Statistical Methods for Data Science

(DS 502/MA 543)

Group 5 Project Proposal

Telco Customer Churn Prediction

Team Members:



Vandana Anand **Kratika Agrawal** **Jiaju Shi**

Introduction

With the increasingly fierce competition in the telecommunications industry, retaining customers becomes a key business indicator for better operation and competitiveness of the company. This key point requires understanding the characteristics of the customer that stopped using the company's services, analyzing the reasons for the customer leaving, predicting whether customers will leave, determining the retention of target users, and making effective plans to retain those target users.

The data are from a **Kaggle dataset** that comes from the IBM database. IBM is an American multinational technology and consulting company headquartered in Armonk, New York. Their primary business model is to produce and sell computer hardware and software as well as supply consulting services for system architecture and network hosting.

The Telecom Customer Churn dataset from IBM is used to determine the behavior of the employer to show the major causes of telecom customer churn and develop a targeted customer retention plan. A customer could churn due to many factors that could include anything from their own demographics to which services they have signed up to use. Our project focuses on pinpointing these factors to see which ones cause customers to churn.

Project Summary

To determine factors that cause customers to churn, the first step we will take is to analyze the dataset. We will look at the type of data available and decide what the variables mean. We will then look at initial visualizations to see if there are any interesting trends or inconsistencies. Next, we will preprocess the data and perform some data cleaning, such as missing values, outliers, or redundancy, to work with the data. We will perform feature engineering to determine which features in the dataset to work with and input in our potential model. Finally, we will look at data resampling methods, such as how to partition our data into training and testing to obtain fair results.

After the statistical data preprocessing step is over, we will start to pick proper classification models to apply to our dataset. Then, we will construct our training model and evaluate it based on its performance. We will then apply our model to our test data to obtain our prediction. We can repeat this process depending on how many models we want to compare. At the end, we will use our evaluation method to compare predictions from each model and determine the best one for our dataset. From this, we can produce features that affect customer churn the most and suggest a retention plan for the company.

We plan to divide the technical work described above among our team members, who will also work on the final paper filling in their parts simultaneously. The following sections serve to highlight the details of the team's plan to work on this project and obtain exact results for the Telecom Customer Churn problem.

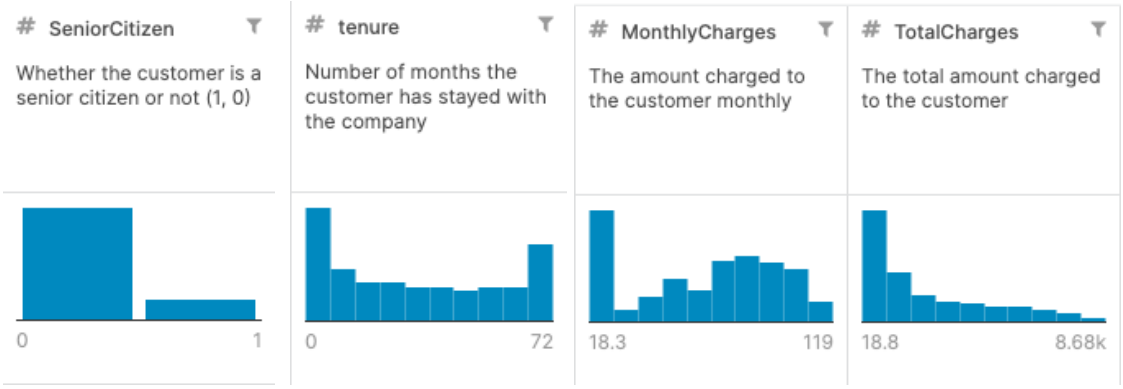
Statistical Data Pre-processing

The original data needs to be processed by handling inconsistent values and cutting or replacing the abnormal part of the data so that the input data becomes normal and reliable. Sampling, analyzing, pre-processing the existing data, and then modeling and evaluating the data can effectively reduce the problem of potential customer loss.

Descriptive Statistics

- The entire data size is about 955MB that consists of 21 columns and 7044 rows.
- Each row stands for a customer and each column holds customer's attributes described on the column metadata.
- Customers who left within the last month – the column called Churn (Target Column).
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they have been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

These are some of the first data visualizations we used to get a gist of the data:

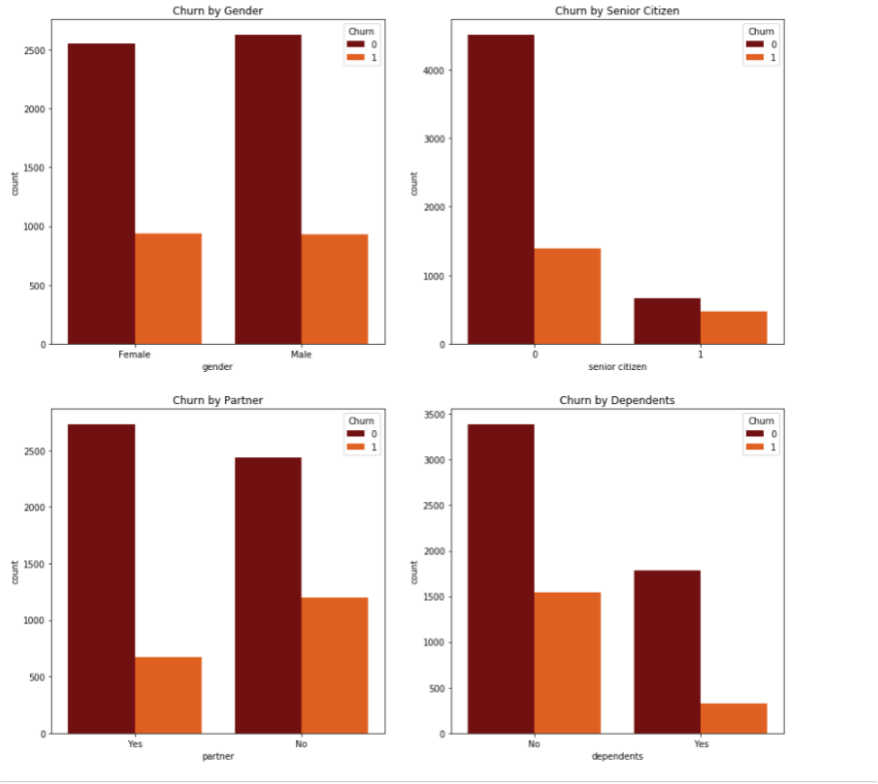


As we can see, most of the customers in the dataset are not senior citizens (marked with 0). The tenure aspect is evenly distributed and the same for the monthly charges. The total charges feature shows the histogram having a left tail and being skewed, meaning that any given customer can charge depending on the plans and services they chose.

Data Cleaning

We will first focus on cleaning and preprocessing the data so that the dataset is ready to apply the classification models. We will make sure to do the following:

- Remove duplicate records
- Handle missing data
- Manage extreme values or outliers
- Check and convert data to proper data type
- Apply data normalization
- Initial data visualizations



The data above shows the gaps and inconsistencies in the dataset. Some categories have more data than others.

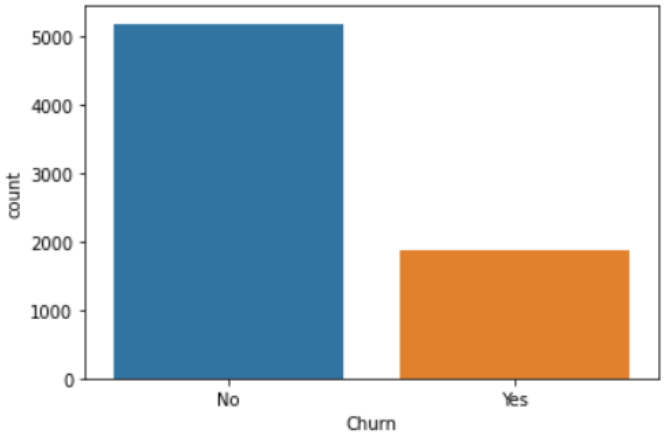
Feature Engineering

The dataset has a combination of numerical and categorical features. For any statistical method to evaluate the data, we would need to transform categorical features to a numerical one using dummy variables. For example, we will assign a column with categorical data with a consecutive number and then replace the data based on that assigned number.

Next, we will perform feature choice techniques to extract key features for evaluation. For instance, this can be based on different services the customer signed up for or demographic information to see which areas or types of customers would churn.

Data Resampling

The dataset at hand is unbalanced with a 3:1 distribution ratio of records for two classes:



Thus, we need to over-sample or under-sample the data to achieve a balance.

We will use resampling methods, such as k-fold Cross-validation, to split the data into training and testing datasets while applying classification algorithms to prevent model overfitting.

Modeling Techniques

General steps

- Setup training and testing data set
- Select classifier algorithm
- Construct training model
- Evaluate training model
- Develop and apply to testing model

Model Selection

We will explore various classification techniques and select the hyperparameters for better accuracy in predicting what factors will lead to customer loss. To process and analyze our data sets, we plan to use:

- Logistic Regression

- K-Nearest-Neighbors
- QDA
- Random forests
- Naïve Bayes

We will then compare the predictions obtained from each to find the best data model. After this process, we will extract the impact of various attributes over customer churn.

Model Evaluation

Since the dataset is unbalanced, we will evaluate the result using the Confusion Matrix and calculate the sensitivity and specificity of our result. Then, we will compare the performance of the various algorithms using the AUC score and ROC curve.

Conclusion

The goal is to apply our classroom learning to practical data which is in a diverse form. We would like to implement various concepts learnt to extract the most important features and best fit a model to make predictions of customer churn.

As mentioned, we plan to divide various tasks evenly among all team members, be open to reasonable suggestions for the modification, and collaboratively work towards the success of our project.