Churn Prediction Estimation Based on Machine Learning Methods

Mykola Malyar
Uzhhorod national University
Uzhhorod, Ukraine
https://orcid.org/0000-0002-2544-1959

Mykola Robotyshyn M.V.
Uzhhorod national University
Uzhhorod, Ukraine
https://orcid.org/0000-0001-6567-6974

Maryana Sharkadi
Uzhhorod national University
Uzhhorod, Ukraine
https://orcid.org/0000-0002-1850-996X

***Abstract*—Customer churn prediction is classic task of machine learning, the relevance of which continues to grow. This is due to the fact that business companies collect more data about their customers and their behavior every year. A model that predicts whether a customer churn will occur in the future allows a business to build an optimal personalized pricing policy to retain a customer. Existing approaches for solving the problem of churn prediction for different areas are analyzed. A strategy for determining the period of customer churn is proposed and the optimal variant of data labeling is selected, which allows to convert the problem to a typical binary classification problem. A set of data from the Prozorro system was chosen for practical application of the approaches. Ensemble tree methods (Random Forest, XGBoost, LightGBM) were chosen as learning algorithms. Customer churn prediction is one of the many applications in today's world. Knowledge of mathematics and machine learning algorithms, along with the correct ability to build a problem statement – are the key skills of a specialist who studies it.**

Keywords—customer churn prediction, machine learning algorithms, ensemble tree methods, binary classification

I. INTRODUCTION

The churn prediction problem is one of the most common in practice Data Science (application of statistics and machine learning to business problems). A popular expression in the industry, which clearly emphasizes the urgency of the problem, is the following: "Finding a new customer is 5 times more expensive than keeping an existing one". The task is quite universal, there are many options for building a problem statement depending on the industry: mobile operators, banks, gas stations, online stores and others.

Given the constant popularity of this problem and the existence of many solutions on the market, we can conclude that there is currently no algorithm that could accurately predict whether the churn will happened or not. Such an algorithm is unlikely to be created, because the result is depends from large number of factors, many of which are unpredictable.

The problem is solved in several stages. The first step - data preprocessing and feature engineering. We process the input to create binary target variable – whether there was churn or not. Based on the input table data we generated additional features that affect the churn. This stage required a detailed analysis of domain data and an understanding of business problem. The next step is to build a model, apply machine learning algorithms to the processed data to find nonlinear patterns between features. The final step is to validate the model on new unseen data and build feature importance graphic to understand which features has strong influence on customer churn.

The churn prediction problem can be formulated as typical binary classification problem[1] and be solved using supervised algorithms. In the related work section we analyzed in which area of business the churn prediction problem can be applied and by what approaches and methods it is solved. In the data preprocessing and feature engineering section a strategy for determining the period of customer churn is proposed and the optimal variant of data labeling is selected to convert problem to binary classification task. The results section presents the results of solving customer churn prediction problem based on real-world data from the public procurement system Prozorro. The optimal churn period is 90 days. Machine learning algorithms based on boosting and bagging ideas are used. AUC ROC is chosen as performance metric. Key features influencing the customer churn are identified.

II. RELATED WORK

The customer churn prediction problem is used in many industries. Initially, the main area of application was mobile operators[2,3], which were among the first to collect and process a large amount of information about their customers.

The problem is popular for banking sector[4], special models are used, which are mainly based on decision tree algorithms [5]. There are a number of studies in the field of online games, which are now widely popular[6,7]. The problem is used also in many other areas[8].

Various machine learning algorithms are used to solve customer churn prediction problem: support vector machines (SVM)[9], boosting algorithms[10], neural networks[11]. One of the main advantages in using decision trees for this problem is a relatively easy interpretation of the model's logic.

Not only tabular data is used to analyze customer behavior. One of the newest approaches to solving the problem is the use of third-party data from telephone calls[12], from social networks[13]. To conduct research in this problem, there are a number of open data and online competitions[14].

The urgency of this problem will grow rapidly in the coming years due to the digitalization of many business companies and their ability to gather more information about behavior of their customers.

978-1-7281-9083-9/20/$31.00 ©2020 IEEE

III. DATA PREPROCESSING AND FEATURE ENGINEERING

A. *Strategy for determining the period of customer churn*

The task of prediction customer churn is a task of classification. Essence which is based on known customer characteristics (features) it is necessary to provide its belonging to the group of those users who will leave or will remain. The task of classification is the task of learning with the teacher. Training and test datasets are required. The main goal of predicting customer churn is to reduce cost of attraction new customers, reducing the cost of human resources for detection of disloyal customers, increase the customer database. The solution to this problem should be the answers to the following questions: "What customers are primarily prone to churn and why?", "Which factors are more common the reasons for the loss of customers?", "What are the trends in churn by structure and what its dynamic?". Thus, an understanding of the reasons why customer go gives companies the ability to predict their churn and develop ways to their retention and return.

The formulation of the customer churn prediction problem can be ambiguous: "Will there be a churn of customer in the future?" or "Will the customer churn happened in the next n days?". The second approach is more deterministic and it is more often used in practice. The solution of the problem begins with determining the number n – the period of customer churn.

The choice of choosing the period depends on the industry and is based on input data. For example, for mobile operators the churn period is always 1 month, because during this period the tariff package is valid. But it is not always possible to clearly define this period. An empirical method for finding the number n is proposed, which is based on the questions: "How often do customers use the services", "What percentage of customers use the services every 30, 60, 90 days on average?". The answer to last question can be visualized by constructing a cumulative distribution function.

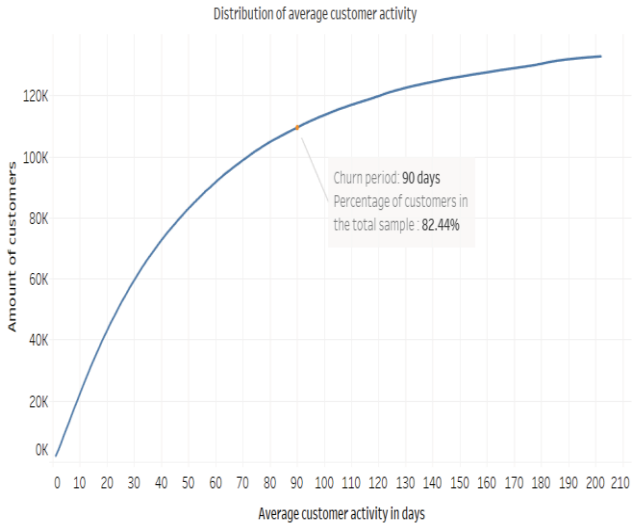

Fig. 1. Distribution of average customer activity

The x-axis shows the average activity of customers in days, the y-axis shows the percentage of customers from the total.

The goal is to find the optimal number of days so that it is not long and to cover as many customers. This figure used Prozorro customer data and shows that after 90 days there is no significant increase in the number of customers. Therefore, it makes no sense to take a larger number. Thus, we cover 82% of customers from total sample.

B. *Process of data labeling*

The next step is the process of labeling for every customer. We need to process the data so that each customer has a binary label: churn or not. Thus, we convert the problem to a binary classification problem for which classical machine learning algorithms can be applied.

For data labeling we divide data into intervals with n days d). In our case, n is equal to 90 days. There are 3 types of intervals:

1. Data history interval – interval to calculate customers features.
2. Data – interval to choose all unique customers, for them we will make prediction whether the churn will happened in next interval.
3. Data label – interval to see whether happened churn or not for every customer who were active in previous interval.

For better understanding of data labeling idea we visualize algorithm step by step as figure.

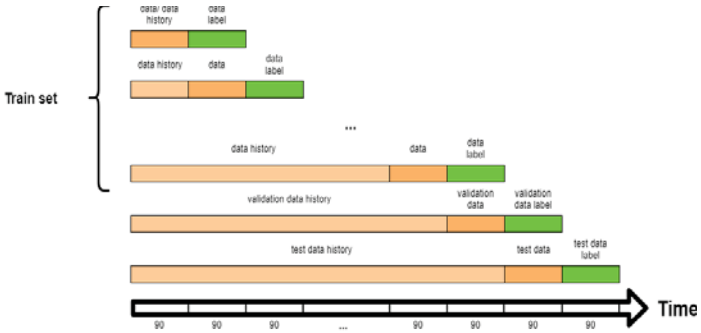

Fig. 2. Step by step visualization of data labeling algorithm.

The algorithm of data labeling is the following:

1. Divide data into 90 days (customer churn period) intervals.
2. For each interval (data + data history from fig.2):
    1. from data interval choose all unique customers;
    2. for each customer determine whether he was active in data label interval (green one in fig.2);
    3. if yes, then put $y = 0$ – churn didn't occur. Otherwise put $y = 1$ – churn occurred.

For each customer there is feature vector that impact its churn. The features are calculated on the basis of historical data. If we are in the first interval, then features are calculated on this interval. Otherwise, features are calculated based on all previous intervals. So, we take into account the previous history of the client.

This data labeling algorithm allows to transform input data into a suitable form for machine learning algorithms. Also, one more advantage that for each customer we create at least one raw. If customer appeared in different time interval, separate raw created. It allows to get more table rows as input to algorithms.

It is recommended to take the penultimate data interval as validation data for choosing hyperparameters of algorithm. After choosing retrain model with validation data and test model on last interval to get the final result.

IV. Results

A. Analysis of domain data

The practical application of the above approaches was implemented on the Prozorro dataset – public procurement system. General information about data[15]:

1. government tenders from 2016 to 2019 years;
2. more than 2 million tenders;
3. amount of data is approximately 12 GB;
4. 140 thousand customers and their open data;

The Prozorro system has customers (sometimes also called as suppliers) who participate in public procurement. It is very often the case that customer stop participating after a certain period of time – there is customer churn. If such churn is predicted before it happened, trading platforms that charge customers for the right to participate in the procurement may make a discount and thus encourage the return of the customer.

The optimal customer churn period was chosen as 90 days according to the empirical method (Fig.1). Data labeling process was performed according to the labeling algorithm (Fig.2). Validation and test data were taken on the basis of the penultimate and last data intervals, respectively.

For each customer 20 features were generated, which can be divided into blocks:

1. Statistical features – based on the history of participation in procurement.
2. Geographical features – related to the geography where the customer participated and where he delivered the goods;
3. Time features – when last time participated in procurement, what is the average activity…;
4. Additional features – favourite host, guarantee amount of money, CPV codes of goods.

It is important to note that the features can be of different types:

1. Categorical – name of the city, month type of business;
2. Binary – favourite host is exist or not, only one type of goods or not;
3. Integer – how many times the customer took part in tenders, how many days have passedsince his last activity;
4. Real – average lot value, amount of saved funds.

Also, one of the advantages of algorithms based on decision trees is that you do not need to normalize the input data and transform them under one scale. Other supervised algorithms such as logistic regression or support vector machines required this step.

Machine learning algorithms determine which features have the greatest impact on the target variable. It is very important to understand what these features are and what information is most relevant to the model. Based on this information additional features can be generated and added to the feature vector for each customer. This process is called "feature engineering" and it is critical when building a model for problems with tabular data[16]. The best way to visualize the importance of features is bar chart.

B. Model results

Algorithms with idea of bagging and boosting become nowadays the first choice when working with tabular data. We apply them to solve customer churn prediction problem based on Prozorro data. Four algorithms were used: decision tree, RandomForest, LightGBM, XGBoost. The first one was as baseline algorithm on initial set of features. Last two implement idea of boosting. And RandomForest is relevant to bagging. Area under the curve (AUC ROC) was chosen as the metric for comparing models and their performance.

Algorithm prediction is applied to one customer (one row of data). The result is the probabilities of the customer belonging to each class $(p_1, p_2) \in R_{[0,1]}$, where p_1 is the probability that the churn of customer happened and p_2 is the probability that the churn of customer did not happened. The algorithm is applied to the entire test sample and the accuracy metric AUC ROC is calculated based on predicted probabilities. The closer the value to 1 the more accurate the result[17]. Based on the value of the metric it is possible to make a comparison between the models to determine whether one model is better than another.

In order to transform the result of the model to the binary value $y^* \in Z_{[0,1]}$, where the value 0 means that the customer churn did not happened and value 1 means that customer churn happened perform the following operations. Set the value of the "threshold" $\theta \in R_{[0,1]}$ and determine y^* from the relation:

$$y^* = \begin{cases} 1, & p_1 \geq \theta \\ 0, & p_1 \leq \theta \end{cases}$$

By default in software implementations the value of "threshold" θ is equal to 0.5, but in practice this value is determined so as to maximize the value of metric calculated on test data.

The best result was shown by the XGBoost algorithm – 0.8322 (AUC score). The worst – the usual decision tree – 0.65 (AUC score). LightGBM result is 0.818 (AUC score) which is quite less than XGBoost. The results show the effectiveness of using an ensemble of trees rather than one tree alone. It should also be noted that the idea of boosting is more effective than bagging on this dataset.

Figure below illustrates top ten features that has the most impact on churn.

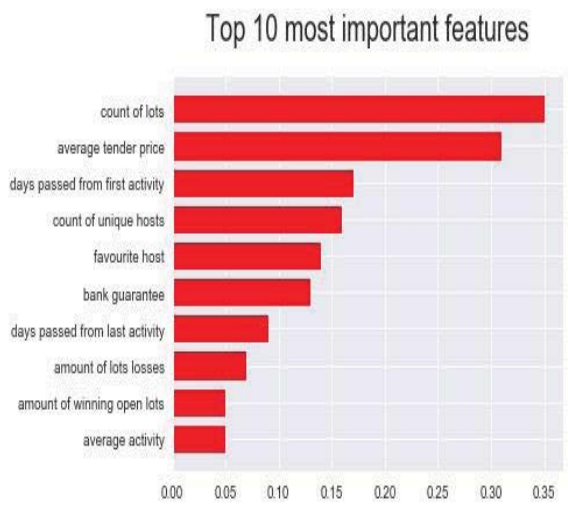

Fig. 3. Feature importance

The most important features that affect the churn are:

1. The count of lots in which customer participated.
2. The average price of tenders in which the customer participated.
3. The number of days that have passed since the first activity of the customer.
4. The number of unique customers to serve goods.
5. Binary variable that indicate whether customer is favourite for others hosts.

Therefore, the result of the algorithms shows that there are strong dependencies in data and there are features that strongly affect customer churn. Adding new datasets such as customer activity on websites, personal data, average market prices for purchased goods probably will improve the result.

To improve the model in the future it is necessary to determine group of customers where the model is wrong, distribution of errors and economic losses for errors and whether the investment in further model improving is cost-effective.

V. Conclusion

The customer churn prediction problem is an urgent applied problem of today, which has practical application in many industries. The paper proposes an approach to solving, which is based on the idea of converting the problem to a binary classification problem and the use of the ensemble algorithms of decision trees as methods of solution. The obtained results show a strong relationship between some features and target variable, but the classification error of models is significant on Prozorro dataset. There are ways how to improve classification error. First and most obvious way is to add additional datasets that were mentioned before. Second way is to collect more customers and do the whole paper pipeline again. Each year brings, at least 40 thousand new customers. So, in next five years the datasets can be doubled.

Prospects for further research are the use of clustering algorithms based on clear and fuzzy logic. The resulting clusters can be used as new features for customers.

References

[1] Sharkadi M.M., Robotyshyn M.V., Malyar M.M., “Machine learning models and methods for prediction problems” // Scientific bulletin UzNU “Mathematics and informatics”. – 2020. - №1(36). pp. 112-122.

[2] C. Archaux, A. Martin, and A.Khenchaf, “An SVM based churn detector in prepaid mobile telephony” in 2004 International Conference on Information and Communication Technologies: From Theory to Applications, 2004. Proceedings, 2004, pp. 459-460.

[3] N. Lu, H.Lin, J. Lu, and G.Zhang, “A customer churn prediction model in Telecom industry using boosting” IEEE Trans. Ind. Inform., vol. 10, no. 2, pp. 1659-1665, May 2014

[4] Jeremy Charlier, Vladimir Makarenkov, “XtracTree for regulator validation of bagging methods used in retail banking”, 5 Apr 2020, https://arxiv.org/abs/2004.02326.

[5] Xi Liu, Muhe Xie, Xidao Wen, Rui Chen, Yong Ge, Nick Duffield, Na Wang “A semi-supervised and inductive embedding model for churn prediction of large-scale mobile games”, https://arxiv.org/abs/1808.06573.

[6] A. Peri ´ a´nez, A. Saas, A. Guitart, and C. Magne, “Churn prediction ˜ in mobile social games: towards a complete assessment using survival ensembles,” in Proceedings of IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2016.

[7] Kumar, Arjun S. and D. Chandrakala. “A survey on customer churn prediction using machine learning techniques.” International Journal of Computer Applications 154 (2016): 13-16.

[8] Benlan He, Yong Shi, Qian Wan, Xi Zhao “Prediction of customer attrition of commercial banks based on SVM model”, Proceedings of 2nd International Conference on Information Technology and Quantitative Management (ITQM), Procedia Computer Science 31 (2014) 423 – 430

[9] Ning Lu, Hua Lin, Jie Lu, Guangquan Zhang “A customer churn prediction model in Telecom industry using boosting”, IEEE Transactions on Industrial Informatics, vol. 10, no. 2, may 2014.

[10] Meng Xi, Zhiling Luo, Naibo Wang, Jianwei Yin , “A latent feelings-aware RNN model for user churn prediction with behavioral data”, https://arxiv.org/abs/1911.02224v1.

[11] Zhong, Junmei & Li, William. (2019). “Predicting customer churn in the telecommunication industry by analyzing phone call transcripts with convolutional neural networks”.

[12] Ahmad, Abdelrahim Kasem, Assef Jafar, and Kadan Aljoumaa. “Customer churn prediction in telecom using machine learning in big data platform.” Journal of Big Data 6.1 (2019): n. pag. Crossref. Web. https://arxiv.org/abs/1904.00690v1.

[13] Searching for churn predi within (search engine). Retrieved from https://www.kaggle.com/search?q=churn+predi

[14] Huang F.J. large-scale learning with SVM and convolutional nets for generic object categorization / F.J.Huang, Y.LeCun // IEEE Computer Society Conference on Computer Vision and Pattern Recognition. – 2006. –P. 284-291.

[15] Data access (BI system). Retrieved from http://bi.prozorro.org/http/sense/app/fba3f2f2-cf55-40a0-a79f-b74f5ce947c2/sheet/HbXjQep/state/analysis#view/pEh

[16] Peter Klauke (2019, March 15) Importance of Feature Engineering (Blog post). Retrieved from https://towardsdatascience.com/importance-of-feature-engineering-methods-73e4c41ae5a3

[17] Sarang Narkhede (2018, June 26) Understanding AUC-ROC Curve: (Blog post). Retrieved from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5