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.

Therefore, it is important to develop a customer retention program by using machine learning to reduce the customer churn rate. Many research confirmed that machine learning technology is highly efficient to predict this situation. This technique is applied through learning from previous data. [1]

1. **Objectives**

Analyse the key traits between churn and non-customers based on customer behaviour data. Then, identify possible solutions to reduce customer churn rate and develop predictive model to predict customer churn. Eventually, estimate on the 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)**

References

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