Time series Analysis & Modeling

DATS 6450

LAB # 11- Survival Analysis

The main purpose of this LAB is to estimate and plot the survival function for real data set using “lifelines” library in Python. The library which needs to be used for this lab are as follow:

```
from lifelines import KaplanMeierFitter
import pandas as pd
import matplotlib.pyplot as plt
```

1- Using Pandas library .read_csv function load the “WA_Fn-UseC_-Telco-Customer-Churn.csv”, as a Dataframe call it df.

2- Plot the first few rows of data set to get a feeling about the dataset. This can be done using df.head()

3- Get more information about Dataframe, i.e. data type and missing values using df.info().

4- Convert “Total Charges” to numeric using the following function:

```
df['TotalCharges']=pd.to_numeric(df['TotalCharges'],errors='coerce')
```

5- Replace yes and no in the churn column to 1 and 0. This can be done as follows:

```
df['Churn']=df['Churn'].apply(lambda x: 1 if x == 'Yes' else 0 )
```

6- Impute the null value of total charges with the median value using the following function:

```
df.TotalCharges.fillna(value=df['TotalCharges'].median(),inplace=True)
```

7- Create an overall Kaplan Meier curve, without breaking it into groups of covariates (groups will be created in the future steps). For this purpose, you need to create Time to event of censored and event data. You also need to create event observed data for customer who has churned (1) and censored (0). This can be done as follows:

```
durations = df['tenure']
event_observed = df['Churn']
```

8- Create a kmf object as km

```
km = KaplanMeierFitter()
```

9- Fit the data into the model

```
km.fit(durations, event_observed,label='Customer Retention')
```

10- Plot the estimated survival curve using:

```
km.plot()
```

11- Interpret the plot created in the previous step.

12- Create Kalan Meier curves for three cohorts:

```
kmf = KaplanMeierFitter()

T = df['tenure']      ## time to event
```

```
E = df['Churn']       ## event occurred or censored


groups = df['Contract']               ## Create the cohorts from the 'Contract' column
ix1 = (groups == 'Month-to-month')    ## Cohort 1
ix2 = (groups == 'Two year')          ## Cohort 2
ix3 = (groups == 'One year')          ## Cohort 3
```

13- Fit the cohort 1, 2 and 3 data and plot the survival curve using the following commands:

```
kmf.fit(T[ix1], E[ix1], label='Month-to-month')
ax = kmf.plot()


kmf.fit(T[ix2], E[ix2], label='Two year')
ax1 = kmf.plot(ax=ax)


kmf.fit(T[ix3], E[ix3], label='One year'
kmf.plot(ax=ax1)
```

14- Interpret the plot created in the previous step. How does the length of contract affect retention?

15- Add the appropriate legend and title to the graph created in the previous step.

16- Define two new cohorts based whether a subscriber “StreamingTV” or not “StreamingTV”. We would like to know how the streaming TV option affect retention. You can create the cohorts as follow:

```
kmf1 = KaplanMeierFitter()


groups = df['StreamingTV']
i1 = (groups == 'No')
i2 = (groups == 'Yes')
```

17- Repeat the procedures in step 13 to fit the cohorts created in the previous step and plot the estimated survival curve. Make sure to assign the correct labels.

18- Interpret the plot created in the previous step. How is the streaming TV affect retention?

Upload the **solution <u>report (as a single pdf</u>**) plus **<u>the .py file(s)</u>** through BB by the due date.