DOI: 10.37943/ILMM7870

Shyngys Akhmetbek
Master's Student of the Faculty of Information Technologies
s_akhmetbek@kbtu.kz, orcid.org/0000-0001-5511-349X
Kazakh-British Technical University, Kazakhstan

FORECASTING CUSTOMER FUTURE BEHAVIOR IN RETAIL BUSINESS USING MACHINE LEARNING MODELS

Abstract: The ability to forecast customers' future purchases, lifetime value, and churn are fundamental tasks in business management. These tasks become more complicated when the relationship between customers and business is not contractual. Therefore, the application of an appropriate method of customer analysis influences the efficiency of company cost management in interaction with their customers. The purpose of this paper is to compare existing solutions of customer lifetime value prediction and provide a new way to predict the future behavior of customers with consideration of the drawbacks of previous works. The method should have the following properties: use data that is available in any retail business; take into account that markets are constantly changing; be more precise than existing solutions. In this paper, we proposed the method of identifying customer churn provided a way to analyze customer behavior associated with churn or retention. In order to understand why customers churn, we used eleven customer behavioral metrics. The relationship of used metrics with churn was proved using churn cohort analysis. The results of training of logistic regression and neural network on prepared dataset showed that their forecast accuracy is in the healthy range for highly predictable churn. Based on predicted churn probabilities, we calculated the customer lifetime value in the future period. Our research results on customer behavior in the retail business confirm the hypothesis that customers who make many purchases are less likely to churn than customers who make few purchases. The main uniqueness of this work is the way of finding customer churn, as no such data was provided in the initial dataset. In addition, the minimum amount of data that most retail companies have was used. This enables the proposed methodologies to be applied to a large number of retail companies.

Keywords: churn, marketing, customer lifetime value

Introduction

Customer lifetime value (CLV) is a metric that evaluates customer importance to a company. This metric helps to predict clients' future monetary value during their interaction with a company. This metric makes it easier to understand a reimbursement of companies' costs on customers' acquisition and retention. Thus, the proper way of calculating this metric plays an important role in business management and marketing costs.

There are many ways to estimate CLV. In most cases, these methods are based on the amount of monetary value from the customer's purchases in the past. According to these methods, if we know how much a client has spent in the past, we can assume that the same client will spend the relatively same amount of money in the future. This assumption is true only if a client's interest in the company's products will not change in the future. But this phenomenon is quite rare, since the client's interest is frequently changed. Therefore, it is necessary to take into account the current activity of the client when calculating CLV. Because the client's future

Copyright © 2022, Authors. This is an open access article under the Creative Commons CC BY license

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 10, JUNE 2022

behavior depends on how often the customer made purchases in the past or how long ago the purchase was made.

We have compared methods of CLV prediction proposed in [1-3] in retail. We have tested all methods on a publicly available Online Retail dataset for two years from 2009 to 2011 [4-5]. This is a transactional dataset of customers of an online store in the UK that sells unique gift items for all occasions. The Online Retail dataset contains main information that companies have in order to analyze their customers.

To compare the models, only a sample of transactions for 2011 from the entire data set was used. We prepared a sample for each method of CLV estimation. We have compared the results of forecasting three models. They are linear regression, neural networks, and the Gamma-Gamma model. All these models were trained on the first three quarters and the predicted values of each model were tested on the values from the last quarter.

We evaluate the correctness of these methods using mean absolute error in our experiment. This metric allows us to estimate the average error between predicted values with real values. The total sum of purchases is represented in pounds sterling in the dataset, so we need to take this into account in interpreting the results.

According to the results, the linear regression makes an error of 190 pounds. The error of probabilistic models was 423 pounds. The error was equal to 811 pounds for neural networks. We see that the linear regression has a low error of CLV prediction in the next 3 months. This model is useful only if we know for sure that the client will come in the future. In order to understand how accurate the probabilistic model is, we need to compare the average amount of purchases predicted by the model with the average monetary value during the test period. The average monetary value during the test period is 309 pounds, and the predicted amount is 655 pounds. As we can see, the difference between the real and predicted values is almost two times. We assume that a large error in neural networks is associated with the non-representativeness of the features used for this task. The features used do not provide sufficient information to predict CLV.

Based on the results of the experiment, we assume that it is necessary to use another method of CLV estimation. In this method, we need to take into account the limitations of previous works. In the rest of our article, we examine how to estimate CLV based on this work [6]. The author demonstrates a lot of useful techniques about how to calculate the probability of customer churn. However, the author of that work considered these techniques in the context of a business where products or services are sold on the basis of subscriptions. Our task is to modify those techniques from that work for the retail type of business and check how accurate the predicted CLV will be on the Online Retail dataset.

Literature review

One of the most common methods of predicting customer behavior is to calculate the probability of future purchases [7]. This probability gives a forecast of whether the next purchase will happen or not in a certain period of time. If we know the probability of future purchases and the monetary value that every customer has spent in the past, we can estimate what monetary value we will receive from this customer in the future. Numerous studies have focused on this method of prediction.

One of the important and modern ways of predicting the future behavior of customers is a group of statistical models [1], [8-10]. This type of model is based on a combination of different probability distributions. This kind of combination simulates customer buying behavior. These models are used in retail businesses, where the process of selling the company's product directly to the customer takes place. Each of these models has its own features and predicts a certain activity of the client.

DOI: 10.37943/ILMM7870
© S. Akhmetbek

All probability models evaluate customer behavior based on four values: the recency of purchase, the frequency of purchases, the length of time during which the customer made its purchases and the average monetary value. Companies can evaluate the future behavior of customers depending on each of these values. For example, if a customer's last purchase was made a very long time ago, then the probability of his next purchase is very low. The more a client purchases and spends money, the more likely it is that the client is interested in the product of the business. This tells us that the customer is most likely to make a purchase again. The longer a customer stays with a company, the more likely they are to be loyal to the company's product. Based on these assumptions, statistical models predict the future probability of purchase, the expected number of purchases, and CLV.

The Pareto/NBD model studied in [8] calculates the estimated number of transactions and the probability of purchase in a certain period of time. In order to calculate these values, the model uses purchase recency and the number of purchases in the observed period. This model is one of the first and a benchmark for later probabilistic models.

In the next paper [9], a new BG/NBD model is studied, which simplifies the process of calculating the estimated number of transactions and the probability of purchase compared to the previous method. The main difference is that the Pareto/NBD model assumes that a customer can become inactive at any time, regardless of whether a repeat purchase has been made or not. This assumption has been changed in the BG/NBD model. The authors suggest that the probability that a customer will become inactive changes after each of his repeated purchases [9]. A repeat purchase is considered to be any purchase after the very first one. If customers make a large number of purchases, then the BG/NBD model can be applied to them. However, if the frequency of purchases is low, then this model does not reproduce customer behavior very well. Nevertheless, due to the simplicity of calculations, this model is widely used by companies to predict the behavior of their customers by taking into account model limitations.

The authors in [10] modified the previous BG/NBD model by taking into account all customers regardless of their number of purchases during the observed period. In this model, the probability of a customer's next purchase is calculated for all customers after each of their purchases. The calculation of these probabilities differs from the probability that was used in the Pareto/NBD model. In the Pareto/NBD model, the likelihood of the next purchase is reduced depending on how long it has been since the previous purchase. However, in this model, this probability does not decrease. It remains the same without changes until the next purchase [10]. Companies can use a modified version of the BG/NBD model to analyze their customers, despite their number of purchases.

We can estimate the number of future purchases using reviewed statistical models. However, we cannot evaluate the customer's importance by using only this value. Since the customer can make purchases very often, but at the same time, the monetary value of his purchases may be small. Assuming that the company has a client who does not make purchases often, but the monetary value of each purchase is very high. It is more important to retain this client for the company, because he spends more money than the client from the previous example [1]. This behavior of customers can be taken into account using the same models that we reviewed before. However, now it is necessary to consider how much money the customer spends on purchases.

The authors suggest to use two values in order to calculate the CLV in work [1]. The first value is the expected number of customers' future purchases obtained from the Pareto/NBD model. The second value is the average monetary value that we can get for each client. The authors suggest to use the Gamma-Gamma model in order to calculate a future monetary value. This model first calculates the average monetary value for all observed clients, then

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 10, JUNE 2022

averages that value for each client. Multiplying the expected number of future purchases by the average monetary value gives a forecast of how much revenue we can get from the client in the future. Thus, CLV can be calculated with the help of probabilistic models.

There are many other ways to predict CLV than those we have reviewed previously. These methods are based on the use of machine learning algorithms. These algorithms are divided into two types depending on which client activity should be predicted. When it is necessary to predict two output values, for example, whether a customer will make a purchase or not, then we need to use models for classification tasks. If it is necessary to predict the monetary value that customers will spend on purchases, then we use the need to use models for regression tasks. In regression problems, continuous numerical values are predicted.

The authors of this paper [2] use a linear regression algorithm and customer purchase history to predict CLV for a certain period of time. CLV was calculated based only on three features in this study. Those features are the total sum of purchases, the average value of purchases and the total number of purchases. These values were calculated quarterly. This solution was applied in order to simulate customers' purchasing behavior from quarter to quarter. By analyzing the customer's past quarterly monetary value, the linear regression predicts the monetary value in the next quarter. This method of predicting the customer's life value is quite simple to implement and apply. According to the authors' results, the predicted values are close to the real values in the test period.

The authors of the following work [3] propose a method of predicting CLV using neural networks. The purpose of their work is to compare the accuracy of neural networks and probabilistic models based on the purchase history of online store customers. First, they divide dataset into two parts. They are calibration and holdout periods. During the calibration period, they aggregate features related to the client's behavior. They calculate the future monetary value of purchases in the holdout period. The future monetary value is considered to be CLV. The authors have used nine features of customer behavior to train neural networks. Those features are the total sum of all purchases, the number of days between the first and the last purchases, the total number of purchases, the number of days between the very first purchase and the end of the calibration period, the average frequency of purchases, the average monetary value and the total number of purchased products. They trained a neural network to predict the future monetary value of customer purchases. The results of their research showed that neural networks more accurately predict CLV during the holdout period than probabilistic models.

Despite the fact that in previous works [2-3] regression models more accurately predicted CLV, they have a significant drawback. These models do not take into account the probability of a future purchase. For instance, if a model provides that an already churned customer will make a purchase in the future, it will not be quite correct. Therefore, these methods are applicable only if we know for sure that the customer will make a purchase in the future time period. However, in real life, we do not know in advance whether the customer will make his purchase or not during the period we are interested in. Therefore, it is necessary to predict CLV by taking into account the probability of customer retention.

The author of the next study [6] proposed the method of CLV prediction based on a customer lifetime. Customer lifetime is the span of time during which a customer makes purchases. We can estimate more precisely how much money each client will spend in the observed future by using their lifetime. If the probability of the next purchase is a monthly forecast, the client's lifetime is 1 divided by that probability in months [6]. Let us assume that the probability of customer's next purchase in future month is 30%. If we divide 1 month by this probability, then we get that the lifetime of this client is three months. We can calculate customer's total sum of purchases per month and multiply it by the predicted lifetime. So, we will get the average amount of purchases that the customer will make in the next three months. The main key of

DOI: 10.37943/ILMM7870
© S. Akhmetbek

this CLV estimation is the probability of the customer's next purchase. The author used logistic regression in order to find that probability. This model allows us to estimate the probability of churn or retention of each customer by taking into account buying behavior. The author examined this method of CLV prediction in companies that sell a subscription-based product or service. The results of author's work showed high accuracy in predicting future purchases of the client. Since, this probability is used in the estimation of a lifetime, we can conclude that the CLV is also accurate.

Research methods

The goal of any retail business is to increase its revenue by attracting new customers and increasing sales. The larger the company's customer base, the more sales are made. Therefore, many companies make great efforts to acquire new customers. However, in order to have a highly profitable business, the company must also engage in customer retention. Since if the number of customers who stop using the company's product is greater than the number of purchased customers, the company will incur large losses. Having a large number of churned clients, companies will have unstable income. Therefore, a decrease in the number of churned customers has a positive impact on customer retention. In addition, it is important to know why your customers stop their purchases. By knowing the reasons or behaviors that affect this kind of customers decision, companies can take certain actions to retain their customers. Consequently, this will have an impact on income growth.

As we can see, the company's revenue depends most of all on understanding why their customers churn or stay. These reasons can be investigated by analyzing historical data of customer purchases. The algorithm of analysis consists of 4 main stages (Figure 1). In the rest of the article, we give a description of the work done at each stage.

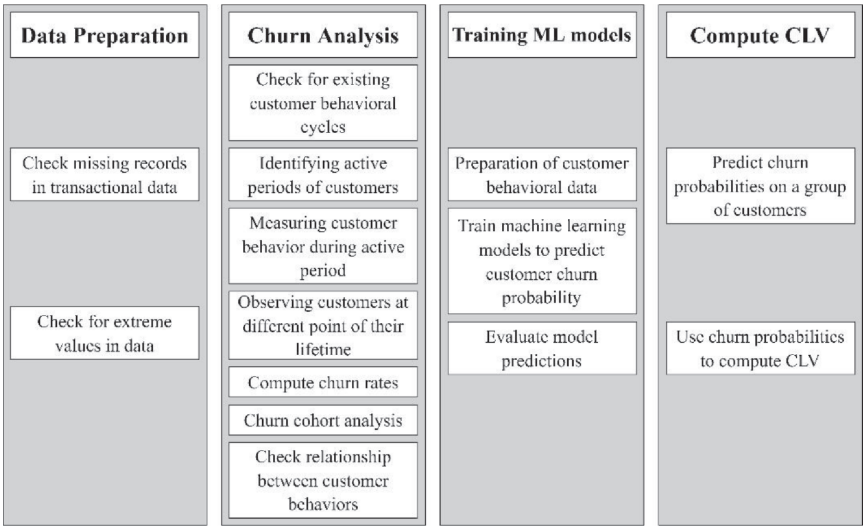

Figure 1. Workflow of research methodology

I. Data preparation

Initial Online Retail dataset contains 8 columns and 1067371 rows. The columns are invoice number, unique product code, product name, the quantities of each product per transaction, the date and time when a transaction was made, product price per unit in pound sterling, unique customer id and the country of transaction. Online Retail dataset consists of two parts. The first part is transactional data between 2009 and 2010, and the second is between 2010 and 2011. Before going deeper into the analysis, a number of preprocessing stages are performed to process the data, as shown below:

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 10, JUNE 2022

1. We have found that there are a lot of duplicated instances after combining two sets of data. Therefore, we have removed them, because they can have a bad effect on the analysis.

2. We checked the presence of missing values. A huge number of the missing values occur in the unique customer id column. Since the customer id is the main key to aggregate data for each customer, transactions without a customer id do not carry any information. So, we have removed instances without a customer id.

3. The authors in the dataset description noted that if the invoice number starts with the letter «C», it means that the transaction was cancelled. Due to the fact that the authors don't provide enough information about how to work with this type of transaction, we didn't examine these instances in our analysis.

4. There are many uncertain product names in the dataset. Such as «ADJUST», «BANK CHARGES», «DOT», «TEST001», etc. We think that these records are not related to customer purchases. So, we do not consider these records.

5. More than 90% of transactions were made in the UK and the remaining parts were made in other countries. We have used only a sample of transactions that were made in the UK in our analysis.

6. One customer's purchase is divided into several transactions in this dataset. Instead of using initial format of transactional data, we have grouped customer transactions by day.

7. We have removed extreme values from the initial dataset.

We made a timeseries summary after preprocessing stages (Figure 2). This summary demonstrates the following things: counting number of purchases per day over 2 years, the total number of sold products per day over 2 years, and total monetary value per day over 2 years. The results of the preprocessing stage show that there are no extreme values left in the data. In addition, we see that the dataset covers almost all two years, and there are no missing periods in the data.

We can see that there are seasonal trends in customer purchases (Figure 2). Also, purchases follow a weekly cycle. No purchases are made on Sundays. Consequently, if we analyze clients during time intervals that are equal to 1 week, we can take into account the entire seasonality of clients' behavior. That is why in the next stages of our analysis we use only a time measurement equal to 1 week.

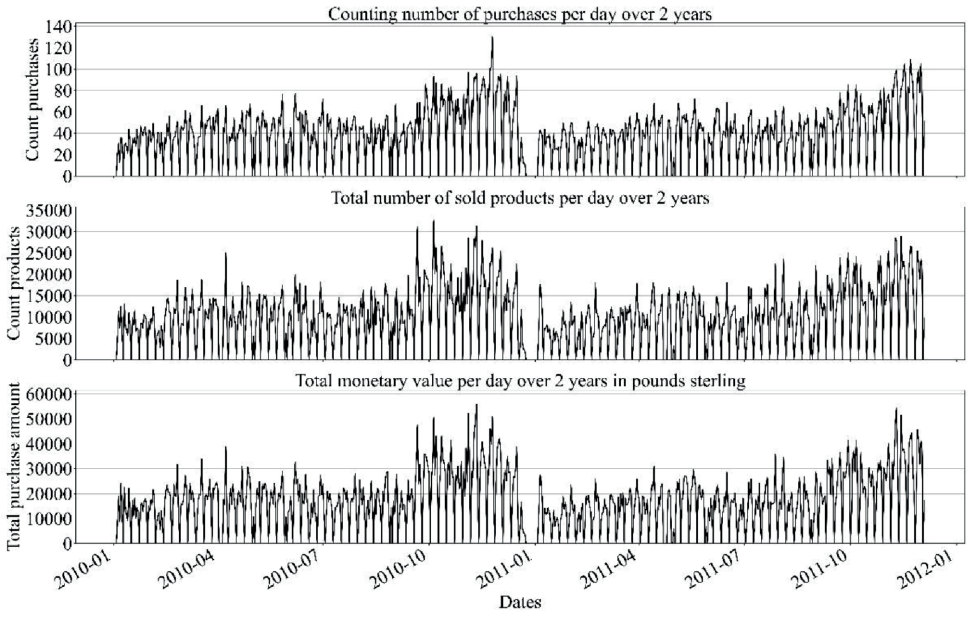

Figure 2. Timeseries summary of the dataset after preprocessing stage

DOI: 10.37943/ILMM7870
© S. Akhmetbek

II. Churn analysis

The next stage of our analysis is to observe customers in the right way. You may ask what does this mean? Observing customer is the analysis of customer behavior at different stages of the customer's lifecycle. Customers are always faced with two decisions during their lifecycle [11]. Such as to churn or continue to purchase the company's products. If we can understand what customer behavior leads to churn or return, we can make better decisions about reducing churn.

III. Active periods of customers

The first step in this process is to find the customer lifecycle. The lifecycle is the period of time that a client makes purchases actively in which none of the purchases are further than the allowed period of inactivity. If we analyze this span of time, we can find out which customer behavior was associated with making repeated purchases or churn. To begin with, we need to determine what period of time we can allow a typical customer to be inactive. We can compute the number of days between customer purchases in order to find the allowed gap. The average number of days between customer purchases on our dataset is equal to 110 days (Figure 3), but we don't use that value. Unlike the average, the median is stable to existing outliers and asymmetric distribution. That's why we have used the median number of days between purchases as the allowed gap of inactivity.

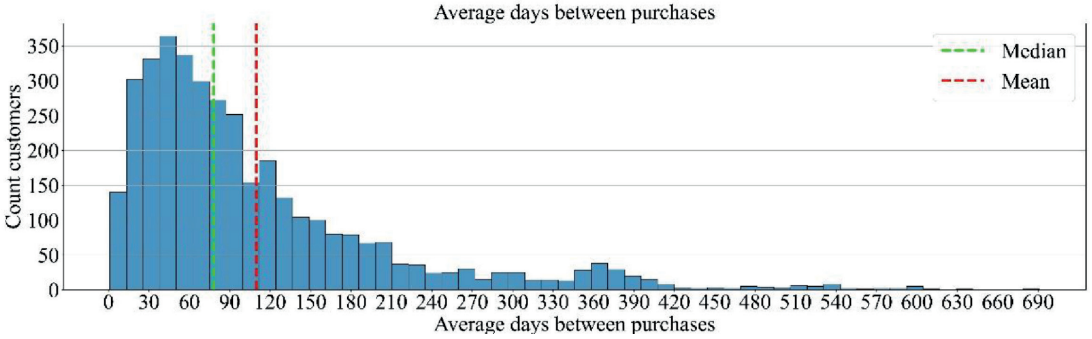

Figure 3. Histogram of the average days between customer purchases

Let us look at an example of determining the active period of a client. A hypothetical customer made a sequence of purchases during his/her lifecycle. Each of the purchases has been made with a gap of one week. This customer made his purchases over seven weeks. After the last purchase was made in the seventh week, customer stopped buying any products. Five weeks later, the customer started shopping again every week. The hypothetical customer has two lifecycles in this example. The first is when he makes purchases within 7 weeks. This active period began in the first week and ended in the seventh week. After the seventh week, the customer didn't make any purchases for a long time. We know that the average number of days between purchases in this case is equal to 7 days. This time is considered as an allowed gap of inactivity for this customer. Therefore, we consider that this client has churned. After five weeks, the customer started shopping again. Now we consider this period of active purchases as a new active period that hasn't ended yet. Based on observed steps of identifying active periods, we have found all customers' lifecycle periods by using an allowed gap of 12 weeks.

IV. Customer behavioral metrics

The next stage of our analysis is to measure customer behavior during active periods. Data that relates to how much money a customer pays for the amount of service consumed or used is one of the most important features in retail business [6]. The number of items per

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 10, JUNE 2022

transaction, the date of transaction and cost of purchase columns of the Online Retail dataset have been used to create features. We have created 11 features to analyze customer behavior based on data from the pre-processing stage (Table 1).

Table 1. The list of features that used in churn analysis

Feature name	Feature description
n_purchase_per_2month	Total number of purchases per 2 months
n_product_per_2month	Total number of purchased items per 2 months
sum_purchase_per_2month	Total cost of purchases per 2 months
avg_n_product_per_2month	Average number of purchased products per 2 months
avg_purchase_per_2month	Average cost of purchase per 2 months
account_tenure	The length of time during which a customer makes frequent purchases
sum_purchase_per_n_product	Average cost of product per 2 months
n_purchase_pcnt_chng_4week	Percentage change in a number purchases compared to the previous month
n_product_pcnt_chng_4week	Percentage change in a purchased number of products compared to the previous month
sum_purchase_pcnt_chng_4week	Percentage change in a total cost of purchases compared to the previous month
days_since_purchase	Number of days since the most recent purchase

The features of customer behavior can be divided into two types. They are simple and advanced features. In our study, simple features are the total number of purchases, the total number of purchased items, the total cost of purchases, the average number of purchased products, and the average cost of purchase. These features have been calculated by counting, summing, and finding the average value of customer purchases. Simple features are good for segmentation, but not for using them to train machine learning models. Because they are too correlated. The high correlation between features makes it hard to create patterns of customer behavior. This is a problem when some customers belong to one cohort of behavior on one feature and another cohort on the second feature. In this case, the model cannot understand which behavior is more important than the other one. That's why it is necessary to create advanced features that have moderate correlation and contain more detailed information about customer behavior. Account tenure, the average cost of a product, the number of days since the last purchase and all features that measure the percentage change in behavior are considered as advanced features in our research [6].

So, as you can see, we have used different measurement periods for basic behavioral measurements. The reason is that customers rarely make purchases in the observed dataset. We have measured the average number of purchases per customer and found out that the average customer makes 1.3 purchases per 1 month. We need to measure customer features over a longer period of time than one month. This way we can cover more customer purchases and compare them by estimated features. Otherwise, if we estimate features for a shorter period than one month, then many people will have 0 in features. Because these people didn't make purchases in such a short period. We have used the rule of the minimum time period during which purchases should be observed in order to make behavioral features proposed in this work [6]. The author suggests using the period to be at least twice the time it takes for an average customer to make one purchase.

Account tenure is the length of time when the customer has made frequent purchases. First of all, we need to find the frequent period of customer purchases to calculate this metric. Then,

DOI: 10.37943/ILMM7870
© S. Akhmetbek

the time between the start and end of this period gives a value of account tenure. Like any customer behavioral feature, account tenure was measured relative to any point of customer lifecycle.

We have used a ratio feature in our analysis, which is called sum_purchase_per_n_product. This feature was estimated by the division of two other features. They are the sum_purchase_per_2month feature and the n_product_per_2month. Usually, these kinds of ratio metrics show how much a customer pays per product. Paying more per product is expected to cause churn.

Also, features that measure the change in customer behavior during an active period are used in our analysis. The importance of these features is that relative changes in customer behavior can influence the decision to churn or continue to make purchases. We can monitor these kinds of changes as a ratio. Percentage change calculated as a division of feature value at the end of measurement window by the feature value at the start.

Time since the last purchase is not a measurement of change in behavior, but it can help to identify customers who have become inactive. For example, if a lot of time has passed since the last purchase, it is most likely that the customer is close to churning.

V. Observing customers during the lifecycle

The purpose of churn analysis is to observe customers who want to churn and renew. If we periodically observe customers during their lifecycle, we will be able to compare behaviors that affect customer churn and retention. A sufficient number of renewals should be observed in order to understand the reasons why customers remain with a company and vice versa. The proportion of churns should be similar to the true churn rate [6]. Usually, a retail business analyzes a monthly churn and retention rates. According to the median of account tenure, a typical customer is active probably 16 weeks (Figure 4).

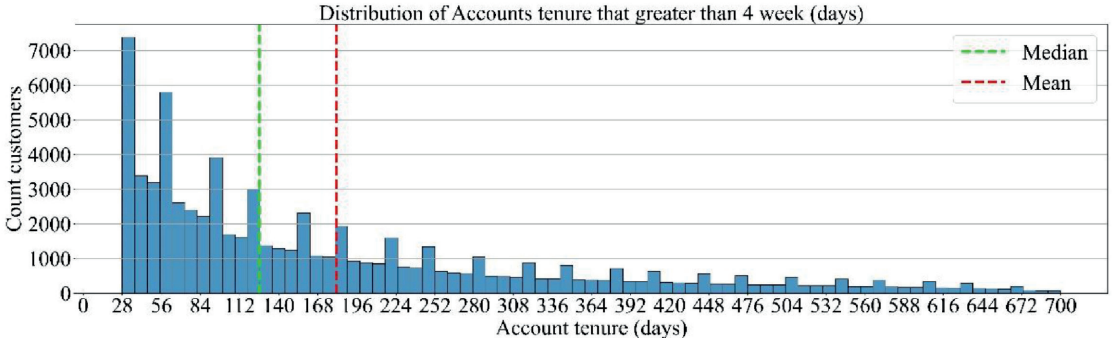

Figure 4. Histogram of account tenure

If we observe customers every 4 weeks, which is equal to 1 month, then the average proportion of time when customers renew will be 75%. We have estimated the monthly retention rates on the entire dataset. On average 80% of customers have continued to make purchases. The churn rate is the proportion of the customers in the start period who stop making purchases. Equation (1) shows the churn rate calculation:

$$\text{Churn rate} = \frac{\text{Number of churned customers}}{\text{Number of customers at the start}}, \quad (1)$$

where the nominator is a number of churned customers at the end of observed period, the denominator is a number of active customers at the beginning of the observed period. Retention rate can be easily calculated from the churn rate and vice versa (2):

$$100\% = \text{Churn Rate} + \text{Retention Rate}. \quad (2)$$

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 10, JUNE 2022

The churn and retention rates for February, 2010 were 18% and 82% (Table 2). This proportion remains relatively same in all months.

Table 2. Estimated churn and retention rates for February 2010

Churn rate	Retention rate	Number of clients at beginning of month	Number of clients at end of month
0.184178	0.815822	809	149

We have observed customer behavior every 4 weeks in order to balance the proportion of customers to the real churn and retention rates. For each observation date, features for the past two months of customer behavior have been measured. If the customer makes purchases between the first observation period and the next period, we consider this as customer retention. We will analyze the customer's behavior that was related to repeated future purchases. If the client does not make any purchases between the first observation period and the next period then we consider this as a customer churn. The features of that customer will indicate what affected the customer's decision to churn.

VI. Churn cohort analysis

The next step of our analysis is to check the relation of features to the churn or retention. We have analyzed this kind of relationship with the help of churn cohort analysis [6]. A customer cohort is a group of individuals that have a similar value in the features. A churn cohort analysis is a comparison of churn rates with different group of customer behaviors. This way we will be able to see what feature values affect the churn rate. The churn rate of cohort of customers is shown on the y-axis on a relative scale. The average value of used feature value is plotted on the x-axis. Each point shows the average feature value and average churn rate for one group of customers.

Overall, the cohort analysis shows that the higher values of features are associated with lower churn rate (Figure 5).

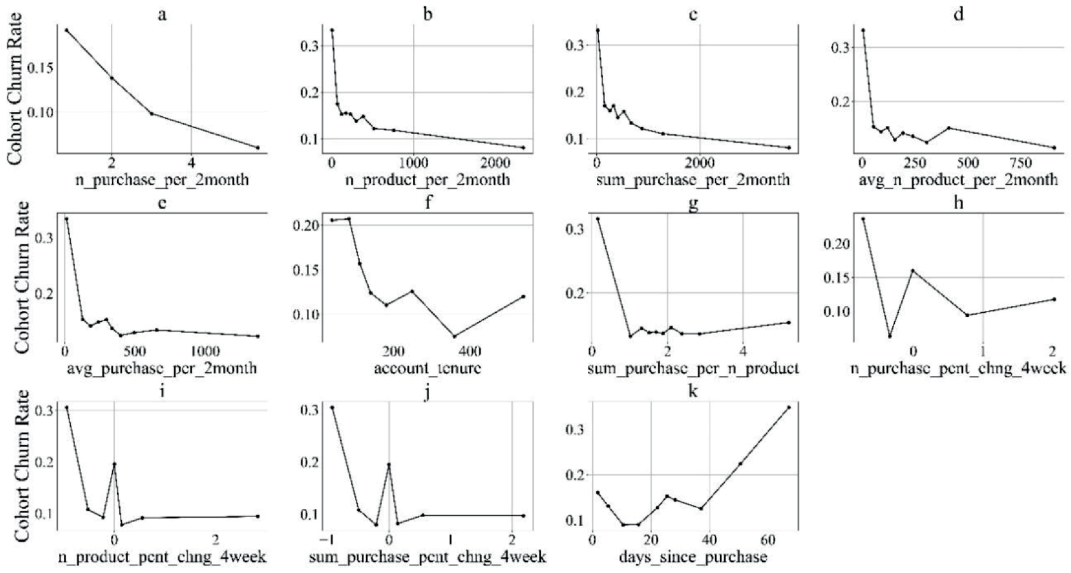

Figure 5. Churn cohort analysis of metrics:
a – n_purchase_per_2month; b – n_product_per_2month; c – sum_purchase_per_2month;
d – avg_n_product_per_2month; e – avg_purchase_per_2month; f – account_tenure;
g – sum_purchase_per_n_product; h – n_purchase_pcnt_chng_4week; i – n_product_pcnt_chng_4week;
j – sum_purchase_pcnt_chng_4week; k – days_since_purchase.

DOI: 10.37943/ILMM7870
© S. Akhmetbek

The results of churn cohort analysis for simple features shows that we can accept the following hypothesis: customers who make more purchases are less likely to be churned (Figure 5. a, b, c, d, e). The churn rate decreases dramatically as the features' value increases. This pattern makes it easy to understand which customer behavior is healthy or not. But a further increase in the feature value after some point shows no changes in churn rate.

The churn cohort analysis of the next feature is the most common (Figure 5. j). Customers who have been customers for a long time churn less than new customers who are just at the beginning of their lifecycle. Customers who have an account tenure higher than 100 days churn less, compared to the new customers. The churn rate for the customers with the longest tenure is less than the group of customers that has a peak churn rate.

The ratio metric shows that it is also related to churn (Figure 5. k). The more customer pays for product item, the less they churn. This kind of behavior may be surprising, but it is common in retail. This behavior is associated with the quality of products that a company sells. That is because expensive products are sold to customers who have a large amount of money, and this type of customers churn less for temporary reasons.

The cohort analysis of metrics that measure a percentage change shows that a large decline in the number of purchases is a significant churn risk (Figure 5. g, h, i). Customers with zero purchased products per month in both past two months have zero change but two times higher churn risk. This contributes to the high risk in the fourth cohort. Also, customers who made fewer purchases than in the previous two months have the highest churn rate. This means that when customers' interest in the product starts to disappear, they start making fewer purchases, which leads to an increase in churn rate.

The results of the cohort analysis of metric that measure days from the last purchase shows that a gap of more than around fifteen days since the last purchase is associated with an increasing risk of churn. The increase in risk is gradual but becomes fairly significant for the cohort with the longest time since purchase. This way, we will be able to analyze what feature values affect the variation of churn rate.

VII. Relationship between customer metrics

After analyzing whether features are related to churn, we need to measure the relationship between them. This kind of analysis is useful, because the goal of churn analysis is to predict customer churn probability. If we know which features are related to each other, we can predict one of them based on the other one. Customer churn prediction may be more accurate with the use of a group of related features. Correlation measures the relationship between numerical features. The correlation coefficient is a measurement of correlation that can be between –1.0 and 1.0. Positive value of correlation coefficient means that an increase in one feature is always associated with the same increase in another feature. It works backwards for negative correlation. A correlation matrix is a table of all of the pairwise correlation coefficients between the features in a dataset. The correlation matrix of our features shows that there are some highly correlated pairs of features, but in general, correlation between features is low (Figure 6).

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 10, JUNE 2022

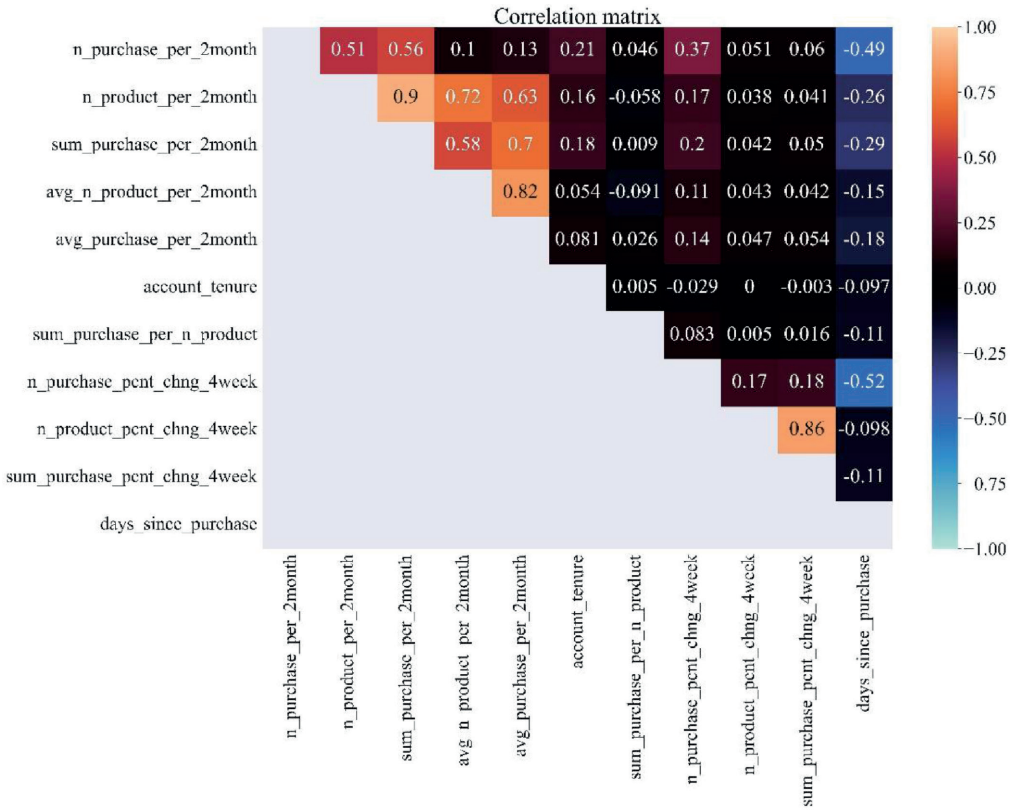

Figure 6. Correlation matrix of features

VIII. Prediction models

In order to solve the problem of forecasting churn probability in retail business, we explored machine learning and deep learning algorithms. The process of forecasting churn starts with the preparation of customer behavioral data to train models on them. Model training means the process of identifying rules based on data from examples of expected results. Customer behavior represents the data on which the models will train. The expected values are the customer's decision to churn or continue shopping. The models considered in this paper show good results only if they have been trained on data that corresponds to the requirements of the models. Those requirements are values in the dataset should be scaled, features' distribution should be not too much skewed and there are no highly correlated pairs features [12]. Scaling values means changing a variety of measurement units and ranges of feature values to a single range. This allows to compare values with each other. If a distribution of feature is skewed when the data includes extreme outliers. The third requirement is that the presence of highly correlated features makes it difficult to train models. All these requirements have been taken into account by the additional preprocessing steps:

1. If the feature was significantly skewed, we have taken the logarithm of feature values [6]. We have used the following equation (3) to use logarithm transformation for positive features:

$$m' = \ln(feature + 1), \tag{3}$$

where m′ is a metric distribution, ln is the natural logarithm function. For features with negative values was transformed with next (4):

$$m' = \ln(feature + \sqrt{feature^2 + 1}). \tag{4}$$

2. The features were scaled according to (5):

DOI: 10.37943/ILMM7870
© S. Akhmetbek

$$\text{scaling(feature)} = \frac{m' - \mu_{m'}}{\sigma_{m'}}, \quad (5)$$

where $\mu_{m'}$ is the mean of transformed distribution m', $\sigma_{m'}$ is the standard deviation of the observed feature distribution m′.

3. Group highly correlated features after scaling by averaging them together.

The final version of the churn dataset consists of six behavioral features and one target which should be predicted (Table 3). According to summary statistics we have observed 18767 instances. The proportion of churns and renewals was close to true rates. The highly correlated features were grouped together. All features were scaled.

Table 3. Summary statistics of churn dataset

Group or metric name	count	nonzero	mean	std	skew	min	1pct	25pct	50pct	75pct	99pct	max
metric_group_1	18767	1,00	0,00	1,09	-2,05	-3,40	-3,40	-0,15	0,23	0,55	1,66	2,35
metric_group_2	18767	1,00	0,00	1,05	1,03	-1,88	-1,88	-0,42	-0,07	0,10	3,33	8,64
n_purchase_per_2month	18767	1,00	0,00	1,00	0,54	-2,05	-2,05	-0,49	-0,49	0,42	2,89	5,68
account_tenure	18767	1,00	0,00	1,00	1,32	-0,91	-0,91	-0,73	-0,37	0,35	3,06	3,42
sum_purchase_pcnt_ chng_4week	18767	1,00	0,00	1,00	-0,05	-2,38	-2,38	-0,39	0,05	0,51	2,52	11,13
days_since_purchase	18767	1,00	0,00	1,00	0,92	-1,29	-1,29	-0,79	-0,09	0,47	2,88	3,28
is_churn	18767	0,16	0,16	0,36	1,88	0,00	0,00	0,00	0,00	0,00	1,00	1,00

Two models were trained to predict the probability of customer future purchases based on a prepared churn dataset. There are the logistic regression and artificial neural network for classification task. In general, both models calculate a membership of the class probability for one of the two classes based on data. The purpose of these models is to find the right combination of parameters for each feature, so that it is possible to divide data into two classes [13]. The main feature of these models is in their method of the decision boundary. Both models use the sigmoid function for binary classification task (Figure 7). As the feature values increase, the function tends to 1. Otherwise, as the feature values decrease to minus infinity, this function tends to 0. This property of the function simulates customer engagement [6]. For example, the most engaged customers are the most likely to be retained, and vice versa. According to the sigmoid function, an average customer has zero engagement.

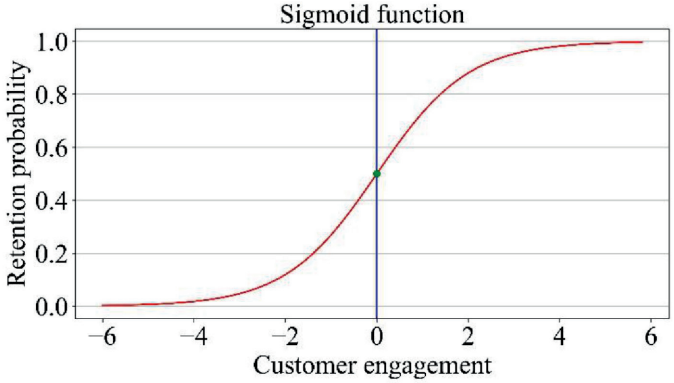

Figure 7. Sigmoid function

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 10, JUNE 2022

The best practice in forecasting customers' churn probability is to use a historical simulation of the data. This technique of model testing is known as out-of-sample testing. The model is trained on a sample of data from the past, and then the predictions of the already trained model are tested on a new sample from the future. This method of model testing was created due to the fact that markets are always changing, so predictive models work differently on randomly shuffled accuracy tests than on real forecasting. Accuracy tests based on realistic historical modeling are best in order to evaluate how the model might have worked if it had been live at the time. We have split the entire dataset into 10 sample sets. The models are trained on the first two months of data and tested on the next two months. In the next split, models are trained on the first four months data, then tested on next two months. The data from the next two months are added into each sequence of training samples. Then the models are tested on the next two months after the last date in the training samples.

We have used two accuracy measurements in order to evaluate the predictions of our models. They are the area under curve (AUC) and the top decile lift. The AUC is the percentage of comparisons in which the model forecasts higher positive probability for a positive class than for a negative class, taking into account pairwise comparisons of all positive and negative classes. If a customer churns, then it will be interpreted as a positive class, and vice versa. The second metric measures how much better the prediction model is at detecting churn compared to a random prediction. We used these metrics because they can give a correct evaluation of predictions when the proportion of classes is imbalanced. For example, the positive-class ratio is 16% in our churn dataset. This kind of churn forecasts proportion cannot be measured with the standard accuracy measurements, because those methods will be dominated by the major class.

Results

The output from a historical simulation shows the lift and the AUC for each out-of-sample test (Table 1). The average AUC of logistic regression was equal to 71.8%. The AUC is a percentage, like accuracy, and 100% is the best possible. The average lift was 2.573. If the list is equal to 1, then it means the model predicts like random guessing. A healthy lift is in the range from 2.0 to 5.0 for businesses with churn rates higher than 20%. Consequently, logistic regression predicts the future churns well.

Table 3. AUC and top decile lift metrics of Logistic regression forecasts for test samples

Metric Name	Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7	Split 8	Split 9	Split 10
AUC	0.725	0.761	0.698	0.704	0.753	0.765	0.746	0.765	0.709	0.559
Top decile lift	2.850	3.213	2.141	2.069	2.328	3.517	2.968	2.914	2.465	1.293

Comparison of neural network accuracy to regression shows that the second model is relatively better in churn prediction (Table 4). There was an increase in the AUC due to the nonlinear structure of neural networks. It makes them more flexible in searching rules compared to logistic regression. The average AUC of the neural network was equal to 72.6%.

Table 4. The AUC metric of neural network forecasts for test samples

Metric Name	Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7	Split 8	Split 9	Split 10
AUC	0.707	0.770	0.695	0.718	0.773	0.766	0.758	0.772	0.740	0.581

DOI: 10.37943/ILMM7870
© S. Akhmetbek

Discussion

To estimate CLV we can use churn forecasts predicted by observed models [6]. The forecasts are the probabilities of customer purchases in the future. Equation (6) shows how to compute the expected customer lifetime:

$$L = \frac{1}{\text{churn probability}}, \tag{6}$$

where L is expected customer lifetime, churn probability is a model forecast and 1 is a predicted time period. Predicted period is the same as the time period for measuring churn. It was one month in our analysis. If the churn probability is 20% per month, the expected customer lifetime is 5 months. We expect that observed customer will make purchases during future five months. To understand how much a customer will worth in the future, we can estimate the past amount spent on purchases (7):

$$\text{Customer value} = \sum_{lifetime} \text{payments}, \tag{7}$$

where $\sum_{lifetime}$ is a summation of all monetary value during the observed customer lifetime. The next step is to compute the expected total profit over the customer's lifetime:

$$CLV = \text{Customer value} \times \text{L} - \text{CAC}, \tag{8}$$

where CAC is a customer acquisition cost. It is the total amount spent on marketing to acquire customer. The CAS value is set by the business itself. Customer lifetime value calculations is not a forecast, because those are known quantities in the sense that you can calculate them from the data. For this reason, we cannot estimate exactly how accurate the calculations will be until we get a profit from the client in the future. However, the customer churn probability is used in estimation of future lifetime, we can conclude that the CLV is also accurate.

Conclusion

In this paper, several customer lifetime value prediction methods were analyzed in the case of UK retail business customers. Our comprehensive study discovered that customers who are more engaged are less likely to be churned, compared to the new customers. In our work, customer engagement was measured based on the properties of the sigmoid function, which is used in two predictive models. These models are logistic regression classifiers and neural networks for the binary classification task. In these models, customer engagement is not measured directly, but by analyzing changes in behavioral metrics related to churn and retention. We have used eleven metrics that measure customers behavior. We have found that these metrics have a strong relation to customer churn. The metrics were derived from data that every retail company has, which makes our method more accessible to use. The accuracy of churn forecasts shows that logistic regression and neural networks can predict future churns well. We consider that the logistic regression model is more suitable for this task than neural networks, despite the fact that neural networks are more accurate. The problem with neural networks is that they require a lot of data. Therefore, depending on the size of the dataset, neural networks must be rebuilt every time. In addition, behavioral metrics that we have calculated can be used not only to find the client's CLV, but also for customer segmentation. Based on metrics values, it is possible to divide clients into groups that are similar in behavior or churn risk. Companies can better conduct activities to attract or retain customers if they can identify these segments.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 10, JUNE 2022

References

1. Mammadzada, A., Alasgarov, E., & Mammadov, A. (2021). Application of BG / NBD and gamma-gamma models to predict customer lifetime value for financial institution. Paper presented at the *15th IEEE International Conference on Application of Information and Communication Technologies*, AICT 2021. doi:10.1109/AICT52784.2021.9620535
2. Hwang, Y.H. (2019). Hands-On Data Science for Marketing (1st ed.). Packt Publishing. https://www.packtpub.com/product/hands-on-data-science-for-marketing/9781789346343
3. Cloud Architecture Center. (2019, February 6). *Predicting Customer Lifetime Value with AI Platform*. https://cloud.google.com/architecture/clv-prediction-with-offline-training-intro
4. Chen, D., Sain, S.L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, *19*(3), 197-208. doi:10.1057/dbm.2012.17
5. Chen, D. (2019). *Online Retail II* (Version 1) [Data set]. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Online+Retail+II
6. Gold, C. (2020). *Fighting Churn with Data: The Science and Strategy of Customer Retention*. Manning Publications Company. https://www.manning.com/books/fighting-churn-with-data
7. Jahromi, A. T., Stakhovych, S., & Ewing, M. (2016). Customer churn models: a comparison of probability and data mining approaches. In *Looking forward, looking back: Drawing on the past to shape the future of marketing* (pp. 144-148). Springer, Cham. doi:10.1007/978-3-319-24184-5_35
8. Bemmaor, A.C., Glady, N., & Hoppe, D. (2012). Implementing the Pareto/NBD Model: A User-Friendly Approach. In *Quantitative Marketing and Marketing Management* (pp. 39-49). Gabler Verlag, Wiesbaden. doi:10.1007/978-3-8349-3722-3_1
9. Bardük, B. (2020, October). Modelling Time Statistics for Customer Churn Prediction. In *2020 28th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE. doi:10.1109/SIU49456.2020.9302329
10. Kim, T., Kim, D., & Ahn, Y. (2022). Instant customer base analysis in the financial services sector. *Expert Systems with Applications, 202*. doi:10.1016/j.eswa.2022.117326
11. Reyes, M. (Ed.). (2019). Consumer Life Cycle and Profiling: A Data Mining Perspective. In *Consumer Behavior and Marketing*. IntechOpen. https://doi.org/10.5772/intechopen.85407
12. Nkikabahizi, C., Cheruiyot, W., & Kibe, A. (2022). Chaining Zscore and feature scaling methods to improve neural networks for classification. *Applied Soft Computing*, *123*, 108908. https://doi.org/10.1016/j.asoc.2022.108908
13. Bharadwaj, S., Anil, B.S., Pahargarh, A., Pahargarh, A., Gowra, P.S., & Kumar, S. (2018, August). Customer Churn prediction in mobile networks using logistic regression and multilayer perceptron (MLP). In *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)* (pp. 436-438). IEEE. doi:10.1109/ICGCIoT.2018.8752982