Springboard Data Science Capstone Project

Churn Analysis of a Telecommunication Company

1.Introduction

Churn rate is the percentage of service subscribers who discontinue their subscriptions within a given time period. For a business whose revenues are based on subscription, understanding the churn rate of its customers is very important for maintaining a steady income flow. Through churn analysis, a company will be able to not only predict the likelihood of churn of a customer, but also to develop a customer retention strategy to decrease churn rate and to increase sales revenues.

Customers whose business models are based on subscription, such as telecommunication companies, newspapers, online entertainment providers, etc. would be interested in churn analysis. In this project, I will conduct a churn analysis by using the data from a telecommunication company.

2. Data Acquisition and Cleaning

The dataset is acquired from Kaggle (the source of the dataset is not specified in Kaggle). It contains the demographic information of the customers, what kind of services they subscribed, their monthly payments, their tenures, etc. The dataset, which is a csv file, contains 21 variables and 7043 observations.

First of all, I loaded it into Pandas DataFrame by using pd.read\_csv syntax. Next, I printed out the first couple rows of the data, to get a sense of the dataset. To explore more information about the dataset, I used dataframe.info to investigate the column information, data type and whether there are missing values. In total, there are 7043 entries and 21 columns in the dataset. Because every column has 7043 entries, I couldn’t tell whether there are missing values or not at this stage. To further explore the dataset, I did value count column by column for the columns whose data type are categorical, and I calculated the statistics (such as min, max, mean, etc.) and plotted histogram for the columns whose data type is integer or float. To summarize, I did not find any outlier in any column, but I did find that a column called “TotalCharge” has 11 missing values. And because of that, the data type of this column, which was supposed to be “float”, became “string” instead.

Subsequentially, I converted the data type of “TotalCharge” column into inter64 by using pd.to\_numeric syntax. Moreover, since many columns in the dataset are categorical, for further analysis, I have to convert them into dummy valuables. Finally, I used assert syntax to confirm that there is no other missing value in the dataset. Since the result is true, everything looks fine at this stage.

3. Data Exploration

By creating a correlation heatmap with all the features in the dataset, I found out that

some of the features are strongly correlated. Specifically, OnlineSecurity\_No internet

service, OnlineBackup\_No internet service, DeviceProtection\_No internet service,

TechSupport\_No internet service, StreamingTV\_No internet service, and

StreamingMovies\_No internet service are all strongly correlated to InternetService\_No.

The correlation coefficient is 1. This is quite self-explanatory because if a customer does

not subscribe internet service, he/she would not have online security, online backup,

device protection, tech support, streaming TV, streaming movies consequentially. To

avoid these collinearities, I removed all the above features (only keep

InternetService\_No).

Moreover, the feature of PhoneService is negatively correlated with MultipleLines\_No

phone service This is also quite self explanatory: a customer who does not subscribe

phone service would not have multiple lines consequentially. To solve this problem, I

removed the feature of MultipleLines\_No phone service.

Furthermore, the feature of TotalCharges is strongly correlated to the feature of Tenure.

This is also logical since the customers who stayed longer with the company would pay

more in total. Similar to above solution, I removed the feature of TotalCharges. In addition,

the feature of InternetService\_Fiber optic is strongly correlated with the feature of

MonthlyCharges, which suggests that whether the Internet is fiber or not would have

strong impact on monthly charges. In this case, I removed the feature of

InternetService\_Fiber to avoid collinearity.

Upon finishing above actions, I re-plot the correlation heatmap and now it looks that there

is no strong correlation between any two features.

Exhibit 1: Features Correlation Heatmap Before Deleting Correlated Features

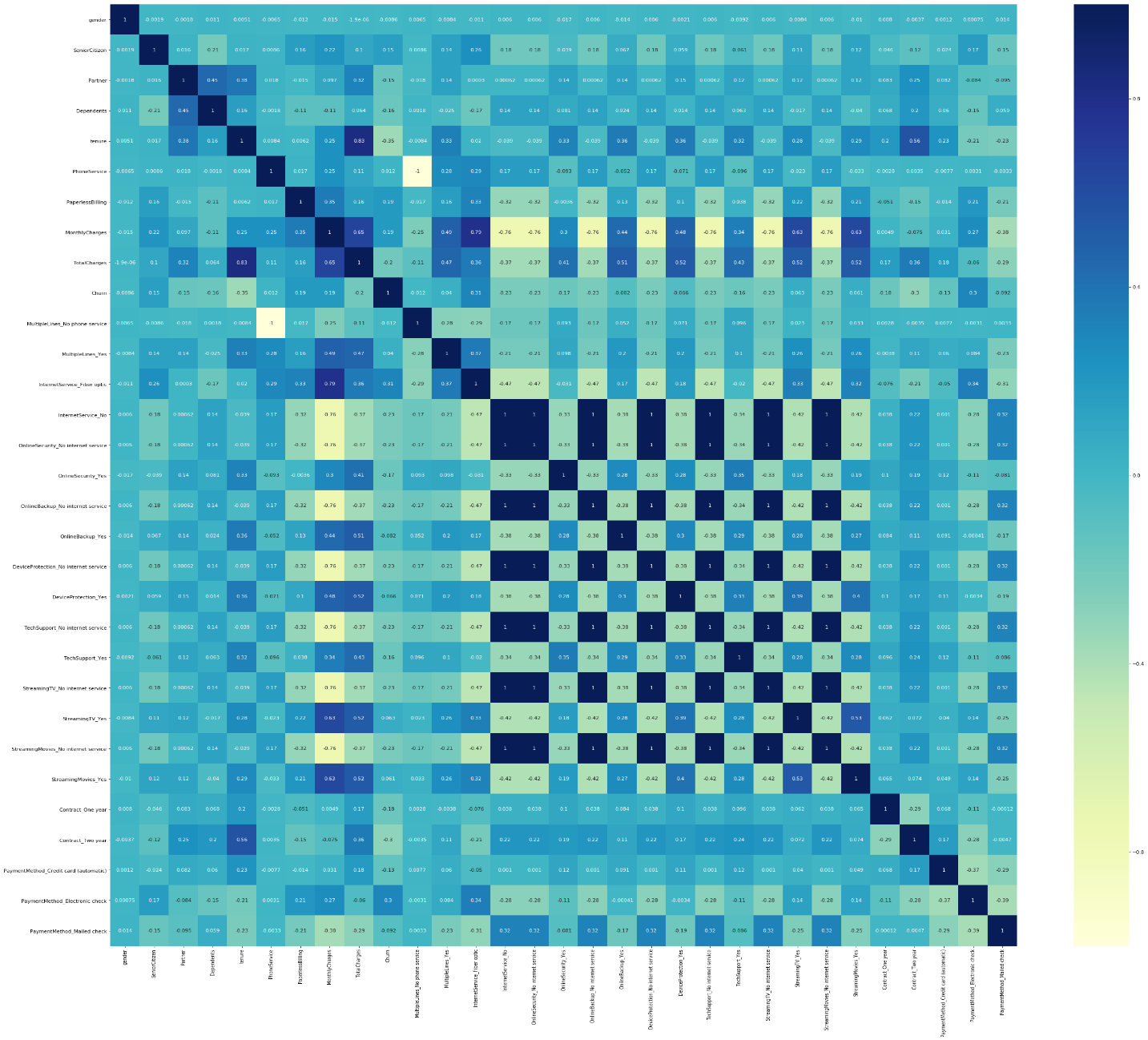


Exhibit 2: Features Correlation Heatmap After Deleting Correlated Features

