Telecom customer churn analyzed by two machine learning models

Chih-Ying Ho

Royal Melbourne Institute of Technology and

Melbourne Technical College, Master of Data Science

Contact details: s3816723@student.rmit.edu.au

Table of Content

Abstract

In this report, building machine learning models is used to analyze what causes customer churn. This study firstly explore data to have an understanding of dataset. Then, this report fits data into decision tree model and logistic regression model and validates two models to choose a better one. The result shows that logistic regression model is more accurate than decision tree classifier. Therefore, it is recommended that logistic regression is an effective way of achieving the goal.

1. Introduction

Customer retention is important to a company, because retaining customer spends less than acquiring new customers (Galetto, 2016). Based on hubspot, if a company rise customer retention rate by 5%, a company would increase its profit by 25%-49%. Therefore, finding out the reason of customer churn is a critical issue. This report will compare two models and provide a more accurate one to analyze telecom customer churn.

2. Methodology

2.1 Tools and Resources

Python is a useful tool to do data analysis because of its powerful packages. It is used to manipulate, visualize and present data. In this report, we use pandas, numpy, sklearm, matplotlib library using Jupyter Notebook Environment.

The report analyzed telco customer churn using from Kaggle. The data includes 21 attributes whose types are object, integer and float (Figure 1).

Attributes	Type	Description
customerID	object	Customer ID
gender	object	Whether the customer is male or female
SeniorCitizen	int64	If the customer is a senior citizen, the data is labeled '1'. If not, it is labeled '0'.
Partner	object	If the customer has a partner, the data shows 'Yes'. If not, it shows 'No'.
Dependents	object	If the customer has dependents, the data is 'Yes'. If not, it is 'No'.
Tenure	int64	How long have the customer stayed with the company

PhoneService	object	If the customer has a phone service, the data is 'Yes'. If not, it is 'No'.
MultipleLines	object	If the customer has multiple lines, the data is 'Yes'. If not, it is 'No' or 'No phone service'.
InternetService	object	Customer’s internet service provider including DSL, Fiber optic and No.
OnlineSecurity	object	If the customer has online security, the data is 'Yes'. If not, it is 'No' or 'No internet service'.
OnlineBackup	object	If the customer has online backup, the data is 'Yes'. If not, it is 'No' or 'No internet service'.
DeviceProtection	object	If the customer has device protection, the data is 'Yes'. If not, it is 'No'.
TechSupport	object	If the customer has tech support, the data is 'Yes'. If not, it is 'No' or 'No internet service'.
StreamingTV	object	If the customer has streaming TV, the data is 'Yes'. If not, it is 'No' or 'No internet service'.
StreamingMovies	object	If the customer has streaming movies, the data is 'Yes'. If not, it is 'No' or 'No internet service'.
Contract	object	The contract term of the customer, including Month-to-month, One year and Two year
PaperlessBilling	object	If the customer has paperless billing, the data is 'Yes'. If not, it is 'No'.
PaymentMethod	object	The customer’s payment method including Bank transfer (automatic), Credit card (automatic), Electronic check, Mailed check
MonthlyCharges	float64	The customer’s monthly fee
TotalCharges	object	The customer’s total fee
Churn	object	If the customer churn,

Figure 1

2.2 Data Preparation

The data is measured including the size and the type and is checked whether there are typos and missing values. In addition, the data is checked if there are outliers using interquartile range (IQR) rule. The data is found that there is whitespace in TotalCharges attribute and the type of this attribute should be float. The whitespace seems to be NaN. Therefore, the problem is solved by removing the rows with whitespace and changing the type to float. Besides, there are typos in the data. The information written ‘No internet service’ is same as ‘No’. Hence, the former is replaced

with the latter.

In order to explore the relationships between customer churn and other attributes, the nominal attributes are illustrated by pie charts and numerical attributes are demonstrated by line charts. Besides, the data is divided into two parts, churn and no churn. Furthermore, the categorical variables are encoded and the numerical variables are standardized to fit data into model.

2.3 Data Exploration – Visualization

Telco customer churn proportion

The percentage of customers churn is about 26%, which is lower than the percentage of customer retention, that is approximately 73% (Figure 2). This shows that higher proportion of customers is likely to stay in the same telecommunications company. In addition, the bar plot indicates that 26 percent of customer stop doing business with their company.

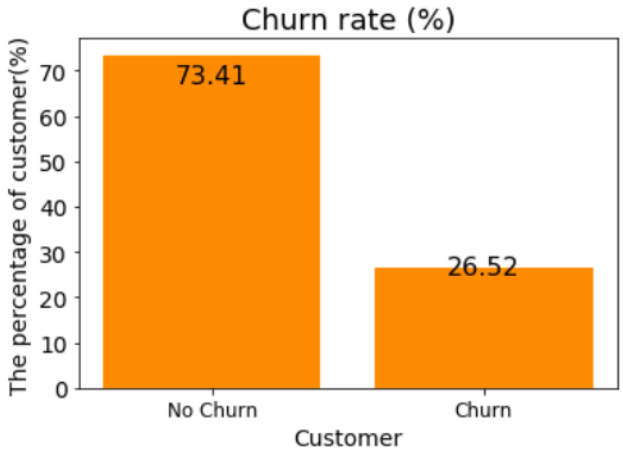

Figure 2

The relationship between churn and gender

In terms of customer churn, the percentages of female and male are close to each other, which are nearly 50%. Likewise, in customer retention, the proportions of men and women are approximately 50%. Therefore, there is no significant relationship between churn and gender.

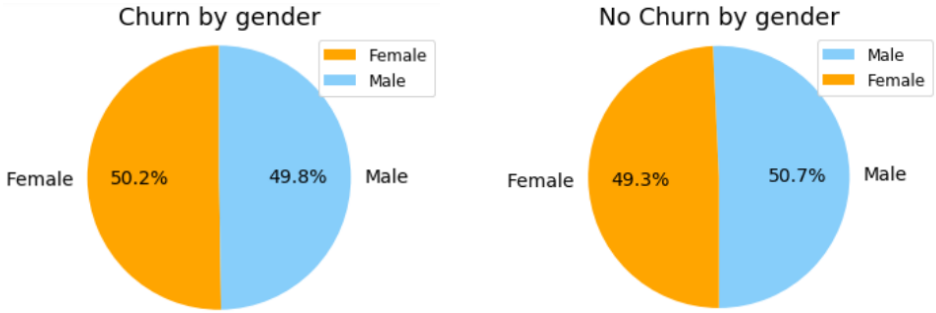

Figure 3. telecom customer by gender

The relationship between churn and seniorcitizen

In Figure 4, the graph shows that more than 25 percent of customer attrition is senior citizen, while the figure is only 12.9 percent in customer retention. It seems that senior citizens make up a small proportion of the customer retention.

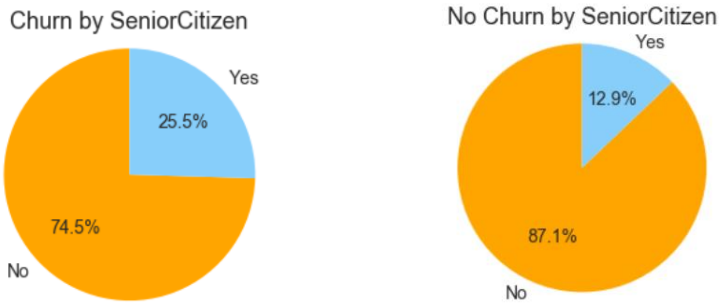

Figure 4. telecom customer by senior citizen

The relationship between churn and dependents

As Figure 5 shown, the large proportion of churn that is caused by customers who have dependents, which is more than 80 percent. Although more than 50 percent of retention is people who support dependents, the figure is still significantly lower than the percentage of clients who need to take care of family.

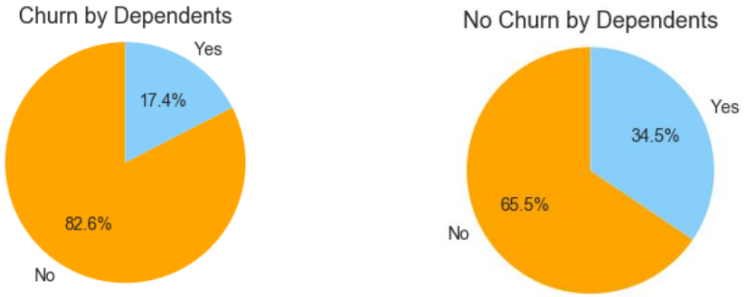

Figure 5. telecom customer by dependents

The relationship between churn and multiple lines

The figure indicates that no matter customers have multiple lines or not, it does not affect customer churn. In customer attention, roughly 45 percent of customer has multiple lines, while the figure is roughly equal to the figure of people who do not use multiple lines. The similar situation happens in customer retention.

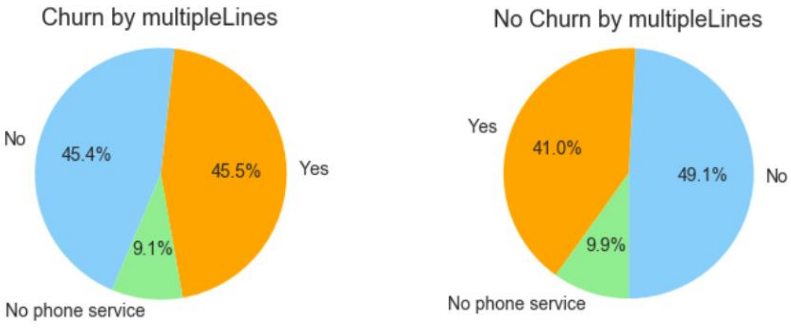

Figure 6. telecom customer by multiple lines

The relationship between churn internet service

Internet service, online security and device protection are important factors of customer churn. As you can see, nearly 70 percent of customer churn use fiber optic, while DSL comprises of the largest proportion in customer retention.

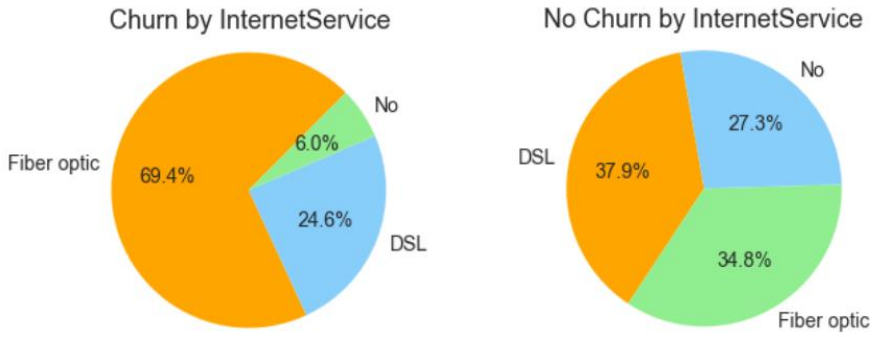

Figure 7. telecom customer by internet service

The relationship between churn and contract distribution

The figure shows 88 percent of customers who has month-to-month contract with telecom is the largest proportion in customer churn, while the figure is only 43 in customer retention. This indicates that if people sign month-to-month contract, the company is more likely to lose the client. In addition, two-year contract seems the best for the telecom because two-year contract accounted for the smallest proportion in customer churn and it is the second highest in customer retention.

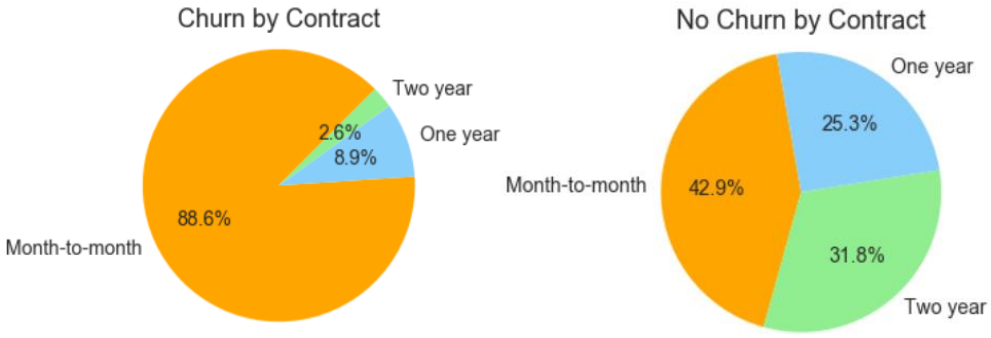

Figure 8. telecom customer by contract

The relationship between churn and monthly charge

As you can see from Figure 9 and 10, clients who pay for less than 60 per month are more likely to be customer retention. In histogram plot, the same situation mentioned happens. However, if people pay for monthly charge between 70 and 100, the chance of making these customers retaining is low. In addition, there is something unexpected happening. It is found that customers who pay for more than 110 tend to keep buying service with the same telecom. Therefore, generally, if the monthly charge is low

enough, the telecom tends to retain clients. The reason probably is that low price helps customer to pay less.

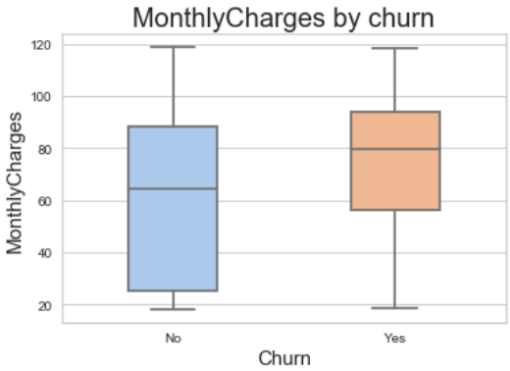

Figure 9. monthly charges boxplot grouped by churn

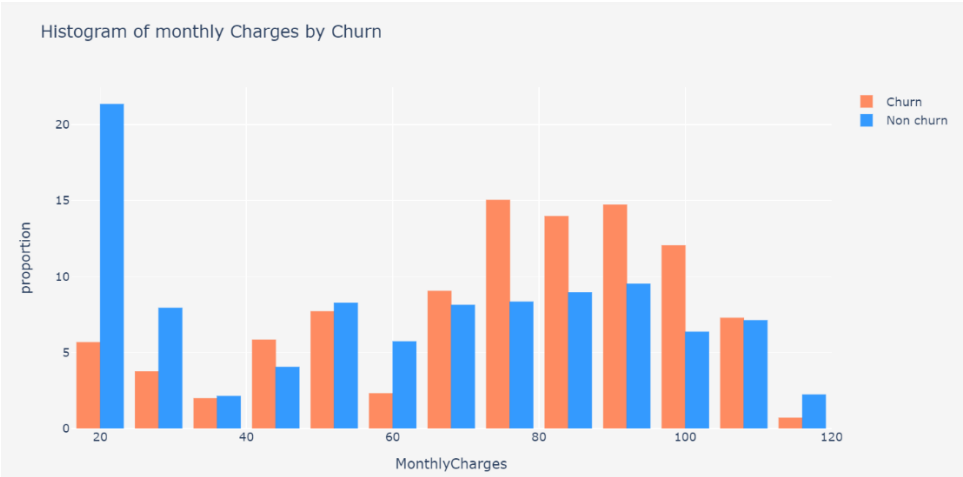

Figure 10. monthly charges histogram grouped by churn

People who pay for higher monthly charge and use Fiber optic or DSL tend to keep having business with the telecom. As you can see from Figure 11, people who pay for more than 60 per month and use DSL or who pay for over 90 and use fiber optic generally are customer retention.

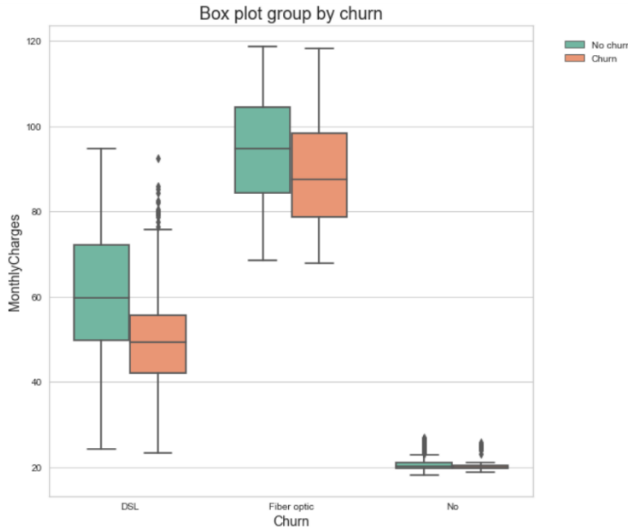

Figure 11. Internet service-monthly charges boxplot grouped by churn

The relationship between churn and tenure

In customer churn, the tenure is between 2 and 30, while in customer retention, the figure is from 15 to 60. This shows that if the tenure of customer is more than 30, the telecom is more likely to retain customers.

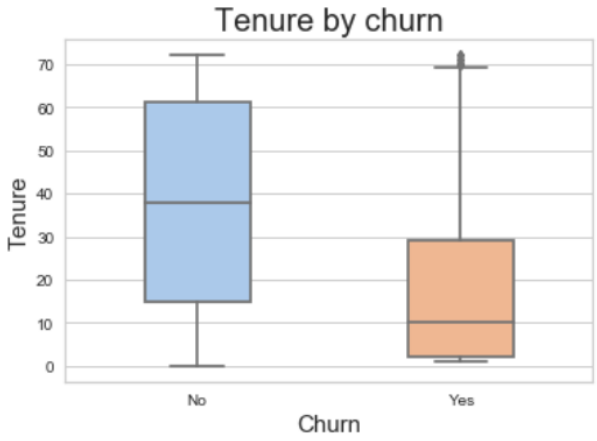

Figure 12. tenure boxplot grouped by churn

As you can see from Figure 13, 20 is boundary. If the tenure is less than 20, it hard for telecom to retain customer. However, if the tenure is more than 20, telecom is more likely to turn customers into repeated customers. In addition, when the tenure is longer, the chance of retaining customers is higher.

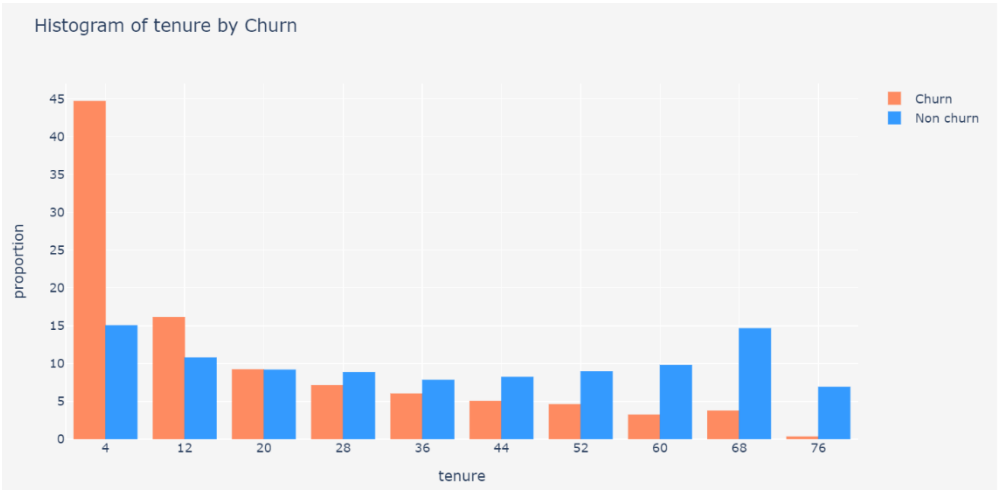

Figure 13. the proportion of tenure histogram grouped by churn

The relationship between churn and total charge

More than half customers whose total charge is lower than 500 make up customer churn. In addition, when people's total charge is more than 3500, the proportion of customer retention is higher than that of customer attrition. Therefore, telecom tends to retain their customers if customer's total charge is high.

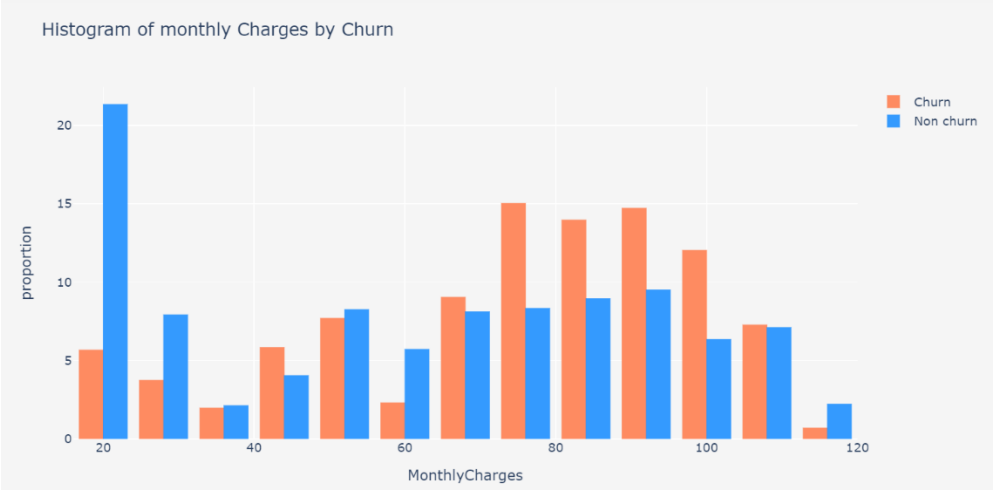

Figure 14. the proportion of monthly charges histogram grouped by churn

2.4 Data modeling

There are two machine learning models created in this report. One is decision tree, which has to prone to prevent overfitting, the other is regression logistic. The first step is feature selection, because some of features are not meaningful to the target variable. After training the models, which is fitting data into the models, the next step is model validation. This shows whether the model is good or not.

In decision tree model, the result shows that the accuracy classification score is 0.68, and the area under the receiver operating characteristic curve (AUC/ROC) is 0.76. In logistic regression model, the accuracy classification score is 0.74, and AUC/ROC is 0.8. Both of the figures are higher than that of decision tree model. This presents that the logistic regression model is better, because if accuracy classification score and AUC/ROC are higher, this indicates that the model is more accuracy. Both

```
Classification report :
              precision    recall  f1-score   support

           0       0.83      0.87      0.85      2061
           1       0.58      0.50      0.54       752

    accuracy                           0.77      2813
   macro avg       0.70      0.68      0.69      2813
weighted avg       0.76      0.77      0.76      2813

ROC/AUC curve                  :  0.683284813197477
Accuracy classification score :  0.7685744756487736
```

Figure 15. The result of decision tree model

```
 Classification report :
               precision    recall  f1-score   support

           0       0.87      0.87      0.87      2061
           1       0.64      0.63      0.63       752

    accuracy                           0.80      2813
   macro avg       0.75      0.75      0.75      2813
weighted avg       0.80      0.80      0.80      2813

ROC/AUC curve                 :  0.7491721897034077
Accuracy classification score :  0.8044792036971206
```

Figure 16. The result of logistic regression model

4. Conclusion

Retaining customers is more cost-effective than acquiring new customers. Based on the analysis, customer churn is mostly caused by money and the Internet. In addition, people care about the quality of internet more than payment. People are willing to pay more if the quality of internet is good. Besides, if customers stay in the telecom longer, the telecom is more likely to prevent customers switching to another company. In terms of building models, if the score is higher, it shows that the model is more accurate. Logistic regression model has higher ROC/AUC curve and accuracy score. Therefore, I suggest that applying logistic regression model is a better way to predict customer behaviors to retain clients.

5. Reference

Bernazzani, S, n.d., Here's Why Customer Retention is So Important for ROI, Customer Loyalty, and Business Growth, https://blog.hubspot.com/service/customer-retention

Galetto, M, 2016, What is Customer Churn? https://www.ngdata.com/what-is-customer-churn/

GeeksforGeeks 2020, Plot a pie chart in Python using Matplotlib https://www.geeksforgeeks.org/plot-a-pie-chart-in-python-using-matplotlib/

Brownlee, 2018, How to Use ROC Curves and Precision-Recall Curves for Classification in Python, https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/

Kaggle, n.d., Telco Customer Churn Focused customer retention programs, https://www.kaggle.com/blastchar/telco-customer-churn

Matplotlib Pie chart https://pythonspot.com/matplotlib-pie-chart/

matplotlib n.d, Barchart A bar plot with errorbars and height labels on individual bars., https://matplotlib.org/gallery/api/barchart.html

Sharma, N 2018, Ways to Detect and Remove the Outliers https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba

sklearn, n.d., sklearn.preprocessing.OneHotEncoder, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html