Survival Analysis on Telco Customer Churn

Problem statement

Customer churn, defined as the percentage of customers that stop using a company's products or services, is one of the most import matrices for a business, as it usually costs more to acquire new customers than it does to retain existing ones.

The objective of this project is to predict if customers are likely to stop doing business and when that event might happen.

The client is the management of the pertaining business. With the model built and data analysis performed, the business owners will be able to segment the customers based on the likelihood to churn and take appropriate actions to prevent that happening, such us launching customer retention program.

One of the key attributes, tenure, indicates that the time period that a customer has been with the business since the contract was signed. Since each customer has different time footings, the traditional linear machine learning algorithms don't perform well in this scenario. This is why survival analysis comes to play as the time-event correlation will be considered.

To perform survival analysis, the non-parametric method Kaplan–Meier estimator can provide a descriptive approach to consider how "tenure" affects the likelihood of customer churn by plotting survival function curve. However, semi-parametric method Cox Proportional Hazard is more powerful to predict whether a customer churn based on not only time factor but also other attributes, which can be seen from the dataset. Furthermore, some advanced techniques such as survival tree and survival random forest will also be explored.

Dataset acquisition

The dataset is for Telco Customer Churn, downloaded from Kaggle competition website. This dataset contains 7,043 unique observations in total, each of which represents a customer. Each column contains customer's 21 attributes.

The data set includes information about:

- Customers who left within the last month – the column is called Churn
- 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've 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

Data wrangling strategy

Data wrangling is the terminology used to describe the process of transforming raw data to a clean and organized format ready for use. The most common data structure used to "wrangle" data is the data frame. For this particular project, we imported the data from CSV file by using Pandas. We will take the following steps to perform data wrangling.

- Describing the data

The first step is to view some characteristics of a DataFrame. One of the easiest things we can do after loading the data is to view the first few rows using head(). We can also take a look at the number of rows and column using .shape attribute of the DataFrame. By using .info attribute or .dtypes attribute, we are able to view data type of each variables.

For any numerical variables, we can use .describe() method to get descriptive statistics. Pandas offers variance (var), standard deviation (std), kurtosis (kurt), skewness (skew), standard error of the mean (sem), mode (mode), median (median), and a number of others. Pandas also provides a large set of summary functions that operate on different kinds of pandas objects (DataFrame columns, Series, GroupBy) and produce single values for each of the groups. When applied to a DataFrame, the result is returned as a pandas Series for each column. For any categorical variables, both .unique and .value_counts are useful for manipulating and exploring the data.

- Dropping duplicate rows

The method .drop_duplicate can help perform this task.

We have determined the date type of each variables. For numerical features, the scaling matters for some machine learning algorithms. By checking the descriptive statistics, we can decide whether the data standardization is needed.

To detect missing data, the functions isnull and notnull return booleans indicating whether a value is missing. If missing data exist, we need to figure out its size and randomness. If the size is minimal and not highly dependent on the target variable, we can probably remove them. Otherwise, we should apply different imputation strategy such as mean or median. To detect outliers, a box-and-whisker or histogram plot helps in this regard. Numpy .percentile() command is also an effective tool to identify the extreme observations and define the upper and lower boundaries. The interquartile range (IQR) method can be used to fix outliers.

Once we have detected the missing values for categorical features, we can apply frequent category imputation strategy, replacing all occurrences of missing values with the most frequent value. Alternatively, we can treat missing data as an additional label or category of the variable so that the importance of 'missingness' can be captured. By using .value_counts() and unique() methods, we can find out the cardinality of a categorical variable and rare labels if possible. Both high cardinality and rare labels may cause overfitting. Removing them can help improve machine learning performance.

Exploratory Data Analysis (EDA)

The first question is what is the distribution of those three numerical variables and how would you explain them?

The distribution of total charges is very positively skewed. The majority of customers are charged under $2000. The distribution of tenure shows a relatively stable trend between 10 months and 60 months. However, significant number of customers stayed with business below 10 months (new customers), or above 60 months (loyal customers).

As for the distribution of monthly charges, above $30 and beyond, it distributes nearly normal. However low-end customers with less than $30 accounts for a large portion.

Another question is that what is the distribution of those three numerical variables? How would you explain them?

From a business perspective, we ignore the first feature customerID since its value has nothing to do with customers churn. For all other categorical features, the largest number of different labels is 4. The cardinality is not high, so that it is not likely cause overfitting. As for the relatively rare labels, we found that 9% for 'NO' values of Phoneservice, and 9% for 'No phone service' of MultipleLines. This is also unlikely to cause overfitting.

Statistical Data Analysis

After we completed data wrangling and data visualization, it is time to use statistical inference tool to explore the relationship between independent variables and dependent variable. For this particular project, we are going to perform survival analysis to predict telecom customers churn. In this context, “tenure” and “churn” will be treated as time-event variables. During the process of data wrangling, we found this dataset has 16 categorical variables and 3 numerical variables that can be used to build model.

First of all, we divide all observations into two groups using ‘churn’, i.e. churned and non-churned. We are interested to know how the numerical variables’ mean differ between these two subgroups? How would we explain them statistically? The result of t-test indicates that statistically significance does exist since p-values are all extremely low. We can confidently reject the null hypothesis and conclude that all these numerical variables differ between two groups. Please also be noted that these three variables were considered individually. In survival analysis, we should treat variable 'tenure' as time factor that may bring censoring issue. This will be addressed in model building and model performance evaluation that follows.

Since the majority of the features are categorical, after we had an idea about cardinality, we are more interested in how each labeled group affect the

performance of time-event in the context of survival analysis. Rather than using traditional chi-square contingence table, we use Log rank test together with Kaplan-Meier curve to figure out how those groups differ from each other. The analysis indicates that out the 16 categorical variables, only "gender" and 'Phone Service" have large value of p-values. We don't have sufficient evidence to reject null hypothesis for them. The statics of other variables is statistically significant. They are more important to predict the customer churn.