Risk Decision and Predicting of Customer Churn Based on Principal Component Analysis

Shiyu Cui[1(✉)], Penghan Lai[2], Yuwei Deng[3], and Xiaojiang Zheng[4]

[1] Information School, Yunnan University of Finance and Economics, Kunming, China
201905001194@stu.ynufe.edu.cn
[2] New College, University of Toronto St George Campus, Toronto, Canada
Penghan.lai@mail.utoronto.ca
[3] The USC Dornsife College of Letters Arts and Sciences, University of Southern California, California, USA
yuweiden@usc.edu
[4] School of Tourism Sciences, Beijing International Studies University, Beijing, China
2018220996@bisu.edu.cn

Abstract. This study will establish a predictive analytics model that uses churn prediction models to anticipate customer churn by evaluating their risk of churn. These models are successful in focusing customer retention marketing activities on the fraction of the customer base that is most prone to churn because they generate a short, prioritized list of probable defectors. We will start with exploratory data analysis in this paper. We will get a quick summary of the data using this way. The data is then further analyzed using feature engineering and feature selection. Finally, the target variables will be visualized using Principal Component Analysis (PCA). Using the Kolmogorov-Smirnov (KS) score test, features are picked after calculating the churn detection rate and comparing it to the average churn rate. The best part is then identified using Cross-Validated Recursive Feature Elimination. The accuracy, auc, and ks of each model were assessed after training, and the Gaussian naive Bayes, Logistic regression, and Neural network were finally picked by comparison. Model stacking is a technique for comparing model performance and ultimately deciding which model to utilize.

Keywords: Customer churn · Exploratory data analysis · feature engineering · feature selection · comparison of models

1 Introduction

Customer churn, defined as the discontinuous usage of service products, has received wide attention in the field of service businesses [10]. This is because customer churn is associated with brand loyalty, as the high customer churn rate is considered as the low brand loyalty. In addition, the customer churn will also have a great effect on the operation's effectiveness. First, the loss of customers could affect a company's valuation.

S. Cui, P. Lai, Y. Deng and X. Zheng—These authors contributed equally.

Z. Zeng et al. (Eds.): ECIT 2022, AHE 11, pp. 693–701, 2023.
https://doi.org/10.2991/978-94-6463-005-3_71

When a company's customer churn rate is high, the potential investors will value the company less. In their eyes, there must be some problems inside the company, and these problems will affect the development of the company. Second, the loss of customers will affect the company's market size. When predicting the potential market size of the company, it needs to exclude the customers who have already lost so that the market size range that the company can expand is gradually reduced. When the growth rate of customers is lower than the loss rate of customers, the company's future development will also have a problem. Third, the loss of customers will bring customers to competitors. When customers need a kind of product, if they do not choose the original company for some reason, they will choose other companies, which means that the customers of our company will become the customers of our competitors.

According to the above points, we can know that the loss of customers will bring many bad effects to the company. Therefore, it is very important to find out the reason for customer loss.

In this paper, we will use Exploratory data analysis first. Through this method, we will first have an overview of the data, and then we will judge the various influencing factors separately and make a more intuitive comparison by drawing some graphs (like a bar chart, pie chart, scatter chart, etc.). By comparing these charts, we can get a rough idea of the factors that influence customer loss. Based on the results of Exploratory data analysis, we further analyze the data through feature engineering and feature selection next. In feature engineering, some new variables will be built first. Secondly, to make the data could be recognized by algorithm models, this notebook adopted the targeted encoding and one-hot encoding to turn the original data into categorical data. Then draw a heatmap and use different colors to represent the correlation between the data. Finally, Principal Component Analysis (PCA) will be used to visualize the target variables. In the feature selection, after calculating the churn detection rate and comparing it to the average churn rate, features are selected using the Kolmogorov-Smirnov (KS) score test. Then, Cross-Validated Recursive Feature Elimination is used to identify the best feature.

According to the above data processing results, we analyzed the results by building a model. However, before building the model, we first assess the fairness of the data and modify the categorical variables before establishing the model. Base model includes logistic regression, Decision tree, K nearest neighbor, Random Forest, Gaussian naïve Bayes, Light GBM, XGboost, Gradient boosting, and Neural network. After the training of each model, the accuracy, AUC, and ks of each model were compared, and the Gaussian naïve Bayes, Logistic regression, and Neural network were finally selected by comparison. Finally, a random forest classifier was used to adjust the optimal performance and re-evaluate the data by calculating the importance of features. Model stacking is used to compare the performance of models and finally select the model to use.

2 Data

First, we get some data. We will get a form (Table 1), which has five rows and 21 columns. And in this form, it has some objects like customerID, SeniorCitizen, and so on. We hope that we can use this data to find out the factors that influence customers to choose the same company, but of course, these factors do not include the product's characteristics.

Table 1. Basic information about the customers.

Customer ID	gender	senior citizen	Partner	Dependents	tenure	phone service	MultipleLines	internet service	…
7590-VHVEG	*Email*	0	Yes	No	1	No	No phone service	DSL	…
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	…
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	…
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	…
9237-HAITI	*Email*	0	No	No	2	Yes	No	Fiber optic	…

Then from the Data Overview (Table 2), we will get some information about the objects, like the datatype, records, populated, and so on. From the data, we can see that the #Unique values of the customerID are 7043, which means that there are 7043 unique customers in this company. There are some similarities and some differences among these customers. Next, we can make a simple grouping of customers according to these characteristics so as to better analyze the influencing factors of customer loss. In data manipulation, we will first inspect data and find out that only the "total charges" have missing values, and the number is 11.

Then we will group the objects by gender, senior citizen, partner, and so on. After that, we can set categorial columns to type objects and print possible values.

And then there are two problems for us to solve.

- The first one is why the "TotalCharges" cannot be directly cast to float type?
- The second one is how we can cast the "TotalCharges" to float type?

To find out why the "TotalCharges" cannot be directly cast to float type, we can define a function. The function is about to check if a variable can be cast to type float. If the variable can be cast to type float, it will return true. Otherwise, it will return false. From the data (Table 3), we can see that there are 11 missing values in "TotalCharges", so that's why the "TotalCharges" cannot be directly cast to float type. For this reason, we can set "TotalCharges" to 0. And then, from the form, we can easily find out that "TotalCharges" is very close to "MonthlyCharges" times "tenure".

$$\text{total charges} \approx \text{MonthlyCharges} * \text{tenure} \tag{1}$$

At last, we will separate churn and not-churn customers. Then we will get an overall churn rate and the number of the churn and not-churn. These data will help us to continue the next part.

- Overall churn rate is 0.2653698707936959
- churn (1869,21)

Table 2. Customer data statistics.

	Data type	# Nonnull records	#Non-zero records	% Populated	# Unique values	Mean	Std
customerID	object	7043	7043	100.0	2	NaN	NaN
gender	object	7043	7043	100.0	2	NaN	NaN
senior citizen	int64	7043	1142	100.0	2	0.16	0.37
Partner	object	7043	7043	100.0	2	NaN	NaN
Dependents	object	7043	7043	100.0	2	NaN	NaN
tenure	int64	7043	7.32	100.0	73	32.37	24.56
PhoneService	object	7043	7043	100.0	2	NaN	NaN
MultipleLines	object	7043	7043	100.0	3	NaN	NaN
InternetService	object	7043	7043	100.0	3	NaN	NaN
OnlineSecurity	object	7043	7043	100.0	3	NaN	NaN
OnlineBackup	object	7043	7043	100.0	3	NaN	NaN
DeviceProtection	object	7043	7043	100.0	3	NaN	NaN
TechSupport	object	7043	7043	100.0	3	NaN	NaN
StreamingTV	object	7043	7043	100.0	3	NaN	NaN
StreamingMovies	object	7043	7043	100.0	3	NaN	NaN
Contract	object	7043	7043	100.0	3	NaN	NaN
PaperlessBilling	object	7043	7043	100.0	2	NaN	NaN
PaymentMethod	object	7043	7043	100.0	4	NaN	NaN
MonthlyCharges	float64	7043	7043	100.0	1585	64.76	30.09
TotalCharges	object	7043	7043	100.0	6531	NaN	NaN
Churn	object	7043	7043	100.0	2	NaN	NaN

Table 3. Data about tenure, MonthlyCharges and TotalCharges.

	tenure	MonthilCharges	total charges
count	*7043.000000*	7043.000000	7043.000000
mean	32.371149	64.761692	64.761692
std	24.559481	30.090047	2265.156206
min	0.000000	18.250000	18.800000
25%	9.000000	35.500000	402.225000
50%	29.000000	70.350000	1400.550000
75%	*55.000000*	89.850000	3786.600000
max	72.000000	118.750000	8684.800000

3 Data Preparation, Feature Engineering, and Feature Selection

3.1 Data Preparation

Exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often using graphics and other data visualization methods. Typically, the EDA starts by looking at each variable separately, searching through the data, checking the shapes of distributions, and looking for outliers and rogue values. After that, the exploratory data analyst explores relationships between pairs of variables before moving on to multivariate relationships [9]. In this part, first, we take an overview of these data, then subdivide them according to the situation of different variables. After that, we split these variables according to whether customers remain existing.

First of all, this paper will draw a histogram to compare the number of customers that churn or not. We can find that the number of customers who did not leave is more than the number left. Next, we analyze the distribution of variables. In the data part, their variables are already split into categories and numerical. After that, we introduce a new variable: churn customers or non-churn customers. Then conduct the same analysis as in the previous step. The difference is that the Y-axis is no longer refers to the number but a percentage. For the categorical fields, we can find that for the elderly and young groups, the customer churn rate is completely different. And for the contract distribution, the longer the contract is, the fewer customers leave, obviously. At last, we connect the three variables in pairs to see their correlation. The scatter plots and 3D visualization graph of tenue, Monthly Charges, and Total Charges will be graphed. For the tenure and total charges, they are correlated because the trend is the same, just the discrete degrees are different. And for the tenure and monthly charges, there is little correlation. Exploratory data analysis is effective and necessary because factorization methods break data into a matrix product that can show outliers, clusters of similar observations, groups of related variables, and crossing relationships between observations and variables [3].

3.2 Feature Engineering

A feature is a numeric measurement from the raw data [7]. Accordingly, feature engineering is defined as an iterative process of creating new features drawing on raw data to improve the learning algorithm's predictive power [6, 7]. The principle of feature engineering is capturing additional information that is not obviously apparent in the raw feature set. In this notebook, the feature engineering is as follows.

Firstly, several new variables were built to clean and generate feature sets from tremendous data. For example, to measure the sum of customers from different services, the variable 'TotalService' was built, which combined the categories of 'PhoneService' and 'InternetService'. Secondly, the notebook jointly adopted the targeted encoding and one-hot encoding to turn the original data into categorical data, which could be recognized by algorithm models. Cerda et al. [4] highlighted that target encoding is the process of replacing features with a blend of the probability of the target with specific categorical values from all the training data. In this notebook, after storing the raw data, the categorical features were paired with other legibly categorical features. Followingly, the categorical variables with two categories were changed from the dummy variables

for performing target encoding. Specifically, to examine the target encoding, the train test sets were split. As for One-hot encoding, it is a procedure of producing one feature per category with each binary [2]. This notebook generates two matrices; one is accomplished by the one-hot encoding, the other is achieved by the target encoding. Consequently, the count, the mean, the minimum value of target encoded variables, and one-hot encoded variables were represented in this notebook. Thirdly, the notebook operated the correlation between each variable through Pearson's correlation coefficient. In this notebook, the heatmap was plotted to present the correlation by color attribution. Fourthly, considering that Principal Component Analysis (PCA) is a procedure of reducing dimension without linear correlation [11], PCA was adopted to visualize the targeted variables and one-hot variables to maximize the dispersion of reflected samples.

3.3 Feature Selection

After the process of feature engineering, this notebook selected the core subset of features for dimensionality reduction, namely feature selection. At first, this notebook calculated the churn detection rate as 0.3, which is higher than the average churn rate of 0.265. By performing a Kolmogorov-Smirnov (KS) score test, which is used to test the distribution of samples [4], features could be selected from the scope of calculated scores. Finally, 27 features were selected by the filter. Followingly, Cross-Validated Recursive Feature Elimination was adopted to rank the features according to the degree of importance. This approach is used to identify the best features by eliminating the lesser important along with cross-validation [12]. This notebook operated recursive feature elimination through the Scikit-learn library to use estimators that have "feature_importances_" attributes. Consequently, the optimal features are the top 20 features, which are used for the following model building section.

4 Results Analysis

Before building any model, fairness of data should assess, and augmentation should be made if necessary. In this case, churn and non-churn have a large difference in terms of available data, which non-churn only accounts for 26.54% of total data. Non-churn data is considered the minority class. To balance the class distribution, a technique known as SMOTE, Synthetic minority oversampling technique was implanted. The mechanism is that by creating synthetic examples from the minority class, but without interfering with existing data, therefore achieve a relatively fair dataset. Since the above method only dealt with numeric variables, categorical variables require modification in order to proceed to model building. One-hot encoding and target encoding were used.

Base model includes logistic regression, Decision tree, K nearest neighbor, Random Forest, Gaussian naïve Bayes, Light GBM, XGboost, Gradient boosting, and Neural network. Each of the models was trained with and without SMOTE set. A summary table was made for target encoding and one-hot encoding, comparing the accuracy, auc, and ks for each model (Table 4). Noted that one-hot encoding dataset performs very similar to target encoding, and SMOTE dataset and original dataset present no significant changes, indicating no significant increase on model performance after upsampling the

Table 4. Summary of accuracy, AUC and ks score on 3 base model.

Model	Accuracy	AUC	ks
Logistic Regression	0.805469	0.846136	0.554803
Gradient Boosting	0.804048	0.847637	0.559285
Neural Network	0.801562	0.843220	0.554689

minority class. According to the performance on both one-hot encoding dataset and target encoding dataset, Gaussian naïve Bayes, Logistic regression, and Neural network were chosen given their performance ware best out of all models on both modified datasets.

Random forest classifier was used to adjust for optimal performance, reassessed numbers of the feature that should be included by calculating the feature importance, total charges, monthly charges, and tenure are the most important features. The next step is to test for abundant features that can be excluded without altering the performance. Although some of the features when dropped, the performance increased. However, dependence on some features impacts the final prediction a lot. At this stage, In terms of turned base models, gradient boosting is the best model by far.

Model stacking is one way to seek performance. After model stacking and turning this model, the accuracy, auc, and ks scores are 0.801919, 0.846658, and 0.553109, respectively. Compared to all the other 3 base models, the performance of the stacking model did not increase significantly.

Evaluating the model on hold-out set is the final step for reassuring model performance no matter whether the overfit or underfit occurs. The turn-stacking model result on hold-out set is 0.794180, 0.827271, and 0.495479, representing accuracy, auc, and ks, respectively. Compared to the cross-validated scores on the training set, the number is slightly lower, which indicates no overfit occurs on the models.

Finally, it's recommended that the logistic regression model with upsampling method is better in terms of computation and interpretability.

5 Conclusion

Customer churn is an important topic for businesses associated with brand loyalty. Previous studies have highlighted that service companies endeavor to reduce the customer churn rate considering the worth brought by a long-term customer compared with a newly recruited customer [10]. In addition, customer churn analysis and prediction are essential to marketing digital transformation for service industries. However, although current studies have investigated customer churn in the logistic industry [5], Business-to-Business (B2B) [8], there remains a paucity of studies focused on telecom companies. Although studies have preliminarily discussed customer churn in telecommunication by focusing on certainty data [1], more comprehensive and accurate models for customer churn in telecommunication industries still need to be refined further. Based on python Jupyter notebook techniques, this study presented the process of data processing, feature engineering, feature selection and jointly investigated the computation and interpretability of different models.

In this study, turnover prediction models are used in predictive analytics to anticipate customer attrition by measuring their risk of churn. These models are successful in focusing customer retention marketing activities on the fraction of the customer base that is most prone to churn because they generate a short-prioritized list of probable defectors. The paper generates several variables and set some missing values to be plausible and find the distribution of churn by different type using histogram, scatter plot and 3D visualization based on the original dataset. Then Creates new variables (such as ratio variables and statistical features) and summarizes them. To discover the attributes, we used a correlation matrix and PCA. They are finding the 13 best factors for further investigation using the KS score and filter. To acquire accuracy, auc, and ks score, we used logistic regression, Decision Tree, KNN, random forest, Gaussian naive Bayes, light GBM, XGBoost, Gradient Boosting, and Neural network in the analysis part. We concluded that the logistic regression model with upsampling method is fitted with the computation and interpretation of customer churn.

6 Future Research Direction

In the future, several directions can be taken for investigating further. Firstly, this study just adopted the customer data from a single case telecom company. Diverse data from different companies or other service industries could be collected to compare different customer churn models further. Secondly, researchers could consider the enhancement of the predictability of existing models as current customer churn models are still needed to refine further by improving the efficacy and accuracy. More nascent models could be considered and compared together.

References

1. Amin A, Al-Obeidat F, Shah B, Adnan A, Loo J, Anwar S (2019) Customer churn prediction in telecommunication industry using data certainty. J Bus Res 94:290–301
2. Bro R, Smilde AK (2014) Principal component analysis. Anal Methods 6(9):2812–2831
3. Camacho J, Rodríguez-Gómez RA, Saccenti E (2017) Group-wise principal component analysis for exploratory data analysis. J Comput Graph Stat 26(3):501–512
4. Cerda P, Varoquaux G, Kégl B (2018) Similarity encoding for learning with dirty categorical variables. Mach Learn 107(8–10):1477–1494. https://doi.org/10.1007/s10994-018-5724-2
5. Chen K, Hu Y-H, Hsieh Y-C (2014) Predicting customer churn from valuable B2B customers in the logistics industry: a case study. IseB 13(3):475–494. https://doi.org/10.1007/s10257-014-0264-1
6. de Melo VV, Banzhaf W (2018) Automatic feature engineering for regression models with machine learning: an evolutionary computation and statistics hybrid. Inf Sci 430:287–313
7. Dong G, Liu H (2018) Feature engineering for machine learning and data analytics. CRC Press
8. Jahromi AT, Stakhovych S, Ewing M (2014) Managing B2B customer churn, retention and profitability. Ind Mark Manage 43(7):1258–1268
9. Maznah Z, Halimah M, Shitan M, Kumar Karmokar P, Najwa S (2017) Prediction of hexaconazole concentration in the top most layer of oil palm plantation soil using exploratory data analysis (EDA). PLoS ONE 12(1):e0166203

10. Qi JY, Zhang L, Liu YP, Li L, Zhou YP, Shen Y et al (2009) ADTreesLogit model for customer churn prediction. Ann Oper Res 168(1):247–265
11. Shoaib M, Naveed MS, Sanjrani AA, Ahmed A (2021) A comparative study of contemporary programming languages in implementation of classical algorithms. J Inf Commun Technol 14(1):23–32
12. Ustebay S, Turgut Z, Aydin MA (2018) Intrusion detection system with recursive feature elimination by using random forest and deep learning classifier. Paper presented at the 2018 international congress on big data, deep learning and fighting cyber terrorism (IBIGDELFT)