CLV as a financial indicator in strategic decision making to provide competitive advantage.

The remainder of this thesis is divided into the following chapters. A sampling of the literature on this subject is included in Chapter 2. Chapter 3 explains the problem definition for both user types and explains the details of the research questions. Chapter 4 discusses the data utilized in this study for both new and existing user types, pre-processing of the data that is used and the performance measures. In Chapter 5 and 6, the models' results are discussed. The findings of the models are described in Chapters 5 and 6. Finally, in Chapter 7, the research discussion and conclusions, limitations, and future study recommendations are presented.

CHAPTER 2

LITERATURE REVIEW

This chapter conducts a thorough evaluation of the available literature to address the research problems outlined in the preceding chapter. We first concentrate on the marketing component of the predicted Customer Lifetime Value (CLV). Second, we lay the theoretical groundwork for the different types of consumers: new and existing users. Third, we provide an overview of the existing user models, Probabilistic methods in the literature. Fourth, we summarize the Machine Learning (ML) models with a special emphasis on multi-output Deep Neural Network (DNN). Finally, we added the contribution of this study to the existing literature.

Customers are the focus of a business' marketing strategies because they generate revenue and increase the business' market worth. Customer relationship management (CRM), according to Kotler and Keller (2015), is an example of how marketing places a strong emphasis on the connections between all processes and activities involved in the creation, dissemination, and providing of value to customers.

Ryals and Knox (2007) states that the major goals of CRM operations are to retain existing clients, forge enduring relationships with them, and bring in new clients. CLV calculation and application are important in this process since the calculation help in categorizing consumers and determine which ones offer the greatest long-term value (Kim, Jung, Suh, & Hwang 2006). The strategic problem of understanding and maintaining customer relationships has become more challenging and more crucial as businesses increasingly fight for consumers' attention, rising the cost of acquisition and increasing the difficulty of retention. At the same time, it has gotten simpler to gather

enormous volumes of client data that, when paired with ever-sophisticated analytical approaches, may yield crucial and useful insights, enabling businesses to innovate and set themselves apart from rivals. However, because of the widespread use of analytics, many of the opportunities are already taken. As a result, isolated, ad hoc analytics projects are unable to establish or maintain a competitive edge. Companies that want to leverage big data to achieve strategic value and decision-making processes must have targeted analytics capabilities that enable agility in adjusting to a changing environment.

According to Eid et al. (2021) CLV is a crucial part of CRM and one of the key factors considered for decision making. With the rapidly increasing amount in volume and variety of the data, companies have big data analytics with CRM systems also needed attention to have a better decision-making process. There have been calls for research of big data analytics' business value and role in supporting more effective decisions (Abbasi, Sarker, & Chiang, 2016; Kitchens et al., 2018). To obtain a competitive advantage strategic decision making is important. Because of this, businesses must use a holistic CLV approach and effectively segment their consumers (Venkatakrishna, Mishra, & Tiwari, 2021). We provide a framework for putting sophisticated customer analytics solutions into practice to predict a holistic CLV for the overall company customers, complete with a variety of constructs for describing and forecasting CLV and help companies to have a better customer relationship with an enhanced decision-making system.

Customers are increasingly viewed as assets whose value should be evaluated and controlled by both academics and marketing practitioners, according to Venkatakrishna et al. (2021). As a result, companies are increasingly moving away from a product-centric strategy toward a customer-centric one (Heldt, Silveira, & Luce 2021). As the

market and customer behaviour change, a customer-centred approach is more likely to help a business grow (Heldt et al., 2021). Customers are being prioritized more, therefore segmenting them is a logical step in creating a marketing strategy that works (Wong & Wei, 2018). The client is the focal point of all business strategies in a customer-centric organization. They are aware of the market segments where their products are offered for sale. In order to improve the customer experience, customer-centric organizations define the full customer journey in addition to focusing on the solution they offer to the consumer. They Companies will concentrate on boosting customer loyalty, which will lead to business expansion. Building customer analytics solutions by focusing on the company's most valuable assets which are customers, offers a major strategic initiative with the opportunity to provide a substantial and long-term competitive advantage. There are chances to establish projects encompassing acquisition, retention, and growth in order to maximize customer lifetime value and equity. A unified framework for creating consumer analytics solutions that takes into business and IT strategy and decision-making account is required to make this a reality. With this, businesses might construct a value-justified infrastructure to support a portfolio of sophisticated consumer analytics capabilities that, when combined, provide substantial strategic value and a sustained competitive advantage (Kitchens et al., 2018). As a result, companies must focus on building solutions that will help them with strategic decision making and making each customer profitable in order to increase CLV and the company’s worth.

2.1 Customer lifetime value (CLV)

CLV is a metric to value the customers at the individual level that considers a client's whole lifetime profit contribution. The present value of a specific customer's projected net cash flows is known as CLV (Fader, 2012a; Pfeifer, Haskins, & Conroy, 2005). As stated in the definition above, CLV determines the customer's net worth. Because CLV is quantified at the individual basis, businesses that have calculated it may study the distribution of their consumers based on the potential value that they will bring. From a business perspective, estimating the CLV has the advantage of giving the manager information about the customer's potential for future profits. Knowing the prospective profit of the customers' that they will bring in the future helps managers target their marketing efforts effectively at the right clients at the right time (Venkatesan & Kumar, 2004). According to Gupta and Zeithaml (2006) advertising methods that were reliant on client values, like CLV, enhanced the organization's financial performance. Businesses need to estimate CLV in order to allocate resources effectively, but each customer-company relationship requires a different strategy for doing so. Differentiating between different types of consumer and company interactions is crucial (Zhang et al., 2015). These interactions could be contractual or non-contractual, according to Bernat (2018). Based on the CLV, researchers have proposed several strategies (client acquisition, retention, etc.) that have been put into practice in a variety of industries where there is a contractual relationship, including banking, airlines, and phone companies. Their investigation reveals that contracts establish a legal connection between the parties (e.g., subscription or membership). In that case, the company will be informed when a consumer stop being active. On the other hand, there are no formal ties between the firm and the client in non-contractual situations, and the business loses track of inactive

customers (e.g., online store purchases). Additionally, purchases can be categorized as discrete or continuing. When compared to discrete-time purchases (like fundraising efforts), continuous-time purchases can be made whenever it is convenient. It is completely up to customers to decide whether to make a purchase from a specific store if they are not contractually required to do so. The term "transactional business model" also applies to this arrangement. Additionally, according to McCarthy, Fader, and Hardie (2017) there should be no contracts with buyer. In a contractual setting, the moment at which a client's connection with the company ceased may be tracked. The moment a client "dies" is not taken into consideration in non-contractual situations. Differentiating between customers who have stopped engaging with the company and those who are merely going through a prolonged period without transactions is challenging in non-contractual contexts (McCarthy et al., 2017). The many kinds of client and corporate relationships must therefore be defined, and a CLV prediction must be made for each one independently (Jasek et al., 2018). Because our data came from a non-contractual, continuous time multi-category e-commerce firm, this thesis only examined CLV models that fit under the proposed 360° framework.

2.1.1 Important CLV features

According to Jasek et al. (2018), traditional practices in relationship management, like brand equity, transactions, and product centricity, have changed to a customer-centric strategy. The consumer is the most important asset of the organization as a result of the recent developments in technology and online shopping (Jasek et al., 2018). Companies have grown more customer-centric, according to Heldt et al. (2021), with the addition of a customer viewpoint in the analysis of projected revenues, which were previously

estimated solely based on anticipated product sales. Even while a client-centric strategy is essential, product quality and variety still important for companies to satisfy consumer demands. Both perspectives are crucial, according to Leone et al. (2006), because branded products are promoted to customers and bought by them. Additionally, the majority of managers want to evaluate and base decisions on consumer perceptions as well as product and brand perceptions (Heldt et al., 2021).

This shows that combined customer and product evaluations would result in better outcomes than each analysis carried out separately.

A company's strategy can be combined with products and clients via cross-selling strategy. A common corporate strategy for boosting customer equity is cross selling (Kitchens et al., 2018). Offering clients more products or product categories, expensive (better) products, or a larger amount of the purchased product increases sales because this practice increases consumer purchase and spending (Villanueva & Hanssens, 2007). When customers buy products from several product categories, they raise their expenditure and CLV proportionately, and they also become more brand loyal (Kitchens et al., 2018). When there is the option of purchasing products from many categories in a single store, customers are less likely to go to other stores since they can obtain everything they need on a single website (González-Benito & Martos-Partal, 2012). Also, the switching costs of finding a new store are high. As a result, when estimating CLV, especially for multi-category e-commerce enterprises, product-related information is crucial.

To obtain a strategic decision-making system, businesses should understand the customers' characteristics that best illustrate why some are more valuable than the others and important factors that lead to make them more valuable. To understand these factors

and to guide the adoption of new constructs in addition to RFM variables Kitchens et al. (2018), designed a framework where they included a group of constructs based on relationship marketing theory to predict the customer behaviour and CLV.

The phrase "relationship marketing" originally debuted in 1983, and most writers adhere to this definition, which defines relationship marketing as "attracting, sustaining, and strengthening client connections in firms" (Perišić Prodan, Cerović , & Ivančić, 2022). Although there are several different definitions present in the literature the majority of definitions of relationship marketing emphasize the significance of creating long-term client connections. Kitchens et al. (2018), created a list of construct categories based on relationship marketing theory to understand the important factors that will create long term client connections which provides the company with a source of enduring competitive advantage and strategic value. The authors briefly discuss each of the detected constructs and their inclusion justifications as in below.

Transactions: Transaction variables are the most fundamental customer relationship construct. As noted, the most prevalent methods for forecasting consumer behaviour depend primarily on RFM variables. In addition to RFM, transaction variables may also include discount levels, promotions, and products purchased etc.. all of which may give important information on customer value and behaviours (Kitchens et al., 2018).

Demographics: The most fundamental qualities regarding clients that businesses can assess are their demographics. Studies on relationship marketing have indicated that consumer demographics have a direct impact on their baseline expectations and thresholds for other variables like satisfaction, which in turn affects their repurchase motivation and behaviour (Kitchens et al., 2018).

Engagement: Beyond just making purchases, customers interact with businesses in many other ways. Conclusions regarding potential behaviour may be made based on the nature of interaction. These interactions might include going to the company website, sending an email, contacting customer care, opening company's app, or writing an online review. Engagement often signifies a customer's desire to keep doing business with a certain organization (Kitchens et al., 2018).

Satisfaction: An essential component of RMT is customer happiness. Previous research have shown that raising customer satisfaction increases customer lifetime value and improves client retention. The main issue with customer satisfaction and associated perceptions is that it might be challenging to quantify them for consumers individually (Kitchens et al., 2018).

Choice: The choice of the product bought by a consumer is a crucial observation that may be obtained from both their first and future transactions. Identifying the product category, a client wants may be valuable in multiple ways especially when a company sells products within multiple categories. From an RMT viewpoint, that the categories of products a consumer purchases might indicate their degree of trust and interest in a given company. Additionally cross-buying increases switching prices as customers grow more knowledgeable about a company's goods and quality. Purchasing from multiple product categories may also provide insight into what items a client may be interested in buying in the future, and product variety is a crucial aspect of customer retention and value (Kitchens et al., 2018).

Channel: The channel via which the client obtains goods or services and continues to connect with the firm provides crucial information about the consumer and the nature of the relationship. Various channels may attract distinct client segments. Studies have

indicated, for instance, that digital channel consumers tend to be more loyal and engaged and they have more chances to develop relationships with the brand (Kitchens et al., 2018).

Messaging: From an RMT standpoint, the messages the client gets from the business may also have a significant influence on subsequent behavior (Kitchens et al., 2018).

Firm and Environment Characteristics: The foundation of a customer's connection with a business is the firm's qualities and the market it serves. This architecture includes several elements that are crucial for consumer analytics (Kitchens et al., 2018).

Leveraging all or some of these constructs may be used to build customer analytics solutions to strengthen customer relationships, increasing customer equity and lifetime value, adding strategic value for the business, and fostering sustainable competitive advantage (Kitchens et al., 2018).

Similar to the constructs above there are other studies that mentions the need of additional factors to take into account while predicting CLV. According to Kumar, Ramani, and Bohling (2004), a number of factors can have an impact on this, and they considered three factors—the frequency of cross-buying by customers, the interval between transactions, and the quantity of product returns—as all worthwhile to investigate. Cross-buying is a significant aspect in CLV prediction, and the number of distinct product categories (DPC) that customers choose to purchase from represents that level of cross-buying (Yılmaz Benk, Badur, & Mardikyan, 2022). According to Kumar et al. (2004), a study by catalog retailers, showed that customers increased their CLV by 5% on average for each additional product category they bought. Additionally, an inverted U-shaped association between the interval between purchases and the lifespan of earnings was found by Kumar et al. in 2004. As an example, opposite to other

consumers, those who make repeated purchases with a modest but consistent interval may see greater profitability over time.

According to Borle, Singh, and Jain (2008), there are a number of dependent factors linked to CLV. The authors found a strong positive relationship between inter-purchase intervals and spend, suggesting that consumers who postpone purchases compensate by spending more when they do so. The duration, spend, and likelihood of customers churn were found by the authors to be interdependent, which supported their combined modeling approach. In designing our framework, we, similar to Kitchens et al. (2018), largely refer to relationship marketing theory especially while predicting the CLV for existing users. We used most of the constructs that was suggested I their study. In addition, by using the suggestions of the other studies that was discussed above we enhanced the RMT based constructs, transactions and choice, by adding two new variables as distinct product category (DPC) and the trend of the amount spent (TAS). Transactions is one of the most fundamental constructs for RMT and most of the studies focused on RFM variables that can be extracted from past purchases. Based on the data in this study, new TAS variable extracted from multiple transactions of each user as an extension of Transaction construct. Similarly, Choice is another important construct in RMT where the choice of the product purchased by a consumer is a crucial observation that may be obtained from their first and future transactions (Kitchens et al., 2018). DPC extracted from the data as an extension of Choice construct. Many CLV-related studies have neglected to take these two new variables into consideration due to the difficulty of integration, even though trend in the amount spent within each successive transaction and the product type offered are crucial components when defining high-value consumers (Villanueva & Hanssens, 2007; Kitchens et al., 2018).

2.1.2 Customer equity

The influence that CLV has on corporate valuation, which influences investor choices, is another crucial aspect of CLV. The Customer Equity (CE) approach, which is closely related to the CLV, is largely used to assess a firm's long-term worth. According to Fader (2012a) and Abdolvand, Albadvi, and Koosha (2014), CE is the total CLV of all present and prospective clients (Blattberg & Deighton, 1996; Gupta, Lehmann, & Stuart, 2004; Jasek et al., 2018; Rust, Lemon, & Zeithaml, 2004; Zhang, Dixit, & Friedmann, 2010). Companies have massive financial advantages when leveraging CLV combined with customer equity (Bolton, Lemon, & Verhoef, 2004; Kumar & Shah, 2009).

Based on the requirements for the data source, derived metrics, fundamental assumptions, and, most crucially, the level of aggregation needed, Kumar and George (2007) distinguish two methods for computing CE. Both of these methods produce several ways to increase customer value: The first is a non-aggregated approach that maximizes CLV by putting individual customer-specific tactics into practice, like the best asset allocation, purchase analysis, and controlling investment ratios between acquiring and retaining clients. The second is an aggregated view that seeks to enhance the elements that maximize CE (Lemon, Rust, & Zeithaml, 2001). As the customer-centric strategy spreads, individual level CLV prediction is becoming more important for determining a company's total value (Heldt et al., 2021). CLV must be calculated holistically for each consumer of the company in order to value the firm, which is dependent on the relationship between CLV and CE.

It should be underlined that prior CLV implementations in the retail industry did not prioritize e-commerce businesses that offer products across a variety of categories or a comprehensive strategy to anticipate CLV for both new and existing users. Applying the

CLV framework to both of the user types and to the multi-category e-commerce companies represents several practical challenges, most important being that; two different types of users; new and existing users require different data and modeling techniques to predict the CLV.

2.2 Existing vs new users

There are not many methods in the literature for determining the best CLV model for a business based on its new and existing users. A new client is defined in this research as an online shopper who makes only one purchase from an e-commerce website, whereas an existing user is an online shopper who makes many purchases from the same website. More specifically, this study distinguishes between a first-time buyer who made only one transaction and a repeat customer who is devoted to a website and makes regular purchases there.

These two categories of consumers require different types of data and modeling: First one is predicting future value for existing users based on transaction history. Since there is more than one transaction for existing users transactions data may include lots of useful details that can be used with the modeling. The most common variables that can be extracted from multiple transactions that existing users have are the RFM variables. In addition to RFM, products and product categories purchased, the total engagement variables, the trend of the spend etc. are variables that may also provide valuable information. The second is a value prediction for new users who have just completed their initial purchase. Since they have only one data point the variables that can be used for the modeling are limited compared to existing users. While majority of the customers are first-time buyers prior studies have largely focused on RFM which is applicable to

existing users and are often not taken new users into consideration when predicting the CLV (Kitchens et al., 2018). Previous studies mostly focused on predicting the CLV for existing users who has repeated transactions (e.g., Chen, 2018; Fader, Hardie, & Lee, 2005a; Gupta, Hanssens, Hardie, Kahn, Kumar, Lin, Ravishanker & Sriram, 2006; Jasek et al., 2018; Schmittlein, Morrison, & Colombo, 1987; Singh, Singh & Gupta, 2020). Because of insufficient data, early models were frequently confined to adopting fundamental parametric statistical models, such as Morrison and David's (1988) negative binomial distribution (NBD) model and the Pareto/NBD model. Traditional RFM models like the Pareto/NBD model, which mainly considers the quantity of purchases made over the course of a lifetime, are frequently used to calculate CLV (Fader et al., 2005a; Fader, Hardie, & Lee, 2005b; Fader, Hardie, & Shang, 2010; Schmittlein et al., 1987).

CLV prediction has traditionally relied heavily on probability models; but, as technology progresses, large-scale e-commerce platforms emerge, and companies begin to gather and retain more data, machine learning (ML) professionals have become more interested. Recent studies show that ML-based CLV models are more effective than probability models (Chamberlain, Cardoso, Liu, Pagliari, & Deisenroth, 2017; Chen, 2018; Salehinejad & Rahnamayan, 2016; Singh et al., 2020; Vanderveld, Pandey, Han, & Parekh, 2016). ML models have the capacity to incorporate more attributes than conventional probability models, which could result in more accurate and better results. For existing users, Gupta et al. (2006) provides a thorough evaluation of CLV techniques. They demonstrate that machine learning techniques like random forest (RF) perform better than the popular models like RFM and Buy Till You Die (BTYD) because ML models can have a variety of additional features.

Creating separate regression models for the average order value (or profit) and purchase frequency of existing users, and then incorporating both metrics into a CLV forecasting model, is another potential two phase approach (Venkatesan & Kumar, 2004). Both the RFM and BTYD frameworks employ this technique. Fader et al. (2005b) makes the assumption that recency and frequency follow an independent Gamma/Gamma distribution, while purchase values follow a Pareto/Negative Binomial Distribution (Pareto/NBD). In contrast, the breakdown is based on the questionable premise that purchase order value is unrelated to the frequency of purchases. As an example, repeat customers can spend less money on each purchase.

CLV prediction of new users, in addition to existing ones, is vital to the advertising industry. However, there are not many studies that look into new customer CLV prediction, which causes businesses to underestimate the full value of their entire customer base and business (Kitchens et al., 2018). Consumers who have made just one purchase from a company are considered new customers. Since the RFM and BTYD model families distinguish clients based on frequency and recency, they do not apply to new users. On the other hand, new customers make the same number of purchases and because of this, predictive signals must be taken from other sources, such as the customer behaviors gathered during registration or sign-up, or the specific service or product type of a new customer's initial transaction. Therefore, the CLV of the new users must be modeled separately.

Customers with a zero CLV are individuals who are new to the company and have only made one purchase. Existing clients with a positive CLV have a wide range of values. This unusual CLV distribution is noted by Chamberlain et al. (2017). The authors deal with the issue by modeling the CLV's percentile rank and then mapping it back to actual

values for usage in downstream procedures A related problem is discussed by Sifa, Runge, Bauckhage, and Klapper (2018) in the context of CLV forecasts for free-to-play video game consumers. Only a tiny fraction of consumers makes repeat purchases, accounting for the vast bulk of income. To improve prediction performance, the authors recommend training a DNN with synthetic minority oversampling (SMOTE; Chawla, Bowyer, Hall, & Kegelmeyer, 2002). SMOTE is an algorithm performing data augmentation that develops synthetic minority entities during the model training phase. This helps in regulation of learning structures and prediction models that represent minority classes.

According to Chamberlain et al. (2017), the key data difficulties that underpin the peculiar distribution of the CLV problem are that a big portion of consumers are one-time buyers who never make another purchase, resulting in a huge number of zero value labels. To address this issue, Vanderveld et al. (2016) and Chamberlain et al. (2017) employ a two-phase hybrid RF model to forecast the CLV of e-commerce visitors. The first phase predicts the user's purchase propensity, which is a binary categorization of whether the user is likely to make a purchase within the specified time period. In the second phase, the monetary value is projected for consumers who were expected to make a purchase in the first stage.

The two-stage strategy provides insights into the multiple elements that influence CLV and is an acceptable method for estimating CLV. However, the increased difficulty of maintaining two models is the most significant disadvantage. As a result, Wang et al. (2019) employed a Deep Neural Network with the ZILN loss function to discover a solution for first-time purchasers with zero LTV values.

Wang et al. (2019) employ supervised regression to forecast CLV of new users. In supervised regression, unlike the BTYD model family, all customer-level features are incorporated. It aims to lower the stated forecast inaccuracy rather than attempting to model the underlying dynamics of recurring business or client attrition. Many common machine learning approaches, including; deep neural networks, support vector machines, random forests, gradient boosting, and logistic regression, are conveniently applicable for this regression problem.

According to Chamberlain et al. (2017), a DNN with enough hidden units can outperform Random Forest in forecasting CLV for new users. The author also demonstrates how the wide-and-deep (Cheng et al., 2016) model produces more exact estimates of customer attrition by combining the benefits of a broad linear model (for memory) and a deep neural network improves performance (for generalization).

The Wide & Deep concept is commonly used in real-world industrial situations. It is divided into two parts: A deep model that discovers nonlinear correlations between features automatically and is equivalent to the Base Model. A broad model that handles cross-product features offered by the user. Wide & Deep requires engineering expertise from the “wide” module. The wide-and-deep model’s cross features are supplied into a linear model that is jointly trained with a DNN model. It works effectively when there are several categorical features, and it is especially useful when there are no history purchase records for new users (Wang et al., 2019).

In summary, CLV for both consumer types necessitates multiple modeling methodologies in order to calculate the total value of the customers and the value of the organization.

2.3 Probability models

Probability models are well-known for anticipating CLV. They are most commonly employed for predicting CLV for existing users since they require past purchase data. These models calculate the likelihood that a consumer will remain active in the following period and anticipate the amount of purchases made by the customer in the following period by using the prior purchasing history of the customers (Jasek et al., 2018). It is critical to note that the retention rate shows the likelihood of a client being active over the following term. Moreover, because it takes the condition of remaining active into consideration, future transaction volume might be seen as a blend of profit and client retention. However, because it is impossible to forecast how much a consumer would spend on each purchase, the total quantity of upcoming purchases does not equal CLV. To solve this, Schmittlein and Peterson developed a submodel in 1994 to anticipate the average future purchase value per client. By multiplying the two factors and making the assumption that the average transaction value and the quantity of transactions are unrelated, it is feasible to forecast a client's total future expenditure. They claim that transaction value can be represented by a normal distribution was debunked by Fader et al. (2005b). Fader et al. (2005b) stated that the organization of transaction values is very distorted and cannot be characterized by a normal distribution. First probability model for customer base analysis was the Negative Binomial Distribution (NBD; Ehrenberg, 1959). Clients, according to this theory, make random purchases at a time-invariant, customer-specific purchase rate. These two assumptions are covered by NBD, which is a Poisson distribution with a gamma-distributed purchase rate. However, the time-invariant purchase rates assumption is rarely correct.

As a result, Schmittlein et al. (1987) created the Pareto/NBD model, that improves upon the earlier model by integrating time-varying purchasing rates.

Because the Pareto/NBD was presented as a generalist Buy Till You Die (BTYD) paradigm which might potentially be employed by any business, we chose it for our investigation. According to Bernat (2018), customers are assumed to be active and continue with purchases for an undefined time before they stopped purchasing and become inctive in the Pareto/NBD model. The NBD model tracks a consumer's purchasing characteristics while they are still live, while the Pareto distribution tracks the time a customer is still alive. According to the Pareto/NBD model, the only data needed to predict a customer's future purchase volume are recency, frequency, and duration of observation. As a consequence, it is not necessary to provide further details, like the precise dates of the purchases. According to Fader and Hardie (2009), despite the model's restricted use of purchase data, it has strong predictive performance.

Glady, Baesens, and Croux (2009), changed the Pareto/NBD model by loosening the notion of the independence between the transaction value and the number of transactions and demonstrated that a relationship between these figures can be used to improve CLV predictions. Fader and Hardie (2009) and Abe (2009), added the variables that are not changing in time into the model, proposed additional modifications to the Pareto/NBD model, which can be especially useful for large data sets with a wide range of customer and purchase characteristics. In conclusion, despite worries about using only a few variables, Pareto/NBD is a popular strategy among probability models for forecasting CLV.

2.4 Machine learning models

Unlike probability models, machine learning (ML) techniques can estimate CLV for both new and existing users. Probability models have been widely used for CLV prediction, but as technology progresses, large-scale e-commerce platforms emerge, and businesses begin collecting and keeping more data, these methods become less successful. ML models offer the advantage of employing more features than probabilistic models, as well as not relying on assumptions about the relationships between underlying data. Machine learning (ML) models, in contrast to classical probability models, can contain a huge number of elements or variables, potentially giving better outcomes with more precision (Bernat, 2018; Chamberlain et al., 2017; Chen, 2018; Gupta et al., 2006; Vanderveld et al., 2016). For example, the CLV of each e-commerce site user was estimated using a two-stage random forest model (Chamberlain et al., 2017; Vanderveld et al., 2016). As an example of ML model comparison Sifa et al. (2018) contrasted the Random Forest (DF), Decision Tree (DT), and Deep Neural Network (DNN) techniques which influenced our study as well. To forecast CLV, Malthouse and Blattberg (2005) used a neural network (NN). Moro, Cortez, and Rita (2015) employed a NN, support vector machine (SVM), logistic regression, and decision tree to forecast the effectiveness of a company's marketing efforts. They discovered that the NN outperformed the other strategies. Wang et al. (2019) used a DNN with a revised loss function, laying claims that it outperformed the previous models. Machine learning algorithms, in general, are good forecasters of the future and universal approximators. The fundamental advantage of these strategies is their flexibility since they do not assume the relationships in the data. However, there are also challenges that are associated with these algorithms which include the frequent

need for considerable hyperparameter tuning, the possibility of overfitting, and their high computational cost.

Despite the fact that ML models are routinely used to combine predictions from other models in order to anticipate CLV, few studies have looked at ML models for directly modeling CLV. The combined forecasts can be used as feature vectors in random forests or neural networks to provide better predictions than individual models.

Vanderveld et al. (2016) implemented the RF in order to discover changes in CLV. According to Singh et al. (2020), Salehinejad and Rahnamayan (2016) developed a model for forecasting consumer shopping behavior utilizing recurrent neural networks (RNNs) and a variety of characteristics such as client loyalty number (CLN), recency, frequency, and monetary values. According to Chen (2018), airlines may classify passengers into high-, medium-, and low-value categories based on their customer lifetime value. Chen (2018), as stated in Singh et al. (2020), employed the XGBoost ML approach to determine the lifetime value of airline passengers in order to better coordinate resource allocation. This resulted in lower resource investments and increased revenue.

Bernat (2018) provided many CLV prediction models. The authors used the gradient tree boosting machine learning method and stated that models in computer science have long been thought to be reliable substitutes for traditional models. All of the ML comparison studies that was discussed above provided us information about the data that was used, best performing models, hyperparameter tuning methods that was leveraged and performance metrics that was used in these studies where we can also leverage in our studies.

The majority of traditional machine learning algorithms, according to Sajjad et al. (2020), are intended for single-output prediction. Predicting many outcomes takes time because each result requires a new training process, and the models perform poorly. Because of the high non-linear relationship between their input and output, predictability, adaptability, and robustness of these approaches can all be improved by leveraging other methods.

One method is using a simple technique, building one neural network model that may predict many outcomes at the same time from the same input data. These models are easier to develop and analyze than dealing with a single-output model for each output attribute independently. To solve these issues, the authors of Sajjad et al., (2020) developed a multi-output deep neural network (DNN) technique that effectively forecasted their two output variables simultaneously. They conducted various experiments that compared single-output ML models to multi-output ML models, and they stated that multi-output model was better than all of the single-output models. Multi-output DNN models were evaluated in chemometrics and found to outperform single-output ML models (Sajjad et al., 2020). According to research published by Mishra and Passos (2022) multi-output deep learning models outperform standard single-output deep learning models. Deep learning algorithms, according to their findings, can do multi-output modeling to forecast data on the physical structure of the fruit by setting the model's output layer to an appropriate number of answers, in this case two. When predicting two independent outcomes in pear fruit, the results showed that the DL model had a smaller root-mean-square error (RMSE) than the standard partial least-squares model (Mishra & Passos, 2022). In summary, similar to the above studies multi-output DNN prediction might be a better model for predicting CLV in conjunction

with other output variables and segmenting clients not only by CLV but also by DPC and TAS.

Explainable Artificial Intelligence (XAI)

Although DNNs have become a popular solution they are black-box models. Users are not aware of the internal dynamics and the main logic of these models, which is a major challenge because it makes understanding how a system thinks and behaves difficult (Muddamsetty, Jahromi, Ciontos, Fenoy, & Moeslund, 2022). The fundamental goals of the artificial intelligence research field known as Explainable Artificial Intelligence (XAI) is to understand ML models and interpret their findings. A variety of ways have been offered to accomplish this. However, with the exception of Lundberg and Lee's (2017) SHAP technique, many of them have a basic flaw: they lack theoretical grounding. SHAP, according to Venkatakrishna et al. (2021), establishes Shapley numbers for the expectation function of initial model, combining the benefits of LIME (Ribeiro, Singh, & Guestrin, 2018) with Shapley values (Shapley, 1953). Game theory backs up the Shapley values. When a trait is introduced to the "coalition," its average contribution to variance is utilized to determine its relevance. Janssens, Bogaert, Bagué, and Poel (2022) state that the (absolute) SHAP values are averaged across all observations to determine a variable's relevance. XAI for DNNs is critical for recognizing the relevance of marketing variables related to CLV prediction in order to provide businesses with a more precise and clear perspective of the model they are utilizing.

Customer Segmentation based on Multiple Outputs

Finally, by correctly segmenting their customer base, e-commerce management teams can develop tailored consumer strategies. Clustering is a technique for organizing

similar objects into tangible or symbolic groups. The K-means approach was developed by MacQueen in 1967 and has since been implemented in a variety of concepts, such as statistical data analysis, data mining, and other business strategies (Cheng & Chen, 2009). Customer segmentation is one of the most common applications of the k-means technique in the literature (Wu, Shi, Lin, Tsai, Li, Yang & Xu 2020).

K-means clustering is a popular strategy for creating and implementing successful marketing programs for valuable clients (Mesforoush & Tarokh, 2013; Wu et al., 2020). Cheng and Chen (2009) used an RFM model with the K-means algorithm to manage customer connections, and the findings of their customer value study were positive. According to Khalili-Damghani, Abdi, and Abolmakarem (2018), an integrated soft-computing method based on clustering, rule extraction, and decision tree methodology was presented to categorize potential clients for customer-centric organizations (Wu et al., 2020).

Management can design customised customer strategies that assist organizations estimate the worth of each unique client by categorizing them. For a long period, CLV-based segmentation has been studied (Benoit & Van den Poel, 2009; Kim et al., 2006; Kumar, 2006). However, for multi-category e-commerce businesses, additional parameters must be considered such as DPC and TAS in addition to CLV while segmenting consumers (Yılmaz Benk et al., 2022).

Businesses can increase their income by focusing their marketing efforts on their most profitable consumers. Furthermore, by identifying less profitable consumers, firms might reassess how they now sell to them, so enhancing their satisfaction and, ultimately, revenues.

Clustering, in a nutshell, is an effective approach for segmenting a company's consumer base. As a result, management of multi-category e-commerce enterprises should employ clustering to group their clients based on a variety of outputs.

2.5 This study

The following is our contribution to the literature using the 360° framework we propose: A holistic framework to predict the CLV for all users of the companies to provide the companies with a strategic decision making and competitive advantage was introduced. Users were divided into two as new users (users purchases just once) and existing users and different strategies have been used and compared to find the best approach to predict the overall customer value and customer equity for the company.

This study includes distinct constructs that have rarely been seen combined in earlier research. Building on relationship marketing theory, two additional constructs, DPC and TAS, were introduced to the literature as an enhancement of the theory. These two variables are incorporated to the model for prediction in addition to CLV, enabling multi-category e-commerce enterprises to concentrate on tactics that would deliver them more value clients for the existing users with numerous purchase records. After the evaluations it was clear that it is necessary to estimate CLV, DPC, and TAS together in order to grasp the real high-value consumers since doing so improves the findings' accuracy. CLV should not be the sole signal used to determine a customer's worth.

The multi-output DNN model is proposed and tested in comparison to building a single-output model for each output independently or against the widely utilized DT and RF algorithms in order to benefit from a more robust model that is simple to maintain and superior in terms of catching the associations among the three outputs, CLV, DPC, and

TAS. CLV, DPC and TAS have suggested as important components of CRM that are used to evaluate the decision making

The multi-output DNN model results were interpreted using XAI for the first time to our knowledge when predicting CLV in order to explain the important features and boost user trust in DNN models, which were previously thought of as black-box models.

Finally, rather than relying just on CLV, client segmentation is carried out based on the three output attributes. Multi-category e-commerce enterprises may utilize segmentation based on the aggregated findings of the three outcomes as a valuable guide to help them concentrate on the correct clients and define their strategy appropriately.

In sum, this study is the first attempt to enhance CLV prediction for the overall multi-category e-commerce company which investigates CLV prediction as a 360° holistic approach. Our findings have management implications for marketers doing consumer analytics and IT managers tasked with supporting projects using big data analytics. We give an all-encompassing platform for building sophisticated consumer analytics with a focus on customer value. Our generally applicable framework adds two new innovative constructs and compares several models for determining the worth of each, therefore facilitating data management choices. Our study also adds to the literature on the use of big data for predictive analytics. Our research concludes by contributing to the literature on providing a holistic CLV prediction framework, with a focus on the importance of segmentation of customers for strategic decision making.

CHAPTER 3

PROBLEM DEFINITION AND MODELS

This chapter starts with a brief description of the two important elements of our study which are different user types: New and Existing users. We describe the two different user types and the importance of CLV prediction separately for each of these user types. The following section is where we outline the model details that have been used in this study for each user type.

Predicting the CLV for all their users is very important for businesses to gain a holistic view for the future and to calculate the overall value of their companies. There are two types of users based on their purchasing behavior which require different data and modeling to predict the CLV. The first one is existing users who have multiple transactions. The other one is new users who just made one purchase during the observation period. Literature shows very few approaches of selecting the most appropriate CLV model for a company based on their new and existing users holistically and this is one of the goals of this study to provide a 360° approach to predict the CLV for every user.

3.1 CLV prediction for new users

Individual-level data such as CLV can help with decisions concerning cost-cutting, revenue and profit gains, return on investment, realigning marketing efforts, client acquisition and retention (Kumar & Reinartz, 2016). Furthermore, according to Malthouse (2013), CLV is the most important basis for assigning sales resources. It is critical to understand and apply the CLV when allocating marketing resources in order

to satisfy customers, create trusted connections, and convert one-time purchasers into repeat customers. In today's highly competitive and global economy, the average consumer has numerous options, thus exceeding customer expectations is critical to a company's success. CLV also has great potential in improving marketing management decision making which makes it very important for businesses (Abdolvand et al., 2014). Given that businesses only have a limited amount of resources, it's critical to attract and keep customers with high CLV from all categories of customers, both new and existing. The majority of literature-based research has focused on existing users' CLV, whereas new user CLV has received significantly less attention and requires additional investigation.

Because it employs frequency and recency to distinguish clients, the conventional probabilistic models cannot be used to predict CLV for new users who have the same frequency and recentness of purchases. Predictive signals must come from somewhere else and be modeled separately. Website activity, customer characteristics collected during client registration or sign-up, or the service or product type of the first transaction for new users are examples of these sources. Forecasting aggregate measures for financial planning is relatively simple for new users. Estimating the CLV of a single consumer, on the other hand, is a far more difficult task. One of the main issues is that many customers buy once and never again, resulting in a large number of labels with zero value. As a result, traditional probability models cannot be used to predict CLV for new users. To predict CLV for new users it is important to get a holistic view of the overall customer value and company value which brings competitive advantage to the companies. This also helps companies to design systems that will help managers with strategic decision making for their overall customers and not just a small group of them

(Kitchens et al., 2018). Therefore, it’s important to design and select the best system while predicting CLV for both existing and new users.

Because of all the reasons above our first research question is Which of the CLV calculation models is most suitable for predicting new users of multi-category e-commerce companies? " And second research question for new users is "how to leverage the CLV to enhance the company’s decision-making strategy?".

For new users, two different models will be used and compared to determine the most effective method of CLV forecasting.

The first method involves using ML models—including DNN—directly to forecast CLV, while the second involves using hybrid ML models to predict whether a new user will make a purchase or not, and then forecasting CLV for those who are expected to. After defining the best models to answer the second question customer segmentation based on CLV values to understand the high value customers will be applied and presented as the first part of holistic framework. This will help managers with a CLV-based approach to enhance their decision making since this will provide a clearer view of the overall value of the customers and company.

According to the literature, most CLV studies have not concentrated on predicting CLV for first-time purchasers, but rather on existing users with a history of many purchases. As a result, our goal is to predict CLV for every user of a company to provide a holistic view of the value of the company.

3.2 CLV prediction for existing users

To gain a competitive advantage over their competitors, businesses must employ a 360-degree strategy that examines all CLV-related factors. To get a sustained competitive

edge, businesses must be flexible in analyzing every factor related to CLV prediction throughout the enterprise and deploying systems that understand overall customer value and respond to customers based on their value to in a dynamic environment. Also complementing the customer insights, in our case this is CLV, with a decision-making approach is important for companies (Abbasi et al., 2016). Therefore, the main goal of this thesis is to develop a CLV prediction system that considers both existing and new users. We are concentrating on existing users as the second component of the innovative 360 framework. Existing users are clients who have made multiple purchases from the same company. Businesses must work hard to retain these existing users because they are more loyal than new users (Kitchens et al., 2018). However, not all existing users are equally valuable, and organizations must understand their CLV in order to effectively allocate resources.

In this thesis, we look at four research questions to evaluate the CLV for the existing users. The first question is, "Which model is preferable to predict CLV for existing users of multi-category e-commerce companies?"

The CLV of existing users has been thoroughly investigated in the literature but to our knowledge there are still research gaps for multi-category e-commerce companies. Predicting CLV for a multi-category company that sell products from a variety of categories need additional techniques to predict the actual CLV of their users and segmenting their users accordingly to create a strategic decision-making process for the company that will help with resource allocation, marketing strategies, valuing the overall company etc. Because of this, new variables needed that will capture the multiple product category related product information and to enhance the RFM based models by adding new transaction related variables The probability models are the ones that are the

most well-known, and the RFM approaches are the ones that are used the most often. All of these RFM parameters, which describe previous purchases made by clients, are significant in CLV computation. A limitation of the RFM method is that it majorly deals with the past. Forecasting future consumer behavior, on the other hand, may be a better method of determining how valuable a customer is. The Pareto/NBD model has been developed to calculate CLV dependent on future projections. This strategy has shown to be more trustworthy. It does, however, necessitate more computational power and is more complicated.

CLV models based on machine learning (ML), on the other hand, have recently been investigated and outperformed probability models. However, there is a scarcity of empirical resources in the literature for comparing various ML-based CLV models. ML models can accommodate more parameters and information than traditional probability models, which could result in better, more accurate results. In our thesis, we compare probability techniques and ML models to compute CLV to see if ML models outperform traditional probabilistic models for existing users.

In addition to comparing multiple alternative models, to respond to the second study question, "How predicting CLV combined with novel relationship-oriented constructs improve the ability to segment customers and improve the accuracy of CLV prediction?" we incorporated two additional variables for predicting CLV by enhancing the relationship marketing theory constructs that was stated by Kitchens et al. (2018). These are distinct product categories (DPC) and trends in the amount spent (TAS). DPC is an enhancement to “Choice” construct. Kitchens et al. (2018) states that when a company sells products within a variety of categories this information is extremely useful as this might be an indicator for the customers’ level of trust and interest in a company.

Knowing the products, they purchased might be an indicator of the next product they are going to buy which will provide additional information and an indicator that customers will increase their CLV. Also, by using this product category related information managers change their strategy and decisions and focus on the right strategies such as cross and upselling to increase the value of customers and the overall value of their company.

Organizations that offer a wide range of product categories should also employ product- and consumer-related indicators when assessing CLV and segmenting their users to acquire a competitive advantage. For businesses that offer multicategory of items, a crucial element is how many different product categories a client eventually purchases (Kitchens et al., 2018). This element represents the ratio of the total count of unique categories bought to the total count of purchases in a particular year and the greater the ratio, the better. The CLV is positively impacted by this ratio. As evidence, Reinartz and Kumar (2000) discovered that consumers who exclusively purchase within one category had a lower CLV compared to consumers who purchase across many categories. González-Benito and Martos-Partal (2012) claim that buying a variety of goods from the same firm displays loyalty, and that loyal consumers spend more over the course of their lifetime, hence this is also an important statistic for estimating a customer's value.

The product price range for multi-category e-commerce businesses is wide because the store sells affordable products in addition to luxury and other high-priced product categories, such as electronics. When a business classifies its customers based on CLV signals only, a challenge will arise since only the total spend of a particular customer is taken into account and not the variety of the product purchased (for example an individual who bought one luxurious item will be categorized as an “important high

value" while an individual who bought from different categories and same spend will also be classified as an "important high value," but this classification will not be accurate). The reason for this is the consumers that purchase from several different categories might become more devoted to the company going forward. As a result, classifying clients only based on the spent at a particular period leaves out individuals who may be more devoted and thus more valuable.

In light of this, we propose that DPC should be anticipated as an extra output in addition to CLV for existing users because it is crucial in deciding on the most valuable customers.

In addition to DPC, TAS is also an enhancement to "Transaction" construct. Most popular variables are RFM variables when a customer have multiple transactions. But it's also important that how the users spend change with each transaction. Previous studies show that the relationship between a user's several purchases in time is important and worth to investigate. It's a crucial metric to understand how the CLV will be affected based on the amount spend with each successive purchase (Kumar et al., 2004; Borle et al., 2008). Because of this TAS variable is extracted from the data and calculated and introduced as an addition to the "Transaction" construct of the RMT. In this study we are leveraging TAS variable as the third output.

The third output, which represents the TAS variable for each user, was derived by applying a regression model to each user's purchases. The dependent variable is the purchase amount, while the independent variable is the number of days between successive purchases. The trend value for each user is then determined by the slope of this regression model. Another important piece of information for determining a customer's true worth is the possibility that two customers with the same amount of total

spent have different trend values. A user's spending may increase with each additional purchase, culminating in a high total value. On the other hand, if we just evaluate the overall amount they spend, a client who spends high amounts on their initial purchase but less on subsequent transactions can be seen as valued even when they are not. This is due to the possibility that a client who consistently spends less may be preparing to part ways with the business. As a result, the value and segment of a client may change as a result of growing, falling, or steady trends.

These two metrics, DPC and TAS, have not previously been researched when segmenting clients despite their significance. We developed these two variables as an enhancement to RMT constructs and defined them as extra outputs for our model and published them in the literature.

Which model is the most effective to use when forecasting extra output variables in addition to CLV is the third problem in the second part of our 360° framework. We require models that can simultaneously predict the 2 extra output features, DPC and TAS, in conjunction with CLV, in order to evaluate the performance of multi-output prediction. A basic and popular technique known as a multi-output model prediction allows one model to predict many outputs from the same input data. As opposed to handling a single-output model for each output variable separately, these models are easier to develop, maintain and test.

These two features are only valid for existing users since they require more than one data point to extract a trend variable and distinct product category. The CLV is a crucial indication for businesses. To our knowledge, there are no other research that have been conducted where the researchers forecast CLV and other related metrics in order to segment clients and understand their value using multi-output prediction models. In this

study, we predicted CLV, DPC, and TAS using a multi-output DNN model and compared the outcomes to those from a single-output DNN model that we independently developed for each outcome. We also used Random Forest and Decision Trees to show that multi-output DNN models are better than other ML models. We used the multi-output variants of the RF and DT models to compare with the multi-output DNN model. DNNs are very accurate, yet some organizations find it challenging to accept them since they are black box models that are complicated for people to understand (Bhargava & Gupta, 2022). The multi-output DNN has been used in conjunction with explainable artificial intelligence (XAI) methodologies to address this challenge and acquire a better understanding of the complete model and the features that are most essential for forecasting the CLV. Our thesis is unique in this area because there is no other research which applied XAI to identify the importance of model variables connected to CLV prediction while using DNN.

The last question we are addressing is “Which segment of existing users will have the highest value in the future and how these segments impact strategic decision making?”. The purpose of this research is to assist organizations in determining the value of each client and give the management of the companies a more strategic decision-making system that will provide them with a competitive edge. By categorizing consumers into distinct categories, management may develop tailored customer strategies. For many years, CLV-based segmentation has been studied by several researchers such as Benoit and Van den Poel (2009) and Kim et al. (2006). In contrast to past research, in our study we are not classifying clients based on only CLV. In addition to CLV, DPC and TAS will also be used for segmentation. K-means clustering was used in this research to categorize and pinpoint the most valuable clients. By concentrating their advertising

strategy on their most valuable customers, businesses may increase revenue. Additionally, understanding which clients are less valuable enables companies to reconsider their current marketing approaches for these clients in order to increase client happiness and, therefore, revenue. Figure 1 shows the proposed framework, including:

i. A group of relationship marketing theory-based constructs and RFM variables to leverage as the input features for the predictive models for both new and existing users
ii. Predictive models to predict CLV both for new and existing users
iii. A system for construct evaluation and providing strategic value for decision making through customer lifetime value.

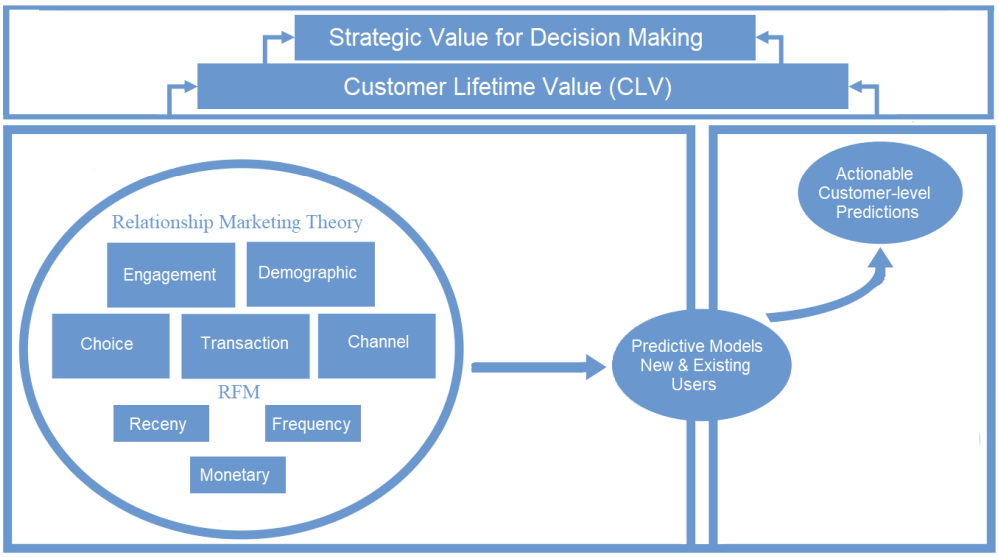


Figure 1. 360° holistic CLV prediction framework

In conclusion, we are modeling new and existing users independently in order to provide a comprehensive solution that addresses every aspect of CLV and to provide a strategic value for decision making to the companies.

3.3 CLV prediction models

This section highlights the various models used to forecast CLV for both existing and new users. Probability models, ML models, and multi-output ML models were chosen for this thesis in alignment with Yılmaz Benk et al. (2022) because they are appropriate for multi-category e-commerce businesses. Python was used to write the code for all the models. All the code was created using Python and the model training, assessment, and testing were carried out on an Nvidia Quadro RTX 5000 Max-Q system with 16 GB of RAM and 3072 CUDA cores in this study. Other tasks were completed with a 64-bit version of Windows 10 with an Intel (R) Core (TM) i9-vPro 10885 H CPU and 128 GB of RAM.

In this study, ML models were largely used, and several of the ML models used for CLV prediction are the same for both new and existing users. To test common ML models for new and existing users, different data sets were employed as inputs. First, an explanation of the common ML models used to calculate CLV for both new and existing users was made. Following that, the models that were used specific to each user type has been discussed in detail below.

3.3.1 Common models

Decision Tree

Decision tree (DT), a supervised learning is used to forecast an explanatory variable using a tree-like architecture. A DT organizes data details by asking a sequence of queries regarding their attributes. It divides the initial collection of data examples into several subsets for each conceivable response to its inquiry. Then, from each of these subgroups, smaller subsets are constructed in the same way. This method is repeated

until a stopping criterion is identified. The response variable is expected to be a constant in each of the final subsets (Bernat, 2018). According to Hastie, Tibshirani, Friedman, and Friedman (2009), after formation of the tree, the latest data details will move from the whole group of data details to the last subgroup of data, where the response variable's worth will be anticipated based on the value assigned to the right subgroup. Decision trees are classified into two forms based on the type of response variable utilized. Regression trees involve the use of continuous response variables, while classification trees work with discrete response variables. The best splitting variable and split point are chosen in different ways in these two types of trees. Regression trees were utilized in this thesis since the response variable in this investigation is continuous.

Random Forest

To handle regression and classification concerns, a supervised machine learning method known as random forest may be utilized. It is a combined technique that involves various decision trees, focuses on each using various replacement sample sets, divides tree nodes based on a few features, takes a forecasted outcome from every decision tree, and tabulates forecasted results' votes. The most accurate final forecast is the predictions' average or maximum voting from each individual tree (Win & Bo, 2020). Random forest regression was utilized in this thesis because the response variable in the investigation is continuous.

Gradient Tree Boosting

This involves a strategy that maximizes differentiable loss functions while training models progressively, additively, and sequentially. Gradient tree boosting is an ML strategy using a set of "weak" decision trees to build a "powerful" predictor for classification and regression tasks (Friedman, 2002). Gradient tree boosting outperforms

a single decision tree in terms of bias and variance. This is accomplished by employing a forward stage-wise modeling method allowing the optimization of a differentiable loss function that quantifies the effectiveness of the model. Because of its exceptional performance, gradient tree boosting is well-known among data scientists, and some of them have used it to predict CLV.

XGBoost Regressor

XGBoost, an algorithm, uses exact estimates to determine which model is more accurate. It computes advanced regularizations (L1, L2) to reduce overfitting and improve model performance and generalization. It computes second-order gradients at the same time to reduce the loss (Hu, Lin & Tang, 2019). XGBoost, which is simple to use, can successfully manage large datasets. It provides great performance and high accuracy in a short period of time.

Deep Neural Network (DNN)

They are machine learning models that are impacted by the human brain. They have input, output, and hidden layers. The number of node in the hidden layer determines the model's complexity or depth. Although deeper networks take longer to train, they are usually more accurate. As a result, maintaining a balance between the period of training and the complexity of the model is a critical component of this method.

A number of models are generated using the neural network framework. The deep neural network (DNN) model, for example, can handle extremely complex input-output interactions due to its multiple non-linear hidden layers (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). DNNs have reportedly been used in data science to forecast both churn and purchases (Kim, Choi, Lee, & Rhee, 2017; Sifa et al., 2018). DNN has also been investigated for use in CLV prediction. In most cases, they are used

as the main method or as a point of comparison. The ability of DNN to recognize non-linear patterns in data is what gives it such a broad range of applications. Because the activation function used in the node can generate both numeric and binary outputs, they are also useful for classification and prediction problems.

A DNN is made up of three layers: an input layer, hidden layers, and an output layer (Schmidhuber, 2015). Predictions are in the output layer, while the features (web analytics and CRM data from each user's previous purchases) are in the input layer. Neurons having nonlinear activation functions are used to connect and construct the layers. Many iterations, known as epochs, are performed during the learning process to improve the neural network. A gradient descent technique changes the weights at each epoch to reduce a preset cost function, including the root-mean-squared error. Because of its ability to predict non-linear relationships, DNNs are our primary technique in this study.

3.3.2 New user specific CLV prediction models

Existing users can infer some characteristics from transactions, such as recency, frequency, and monetary value, but new users can only deduce features from their initial purchase, necessitating the employment of alternative modeling techniques. Because of this reason, other models are also tested for new users.

In addition to the ML models mentioned above, the second type of model studied in this study for new users uses two stage hybrid ML models to predict whether a new user will make a purchase or not, followed by forecasting the CLV for those who were projected to make a purchase.

Hybrid Model for New Users

Many algorithms, specifically distinct machine learning techniques, are combined in hybrid machine learning techniques. Our literature review indicates a wide range of hybrid ML techniques in numerous disciplines, including bioinformatics, recommender systems, medical diagnosis and, most importantly related to this work, customer segmentation and financial forecasting. Hybrid models can deliver better predicted results while eliminating defects since they efficiently deal with high-dimensional datasets and incorporate the best aspects of numerous models (Brazdil, Carrier, Soares, & Vilalta, 2008).

In the literature, hybrid models are frequently employed to predict CLV. Chamberlain et al. (2017) and Vanderveld et al. (2016) used them to forecast the CLV for first-time buyers. The CLV of each e-commerce site user was estimated using the same method, which employed a two-stage random forest model. The first stage uses logistic regression, a binary classifier, to assess the likelihood that a user will make a purchase during the specified time period. In the second stage, the monetary value is projected for customers who were expected to make a purchase in the first stage. We also used this method to forecast CLV for new users because our major goal is to estimate CLV for every user of the company.

3.3.3 Existing user specific CLV prediction models

After determining the initial purchase date within our observation period, we included existing users who had made at least one purchase during the three months before the first transaction. In the database we used for this investigation, existing users had a considerable number of transactions. By extracting numerous factors from existing

customers' transactions, such as RFM value, we were able to apply various techniques including Pareto/NBD and Multi-Output DNN Models, in addition to the ML Models stated above.

3.3.3.1 Pareto/NBD model

This model is an appropriate benchmark model for noncontractual settings where transactions occur under any circumstances.

The Pareto/NBD model was one of the most prominent probabilistic models. Fader et al.'s (2005b) model, which models expenditure per transaction as a Gamma-Gamma distribution, was employed for the purposes of this investigation.

3.3.3.2 Multi-Output models

Deep Neural Network (DNN)

It is common to develop a DNN model for a classification or regression tasks separately for each prediction of interest, however, it may be advantageous for some predictive modeling applications to create one model that can make two or numerous predictions at once. A solution to solve this challenge is to develop a new model capable of producing independent forecasts for several outputs from a single input. This is a NN model with many outputs.

Figure 2 is an example of multi-output model with three outputs and two hidden layers.

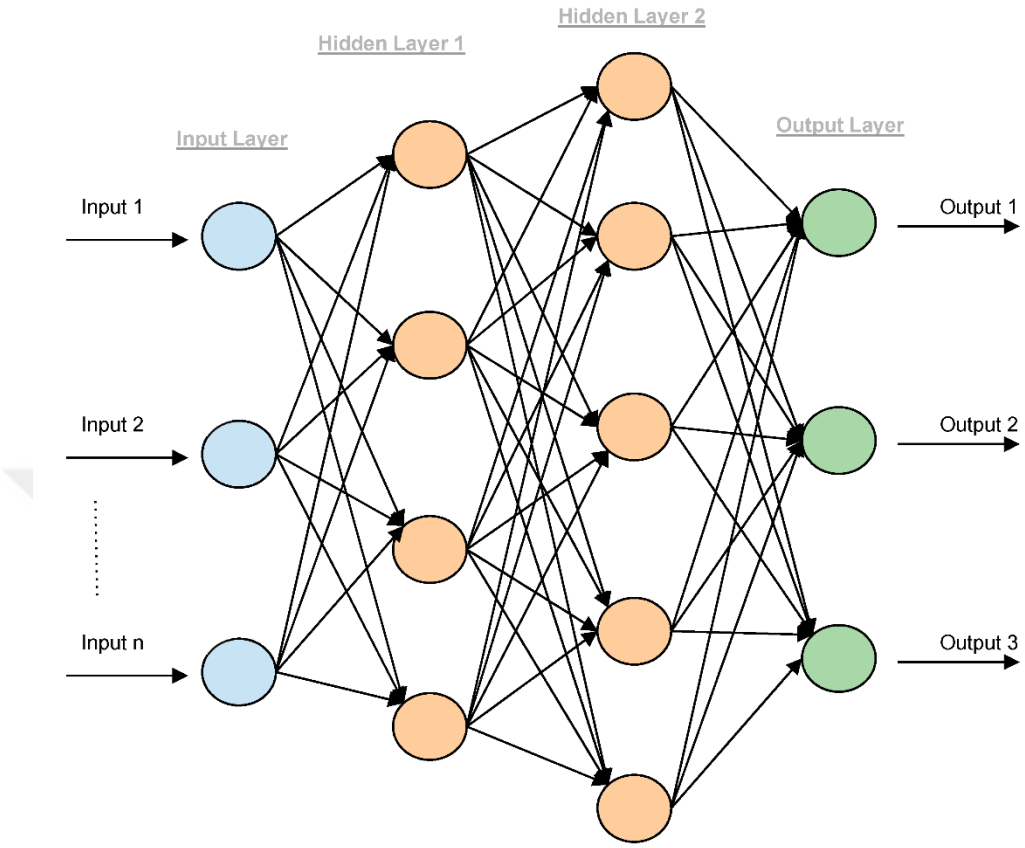


Figure 2. Multi-Output DNN structure

One advantage of this sort of model is that it only needs single model to be built and maintained, as opposed to several models, and training the model on all output types at the same time may result in improved forecasting stability across all output types. This structure, which is consistent with multi-output learning, helps the model generalize more on each task, according to Ruder (2017).

Neural network models can handle multi-output regression because they train a continuous function that can simulate a more elegant link between input and output changes.

Multi-output regression can be performed by neural networks by describing the number of output labels as the number of nodes in the output layer. For a job involving three

output variables, for instance, a neural network output layer with three nodes, each with an activation function, is needed.

A DNN model capable of predicting three regressions from the same input data was employed in this study. We set out to estimate the CLV as one of the outputs, along with the ratio of distinct product categories (DPC) purchased and the trend in the amount spent (TAS), in order to understand how these elements, affect each user's value.

The initial stage was to develop a multi-output model in which all three output labels were treated as continuous variables. Then the second and baseline models were used to forecast each output variable separately using a single-output DNN model, the results were compared to those of the multi-output model.

Aside from DNNs, many machine learning approaches already support multi-output regression. Decision trees and decision tree ensembles, such as random forest, are widely used methods. The disadvantage of utilizing decision trees for multi-output regression is that the connections between inputs and outputs may be too blocky or overly structured depending on the training data.

Finally, multi-output ML models were employed to investigate the three linked regression tasks for existing users.

Model Structure for Multi-Output Deep Neural Network

This research used a neural network with two hidden layers (144 and 45 units) and input layers (64) where the parameters were set through Bayesian optimization-based hyperparameter tuning. The number of nodes and parameters in single-output models varies based on the hyperparameter configuration. In single output models, just one output node is included to predict the output variable. Only one output node is included in single output models to predict the output variable. In contrast, three output nodes

were predicted in a multi-output model, and the concatenated weights of every pair of output nodes were then supplied to the third output to reflect the relationship between the three output variables. Model architecture is shown in Figure 3.

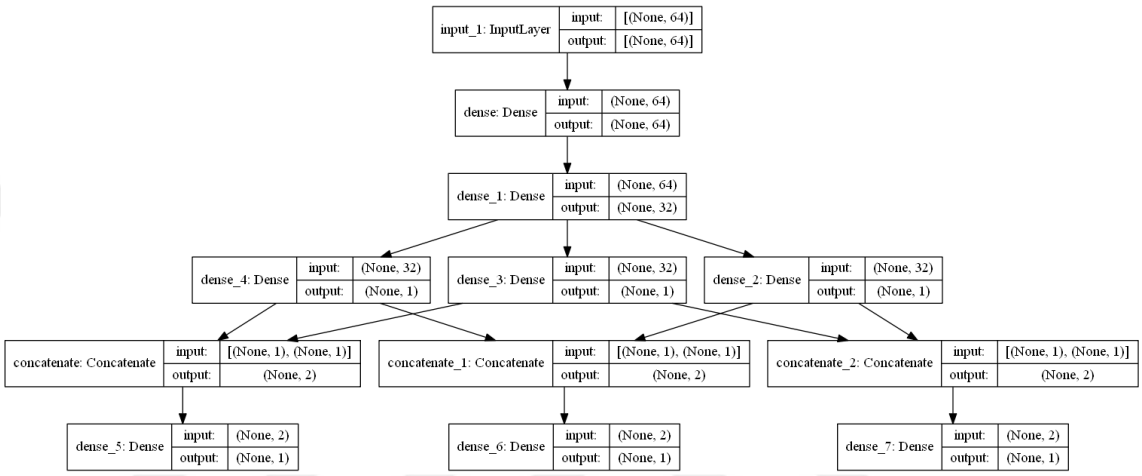


Figure 3. Model architecture for multi-output DNN

To discover the best and most successful settings for single and multi-output DNN models, epochs, the number of nodes, learning rate parameters, and batch size were changed and tested via hyperparameter tuning. Table 1 shows the top hyperparameters used for multi and single output DNNs.

In the hidden layers, a well-known activation function which is rectified linear unit (RELU) is used. For all outputs, the Adam version of stochastic gradient descent and the Mean Squared Error (MSE) loss are utilized.

Table 1. Hyperparameter Ranges and Optimal Hyperparameters for DNN Models

Model Configuration					
		Single-Output DNN			Multi-Output DNN
Name	Range	Optimal Value			Optimal Value
		CLV	DPC	TAS	CLV, DPC, TAS
Batch	(512,1024)	916	948	734	960
#ofNodes	(32,128)	100,51	94, 50	72, 36	114, 45
Learning Rate	(0.1, 0.999)	1e-2	1e-2	1e-2	1e-2
Epochs	(100,200)	143	133	148	136
Optimizer	Adam	Adam			Adam
Loss Function	MSE	MSE			MSE

Multi-Output Decision Tree

These are supervised learning regression and classification algorithms. The objective is to create a model that forecasts target variable values by studying easy decision rules generated from data properties. Decision Trees (DT) can be easily understood and interpreted. They are convenient in forecasting multiple outputs, making them a great fit in our case to forecast CLV, DPC, and TAS simultaneously and compare to multi-output DNN models. Given the high nonlinearity and multidimensionality of our data, decision tree regression looks to be a suitable alternative to testing with our proposed model.

A primary advantage of decision tree implementation is that it's very easy to tune. Max features (total number of variables while attempting to determine the best split), min samples leaf (the least samples needed for a node to change into a leaf), min samples split (the minimum number of samples needed to divide), and max depth are the hyperparameters utilized in this study's DT. Table 2 demonstrates the ideal hyperparameters and ranges for multi-output DT.

Table 2. Optimal Hyperparameters and Range of Hyperparameteres for Multi-Output DT

Model Configuration		
Name	Range	Optimal Value
maximum_depth	(3,30)	4
maximum_feature	(0.1, 0.999)	0.1
minimum_samples_leaf	(2,30)	3
minimum_samples_split	(3,40)	34

Multi-Output Random Forest

The ensemble learning technique is used to improve forecast accuracy while avoiding overfitting. Random Forest (RF) mixes numerous decision trees on dataset subsamples (Breiman, 2001). To divide a leaf node sample into two groups, the model uses every variable from the training set and attempts every conceivable combination. The mean MSE between each sample and the dividing line is calculated for each feasible split, and the lowest MSE split is chosen. When the number of leaf samples falls below a predefined minimum, the process of separating leaf nodes comes to a halt. The last forecast is made using the mean tree estimates (Breiman, 2001). RFs can also be used for regression and classification. This study addresses a multi-output regression issue using continuous variables such as CLV, DPC, and TAS. The CLV, DPC, and TAS outputs are all predicted using a multi-output Random Forest Regressor in this work. RF, like DTs, is an effective algorithm for forecasting a wide range of outcomes. González, Mira, and Ojeda (2016) used multi-output RF to estimate electricity demand and pricing at the same time. According to the study, multi-output RF is better than single-output RF based on accuracy. We compare our proposed multi-output DNN model to multi-output RF in this situation. We also performed tuning of hyperparameters to get the best values for multi-output RF. The four multi-output DT algorithm hyperparameters that was

stated above were used here as well. Other than the four, we provided the n_estimators option, which defines the total trees. Table 3 shows the ideal values and ranges for configuring the hyperparameters for multi-output Random Forest.

Table 3. Optimal Ranges and Hyperparameters for Multi-Output RF

Model Configuration		
Name	Range	Optimal Value
maximum_depth	(5,30)	5
maximum_feature	(0.1, 0.999)	0.9
minimum_samples_leaf	(2,30)	11
minimum_samples_split	(3,40)	29
n_estimator	(10,100)	37

Hyperparameter Tuning

To acquire reliable results, machine learning models' hyperparameters must be optimized. Parameters in a model are variables obtained during training that show the translation of input data into predicted output, as opposed to hyperparameters, which govern the model's structure and may have an impact on model accuracy and computation efficiency (Win & Bo, 2020). Hyperparameter tuning is the process of selecting the best set of parameters from among the many hyperparameters that may be included in a model (Win & Bo, 2020).

For all models, we employed Bayesian Optimization Hyperparameter Tuning to fine-tune the hyperparameters. Bayesian optimization, according to Wu, Chen, Zhang, Xiong, Lei and Deng (2019), is an effective strategy for conducting computationally challenging functions. The foundation of Bayesian optimization is the Bayesian theorem. They conducted a number of experiments using standard test datasets. The results from an experiment indicate the recommended approach can determine the optimal hyperparameters for well-known ML models such as neural networks and the

RF algorithm (Wu et al., 2019). Another technique used in applied machine learning to test a model's performance on a new dataset is cross-validation. In other words, when predictions are produced using data that was not used during training, a very limited sample size is employed to assess the model's overall performance. We used Bayesian Optimization hyperparameter tuning with repeated 5-fold cross-validation to avoid overfitting and discover the best hyperparameters, as indicated by Kumar, Batra and Raman (2021).

The values of hyperparameters have a substantial impact on the structure and training efficiency of DNN models. As a result, selecting the optimum hyperparameter settings is critical to realizing DNNs' full potential (Dobslaw, 2010).

Four critical DNN hyperparameters were examined in our study: the total count of nodes in every layer, epochs, learning rate and batch size, corresponding to the study of (Kapanova, Dimov, & Sellier, 2018; Ulrich & Korn, 1999).

We chose the hyperparameters and tuning ranges for the multi-output RF and multi-output DT algorithms based on the part of the previous research in (Lee, Bae, Hong, Yang, Panagos, Borelli, Yang, Kim & Lim, 2022). Testing every potential parameter value combination for every model is computationally costly, thus certain parameters are changed while others remain unchanged. Prior research is used to choose the most common parameters for each type of model.

After evaluating the effectiveness of the machine learning models and selecting the optimal hyperparameters, we tested and compared models’ accuracy.

To accomplish this, we equip the models on all of the training data using 5-fold cross validation. The outcome of our models on the unknown test set was then evaluated and compared for each model using the best hyperparameters we could find.

CHAPTER 4

DATA ANALYSIS AND PRE-PROCESSING

CLV has been investigated by many researchers for a long time, but still the most prevalent CLV model approaches nowadays use monetary values and ignore perceptual or other non-monetary factors. According to Larivière (2008), studies have so far looked at the direct relationship between customer behaviors and customer financial metrics. Sifa et al. (2018) discusses how customer value perception needs to take into account many consumer characteristics and factors. Additionally, to include multiple important factors into one model, their study suggests additional variables pertinent to an e-commerce environment. Kitchens et al. (2018) referenced the relationship marketing theory (RMT) while designing their advanced customer analytics framework. They stated that customer relationships provide the company with a source of long-term competitive advantage and strategic value. Kitchens et al. (2018) used a group of relationship-based constructs while building their advanced analytics systems to predict churn and CLV. The constructs that were used in their study were listed in the literature review section. They evaluated the different constructs to understand which one provides more value.

Comprehensive information about consumers' purchasing habits is needed for CLV calculation. As a result, companies track and store analytics data on customers' order histories and other actions, such as customers' engagement with content on the company's websites and apps and their click-through rates. With the use of this data, models that reflect the typical purchasing patterns of customers may be developed. Then, future revenues and associated costs for consumers, or CLV, may be predicted

using these models. We leveraged web analytics + CRM data sources for the training of the model and prediction of the CLV for one of the biggest Online Multi Category Turkish E-commerce Retailers.

4.1 Data

Data collected by a company's website or mobile app through a web analytics platform collects information about how users interact with the site or app.. According to Chamberlain et al., (2017), web and app activity logs are among the most complete and comprehensive data sources accessible for CLV prediction. Using a user's behavior on a website, Bucklin (2008) investigated how to predict whether they would make a purchase. Interactions, content, and data gathered from many marketing channels engagement, intent to buy, product affinity, marketing channel and content preferences, preferred device, and recency, frequency, and monetary (RFM) metrics are all part of web analytics. In order to manage data for profitable consumer categories and incorporate these indicators into advertising networks, analytical platforms may also act as a target platform (e.g., integration of Google Analytics, Google BigQuery with Google Ads). The web analytics platform that was used in this study where the data was collected was Google Analytics and that data was exported and stored inside Google BigQuery. Also, CRM data uploaded into BigQuery by the company that was investigated in this study.

The data was collected from a Turkish e-commerce company that offers a wide range of categories of products. One of the main reasons why we collected data from this company in this study is focusing on multi-category e-commerce retailers and there are limited number of multi-category e-commerce websites in Turkey. We reached all those

companies and only this company allowed access to their data. The constructs that were mentioned above were shared with the companies in advance and the company we collected the data from shared any data they have matching with the constructs we asked for. Client shared the 2 years of data between July 4, 2018, to August 1, 2020. There are a total of 150,421,192 transactions included in the dataset. In addition, there are 8,900,215 users. Web and app analytics data that can be matched by unique User Id that is assigned to each user are also included in this dataset.

Covid-19 Pandemic and Impact on E-Commerce

Coronaviruses are a group of viruses that may infect the lower airways and start more serious illnesses including pneumonia (Bhatti, Akram, Basit, Khan, Raza & Naqvi, 2020). People infected with this virus are susceptible to both deadly and benign viral diseases. On December 12th, 2019, the COVID-19 virus was first detected in Wuhan, China. Within a few weeks, millions of individuals around the world had contracted this illness. The economy of 205 nations is also affected with this illness. The worldwide economic climate was difficult, and this had an impact on e-commerce patterns. On March 10th, 2020, COVID-19 epidemic's first case discovered in Turkey. On March 15th, 2020, the virus's first fatality occurred.

The majority of the world, including Turkey, has been socially isolated during COVID-19 and people stay inside until absolutely necessary. As a result, COVID-19 has had far-reaching effects on sectors as diverse as e-commerce, technology, business travel, and the economy. As a result of the increase in both instances and deaths, measures such as curfews, mall closures, and the transfer to distance learning were enacted.

There are studies which show how e-commerce grew because of the pandemic. Bhatti et al. (2020) look into the hypothesis that the spread of the coronavirus is responsible

forthe expansion of the online retail sector. In this scenario, where conventional shopping takes place primarily in offline stores, e-commerce has emerged as a primary alternative.

Similar to the scenario in other countries, when a curfew was imposed in Turkey, people started mainly shopping online to fulfill their purchases. Additionally, customers have turned to e-commerce sites when conventional shopping options, such as malls and trade centers, are closed or unsafe (Kayakuş & Çevik, 2020)

Data period (July 2018-August 2020) in this study contains the time when Covid-19 pandemic started in Turkey, which is March 11th, 2020. To remove the impact of pandemic from our study we excluded the data after March 31st, 2020, which reduced the period of time that can be used in the analysis from 24 to just 18 months.

Time Horizon of the Prediction

To reduce the influence of seasonal changes, the forecast horizon should be a set number of years (Wang et al., 2019). Long prediction models are often impractical because they need a lot of past data to create training labels. For instance, multiple authors conducted CLV forecasts in a span of 1 year (Chamberlain et al., 2017; Vanderveld et al., 2016). Different sources might contribute to the input features of a model. In this study we referenced RMT and used the available suggested constructs in Kitchens et al. (2018)'s study as our input features. As the first data source we used the transaction construct where the purchase history of the customers is available inside web-analytics platform. Purchase history is typically the main source utilized when it is available. Demographics construct where the sex of the customers available coming from CRM was also leveraged. According to Wang et al. (2019), client demographics, customer groups, return histories, and customer service quality measurements are also other common

sources. Engagement construct, which includes information regarding website visits coming from the web analytics was another group that we used as the input features. Website behavior has also been leveraged by some researchers. In order to forecast consumers' lifetime value, Vanderveld et al. (2016) studied how the consumers interacted with an e-commerce website before making a final purchase. These interactions included visits to websites and clicks on marketing emails, time spent on the website, and product interest information. According to Wang et al. (2019), Sifa et al. (2018) used interaction metrics such as the number of sessions, matches, and days played as well as the amount of in-game money spent, to estimate the lifetime value of users of free-to-play games. In addition to longitudinal patterns of activity like the period between the first and final session, meta-features including the user's origin country, device kind, operating system version, and channel of collection were also utilized. The “choice” of which stores the purchased product related information was another important construct that we leveraged based on the available data we have. Product related input features were also used in other previous studies such as by Chamberlain et al. (2017) where product embeddings were used in unsupervised neural networks and mixed with manually constructed attributes learned from session and app logs. And lastly, we used channel construct which shows where the company first acquired the client. Satisfaction, messaging, and firm related constructs were not included in the dataset that was used in this study because they were missing on the dataset we collected.

The model also included the number of visits, page views, duration on the website, and other aspects of online activity in addition to spending amounts, the quantity of items purchased, and the seller and product category of each individual product. All of these

variables were available inside the data we collected from the client and all of them can be matched with the constructs of the relationship marketing theory. Therefore, we added all of these as our input features.

We address the issue of calculating the total spend value of each client over the next 12 months after the initial purchase. Since we excluded post Covid-19 period we restrict our observation period between 2018-07-04 and 2020-03-31. The initial purchase period is the one right before the full 12-month cycle which is between 2020-03-31 and 2019-03-31. The 6-month period before 2019-03-31 was used as the initial purchase period. We designed two databases separately, one for new and one for existing users. To determine the existing users, we used the initial 6-month purchase period between 2019-03-31 and 2018-10-01 and used the 3 months period between 2018-10-01 and 2018-07-01 to determine the existing and new users. For the existing users, after finding the first purchase date between 2019-03-01 and 2018-10-01 if this user purchased within the 3 months prior to the first purchase period than we counted that user as existing user and if the user does not have another purchase before the first purchase during observation period than that user counted as a new user. As a result, the number of new customers is 1,394,947 and the number of existing users is 1,715,519 between 2019-03-31 and 2018-10-01. There might be other previous purchases which happened before our analysis period for these same customers but for the sake of the study and since we only have data for this two-year period, we assume that they are first time buyers within the training period.

4.2 Data analysis

In order to predict the CLV holistically we are focusing on two different user types which require distinct data and models. The detailed explanation of the data that was used in the study is shown in Appendix A (Table A1).

4.2.1 Data analysis for new user

There were 27 features that were extracted from the whole data. 20 of them are Categorical variables and 7 of them are continuous variables.

In order to utilize the training data for the model, the categorical variables are subjected to the following data transformations.

Categorical Features

Operating System: Due to the fact that the dataset contained over 15 different operating systems, only the four most common systems (Android, IOS, Windows, Macintosh) were retained, whereas the remainder were grouped under the category “Other”. iOS and Android are the most popular mobile operating systems, with Windows following closely behind.

Data Source: The dataset contained both app and web data and 82 percent of users were using the app.

Campaign: The campaign variable corresponded to over 9000 distinct values in the dataset. Instead of grouping the campaigns, we decided to create two new variables that showed whether campaign information was stated or not. “*Campaign*” indicates whether the user clicked on a paid advertising campaign or not. This information is important in order to understand if there is any difference between the CLV of users coming from

paid searches or not, which will help companies to better understand their marketing efforts.

Source: The different channels that customers chose while visiting the website when making their initial purchase were displayed in the "*Source*" variable. We were able to narrow this down to three key sources. Because the dataset contained over 532 different sources, the two most common sources (google, direct) were recorded, whereas the rest were grouped under the category "Other".

Medium: This variable shows which medium was used when a user came from a paid campaign. There were more than 120 distinct mediums. Based on the distribution of distinct categories there were five main groups of mediums with almost equal sizes and the sum of all the rest was also equal to the previous five. Therefore, we retained the five main mediums (cpc, organic, none, referral, retargeting) and added the others under the "Other" bucket. Most of the users came from apps and were not assigned a medium, which is the reason why the majority of users had a value of "none" for the medium variable.

Sex: There were different values for sex. Values for women had 3 distinct values—"female", "Female", and "0". The same result was obtained for the men's values. Therefore, we represented all the different values using 0 and 1, with 0 for women and 1 for men. More than half of the users in this database were women. Furthermore, 7 percent of the users did not disclose their gender.

Device Type: There were three different device types—mobile, desktop, and tablet—but the majority of customers used mobile devices, in line with the fact that the dominant data source was apps.

Seller: Since this was a multi-category e-commerce company, there were also different sellers, called sellers. The dataset contained almost 1500 different sellers. We matched the seller with seller categories coming from the company's website with a script. Based on the frequency of each category only the eight most frequent seller categories were kept and linked to users, with the rest of the users being grouped under the category 'Other'. The categories under other were very rare and did not provide additional information to the models so they combined under the other group. Nine dummy variables were created from this variable.

Product category: There were ten different main product categories in the company. However, these categories were not stated explicitly in the database. Instead, there were 528,460 distinct subcategories. These subcategories were mapped via a python script to match the subcategories with the main categories by keyword mapping. We used ten main categories from the company's website to make the analysis easier and more suitable for the algorithms; otherwise, it may have required a lot of processing power. Based on the data, the "shoes and bags" category seemed to be the category with the most purchases but most of the revenue came from the "electronics" category.

Figure 4 shows the count of categories and the revenue distribution by Product Category can be seen in Figure 5.

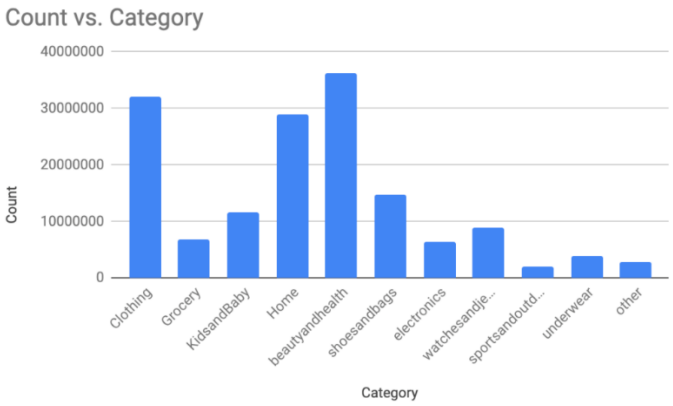


Figure 4. Product category distribution by count

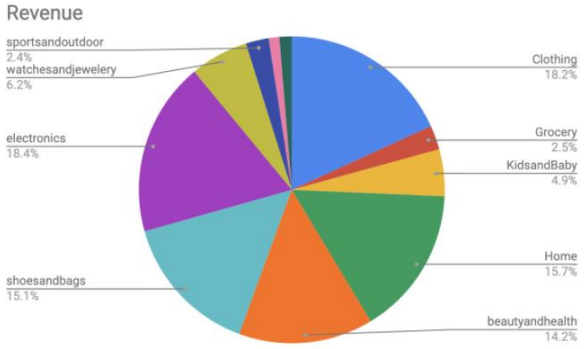


Figure 5. Product category revenue distribution by percentage

Continuous Variables

Total revenue ("totalRevenue") is the output label which is the total amount of customer spend within one year after the initial purchase. The detailed explanation of the continuous variables that was used in the study is shown in Appendix A (Table A1). Also, an overview of the data description for continuous variables can be seen in Appendix A (Table A2).

4.2.2 Data analysis for existing user

There were 50 features that were extracted from the whole data for existing users. Nine of them are Categorical variables and 41 of them are continuous variables. Recency, frequency, monetary values are extracted both for Pareto/NBD and multi-output ML models. Seller and product category related features such as total product category purchased ("*totalproductcategorypurchased*"), total seller purchased ("*totalsellerpurchased*") and distinct number of sellers and product category that was purchased by each user are also calculated and added as additional features. Lastly a trend variable calculated for each of the existing users. If a user has at least two purchases during the observation period, we run a regression model where the independent variable is time as the number of days after the first purchase and the value of the purchase is the dependent variable. Then the regression model's slope for each user is our trend value.

Existing customers may have made several purchases throughout the period of observation. In order to ensure no data is lost, continuous variables are computed as the average of each user's prior transactions. For each unique Categorical feature value, one column is created. This turned 7 categorical features into 21 features.

For the multiple models used in this research, the input and output variables needed to be in unique formats. Pareto/NBD model requires recency, frequency, and monetary values as the only inputs. The output variables of the single-output and multi-output ML models required distinct preprocessing. Nevertheless, for the input characteristics, all models were subjected to the identical preprocessing steps:

i. The first output label was a new variable named "totalrevenue" that represented the sum of the customer's annual expenditure.

ii. Given the possibility of numerous transactions being performed, continuous variables were calculated using the average of all purchases made by every user throughout the observation period.

iii. One column was produced for each unique value of the category variables. Then, that column was given the proportion of one value over all transactions for each user. One example of these was the feature “opSystem”, which displayed the operating system the user was using to access the website. There were 5 unique values for the operating system variable (Macintosh, IOS, Windows, Android, and other). For each different one, one unique new column was created. Let's say a person completes ten distinct purchases, four of which occur on a Mac device, four on an iOS, and two on an Android device. For Mac, iOS, and Android, the numbers would be 0.4, 0.4, and 0.2, respectively. The values for Windows and other operating systems will be 0.

iv. Various approaches have been used to fill in the missing data for the variables sex, “timeonsite”, “pageviews”, and “prevtrend”. Sex was encoded with two values: 0 for "women" and 1 for "men." The missing values for Prevtrend and Pageview are then imputed by the mean of the available values for these two variables, and the missing values for Sex are then filled by 2. Additionally, the whole dataset was cleaned up of any time-on-site data that was missing.

v. "distinctcategoryp" is a new input variable that counts how many times purchases were made from different categories during each transaction.

vi. Input variables total purchased category and total purchased seller were introduced to as the sum of all product categories and vendors used in a given transaction respectively.

vii. In addition, we converted categorical data to a numerical representation using one-hot encoding. This method divides the contents of a column among a large number of flag columns and gives each one a value of 0 or 1. Win and Bo (2020) say that these binary values give information about how consolidated and encrypted columns are linked.

We also obtained three more characteristics for each consumer, including recency, frequency, and monetary value. This data was extracted from our dataset's past transaction records. Recency refers to the number of transactions made between the user's most recent transaction and the conclusion of the observation period, which was March 31, 2019.

The total count of repeat purchases made throughout the observation period is known as frequency, and the average amount of each purchase is known as monetary value. Consumers who purchased something on the final day of the observation period had a recency of one, whereas consumers who placed an order on the first day of the observation period had a recency of 365 (Bernat, 2018).

The average values of prior purchases were added as a separate variable for existing users with the name of “avgValue”. Due to the lack of historical purchase data, one-time customers were omitted from the existing user dataset.

4.3 Pre-Processing

Pre-Processing for Pareto-NBD Models

We begin by gathering the information needed for the Pareto/NBD analysis. For each existing client, the Pareto/NBD and Gamma Gamma models use RFM based three

variables: recency, frequency, and monetary value. This comes from the transaction data in our dataset. The recency is the number of days between the last purchase a customer made and the end of the observation period, which is March 31, 2019. Frequency is the average number of purchases made during the observation period. Monetary value is the average amount spent on each purchase.

The one-time purchasers must be excluded from the dataset since they lack a history of previous purchases. These models are designed to forecast the future spend of each client. The revenue variable, which displays the revenue produced by client orders, may be used to calculate this figure. A monetary value is created by averaging the costs of all transactions for each consumer.

Pareto/NBD models just need transaction data, which should include the customer ID, the day and time the consumer made the purchase, and the spending. It is necessary to translate this data into Recency, Frequency, and Monetary Value (RFM) data along with the duration (T) since the consumer made their first purchase. The Lifetimes package, which optimizes the likelihood functions, is used to derive the parameter functions. Additionally, this package takes advantage of these parameters by implementing them in the appropriate functions to get forecasts for each customer's expected transaction volume and dollar value.

Pre-Processing for Machine Learning Models

The data are prepared in the following ways before being fed into the Machine Learning models: Each existing customer's recency, frequency, and monetary value are calculated by parsing the data set. Customers who shopped on the final day of the observation period have a recency of one, while those who shopped on the first day had a recency of 365.

Pre-Processing for Multi-Output Machine Learning Models

In addition to CLV, we predicted two more output characteristics using a multi-output deep neural network. Along with CLV, we also looked at the ratio for distinct product categories, the trend in the amount spent by each user, and other factors. Following the aforementioned preprocessing, our multi-output model generated two additional variables.

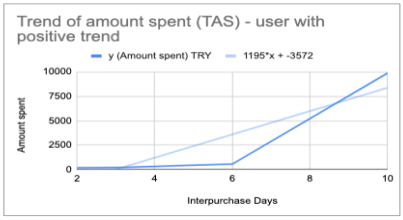
As the second output label, the distinct category bought (DPC) variable was introduced. The proportion between the total number of product categories purchased in a given year and the total number of distinct product categories purchased is shown by this variable. Both the observational period and the test period were used to establish this value.

The third output label was determined for each user, revealing the trend in the amount spent. We ran a regression model for each user's purchases individually, with the number of days between consecutive transactions as the independent variable and the difference in purchase amount between consecutive transactions as the dependent variable. Two clients with the same CLV may have different trends, which would provide us with yet another important piece of information for figuring out a customer's genuine value. The validation data sample purchase amounts were regressed on prior purchase amounts, interpurchase times, and cumulative purchase frequency. The predicted purchase volumes were then calculated using the coefficients. Figure 6 displays the prior transactions, interpurchase day, and transaction amounts for two hypothetical consumers. For both users, the overall sum was comparable and substantial. The first consumer had a trend of 1195 and made four transactions. With each purchase, the amount spent increased. The second consumer made six transactions, with a revenue decline in each purchase, with a downward trend of -1239. Based on this, the second

individual with the negative trend may be churning even if both users have the same total amount spent.

Customer purchases - with positive trend

X (Interpurchase Day)	y (Amount spent) TRY	Total Revenue After First Purchase
2	171.48	10,629.91
3	205.94	
6	548.97	
10	9875	



Customer purchases - with negative trend

X (Interpurchase Day)	y (Amount spent) TRY	Total Revenue After First Purchase
2	3100	10,114
2	3344	
3	3320	
3	3300	
4	120	
5	30	

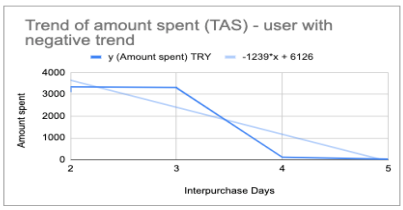


Figure 6. Example transactions of two users one with positive trend one with negative trend

After standardizing the trend, a category segment variable was produced based on the trend values, using K-means clustering to better comprehend the trend value. The clusters were then classified as positive, stable, or negative trend value clusters. Positive trend shows that this group included consumers who had greater trend values, which means that they spent less at first but then raised the amount that they spent on their respective purchases. A stable trend means customers in this category had a consistent purchase tendency of about 0. Lastly the negative trend group included users who spend a lot initially and then spent less on their purchases. Customers were divided into three categories based on their purchase habits, and this variable is referred to as Segment. The boxplots of total revenue after the initial purchase for each trend section are shown in Figure 7 below. The right graphic, for example, indicates that the mean value of total income within one year in the positive trend class was larger than the means in the stable and negative trend segments.

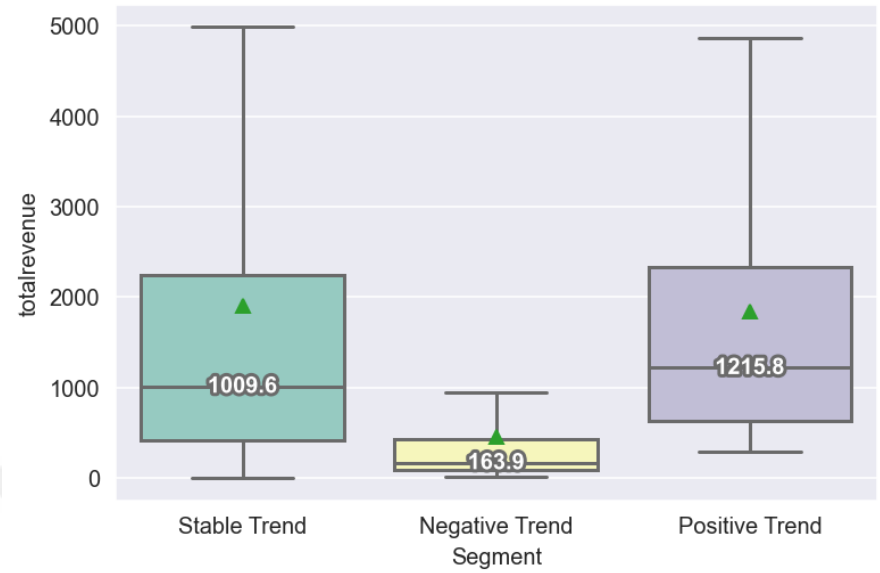


Figure 7. Trend segment by total amount spent (totalrevenue)

4.4 Performance measures

There are many different performance metrics that may be used with different models. We looked at earlier research on this subject.To assess the model's anticipated accuracy, Tsiptsis and Chorianopoulos (2010) highlighted the need of examining a wide range of residual diagnostic plots and data. Mean errors for all studied records are a common form of error measurement. The mean-squared error (MSE), mean absolute error (MAE), and root mean-square error are examples of these error measurements (RMSE). CLV projections are typically evaluated using MSE, MAE, or RMSE, as stated by Berka (2005) and Tsiptsis and Chorianopoulos (2010). Using the data from the training set, Glady et al. (2009) calculated the RMSE and MAE to see how closely their predictions came to the actual CLV in the test set.

In addition to the aforementioned metrics, we wanted to demonstrate that the models could discriminate high-value consumers from others. We picked the hit-ratio statistic

because we wanted to categorize users of e-commerce businesses, and to see the capability to identify high-value individuals is crucial when forecasting numerous outputs with the same model. A hit-ratio statistic was developed, which is the percentage of consumers who's anticipated CLV is the same as their actual CLV, was introduced by Donkers, Verhoef, and de Jong in 2007. The hit ratio calculates the percentage of these consumers have a forecasted CLV over that amount if the most valued 20% of consumers have a CLV over a particular threshold. In Moro et al. (2015), an ordering-based hit ratio was also taken into account. Chiang and Yang (2018) used the hit ratio to assess how well the CLV prediction model performed.

The performance measurements used in this research were the MAE, MSE and RMSE. In addition, the hit-ratio was also employed to compare the models' capability to segment users.

CHAPTER 5

NEW USER MODEL RESULTS

This chapter presents the model results of the current study for new users. We begin this chapter with the results of different ML models that were used in this study for the CLV prediction and following Hybrid Model results for new users.

We conclude this chapter with the comparative analysis of the results of the models that were used for new users.

5.1 ML models

This section summarizes the results of different ML models; Decision Tree, Random Forest, XGboost, Gradient Boosting, and DNN used to predict the CLV for new users. We begin each subsection by reporting the optimal hyperparameters regarding data representation and the model. Then we continue with the two stage hybrid ML models. Lastly, a comparative analysis highlights the relative performance improvement upon the CLV prediction models achieved to predict CLV for new users. Since it would take a very long time to test every possible parameter value combination for every model, certain parameters are altered while others are left alone. For each sort of model, the most typical parameters are selected based on prior research (Lee et al., 2022).

We examined and compared model accuracy after assessing the machine learning models' efficacy and choosing the best hyperparameters. In order to do this, we also use 5-fold cross validation to equip the models with all of the training data. Then, using the best hyperparameters we could uncover, we tested and compared each model's performance on the unknown test set for each different model.

5.1.1 ML model results for new users

We initially divided the new user data set into training (%70) and test (%30) sets in order to apply the machine learning models. Multiple Machine Learning models, including Decision Tree, Random Forest, XGboost, Gradient Boosting and DNN, were examined and compared in order to determine which one would provide the maximum accuracy. Grid search with five-fold cross validation was used after the first run of the ML models, and hyperparameter tuning was utilized on the models to discover the optimal parameters.

Decision Tree

The Decision Tree Regressor was run with the following parameters: maximim_depth=9, maximum_features=25, minimum_samples_leaf=2, minimum_samples_split=7, and n_jobs=-1. Scikit-Learn API and the model.fit() function are used to fit the model. Table 4 illustrates the outcomes that were attained.

Table 4. CLV Prediction with Decision Tree for New Users

Model Results			
	MAE	MSE	RMSE
Decision Tree	2.75	8.83	2.97

Random Forest

After splitting the datasets, in order to identify the most relevant features that contribute most to CLV prediction, feature selection is performed on the training set before moving on to the testing set for evaluation. To train our model, we employed these features. Figure 8 shows the feature importance graph for Random Forest.

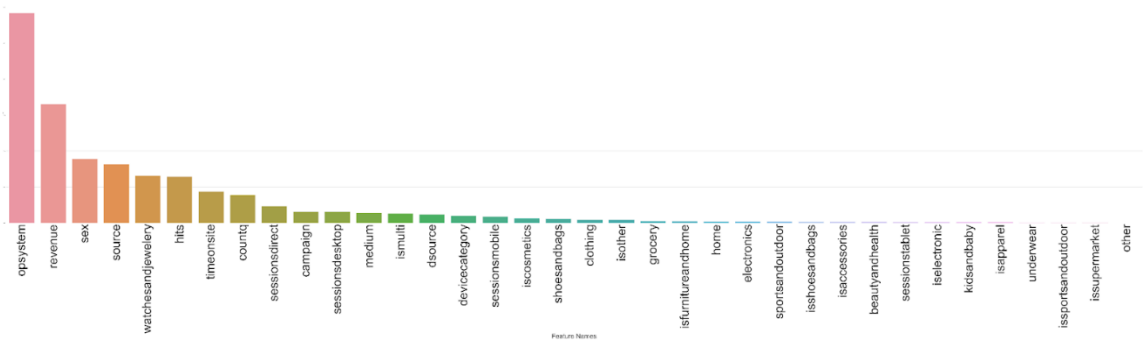


Figure 8. Random Forest feature importance

We started by training our model with the RF Regressor algorithm's default settings for its hyperparameters. To learn the model's parameters, the training target and features are utilized. To make predictions about the target variables in the testing phase, the model for testing features is tuned using the parameters learned in the training phase. After that, the best hyperparameter values of Random Forest are determined via 5-fold cross validation using grid search as stated above. For this RF Regressor run, we used the following parameter settings: maximum_depth = 9, maximum_features = 25, minimum_samples_leaf = 2, minimum_samples_split = 7, and n_jobs = -1. Table 5 shows the results that were obtained:

Table 5. CLV Prediction with Random Forest for New Users

	Model Results		
	MAE	MSE	RMSE
Random Forest	2.75	8.81	2.97

Gradient Tree Boosting

Initially, this model is trained with the Gradient Boosted Regressor and its default hyperparameters. To learn the model's parameters, the training target and features are utilized. For the testing phase, the model's testing features are used to make predictions

about the target variable using the parameters learned during training. After that, the best hyperparameter values of Gradient Boosted Regressor are determined using Grid Search 5-fold cross validation. Maximum depth was set to 12, maximum features were set to 28, minimum leaf sample size was 2, minimum split sample size was 9, and number of jobs was set to -1. The results obtained can be found in Table 6.

Table 6. CLV Prediction with Gradient Boosting for New Users

	Model Results		
	MAE	MSE	RMSE
Gradient Boost	2.75	8.79	2.97

XGBoost Regressor

The xgboost package was imported and XGB Regressor was used to for the model. The parameters used for the XGB Regressor for this, are set as the learning rate to 0.1, the maximum depth to 9, the minimum child weight to 4, the number of estimators to 70 and n_jobs=-1. Table 7 shows the result for XGBoost Regressor.

Table 7. CLV Prediction with XGBoost Regressor for New Users

	Model Results		
	MAE	MSE	RMSE
XG Boost	2.74	8.78	2.96

5.2 Hybrid model results for new users

Since predicting CLV for every user of a company is our primary objective, we also used this particular method to predict CLV for new users. We used the two stage Hybrid model as one of the main models while predicting CLV for new users. The first stage predicts the likelihood of purchases where we used Logistic Regression as our binary classification method for predicting the user whether to purchase for the specified time

window or not. Then, in step two, the CLV is forecasted for consumers who were anticipated to buy in stage one. Table 8 summarizes the result of the first model which is Logistic Regression.

Table 8. Logistic Regression Phase One of Hybrid Model

Model Results			
	precision	recall	f1-score
0	0.57	0.60	0.58
1	0.58	0.56	0.57
accuracy			0.58
macro avg	0.58	0.58	0.58
weighted avg	0.58	0.58	0.58

Then the next step is using the outputs with the other models. The output of the logistic regression those who were predicted to purchase was used with the hybrid ML models and DNN.

5.3 Comparison of model results for new users

The results described in this section allow to determine which models is the most accurate in CLV for new users. DNN model was shown to be good and stable for multi category e-commerce enterprises while predicting CLV, as it outperformed the other selected models consistently across all evaluation metrics.

Performance of the Hybrid and ML models summarized in Table 9. Results are evaluated using the Mean Absolute Error (MAE), Mean Standard Error (MSE), and Root Mean Squared Error (RMSE) as outlined in the previous chapter under Performance measures.

Table 9. ML and Hybrid Model Comparison for New Users

	MAE	MSE	RMSE
DT	2.74	8.83	2.97
RF	2.73	8.81	2.96
XGBoost	2.72	8.77	2.95
GB	2.73	8.79	2.96
DNN	2.67	2.93	1.71
Hybrid DT	2.43	7.96	2.82
Hybrid RF	2.43	7.93	2.81
Hybrid GB	2.44	7.88	2.80
Hybrid XGBoost	2.42	7.83	2.79
Hybrid DNN	2.42	2.83	1.68

The Hybrid DNN model's MAE and RMSE values were found to be 2.42 and 1.68, respectively, and they represent the best performance metrics. Significant differences between DNN and other ML models were found, and Hybrid DNN consistently outperforms all the other models. DTs were shown to have the lowest MSE and MAE values in both the ML and Hybrid ML models.

Calculating the hit-ratio as one of the performance metrics was also in scope of this study. With this strategy, we want to demonstrate that in addition to the aforementioned metrics, the model created for new users is also capable of differentiating high-value clients from other clients.

The hit ratios of the ML models and the hybrid ML models were compared and are shown in Table 10.

Table 10. Hit-ratios (Top 20%) for ML and Hybrid Model Outputs

	Hit-ratio
DT	0.31
RF	0.34
XGBoost	0.25
GB	0.29
DNN	0.38
Hybrid DT	0.30
Hybrid RF	0.40
Hybrid GB	0.39
Hybrid XGBoost	0.41
Hybrid DNN	0.61

By comparing the predicted and real CLV of top 20% of new users, the hit-ratio for CLV prediction was obtained. The hybrid DNN model's hit-ratio outperformed other ML models. The DNN model was followed by the hybrid DNN model in terms of hit-ratio. The Hybrid DNN model performed well, with an accuracy rate of 61%.

The main findings showed that DNN models perform better for new users than other ML models across all performance measures, and it is recommended that multi-category e-commerce businesses employ hybrid DNN models to forecast the CLV for new users.

5.4 Customer segmentation for new users

Similar to the case of new users requiring separate prediction techniques, they also need to be segmented differently than existing users. In this case, only CLV values are used for segmenting new users. First, the users who are predicted as they are not going to

make another purchase excluded from the segmentation and they counted as a separate segment called ‘not coming back’ segment. Then the rest of the group segmented by using k-means clustering and elbow method. The dataset was separated into two clusters. Table 11 shows the final cluster centers for the two groups and the excluded “not coming back” segment.

Table 11. Cluster Results for New Users

Cluster No	Value	Number of Clusters
High Likely to Repeat	3866.98	387,096 – 27%
On the Edge	1693.75	699,554 – 50%
Not Coming Back	867.20	308,297 – 22%

As illustrated in Table 13, most of the new users are “On the edge”. They purchased once from the company and the amount of their spending is not high. The second biggest group with the highest amount of value is “High Likely to repeat” group. They spend a lot and are likely to become repeat purchasers. The last group is the one as predicted not likely to purchase from the company. They are churners and ‘Not coming back’.

CHAPTER 6

EXISTING USER MODEL RESULTS

This chapter presents the model details and results of the current study for existing users. We begin this chapter with the different models that have been used in this study. Next, we report the results of each different model for existing users. We conclude this chapter with the comparative analysis of the results of the models for existing users.

6.1 Pareto/NBD models

After preparing the data and calculating its recency, frequency, monetary value, we fit the data using the Python beta geo fitter (BGF). The BGF module which is based on the Pareto/NBD model is used. This model forecasts the count of anticipated future transactions.

The Pareto/NBD model forecasts the quantity of future purchases, but not their spend value, which is necessary to predict the CLV. The value was predicted using the extended Pareto/NBD Gamma Gamma submodel, which forecasts the average spent per transaction by each consumer. By multiplying the average amount spent and predicted number of future purchases, we can forecast the CLV for each consumer.

Table 12 below shows the result of the model.

Table 12. Pareto/NBD Model Results

	MAE
Decision Tree	1.50

6.2 ML models

Since testing every potential parameter value combination for every model would take a very long period, certain parameters are changed while others are kept alone. The most common parameters for each kind of model are chosen based on previous research (Lee et al., 2022).

After evaluating the effectiveness of the machine learning models and selecting the ideal hyperparameters, we looked at and compared model accuracy. To do this, we employ Grid Search with 5-fold cross validation. Then, we evaluated and compared the accuracy of every model on the unknown test set using the best hyperparameters we could find.

6.2.1 Different ML models for existing users

Linear Regression

The second model to come to a right CLV determination is a model that predicts the revenue of the future periods. The dependent variable of this model is therefore a continuous variable, namely total revenue after the first purchase. To model the impacts of several independent factors on a continuous dependent variable, a multiple linear regression is suitable. A multiple linear regression presumes that one dependent variable Y, relies on multiple independent variables X. The dependent variable is the total revenue after the first purchase, the independent variables are the ones that were explained above.

Decision Tree (DT)

A DT is a supervised ML approach that forecasts a dependent variable using a tree-like design based on numerous input variables. Based on the kind of dependent variables used, there are two different forms of decision trees. Classification trees are concerned

with discrete dependent variables, although regression trees are focused on continuous dependent variable. The optimal separating factor and split point are determined differently for these two tree types. Due to the continuous nature of the response variable in this research, regression trees were employed in this study. Maximum_depth=9, maximum_features = 25, minimum_samples_leaf = 2, minimum_samples_split = 5, n_jobs=-1 were the parameters used for the Decision Tree Regressor. Table 13 shows the results for Decision Tree model for existing users.

Table 13. CLV Prediction with Decision Tree for Existing Users

Model Results			
	MAE	MSE	RMSE
Decision Tree	0.63	0.81	0.90

Random Forest (RF)

A supervised machine learning algorithm called random forest can be applied to both classification and regression problems. Random Forest regression was used in this thesis because the study's response variable is continuous.

After dividing the datasets into training and testing, a feature importance table is generated. The training set is then used to select the features that are most closely related to and contribute to our predictand variable.

The importance values of the features are mapped in Figure 9 below. These variables were leveraged while training model.

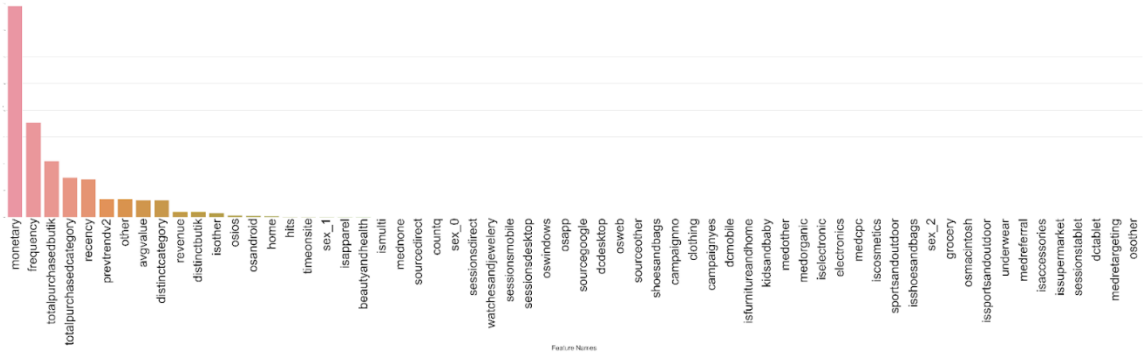


Figure 9. Random Forest feature importance for existing users

The RF Regressor algorithm is initially used to train the model, using the default hyperparameters. To learn the model's parameters, the training target variable and features are used. The model for testing features uses the parameters from the training period to predict the target variable for the testing period. After that, grid search 5-fold cross validation is used to identify the best Random Forest hyperparameter values for optimum accuracy. The RF Regressor's parameters were as follows: n jobs=-1, minimum_samples_leaf = 2, minimum_samples_split = 7, maximum_depth = 10, maximum_features = 25. The results that were obtained can be seen in Table 14.

Table 14. CLV Prediction with Random Forest for Existing Users

	Model Results		
	MAE	MSE	RMSE
Random Forest	0.63	0.80	0.89

Gradient Boosting

We trained the model with the Gradient Boosting Regressor algorithm and its default hyperparameters to see how well it performs. To learn the model's parameters, the training objective and features are utilized. To get the ideal hyperparameter settings for Gradient Boosting Regressor, we employed Grid Search 5-fold cross validation.

Maximum depth was set to 9, maximum features were set to 20, minimum leaf samples were set to 2, minimum split samples were set to 5, and the n_jobs was set to -1. Results the one can be seen in Table 15.

Table 15. CLV Prediction with Gradient Boosting for Existing Users

	Model Results		
	MAE	MSE	RMSE
Gradient Boost	0.62	0.77	0.88

XGBoost

It is an ensemble Machine Learning method built on decision trees, and it performs well when dealing with both classification and regression problems. It offers excellent performance in a short amount of time with good precision. The model was classified using the XGB Regressor and XGB Package. The learning_rate = 0.1, maximum_depth = 9, minimum_child_weight = 4, n_estimators = 70, and n_jobs = -1 parameters were utilized for the XGB Regressor. Table 16 shows the results for XGBoost Model for existing users.

Table 16. CLV Prediction with XG Boost for Existing Users

	Model Results		
	MAE	MSE	RMSE
XGBoost	0.62	0.77	0.88

6.3 Comparison of model results for existing users

Comparison of the ML Models vs Pareto/NBD Gamma Gamma Models

Model evaluation metrics that were used to compare the performance of different models are by MAE, MSE and RMSE. The MSE and RMSE calculations look quite similar, where the RMSE is the rooted formula of the MSE. Both measures assess the

error in different ways. An optimization of the MAE seeks a value that splits the data in two equal parts, so it seeks the median forecast. This is different for the MSE and in turn the RMSE. Optimizing the MSE aims to give a prediction that is on average correct. In summary the MAE aims at the median prediction and the (R)MSE at the average prediction.

Performance results for the Pareto/NBD and ML methods are summarized in the table below. When considering the median prediction as most important, RF, XGBoost, Gradient Boosting and DNN have identical performances. However, when considering the prediction on average, concerning the (R)MSE, the DNN performs better. Also, as can be derived from the Table 17, Pareto/NBD performed worst across all metrics which shows that when there are more input features ML models outperforming the probability models because of their power in capturing the non-linear relationship between input variables.

Table 17. ML and Pareto/NBD Model Comparison for Existing Users

	Model Results		
	MAE	MSE	RMSE
Pareto/NBD	1.50	5.44	2.33
DT	0.63	0.86	0.92
RF	0.62	0.86	0.92
XGBoost	0.62	0.84	0.91
LR	0.82	2.36	1.53
GB	0.63	0.83	0.91
DNN	0.61	0.82	0.90

In addition to the comparison above, the proposed multi-output DNN model, which predicts CLV, DPC, and TAS simultaneously, was also compared to a single-output

DNN model created for each of the output models independently, as well as to multi-output DT and multi-output RF models, using the performance evaluation metrics indicated above. Table 18 demonstrates that the multi-output DNN model performed the best across all metrics. The results demonstrated the significance of the association between CLV, the DPC, and the TAS. It yields better results when the output variables are predicted simultaneously, and the suggested DNN method captures this connection more accurately.

Table 18. Multi-Output and Single-Output Model Results Based on MAE, MSE and RMSE

	Model Results								
	CLV			DPC			TAS		
	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
Single-Output DNN	0.50	0.44	0.66	0.31	0.26	0.51	1.76	12.38	3.51
Multi-Output DNN	0.48	0.39	0.62	0.11	0.02	0.14	1.51	10.74	3.27
Multi-Output DT	0.56	0.63	0.79	0.13	0.03	0.17	1.58	12.70	3.56
Multi-Output RF	0.52	0.45	0.67	0.14	0.03	0.17	1.52	11.09	3.33

In addition to the aforementioned metrics, we wanted to demonstrate that the model could discriminate high-value clients from other consumers. We chose the hit-ratio metric because the ability to identify high-value consumers is crucial when segmenting e-commerce company users into numerous groups.

In this research, multi-output DNN model hit ratio was compared to single-output DNN model hit ratio, multi-output DT model hit ratio, and multi-output RF model hit ratio. The hit ratios for the CLV, DPC, and TAS forecasts are shown in Table 19.

Table 19. Hit-ratios (Top 20%) for Model Results

	Model Results		
	CLV	DPC	TAS
Single-Output DNN	0.70	0.30	0.28
Multi-Output DNN	0.71	0.64	0.42
Multi-Output DT	0.63	0.41	0.28
Multi-Output RF	0.68	0.32	0.41

Hit-ratio for the three outputs was predicted by comparing the forecasted top 20% of consumers to the actual top 20% for each output variable. As can be shown in Table 20, the multi-output DNN model's hit-ratio was higher than that of any of the individual models. The highest hit-ratio was found for CLV, followed by DPC and TAS. In particular, the multi-output DNN model for CLV failed to identify 29% of the top 20% of customers by CLV. The multiple-output DNN model did quite well, with an accuracy of 71%.

For the multi-output DNN model's forecasts of DPC and TAS outputs, the hit ratios were 64% and 42%, respectively, indicating weaker results compared to CLV prediction. However, the percentage of customers whose genuine CLV, DPC, and TAS are in the top 20% based on the computed values was less accurately predicted by the single-output DNN model than by the multi-output DNN model. CLV, DPC, and TAS hit rates for single-output models were 70%, 30%, and 28%, respectively. However, single-output DNN outperformed multi-output DT and RF in CLV prediction. We also used this hit-ratio criteria to compare the outcomes of our suggested model to other studies that had also used it. CLV prediction using a hit-ratio rule (20-55) was established by Malthouse and Blattberg (2005). Based on their calculated CLV, they separated their customers into four groups of roughly the same size. They applied the same thing to their clients' estimated CLV. The hit ratio was then computed by

determining the proportion of customers whose expected and actual CLV clusters are identical. What they discovered was that roughly 45% of the true top 20% would be correctly ranked. Chiang and Yang (2018) similarly utilized the hit ratio to assess the accuracy of their CLV model prediction. 35% of the top 20% of most valuable clients (top 20%) could not be recognized by their CLV prediction algorithm. Their study results exceed previous research with a correct classification ratio of 65%. Based on the outcomes, our multi-output DNN model's hit ratio for the CLV output, which has a 71% true classification rate, is highly competitive in comparison to Malthouse and Blattberg's (2005) and Chiang and Yang's (2018) research. Table 20 displays the comparison between our suggested model and the two-prior research.

Table 20. Hit-ratios (Top 20%) for Multi-Output DNN Model, Malthouse and Blattberg's (2005) Study and Chian and Yang's (2018) Study

	Multi-Output DNN Model	Malthouse and Bratterbeg's Model	Chiang and Yang's Model
Hit-ratio	0.71	0.45	0.65

Calibration plots (decile charts) were also included, which are graphs of fit that show the expected probability and the percentage of correct labels (Wang et al., 2019). As an example, if it is expected that 20% of our customers to be high-value customers, then the frequency of high-value consumers in the data should also be close to 20% (Wang et al., 2019). It is expected that the forecasts will fall along the 45-degree line.

In order to examine the discrepancy between average prediction and average label, we used decile charts to depict the labels of forecasts by decile.

If the model has been calibrated well, the predicted mean for each decile of predictions should be near to the true mean.

In Figure 10, we can see how the highest performing multi-output DNN model and the baseline single-output DNN models calibrated their respective three outputs. As can be seen in the graphs, the multi-output DNN model's predicted CLV values were quite close to the observed values. Predictions from the DPC for the highest and lowest 10% of ratios in the multi-output model have worsened. However, the single-output DNN model's predictions were higher than the actual values for the DPC. The trend varied from negative to zero to positive. Both the single- and multi-output DNN models had trouble calibrating the trend, especially the single-output model and the top and bottom 10% of the distribution. The trend output variable's anticipated mean, however, is very near to the actual mean, as predicted by the multi-output DNN model.

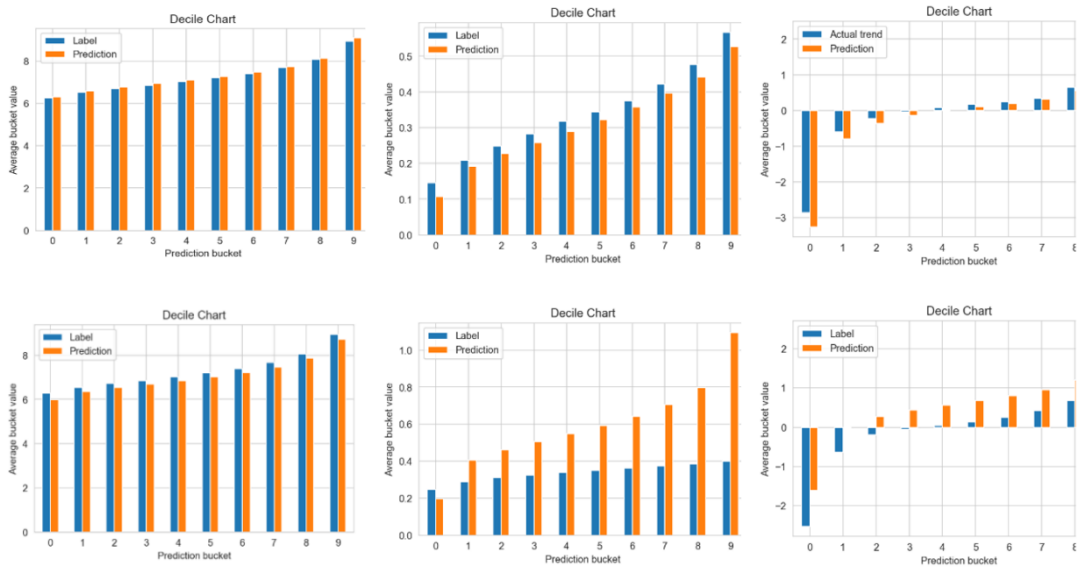
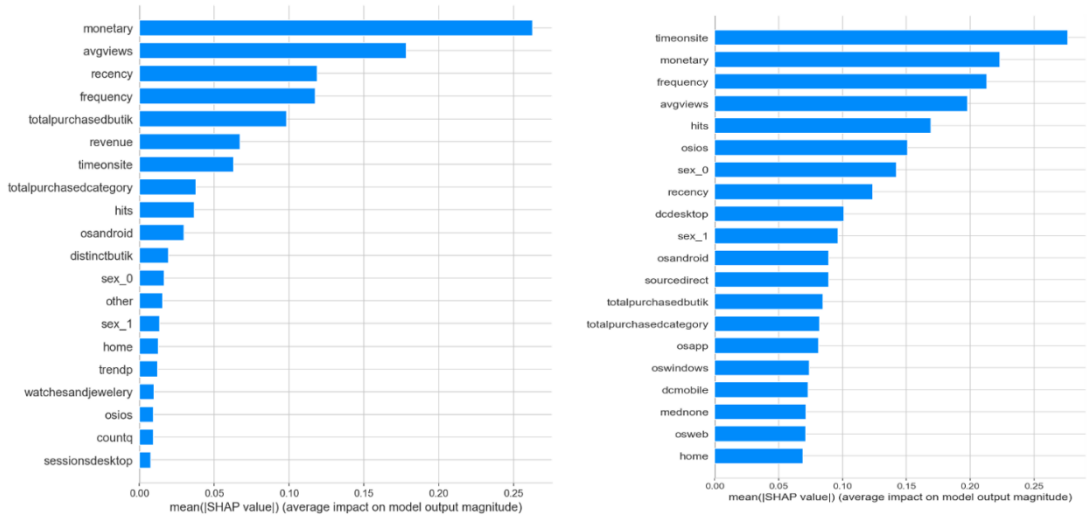


Figure 10. Decile charts for CLV, DPC, and TAS respectively (top row: multi-output model, bottom row: single-output model result).

Figure 11 depicts the SHAP feature significance charts for the CLV output of the multi-output DNN model and the single-output DNN model. Even though both models predicted the same outcomes, the multi-output and single-output models had distinct key values. According to multi-output models, the RFM variables recency, monetary value, and frequency are the most relevant for predicting CLV. The single-output approach, on the other hand, focuses higher emphasis on online behaviors, such as duration spent on site and quantity of interactions on the website. Product category and seller information are of more value in the multi-output model, but these variables are not so relevant for the single-output model. This highlights the significance of association between the many outputs.



(a) (b)

Figure 11. Feature importance levels defined by average SHAP values for CLV (a) multi-output model CLV SHAP values, (b) single-output model CLV SHAP values.

Appendix B's Figure B1 shows how the multi-output model for the DPC ratio relied heavily on the total number of previously purchased distinct categories from which customers made purchases (*distinctcategory*). However, for the single-output model, website behavior, namely *timeonsite*, was once again of more value. Also there were different metrics for website performance (such as *hits* and *timeonsite*) ranked in different orders of priority. In the single-output model, these factors were more important than in the multi-output model. Furthermore, seller information and previously purchased distinct product categories are unquestionably important for both models, but these are more significant to multi-output models, demonstrating the power of predicting these features simultaneously, especially for multi-category firms, which have many product categories and sellers. Furthermore, the models' ability to predict DPC indicates that the operating system (*osios* and *osandroid*, whether acquired from app via IOS or Android operating systems) is crucial. Customers who use a company's mobile application are more likely to make purchases across different categories and are more loyal to the brand overall (Dirsehan & Cankat, 2021).

The TAS is depicted in Appendix B, Figure B2. Despite initial impressions that the multi-output algorithm learned a subset of data differently than the single-output model, both models surprisingly placed significant importance on very similar input variables, with the single-output model's highest value going to seller information. Both models found that the length of time a user spent on the company's website, the level of user engagement, and trends of the previous transaction amount were the most relevant factors in determining whether a user will exhibit a rising, declining, or constant spending trend in future purchases.

Every output had its own feature importance table, as shown in Appendix B (Figure B1), and Appendix B (Figure B2). The model parameters for DNNs have been a mystery for decades, thus this is generally a good practice and a solid feature that is helpful to reveal them.

6.4 Customer segmentation for existing users

Our analysis of the existing user base utilized k-means clustering to categorize users according to their CLV, DPC, and TAS. Using the elbow technique, the data set was divided into five groups. The final clusters centers are shown in Table 21.

Table 21. Cluster Results

Cluster No	CLV	DPC	TAS	Number of Clusters
1	4386.93	0.46	-0.12	412,096 – 30%
2	3362.35	0.39	17.23	18,280 – 1%
3	2647.50	0.18	-19.83	12,897 – 1%
4	1073.78	0.30	0.03	612,885 – 45%
5	551.57	0.61	-0.08	319,291 – 23%

Thirty percent of customers fell into Cluster 1 (High-Value Watchout), which included individuals who spent the most overall and had the second greatest different product category ratios but a declining purchase amount trend. Nearly 45% of all consumers belong to Cluster 4 (Promising Purchasers), which contains clients with a moderate level of spending, the second-lowest percentage of purchases from diverse product categories, and a consistent spending pattern.

23% of consumers fall towards Cluster 5 (Low Value Distinct Product Lovers). This group of buyers has the steadiest spending patterns and spending less overall despite making purchases from a wider variety of product categories with a higher DPC ratio.

Finally, 2% of all consumers fall into either cluster 2 (Curious Spenders) or 3 (Risky). The two of them have the second- and third highest spend, respectively. The highest and lowest trend values, respectively, are also seen in these two groups. The highest rising trend is shown in cluster 2, where buyers initially spend a small amount before making large subsequent purchases. In contrast, Cluster 3 exhibits the steepest declining tendency, with a high amount of first purchase spend followed by a significant slowdown in subsequent ones.

As a summary, we show that when compared to the Pareto/NBD and other single-output ML models, the single-output DNN model in which we only predicted CLV had significantly better outcomes. The impact of predicting two additional outputs in addition to CLV was then examined by multi-output DNN model being compared to single-output models and multi-output DT and RF models. We found that the multi-output DNN model performed better than the single-output DNN model, multi-output DT model, and multi-category RF model across all evaluation metrics including hit-ratio and as a result, it is dependable and appropriate for multi-category e-commerce businesses. Moreover, clustering based on the findings of the multi-output DNN model may give multi-category e-commerce firms beneficial guidance.

CHAPTER 7

DISCUSSION AND CONCLUSION

This chapter will provide a detailed analysis of the study`s findings in two sections. In the first section, we explain how we divided the framework into two and contributed to the understanding of the best models to predict CLV for every user type (new and existing users), the second will elaborate on how and why new features and methods required for developing better marketing strategies as a holistic approach and how these methods are different from the previously developed methods while predicting CLV. These sections explain how we see the results from a theoretical and an empirical point of view. The focus of this chapter is on how well the current study matches (or does not match) the research we looked at in the literature review. Then the chapter concludes the study by focusing on the research and business implications. We continue with the limitations of this study. The study concludes with recommendations for further studies and a discussion of this thesis's contribution.

The first part focuses on developing the right methods that captures the CLV for every user of a multi-category e-commerce company. Managers need to look at both the value of new and existing users to figure out the overall worth of their users and the worth of the company. Having a holistic CLV prediction approach also important to help with the strategic decision making which will provide businesses with competitive advantage. For this, we first introduced different user types and different models specific to both new and existing users separately to understand the best method to predict the CLV. This strategy aligns with Kitchens et al. (2018) who stated the need of a holistic approach that will not only take into account users with nultiple transaction bu also will

focus on users with single transaction. The second part mainly focuses on the managerial impact and how companies can leverage the CLV to segment their users that will provide them with a competitive advantage within online retailers. This may boost their revenues by identifying the most lucrative clients and devoting marketing resources to them, as well as by constructing a distinct marketing strategy to develop strong, long-lasting connections.

Using CLV, a company can see its customers in a way that is both profitable and sustainable over the long term. But so far, previous research hasn't paid much attention to the idea of applying CLV to every user as a whole, instead focusing mostly on users who have made multiple purchases in the past. We focused on predicting CLV for every user of the company to have a holistic approach to provide a strategic decision-making approach to company managers in contrast with the researchers mostly focused on predicting the CLV for existing users who had already made several purchases (e.g., Chen, 2018; Fader et al., 2005a; Jasek et al., 2018; Salehinejad & Rahnamayan, 2016; Schmittlein et al., 1987; Singh et al., 2020). We proved that having a holistic approach and knowing the overall companies' worth is a powerful tool for the management of multi-category e-commerce companies in line with study of Bolton et al. (2004) and Kumar and Shah, (2009).

Also, not taking into account the product-related information and the amount spent with each purchase while predicting the CLV has been an issue, especially for companies that are selling products within several categories. Additionally, not having a customer relation based decision-making system that covers every user of a company is a limitation of prior studies. By offering a structural method to measure the CLV metric of both new and current users, considering the difficulties of model development for

different user types as well as by considering the importance of other factors while predicting CLV, enhancing the relationship marketing theory (RMT) with additional constructs, segmenting the users accurately and leveraging this output for decision-making we are making an effort to close these gaps. The research presents significant findings by bridging the gap between CLV and the multi-category e-commerce retailer context for decision making.

Even though CLV has been adopted by several industries, with a focus on relationship marketing, it is still a problem that not every type of user is taken into account. Managers of multi-category e-commerce companies struggle to develop a long-term marketing plan with a focus on long-term profitability when they only use CLV predictions for some of their users or only use CLV predictions and not other factors. In this study, we solve this problem by suggesting a 360° holistic approach where we are focusing the prediction of CLV for all customers of online retailers that are selling products from multiple categories to quantify individual customer lifetime value. We apply our modeling approach to transaction data and website and app activity data acquired from a multi-category e-commerce firm. The data we used is based on RMT constructs and we derive insights from them to enhance the management's decision making.

We focus on the different types of users in the first part of the framework. We found that even though the ratio of existing users is high compared to new users the amount of new users in a company is also substantial and predicting both new and existing users' CLV provides the company with a holistic approach. The framework we used proved that a separate focus is required for different user types to predict CLV. However, this result contradicts with Chamberlain et al. (2017), Vanderveld et al. (2016) and Wang et al.

(2019). For the new users, since there are limited features that can be used for the modeling, we only used website and app behavior and the first purchase related information for the models that we tested and compared. Also because of the same limitations we implemented ML models and Hybrid ML models for new users unlike the existing users where we implemented traditional probabilistic models and multi-output models in addition to the ML models. For the new users the results showed that the best performing model is the Hybrid DNN model when evaluated with MAE and RMSE. ML models performed poorly when evaluated and compared with Hyrbid ML models since there were limited input features extracted from the website behavior and there were users with zero CLV.

We discovered that the hybrid DNN model designed for new users had the potential to separate valuable clients from regular ones using the hit ratio. For example, hit-ratio for hybrid DNN is the highest with 61%, but this number was much lower with 41% for the second best Hybrid XGBoost model. These types of results are very helpful for the managers of multi-category ecommerce companies who can now design the best system that will predict the CLV for new users and allocate marketing spend accordingly by segmenting and differentiating their first-time purchasers. Comparing our results to those of previous research (Chamberlain et al., 2017; Vanderveld et al., 2016), we found that hybrid DNN models work best for new users. According to Chamberlain et al. (2017) with the right hyperparameters and enough hidden layers, DNNs perform similarly to Random Forest, and in our case, it even did better.

The main results showed that DNN models work better than other ML models for new users, and it was suggested that multi-category e-commerce companies should use Hybrid DNN models to predict the CLV for new users. Also, the segmentation was

based on the new users' total value. There are three clusters based on the results. The first one is "On the Edge" group. These customers can be turned into high value customers with the right methodologies such as remarketing. The second group with the highest amount of value is "High Likely to repeat" group. They spend high and may become a repeat purchaser with the right strategies. So, the company needs to focus on up-sell and cross-sell techniques and strategies for this group of users. Also remarketing to showcase other products and unique repeat customer discounts directed to this group of customers make them purchase again from the company. Last group is not likely to purchase from the company. Therefore, this group can be excluded from marketing strategies and companies instead focus on the other group of clients that have more room to grow through upselling and cross-selling rather than putting money into these clients. This overall information provides managers and enhanced decision-making capability to rethink and manage their strategies for each client and client segment separately. These results were consistent with the findings of Villanueva & Hanssens, (2007) and Kitchens et al. (2018).

The other part of the discussion is predicting CLV for existing users. For the existing users unlike the previous studies, we not only focused on finding the best method to predict CLV but also focused on other important factors such as product related DPC and customer related TAS that companies should also take into account to get a better understanding of the values of their customers and value of their company. This section was divided into four to address the research questions that were stated in the introduction chapter.

First, we compared the traditional probability models with the ML models to predict CLV for the existing users of e-commerce platforms. We not only used the past

transaction information but also used the website behavior data as our input features. As a result, we found many similarities with the studies which claim ML-based CLV metrics can outweigh probability models (Chamberlain et al., 2017; Chen, 2018; Salehinejad & Rahnamayan, 2016; Singh et al., 2020; Vanderveld et al., 2016). We found that ML models can take into account a lot of parameters or features, which could lead to better results with more accuracy.

Secondly, the proposed framework was developed, which enables us to forecast, in addition to CLV, the DPC and TAS. It was proved that the RMT constructs improves the customer-relations when combined with big data similar to what was stated in Kitchens et al. (2018). The two variables that were introduced literature as an extension of the RMT constructs added additional value while predicting CLV. CLV may be used to identify a company's top clients, but we found that this metric alone is not enough for companies that sell products in more than one category. In the past, most implementations for existing users only looked at how much a customer spent to figure out how valuable they were. This is because the main goal was to use CLV and customer-centered ideas. A significant gap is that the multi-category e-commerce companies need to consider the products they sell while predicting the CLV. Even now, most e-commerce managers only use CLV to measure how effective their marketing strategy is. Even though this approach has been viable in the past, it is unsustainable for companies selling products from multiple categories. By relying only on CLV for their existing users, multi-category companies are not deriving the best outcomes for their clients. As a new approach, we proposed forecasting DPC, TAS in addition to CLV. We found a lot of evidence that is consistent with the findings of Kitchens et al. (2018) who found that using RMT based constructs improve the relationship with customers. We leveraged relationship marketing theory

constructs to add on top of it and to implement the new variables DPC and TAS. DPC is closely related to "Choice" construct where the product category information is very important for relationship marketing. We evaluated this metric and proved that product category information increases the accuracy of CLV prediction. In addition, we extracted the trend information from the big data we have and added it as an extension to "transaction" construct of RMT. Transaction construct includes everything related to past purchases of each customer. We extracted the trend information in addition to RFM based variables and used that as the additional variable while predicting CLV. We showed that measuring these values in conjunction with CLV and segmenting clients based on these variables elevates the accuracy of CLV prediction in line with Kitchens et al. (2018)'s study. Moreover, we found that when these variables were predicted together, they gave more accurate results.

The other component was determining the best approach to predict CLV for a company that sells products from multiple categories. First, we implemented and compared ML models to predict only CLV. Then, to predict the CLV and other output variables together, we evaluated and compared various multi-output models, which take a single set of inputs and generate predictions for multiple variables. The new method for CLV prediction with multiple outputs has never been used before. This is a more effective technique to identify the correlations between output variables and is a robust solution for businesses who want to build and manage ML models more simply which is multi output model prediction. Multi output models have been used in some other research areas such as in the studies like Mishra and Passos (2022) and Sajjad et al. (2020). They showed that multi-output models outperform classic single-output deep learning models. Similar to their studies we discovered multi-output DNN models performed better

compared to modeling each output variable as a separate single-output model or compared to multi-output RF and multi-output DT model. Additionally, to develop an empirically valid model, we analyzed the hit-ratio output with the results in the prior studies. When the multi-output DNN model's performance was evaluated compared to that of Malthouse and Blatterberg (2005) and Chian and Chen (2017), it was also found to be very competitive.

Another important result we identified is that a concise interpretation of the results from the multiple-output DNN is essential in elevating the confidence levels of the managers to promote their use of DNN. Like what Carvalho, Pereira, and Cardoso (2019) said, providing a rationale behind the integrity of a model will underpin its functionality. We validated this hypothesis by explainable artificial intelligence (XAI) method, Shapley values methodology which was used to interpret DNN results. We discovered that companies might benefit from this given that they will know which factors are more important and the key drivers for increasing the value of their customers. This is also another addition to have an holistic decision-making system we are suggesting to the companies since they will have the visibility into the key drivers of high value users again which will provide them competitive advantage which is consistent with the Kitchens et al. (2018)’s findings.

The final component within the strategy is to take on a client segmentation approach for both user types separately. To segment new users, we used segmentation based on the value and divided the new users into three groups. This contrasts with the CLV-based customer segmentation approach taken by Benoit & Van den Poel, (2009), Kim et al. (2006) and Wu et al. (2020) for the existing users, where we divided them based on the output variables which involve the three outputs: DPC, TAS, and CLV. Additionally, we

discovered that the product and spend-related information for e-commerce institutions gives management a more accurate snapshot of their existing user values. Supporting the findings of González-Benito and Martos-Partal (2012), our results proved that making purchases from different categories is considered a form of loyalty and related to high spending. Also, spending more with each consecutive purchase does not always translate to high value users. Because of these reasons, using these two output features next to CLV is aggressively critical to understand the real value of customers. This is also crucial to build a more accurate strategic decision-making solution for the companies. Based on the results of the correct segmentation companies can allocate resources accordingly and apply marketing strategies more effectively. We found different groups of customers behaving differently and for those groups different strategies need to be implemented differently to each cluster. The study of Kitchens et al. (2018) aligns with our findings. As an example, Cluster 1 (High-Value Watchout), which included individuals who spent the most overall and had the second greatest different product category ratios, but a declining purchase amount trend is anticipated that they will spend much more than usual across a wide range of distinct product categories, however their spending is anticipated to gradually diminish over time. Therefore, multi-category e-commerce businesses may use retention-based approaches and strategies that are attractive to them, such as sending birthday and anniversary greetings, as well as customized price incentives and bulk promotions, to keep these crucial consumers around. Cluster 4 (Promising Purchasers) 4 is the company's core group, thus various strategies, such as discounts, should be put into place to prevent turnover. By using cross- and upselling techniques to increase the particular product category from which each user is buying, leading to a greater average order value, the multi-category e-

commerce company must establish a plan to attract these sorts of clients. Through further research, businesses may provide more appealing upselling and cross-selling potential for their target consumers.

Another group, Low Value Distinct Product Lovers are more loyal than other clusters according to the DPC ratio, it is clear that they can spend more and increase their buying trend with each successive purchase if the right strategies are put in place. Direct messaging may be used to cross-sell high value items in order to raise their spending with each consecutive purchase.

Finally, Curious Spenders and Risky groups have the second- and third highest spend, respectively, but they have low DPC and neative and positive trends respectively again. Multi-category e-commerce enterprises should evaluate the costs and benefits of retaining clients from Risky group, who may be more costly to serve than customers from Clusters 1, 2, 4, and 5. They may decide that Clusters 1, 2, 4, and 5 are more deserving of funding than Risky group. On the other hand, a smart customer support strategy, an emphasis on client service or not charging this group of consumers for product returns, might help to retain the Curious Spenders group, which is on the rise.

Finally, we demonstrated that the anticipated results for the DNN need to be leveraged in the segmentation effort instead of relying only on CLV. This discovery stands in contrast with prior research that clustered customers only based on CLV (Benoit & Van den Poel, 2009; Kim et al., 2006; Kumar, 2006).

In general, this study has made an important contribution not only to CLV prediction methods but also to the understanding of how differential factors and holistic apporach is needed by management of the multi-category e-commerce companies. For the former, it clearly shows the importance of holistic approach and modeling CLV for every users of

the companies not just a part of the users. Another important contribution to the former is to show how important it is to select the best performing model accurately.
It clearly shows the superiority of Hybrid models for new users.
For the latter, it showed that how the new suggested relationship marketing theory based factors DPC and TAS contribute to changes and to the selected algorithm. For instance, with our novel new output features and with multi-output DNN algorithm we attained an average improvement of 21% in CLV prediction. Similarly, these two output features and multi-ouput model prediction largely contributed to the performances overall CLV predictions and its impact on the segmentation of customers which provided an improved decision-making approach. Lastly XAI also added huge benefit to the adoption of DNN models. In short, this study developed a holistic, theory-driven CLV prediction methods using different algorithms for different user types. All these findings align with our results. The results are quite promising for the management of the multi-categories e-commerce companies as evident from the improved predictive performances of the models and a great guide for a better decision making that will provide competitive edge.

7.1 Conclusion

In this study, we created a novel 360° methodology for calculating CLV for online retailers dealing in a wide range of product categories in order to anticipate CLV and segment consumers in a variety of ways. Generating a solution like this can potentially elevate profits for organizations since they can direct their attention to high-level clients. Our study was motivated by two main components.

We first divided the users into two as new and existing users and build different approaches that predicts CLV for each user type differently. We tested this approach on a multi category e-commerce retailer's dataset that includes both the new (n=1,394,947) and existing users (n=1,715,519) in between 2019-03-31 and 2018-10-01.

The methodology starts with model selection which is applicable for each user type seperately and preparing the data accordingly where we can compare and decide on the model with the highest accuracy. For the new users where there is no historical information, we only used the first transaction information and website and app behavior so that we tested ML models and Hybrid ML models. The major findings of the model comparison for new users demonstrated that Hybrid DNN models outperformed the other ML models and Hybrid versions of these models. We also showed that the hybrid DNN model designed for new users could separate high-value clients from others according to the hit-ratio metric. For example, the hit ratio for hybrid DNN is the highest at 61%, but this number was much lower at 41% for the second-best Hybrid XGBoost model.

Next, for the existing users in addition to ML models we also tested Pareto/NBD model which requires information extracted from previous transactions such as recency, frequency, etc. We enhanced the RMT constructs by adding two new variables. Also, we evaluated these variables, DPC and TAS, through predicting next to CLV for the multi-category e-commerce companies. To prove that a combination of CLV, DPC, and TAS delivers more integral results, we utilized multiple-output ML models and measured the results against single-output techniques. This is the first-ever suggestion and testing of a multiple-output DNN methodology to predict CLV. The result from the study shows that multi-output approaches outperform single-output methods based on all of the

performance measurement features (MSE, RMSE and MAE) and predicting high-value customers (top 20%) compared to previous studies. The results revealed that predicting CLV together with DPC and TAS performed better. The association among CLV, DPC, and the TAS is essential, and this new method captures the relationships better by generating more efficient results using the multi-output DNN. The results demonstrated that the multi-output DNN model performed better than other models for all output variables. In comparison to single-output DNN modeling for CLV, DPC, and TAS, the multi-output DNN model that predicted CLV, DPC, and TAS simultaneously lowered the RMSE by up to 6%, 70%, and 7%, respectively. The results were more promising for the DPC values, which improved by a huge extent within the combined model. Moreover, the RMSE for CLV prediction was reduced by 21% compared to multi-output DT and 7% compared to RF methods. In addition, we used the hit ratio for these measures to illustrate that the model could differentiate high-value clients from other consumers. The multi-DNN model also produced consistent results when applied to predict the hit-ratios, resulting in a 71% ratio which indicated that there was only a 29% misclassification. The hit ratio revealed essential findings in that the DPC and TAS recorded poorer results at 64% and 41%. Additionally, we proved that such a single model could be easily managed, updated, and transferred by the management, which provides better advantages.

Another novel approach is the integration of XAI, Shapley values which was performed for the first time to our knowledge while predicting CLV. This led to the determination that the relative importance of the model variables changes between single and multi-output models. For example, for CLV prediction, the single-output model emphasizes the importance of web analytics metrics, whereas the multi-output model focuses more on the

product- and seller-related features. An alternative example is that even though the time spent at the website was vital, the most pertinent aspect across the categories was the different sellers that the customers shopped at. This shows that focusing on the seller variability may increase the products ordered.

Finally, we applied different segmentation techniques for different user types. For the new users we segmented the users into three groups as “On the edge”, “High Likely to repeat” and “Not coming back’”. “On the edge”. customers can be turned into high value customers with the right methodologies such as remarketing. “High Likely to repeat” group spend high and very likely to become a repeat purchaser. As a result, the company needs to focus on up-sell and cross-sell techniques and strategies for this group of users. Last group, ‘Not coming back’ is the one as predicted not likely to purchase from the company. They are churners therefore this group can be excluded from marketing strategies (Kitchens et al., 2018). For the existing customers, we proposed to categorize them according to the suggested variables: DPC, TAS, and CLV. We illustrated that the anticipated findings of our multi-output representation could be utilized to classify customers. The main result we have learned is that each cluster needs its own marketing plan and set of customers. For example, clients in Cluster 1 (High Value Watchout Group) were expected to have a high CLV and a high ratio of unique products purchased, but they would spend less in each consecutive purchases. To get these customers to spend more, managers of the companies can use up-selling and cross-selling techniques. In addition, the data demonstrated that clients spend more money when they purchase a greater variety of product categories. Therefore, while segmenting their consumer base, multi-category e-commerce enterprises should pay attention to both CLV and product information. Cluster 1 exhibited negative trend values, suggesting that

if management does not come up with particular tactics for this group, customers who spent a lot of money and purchased items from numerous categories may eventually cut down on their purchases. Therefore, multi-category e-commerce companies should closely monitor these important customers and may use retention-based techniques and strategies that appeal to them, like wishing them a happy birthday or happy anniversary. Cluster 4 (Promising Purchasers) is made up of customers who spend a moderate amount, buy products from the second fewest different categories, and have a stable spending pattern. The company should make sure that there is no turnover in Cluster 4. They must devise a strategy to acquire these clients by utilizing cross-selling and upselling to expand the number of product category each user purchases from which will raise the average order value. Through more analysis, companies should find more compelling ways to upsell and cross-sell to their targeted customers. Companies can utilize these results in conjunction with recommendation systems to discover upselling and cross-selling opportunities. Cluster 5 (Low Value Distinct Item Lovers) represents individuals who spend low. They have the most substantial DPC ratio and the steadiest spending patterns. Proper techniques will allow them to expand their procuring habits and spending since they have greater loyalty compared to other clusters according to the DPC ratio. Direct messaging should be utilized to elevate their steady buying and the amount of total spend by cross selling more expensive products.

Last but not least, clusters 2 (Curious Spenders) and 3 (Risky) make up 2% of all customers. Their average spending is second and third highest, respectively. The highest and lowest trend values are found in these two groups. With a modest first purchase spend and a sharp rise in subsequent purchases, Cluster 2 has the greatest increasing trend. On the other hand, Cluster 3 has the highest decreasing trend. This means that the

amount spent on the first purchase is high, but the amount spent on subsequent purchases goes down a lot. Customers in Cluster 3 may cost more to keep than customers in Clusters 1, 2, 4, and 5; hence, multi-category e-commerce company managers should use a cost-benefit assessment of these clients. They should prioritize allocating resources to the other clusters first instead of Cluster 3. Cluster 2 is the High-Value group, and they have an increasing trend, so a good customer support strategy such as prioritizing customer service (for example, giving a caller priority response on phone support or in support chats) or not charging this group of customers for product returns, could help to keep them. Finally, the primary contribution of this study is the proposal of CLV as a comprehensive financial indicator for strategic decision making. Having a clear view on the overall value of customers may facilitate more precise strategic decision making in company strategies and relationship marketing. Customers are often cited as an organization's most significant asset; hence, management should take into account the actual value of the customers.

In short, we developed a theory-driven 360° holistic framework in this study. The results we obtained were quite promising in terms of enhancing the existing methods that are heavily reliant upon past purchase data and just one type of users of the companies. In that sense, the developed system caters to management of multi-category e-commerce retailers for increasing the overall value of their businesses as well as achieving operational efficiency.

7.2 Research implications

There are several research ramifications of this thesis. The primary contribution of this thesis is building a new framework for multi category e-commerce enterprises to use a

360° holistic CLV projection method and to enhance their decision making strategies by leveraging the values of this framework. For example, it offered solid empirical support for the requirement of a holistic CLV prediction approach where the companies consider every user while predicting CLV, as implied by (Chamberlain et al., 2017; Vanderveld et al., 2016; Wang, 2019). Likewise, a specific algorithm may be significant for one type of user, but it can be completely irrelevant for another type of user in another context. Furthermore, the results demonstrated that while the Hybrid ML model works best for new users while predicting CLV, multi-output DNN model provides better accuracy for existing users. So not using the same type of model for every user and focusing on customized solutions while predicting CLV is a new area where researchers can dive in. We also discovered when determining high-value customers predicting only CLV for multi-category e-commerce companies might be misleading. Because of this we introduced two new factors to the literature, distinct product category (DPC) and trend amount spent (TAS) as the two novel factors to consider alongside CLV while predicting the expected customer lifetime value and defining user segments based on these three values. These factors are the enhancement to RMT constructs that was stated by Kitchens et al. (2018). This conclusion was consistent with Villanueva and Hanssens (2007), who similarly stated that although the purchased product category information and the change of the spend between each consecutive purchase are important indicators to calculate the actual value of the clients, most CLV-associated works had not succeeded in utilizing these components.

These results show how important it is to make a framework that includes these two new outputs and to use a multi-output model to predict all these values at once.

Another research implication is the fact that multi-output DNN algorithms are superior to single-output DNN, multi-output Random Forest, and multi-output Decision Tree when it comes to predicting CLV. When compared to developing a single-output model for each output separately or employing the popular Decision Tree (DT) and Random Forest (RF) algorithms, this model is more robust, simpler to manage and accurate in documenting the links between the three outputs (CLV, DPC, and TAS). As a result, this study emphasized the efficiency of multi-output DNN algorithms and suggests as a new approach in this CLV concept. Researchers can also leverage multi-output models in other research areas as well when there are multiple variables that need to be predicted. One last research implication is that this study demonstrates the feasibility of using XAI to boost user confidence in DNN models previously viewed as black boxes by providing an explanation of the model's results, the details of feature importance, and the model's performance. After leveraging this, the researchers can apply this to any DNN based study to help the explanation of the model.

7.3 Managerial implications

CLV-associated marketing techniques enable the enterprise to see the customer as a key resource of the company which allows them to evaluate the effect of marketing strategies on the customers. Our proposed 360° holistic CLV implementation aligns properly with the client-centered paradigm where the main goal is to meet the client's needs and improve their profitability. The 360° holistic CLV framework focusing on every user (new and existing) of companies, allows administrators to disparately align marketing spend to specific client clusters according to their profits. Due to this, various companies have been innovative in constructing marketing techniques to enhance profits. CLV-centered

marketing methods allow marketers to recognize their most vital clients and develop a close association focusing on retention. Regulating acquisitions and retention are examples of proper practices (Reinartz, Thomas, & Kumar, 2005). Additionally, to succeed financially, CLV may be used to make flexible decisions about resource distribution. When a CLV based framework is used, businesses may change their marketing efforts and spending to find the best marketing mix. CLV can also be used to engage clients in other ways, such as by sending direct messages to people based on what they value, giving vouchers or coupons to higher-tier clients, offering discounts based on customer value to keep customers longer, creating a personalized experience, such as making product recommendations based on customer value and DPC, creating a personalized website, and having customized support systems, such as giving priority to customers, among other things.

Adopting CLV as a whole is very important to the success of the things above. In this study in addition to CLV we introduced two important concept DPC and TAS to calculate CLV in a more accurate way by considering and capturing the product and spend related information similar to Yılmaz Benk et al. (2022).

The varied utilization of segmentation according to DCP, CLV and TAS includes choosing client service preferences and marketing strategy allocation. The management of the companies should formulate some strategies unique to each different clusters which helps to decide which customers to focus on in their campaign strategies or which group to exclude. As mentioned, our proposed model has clear benefits that managers should think about when evaluating their strategies.

In the context of a multi-category e-commerce firm, the multi-output modeling approach generated outstanding results by not only focusing on CLV but also unique product

categories and spending patterns for forthcoming purchases. Therefore, while installing and executing CLV models, it is essential to take into account CLV drivers as well as other important product-related elements. This is because product-specific data serves as management's go-to resource for developing cross- and up-selling tactics. This tactic is essential for keeping valuable customers and preventing churn. However single-output models didn't work well, and companies had to develop and maintain a lot of models with the same input variables, which was expensive and inefficient.

As can be seen from the results and comparisons provided, the multi-output DNN model outperformed the single-output DNN model, as well as the multi-output DT and multi-output RF models. Having a single model to manage is also more convenient and less expensive than having separate models for different outputs. Therefore, multi-category e-commerce business management that offer online shopping may use the multi-output DNN model as a better choice.

Lastly, previous studies have highlighted that CLV is one of the most useful techniques for enhancing the worth of a company. Customers must be seen as assets that produce income because customer interactions must be seen as investment strategies. The CLV-based metrics not only increase cash flow but also decrease the costs of client retention and attrition, hence boosting the value to shareholders (Stahl, Matzler, & Hinterhuber, 2003). Investors that respect the aforementioned qualities find customer-centric firms appealing because of these benefits. The use of CLV/CE-based measurements in financial reporting has been shown in prior studies to help achieve this goal for businesses. Sunder, Kumar and Zhao (2016) cited in their study that to better connect customer management with company aims and investor viewpoints, companies should disclose CLV to investors. CLV shows how healthy a company is, which makes

investors feel better about it. Another unique aspect of the research is that it not only quantifies CLV for every business type but also refers to other essential factors while predicting CLV. We also draw associations between the additional outputs that we proposed based on the RMT and customer segmentation. With this approach as a contribution of our study we enhanced the strategic decision making where executives can leverage the above information to determine (a) the future profitability potential of their organizations and (b) techniques to attract and retain high-value CLV clients, leading to sustained financial performance.

In conclusion, CLV is a metric that should be calculated for the overall company customers and different methods need to be implemented to calculate the CLV for each different customer type. Therefore, multi-category e-commerce companies would benefit greatly from taking CLV and other factors into account when making decisions. This would help them grow and stay ahead of the competition in the long run.

7.4 Limitations

There are several limitations with this research. First and foremost, we used real data from an e-commerce business which sells products from multiple categories. Due to this, the outcomes were reliant on this single entity which could mitigate generalizability. The study results ideally be backed by statistics from other multi-category e-commerce businesses to elevate generalizability.

Second, the number of model-related parameters we experimented with the hyperparameter tuning is another limitation of our study. The exponentially growing number of model configurations in our analysis is the main reason why we did not choose to experiment with additional parameters such as dropout rate, etc. because

testing with more hyperparameters is time consuming and costly. Therefore, we strongly encourage replication studies in different hyperparameter tuning settings. Finally, the study is limited to non-contractual businesses, which confines the study's generalizability to other industries. As a result, we also strongly encourage replication studies in different industry settings.

7.5 Future research

There is also a need for additional study to ascertain the theoretical and practical outcomes of incorporating the CLV and other attributes into customer-centric activities. Along with the CLV, researchers may look for new forecasting features that will help them predict and divide high-value clients. This study also examines how to forecast the ratio of various product categories. To determine which product categories should be cross-sold to each group, it would be beneficial for businesses to forecast category-based CLV indicators and integrate this data into their strategy.

The basis for our findings is analyzing a genuine dataset from a multi-category online e-commerce site and comparing several methods for calculating the CLV, among other things. As a follow-up to the current study, it would be beneficial to investigate multi-category organizations in other industries. Finally, because similar empirical analysis performed so well in this study, implementation of multi-output DNN models in other sectors are suggested as a future research path.

APPENDIX A

THE FEATURES EXTRACTED FROM THE DATASET

Table A1. Features extracted from the dataset.

Field Name	Type	Description
totalRevenue	FLOAT	Total transaction revenue.
sessionsDirect	INTEGER	True if the source of the session was direct (meaning the user typed the name of the website URL into the browser or came to the site via a bookmark).
sessionsMobile	INTEGER	The type of device (mobile).
sessionsDesktop	INTEGER	The type of device (desktop).
sessionsTablet	INTEGER	The type of device (tablet).
source	STRING	The source of the traffic source. Could be the name of the search engine, the referring hostname, or a value of the utm_source URL parameter. 3 new features generated from the source feature after preprocessing: sourceother, sourcegoogle, sourcedirect
medium	STRING	The medium of the traffic source. Could be “organic”, “cpc”, “referral”, or the value of the utm_medium URL parameter. 6 new features generated from the medium feature after preprocessing: medorganic, mednone, medretargeting, medreferral, medcpc, medother
campaign	STRING	The campaign value. Usually set by the utm_campaign URL parameter. 2 new features generated from the campaign feature after preprocessing: campaignyes, campaignno
browser	STRING	The browser used (e.g., “Chrome” or “Firefox”).
operatingSystem	STRING	The operating system of the device (e.g., “Macintosh” or “Windows”). 5 new features generated from the operatingSyste, feature after preprocessing: osios, osandroid, oswindows, osmacintosh, osother
mobileDevice	STRING	The brand or manufacturer of the device.
deviceCategory	STRING	The category of the device. 3 new features generated from the deviceCategory feature after preprocessing: osMobile, osDesktop, osTablet
country	STRING	The country from which sessions originated, based on IP address.
productCategory	STRING	The name of the product category purchased. 10 new features generated from the productCategory feature after preprocessing: clothing, kidsandbaby, home, grocery, beautyandhealth, shoesandbacgs, watchesandjewelery, electronics, sportsandoutdoor, other
productBrand	STRING	The brand associated with the product.
productPrice	STRING	The price of the product, expressed as the value passed to Analytics.

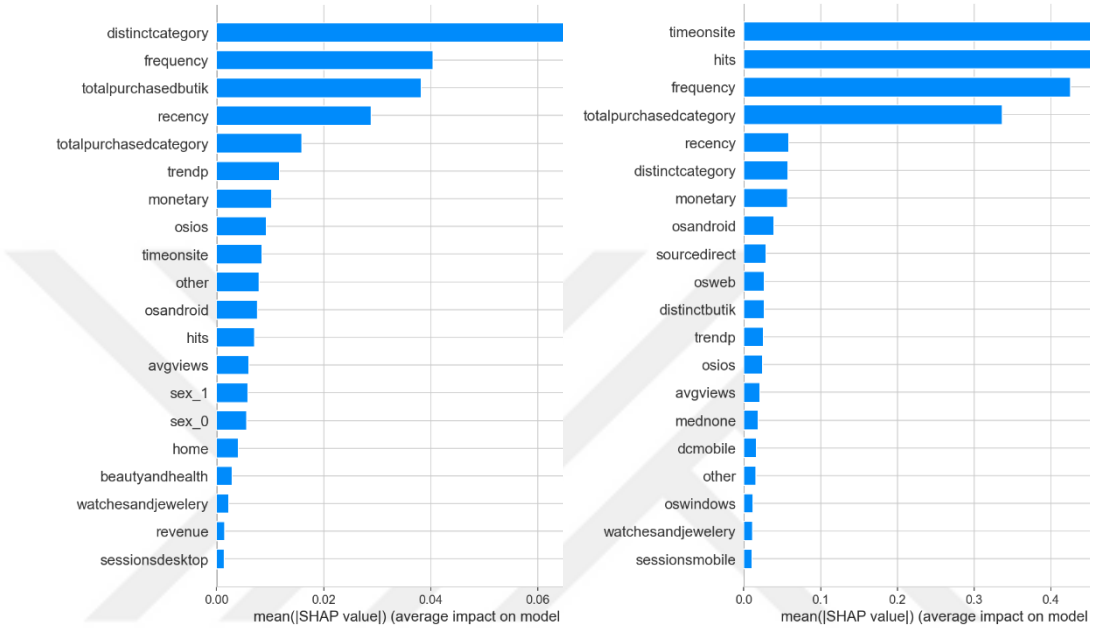
dataSource	STRING	Web or app. 2 new features generated from the source feature after preprocessing: dsapp, dsweb
Transactions	STRING	Total number of e-commerce transactions within the session.
hits	INTEGER	Total number of specific interactions between a user and your website within the session.
timeOnSite	INTEGER	Total time of the session, expressed in seconds.
sessions	INTEGER	The number of sessions. This value is 1 for sessions with interaction events. The value is null if there are no interaction events in the session.
seller	STRING	Seller that sells products under the multi category e-commerce company. 9 new features generated from the seller feature after preprocessing: isapparel, iselectronics, ismulti, isfurnitureandhome, isaccesories, issupermarket, isshoesandbags, iscosmetics, issportsandoutdoor, isother
totalsellerpurchased	INTEGER	Total number of distinct sellers from which clients purchased.
distinctProductCategory	INTEGER	Total number of distinct product categories from which clients purchased.
trend	INTEGER	Trend in amount spent with each successive purchase.
recency	INTEGER	The number of days between the customer's most recent purchase and the end of the observation period.
frequency	INTEGER	The overall number of repeat purchases over the observation period.
monetary	INTEGER	The mean purchase value.
Sex	INTEGER	Female—0, Male—1, Other—2.
countq	INTEGER	Number of products purchased
revenue	FLOAT	First purchase revenue

Table A2. Descriptive Statistics for Continuous Variables

	Descriptive Statistics										
	revenue	totalrevenue	sessiondirect	sessionmobil	sessiondesktop	sessiontablet	sessions	pageviews	hits	timeonsite	Countq
count	1,394,947	1,394,947	1,394,947	1,394,947	1,394,947	1,394,947	1,394,947	1,394,947	1,394,947	1,394,947	1,394,947
mean	205.47	474.92	0.85	0.73	0.37	0.01	1.11	47.07	253.88	2647.47	1.79
std	553.46	4663.51	0.58	0.62	0.55	0.09	0.37	54.93	348.51	2807.98	1.35
min	0.10	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00
25%	73.48	0.00	1.00	0.00	0.00	0.00	1.00	15.00	58.00	832.00	1.00
50%	121.99	104.00	1.00	1.00	0.00	0.00	.100	29.00	136.00	1824.00	1.00
75%	208.98	474.00	1.00	1.00	1.00	0.00	1.00	59.00	309.00	3486.00	2.00
max	398,008.99	2,785,607	14.00	14.00	8.00	6.00	14.00	3132.00	14,211	120,554	16.00

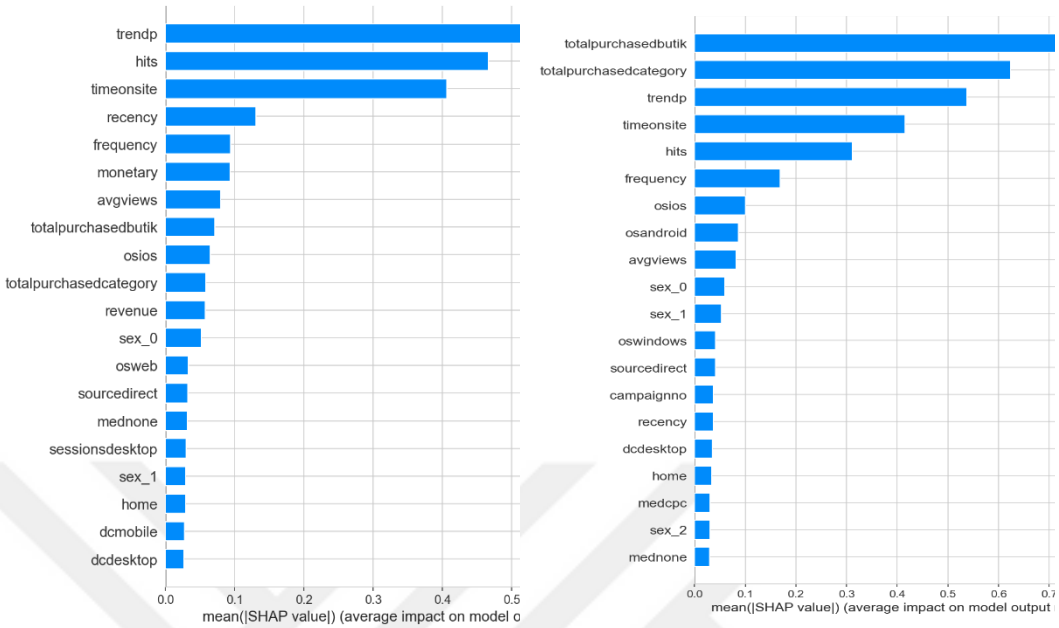
APPENDIX B

SHAP VALUES FOR DPC AND TAS



(a) (b)

Figure B1. Feature importance levels defined by average SHAP values for Distinct Product Category (DPC). (a) multi-output model DPC SHAP values, (b) single-output model DPC SHAP values.



(a) (b)

Figure B2. Feature importance levels defined by average SHAP values trend in amount spent (TAS) (a) multi-output model TAS SHAP values, (b) single-output model DPC TAS values.

REFERENCES

Abbasi, A., Sarker, S., & Chiang, R. H. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the association for information systems*, *17(2)*, 3.

Abdolvand, N., Albadvi, A., & Koosha, H. (2014). Customer lifetime value: Literature scoping map, and an agenda for future research. *International Journal of Management Perspective*, *1(3)*, 41– 59.

Abe, M. (2009). "Counting your customers" one by one: A hierarchical Bayes extension to the Pareto/NBD model. *Marketing Science*, *28(3)*, 541–553.

Belsky, J. (2017). Advertisers' biggest misconceptions about silicon valley's two giants. Retrieved November 10, 2021, from https://www.adexchanger.com/data-driven-thinking/advertisers-biggest-misconceptions-silicon-valleys-two-giants/

Benoit, D. F., & Van den Poel, D. (2009). Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services. *Expert Systems with Applications*, *36*, 10475–10484. https://doi.org/10.1016/j.eswa.2009.01.031

Berka, P. (2005). *Dobývání Znalostí z Databází*. Academia.

Bernat, J. R. (2018). *Modelling customer lifetime value in a continuous, non-contractual time setting* (Published Master's thesis). Erasmus University, Rotterdam.

Bhargava, D., & Gupta, L. K. (2022). Explainable AI in neural networks using Shapley values. In A. Khamparia, D. Gupta, A. Khanna, & V. E. Balas, (Eds.), Biomedical data analysis and processing using explainable (XAI) and responsive artificial intelligence (RAI). Intelligent Systems Reference Library, *vol. 222* (pp. 59–72). Singapore: Springer. https://doi.org/10.1007/978-981-19-1476-8_5

Bhatti, A., Akram, H., Basit, H. M., Khan, A. U., Raza, S. M., & Naqvi, M. B. (2020). E-commerce trends during COVID-19 Pandemic. *International Journal of Future Generation Communication and Networking*, *13*, 1449–1452.

Blattberg, R.C. & Deighton, J. (1996). Manage Marketing by the Customer Equity Test. *Harvard Business Review*, *74(4)*, 136–144.

Bolton, R. N., Lemon, K. N., & Verhoef, P. C. (2004). The theoretical underpinnings of customer asset management: A framework and propositions for future research. *Journal of the Academy of marketing science*, *32(3)*, 271–292.

Borle, S., Singh, S. S., & Jain, D. C. (2008). Customer lifetime value measurement. *Management Science*, *54*, 100–112. https://doi.org/10.1287/mnsc.1070.0746

Brazdil, P., Carrier, C. G., Soares, C., & Vilalta, R. (2008). *Metalearning: Applications to data mining*. Springer Science & Business Media.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Bucklin, R. E. (2008). Marketing models for electronic commerce. In B. Wierenga (Ed.). *Handbook of marketing decision models* (pp. 327–369). Springer. https://doi.org/10.1007/978-0-387-78213-3_10

Capon, N., & Go, F. (2016). *Frameworks for market strategy: European edition.* Routledge.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics, 8(8),* 832. https://doi.org/10.3390/electronics8080832

Chamberlain, B. P., Cardoso, A., Liu, C. B., Pagliari, R., & Deisenroth, M. P. (2017). Customer lifetime value prediction using embeddings. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1753–1762). New York, NY. doi:10.1145/3097983.3098123

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16,* 321–357.

Chen, S. (2018). Estimating customer lifetime value using machine learning techniques. *Data Mining.* https://doi.org/10.5772/intechopen.76990

Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications, 36(3),* 4176–4184. https://doi.org/10.1016/j.eswa.2008.04.003

Cheng, H. T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., ... & Shah, H. (2016, September). Wide & deep learning for recommender systems. *In Proceedings of the 1st workshop on deep learning for recommender systems* (pp. 7-10).

Chiang, L. L., & Yang, C. S. (2018). Does country-of-origin brand personality generate retail customer lifetime value? A big data analytics approach. *Technological Forecasting and Social Change, 130,* 177–187. https://doi.org/10.1016/j.techfore.2017.06.034

Dirsehan, T., & Cankat, E. (2021). Role of mobile food-ordering applications in developing restaurants' brand satisfaction and loyalty in the pandemic period. *Journal of Retailing and Consumer Services, 62,* 102608. https://doi.org/10.1016/j.jretconser.2021.102608

Dobslaw, F. (2010). A parameter tuning framework for metaheuristics based on design of experiments and artificial neural networks. *World Academy of Science, Engineering and Technology, Open Science Index 40, International Journal of Computer and Information Engineering, 4(4)*, 684 - 687. doi.org/10.5281/zenodo.1058561

Donkers, B., Verhoef, P. C., & de Jong, M. G. (2007). Modeling CLV: A test of competing models in the insurance industry. *Quantitative Marketing and Economics, 5*, 163–190. https://doi.org/10.1007/s11129-006-9016-y

Ehrenberg, A. S. (1959). The pattern of consumer purchases. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 8(1)*, 26-41.

Eid, M. B. A. H., Robin, R., & Wierdak, N. (2021). An empirical investigation of the factors affecting customer lifetime value. *International Journal of Quality & Reliability Management.*

Fader, P. (2012a). Customer centricity: Focus on the right customers for strategic advantage. Wharton Digital Press.

Fader, P. (2012b, June 18). How can customer centricity be profitable? Google Think Insights 2012 Israel [Video file]. Retrieved from https://youtu.be/TlTTHUjaUng

Fader, P. (2017). Womenswear retailer Dressbarn saves millions on retention with CLV insights. Retrieved from https://www.zodiacmetrics.com/dressbarn-casestudy/

Fader, P. S., & Hardie, B. (2009). Probability models for customer-base analysis. Journal of Interactive Marketing, *23(1)*, 61–69. https://doi.org/10.1016/j.intmar.2008.11.003

Fader, P. S., Hardie, B., & Lee, K. L. (2005a). "Counting your customers" the easy way: An alternative to the pareto/NBD model. *Marketing Science, 24*, 275–284. https://doi.org/10.1287/mksc.1040.0098

Fader, P. S., Hardie, B., & Lee, K. L. (2005b). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research, 42*, 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415

Fader, P. S., Hardie, B., & Shang, J. (2010). Customer-base analysis in a discrete-time noncontractual setting. *Marketing Science, 29,* 1086–1108. https://doi.org/10.1287/mksc.1100.0580

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38(4),* 367–378. doi:10.1016/S0167-9473(01)00065-2

Glady, N., Baesens, B., & Croux, C. (2009). A modified Pareto/NBD approach for predicting customer lifetime value. *Expert Systems with Applications, 36*, 2062–2071. https://doi.org/10.1016/j.eswa.2007.12.049

González-Benito, Ó., & Martos-Partal, M. (2012). Role of retailer positioning and product category on the relationship between store brand consumption and store loyalty. *Journal of Retailing, 88*, 236–249. https://doi.org/10.1016/j.jretai.2011.05.003

González, C., Mira, J.M. & Ojeda, J.A. (2016). Applying multi-output random forest models to electricity price forecast. *Preprints*, 2016090053. doi:10.20944/preprints201609.0053.v1

Gupta, S. & Zeithaml, V. A. (2006). Customer metrics and their impact on financial performance. *Marketing Science, 25(6)*, 718–739. https://doi.org/10.1287/mksc.1060.0221

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., … Sriram, S. (2006). Modeling customer lifetime value. *Journal of Service Research, 9*, 139–155. https://doi.org/10.1177/1094670506293810

Gupta, S., Lehmann, D., & Stuart J. A. (2004). Valuing customers. *Journal of Marketing Research, 41(1)*, 7–18. https://doi.org/10.1509/jmkr.41.1.7.25084

Hasanat, M. W., Hoque, A., Shikha, F. A., Anwar, M., Hamid, A, & Tat, H. H. (2020). The impact of coronavirus (COVID-19) on e-business in Malaysia. *Asian Journal of Multidisciplinary Studies, 3(1)*, 85–90

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). New York: Springer.

Heldt, R., Silveira, C. S., & Luce, F. B. (2021). Predicting customer value per product: From RFM to RFM/P. *Journal of Business Research, 127*, 444–453. https://doi.org/10.1016/j.jbusres.2019.05.001

Hu, Y., Lin, L., & Tang, S. (2019). Second-order information in first-order optimization methods. *ArXiv*, abs/1912.09926.

Janssens, B., Bogaert, M., Bagué, A., & Poel, D. (2022). B2Boost: Instance-dependent profit-driven modelling of B2B churn. *Annals of Operations Research, 310*, 1–27. https://doi.org/10.1007/s10479-022-04631-5

Jasek, P., Vrana, L., Sperkova, L., Sperkova, L., Smutny, Z., & Kobulsky, M. (2018). Modeling and application of customer lifetime value in online retail. *Informatics, 5*, 2. https://doi.org/10.3390/informatics5010002

Jiang, L. A., Yang, Z., & Jun, M. (2013). Measuring consumer perceptions of online shopping convenience. *Journal of Service Management*, *24*, 191–214. https://doi.org/10.1108/09564231311323962

Kapanova, K.G.; Dimov, I., & Sellier, J.M. (2018). A genetic approach to automatic neural network architecture optimization. *Neural Computing and Applications, 29*, 1481–1492. https://doi.org/10.1007/s00521-016-2510-6

Kayakuş, M., & Çevik, K. (2020). Estimation the number of visitor of e-commerce website by artificial neural networks during Covid19 in Turkey. *Electronic Turkish Studies, 15*(4).

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science, 3*, 57–63. https://doi.org/10.1016/j.procs.2010.12.011

Khalili-Damghani, K., Abdi, F., & Abolmakarem, S. (2018). Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. *Applied Soft Computing, 73*, 816–828. https://doi.org/10.1016/j.asoc.2018.09.001

Kim, S., Choi, D., Lee, E., & Rhee, W. (2017). Churn prediction of mobile and online casual games using play log data. *PLoS ONE, 12*, e0180735. https://doi.org/10.1371/journal.pone.0180735

Kim, S., Jung, T. S., Suh, E. H., & Hwang, H. S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert Systems with Applications, 31*, 101–107. https://doi.org/10.1016/j.eswa.2005.09.004

Kitchens, B., Dobolyi, D., Li, J., & Abbasi, A. (2018). Advanced customer analytics: Strategic value through integration of relationship-oriented big data. *Journal of Management Information Systems, 35(2)*, 540-574.

Kotler, P., & Keller, K. L. (2015). *Marketing management* (15th ed.). Prentice Hall.

Kumar, V. (2006). Customer lifetime value. In R. Grover (Ed.), *The handbook of marketing research* (pp. 602–627). Sage Publications. https://doi.org/10.4135/9781412973380.n29

Kumar, V., & George M. (2007). Measuring and maximizing customer equity: A critical analysis. *Journal of the Academy of Marketing Science, 35(2)*, 157–171. https://doi.org/10.1007/s11747-007-0028-2

Kumar, V., & Reinartz, W. (2016). Creating enduring customer value. *Journal of Marketing, 80*, 36–68. https://doi.org/10.1509/jm.15.0414

Kumar, V., & Shah, D. (2009). Expanding the role of marketing: From customer equity to market capitalization. *Journal of Marketing, 73(6)*, 119–136. Retrieved from http://www.jstor.org/stable/20619064

Kumar, V., Ramani, G., & Bohling, T. (2004). Customer lifetime value approaches and best practice applications. *Journal of Interactive Marketing, 18*, 60–72. https://doi.org/10.1002/dir.20014

Kumar, P., Batra, S., & Raman, B. (2021). Deep neural network hyper-parameter tuning through twofold genetic approach. *Soft Computing*, 25(13), 8747-8771.

Larivière, B. (2008). Linking perceptual and behavioral customer metrics to multiperiod customer profitability: a comprehensive service-profit chain application. *Journal of Service Research, 11(1)*, 3–21. https://doi.org/10.1177/1094670508319092

Lee S., Bae, J. H., Hong, J., Yang, D., Panagos, P., Borrelli, P., … Lim, K. J. (2022). Estimation of rainfall erosivity factor in Italy and Switzerland using Bayesian optimization-based machine learning models. *Catena, 211*, 105957. https://doi.org/10.1016/j.catena.2021.105957

Lemon, K. N., Rust, R. T., & Zeithaml, V. A. (2001). What drives customer equity? *Marketing Management, 10(1)*, 20–25.

Leone, R. P., Rao, V. R., Keller, K. L., Luo, A. M., McAlister, L., & Srivastava, R. (2006). Linking brand equity to customer equity. *Journal of Service Research*, *9*, 125–138. https://doi.org/10.1177/1094670506293563

Lessmann, S., & Voß, S. (2008). Supervised classification for decision support in customer relationship management. In A. Bortfeldt, J. Homberger, H. Kopfer, G. Pankratz, & R. Strangmeier (Eds.), *Intelligent decision support* (pp. 281–297). Gabler. https://doi.org/10.1007/978-3-8349-9777-7_14

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *In Proceedings of Advances in neural information processing systems* (pp. 4765–4774). MIT Press. Long Beach, CA, USA,

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In N. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297). Berkeley: University of California Press.

Malthouse, E. C., & Blattberg, R. C. (2005). Can we predict customer lifetime value? *Journal of Interactive Marketing*, *19*, 2–16. https://doi.org/10.1002/dir.20027

Malthouse, E. C. (2013). Segmentation and lifetime value models using SAS®. *SAS Institute Inc.*

Mesforoush, A., & Tarokh, M. J. (2013). Customer profitability segmentation for SMEs case study: Network equipment company. *International Journal of Research in Industrial Engineering*, *2*, 30–44.

McCarthy, D., Fader, P. S., & Hardie, B. (2017). Valuing subscription-based businesses using publicly disclosed customer data. *Journal of Marketing, 81(1)*. https://doi.org/10.1509/jm.15.0519

Mishra, P., & Passos, D. (2022). Multi-output 1-dimensional convolutional neural networks for simultaneous prediction of different traits of fruit based on near-infrared spectroscopy. *Postharvest Biology and Technology, 183*, 111741. https://doi.org/10.1016/j.postharvbio.2021.111741

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. https://doi.org/10.1016/j.dss.2014.03.001

Moro, S., Cortez, P., & Rita, P. (2015). Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. *Neural Computing and Applications*, *26*, 131–139. https://doi.org/10.1007/s00521-014-1703-0

Morrison, D. G., & David, C. S. (1988). Generalizing the NBD model for customer purchases: What are the implications and is it worth the effort? *Journal of Business & Economic Statistics*, *6*, 145–159. https://dx.doi.org/10.1080/07350015.1988.10509648

Muddamsetty, S. M., Jahromi, M. N., Ciontos, A. E., Fenoy, L. M., & Moeslund, T. B. (2022). Visual explanation of black-box model: Similarity Difference and Uniqueness (SIDU) method. *Pattern Recognition, 127*, 108604.

Perišić Prodan, M., Cerović, M., & Ivančić, I. (2022). The Impact of Relationship Marketing on Customer Satisfaction and Loyalty in Luxury Hotels. *Ekonomska misao i praksa*, *31(1)*, 189-210.

Pfeifer, P. E., Haskins, M. E., & Conroy, R. M. (2005). Customer lifetime value, customer profitability, and the treatment of acquisition spending. *Journal of Managerial Issues, 17*, 11–25. Retrieved from https://www.jstor.org/stable/40604472

Reinartz, W., & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing, 64*, 17–35. https://doi.org/10.1509/jmkg.64.4.17.18077

Reinartz, W., Thomas, J. S., & Kumar, V. (2005). Balancing acquisition and retention resources to maximize customer profitability. Journal of marketing, 69(1), 63–79, doi:10.1509/jmkg.69.1.63.55511.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence, 36(11)*, 1527–1535.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *ArXiv*, arXiv:1706.05098. Retrieved from http://arxiv.org/abs/arXiv:1706.05098

Rust, R. T., Lemon, K. N., & Zeithaml V. A. (2004). Return on marketing: Using customer equity to focus marketing strategy. *Journal of Marketing, 68(1)*, 109–127. doi:10.1509/jmkg.68.1.109.24030

Ryals, L., & Knox, S. (2007). Measuring and managing customer relationship risk in business markets. *Industrial Marketing Management*, *36*, 823–833. https://doi.org/10.1016/j.indmarman.2006.06.017

Sajjad, M., Khan, S. U., Khan, N., Haq, I. U., Ullah, A., Lee, M. Y., & Baik, S. W. (2020). Towards efficient building designing: Heating and cooling load prediction via multi-output model. *Sensors*, *20*, 6419. https://doi.org/10.3390/s20226419

Salehinejad, H., & Rahnamayan, S. (2016). Customer shopping pattern prediction: A recurrent neural network approach. *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–6. doi: 10.1109/SSCI.2016.7849921

Schauer, T. L. (2022). Understanding Millennials Regarding Online Versus In-Store Grocery Shopping Preferences. *Atlantic Marketing Journal, 11(1)*, 6.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, *61*, 85-117.

Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting your customers: Who are they and what will they do next? *Management Science, 33*, 1–24. https://doi.org/10.1287/mnsc.33.1.1

Schmittlein, D. C., & Peterson, R. A. (1994). Customer base analysis: An industrial purchase process application. *Marketing Science, 13(1)*, 41–67. https://doi.org/10.1287/mksc.13.1.41

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games II (AM-28)* (pp. 307–317). Princeton University Press.

Sifa, R., Runge, J., Bauckhage, C., & Klapper, D. (2018). Customer lifetime value prediction in non-contractual freemium settings: Chasing high-value users using deep neural networks and smote. *Hawaii International Conference on System Sciences (HICSS)*. doi:10.24251/HICSS.2018.115

Singh, N., Singh, P., & Gupta, M. (2020). An inclusive survey on machine learning for CRM: A paradigm shift. *Decision*, *47*, 447–457. https://doi.org/10.1007/s40622-020-00261-7

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*, 1929–1958.

Stahl, H. K., Matzler, K., & Hinterhuber, H. H. (2003). Linking customer lifetime value with shareholder value. *Industrial Marketing Management, 32(4)*, 267–279. https://doi.org/10.1016/S0019-8501(02)00188-8

Sunder, S., Kumar, V., & Zhao, Y. (2016). Measuring the lifetime value of a customer in the consumer-packaged goods industry. *Journal of Marketing Research*, 53*(6)*, 90

Tsiptsis, K., & Chorianopoulos, A. (2010). *Data mining techniques in CRM: Inside customer segmentation*. West Sussex, United Kingdom: Wiley.

Ulrich, A., & Korn, O. (1999). Model selection in neural networks. *Neural Networks, 12*, 309–323. https://doi.org/10.1016/S0893-6080(98)00117-8

Vanderveld, A., Pandey, A., Han, A., & Parekh, R. (2016). An engagement-based customer lifetime value system for e-commerce. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 293–302). New York, NY. https://doi.org/10.1145/2939672.2939693.

Venkatesan, R., & Kumar, V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing, 68(4)*, 106–125. https://doi.org/10.1509/jmkg.68.4.106.42728

Venkatakrishna, R., Mishra, P., & Tiwari, S. P. (2021). Customer lifetime value prediction and segmentation using machine learning. *International Journal of Engineering Science, 9*, 36–48.

Venkatesan, R. (2015). Customer-lifetime-value–based resource allocation. *In Handbook of Research on Customer Equity in Marketing* (pp. 283-305). Edward Elgar Publishing.

Villanueva, J., & Hanssens, D. M. (2007). Customer equity: Measurement, management and research opportunities. *Foundations and Trends in Marketing, 1*, 1–95. https://doi.org/10.1561/1700000002

Wang, X., Liu, T., & Miao, J. (2019). *A deep probabilistic model for customer lifetime value prediction*. Retrieved from https://arxiv.org/abs/1912.07753.

Win, T. T., & Bo, K. S. (2020). Predicting customer class using customer lifetime value with random forest algorithm. *Proceedings of the 2020 International Conference on Advanced Information Technologies (ICAIT)* (pp. 236–241). Yangon, Myanmar. https://doi.org/10.1109/ICAIT51105.2020.9261792

Wong, E., & Wei, Y. (2018). Customer online shopping experience data analytics: Integrated customer segmentation and customised services prediction model. *International Journal of Retail & Distribution Management, 46(3)*. doi:10.1108/IJRDM-06-2017-0130

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., & Deng, S.-H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology, 17*, 26–40. https://doi.org/10.11989/JEST.1674-862X.80904120

Wu, J., Shi, L., Lin, W.-P., Tsai, S.-B., Li, Y., Yang, L., & Xu, G. (2020). An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm. *Mathematical Problems in Engineering*, 1–7. https://doi.org/10.1155/2020/8884227

Yılmaz Benk, G., Badur, B., & Mardikyan, S. (2022). A new 360° framework to predict customer lifetime value for multi-category e-commerce companies using a multi-output deep neural network and explainable artificial intelligence. *Information, 13*(8), 373. https://doi.org/10.3390/info13080373

Zhang, Y., Bradlow, E. T., & Small, D. S. (2015). Predicting customer value using clumpiness: From RFM to RFMC. *Marketing Science, 34*, 195–208. https://doi.org/10.1287/mksc.2014.0873

Zhang, J., Dixit, A., & Friedmann R. (2010). Customer loyalty and lifetime value: An empirical investigation of consumer-packaged goods. *Journal of Marketing Theory and Practice, 18(2),* 127–140. Retrieved from https://www.jstor.org/stable/27821047