A picture containing logo

Description automatically generated

Fall 2020 IST 718 Project Proposal

**Telecom Customer Churn Analysis**

Group 13 | Tian Jiang, Wanyue Xiao, Yiyuan Cheng

**Objective**

Customer churn, also known as customer attrition, is a measure of the loss of individuals switching from one collective group to another during a specific period. As a measurement of damage inflicted, it is gaining more attention from big companies due to its direct effect on revenues. Therefore, it is necessary for a company to find out factors that give rise to customer churn. The purpose of this project is to analyze all relevant customer data, identify the critical features that correlate with churn rate, construct customer churn prediction model, and finally provide business insights that prevent customers from leaving.

**Data Set Description**

The Telecom Customer Churn Dataset containing customers personal and phone plans is provided by the telecom company. Specifically, there are 7043 rows and 21 columns in the dataset, where each row represents a customer, and each column is the customer's attribute. The target variable, *churn*, indicates whether a customer left within a month. The predictors are divided to three groups:

* Customer private information includes *gender*, *seniorCitizen*, *partner*, and *dependents*. Particularly, *seniorCitizen* indicates whether a customer is a senior citizen.
* Customer account information includes *tenure*, *contract*, *MonthlyCharge*, *TotalCharges*, etc. For instance, *tenure* is a numeric variable which represents the number of months that a customer sticks to the company.
* Services available, including phone, multiple lines, online security, online backup, etc. Each service attribute possibly has a value of “Yes/No” denoting whether the customer has signed up for this service or not, or “No Service”, which stands for no such service provided.

The original data can be retrieved from: <https://www.kaggle.com/blastchar/telco-customer-churn>.

Something obvious is that the number of categorical variables overweighs that of numeric variables. In addition, it is worthwhile to mention that the total charges per customer varied significantly, ranging from 18.8 to 8684.8. Apart from that, it is also interesting that the dataset uses *seniorCitizen* to represent age instead of using numbers or age groups.

**Preliminary Data Exploration**

*Data Preprocessing*

After checking the dataset, 11 missing values in total are identified and all of them are located in *TotalCharges*. They are replaced by 0 since they result from customers not paying bills. Three attributes, *tenure*, *MonthlyCharges*, and *TotalCharges*, are transformed from string type to numeric for further analysis.

*Summary Statistics of Numeric Data*

Table

Description automatically generated

*Summary Statistics of Categorical Data*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Attribute** | **Value** | **Count** | **Attribute** | **Value** | **Count** | **Attribute** | **Value** | **Count** |
| **gender** | female | 3488 | **Senior-Citizen** | 0 | 5901 | **Partner** | No | 4933 |
| male | 3555 | 1 | 1142 | Yes | 2110 |
| **Partner** | No | 4933 | **Dependents** | No | 4933 | **Phone-Service** | No | 682 |
| Yes | 2110 | Yes | 2110 | Yes | 6361 |
| **Multiple-Lines** | No Service | 682 | **Internet-Service** | Fiber optic | 3096 | **Device-Protection** | No | 3095 |
| No | 3390 | No | 1526 | Yes | 2422 |
| Yes | 2971 | DSL | 2421 | No Service | 1526 |
| **Online-Security** | No | 3498 | **Online-Backup** | No | 3088 | **Tech-Support** | No | 3473 |
| Yes | 2019 | Yes | 2429 | Yes | 2044 |
| No Service | 1526 | No Service | 1526 | No Service | 1526 |
| **Multiple-Lines** | No Service | 682 | **Internet-Service** | Fiber optic | 3096 | **Device-Protection** | No | 3095 |
| No | 3390 | No | 1526 | Yes | 2422 |
| Yes | 2971 | DSL | 2421 | No Service | 1526 |
| **Streaming-TV** | No | 2810 | **Contract** | No | 3875 | **Payment-Method** | Credit card (automatic) | 1522 |
| Yes | 2707 | Yes | 1473 | Mailed check | 1612 |
| No Service | 1526 | No Service | 1526 | Bank transfer | 1544 |
| **Paperless-Billing** | No | 2872 |  | | | Electronic check | 2365 |
| Yes | 4171 |

*Categorical Variables vs. Churn*

The plots below show the distributions of churn rate based on different conditions. For customers with dependents, the churn ratio is a little bit lower. Additionally, with a longer contract (more than one year) or additional services like tech support or internet service, the customers are less likely to churn.

Chart, bar chart

Description automatically generated

*Numeric Variables vs. Churn*

The plots below show the distributions of the three numeric attributes by different churn types. It could be estimated that the less time the customers use the telecom companies’ service, the more likely they churn. Lower monthly charges would help reduce churn. Total charges might have limited influence on customer churn since the trend of customers who churn and don’t churn are similar.

A picture containing histogram

Description automatically generated

**Predictions**

Predict customer leaving given the information about customer private information, account information, and the services they signed up for by:

* Building machine learning models: K-Means Clustering, Logistic Regression, Linear Support Vector Machine, Random Forest model, and Gradient Boosting model
* Comparing model performance and selecting the most reliable prediction model

**Inference**

* How is the customer churn influenced by the number of services registered?
* How does the account information including tenure and monthly charges indicate customer churn?
* How do personal characteristics such as gender and dependents affect customer churn?
* What are the most important features in predicting customer leaving (churn = yes)?
* What are the most important features indicating a loyal customer (churn = no)?
* Is a linear model enough to capture the relationship?

Based on the inference output, suggest business insights to reduce customer churn:

* What services should the telecom company consider improving to diminish customer’s willingness to leave?
* Which group of customers should be targeted to enhance their loyalty?

**Non-Spark Packages**

Temporarily none other than spark, numpy, pandas, matplotlib, and seaborn.