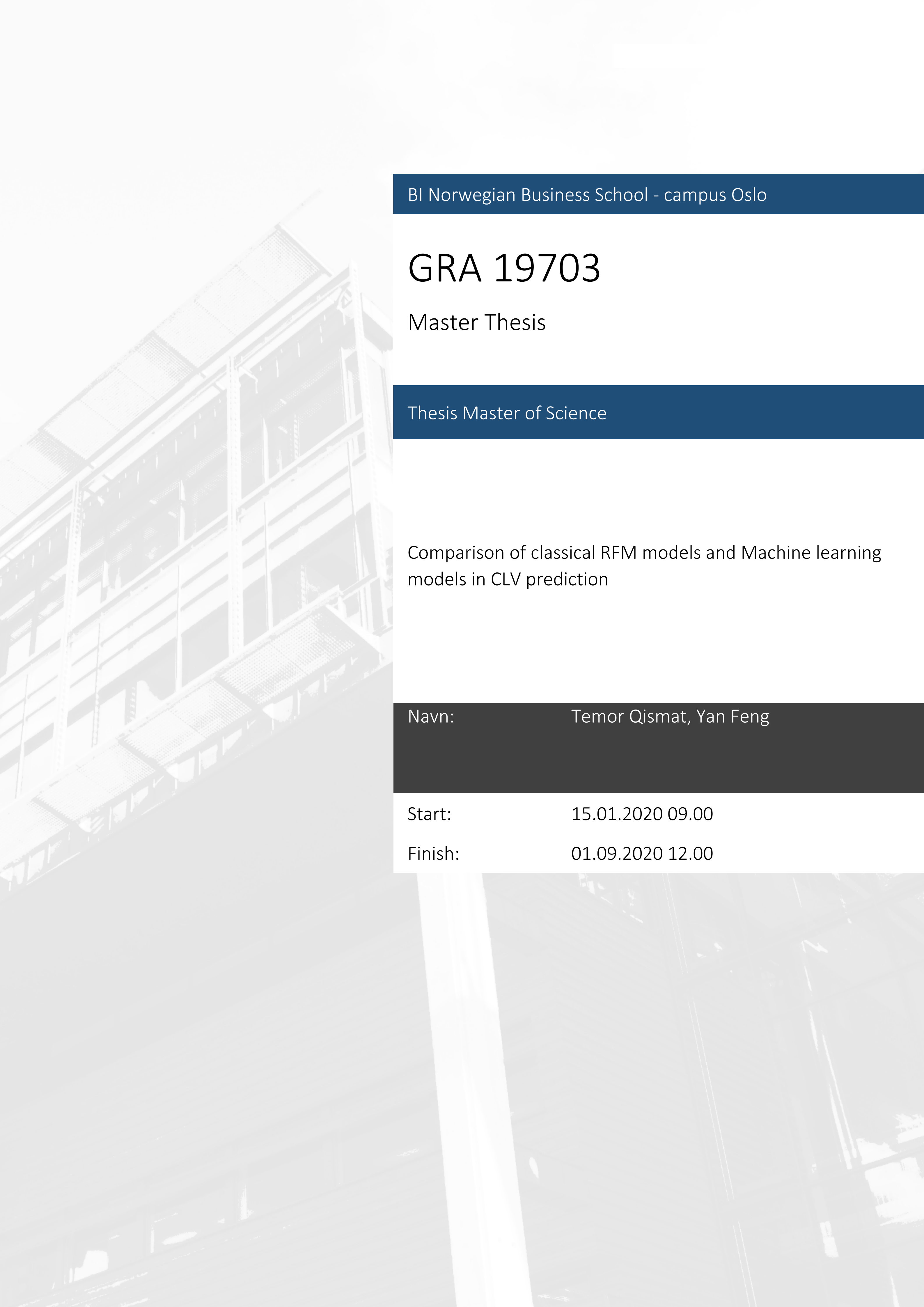


BI Norwegian Business School - campus Oslo

GRA 19703

Master Thesis

Thesis Master of Science

Comparison of classical RFM models and Machine learning models in CLV prediction

Navn: Temor Qismat, Yan Feng

Start:	15.01.2020 09.00
Finish:	01.09.2020 12.00

Comparison of classical RFM models and Machine learning models in CLV prediction

Hand-in date:

01.09.2020

Campus:

BI Oslo

Examination code and name:

GRA19703 - Master Thesis

Programme:

Master of Science in Business Analytics

Table of Contents

Acknowlegdment

This master thesis marks the end of two years at BI Norwegian Business School and was finalized on September 1st, 2020.

As the first batch of the Business Analytics program, we were both lucky and unlucky to write a thesis in this field. Lucky, because we are the first batch that got to study the exciting area of analytics in combination with business. Unlucky, in the sense that there are no or little precedent on how to use this knowledge to produce a master thesis. Therefore, we had to face many challenges when choosing a topic and how we would solve the practical side of our study. These choices would depend on some criteria's that had to be met, whereas the most crucial one was the data collection. Because of this, we found ourselves in a situation of trying and failing and even began to doubt our choices and at one point even considered changing the topic.

However, we persevered and found solutions to our problems. The thesis has therefore uplifted our knowledge-base and enabled us to understand the topic even deeper and appreciate the complexity of the field of Marketing and Costumer Behavior.

The writing of the thesis was especially challenging because of the ongoing corona pandemic in the world, which affected our personal lives and thus the progress of the thesis. This however made us stronger and we feel like we are ready to face any such challenges that may come in our professional lives.

Last, but not least, we would like to say thank you to our thesis supervisor Matilda Dorotic for providing us with guidance and valuable feedback along the way.

Thank you.

Abstract

This study analyzes the difference between classical RFM models and Machine Leaning (ML) models when calculating CLV with transactional data.

In this paper we run different programs to analyze the CLV value by using both methods. Based on the results, the researchers found out that Pareto/NBD model have better predictive power of performing CLV predictions than ML models. Lastly, the findings proved the effectiveness of the Pareto/NBD method of calculating CLV.

1. Introduction

1.1 Area of study

Companies use great amounts of time on customers by investing in marketing campaigns, ads, acquisitions, promotions, discounts, etc., to generate revenue and be profitable. Some customers respond to these actions and, in turn, increase profitability, which makes them valuable. At the same time, some customers do not respond to these actions, and therefore, decrease the profitability, which makes them not as valuable. Therefore, identifying customer behavior patterns is important to target the 'right' customers with the 'right' actions, and as such increase, the response rate, decrease costs, and increase profitability.

To identify customers' behavior, companies may look at customers' past purchase history, which can help companies understand customers' buying behavior. Understanding the behavior of a customer can then help us to determine how much value a customer will generate through the course of their interaction with the company, and in turn, determine how much a customer is worth. The term for this is Customer Lifetime Value (CLV), which is widely used when deciding which customers are valuable and which are not.

There has been a lot of research and many articles written in the field of CLV. Therefore, various methods have been developed to accurately calculate a customer's CLV. The most well-known method is RFM-method, which stands for recency, frequency, and monetary-value. All of these parameters describe customers' past purchases and are used to calculate the CLV.

One of the problems with the RMF method is that it only looks to the past. However, a better determination of a customer's value may be to predict future purchase behavior. That is why methods, such as the Pareto/NBD model, have been developed to determine CLV based on future predictions. This method has proven to be more reliable. However, it is more complex and can be computationally intense.

In recent years, Machine Learning (ML) has come a long way. ML models can incorporate many parameters or features as opposed to the classical models, and therefore can potentially yield better results with higher accuracy. In our thesis, we are therefore interested in using both the classical method and ML to calculate CLV and compare results, in order to find whether this is a better alternative to the classical models.

There has been some research about whether ML models perform better than the classical models. To our knowledge, no such studies have been conducted where the researchers compare the predictive accuracy between traditional CLV models and ML models, such as Pareto/NBD and ML models.

1.2 Motivation of the Research

The thesis focuses on comparing classic RFM and ML methods to CLV prediction. The reasons why this topic was chosen was the following:

1) The present issue of testing out the performance of both methods in CLV prediction is not as widely researched, and empirical studies are limited.

2) This thesis will provide valuable information for future CLV studies of customer behavioral analysis.

1.3 Research questions

Our thesis consists of two parts, where we first look at customer segmentation by using classical RFM methods, with transactional data acquired from Kaggle. We will then use the superior Pareto/NBD model to calculate CLV. The second part will focus on using ML techniques to calculate CLV, where we will use the same transactional data from Kaggle, including a few extra variables.

Our research question is, therefore, the following:

‘‘*Does predicting CLV based on Machine Learning give better predictive accuracy than the classical Pareto/NBD model?*’’

1.4 Value of research

This thesis will assess the performances of the models by using different CLV prediction methods and analyzing the challenges of measurements. This paper will give ideas to commercial companies who have an interest in using the technology of customer behavioral analysis. In addition, the analysis of this thesis will show the effectiveness, stronger predictive power, time-saving aspect of using a better method in the CLV prediction.

1.5 Research structure

The following figure presents the research structure of this paper. The first part includes the area of the study and why this topic was chosen. Second, the theoretical foundation for the research: the different CLV prediction methods and measures of different methods. Then, the research methodology to the theoretical framework and the practicality performance of models. Next, the results of the models will be compared and discussed. Last, the conclusion and suggestions for future research will be presented.

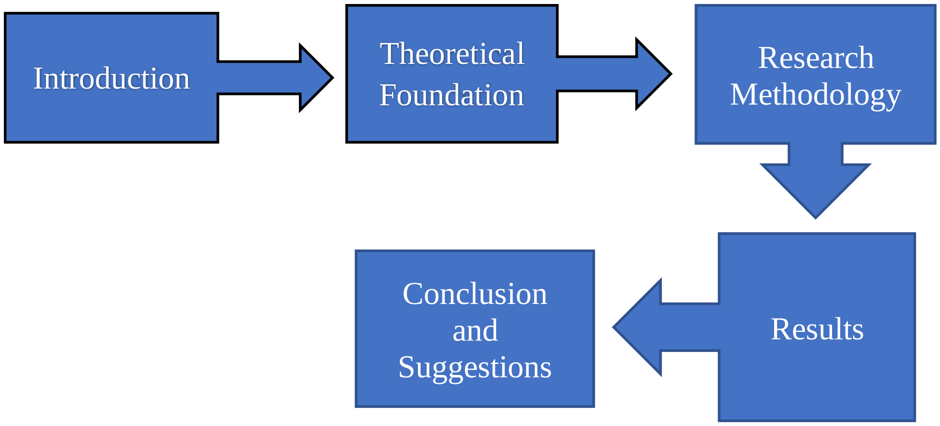


Figure 1 Research Structure

2. Literature Review

2.1 Aggregate CLV

Companies are focusing on customer retention-oriented marketing developments and maintaining long-term relationships with valuable customers instead of focusing on individual transactions in the recent marketing perspective. It is important to not only building a relationship with customers but also targeting the right customers who will generate the most profit to the firm in the long term. Thus, the firm could establish strong retention through customer satisfaction (Petter Flordal & Joakim Friberg, 2013). It is shown in many studies that customer retention management (CRM) plays a significant role in many service-oriented companies. Maximizing customer loyalty creates satisfaction and happiness in the customers (Kincaid, 2003; Ngai et al., 2009; Parvatiyar and Sheth, 2001; Umayaparvathi and Iyakutti, 2012). The longer the customer relationship, the higher the customer's worth (Reichheld & Kenny, 1990).

A small change in the level of customer retention could lead to significant changes in the shares and the profits of the organization (Agarwal et al., 2004; Kotler and Keller, 2006; Swift, 2001; Van den Poel and Larivie're, 2004). The advanced technology allows the change in marketing approaches and consideration of customers as the real focus of the market.

Therefore, many commercial enterprises start to shift focus towards products to customers and predict their behaviors to prevent them from leaving and to obtain the highest profits and revenues for organizations (Dick and Basu, 2003; Lai, 2009; Payne and Frow, 2004).

Predicting the calculate lifetime value (CLV), which drives firm profitability and value, is the heart of predicting retention. (Berger PD, Nasr NI, 1998; Fader PS, Hardie BGS, 2012; Gupta S, Lehmann DR, Stuart JA, 2004; Venkatesan R, Kumar V, 2004; Gutha Jaya Krishna, Vadlamani Ravi , 2016). Often a primitive or vague notion of the “customer lifetime value” (CLV) of customers surfaces in the evaluation of customer acquisition strategies or media tests (Dwyer, 1997).

Customer Lifetime Value (CLV) is a way of forecasting all the worth a business will get associated with a customer from the time the customer starts dealing with the business until the time the customer attrite (Eva Ascarza et al., 2018). It is difficult to know exactly the relationship between the company and each individual customer, a decent estimate and state CLV prediction is therefore needed. CLV prediction helps to create a sustainable relationship with selected customers, generating higher revenue that, in turn, enhances business growth—determining the dynamic view of customer behavior, future marketing strategies and fostering customer loyalty (Aslekar Avinash, Piyali Sahu, Arunima Pahari, 2019). Customer lifetime value (CLV) is an expectation of the net revenue (Greenberg, 2009), represents the present value of the expected benefits (e.g., gross margin) less the burdens (e.g., direct costs of servicing and communicating) from customers (Dwyer, 1997).

When predicting future CLV, there are two kinds of customer-company relationships to be counted: contractual or non-contractual (Reinartz and Kumar,2000). A contractual relationship implies a legal relationship between the customer and the company, which the company knows exactly when a customer becomes inactive. At the same time, a non-contractual relationship implies a non-legal binding relationship that the company does not know when a customer drops out. On top of that, customers could make both discrete purchases, which only occur at a certain time, and continuous purchases, which occur at any point in time. The customer-company relationship of our dataset is a non-contractual and continuous one.

There are two strategies which require different data and modeling:
First, predict the future value for existing customers whose transaction history has been collected. Second, predict the future value for new customers who just made their first purchase.

CLV is most commonly estimated by the recency, frequency, and monetary (RFM) aspects of the customer’s association with the company. There are some CLV models, such as simple basic models, retention models, migration models, and Markov models, were presented by Petter Flordal & Joakim Friberg in their report.

2.2 Simple basic model

The following formula computes CLV at the individual level:

$\mathrm{CLV_i}=\sum_{t=1}^{T}\frac{(P_{it}-C_{it})r_{it}}{(1+d)^t}-\ AC_i$

Where I represent the customer index, P_{it} represents the price paid by customer I at time t, C_{it} represents the direct cost of servicing customer I at time t, d is the discount rate, $r_{it}$is the probability of customer I being active at time t, AC_I is the acquisition cost of customer I, and T is the forecast horizon for estimating CLV.

2.3 RFM/ probabilistic models

Probabilistic models are using the RFM values that are computed from a list of purchase transactions. Each transaction consists of an individual customer ID, an order date, and an order value. Different models are used for different customer relationships. RFM analysis is a method to predict which customers are likely to respond as well as that will not respond to firms’ offers. It provides guidance to the companies for analyzing the customer value (Inderpreet Singh &Sukhpal Singh, 2017).

Aslekar Avinash and his cowriters believe that the simplest RFM approach to calculate CLV is historical customer lifetime value, which does not account for time. This method is based on the past transactions to find the customers who have the same behavioral pattern and over the same time span. The CLV prediction is, therefore, based on the recency of the last purchase. However, it does not help when predicting the future activities of the customers. The propose of predicting CLV by using RFM models is to identify the customer behaviors in order to therefore predict the future activities of them (Aslekar Avinash, Piyali Sahu, Arunima Pahari, 2019).

Beta-geometric (BG/NBD) Model is used for non-contractual situations in which customers can make purchases at any time (Fader and Hardie, 2009). The BG/NBD model uses four parameters to describe the rate at which customers make purchases and the rate at which they drop out. The NBD model is the first customer-based probability model created by Ehrenberg in 1959. Ehrenberg assumed that customers purchased randomly at an individual level, with a different time-invariant purchase rate per customer and captured by a Poisson distribution with a gamma-distributed purchase rate. However, the BG/NBD model is easier to implement than Pareto/NBD and runs faster. In 1987, Schumittlein et al. introduced the extended model by allowing a time-variant purchase rate, which is called Pareto/NBD model.

Pareto/ negative binomial distribution (NBD) Model is one of the most classic and commonly used RFM models of CLV calculation (Schmittlein, Morrison, and Colombo, 1987), focusing only on the number of purchases counted throughout a lifetime. This model assumes that customers are first ‘alive’ (actively purchasing) for an unobserved period of time before they ‘die’ (permanently inactive). When a customer is alive, he/ she is captured by an exponential distribution with a gamma-distributed dropout rate, which is also known as the Pareto distribution. The recency, frequency, and length of observation period are used to predict the number of customers’ future purchases. Researchers like Fader and Hardie were impressed by the performance of this model. Therefore, the model was expanded by Fader and his cowriters (Peter S. Fader, Bruce G. S. Hardie, Ka Lok Lee, 2004).
Later, Bruce Hardie (2007) have modified and developed the model, too. He also created a multitude of tutorials for using probabilistic models in a marketing context. In 2009, Abe developed a Bayesian method that incorporates time-invariant covariates in the model, which performs well with large size of data that includes different characteristics. And in the same year, Glady et al. (2009) relaxed the assumption of independence between the number of purchases and the average purchase amount. Their research also showed that the dependency between these values could be exploited to improve the performance of the prediction.

Gamma-Gamma Model is the extension of the Pareto/NBD model. (Aslekar Avinash, Piyali Sahu, Arunima Pahari, 2019). Pareto/NBD does not focus on the monetary value component. However, the Gamma-Gamma model assigns the monetary value to each of these future purchases. The Gamma-Gamma model is a good approach because it considers the economic component of each transaction and then estimates the customer's probability of staying (Aslekar Avinash, Piyali Sahu, Arunima Pahari, 2019).
Schumittlein and Peterson (1994) assumed that the number of future purchases and the average future purchase value is independent, and they created a sub-model based on the Pareto/NBD. The model they created was used to predict the average future purchase value per customer. However, later in 2005, Fader et al. disagreed with the Schumittlein and Peterson's assumption that the purchased value follows a normal distribution.

RFMTC model is developed by C.H. Cheng and Y.S. Chen as an augmented RFM model, taking consideration of recency, frequency, monetary value, time since first purchase, and churn probability (C.H. Cheng and Y.S. Chen, 2009). They used the Bernoulli sequence in probability theory, which improved the performance compared with the traditional RFM model.

These models, which are based on customers' purchasing behavior, fit a probability distribution to the observed RFM value of customers. RFM analysis could be used to quantitatively determine the recency of the purchases made by the customer, frequency of a customer's purchases, and the expense spent by the customer (monetary) from this, the probability of customers can be predicted. Customers are then given the ratings on the basis of these three input parameters. Customers who have RFM scores are more likely to be the best customers in the future (Inderpreet Singh &Sukhpal Singh, 2017).

2.4 Machine learning (ML) techniques

Data analytics makes it easier to find unknown patterns and market trends, correlations, association, customer preference, and other actionable insights. The analytical results can lead to better decision making, more effective marketing strategies, new revenue opportunities, improvement of operational efficiency. It gives a competitive advantage over its competitors and provides business benefits. Big data is used with related concepts of artificial intelligence (AI), business intelligence (BI), and data mining. A large amount of data associated with customer attrition, objectives, diverse products and services, customer characteristics, customer psychography, customer geography has been taken into consideration when doing analysis. Big data analytics thereby provides efficiency to resource management and allocation for individual customers (Aslekar Avinash, Piyali Sahu, Arunima Pahari, 2019).

With the increase in customer transaction data, it has been quite an interest to estimate the value of customers or the assessment of customers. This is an important trend in the disciplines, such as - accounting, finance, and especially marketing catering to various sectors (Petrison, Blattberg & Wang, 1993).

The combination of machine learning algorithms and customer behavioral predictive modeling pervade data science and technology to a much higher degree than a casual inspection (Arick, 2019). This is largely done due to its characteristics in data-driven, computational, and cross-disciplinary nature. As a result, a more accurate prediction of customer data, on the base of machine learning algorithms, can help people gain actionable insights and make better data-driven strategic decisions to create customer lifetime value.

2.4.1 ML models

ML models are known as suitable alternatives to the probabilistic models. These models are a widely used class of statistical models in which the parameters are fitted to the data by training with stochastic gradient descent. The advantages of ML models are that they can make use of more features than the probabilistic models, and they are very flexible because they make no assumptions on underlying relationships in the data. However, they have disadvantages as well. The ML models require more parameter tuning, could possibly get overfitting problems, and with computational high costs.

There are few studies that have considered machine learning techniques for modeling CLV.

Logistic regression (LR) is a widely used statistical technique for modeling a dependent variable by a linear combination of one or more independent variables (Hosmer, D.W., and Lemeshow, S., 2000). LR deals with binomial or multinomial classification problems, which are aiming at predicting the probability of an event by fitting data into a logistic function, thereby allowing inputs with any values to be transformed and confined to values between 0 and 1. LR follows the same assumptions of traditional regression analysis.

Artificial neural networks (ANN) is a supervised machine learning network that imitates the functionality of biological neural systems (Smith, K.A. and Gupta, J.N.D., 2000) and solves complex problems (Huang et al., 2010; Hung and Wang, 2004; Vafeiadis et al., 2015). ANN includes neuron nodes which are used to simulate the structures of neurons of the organism, and the linkage between each node is similar to the synapse of the neurons (Haykin, 1999).

Decision tree (DT) is one of the most popular supervised learning classification techniques (Quinlan, 1986). The flexibility and simplicity make it popular among many learning techniques (Tamaddoni Jahromi, A., Stakhovych, S. and Ewing, M., 2014). DT is used for creating a binary tree (Huang et al., 2012). DT C5.0 is the most widely used supervised machine-learning machine for creating DT due to its high accuracy (Juan et al., 2007; Williams et al., 2012).

A support vector machine (SVM) is a supervised learning method that is able to solve both linear and non-linear classification problems (Crone et al., 2006; Guo-en and Wei-dong, 2008). Polynomial and RBF cores are normally selected as the core functions which help a lot in decision making (Vapnik,1995; Huang et al., 2010, 2012; Vafeiadis et al., 2015; Yu et al., 2011).

Extreme gradient boosting (XGBoost) is an implementation of gradient boosted decision trees designed for speed and performance. It is also an integration model formed by continuous iterations of a week classifier (Jing Zhou, Wei Li, Jiaxin Wan, Shuai Ding, Chengyi Xia, 2019). There is k number of regression trees that are created XGBoost model to ensure that the prediction of the tree cluster is as close to the actual value as possible and that the generalization capability is as high as possible.

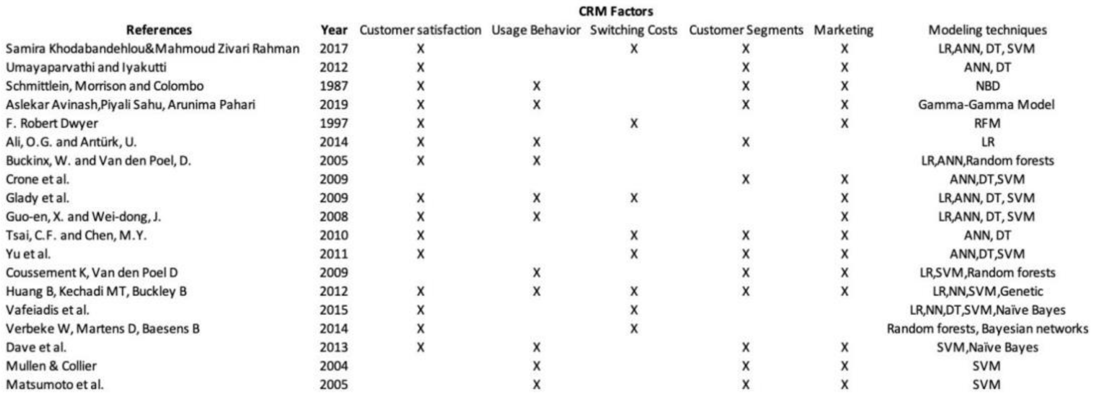
Optimizing the objective, gradient descent is used to lead to a problem of finding an optimal structure of the successive tree.

There are many other useful methods of prediction which give a good performance as well. Selecting an appropriate algorithm for prediction is one of the most important steps (Buckinx, W., and Van den Poel, D., 2005). Above practical techniques for customer prediction are briefly introduced, whose reliability, performance, and functionality have been proved in a great number of studies (Buckinx and Van den Poel, 2005; Freund and Schapire, 1996; Hung and Wang, 2004; Keramati et al., 2014; Qureshi et al., 2013; Runge et al., 2014; Vafeiadis et al., 2015; Yu et al., 2011).

Machine learning techniques can also be used to combine predictions from different models, resulting in better predictions than individual models.

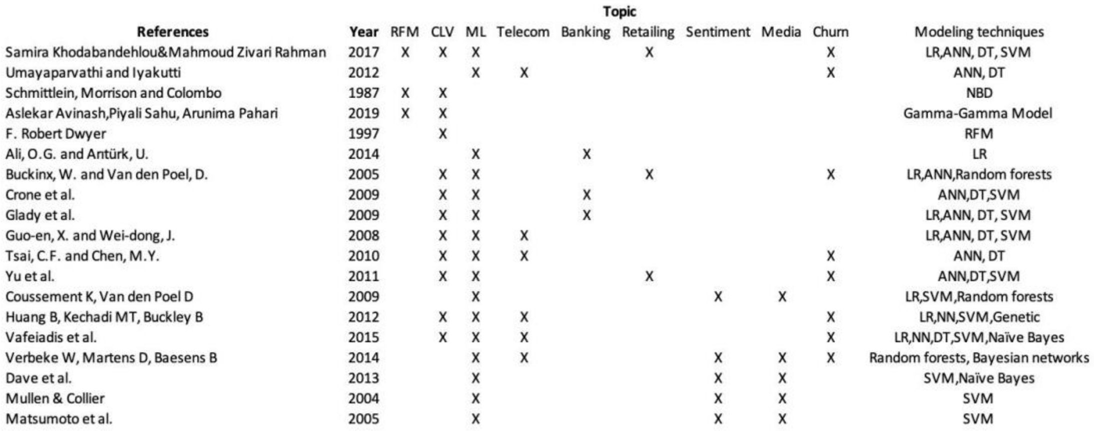
2.5 Related work

The tables below represent a summary of the previous work that were done with either RFM or ML methods of predictions in different industries.



		CRM Factors					
References	**Year**	Customer satisfaction	Usage Behavior	Switching Costs	Customer Segments	Marketing	Modeling techniques
Samira Khodabandehlou&Mahmoud Zivari Rahman	2017	X		X	X	X	LR,ANN, DT, SVM
Umayaparvathi and Iyakutti	2012	X			X	X	ANN, DT
Schmittlein, Morrison and Colombo	1987	X	X		X	X	NBD
Aslekar Avinash,Piyali Sahu, Arunima Pahari	2019	X	X		X	X	Gamma-Gamma Model
F. Robert Dwyer	1997	X		X		X	RFM
Ali, O.G. and Antürk, U.	2014	X	X		X		LR
Buckinx, W. and Van den Poel, D.	2005	X	X				LR,ANN,Random forests
Crone et al.	2009				X	X	ANN,DT,SVM
Glady et al.	2009	X	X	X		X	LR,ANN, DT, SVM
Guo-en, X. and Wei-dong, J.	2008	X	X			X	LR,ANN, DT, SVM
Tsai, C.F. and Chen, M.Y.	2010	X		X	X	X	ANN, DT
Yu et al.	2011	X		X	X	X	ANN,DT,SVM
Coussement K, Van den Poel D	2009		X		X	X	LR,SVM,Random forests
Huang B, Kechadi MT, Buckley B	2012	X	X	X	X	X	LR,NN,SVM,Genetic
Vafeiadis et al.	2015	X		X			LR,NN,DT,SVM,Naïve Bayes
Verbeke W, Martens D, Baesens B	2014	X		X			Random forests, Bayesian networks
Dave et al.	2013	X	X		X	X	SVM,Naïve Bayes
Mullen & Collier	2004		X		X	X	SVM
Matsumoto et al.	2005		X		X	X	SVM

Table 1 Related work 1



		Topic									
References	**Year**	RFM	CLV	ML	Telecom	Banking	Retailing	Sentiment	Media	Churn	Modeling techniques
Samira Khodabandehlou&Mahmoud Zivari Rahman	2017	X	X	X			X			X	LR,ANN, DT, SVM
Umayaparvathi and Iyakutti	2012			X	X					X	ANN, DT
Schmittlein, Morrison and Colombo	1987	X	X								NBD
Aslekar Avinash,Piyali Sahu, Arunima Pahari	2019	X	X								Gamma-Gamma Model
F. Robert Dwyer	1997		X								RFM
Ali, O.G. and Antürk, U.	2014			X		X					LR
Buckinx, W. and Van den Poel, D.	2005		X	X			X			X	LR,ANN,Random forests
Crone et al.	2009		X	X		X					ANN,DT,SVM
Glady et al.	2009		X	X		X					LR,ANN, DT, SVM
Guo-en, X. and Wei-dong, J.	2008		X	X	X						LR,ANN, DT, SVM
Tsai, C.F. and Chen, M.Y.	2010		X	X	X					X	ANN, DT
Yu et al.	2011		X	X			X			X	ANN,DT,SVM
Coussement K, Van den Poel D	2009			X				X	X		LR,SVM,Random forests
Huang B, Kechadi MT, Buckley B	2012		X	X	X					X	LR,NN,SVM,Genetic
Vafeiadis et al.	2015		X	X	X					X	LR,NN,DT,SVM,Naïve Bayes
Verbeke W, Martens D, Baesens B	2014			X	X			X	X	X	Random forests, Bayesian networks
Dave et al.	2013			X				X	X		SVM,Naïve Bayes
Mullen & Collier	2004			X				X	X		SVM
Matsumoto et al.	2005			X				X	X		SVM

Table 2 Related work 2

2.6 Performance measures

2.5.1 RMSE &MAE

In 2009, Glady et al. divided the data set into a three-year training set and a two-year test set. They predicted CLV using data from the training set and compared their predictions with the actual CLV from the test set and measured by both the root mean squared error and the mean absolute error. At the same time, Glady et al. ranked each individual by their predicted CLV and compared them with their true ranking by measuring the strength of a monotonic relationship between two variables. This measurement is called Spearman's correlation coefficient.

2.5.2 Ranking best to worst

Venkatesan and Kumar (2004) also spilt the data into training and test set, but they took 2.5 years as a training set and 1.5 years as a test set, then they ranked the customers from best to worst as well. The ranking was according to their models and then compared the actual sales, costs, and profits from the predicted top 5%, 10%, and 15% of customers.

A similar method was used by Malthouse and Blattberg (2005), who divided their data into a training and test set of almost the same size. Then, they ranked the customers from best to worst. They compared customers in the predicted top 20% with the actual top 20%. They then computed the accuracy, false-positive rate, and false-negative rate of the predictions.

3. Methodology

In this section, we present our approach to the research methodology for our paper. Our overall research is characterized as quantitative, and the sections below explains how we conducted the research. It describes the general research outline and elaborates on how the data and models are used to get CLV predictions.

3.1 Data Collection

The data is collected from Kaggle.com, an online site where companies create competitions and upload their data. The people that enroll in those competitions use the data to provide the desired analyzes or prediction models to these companies. The data that we collected was from a 6-year-old competition. Therefore, the data is now publicly available and can be used by anyone.

The data consists of almost 350 million rows of completely anonymized transactional data from over 300,000 shoppers. As the size of the data was too large, we had to prepare the data so that we would have a transactional history of customers from 1 shop, which is satisfactory for our research aim.

The data set, called *transaction*, contains 7,964,915 rows, where each transaction is represented by a row. The observation period represents transactions between *2012-03-2* and *2013-07-20.*

The dataset has the following attributes:

- ID: A unique id representing a customer
- Chain: An integer representing a store chain
- Dept: An aggregate grouping of the Category (e.g. water)
- Category: The product category (e.g. sparkling water)
- Company: An id of the company that sells the item
- Brand: An id of the brand to which the item belongs
- Date: The date of purchase
- Product Size: The amount of the product purchase (e.g. 16 oz of water)
- Product Measure: The units of the product purchase (e.g. ounces)
- Purchase Quantity: The number of units purchased
- Purchase Amount: The dollar amount of the purchase

3.2 Research process

In this section, we elaborate on how the observation period is handled and which performance measures were used to evaluate the Pareto/NBD model and Machine Learning model.

3.2.1 Observation Period

As mentioned, our observation period is from *2012-03-02* to *2013-07-20*. This is roughly 1.5 years. Usually, one would want to change the period to yearly observation and account for whole years. This often simplifies the statistical models, and it is easier to interpret the results afterward. However, our data is limited, and we do not have observations for the whole year in 2012 and 2013.

To evaluate the models, one has to split the data into training and test sets. The models are then fitted to the training set, which then produces predictions for the future period. The prediction is then tested on the test set, in other words, compared to the actual values.

The training and test split datasets are derived from splitting the period. However, this does not apply to both models. This is due to the nature of the models. The Pareto/NBD model, which is a probability model, is quite flexible when it comes to the length of the period in the training set, as it is independent of the prediction period. This means that one can train the model for any length in the training set and predict for any length of the future period in the test set. However, in the Machine Learning model, every row in the observation period is dependent on the data preceding it.

In other words, the length of the period in the training set is dependent on the length of the prediction period. This is due to the need for labeling values in the training data that corresponds to the values in the future period.

To account for this, we split the data into 12 months and 5 months sets. The first 12 months of the data are used to label the next 5 months. Further, the 5 month set is then used in the training model to predict values for the future period.

3.2.2 Evaluation Methods

We primarily use three metrics for comparing the Pareto/NBD and Machine Learning models;

- Root Squred Mean Error (RSME)
- Mean Absolute Error(MAE)
- R-Squared (R2)

These metrics are chosen primarily to account for the different nature of our prediction models. The Pareto/NBD model predicts CLV for each customer. This is beneficial if a company wants to target a very specific group. For example, a company may be running a new marketing campaign which is only profitable for a very specific group of customers, and if aimed at the wrong customers, they lose money.

The Machine Learning model, however, predicts CLV based on segments or rather predicts which segment a customer belongs to. In other words, it classifies which customers are the most profitable and which are not. An example would be a company that wants to run a marketing campaign to their most loyal customer group segment.

In order to get a complete picture of which of the models performs best, we evaluate them based on all of the three metrics.
In addition, the following metrics are used to evaluate the Machine Learning model;

- Area Under The Curve (AUC)
- Precision, Recall and Accuracy (P/R/A).

The AUC and P/R metrics can not be used for the comparison between the Pareto/NBD and Machine Learning model because of the different natures. In that case, one might ask why we use them at all. The reason is that AUC and P/R are metrics that can help one to interpret the ML model results and can be further used to tweak the model in order to get higher accuracy. Each of the mentioned metrics is furtherly discussed in more detail, except for AUC and P/R/A. The reason being that these metrics generally well known and are not used for the comparison of the two models. Therefore, they are not that relevant to discuss in detail.

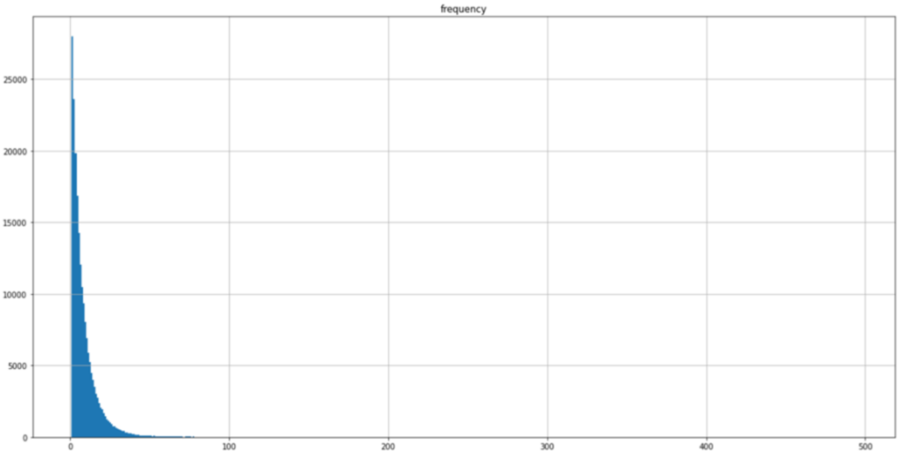
RMSE & MAE

Root Mean Squred Error (RMSE) is a performance measure that is given by

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{CLV}_{i,5} - CLV_{i,5})^2},$$

N is the number of customers, $\widehat{CLV}_{i,5}$ is the predicted CLV of customer i for year 5 and $CLV_{i,5}$ is the actual CLV of customer i for year 5. The problem with RSME is that it can inflate if is there are some extreme outliers in the data. CLV is often heavily right skewed with a long tail as shown in figure NO. To account for this MAE is used which is given by

$$\frac{1}{N}\sum_{i=1}^{N}|\widehat{CLV}_{i,5} - CLV_{i,5}|.$$



P/R/A

Precision, recall and accuracy could be viewed as separated metrics on their own, however is most useful when used in combination with each other.

Precision is the fraction of the relevant instances among the retrieved instances.

$$Precision\ = \frac{True\ positives}{True\ Positives + False\ Positives}$$

Recall is the fraction of the total amount of relevant instances that were actually retrieved.

$$Recall\ = \frac{True\ positives}{True\ Positives + False\ Negatives}$$

Accuracy is the fraction of the total amount instances that were actually retrieved.

$$Accuracy\ = \frac{True\ positives + True\ negatives}{Total}$$

Put in simple terms, these metrics measure how well the model classifies a value to a certain segment.

AUC

The AUC curve is a graph that visually shows the performance of a model across all possible classification thresholds. This is an indirect visualization of the R/P/A measures. It helps one to visually understand the performance of the model.

3.3 Transformation

As mentioned, the data set was collected through Kaggle.com and was preprocessed by including only the transactions from one of the 300 shops. We noticed that some transactions had a Purchase Amount that was equal to 0 or less. These values were removed. We assume these customers have received the product for free, and in cases were the Purchase Amount is negative, we assume these are products that were returned. These customers are, therefore, to be seen as outliers. We do this for both the Pareto/NBD model and the ML model. Besides that, the data was quite clean and did not need further preprocessing that were applicable for both the models.

3.3.1 Transforming for Pareto/NBD

We start by first preparing the data for the Pareto/NBD. Luckily, the Pareto/NBD model requires only three variables, recency, frequency, and monetary value for every customer. This is extracted from the variables in our dataset. Recency is the number of days between the customer's last purchase and their initial purchase. Frequency is the total number of repeat purchases in the observation period, and monetary value is the mean of the purchase value in dollars. The extraction of these variables was done with the help of the package lifetimes in python. This is a package that automatically calculates all of the three variables, including the total observation period T.

3.3.2 Transforming for Machine Learning

For the Machine Learning model, we do need to preprocess quite a bit. We start by splitting the data into two periods. 12 Months and 5 months. The recency, frequency and monetary values are then calculated for the first 12 months. Further these variables are used to segment the customers into 7 segments;

- Best Customers
- Big Spenders
- Loyal Customers
- Almost Lost
- Lost Customers
- Lost Cheap Customers
- Others

We start by first assigning scores to recency, frequency and monetary value. We use quantiles for this purpose. We split each of the variables into 25, 50, and 75 quantiles.

- For recency we assign value 1 = 25 quantile, 2 = 50 quantile and 3 = 75 quantile.
- For frequency we assign 4 = 25 quantile, 3 = 50 quantile and 1 = 75 quantile.
- For monetary value we assign 4 = 25 quantile, 3 t= 50 quantile and 1 = 75 quantile.

Recency are given scores from low to high, because the lowest recency corresponds to a customer purchasing something very recent.

Both Frequency and Monetary Value are given score high to low, because highest frequency and monetary value corresponds to customer purchasing many times and high spending.

These scores are further used to create the mentioned segments above.
We do this by assigning the one, two or the combination of all three scores from recency (R), frequency (F) and monetary value (M) to each segment:

- Best Customer: R = 1, F = 1, M = 1
- Big Spenders: M = 1
- Loyal Customers: F =1
- Almost Lost: R = 3, F = 1, M = 1
- Lost Customers: R = 4, F = 1, M = 1
- Lost Cheap Customers: R = 4, F = 4, M = 4
- Others: Every Other Combinations

We call this variable simply Segment

We repeat this step and create the same variable. However, this time, the segments are represented by numbers where;

- Best Customer (0): R = 1, F = 1, M = 1
- Almost Lost (1): R = 3, F = 1, M = 1
- Lost Customers (2): R = 4, F = 1, M = 1
- Lost Cheap Customers (3): R = 4, F = 4, M = 4
- Big Spenders (4): M = 1
- Loyal Customers (5): F = 1
- Others (6): Every Other Combinations

This second segment variable is called Cluster.

The reason for creating a second segment variable is because we will transform the first segment variable into seven dummy variables, while the dummy variables will be used to visualize the classification results. This is due to that the Machine Learning model does not handle variables with characters.

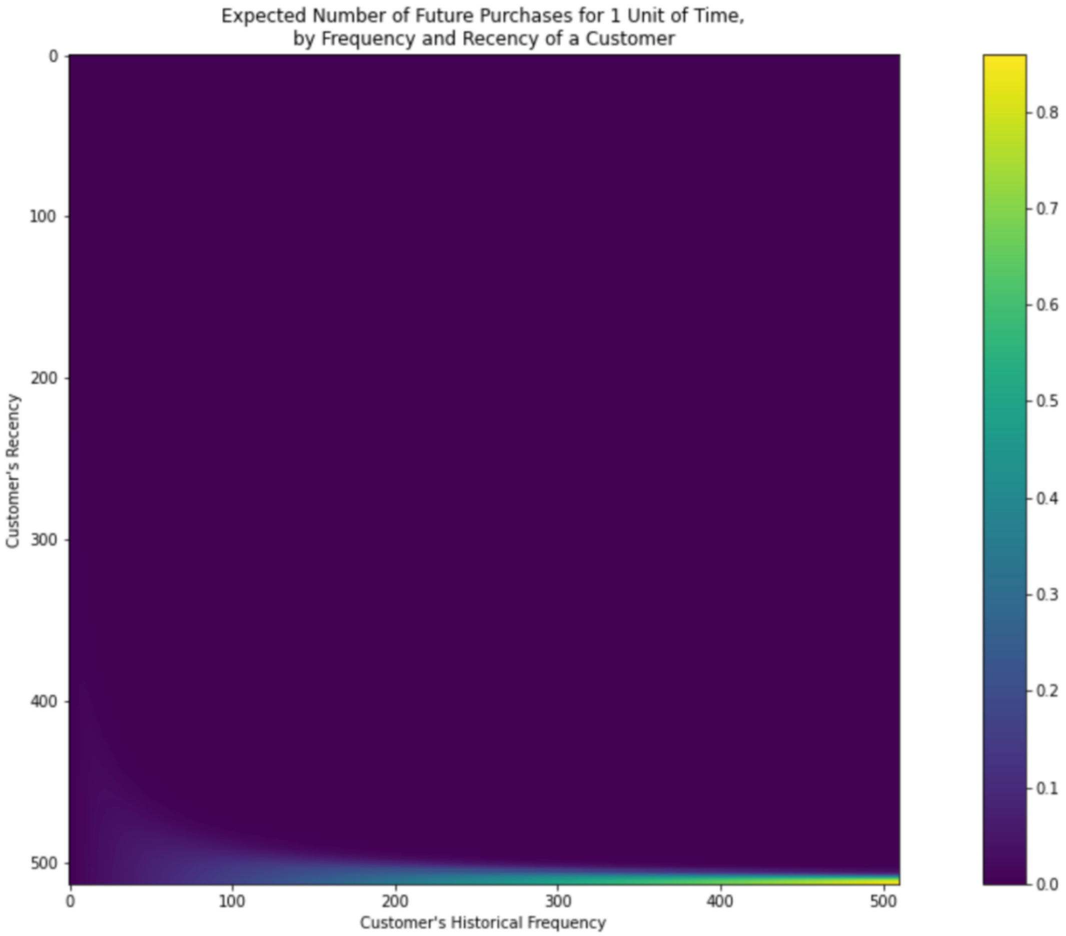
As mentioned in section 3.2.1, with machine learning, one has to account for dependencies in a time series prediction. That is why we repeat the same process for the 5-months data set until the extraction of the recency, frequency, and monetary value. We then merge the monetary value called monetary_value5M from the 5-month data set with the 12-month data set by ID variable. This way, we will have labeled values in the training data that corresponds to the values in the future period.

Further, we use k-means clustering to predict segments for the future 5-month period. K-mean clustering takes all the values in a variable and fits a model based on those values. The fitted model is then used to predict segments for each value in the input variable. The input variable is monetary_value5M, and the predicted segments for this variable are called CLVCluster. The data should now be ready for estimation with the Machine Learning model.

3.4 Pareto/NBD Estimation

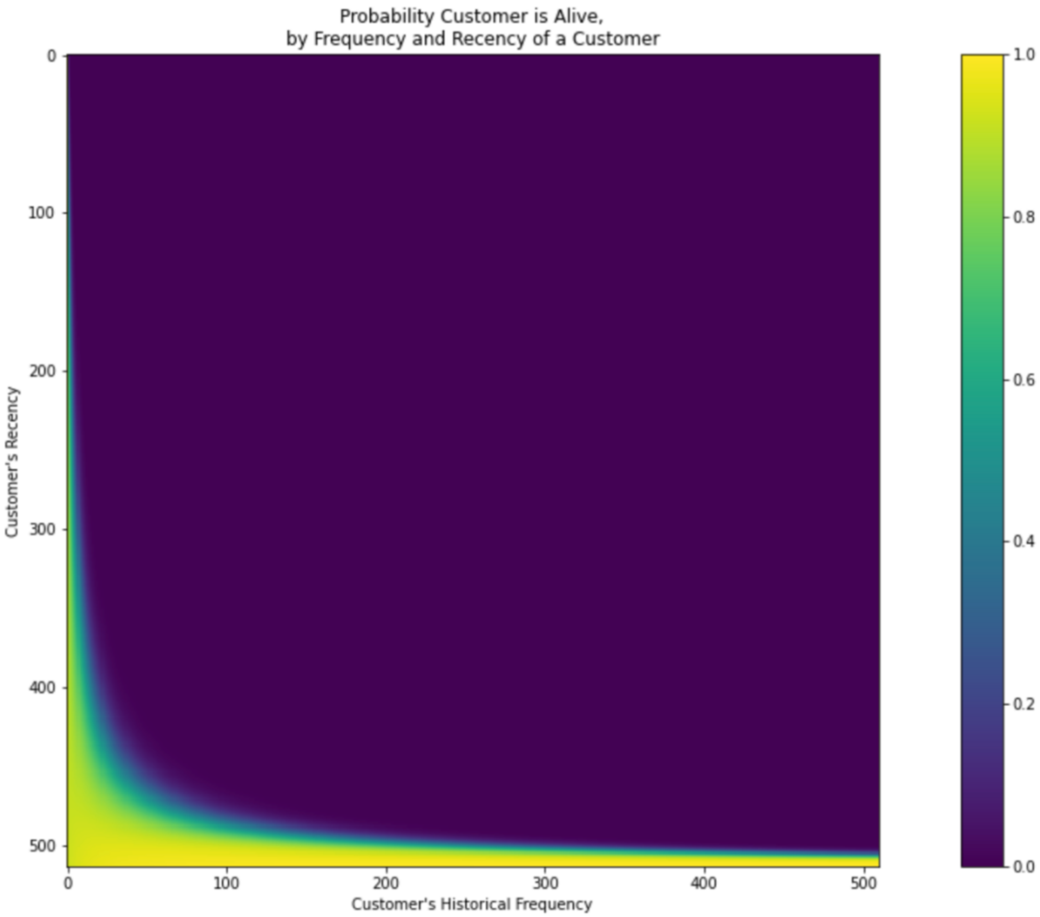
We will continue with were we left of for Pareto/NBD model. After the preparation of the data and the calculation of the recency, frequency, monetery value and T, we fit the data to what's called the betageofitter (BGF) in python. The BGF is a module in the lifetimes packages that is based on the Pareto/NBD model. This model predicts the expected number future purchases.

We can plot the model with a recency/frequency plot shown below in figure NO.



This plot show the expected number of future purchases for the next day, which is. From the plot we can see that customer with frequency above 400 and recency above 500 have an expected number purchases of 0.7 - 0.8.

We can also plot the model with a ‘Still’ Alive plot as shown below in the figure NO.



This plot shows the probability of a customer still being ´´alive´, in other words, the probability of whether they will still buy something in the future. From the plot we can see that customer with frequency of about 50 and recency of about 450 have a probability of about 90% of being alive.

Further, the test the model and predict future values, we split the dataset into two periods, calibration period, and a holdout period. In the calibration set, the period end is set to 2013-02-28, while in the holdout set, the period end is set to 2013-07-28. We then fit the BGF model on the holdout set and get predictions.

The Pareto/NBD model predicts the number of future purchases, and it does not predict the spending amount of future purchases, which is needed in order to predict the customers' CLV. Luckily, there is a solution, the extended Pareto/NBD model, which predicts the average spending amount per purchase for each customer. By multiplying the average spending amount and future expected purchases, we can predict CLV for each customer. This is done with the help of the gamma gamma submodel.

To use the gamma gamma submodel, we create a summary data from our original RFM data set to include or account for economic value, such as the monetary value. Note that in BGF, the monetary value was not used.

3.5 Extreme gradient boosting (XGBoost) Estimation

To estimate the machine learning model, we first split the data into training and test set. To decide wich ML model would give us the highest accuracy we create write wrote a code in python, where multiple Machine Learning models were appended, such Decision Tree, XGboost, SVM, KNeighreistNeighbors, RandomForest, LogisticRegression and AdaBoost. The following results were obtained:

LR	0.77987
XGB	0.95577
KNN	0.75297
DT	0.95532
RF	0.95451
ADA	0.92173
SVC	0.77987

We found that XGboost performance best out of those Machine Learning models. We then proceed to fit the model to the training set and test on the test set. The result from both the extended Pareto/NBD model and the XGBoost model are further discussed in the next section.

4. Results

In this section, both models are discussed, and the performances are compared according to different measurements. First, the Pareto/NBD model and its extended gamma-gamma model are compared. Second, the results of the XGBoost model is shown and discussed. Lastly, a comparison of the overall performance for both Pareto/NBD models and ML model is presented and measured by R squared, MAE, and RMSE.

4.1 Pareto/NBD model

Then fitted the Pareto/NBD model discussed in section 3.4 is evaluated with the help of some plots.

The plot (Figure 2) below visualizes the comparison of the repeat purchases between the actual and the model's predictions in the calibration period. The smaller gap between the actual values and predictions, the better. The blue area represents the actual values, and the orange area is the predictions. This plot shows that the Pareto/NBD model fits the data well. Even though it underestimated the frequency a bit from the beginning and more and more from the period more than 50 days, it still catches most of the trend of repeat purchases in the period.

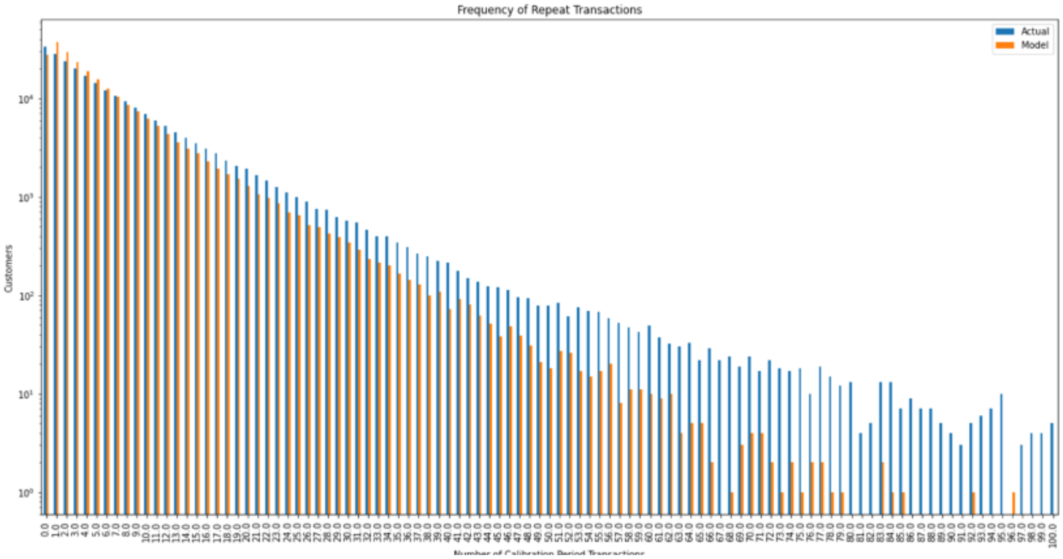


Figure 2 Frequency of repeat purchases according to Standard Pareto/NBD model

In addition, Figure 3 shows the difference between actual and predicted cumulative number of purchases for Pareto/NBD model. The red line is the boundary between the calibration period on the left and the holdout period on the right. The blue curve represents the actual value and the orange curve represents the estimated value. Here, the Pareto/NBD model is accurate until it overestimates at the end in the holdout period. In the calibration period, it performs quite well before 100 days, and underestimates slightly between 200 days and 400 days, and the gap between prediction and actual becomes bigger after 400 days. In the holdout period, it overestimates the trend a bit before 430 days, and overestimated after that.

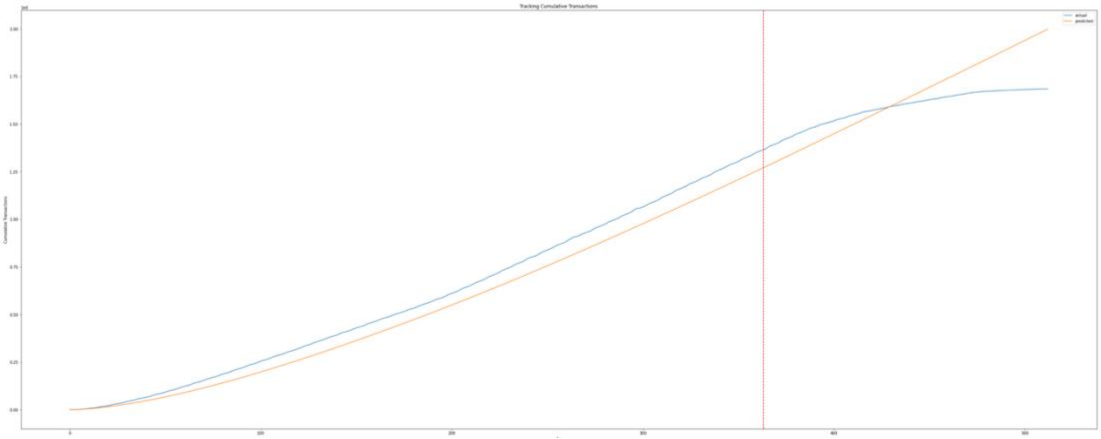


Figure 3 Actual and predicted cumulative number of purchases for standard Pareto/NBD model

Figure 4 shows incremental transactions. The plot shows that the model does a decent job capturing general trends in the data. But the same overestimation is also captured here.

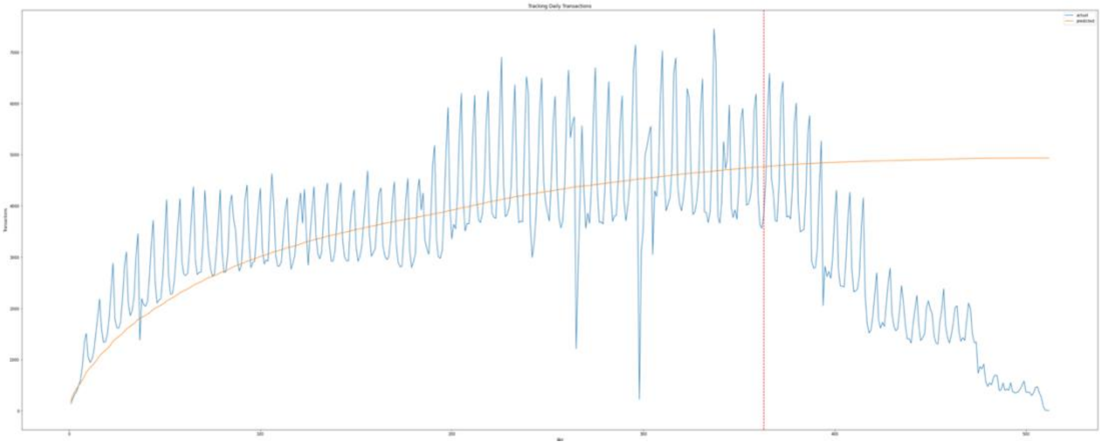


Figure 4 Actual and predicted incremental number of purchases for Pareto/NBD model

Figure 5 below represents the difference between purchases in calibration period and the holdout period. The model performs not bad up to about 150 days with small difference between actual and predicted values, but not very satisfying after that.

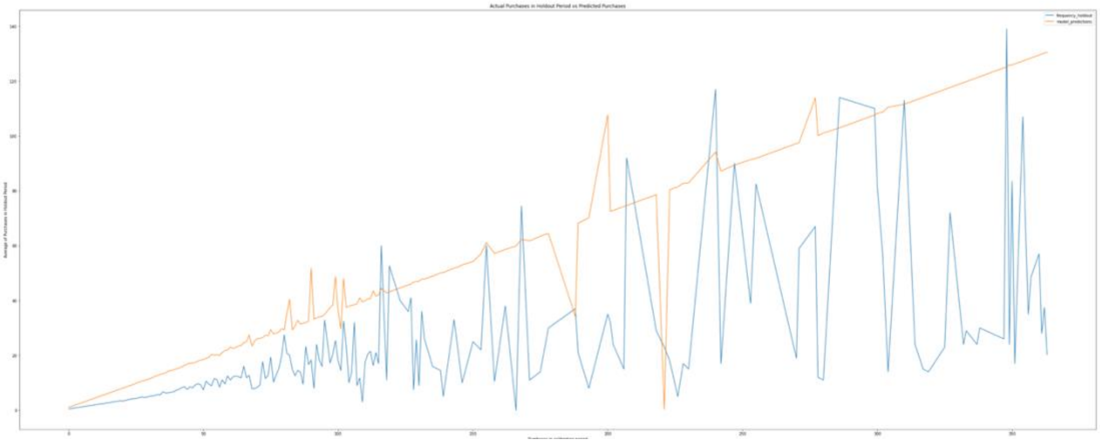


Figure 5 Calibration purchases VS holdout purchases for standard Pareto/NBD model

In order to see the prediction of individual future purchases, some single customers are picked up to check their alive rate. Given a customer’s transactional history, their historical probability of being alive can be computed according to the trained model.

Thus, a random customer is picked up and the figure below shows the probability of this customer being alive. The probability of this customer being alive decreases when this customer does not have very many purchases in a period and increases right after this customer makes a new purchase and the probability dropps when there is about one month after the customer stopps purchasing.

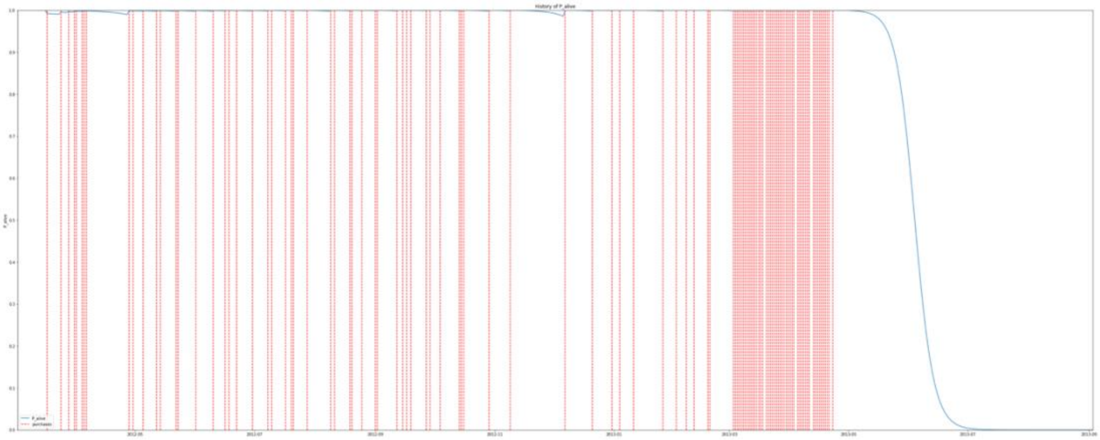


Figure 6 A random individual's probability of being alive – Pareto/NBD model

It seems that the Pareto/NBD model without covariates is somewhat realistic when predicting customer purchase behavior, however can give overestimated predictions.

To include the economics value of each transaction we proceed the a Gamma-Gamma sub-model. This model assumes that there is no relationship between the monetary value and the purchase frequency. If the Pearson correlation between the two vectors is close to 0, this model can be applied. When tested with the Person correlation the result was (0.164). The correlation is a bit high and is worrisome. However, we assume this correlation is the result of the nature of our data, where most of the customers have high number of repeated purchases. We can then assume that assume that because of this, the correlation we see are spurious correlation. We therefore decided to proceed in training the Gamma-Gamma sub-model and predict the conditional, expected average lifetime value of the customers. The calculated expected conditional average revenue is around 15, average revenue is around 13.

By plotting the Gamma Gamma sub-model, we can see that there is a very slight improvement at the beginging of the period and up to 50 days. However, it does not as well as the standard Pareto/NBD model after 50 days.

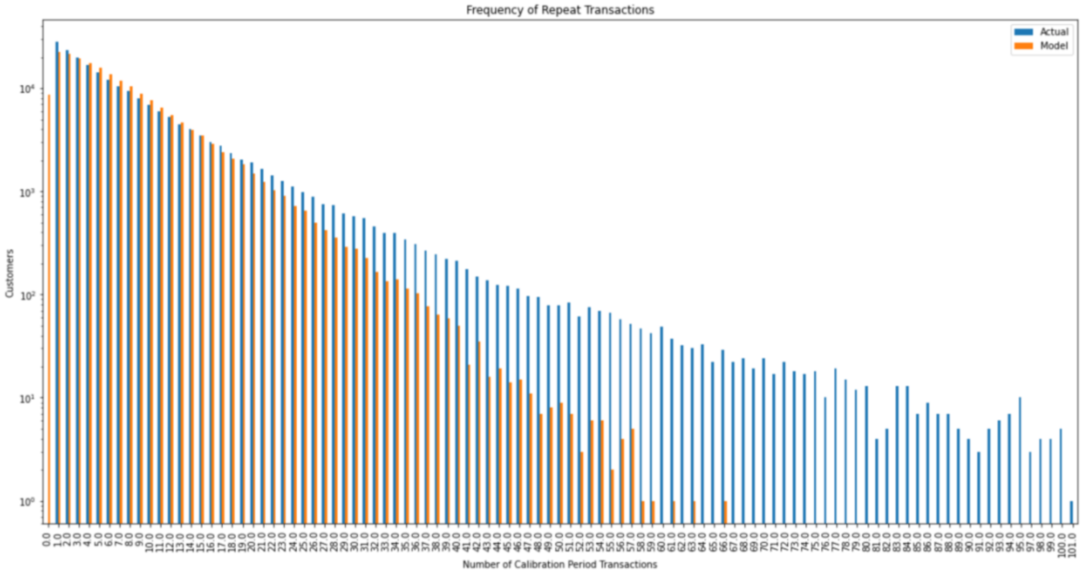


Figure 7 Frequency of repeat purchases according to Gamma-Gamma model

In terms of cumulative number of purchases, Gamma-Gamma performs better in the holdout period. We can see that it still follows the trend in the holdout period as opposed to the standard Preto/NBD model. In the calibration period, it performs quite well before 100 days as well compared with the Pareto/NBD model, and underestimates slightly between 100 days and 200 days, but the gap starts to grow between 200 days to 400 days.

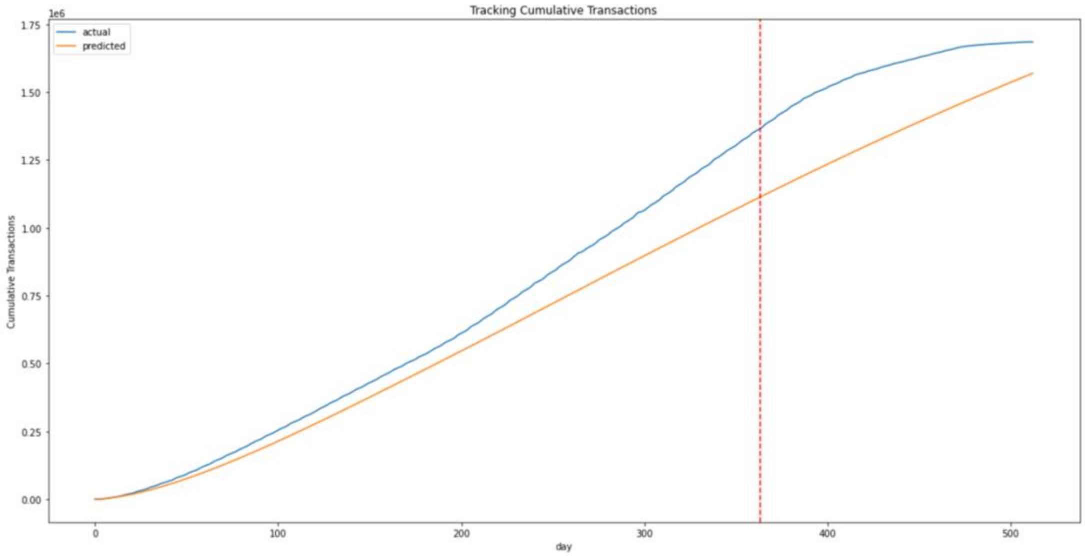


Figure 8 Actual and predicted cumulative number of purchases for Gamma-Gamma model

Figure 9 shows that the prediction of Gamma-Gamma model respect to incremental transactions has been improved camparing with stardard Pareto/NBD model.

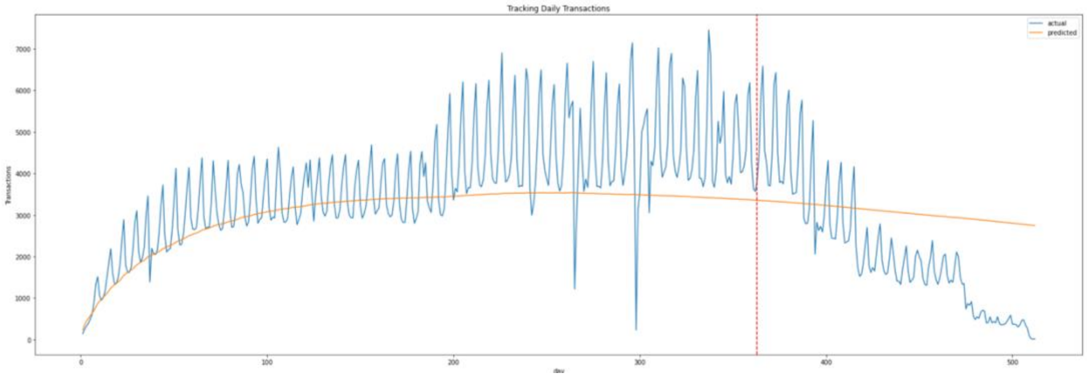


Figure 9 Actual and predicted incremental number of purchases for Gamma-Gamma model

Here, the plot below illustrates the difference between purchases in training set and the test set. The performance of model is quite good up to about 100 days, and then it underestimated the value for the rest of the training period. However, it improves the result from Pareto/model.

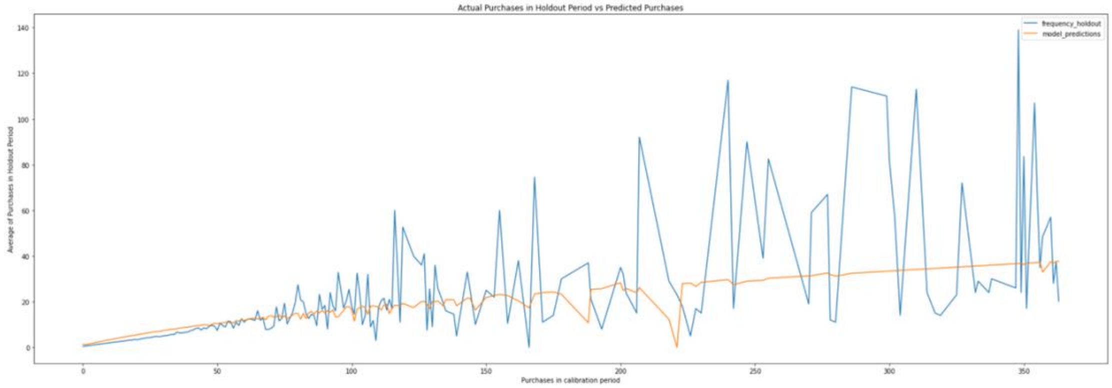


Figure 10 purchases VS holdout purchases for Gamma-Gamma model

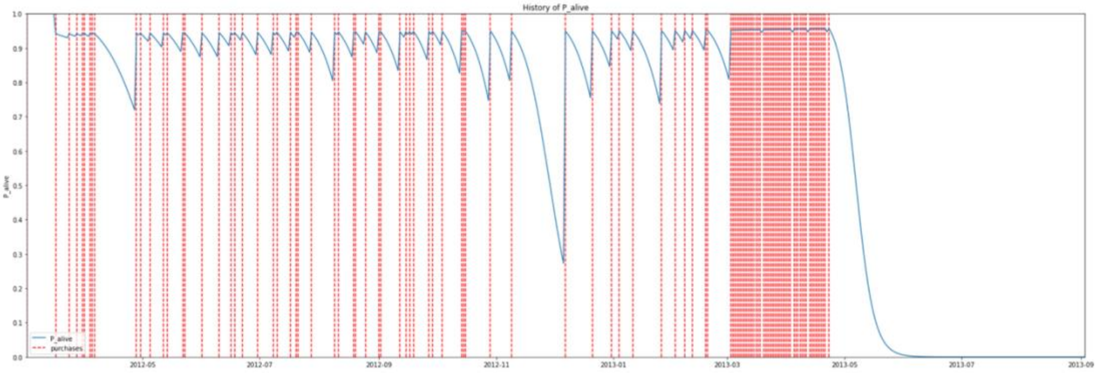


Figure 11 A random individual's probability of being alive – Gamma Gamma model

In general, the Pareto/NBD model and its extended Gamma-Gamma model illustrate well the real-life situation of the customer behavior in the observation period. But we can conclude that the Gamma Gamma sub-model in general is better at following a trend therefore performs better than the standard Pareto/model. By looking at the Single customer probability histories plot, we can see that the probability of the customer staying alive after each purchase is more realistic than in figure NO. This further support our conclusion.

4.2 XGboost

As it is discussed in section 3.5, the model with the best performance was XGBoost. Before we dive into the prediction, we try to tune the model.

XGboost has many parameters that can be tuned. In our case we chose to tune three of these hyperparameters: learning_rate, max_depth and subsample. By appliying a grid search we were able to produce plots that showed optimal values for learning_rate, max_depth and subsample.

Figure 12 visualizes the optimal value of max_depth calculated by grid search method. The plot shows that max_depth= 3 have the best best performance.

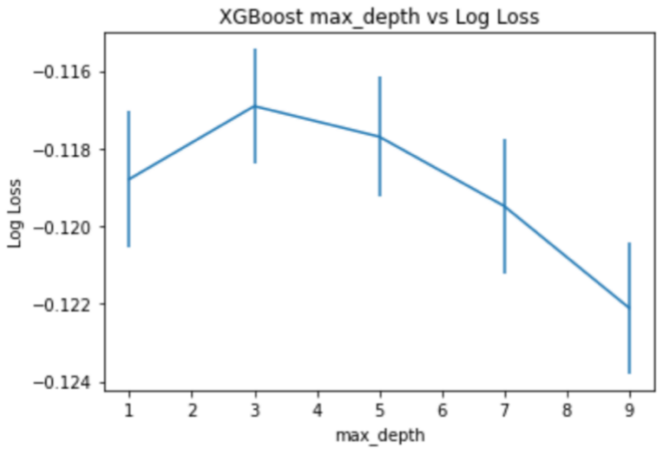


Figure 12 XGBoost max_depth VS log loss

Figure 13 visualizes the optimal value of max_depth and n_estimators. The top results are max_depth= 2 with n_estimator of 200 and the second best is max_depth = 3 with n_estimators of 100. However, when tested in the model, we found that max_depth =3 with n_estimators of 100 give better performance than max_depth = 2, with n_estimators of 200 or lower.

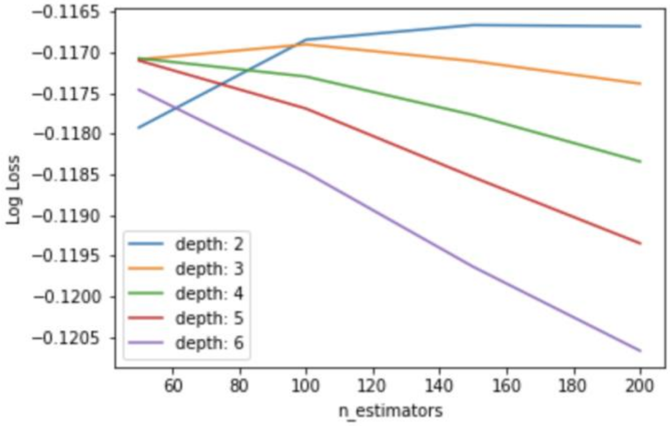
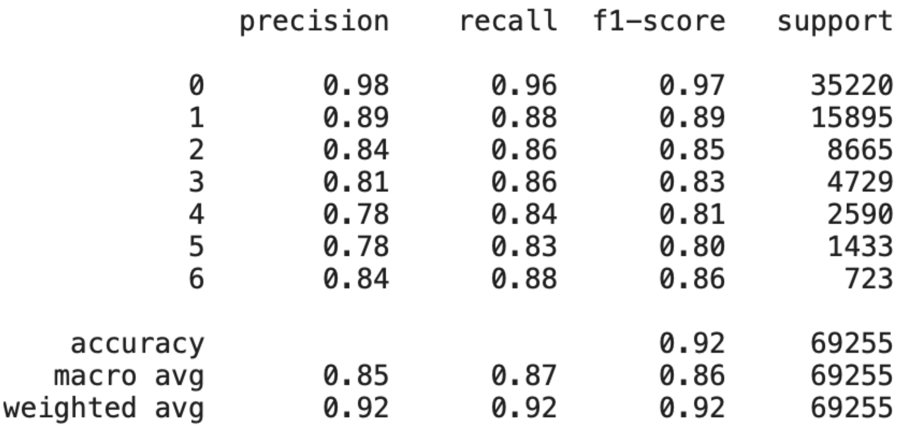


Figure 13 Optimal max_depth of XGBoost

The model is then fitted based on the train and test set and by using the precision and recall function we can evaluate the model's performance.

To evaluate the model, we first use P/R/A to. From Figure 14 we can see that the overall accuracy of the model is a very high 92%. The precision and recall for each of the segments are high as well. This tells us the model performs extremely well when classifying customers to these segments.



```
              precision    recall  f1-score   support

           0       0.98      0.96      0.97     35220
           1       0.89      0.88      0.89     15895
           2       0.84      0.86      0.85      8665
           3       0.81      0.86      0.83      4729
           4       0.78      0.84      0.81      2590
           5       0.78      0.83      0.80      1433
           6       0.84      0.88      0.86       723

    accuracy                           0.92     69255
   macro avg       0.85      0.87      0.86     69255
weighted avg       0.92      0.92      0.92     69255
```

Figure 14 Sensitivity and specificity analysis of XGBoost- 7 clusters

Below in Figure 15 we can also see an ROC plot of which visualy represents how much of the data is classified correctly. Here also, we see that AUC for all the segments is between 0.99 and 1.00.

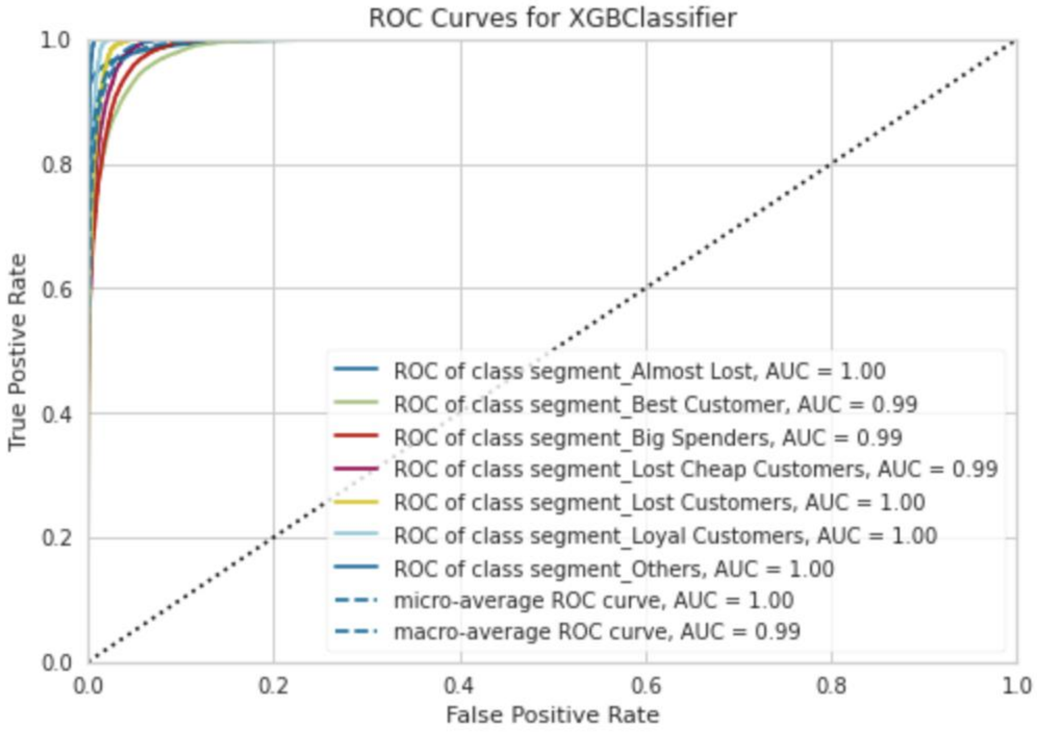


Figure 15 ROCAUC- 7 clusters

Besides the analysis above, a class prediction error for XGB Classier is shown in the Figure 16 below plot. It looks that the model with seven clusters succesufuly targets the right customer segmentation.

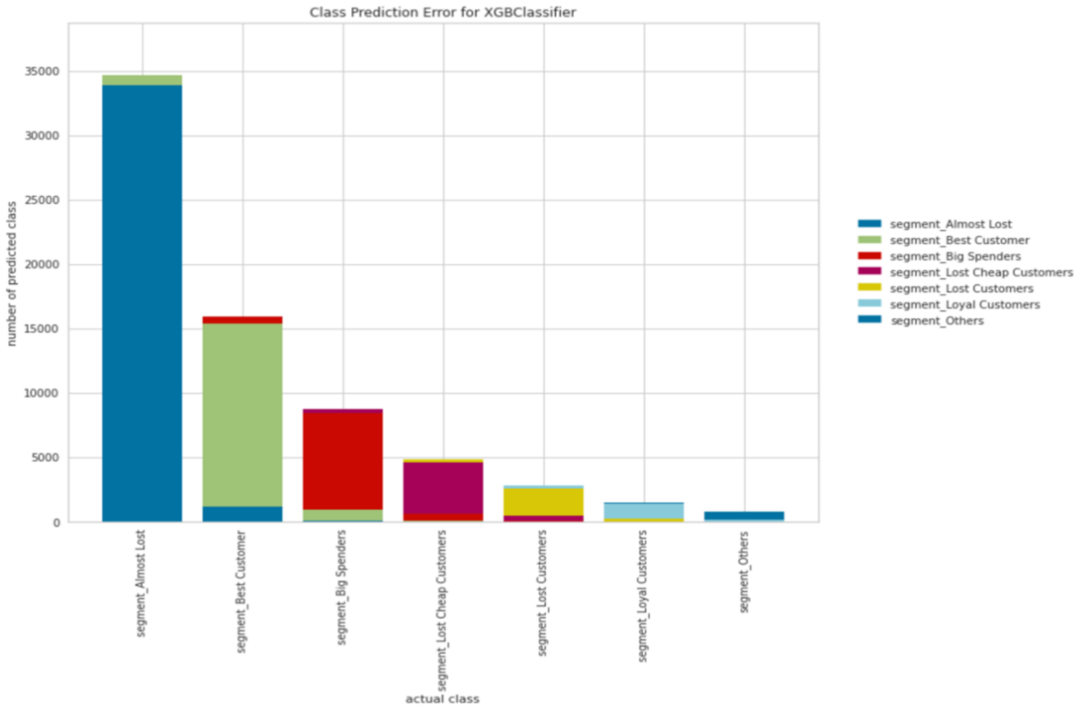


Figure 16 Class prediction error for XGB Classier

It is a little worrisome that the accuracy is as high as we see, as these are often dreamlike results, and almost never achieved. However, we have to account for the low number of segments, which is the reason for the such a high accuracy of the model. Increasing the number of segment would definitely decrease the accuracy, however in the context of our research aim, we believed this result makes the point across that machine learning models might be a good alternative to the pareto/NBD model. This is further discussed in the next sextion.

4.3 Comparison of the Pareto/NBD and XGboost models

The table below is a summary of the performance according to the test set with both traditional and ML methods. The performances are measured by MAE, MSE, RMSE and R-squared of the customers.

	MAE	RMSE	R - Squared
Pareto/ NBD& Gamma-Gamma	0.49421	4.28397	0.98131
Xgboost	0.09124	0.33172	0.94004

Table 3 Comparison of performance

The XGboost model have the best scores for MAE and RSME. The Pareto/NBD perfomers better when evaluted with R squared.

The XGboost clearly outperformes the Pareto/NBD model. However, the XGboost model predicts for only 7 segments, and this is therefore no surprise. As the difference in the MAE and RMSE score are not drastic, it is not unreasonable to think that the Pareto/NBD model would outshine the XGboost, if the number of segments increases considerably.

5. Conclusion

The right CLV prediction creates significant value in identifying the most profitable customer, which increases the firms' profits. The purpose of this research is to compare the probabilistic models with the ML models in order to find an optimal method of calculating CLV. Different models are presented and compared according to predictive power. Pareto/NBD, Gamma-Gamma, and XGBoost models are used in this paper, which represents the competition between traditional and ML methods in CLV prediction. XGBoost models have stronger predictive power than probabilistic models; however, only when predicting which customers belong to which segments. Based on the evolution measures, MAE and RSME, we believe that the Pareto/NBD model would perform far better if compared with a Machine Learning model that predicts the CLV directly, rather than the segments. If the aim of a company is to find the most profitable group, we believed Machine Learning models would do a better job. If the aim is to find a very specific group for a marketing campaign, the Pareto/NBD model seems to perform better.

The weakness of this research paper is the limited data and the possibly unfair comparison of the models. Each of the models performs best in their own domain, and therefore results of the comparison of the models are not very clear.

A future direction for further work on this topic might be to use data with a longer period, either increase the number of segments considerably or use a machine learning model to directly predict each customer's CLV. We believe this would result in a much fairer comparison, as both of the models predict the same target, and it will belong to the same domain.

Bibliography

Petter Flordal & Joakim Friberg. (2013). Modeling Customer Lifetime Value in the Telecom Industry . Lund University.

Dwyer, F. R. (1997). Customer Lifetime Valuation to Support Marketing Decision Making .

Eva Ascarza et al. (2018). In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions .

Aslekar Avinash,Piyali Sahu, Arunima Pahari . (2019). Big Data Analytics for Customer Lifetime Value Prediction.

Peter S. Fader ,Bruce G. S. Hardie, Ka Lok Lee. (2004). "Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model.

Fader and Hardie. (2009). Probability Models for Customer-Base Analysis. Advanced Research Techniques Forum.

C.H. Cheng and Y.S. Chen . (2009). Classifying the segmentation of customer value via RFM model and RS theory. Expert systems with Applications, 4176-4184.

Inderpreet Singh &Sukhpal Singh. (2017). Framework for Targeting High Value Customers and Potential Churn Customers in Telecom using Big Data Analytics.

Arick, D. (2019). Retrieved from https://itsecuritycentral.teramind.co/2019/11/01/using-machine-learning-for-behavioral-analysis-to-detect-insider-threats/

Hosmer, D.W. and Lemeshow, S. (2000). Applied Logistic Regression.

Smith, K.A. and Gupta, J.N.D. (2000). Neural networks in business: techniques and applications for the operations researcher. Computers & Operations Research, Vol. 27 Nos 11/12, pp. 1023-1044.

Haykin, S. (1999). Neural Networks: A Comprehensive Foundation, 2nd ed. Prentice-Hall.

Quinlan, J. (1986). Induction of decision trees. Machine Learning, Vol. 1 No. 1, pp. 81-106.

Tamaddoni Jahromi, A., Stakhovych, S. and Ewing, M. . (2014). Managing B2B customer churn, retention and profitability. Industrial Marketing Management, Vol. 43 No. 7, pp. 1258-1268.

Buckinx, W. and Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. European Journal of Operational Research, Vol. 164 No. 1, pp. 252-268.

Jing Zhou, Wei Li , Jiaxin Wan, Shuai Ding, Chengyi Xia. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. Elsvier.

Kenny, R. &. (1990). The hidden advantages of customer retention. Journal of Retail Banking, 12(4), 19–23. .

Dick, A. a. (2003). Customer loyalty: toward an integrated conceptual framework", Journal of the Academy of Marketing Science, Vol. 22 No. 2, pp. 99-113.

Lai, X. (2009). Segmentation study on enterprise customers based on data mining technology", IEEE First International Workshop on Database Technology and Applications, Wuhan, Hubei, pp. 247-250. .

Payne, A. a. (2004). The role of multichannel integration in customer relationship management", Industrial Marketing Management, Vol. 33 No. 6, pp. 527-538. .

Kincaid, J. (2003). Customer Relationship Management: Getting It Right, Prentice-Hall PTR, New Jersey, NJ. .

Ngai, E. X. (2009). "Application of data mining techniques in customer relationship management: a literature review and classification", Expert Systems with Applications, Vol. 36 No. 2, pp. 2592-2602. .

Parvatiyar, A. a. (2001). "Customer relationship management: emerging practice, process and discipline", Journal of Economic and Social Research, Vol. 3 No. 2, pp. 1-34.

Umayaparvathi, V. a. (2012). “Applications of data mining techniques in telecom churn prediction”, International Journal of Computer Applications, Vol. 42 No. 20, pp. 0975-8887. .

Agarwal, A. H. (2004). “Organizing for CRM”, McKinsey Quarterly, Vol. 3, pp. 80-91. . Kotler, P. a. (2006). Marketing Management, 2nd ed. New Jersey, NJ. : Pearson Prentice Hall. Swift, R. (2001). Accelerating Customer Relationships Using CRM and Relationship

Technologies. New Jersey, NJ. : Prentice-Hall PTR.
Van den Poel, D. a. (2004). Van den Poel, D. and Larivie’re, B. (2004), “Customer attrition analysis for financial service using proportional hazard models”, European Journal of Operational Research, Vol. 157 No. 1, pp. 196-217.
Berger PD, Nasr NI. (1998). Customer lifetime value: marketing models and applications. J

Greenberg, P. (2009). CRM at the Speed of Light: Social CRM 2.0 Strategies, Tools, and Techniques for Engaging your Customers. United States of America: McGraw Hill Professional.

Gutha Jaya Krishna, Vadlamani Ravi . (2016). Evolutionary computing applied to customer relationship management: A survey.

Eva Ascarza et al. (2018). In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions .

Aslekar Avinash,Piyali Sahu, Arunima Pahari . (2019). Big Data Analytics for Customer Lifetime Value Prediction.

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), 1-24. .

Peter S. Fader ,Bruce G. S. Hardie, Ka Lok Lee. (2004). “Counting Your Customers” the Easy Way: An Alternative to the Pareto/NBD Model.

Hardie, B. (n.d.). Retrieved from Tutorials and excel spreadsheets of NBD model: http://www.brucehardie.com/talks.html

Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozzafari, M. and Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining technique. Applied Soft Computing, pp. 994-1012.
Fader and Hardie. (2009). Probability Models for Customer-Base Analysis. Advanced

Research Techniques Forum.
Petter Flordal & Joakim Friberg. (2013). Modeling Customer Lifetime Value in the Telecom

Industry . Lund University.
C.H. Cheng and Y.S. Chen . (2009). Classifying the segmentation of customer value via RFM model and RS theory. Expert systems with Applications, 4176-4184.

Inderpreet Singh &Sukhpal Singh. (2017). Framework for Targeting High Value Customers and Potential Churn Customers in Telecom using Big Data Analytics.

Arick, D. (2019). Retrieved from https://itsecuritycentral.teramind.co/2019/11/01/using- machine-learning-for-behavioral-analysis-to-detect-insider-threats/

Hosmer, D.W. and Lemeshow, S. (2000). Applied Logistic Regression.
Smith, K.A. and Gupta, J.N.D. (2000). Neural networks in business: techniques and applications for the operations researcher. Computers & Operations Research, Vol.

27 Nos 11/12, pp. 1023-1044.
Haykin, S. (1999). Neural Networks: A Comprehensive Foundation, 2nd ed. Prentice-Hall.

Quinlan, J. (1986). Induction of decision trees. Machine Learning, Vol. 1 No. 1, pp. 81-106. Buckinx, W. and Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. European Journal of Operational Research, Vol. 164 No. 1, pp. 252-268.

Freund, Y. and Schapire, R.E. (1996). Experiments with a new boosting algorithm. 13th

International Conference on Machine Learning (pp. pp. 148-156). NJ: Murray Hill. Hung, S.Y. and Wang, H.Y. . (2004). Applying data mining to telecom churn management.

Pacific Asia Conference on Information Systems (PACIS).

Qureshi, S.A., Rehman, A.S., Qamar, A.M., Kamal, A. and Rehman, A. (2013). Telecommunication subscribers' churn prediction model using machine learning. Eighth International Conference on Digital Information Management (ICDIM) (pp. pp. 131-136). Islamabad: IEEE.

Runge, J., Gao, P., Garcin, F. and Faltings, B. (2014). Chum prediction for high-value players in casual social games. Proceeding of 2014 IEEE Conference on Computational Intelligence and Games. Washington, DC: IEEE Computer Society Press.

Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G. and Chatzisavvas, K.C. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, Vol. 55, pp. 1-9.

Yu, X., Guo, S., Guo, J. and Huang, X. . (2011). An extended support vector machine forecasting framework for customer churn in e-commerce. Expert Systems with Applications, Vol. 38 No. 3, pp. 1425-1430.

Tamaddoni Jahromi, A., Stakhovych, S. and Ewing, M. . (2014). Managing B2B customer churn, retention and profitability. Industrial Marketing Management, Vol. 43 No. 7, pp. 1258-1268.

Juan, G., Luo, S., Jia, H., Zhang, T. and Han, Y. . (2007). Type 2 diabetes data processing with EM and C4.5 Algorithm. IEEE Complex Medical Engineering Internaonal Conference, (pp. pp. 371-377). Beijing.

Williams, P.H., Eyles, R. and Weiller, G. . (2012). Plant microRNA prediction by supervised machine learning using C5.0 decision trees. Journal of Nucleic Acids.

Crone, S.F., Lessmann, S. and Stahlbock, R. (2006). The impact of preprocessing on data mining: an evaluation of classifier sensitivity in direct marketing. European Journal of Operational Research, Vol. 173 No. 3, pp. 781-80.

Guo-en, X. and Wei-dong, J. . (2008). Model of customer churn prediction on support vector machine. Systems Engineering–Theory and Practice, Vol. 28 No. 1, pp. 71-77.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer Verlag.

Ali, O.G. and Arıtürk, U. (2014). Dynamic churn prediction framework with more effective use of rare event data: the case of private banking. Expert Systems with Applications, Vol. 41 No. 17, pp. 7889-7903.

Glady, N., Baesens, B. and Croux, C. (2009). Modeling churn using customer lifetime value. European Journal of Operational Research, Vol. 197 No. 1, pp. 402-411.

Tsai, C.F. and Chen, M.Y. (2010). Variable selection by association rules for customer churn prediction of multimedia on demand. Expert Systems with Application, Vol. 37 No. 1, pp. 2006-201.

Huang B, Kechadi MT, Buckley B. (2012). Customer churn prediction in telecommunications. Expert Systems with Applications, Vol. 39 No. 1, pp. 1414–1425.

Abinash Tripathy , Ankit Agrawal, Santanu Kumar Rath. (2016). Classification of sentiment reviews using n-gram machine learning approach. Expert Systems With Application.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.

Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82–89 .

Gautam, G., & Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. Contemporary computing (IC3), 2014 seventh international conference (pp. pp. 437–442). IEEE.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. Springer.
Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: opinion

extraction and semantic classification of product reviews. 12th international

conference on World Wide Web (pp. pp. 519–528). ACM.
Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with

diverse information sources. EMNLP, 4 (pp. 412–418.
Matsumoto, S., Takamura, H., & Okumura, M. . (2005). Sentiment classification using word sub-sequences and dependency sub-trees. Springer, pp. 301–311.

Bryman, A., and Bell, E. (2015). Business Research Methods. 4th ed. Oxford University Press.

Machine Learning in CLV Prediction. Retrieved from

https://cloud.google.com/solutions/machine-learning/clv-prediction-with-offline-

training-intro

Merton, Robert K. (1973). The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago Press

Neil J. Salkind (2010). Encyclopedia of Research Design. Sage Publications, Inc.

Cheryl B. Thompson, Edward A. Panacek (2006). Basics of research. Air Medical Journal