Title: Telco Customer Churn Analysis

Year:

1. **Introduction**

Telecommunication sector has become one of the main industries in developed countries. With the advance technical progress and the competitive environment, companies need to work hard to survive. There are few popular strategies have been used widely in order to generate more revenues such as 1) acquire new customers, 2) upsell existing customers and 3) increase retention period of customers. However, considering the value of return on investment (ROI), retaining existing customers has the highest potential to generate substantial revenue with lower costs and efforts as compared to acquiring new customer.

Therefore, it is important to develop a customer retention program by using machine learning to reduce the customer churn rate. Many researches confirmed that machine learning technology is highly efficient to predict this situation. This technique is applied through learning from previous data [1]. Several variables related to the user and their respective subscription are used to perform the prediction such as age, tenure, and monthly charges, to name a few.

1. **Objectives**

The objectives of this analysis are as follows:

* 1. Analyse the key traits between churn and non-customers based on customer behaviour data.
  2. Identify possible solutions to reduce customer churn rate and develop predictive model to predict customer churn.
  3. Estimate potential profit with the implementation of customer churn model.

1. **Dataset**

The telco churn customer data were collected from Kaggle [2]. The dataset contains 7043 records and 21 attributes. **Table 1** describes the description and type of attributes. There are 20 attributes that feature in telco customer prediction and one attribute serves as the output of the predicted attribute for the presence of customer churn.

1. **Methods**

In this project, RStudio was used to conduct the analysis and predictive modelling development because it supports open source innovation and availability. The project starts from the data understanding and pre-processing phase, followed by explanatory data analysis (EDA), to analyse the cause of customer churn and identify key features that affect customer churn, to create models for prediction based on different machine learning approach such as SVM, Decision Tree, Neural Network, Random Forest, and Logistic Regression. The output and performance of each model created will be evaluated and eventually to select the best performance predictive model.

**Table 1.** Description of attributes from Kaggle Dataset.

|  |  |  |
| --- | --- | --- |
| Attribute | Description | Type |
| Customer ID | Customer Identification |  |
| Gender | Gender of customer (Male, Female) | Nominal |
| Senior Citizen | Whether customer is a senior citizen (1 for yes, 0 for no) | Nominal |
| Partner | Whether customer has a partner (Yes, No) | Nominal |
| Dependents | Whether customer has dependents (Yes, No) | Nominal |
| Tenure | Number of months customer has stayed with the company | Numeric |
| Phone Service | Whether customer has a phone service (Yes, No) | Nominal |
| Multiple Lines | Whether customer has multiple lines described with 3 values;   1. No phone service 2. No 3. Yes | Nominal |
| Internet Service | Customer’s internet service provider described with 3 values;   1. No 2. DSL 3. Fiber optic | Nominal |
| Online Security | Whether customer has online security described with 3 values;   1. No 2. No internet service 3. Yes | Nominal |
| Online Backup | Whether customer has online backup described with 3 values;   1. No 2. No internet service 3. Yes | Nominal |
| Device Protection | Whether customer has device protection described with 3 values;   1. No 2. No internet service 3. Yes | Nominal |
| Tech Support | Whether customer has tech support described with 3 values;   1. No 2. No internet service 3. Yes | Nominal |
| Streaming TV | Whether customer has streaming TV described with 3 values;   1. No 2. No internet service 3. Yes | Nominal |
| Streaming Movies | Whether customer has streaming movies described with 3 values;   1. No 2. No internet service 3. Yes | Nominal |
| Contract | The contract term of customer described with 3 values;   1. Month-to-month 2. One year 3. Two year | Nominal |
| Paperless Billing | Whether customer has paperless billing (Yes, No) | Nominal |
| Payment Method | Whether customer has streaming movies described with 4 values;   1. Electronic check 2. Mailed check 3. Bank transfer (automatic) 4. Credit card (automatic) | Nominal |
| Monthly Charges | Amount charged to customer monthly | Numeric |
| Total Charges | Total amount charged to customer | Numeric |
| Churn | Whether customer churned or not (Yes, No) | Nominal |

* 1. **Data Pre-processing**

The data were pre-processed after collection. There were 11 records that have missing values in the dataset with attribute name: Total Charges and anomalies record in attribute name: Tenure with values 0. All the records that have missing values in total charges were replaced with the median value due to the distribution of data is positively skewed. The 0 value records in tenure are replaced with mean value as the distribution is more towards normal. Next, all the factor categorical variables are converted to factor numerical variables for the ease of model development. After the replacement and transformation, data is cleaned and used for analysis and modelling.

**4.1 Explanatory Data Analysis (EDA)**

**A screenshot of a social media post

Description automatically generated**

**Figure :** Customer tenure boxplot

From the graph, it can be observed that the first quartile of the distribution is at 10 months, the third quartile is around 55 months and the median is around 30 months. Hence, majority of the customer’s tenure duration lies within the interquartile range, which is about 45-month difference. The median is relatively at the centre of the interquartile range, which infers a mildly skewed tenure distribution, to the left.

**A screenshot of a social media post

Description automatically generated**

**Figure :** Monthly charges boxplot

From the graph, it can be observed that the first quartile of the distribution is around $35 per month, the third quartile is around $90 per month and the median is around $70 per month. Hence, majority of the customer’s tenure duration lies within the interquartile range, which is about 45-month difference. The median is relatively at the centre of the interquartile range, which infers a mildly skewed tenure distribution, to the left.

**A screenshot of a social media post

Description automatically generated**

**Figure :** Monthly charges boxplot

From the graph, it can be observed that the first quartile of the distribution is around $35 per month, the third quartile is around $90 per month and the median is around $70 per month. Hence, majority of the customer’s tenure duration lies within the interquartile range, which is about 45-month difference. The median is relatively at the centre of the interquartile range, which infers a mildly skewed tenure distribution, to the left.

**A close up of a logo

Description automatically generated**

**Figure :** Monthly charges density plot vs churn output

From the figure above, it was observed that the lower the monthly charges for a customer, the higher the likeliness for the customer to continue subscription. The reason behind this observation is that, lower monthly charges means lower monthly commitment for a customer, which would not burden them financially if they choose to continue subscribing to the telco service. On the other hand, when the monthly charges are higher, the total amount of charges accrued throughout the whole tenure duration would be of a higher magnitude. This sole reason can discourage users from continuing to subscribe to the telco plan, hence churning.

**A close up of a logo

Description automatically generated**

From the figure above, it was observed that low tenure duration resulted to a higher churn density. This phenomenon happened likely due to uncemented customer preferences, as many customers are still acclimating with the new telco experience, and they do not feel compelled to stay subscribing to the service just yet. Customers with longer tenure have a lower churn density, likely due to already being satisfied with the service and changing to a new telco provider would be a hassle to them.

**A close up of a logo

Description automatically generated**

The figure above portrayed an interesting observation pertaining to the total charges versus churn probability. The churn density surpasses the non-churn density. However, both churn and non-churn density are denser towards the lower value of total charges. This fits the boxplot visualization where most of the datapoints are closer to the minimum value. An explanation to this result is that most of the customers either have a combination of small monthly charges and long tenure duration or the opposite, large amount of monthly charges and short tenure duration. This inference is supported by observation made on Figure (monthly charges vs churn) where high monthly charges result to customer churning and vice versa.

References

1. <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6>
2. <https://www.kaggle.com/blastchar/telco-customer-churn>