Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/17/20230938

Predicting and visualization analysis of customer churn in telecommunications leveraging decision tree model

Ruiyang Luo

Zhengzhou Foreign Language School, Zhengzhou, Henan, 450000, China

rl5105@nyu.edu

Abstract. Nowadays, how to keep more customers is one of the most crucial problems to be solved facing communication companies in the increasingly competitive market. In these companies, one of their most important business indicators is customer churn for it is significantly less expensive to keep a current client than to find a new one. In order to carry out truly effective methods to minimize customer loss, build a prediction model is a reasonable way to analyze the characteristics and reasons of possible lost customers. In this research, after creating a Decision Tree Model and visualize it, it could be concluded that there was little difference between male and female, but the aged, unmarried and customers who are economic dependent have a great possibility to lose. Moreover, signing duration matters in the prediction. The longer the contract term is, the less likely the users to be lost. According to mentioned and even more analysis below, certain targeted suggestions are introduced to improve the situation. This research would like to provide research thought and analytical method for the prediction of customer churn.

Keywords: customer churn, decision tree, machine learning.

1. Introduction

With the increase of market saturation in the telecommunication industry, communication companies are facing fierce competition for customers because who wins the customers wins the world [1,2]. It is generally accepted that keeping an existing customer costs substantially less than getting a new one. Therefore, for these companies, reducing customer churn is the top among all priorities. A study shows that a 5% reduction in customer churn can increase the company's profit by 25% to 85%, which has a huge effect on company operations [3,4]. So, it is urgent for these companies to figure out solutions to increase user stickiness and prolong the user life cycle.

Building a prediction model is a vital way to analyze the characteristics and reasons for the potential loss of customers. In this way, companies are able to develop a targeted retention plan to improve user retention. Machine learning is part of artificial intelligence, which to a certain extent, can help people complete some data prediction, automatic decisions, and other tasks that initially replace brain power [5,6]. Compared with other models like Naive Bayes and Random Forest, Decision Tree's intelligibility and simplicity attracts researchers' attention. Its visualization is straightforward to understand and explain [7]. Also, it can make feasible and effective results on a large number of data sources in an extremely short time. As a consequence, in this research, the research chooses to use Decision Tree Model to make a prediction, whereafter, make an analysis and draw a conclusion.

DOI: 10.54254/2755-2721/17/20230938

2. Method

2.1. Dataset

The chosen dataset is about communication customers in China Telecom Company from iDataScience website [8]. In this dataset, there are 5986 samples in total which is large enough to support the results. Also, lots of parameters in this dataset offer me a thorough consideration of the situation to better figure out the character of lost customers.

Parameters include gender (Female & Male), Senior Citizen (the aged or not), Partner (married or not), Dependents (economic independent or not), Contract (ways to sign contract), Paperless Billing (open electronic account or not), Payment Method (approaches to pay), Phone Service (open phone service or not), Multiple Lines (have multiple lines or not), Internet Service (open internet service or not), Online Security (open this service or not), Online Backup (open this service or not), Device Protection (open this service or not), Tec support (open this service or not), Streaming TV (open this service or not), Streaming Movies (open this service or not), Tenure (time online), Monthly Charges(fees per month), Total Charges (fees in total), Churn (whether lose or not).

2.2. Data processing

In order to analyze and visualize the data more efficient, data is processed first. The first step, all data are imported and are screened to verify the correctness, getting to know the type of each parameter and whether there are some missing values. The second step, data polishing is conducted. Moreover, astype function is leveraged to transfer "TotalCharges" from object type into float type. Then, the mean value of a specific feature is leveraged to fill the missing value in "TotalCharges". The third step, for machine learning, changing data into integer to do data normalization is super necessary. At last, thanks to upper processing, data could be further analysed and visualized.

2.3. Decision tree

Given their simplicity and comprehensibility, this method is one of the most widely used in machine learning. A supervised learning method used in data extraction, statistics and machine learning is decision tree learning [9]. This philosophy bases its interpretation of a collection of observable observations on a predictive model, such as a classification or regression decision tree. For learning in this project where the goal variable can take a discrete set of values, the decision tree approach is chosen.

This approach produces a tree structure that resembles a flowchart where each leaf node represents the conclusion, the branch represents a decision rule, and the inside nodes indicate features [10]. A decision tree is produced by segmenting the dataset into subsets in accordance with a combination of fractionation rules based on categorization criteria. Recursive partitioning is a method for learning decision trees from data that repeats this procedure on each derived subset. The most common approach for doing this is by far top to down induction of decision trees.

Any decision tree algorithm's fundamental premise is as follows: Firstly, to divide the records, use Attribute Selection Measures (ASM) to choose the best attribute. Secondly, dividing the dataset into smaller sections and making that attribute a decision node. Thirdly, recursively repeat this technique for each kid to begin growing the tree until one of the prerequisites is satisfied: There are either no more attributes or instances, or all of the tuples have the same attribute value.

For ASM, the attribute selection metric is heuristic to select the fractionation criterion which best distributes the data. Since it enables users to identify tuple breakpoints on a certain node, it is also known as splitting rules. In this model, selection measures used include Information Gain, Gini Index, Chi-square. The formulas are demonstrated as follows:

$$Info = -\sum_{i=1}^{m} p_i log_2 p_i \tag{1}$$

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2 \tag{2}$$

DOI: 10.54254/2755-2721/17/20230938

$$x^2 = \sum \frac{(O-E)^2}{E} \tag{3}$$

In these formulas, x^2 denotes the Chi-square, O refers the observed score, and E represents the expected score.

Before create the model, the dataset is separated into two parts: training part & test part. 70% of data are classified as training part randomly, the rest 30% are classified as test part. For training part, these data will repeat the process to create the decision tree Model. For test part, these data are used for evaluating the model. As for the model evaluation, accuracy is leveraged as the index.

3. Result

3.1. Visualization analysis of basic customer information

Form the results illustrated in Figure 1, more than 25% of customers are lost. This situation is serious enough to be noticed.

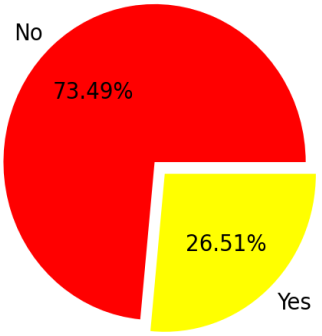

Figure 1. Proportions of customer churn.

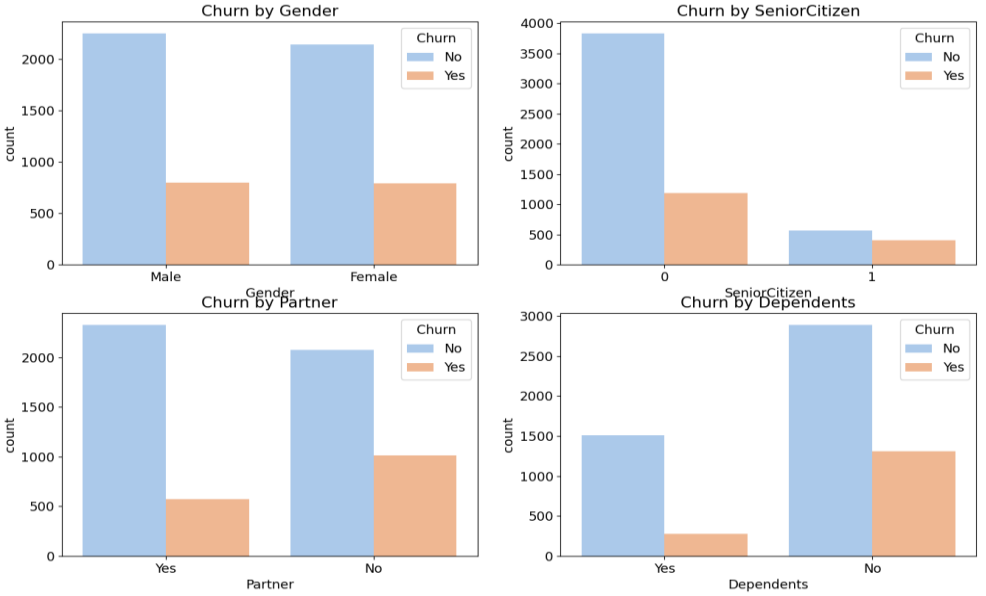

Figure 2. Distributions of customer lost across gender, senior citizen, partner, and dependents.

To further demonstrate the detailed distribution of various features, Figure 2 illustrates detailed distribution. For gender, there is little difference in churn between male and female users. For Senior Citizen, the loss of elderly users is higher than that of non-elderly users. For Partner, the number of

DOI: 10.54254/2755-2721/17/20230938

unmarried defectors is twice as high as the number of married. For Dependents, the loss of users without economic independence is much higher.

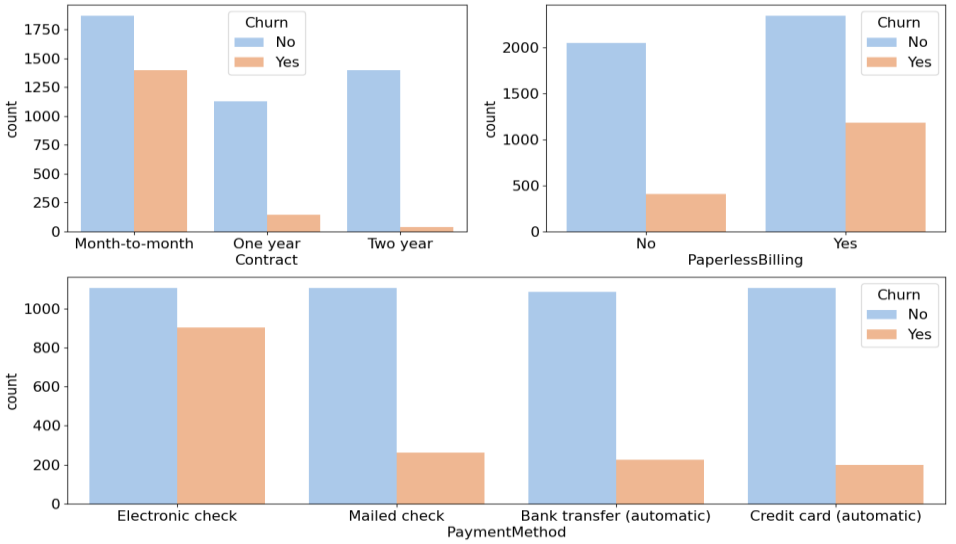

Figure 3. Distributions of customer lost across contract, paperless billing, and payment method.

Three features and their corresponding distributions are demonstrated in Figure 3. For Contract, the longer the contract duration, the less likely to lose users. For Paperless Billing, users are more likely to churn when paperless billing is adopted. For Payment Method, it is obvious that electronic payment is more likely to lose customers, while the other three payment methods' user loss is basically the same.

Distribution of other features are illustrated in Figure 4. For Phone Service, whether to open or not has no effect on the loss of users. For Multiple Lines, it has little impact on customer churn. For Internet Service, customers with Fibber optic service are more likely to leave. For Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, the churn rate of No internetserive customers is relatively low. Possible reason for these six is that these six factors affect customers only when they use Internet services.

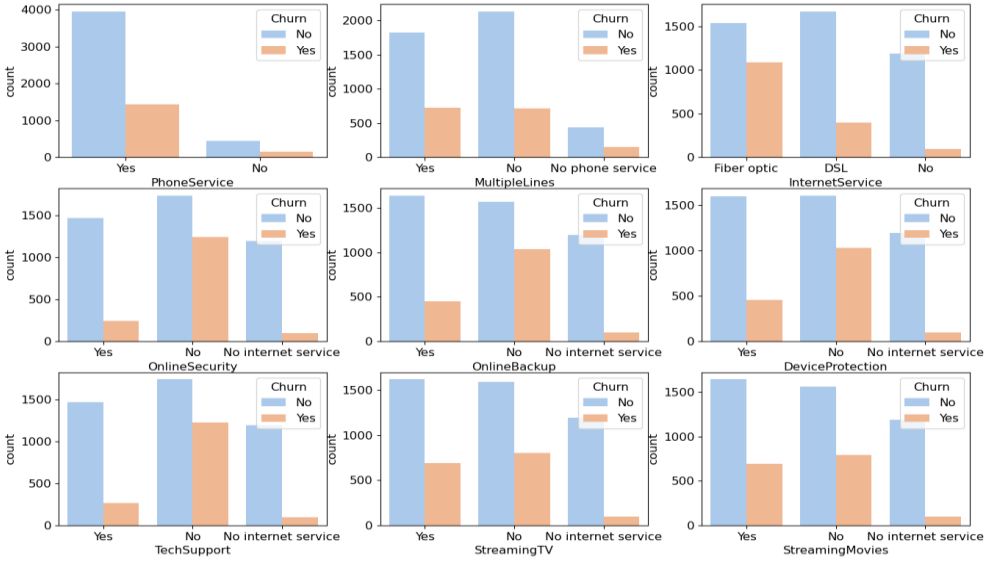

Figure 4. Distributions of customer lost across other features.

DOI: 10.54254/2755-2721/17/20230938

3.2. Visualization analysis of numerical characteristics of data

The numerical characters are illustrated in Figure 5,6 and 7 respectively. The line with shadow refers to remining customers and that without shadow illustrates the customer churn.

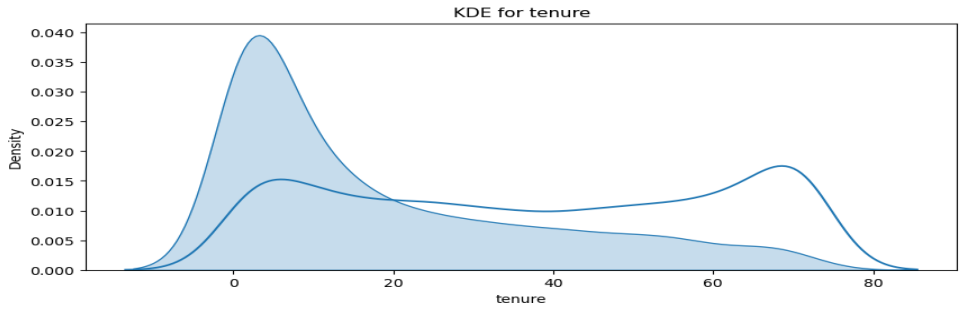

Figure 5. Customer distribution of tenure.

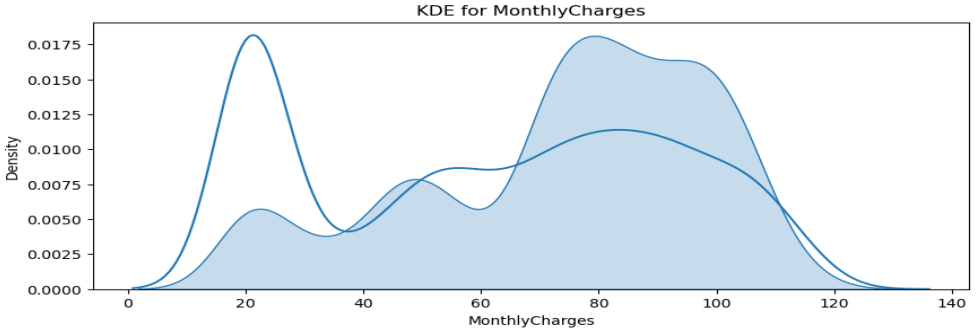

Figure 6. Customer distribution of monthly charges.

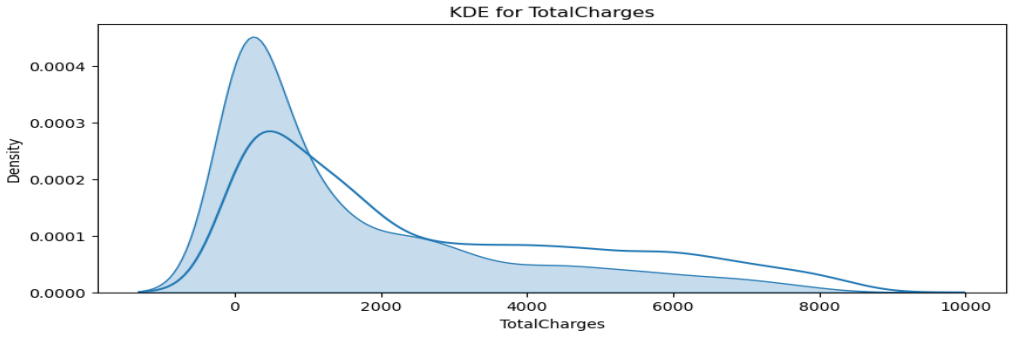

Figure 7. Customer distribution of total charges.

For Tenure, the user churn rate decreases with the increase of time spent on the network. For Monthly Charges, the higher the monthly fee, the more likely it is to lose customers. For Total Charges, customers with lower total charges are more likely to be lost.

3.3. Visualization analysis of numerical characteristics of data

According to the result of Chi-Square Test shown in Table 1, gender, Payment Method, Phone Service and Multiple Lines are not that relevant to the customer churn. Which means they will be removed from the model in order to improve the accuracy.

DOI: 10.54254/2755-2721/17/20230938

Table 1. P-value of Chi-Square test results.

	Gender	Senior Citizen	Partner	Dependent	Contract	Paperless Billing	Payment Method	Phone Service
P	0.478	5.640	9.151	4.151	4.526	3.124	0.198	0.497
	Multiple Lines	Internet Service	Online Security	Online Backup	Device Protection	Tech support	Streaming TV	Streaming Movies
P	0.981	1.056	9.849	1.027	2.436	8.334	2.611	8.324

3.4. Decision tree model

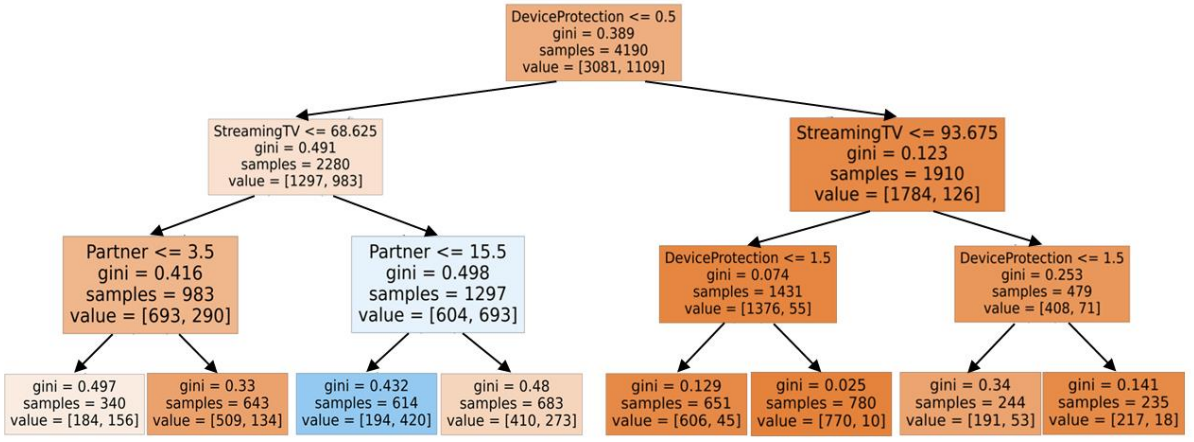

Figure 8. Visualization of the learned decision tree model.

The decision tree architecture is demonstrated in Figure 8. Using dichotomy to split, a lot of details and information could be achieved from the conditions that make the separation and percentage of customer churn in different subsets. Left of the value are numbers of customers that are remained, right of value are numbers of customers that are lost.

The ultimate accuracy is merely 78.11%. Maybe in the future, more samples could be leveraged and more machine learning techniques could be used for learning.

4. Discussion

Based on the above chart and information, the characteristics of users who are easy to lose can be basically outlined. From users' characteristics, according to user characteristics, older users, single users, and users who depend on money have a higher chance of losing. From service attributes, Users who have not opened related additional value-added services have more potential to turnover. From contract attributes, the shorter the contract term is, the more easily the users who use electronic payment will lose. The following traits continue to exist independently, while other factors have little influence on user loss.

The related recommendations are presented in light of the aforementioned results. Firstly, based on a predictive model, identify high turnover customers. User research is used to launch a minimal viable product, which seed consumers are then invited to test out and provide feedback on. Secondly, develop targeted strategies based on disparate classifiers. For elderly users, unmarried users and economically dependent users, diversified exclusive packages can be developed for them according to their preferences and characteristics, to improve their user experience. For new users, gift coupon will be pushed to tide over the peak of user loss. Meanwhile, relevant value-added service experience qualification can be presented at the time of purchase to enhance users' awareness and sense of experience of value-added services. For existing customers, according to their behavior to find out suitable products and focus on advertising them. Thirdly, when signing contracts with customers, the

DOI: 10.54254/2755-2721/17/20230938

corresponding price preferential policies can be implemented to encourage customers to sign long-term contracts.

5. Conclusion

In this research, a machine learning algorithm, specifically a decision tree model, has been thoroughly leveraged and constructed, aiming to figure out the true reasons for customer churn and come up with a series of efficient programs to settle the matter as well as gain a better reputation and profit for the company. After data visualization and model creation, some significant features of lost customers are captured. The aged, unmarried, and economically dependent customers have more potential to churn. Also, users without related additional value-added services are free to leave because they do not have a great number of issues to consider. Moreover, contract duration matters. The longer the term is, the lower the possibility of customer churn. According to these, a number of targeted strategies can be published. First, diversified exclusive packages in terms of users' preferences are suitable to be developed. Second, relevant value-added service experience qualifications can be presented for free in the first few months to deadlock the business. Third, the corresponding price preferential policies can be implemented to encourage customers to sign long-term contracts. Hopefully, in the future, other more enforceable suggestions can be introduced thanks to more comprehensive and more in-depth research.

References

[1] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 1-24.

[2] Al-Weshah, G. A., Al-Manasrah, E., & Al-Qatawneh, M. (2019). Customer relationship management systems and organizational performance: Quantitative evidence from the Jordanian telecommunication industry. Journal of Marketing Communications, 25(8), 799-819.

[3] Sudharsan, R., & Ganesh, E. N. (2022). A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy. Connection Science, 34(1), 1855-1876.

[4] Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. Computing, 1-24.

[5] Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. Electronic Markets, 31(3), 685-695.

[6] Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. American journal of epidemiology, 188(12), 2222-2239.

[7] Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research, 70, 245-317.

[8] iDataScience (2021). URL: http://www.idatascience.cn/

[9] Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 18(6), 275-285.

[10] Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.