Predict Customer Churn based on Machine Learning Algorithms

Zhiqing Liang *

School of Business, University of Sydney, Sydney, Australia

* Corresponding Author Email: zlia7671@uni.sydney.edu.au

Abstract. With the increasingly fierce competition in the market, it is inevitable for companies to find ways to make customer churn analysis and prediction in order to maximize benefits and save costs. This review will start with the application of machine learning in customer churn prediction, introduce the industries using churn analysis, state the principles and processes of prediction and summarize and compare some of the mainstream algorithms used in the prediction such as naive bayes, logistic regression, support vector machine, decision tree, random forest, extreme gradient boosting, adaptive boosting, k-nearest neighbor and artificial neural networks. Even though there are still many shortcomings in the various algorithms that need to be improved, machine learning still plays an important role in making customer churn analysis and prediction. As technology continues to develop and model research gradually deepens, the accuracy of customer churn prediction will continue to improve in the future, providing more comprehensive guidance for companies to develop customer retention programs.

Keywords: Customer Churn; Retention; Machine Learning; Algorithm.

1. Introduction

In recent years, the market has become very competitive, and companies are looking for ways to balance the retention of existing customers with the expansion of new ones. It is well known that retaining existing customers requires less costs [1]. Therefore, companies need to have appropriate systems and mechanisms in place to analyze and predict customer churn. Customer churn prediction is the prediction of the likelihood of a customer leaving a product or service, which means that the customer chooses to move to a competitor of the company in a highly competitive market. Its purpose is to predict the possibility of such behavior occurring before it happens and to take some action in time [2]. As more and more companies focus on customer retention, the development and application of smarter machine learning models for analysis and prediction is also becoming a major trend. In an ever-changing market, product quality, service, and market fluctuations can cause loss of customer. Identifying early signs of churn and potential lost customers can help companies reduce losses to a certain extent [3]. In recent years, the research community has also conducted a lot of research and proposed some mainstream machine learning algorithms in related fields.

2. Customer Churn Analysis in Business Application and Process

2.1. Business application and industry

Nowadays, customer churn prediction is widely used, mainly in industries where the cost of acquiring new customers is higher than that of retaining existing customers. These industries often require to establish good customer relationship management systems, and companies in these industries that do not make customer churn prediction often find that rivals that can use retention machines well will quickly outpace them. The common customer churn prediction literature focuses on industries including mobile and telecommunication, banking, insurance, retail market, gaming, logistics, e-commerce and media [4], where the market is highly competitive and the need to improve customer satisfaction, loyalty and retention is high.

2.2. Regular Process

Machine learning is a commonly used approach to customer churn prediction. Its modeling principle for predicting customer churn is usually through the use of classification algorithms that observe different patterns of churners and non-churners. Figure 1 shows the basic customer churn prediction process. By collecting historical data and relevant information, including basic information of customer such as total amount, discount amount, return number and information that can improve the accuracy of forecasting such as social surroundings [5]. Next, make some feature engineering, such as data exploration (manually observing and eliminating useless features), feature selection (using algorithms to select and determine the impact of features on customer churn), data mapping (converting text into numbers), and data processing and transformation (data normalization, handling duplicates, outliers, missing values). Then comes the machine learning phase, which tries to select and apply some relatively mature and reasonable algorithms to predict, confirm the model parameters, and the models are trained and saved in test set. Finally, the models are evaluated, and the accuracy of each model's prediction is obtained by comparing the prediction of customer churn with the real value [6]. Based on the results, optimal model is selected and used for prediction. In the process, the company can make both relatively accurate forecasts of lost customers and analyze the important factors of customer churn, allowing the company to better find appropriate solutions and responses.

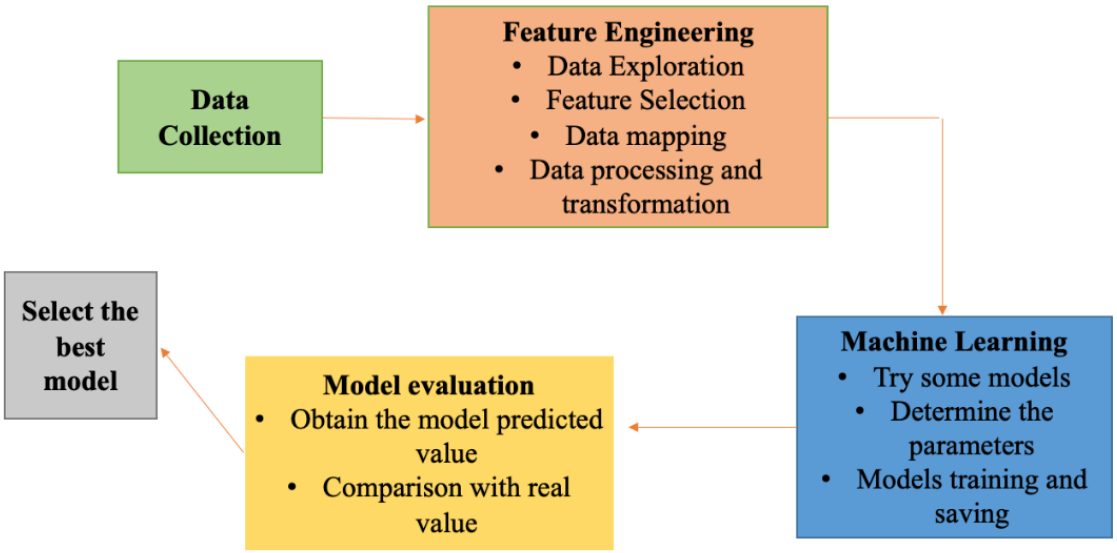


Fig 1. Basic Customer Churn Prediction Process, Photo credit: Original

3. Algorithms Used in Customer Churn Analysis

3.1. Classification Algorithm definition

The algorithms used in customer churn prediction are mostly classification algorithms. The classification algorithm ultimately classifies the customer into two or more categories and is used to predict whether the customer will leave the provider. From a mathematical point of view, the classification problem can be defined as follows:

Knowing the sets: S = $\{y_1, y_2, \ldots, y_n\}$ and I = $\{x_1, x_2, \ldots, x_m, \ldots\}$, the mapping rule is $y = f(x)$, make any $x_i \in$ 'I' and there is one and only one $y_i \in$ 'S', so that $y = f(x)$ holds. 'S' denotes the collection of categories, in which each individual element is itself a category. 'I' is the set of items, and each individual element is an object that needs to be categorised. 'f' is the classifier, and the objective of the classification algorithm is to produce 'f'. The quality of the classifier is related to many factors such as the classifier construction method, the characteristics of the data to be classified, and the number of training samples [7].

3.2. Naive Bayes

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)} \quad (1)$$

Naive Bayes is a basic but practical classification algorithm. The aim behind it is to anticipate a categorization based on the probability of features. As shown in (1), P(A) is the prior probability of A, the conditional probability of A given the occurrence of B is denoted by P(A|B). The conditional probability of B given the occurrence of A is denoted by P(B|A), and P(B) is the prior probability of B, it is also used as a scalarization constant. The main use of Bayes' formula is to infer the fourth probability from the known three probabilities. Bayes' steps are: 1. Determine the feature attributes as needed and segment each characteristic in its proper way, Next, manually categorize some of the items to be categorized to create a training sample set; 2. The conditional probability estimate for each category and the frequency of occurrence of that category in the training sample are computed and stored for each feature attribute class; 3. Make use of the classifier to sort out the items that need to be categorized. It has the advantages of simple algorithm, few estimation parameters required, and less sensitive to missing data. However, it has the disadvantage that it cannot learn the interaction between features, does not classify well whenever either the quantity of attributes or the degree of correlation between them is high, and is sensitive to the expression form of the data [8].

3.3. Logistic Regression

$$y = \frac{1}{1+e^{-\omega^T x}} \quad (2)$$

As a machine learning technique, logistic regression is useful for addressing questions of binary classification, which helps estimate how likely something is to occur. As shown in (2), ω^T is the parameter that needs to be calculated using the training set, and the formula is used to estimate and predict the value of the dependent variable (y) when independent variable (x) is known. Logistic regression estimates the parameter ω^T in the training set using maximum likelihood estimation. Routine steps of logistic regression problem solving: 1. Select an appropriate prediction function, typically the h function (hypothesis function) is used; 2. Create a Cost function that quantifies the discrepancy between prediction function (h) and the class of the training data(y). Obtain the total or mean value of the Cost, denoted as $J(\theta)$; 3. Find a way to minimize $J(\theta)$ and find the optimal parameters. Logistic regression can predict the probability of an outcome bounded between 0 and 1. It has the pros that it can be applied to continuous independent variables and is easy to use and interpret. However, multicollinearity among the model's independent variables can severely affect the results of logistic regression, and the "S" shape of the prediction results leads to some intervals where the threshold cannot be determined [9].

3.4. Support Vector Machine

Support Vector Machine (SVM) is a two-class classification model. It works by maximizing the distance from the nearest point (support vector) of the hyperplane to that plane. Since the steps of SVM need to go through a more complex mathematical derivation, there is only a brief introduction to its principle. SVM uses interval maximization as its learning approach, which can be formulated as a convex quadratic programming problem. An optimization method that can be used to solve convex quadratic programming is the process of the SVM learning algorithm. It has the advantage that it can map to higher dimensional space, solve nonlinear classification, and the classification idea is simple and better. However, at the same time, it is challenging to put into practice for large-scale training sets. Multi-classification issue and high sensibility to the absence of data and parameters are also difficulties that need to overcome [10].

3.5. Decision tree

Decision tree algorithm is a supervised learning algorithm that classifies samples into different classes by the value of a feature in the sample. It is based on a tree structure to make decisions, and

the leaf nodes are the results of the classification. The main steps of the decision tree are: 1. Make feature selection, screening out the features that have higher correlation with the classifying outcomes and stronger classification ability; 2. Generate decision trees, it all starts at the root node and calculates the information gain of each feature for each node. Node feature is determined by the one that provides the most information gain, and offspring nodes are made to correspond with the feature's various values, when there is no longer any information gain or features to select, the algorithm stops producing new child nodes in the same way it did before; 3. Pruning decision trees, eliminating certain branches in order to decrease the possibility of overfitting. Decision trees have the advantages of being easy to understand and visualizable. It can handle samples with missing attributes, uncorrelated features and both nominal and numerical data and run fast. However, it also faces the problem of being prone to overfitting and overlooking the interconnectedness of attributes in the dataset [9].

3.6. Random Forest

An ensemble algorithm known as random forest is made up of decision trees, it is also a supervised machine algorithm. The many individual decision trees that make up a random forest are unrelated to one another. The main steps in the operation of random forest are: 1. For an N-sample-size-minimum sample, N samples are taken for N times, and getting an N sample is created by taking a single sample each time; 2. When there are M attributes per sample and each node of the decision tree needs to be partitioned, then m attributes will be chosen at random (meeting the criterion of $m \leq M$). Then, one of these m attributes is chosen as the node's splitting attribute using some method; 3. During the decision tree-building process, each node is divided in accordance with the precious step; 4. Follow the above procedures to construct a huge number of decision trees. Random forest can be applied to highly dimensional data, judge the significance of features and the interaction between different features, balance the error for unbalanced data sets, and can maintain accuracy for missing features. Compared to decision trees, it is less prone to overfitting. However, it has been shown to overfit on certain noisy classification or regression problems [11].

3.7. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a special case of Boosting and is also based on decision trees, where the error is minimized by the gradient descent algorithm. The steps of the XGBoost algorithm are similar to CART, first initialized to a constant, and based on the first and second order derivatives, the base learner is generated iteratively and the summation updates the learner. XGBoost can make the learned model simpler, more flexible, can automatically handle missing values, easy to get the optimal number of iterations, and not easy to fall into the local optimal solution. However, it is only appropriate for processing structured data due to its excessive number of parameters, not effective for using with extremely high-dimensional feature data [12].

3.8. Adaptive Boosting

Adaptive Boosting (AdaBoost) is a common algorithm in integration learning. It can form a powerful model by combining models with poor performance. The reference steps of AdaBoost are: 1. Constructing a weak learner and calculate the error of this weak learner; 2. Calculating the weights of the weak learner and adjust the sample weights; 3. Doing second run: use the sample weights to generate a new dataset; 4. Repeating the process until the termination condition is reached and finally combine the weak learners into an integrated model. AdaBoost has good accuracy and takes full account of the weight of each classifier compared to random forest, but its number of weak classifiers is not well set, and it is easy to cause accuracy degradation when the data is not balanced [13].

3.9. K-nearest neighbor

K-nearest neighbor (KNN) is a supervised algorithm that uses proximity to classify or predict groups of individual data points. The KNN method consists of the following steps: 1. Determining the distance between points that need to be categorized and points that already belong to categories

that are known; 2. Sorting in ascending order of distance; 3. Choosing K points that are closest to the points that need to be categorized; 4. Finding out how many times each category appears in the first K points and tallying those results; 5. The prediction classification of the points to be categorised is the category in which the first K points appear most frequently. KNN can be applied to nonlinear classification, with lower time complexity than support vector machine, and higher accuracy than Bayes. However, it is not very explanatory compared to decision trees, less suitable for classification with small sample size, and has low prediction accuracy when the sample is not balanced [14].

3.10. Artificial Neural Networks

Artificial Neural Networks (ANN) is an information processing system draws on the concept of biological neural network. The general process of ANN is: 1. Place the training set into the neural network and obtain the output; 2. Compare the output with the desired output and calculate the error using the cost function; 3. Decide how to change the weights in order to minimize the error value based on the error value and the cost function used; 4. Repeat the process until the error is minimized. ANN has self-learning capability and the ability to quickly find optimal solutions, but sometimes it cannot explain the reasoning process and the basis [15]. The algorithms described above are all frequently used methods in the existing literature, and the algorithms used in many studies make some innovations and adjustments based on these algorithms. More ensemble algorithms have also emerged nowadays, and the accuracy of the ensemble algorithms is often higher than that of the individual classifiers.

4. Conclusion

Overall, using machine learning to do customer churn analysis and prediction is an effective method to help companies make a good maintenance in customer relationship management, and different companies and industries can also adjust and improve the models according to their different situations. However, it is also suggested in some literature that the analytical predictions of ordinary machine learning and business goals are sometimes not very well aligned. These common models do not really take into account the financial costs and benefits. Meanwhile, there are many factors that influence customer churn, and more factors that can improve prediction accuracy need to be tried and considered. In addition, different models and algorithms of machine learning have their own limitations. As the research progresses, these limitations need to be improved and enhanced better. The future development of customer churn prediction can strengthen the work and framework structure of customer churn management, more closely integrated with the company's business interests; more accurate recording and analysis of data on feature selection; and continuous improvement of the model through further research. At the same time, a clearer division of customer loyalty and satisfaction are needed, making the accuracy of customer churn forecasting to be improved and developed.

References

[1] De S, Prabu P, Paulose J. Effective ml techniques to predict customer churn. 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 895-902, 2021.

[2] Çelik O, Osmanoglu U O. Comparing to techniques used in customer churn analysis. Journal of Multidisciplinary Developments, 2019, 4(1): 30-38.

[3] Hadden J, Tiwari A, Roy R, et al. Computer assisted customer churn management: State-of-the-art and future trends[J]. Computers & Operations Research, 2007, 34(10): 2902-2917.

[4] Vafeiadis T, Diamantaras K I, Sarigiannidis G, et al. A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 2015, 55: 1–9.

[5] Richter Y, Yom-Tov E, Slonim N. Predicting customer churn in mobile networks through analysis of social groups. International conference on data mining. Society for Industrial and Applied Mathematics, 2010: 732-741

[6] Lalwani P, Mishra M K, Chadha J S, et al. Customer churn prediction system: a machine learning approach[J]. Computing, 2022, 104(2): 271-294.

[7] Cui C X, Research on classification algorithms for imbalanced data. Taiyuan: Shanxi University, 2022.

[8] Rish I. An empirical study of the naive Bayes classifier. Workshop on empirical methods in artificial intelligence. 2001, 3(22): 41-46.

[9] Dalvi P K, Khandge S K, Deomore A, et al. Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. 2016 symposium on colossal data analysis and networking (CDAN). IEEE, 2016: 1-4.

[10] He K, Guan Y Q, Gong R. Deep learning and support vector machine-based text classification model. Computer Technology and Development, 2022, 32(07):22-27

[11] Qiao J, Zhu J H, Yan K H. A telecom customer churn prediction model based on random forest CART feature selection improvement algorithm. Telecommunications Engineering Technology and Standardization, 2022, 35(03): 78-82.

[12] Liao K J, Zou K X, Zhuang Y Y. Improved XGBoost-based e-commerce customer churn prediction. Computer and Digital Engineering, 2022, 50(05): 1115-1118+1125

[13] Feng D C, Liu Z T, Wang X D, et al. Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach. Construction and Building Materials, 2020, 230: 117000.

[14] Peterson L E. K-nearest neighbor. Scholarpedia, 2009, 4(2): 1883.

[15] Yu M X, Zheng Y Y. Telecom customer churn prediction based on neural network algorithm. Modern Information Technology, 2023, 7(02):30-33