Application of Machine Learning to Customer Churn Risk Prediction

Chuyan Xiao *

School of Business and Management, Jilin University, Changchun, China

* Corresponding author: xiaocy0921@mails.jlu.edu.cn

Abstract. Accompanied by technological globalization and upswing of telecommunication industry in the 21st century, the number of operators is springing up like mushrooms after rain in the market and that intensifies the industry competition environment due to the unprecedented growth trend and the challenges. As a research hotspot in the field of business analysis, prediction of customer attrition risk possesses some extensive range pertaining to applications within global marketing, telecommunications and other fields. Due to the complex relationship between customer information and the products used, it is difficult for merchants to conduct effective risk assessments. Nevertheless, customer churn risk prediction has made such rapid progress over the past few years with the development within computer science and machine learning so that the researchers would be able to establish more potential connections and achieve more accurate predictions about customer churn. This article summarizes customer churn risk prediction in the order of time and technology iteration, and mainly introduces 9 classic prediction methods based on machine learning. Additionally, four relevant performance measurement metrics and one data set are explored. Eventually, in view of the problems existing in the current customer churn risk prediction methods, prospects for future research are proposed.

Keywords: Machine Learning, Customer Churn Risk, SVM, Random Forest.

1. Introduction

The Churn Prediction (CP) a method of forecasting corresponding circumstances clients have the potential to stop using a company's services, and it manages to solve commonly addressed predictive task within the telecommunication sector [1]. Accompanied by technological internationalization, together with upswing of the telecommunications sector, the operators' quantity is springing up like mushrooms after rain in the market and that intensifies the industry competition environment [2]. In this era of increasing competition, the mechanism to periodically and continuously retain the customers and additionally maximize the hard-won profit has been significant. Merchants generally adopt three strategies, namely, obtaining brand-new customers, enhancing the retention intention of current clients and up-selling the existing customers. Nevertheless, within the range of procedures, enhancing the maintaining of current clients ought to be the cheapest. Since the main reason for customer churn refers to customer dissatisfaction with the services offered by merchants to clients [3], some focus of solving this problem is to predict customers at risk of churn in advance [4-6] and optimize services for them. However, the different consumption habits of customers pose a huge challenge to forecasting or predicting churn risks.

Fortunately, together with the development related to the artificial intelligence, machine learning models or techniques have made great breakthroughs in tasks such as image, speech, and text. Some researchers have introduced machine learning into customer churn prediction and obtained better results. Praveen et al [7] investigated the effect of enhancement algorithms on accuracy in classification, which can be enhanced with the application of feature selection techniques. Y. Huang et al [8] experimented with different classifiers on churn prediction dataset and proposed optimization techniques for feature extraction to improve the accuracy. P. Kisioglu et al [9] proposed the use of Bayesian Belief Network (BBN) to predict customer churn.

However, there are fewer comparisons of different algorithms in these studies, which brings some difficulties for practical use. Based on this, this paper introduces customer churn prediction from two perspectives: data and algorithm. Firstly, it introduces the commonly used datasets and model

evaluation indexes in forecasting customer churn. Then, the algorithms were associated with the field of machine learning are introduced, including Logistic Regression, Random Forest Classifier, Extra Tree Classifier, Boosting Algorithm such as Ada Boost, XGBoost, CatBoost, Naive Bayes, Support Vector Machine and Decision Trees. Eventually, some effects pertaining to these algorithms are compared experimentally and future research directions are envisioned.

2. Datasets

Selecting a good dataset is a solid foundation in the context of customer churn risk estimation research. In the relevant field of customer segmentation, Sandra Mitrović et al [1] came up with a dataset called "SyriaTel data". All data pertaining to customer's services and contract information is contained within it. Moreover, it comprises information as well derived from the system named CRM, such as GSMs, subscription kind, birthday, gender, place pertaining to residence, etc.

The performance evaluation method of a model is the key to compare the effect of different models. Commonly used analysis indicators include TP, TN, FP, FN, and the meanings of these indicators are as follows.

• True Positive (TP): The count of customers classified as churners, and accurately forecast by the predictive model.

• True Negative (TN): The count of customers classified as non-churners, and accurately predicted by the predictive model.

• False Positive (FP): The count of customers classified as non-churners, but mistakenly labeled or identified as churners by the predictive algorithm.

• False Negative (FN): The count of customers classified as churners, but mistakenly labeled or identified as non-churners by the predictive model.

Based on the above analytics metrics recall, precision, accuracy, and F1-score can be calculated. Specifically, recall signifies the ratio of accurately estimated positive samples to the complete sum of positive samples. Recall rate is a measure of the model's proficiency to correctly identify actual churn customers, i.e., the proportion of actual churn customers correctly predicted by the model. Precision is a measure of the proportion of actual churned customers that are predicted to be churned. A larger precision value signifies that the model is more accurate in predicting churned customers and reduces false positives. Accuracy is meant to measure the predictive accuracy of the model as the proportion of correctly categorized instances to total instances. Accuracy provides an overall metric for the assessment of the model's predictive correctness. F1-score is utilized for the evaluation of the model's proficiency in accurately identifying churned customers. It provides an overall assessment of the model's predictive performance by balancing meticulousness and recall. A larger F1-score signifies that this pattern is large in not only meticulousness but also recall. The specific calculation formula for these four indicators is as follows.

$$Recall = \frac{T_p}{T_p + F_N} \tag{1}$$

$$Precision = \frac{T_p}{T_p + F_p} \tag{2}$$

$$Accuracy = \frac{T_p + T_n}{T_p + F_P + T_n + F_n} \tag{3}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

3. Method

The structure of consumer churn prediction is shown in Fig 1. The data has undergone pre-processing and then trained, and that model together with training method used are important steps in customer churn prediction. Next this section first introduces nine machine learning algorithms from the algorithmic principles and then compares their strengths and weaknesses by analyzing the effectiveness of these algorithms.

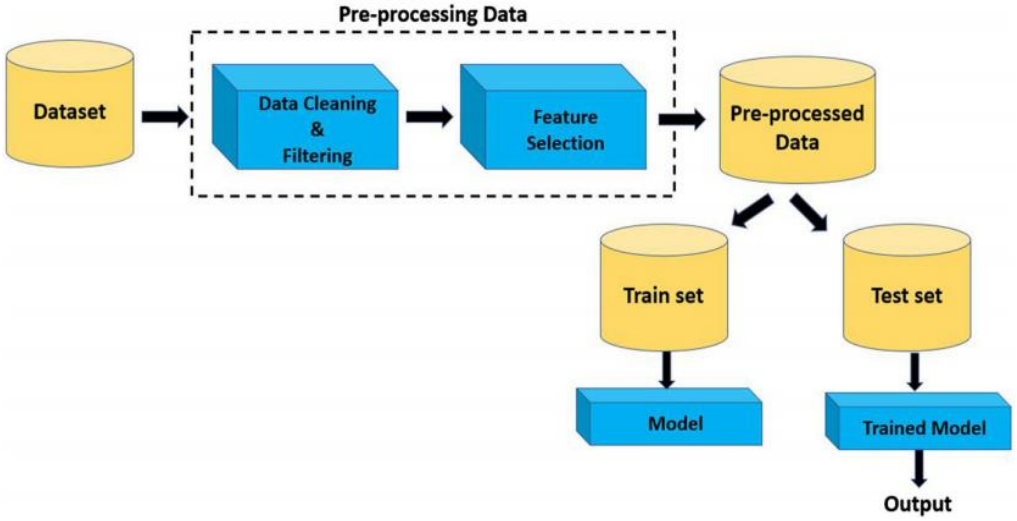

Figure 1. Customer churn forecasting architecture [6]

3.1. Logistic Regression

Logistic Regression (LR) is a statistical model used for binary classification which predicts the likelihood of an event happening based on a given independent variable and is commonly used in classification tasks. Logistic Regression uses a cross-entropy loss function to find the optimal classification hyperplane through great likelihood estimation. However, logistic regression, although simple, is less effective when faced with complex classification tasks. When performing model training, discretizing the features is beneficial to improve the stability of the model.

3.2. Naive Bayes

The Naive Bayes (NB) classifier is regarded as a methodology based on probabilities that assumes that every vector feature is unrelated to any other feature. The formula for calculating the conditional probability of plain Bayes is shown in (5).

In addition, plain Bayes treats individual features as equally important and task features are independent of each other, making it difficult to learn the underlying relationships before the data. Meanwhile, Park Bayes is insensitive to outliers and missing values, and has a high classification accuracy.

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{5}$$

3.3. Support Vector Machine

Support Vector Machines (SVMs), alternatively called or referred to as Support Vector Networks, were introduced by Boser, Guyon, and Vapnik [10]. In machine learning, they are methodologies for supervised learning that employ associated learning algorithms to perform classification and regression analysis. The major target of SVM is to separate the data into two parts and split the two types of data by learning a maximally spaced hypersurface. For some linearly indivisible tasks, which can be solved by kernel tricks due to the possible existence of hypersurfaces in the feature space. The

commonly used kernel functions are linear kernel, polynomial kernel, RBF kernel, Sigmoid kernel. SVM is robust, insensitive to noise, outliers in the data, and computationally efficient, its disadvantage is that it is very time consuming to train for higher latitude data.

3.4. Decision Trees

The decision tree (DT) algorithm, operates on a strategy based on greed and employs a collection of guidelines for categorization. A decision tree is structurally characterized by a tree-like structure where every internal node signifies a test on an attribute, each branch represents the outcome of a test, and each leaf node signifies a category. Decision Trees are easy to understand and implement and have high interpretability. Decision Trees are difficult to implement for continuous probabilistic prediction because the branching structure discretizes the results.

3.5. Random Forest

The Random Forest (RF) algorithm is an assemblage learning technique that consolidates multiple decision trees to produce accurate estimations. The theoretical foundation of this algorithm lies in the concept of bagging and random feature selection. The decision trees in the forest are built using bootstrap aggregating, where each tree is trained on a randomly divided segment of the training data. Additionally, a random subset of features is considered at each split, further enhancing the diversity of the forest. This approach reduces overfitting and improves generalization. The characteristics of Random Forest include robustness to noise and missing data, capability to handle both nominal and numerical variables, and this provision pertaining to feature importance measures for feature selection purposes.

3.6. Extra Tree Classifier

The Extra Tree Classifier is an assemblage learning method that extends the concept of decision trees. In theory, Extra Trees utilize random subsets of the training data and random feature selection, resembling Random Forest. Nevertheless, it varies in that that case employs random splitting points instead of finding the best split. This randomization reduces computational complexity and increases diversity among the trees. In terms of characteristics, the Extra Tree Classifier is known for its efficiency, making it suitable for large datasets. It is robust to noise, can handle both categorical and continuous variables, and provides feature importance measures for feature selection purposes.

3.7. Adaboost Classifier

The Adaboost Classifier is a popular ensemble learning algorithm that integrates multiple weak classifiers to create a superior classifier. In theory, Adaboost amplifies the impact of misclassified instances through higher weights, empowering subsequent weak predictors to focus on these misclassified instances. This iterative process continues until a strong classifier is obtained. Adaboost is based on the concept of boosting, where each weak learner is trained sequentially, with more emphasis placed on instances that were previously misclassified. One key advantage of Adaboost is its ability to handle complex datasets and achieve high accuracy. It is widely used in various applications, including face detection, text categorization, and speech recognition, where performance and accuracy are crucial. Adaboost is also known for being less prone to overfitting compared to other classifiers. Furthermore, Adaboost can handle both categorical and continuous variables, making it versatile for different types of data. However, it is important to note that Adaboost is sensitive to noisy data, which can adversely affect its performance. Therefore, data preprocessing and cleaning are essential to ensure optimal results when using this classifier. Overall, Adaboost offers a powerful technique for classification tasks, with the ability to improve the performance of weak learners and achieve robust classification results.

3.8. XGBoost Classifier

The XGBoost Classifier is an advanced gradient boosting algorithm that combines multiple weak classifiers to create a strong classifier. It uses a regularized objective function and a second-order approximation to capture intricate structures in the data. XGBoost can handle missing values and categorical variables, and it is known for its speed and efficiency. It implements parallel processing and tree pruning techniques to accelerate training and achieve faster predictions. XGBoost also provides built-in cross-validation and early stopping capabilities to optimize model performance. Overall, the XGBoost Classifier offers a powerful and efficient solution for regression and classification tasks.

3.9. CatBoost Classifier

The CatBoost Classifier is a gradient boosting algorithm designed to work with non-numeric features efficiently. It uses an innovative method called Ordered Boosting, which allows it to process categorical variables without the need for one-hot encoding. CatBoost incorporates a combination of random permutations and ordered boosting to handle categorical features effectively. It also introduces a novel gradient calculation method for categorical variables, improving accuracy and reducing overfitting. CatBoost is known for its robustness to noisy data and its capacity to handle extensive datasets. It delivers built-in cross-validation and early stopping capabilities to optimize model performance. Overall, the CatBoost Classifier offers a powerful solution for classification tasks with categorical features.

3.10. Comparative analysis

The results obtained from testing the pre-processed data [6] using various algorithms, including Logistic Regression, Decision trees, Random Forest, Naive Bayes, Adaboost Classifier, KNN Classifier, SVM Classifier Linear, Logistic Regression (Adaboost), Adaboost Classifier (Extra tree), Random Forest (Adaboost), SVM Classifier Poly, SVM (Adaboost), XGBoost Classifier, and CatBoost Classifier, are documented in the Table 1.

Table 1. Analyzation of these machine learning methods [6]

Model	Accuracy	Recall	Precision	F1-Score	AUC Score
LR	0.8045	0.8023	0.7911	0.7889	0.82
LR (Adaboost)	0.7657	0.7557	0.5661	0.6471	0.78
DT	0.8014	0.801	0.7881	0.7889	0.83
Adaboost	0.8171	0.8121	0.8014	0.8028	0.84
Adaboost (ExtraTree)	0.8114	0.8164	0.8057	0.8060	0.72
RF	0.7804	0.7868	0.7754	0.7791	0.82
RF (Adaboost)	0.8121	0.8128	0.8019	0.8029	0.82
NB (Gaussian)	0.7707	0.7712	0.7760	0.7731	0.80
SVM (Linear)	0.7914	0.7989	0.7867	0.7886	0.79
SVM (Poly)	0.8021	0.8064	0.7966	0.7811	0.80
SVM (Adaboost)	0.7407	0.7443	0.5491	0.6317	0.80
XGBoost	0.808	0.807	0.803	0.787	0.84
CatBoost	0.818	0.822	0.812	0.796	0.82

According to the result, Adaboost Classifier, which utilizes ensemble learning, achieved the highest accuracy among others, reaching 81.71%. Additionally, it demonstrated to achieve a high recall rate 81.21% and yielded high precision together with F1-Score. Furthermore, it obtained an AUC score of 84%. As a result, both Adaboost Classifier and XGBoost Classifier deliver the utmost notable outcomes.

4. Summary

In the final analysis, through this research paper, a comparative analysis of popular machine learning algorithms used for forecasting consumer churn in the modern business is concentrated on. First of all, the relative concepts of the customer churn and some basic application scenarios are comprehensive narrated, such as in telecommunication field. In addition, this paper provides a classic dataset, SyriaTel data, which is most frequently adopted by the researchers, together with four common performance evaluation indicators: precision, recall, accuracy, and F-measure. Last but not least, following time development and technological iteration development, it introduces and sorts nine well-known machine learning methods, including Regression Analysis-logistic Regression Analysis, Naive Bayes, Support vector machine, Decision trees, Random Forest classifier, Extra tree classifier, Adaboost, XGBoost classifier and CatBoost classifier, and accentuates some difference in effectiveness between them of those methods.

According to the obtained outcome analysis shown above, it seems that Adaboost Classifier and XGBoost Classifier give the most notable outcomes in customer churn prediction and they are the most suitable methods in dealing with such the problems. Nevertheless, there is no doubt that some limitations still exist in these two methods. Hence, a more comprehensive combination and utilization of them is the best move in customer churn prediction.

Although these machine learning algorithms have achieved good results, there is still ample opportunity for enhancement. Over the last few years, with the emergence of deep learning and especially big model related techniques, the study of customer churn will see new technological innovations.

References

[1] Mitrović S, Baesens B, Lemahieu W, et al.On the operational efficiency of different feature types for telco Churn prediction. European Journal of Operational Research, 2018, 267 (3): 1141 - 1155.

[2] Verbeke W, Dejaeger K, Martens D, et al. Customer churn prediction: does technique matter? Joint Statistical Meeting. 2010.

[3] Hadden J, Tiwari A, Roy R, Ruta D., Computer assisted customer churn management: State-of the-art and future trends. Computers & Operations Research, 2010, 34 (10): 2902 – 2917.

[4] Lejeune, M.A. Measuring the impact of data mining on churn management. Internet Res, 2001, 11: 375 - 387.

[5] Rajamohamed R, Manokaran J., Improved credit card churn prediction based on rough clustering and supervised learning techniques. Cluster Computing, 2018, 21 (1): 65 – 77.

[6] Lalwani, P., Mishra, M.K., Chadha, J.S. et al. Customer churn prediction system: a machine learning approach. Computing, 2021, (prepublish): 1 - 24.

[7] Dahiya, K., Bhatia, S.: Customer churn analysis in telecom industry. In: 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 2015. 1 – 6.

[8] Kisioglu P, Topcu YI. Applying bayesian belief network approach to customer churn analysis: A case study on the telecom industry of turkey. Expert Systems with Applications, 2011, 38 (6): 7151 - 7157.

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

[10] Sokolova, M., Japkowicz N., Szpakowicz, S.: Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: Australasian joint conference on artificial intelligence, 2006, 1: 1015 – 1021.