The George Washington University

Laboratory Ten:

Exploring Survival Analysis

Fernando Zambrano

DATS 6450: Multivariate Modeling

Dr. Reza Jafari

18 April 2020

Abstract:

Quick dive into exploring the application of survival analysis to customer retention with Python implementation.

Introduction:

Initially the field of survival analysis was developed in the medical and healthcare industry to give estimates of lifespan after certain diagnosis were determined such as cancer or other terminal illnesses. However contemporary applications include customer retention and machine durability in the business sector or determining premiums in the insurance industry. In general, survival analysis is statistical tool that answers, "how long would it be before a particular even occurs?" This type of methodology is often referred to as "time to event analysis."

Methods & Theory:

There are certain criteria and assumptions in order to perform survival analysis. The foundational assumption is that the time until a certain event occurs, T, is a non-negative continuous random variable that includes 0. T has the following probability density function and probability distribution function:

Random variable T has a **probability density function** $f(t)$ and **probability distribution function** $F(t)$.

$F(t) = P(T < t)$

$$\int_0^t f(x)dx$$

The survival function determines the probability that the event has not occurred by a given time t, in other words "survived." Below is the survival function equation.

$$S(t) = 1 - F(t) = P(T \geq t)$$

From a different point of view, the survival function can give insight into the hazard function, which is the amount of risk that will be gained at given time t. This can be derived by dividing the survival function over its derivative.

Hazard Function: h(t):

$$h(t) = \lim_{dt \to 0} \frac{S(t) - S(t + dt)}{dt} \times \frac{1}{S(t)}$$
$$= \frac{S(t)'}{S(t)}$$

S(t) – S(t + dt) is the proportion of cases that experienced the event, or "die," at time dt out of the cases that survive at time t. To find the actual number of cases that died, the population of all cases (P) needs to be implemented as well. Hence, the actual value of cases that survive or die can be represented by the following equation: *(S(t) – S(t+ dt))P*. The following equation: *S(t)P*: determines the actual number of cases that surrvied.

Survival analysis can be applied similarly across different industries, but the terminology and definitions of variables can vary. Below are some important definitions that are useful to know if for any applications within that field.

Predictive maintenance applies survival analysis to estimate how long a machine will last. Events are defined as the point in time when a machine ceases to work. Time begins when the machine is first used. What is important to know is the time scale, whether time is measured in seconds, minutes, hours, days, etc. The difference between the events and time of origin is called the lifetime.

Customer analytics aims to understand customer retention, or the lifetime value of a customer for a company. The event in these scenarios is when a customer *churns* or unsubscribes from the business service. Time of origin starts whenever the customer starts the service with the company. The time scale may be in days, weeks, months etc.

Marketing analytics are interested in retention rates of customers based upon the marketing channels which customers subscribe to. The event in this case is when the customer subscribes to the marketing channel in question. Time of origin is when the customer subscribed. Time scale can range from days to years.

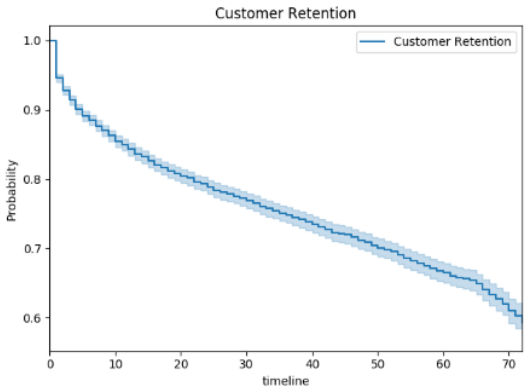
Actuaries are the probabilities that a population is at risk of dying within a certain time frame.

One important note to be cognizant of is that survival analysis is based off samples, and not working with population data. Hence the results of survival analysis are estimates and should be validated with confidence intervals.

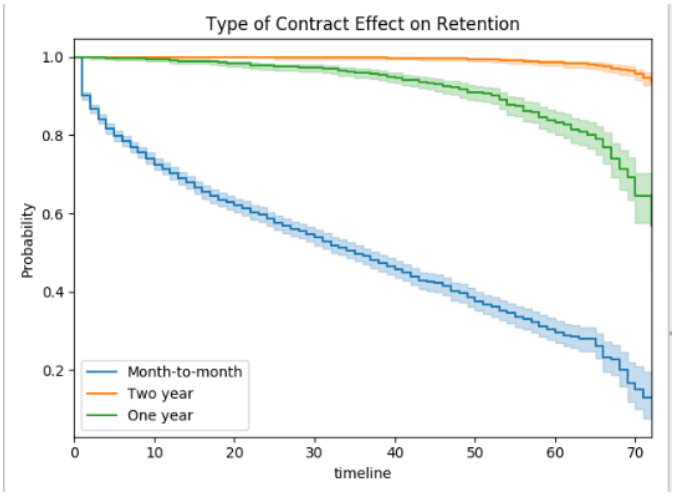
Implementation & Results:

The following examples will by implemented in Python using the lifelines package to estimate Kaplan Meir curves of different cohorts of a telecommunication company.

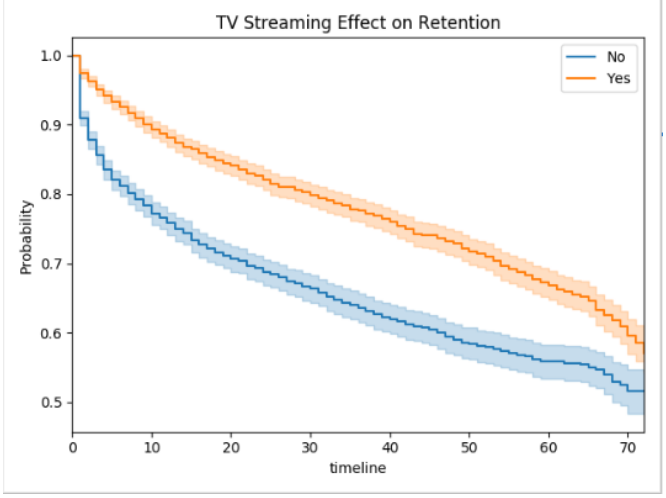
1. The first example illustrates the overall retention of customers from time = 0 to 70. The retention probability drops from 1 to almost 0.5 within the entire time frame. The probability initially decreases at an increasing rate but then begins to decrease to an almost constant rate from time = 25 to 65, and then decreases at an increasing rate after time = 65.



2. The next example will examine retention rates of different cohorts based upon customer contract types with the telecommunication company. Customer can choose three different types of contracts: month-month, one-year, and two-year. Customers who have two-year contract have a clearly higher probability of retention compared to 1-year and month-month. These contracts begin to lose retention towards the end of the timeframe. One-year contracts start off similar to 2-year contracts but begin to lose retention at around time =12 at which retention rate date decreases at an increasing rate. Month-Month contracts have the worst retention rates as they begin losing retention almost at the time of origin at increasing rate.



3. The last example will compare customer retention based upon whether that customer streams TV service or does not. There is clear distinction in retention rates between customers who stream TV compared to those that do not. Those that do stream TV have a much higher retention rate between time of origin and the time = 70. This is due to retention rates of customers who do not stream TV having a faster decreasing retention rate than those that do stream TV. However, by the end of the time frame the two cohorts have converged to similar retention probabilities.



Conclusion:

Overall, survival analysis is an intuitive and practical method to demonstrate how events effect different cohorts within the same population. This can give insight to which cohorts require more attention because they exhibit a higher risk or which ones are bringing in the most money because they have a lower risk depending on the application.

Appendix:

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



# q1-2
df = pd.read_csv("WA_Fn-UseC_-Telco-Customer-Churn.csv")
df.head()
# q3

df.info()
# q3

df.info()
df.isnull().sum()
# Question 4
df['TotalCharges']=pd.to_numeric(df['TotalCharges'],errors='coerce')
# Question 5
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
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()
plt.title('Customer Retention')
plt.ylabel("Probability")
plt.show() ;
# 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)
plt.title("Type of Contract Effect on Retention")
plt.ylabel("Probability")
plt.show();
#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.


# Those who have Two year contract have a clearly higher probability of retention compared to 1-year and month-month.
# One yearr contracts start off similar to 2-year contracts, but begin to lose retetntion at around t=12 and begins to
```

```
#
```

```
#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')

kmf1.fit(T[i1], E[i1], label='No')
ax = kmf1.plot()
kmf1.fit(T[i2], E[i2], label='Yes')
ax1 = kmf1.plot(ax=ax)
plt.title("TV Streaming Effect on Retention")
plt.ylabel("Probability")
plt.show();
```

```
#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?


# Based upon the survival analysis of whether customers stream TV service, there is a clear disticntion between those who do and those that do not stream TV from time 0 - 70.  At t = 70 the probablity of retention for both groups converge and becomes similar.
```

Sources: