Project Report

Customer Churn Prediction

Cheryl Ye

Introduction

This dataset from Kaggle (https://www.kaggle.com/blastchar/telco-customer-churn) is about the customer behavior of a telecommunication industry, and the goal is to study about what makes the customer churn. The dataset has 21 columns to illustrate each client's situation. This project will use KNN, Naïve Bayes, Decision Tree, Random Forests, and Logistic Regression to find out which variable make the clients to unsubscribe the telecommunication service. This project provides business insights through analyzing the related variables of clients. By knowing what variable will make customers unsubscribe the telecommunication service, the telecommunication companies can formula strategies to prevent clients churning.

Data Exploration

First of all, after reading the dataset from Kaggle website, I printed the numbers of rows and columns of the dataset. Results are shown as below:

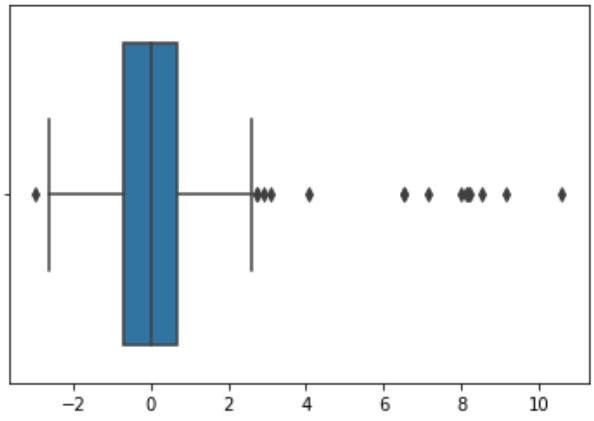
Number of rows: 7043

Number of columns: 21

This step helps us to have a basic roughly overview of the dataset size. So we can see there are 7043 clients' information, and 21 variables to describe the clients' situation.

Data Cleaning

I checked there are no duplicated value in the dataset. And then I check for the missing value. I decide to drop the rows if there is a missing value. To find out the outliers, I draw the boxplot as below:



Even though the boxplot shows this dataset has outliers, I decide no to drop any of them. Because to define an outlier requires sufficient domain knowledge and industrial experience.

Feature Preprocessing

For feature preprocessing, I assigned column “Churn” as the y value. In this column, there are two types of values: “Yes” and “No”. I assigned “Yes” as 1, and “No” as 0. And then I think the column “customerID” has no numerical relationship with the prediction, so I dropped this irrelevant column. For the columns that has “Yes” and “No”, I transform yes and no to True and False. For the columns that has “Yes”, “No” and another third value; I transform no to 0, yes to 1, and the third value to 2. For the columns that has various values, I used label encoding to encode them into numerical values. Moreover, I made sure all data type are ready for model training.

Model Training & Evaluation

In this project, I used five classifiers to train the dataset: K Nearest Neighbor, Naïve Bayes, Decision Tree, Random Forests, and Logistic Regression. Before doing the training, I do standard scalar for the dataset. Because standard scalar can speed up gradient descent and train

the dataset in the same scale. While training the dataset, I set 5-fold cross validation to reduced bias. Below are the results

[0.78419183 0.75377778 0.74844444 0.77066667 0.77046263]

Model accuracy of KNN is: 0.7655

[0.75222025 0.73955556 0.77333333 0.75288889 0.75800712]

Model accuracy of Naive Bayes is: 0.7552

[0.75133215 0.72444444 0.72711111 0.74666667 0.72953737]

Model accuracy of Decision Trees is: 0.7358

[0.79040853 0.77511111 0.78311111 0.78577778 0.79448399]

Model accuracy of Random Forest is: 0.7858

[0.79928952 0.78577778 0.80177778 0.80088889 0.81405694]

Model accuracy of Logistic Regression is: 0.8004

The first line of the result is the accuracy that generated by 5-fold cross validation, and the second line is the average result. From the result we can see Logistic Regression has the highest score.

In our case, we value the model performance by accuracy. Because in this case, the impact of false positive and false negative is the same. The false positive situation is we predict the client will churn but in the reality the client does not churn. This will end up our company give out too much promotions to the client in order to keep the clients. The false negative situation is we predict the client will not churn but in the reality the client does churn. Since we have the wrong prediction, we did not take any action to prevent clients to churn. As a result, we

can see the impact of these two errors are the same. The formula of accuracy is true positive / (true positive + false negative + false positive + true negative), which can evaluate all situations.

Features Selection

After deciding to use Logistic Regression as our prediction model, I did grid search to find out whether we should input L1 or L2 into the model. The grid search result is

Best score: 0.8004

Best parameters set:

C: 1

penalty: 'l2'

So when we use C = 1 and L2 in the Logistic Regression model, we will reach the best result of Logistic Regression model. The highest accuracy is 0.8004.

In the Feature selection part, I rank the futures' importance from high to low. The result of ranking is shown as below:

Logistic Regression (l2) Coefficients

tenure: -1.4026

TotalCharges: 0.6469

Contract: -0.58

InternetService: 0.4512

OnlineSecurity: -0.372

TechSupport: -0.3577

MonthlyCharges: 0.2485

PaperlessBilling: 0.1792

OnlineBackup: -0.1667

MultipleLines: 0.1436

StreamingMovies: 0.1114

PhoneService: -0.1077

StreamingTV: 0.1067

SeniorCitizen: 0.088

DeviceProtection: -0.08

Dependents: -0.0751

PaymentMethod: 0.0525

gender: -0.0133

Partner: 0.0072

We can see the top five related variables are: tenure, total charges, contract, internet service, and online security. This means a client that has shorter tenure, higher charges, shorter contract, no internet service and no internet service is more likely to churn.

Business Insights

To prevent client churning from the telecommunication company, the company can give promotion for the clients who have longer tenure to attract clients to stay longer with the company. The company could also give out promotions to those predicted churn clients on their total charges to prevent them churning. Moreover, since the shorter contract clients will easier churn, the company can give out promotion for clients who are willing to sign a longer contract.

Conclusion

Through conducting the data exploration, feature preprocessing, model training and evaluation, we are able to get the feature ranking and business insight of the project. This project analyzes the factors that make the client unsubscribe the telecommunication service. The decision makers in the company can use the feature ranking list to coordinate with the marking department to formula the pricing and service strategies to prevent the clients churning from the company.