

**TIME SERIES MODELING & ANALYSIS**

**Instructor Name:** Reza Jafari  
**Lab#:** 11  
**Submitted by** Dinesh Kumar Padmanabhan  
**Date:** 03-Dec-2020

# ABSTRACT

The main purpose of this LAB is to estimate and plot the survival function for real data set using “lifelines”

library in Python. Using survival analysis we can study the fundamental questions like How long will

particular customer remains with your business? How long will this machine last, after successfully

running for a year? What is the relative retention rate of different marketing channels? What is the

likelihood that a patient will survive after being diagnosed ? All of above questions can studied under

Survival Analysis topic.

# INTRODUCTION

**Theory**

Survival analysis is a set of statistical tools which answer the following question: How long would it

be, before a particular event occurs or a particular event occurs? In other words we can also call it as

a ”time to event analysis”. This method is called survival analysis because it was developed by

medical researches and they were more interested in finding expected lifetime of patients in

different cohort. The method can be further applied to not just traditional death events, but to many

different types of events of interest in different businesses domains like

Predictive maintenance

Customer Analytics:(customer retention)

Marketing Analytics:(Cohort Analysis)

**Mathematical Intuition**

**Survival function:**

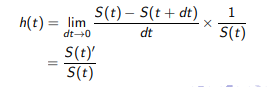
S(t) = 1 − F(t) = P(T ≥ t)

S(t) gives us the probability that the event has not occurred by the time t.

In simple words, S(t) gives us the proportion of population with the time to event value more than t.  


Along with the survival function, we are also interested in the rate at which event is taking place, out of the surviving population at any given t.

**Hazard Function: h(t):**



# METHOD, THEORY & PROCEDURES

**Method:**

1. Programming Language: Python

*Libraries used:* Some basic libraries used for analysis & model building are mentioned below *library(Numpy)* - large collection of high-level mathematical functions to operate on these arrays. *library (Pandas)* – For Data manipulation and analysis  
*library(Matplotlib)* – is a system for declaratively creating graphics  
*library(Math) –To Compute mathematical calculations*  
library (statsmodels) – Import statistical models  
library (scipy) – Scientific Computations  
library lifelines – statistical computations

***Theory***:

To estimate and plot the survival function for real data set.

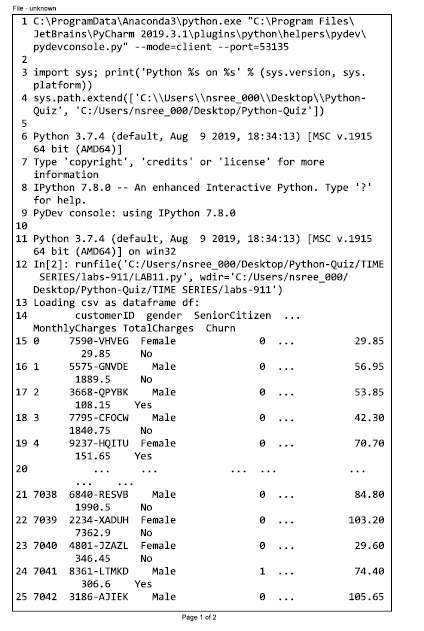
**Procedure:**

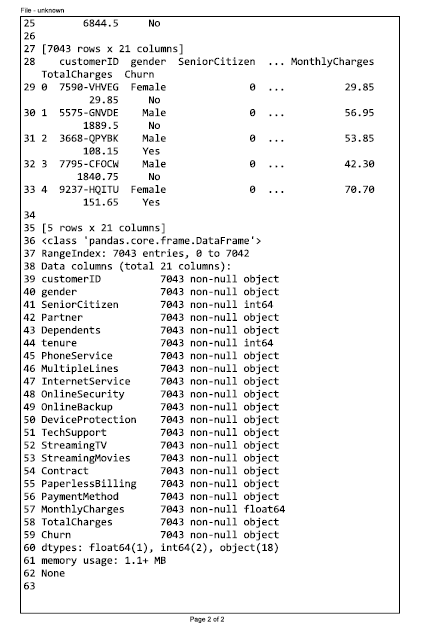
I shall be looking at the results for WA\_Fn-UseC\_-Telco-Customer-Churn.csv using survival function. Perform various plots and infer about it in my analysis. And through my exploration I shall try to identify which methods perform better and draw inferences.

The Dataset will be explored in following stages:

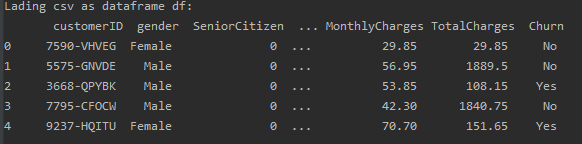
1. **Data Exploration (EDA)** – looking at the models and making inferences about the data.
2. **Data Visualization** – Plotting different time series plots for the regression method and forecast accuracy.
3. **Testing** – Running Autocorrelation, Pearson correlation test to identify the correlation between errors.

**ANSWERS TO QUESTIONS**

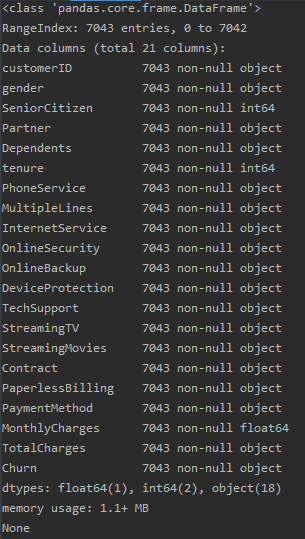




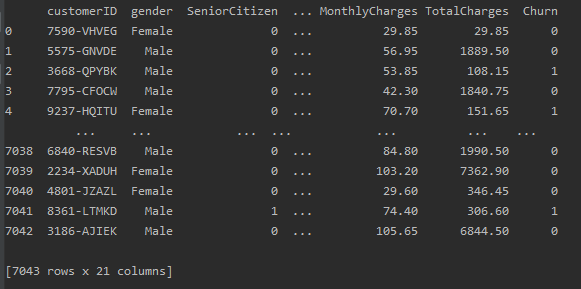
1. Using Pandas library .read\_csv function load the “WA\_Fn-UseC\_-Telco-Customer-Churn.csv”, as a Dataframe call it df.
2. Plot the first few rows of data set to get a feeling about the dataset. This can be done using df.head()



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

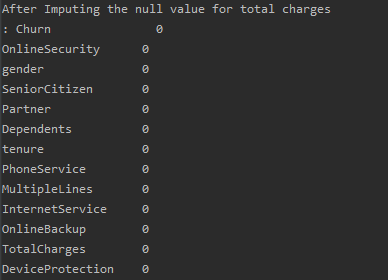


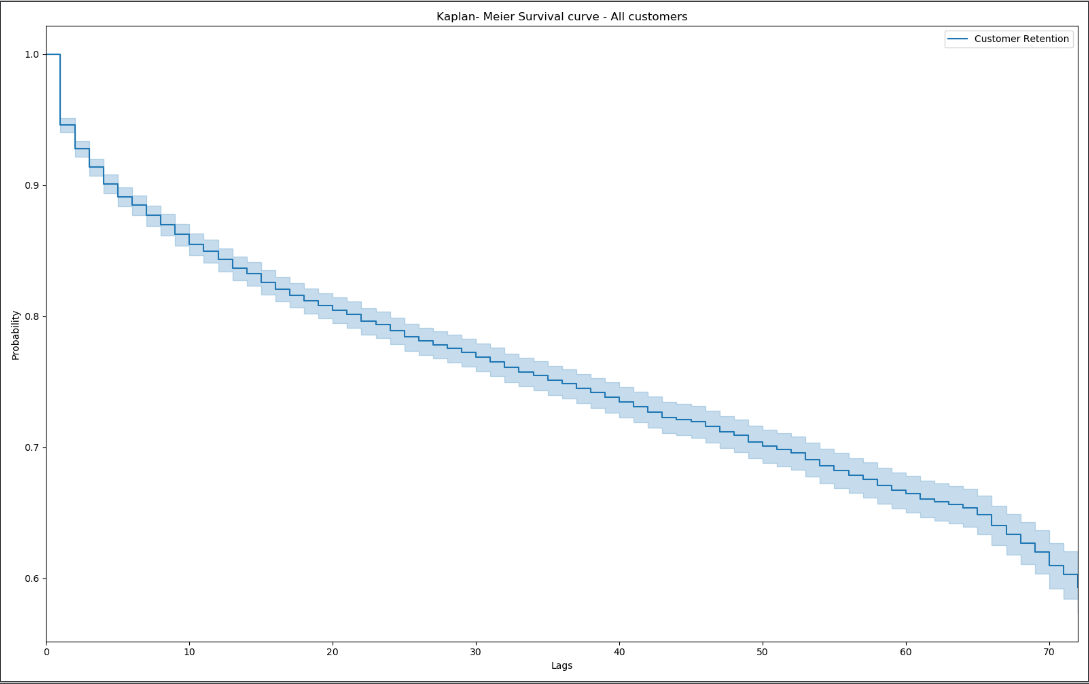
1. Convert “Total Charges” to numeric using the following function:
2. Replace yes and no in the churn column to 1 and 0. This can be done as follows:



1. Impute the null value of total charges with the median value using the following function:



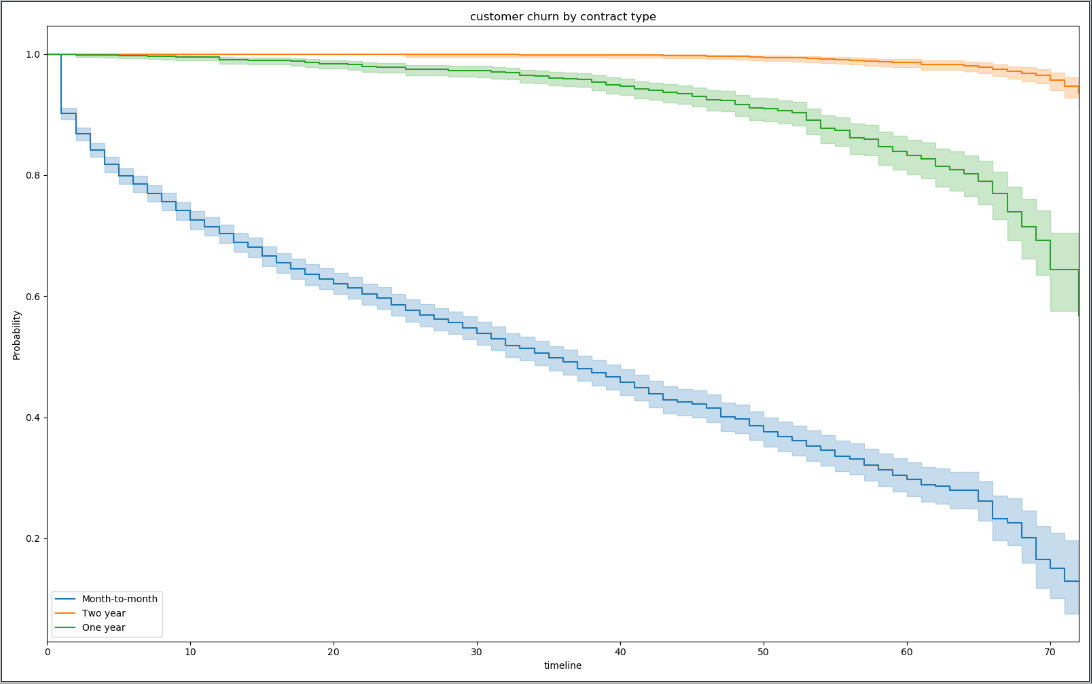


10. Plot the estimated survival curve using: 

11. Interpret the plot created in the previous step.

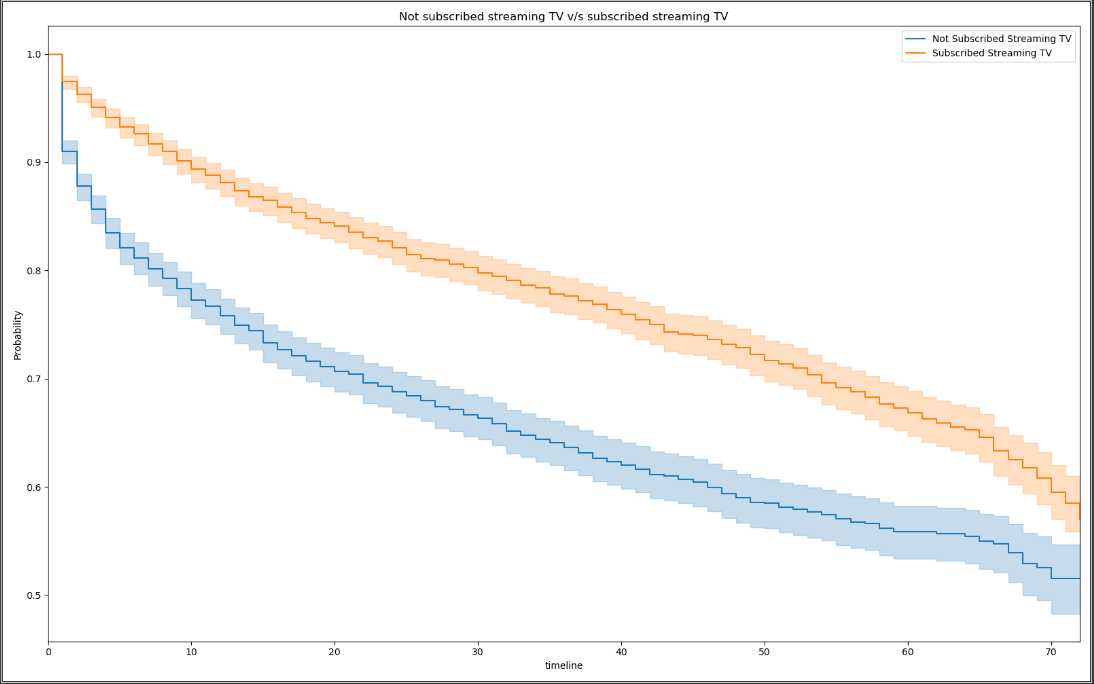
The above should give us some basic intuition about the customers. As we would expect for telecom, churn is relatively low. Even after 72 months, the company is able to retain 60% or more of their customers.

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



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

The above should give us some basic intuition about the customer churns by contract type. churn is relatively low for month-month type the company is able to retain about less than 0.2%probability of their customers. The most important feature, by far, is the presence of a 1 or 2 year contract. Customers are .65 and .9, respectively, times as likely to retain their service if they are in 1 and 2 year contract.

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?

The above should give us some basic intuition about the customers churn who are subscribed to streaming TV and with those who are not subscribed to streaming TV. It looks like the rate is relatively low for those customers that are not subscribed streaming TV vs to those with the subscription.

CONCLUSION

Thus it was estimated and ploted the survival function for real data set using “lifelines” library in Python.

Using survival analysis studied the fundamental question How can our telecom company reduce

customer churn? We can make recommendations along three dimensions: contract specification,

customer selection, and payment systems. To visualize some of our findings, we will fit categorically

based Kaplan-Meier curves and plot them, allowing us to see difference in churn rate between customer

categories.

CHALLENGE

None

APPENDIX

from lifelines import KaplanMeierFitter

import pandas as pd

import lifelines

import matplotlib.pyplot as plt

from pandas.plotting import register\_matplotlib\_converters

import warnings

warnings.filterwarnings("ignore")

register\_matplotlib\_converters()

#%%==================================================================================

# 1. Using Pandas library .read\_csv function load the “WA\_Fn-UseC\_-Telco-Customer-Churn.csv”, as

# a Dataframe call it df

# %%------------------------------------------------------------------------------------------------------------

df = pd.read\_csv('WA\_Fn-UseC\_-Telco-Customer-Churn.csv')

print("Loading csv as dataframe df:\n", df)

#%%==================================================================================

# 2. Plot the first few rows of data set to get a feeling about the dataset. This can be done using

# df.head()

# %%------------------------------------------------------------------------------------------------------------

print(df.head(5))

#%%==================================================================================

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

# %%------------------------------------------------------------------------------------------------------------

print(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 )

# print(df)

#%%==================================================================================

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

# %%------------------------------------------------------------------------------------------------------------

# df = df.isnull().sum().sort\_values(ascending = False)

# print('Before Imputing the null value for total charges\n:', df)

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

# df = df.isnull().sum().sort\_values(ascending = False)

# print('After Imputing the null value for total charges\n:', df)

#%%==================================================================================

# 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')

# %%------------------------------------------------------------------------------------------------------------

km.fit(durations, event\_observed,label='Customer Retention')

#%%==================================================================================

# 10. Plot the estimated survival curve using:

# %%------------------------------------------------------------------------------------------------------------

fig, ax = plt.subplots(figsize=(16,10))

km.plot()

plt.ylabel('Probability')

plt.title("Kaplan- Meier Survival curve - All customers")

plt.show()

#%%==================================================================================

# 11. Interpret the plot created in the previous step.

# %%------------------------------------------------------------------------------------------------------------

''' The above should give us some basic intuition about the customers.

As we would expect for telecom, churn is relatively low. Even after 72 months, the company is able to retain 60% or

more of their customers.

'''

#%%==================================================================================

# 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:

# %%------------------------------------------------------------------------------------------------------------

fig, ax = plt.subplots(figsize=(16,10))

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)

plt.ylabel('Probability')

plt.title("customer churn by contract type")

plt.show()

#%%==================================================================================

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

# retention?

# %%------------------------------------------------------------------------------------------------------------

'''

The above should give us some basic intuition about the customer churns by contract type.

churn is relatively low for month-month type the company is able to retain 18% of their customers.

The most important feature, by far, is the presence of a 1 or 2 year contract. Customers are .25 and .02, respectively,

times as likely to cancel their service if they are under contract. Cancellation fees are a possible underlying cause.

As long as these fees do not prohibit new sales, we would recommend continuing to put them into as many contracts as

possible.

'''

#%%==================================================================================

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

# %%------------------------------------------------------------------------------------------------------------

'''  
Added titles and legends

'''

#%%==================================================================================

# 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.

# %%------------------------------------------------------------------------------------------------------------

fig, ax = plt.subplots(figsize=(16,10))

kmf1.fit(T[i1], E[i1], label='Not Subscribed Streaming TV')

ax = kmf1.plot()

kmf1.fit(T[i2], E[i2], label='Subscribed Streaming TV')

ax1 = kmf1.plot(ax=ax)

plt.ylabel('Probability')

plt.title("Not subscribed streaming TV v/s subscribed streaming TV")

# plt.figure(figsize=(16,10))

plt.show()

#%%==================================================================================

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

# %%------------------------------------------------------------------------------------------------------------

'''  
Customers with a partner or dependents are .82 and .91 times as likely to cancel as normal customers.

Families and other large households seem to be less likely to change providers.

This could be due to higher incomes, less time to consider options, or another combination of factors.

'''

REFERENCES

[https://otexts.com/fpp2/#](https://otexts.com/fpp2/)