Aalto University
CS-C3240 - Machine Learning

# Survival Analysis for the Better Scooter Company

November 2022

# 1. Introduction and motivation

## 1.1. Introduction

The number of scooter rental companies have dramatically increased over the lest three years due to the high demand for lower carbon emissions scooter rental. As a result, the competition in this business become intense, leading to the idea that the company needs to minimum the cost.

It is noticed that the cost of repairing scooters increases exponentially with their survival rate. The project is to answer these questions: how long a scooter need to be repaired, how many scooters in inventory will survive after a period of time, at what rate one could work or need to be repaired. To answer all these, a survival analysis must be studied. This project is to analyze the survival rate of scooters in inventory

Due to the high demand for sustainability and flexible mobility, there are many companies that rent scooters that have lower carbon emissions. Competition in this business is fierce! To make profits, the Better Scooter company hires you as a data analyst to help them improve their business.

## 1.2. Motivation

The company Better Scooter need a survival model to predict the survival rate for scooter inventory and thus optimize the cost and maximum profit for its business.

The given data provides the historical information related to scooter inventory, including manufactures they were purchased, accident record, average number of complaints in the last three months, and the accumulated miles they traveled. Since the repair cost and survival rate have a strong linear positive relationship, the company will have many benefits from predicting the survival rate for scooter inventory including 10 scooters.

# 2. Discussion of data

## 2.1. General information

The dataset extracted from the file **cleaned_better_scooter.csv** contains 283 rows/ observations with 7 columns/variables related to the scooters belonging to the Better Scooter company. Each datapoint corresponds to a scooter identified by a unique ID. More specifically, the properties of this dataset present historical information of the scooters, including scooter id (id), time to reparation (tte), if it needs to repair (need_repair), the number of days the scooter has been used, the manufacturer (manufactor), the average number of complaints in the last three months (avg_complains) and the accumulated miles (riding miles).

The features recorded for each datapoints are explained in the Table 1. There are 5 numeric and 1 categorical feature variables. The label for each observation is binary Boolean value “need_repair”, which is True if the scooter was repaired, False otherwise.

The other dataset from the text file **ten_scooters.csv** will be used to predict survival rate.

## 2.2. Descriptive Investigation

Data Investigation or exploratory data analysis is the crucial process of using statical summary and graphical representations to have a better understanding of the dataset to uncover patterns, detect anomalies, test hypothesis, and verify assumptions. This step is significant before modelling the data since it not only help analysts see what data may disclose outside of formal modelling (outliers, missing/NULL values, etc…) but also gives an insight into variables and relationship between them. The summary of descriptive investigation is described as following:

### 2.2.1. Anomalies Detection

The figure 1 shows that the dataset has no missing values.

There is no sign of anomalies detected on the columns name and values. The statistical information in the next section will confirm it better if there is anything abnormal about the dataset.

### 2.2.2. Statistical information

The figure 2 gives us some statistical information like the total number of rows, mean, standard deviation, minimum value, 25th percentile, 50th percentile, 75th percentile, and maximum value for each column. The visualisation will better help detect outliers and provide detail data distribution.

### 2.2.3. Numeric data distribution

It gives us a general visualisation about how data is distributed. In the graph 4, it is highlighted that most scooters received only one or two complaints on average, which is good. Most scooters have been used between 45 days and 55 days and ride from 25 miles to 120 miles. There is some outliers in the ride_miles variables with the riding miles of over 1000 miles.

# 3. Methods

## 3.1. Kaplan-Meier estimator and KM Curve

The Kaplan-Meier estimator is a non-parametric statistic technique to estimate the survival function from lifetime data. The conception of Kaplan Meier Estimator lies in estimating the survival times for a certain time of an event (a time to death, failure of machine parts or any major significant event).

The true survival function, which return the probability of survival beyond time t, approximated by the estimator according to the following formular:

$$\hat{S}(t) = P(T > t) = \frac{numbers\ of\ sample\ surving\ beyone\ time\ t}{total\ numbers\ of\ sample}$$

The Kaplan-Meier curve is a graphical representation of the survival function occurring at time points where one or more scooters were repaired. The curve is constructed by plotting the survival against time.

Since the survival time are subject to right-censoring, the status of reparation in addition to survival time (tte) need to be considered. The two variables "need_repair" indicates whether the actual survival time was observed or censored and "tte" denotes time to event/survival time, which is the time of reparation (if need_repair = True). The Kaplan-Meier estimator is chosen to conduct the survival analysis since it is valid if the survival times are right-censored. According to the figure 5 shows, there are 140 survival times are right-censored (need_repair = True or they were repaired during the study period) and 143 survival times are uncensored (need_repair = True or there are working without repairing).

### 3.2. Log-rank test

The log-rank test is a hypothesis test that is used in survival analysis to compare the survival distribution of time to event occurrence of two or more independent samples. Log-rank test is the most powerful for detecting alternatives that correspond to proportional hazards (related to Cox regression). In this task, the log-rank test is used to detect if there is difference among manufacturers regarding survival rate.

### 3.3. Cox proportional-hazards model

The Cox proportional-hazards model is a regression model commonly used to investigate the relation between the survival time and one or more predictor variables. Unlike Kaplan-Meier curves and log-rank test which are only only for categorical predictor variable is categorical,

The **Cox proportional-hazards model** is essentially a regression model commonly used statistical in medical research for investigating the association between the survival time of patients and one or more predictor variables, Cox proportional hazards analysis works for both quantitative and categorical predictor variables and assess the effect of serval risk factor on survival time.

### 3.4. Random Survival Forests

Random Survival Forests (RSF) is another non-parametric method for the analysis of right-censored survival data, beside Kaplan-Meier estimator. The RSF applies the same principle as Random Forest algorithm as: (a) it uses bootstrapped data to grow survival trees; (b) it uses split criterion as random feature selection when splitting tree nodes; (c) Predictions are formed by averaging predictions of individual trees in the ensemble.

The task is to compare C-index from two model Random Survival Forest and Cox proportional-hazards model. The interpretation of c-index is the same as AUC (area under curve) for binary classification. A value of 0.5 denotes a random model

The interpretation is identical to the traditional area under the ROC curve metric for binary classification: - a value of 0.5 denotes a model the predicts the outcome by random chance, a value of 1.0 denotes a perfect model, and a value of 0.0 denotes a perfectly wrong model.

## 4. Results

### 4.1. Survival analysis with Kaplan-Meier

In the plot 5, the y-axis represents the probability a scooter is still working without reparation after $t$ days, where $t$ days is on the x-axis. The figure shows that around 60% scooters survive after 350 days, and 40% scooters survive more than 700 days.

### 4.2. Kaplan-Meier curve by manufacturers

The plot 6 visualizes the survival rate in the different period. According to the graph, around 20%, 40%, and 60% scooters from manufacture C, B, and A respectively survive after 500 days. The scooters from the manufacture A seem to have higher chance to survive than that of the manufacture B, and C given the other risk factors are similar. Therefore, it is recommended that the company purchase scooters from the manufacture A.

### 4.3. Alternative hypothesis with log-rank test

The log-rank test yielded a p value of ***5.503522413332893e-14*** (greater than the significance level of 0.01), indicating that the difference in survival rate between the different manufactures was statistically significant. Therefore, it is assumed that the null hypothesis is wrong. In other words, manufactures plays significant role in a scooter's survival time based on both visualisation and testing.

### 4.4. Cox proportional-hazards model or Random Survival Forest

The c-index in the Cox proportional-hazards model and Random Survival Forest are **0.*75*** and ***0.77*** respectively, which means the Random Survival Forest on this dataset performs a bit better than Cox model.

The function **permutation_importance** of skicit-learn is used to estimate feature importance. The result from table 2 shows that the average number of complaints is by far the most important feature. If its relationship to survival time is removed, the c-index on the test data drops on average by 0.139939 points. It is interpreted intuitively that most complaints are on the survival time of scooters.

### 4.5. Prediction and Survival Rate Plotting

The result in the table 3 is a vector of coefficients using Cox proportional-hazards model, one for each variable, where each value corresponds to the log hazard ratio. For example, the coefficient for the variable "usage_length_days" is 0.005, corresponding to the hazard ratio exp(0.005) = 1.005. The hazard ratio 1.005 means the scooter with one more day used will increase 0.5% risk of event occurring (need to repair). The manufacture C has hazard ratio is exp(0.61) = 1.84, which mean the scooter produced by manufacture C is 1.84 times as likely to be repaired.

The graph 7 plotted the survival rate for 10 scooters in the dataset. The scooter 9 and scooter 4 have the highest survival rate while the scooter 5 has the lowest survival rate. According to the data set, the scooter 9 and scooter 4 have the lowest average number of complaints and both were purchased from manufacture A, verifying the accuracy of the result.

## 5. Conclusion and Discussion

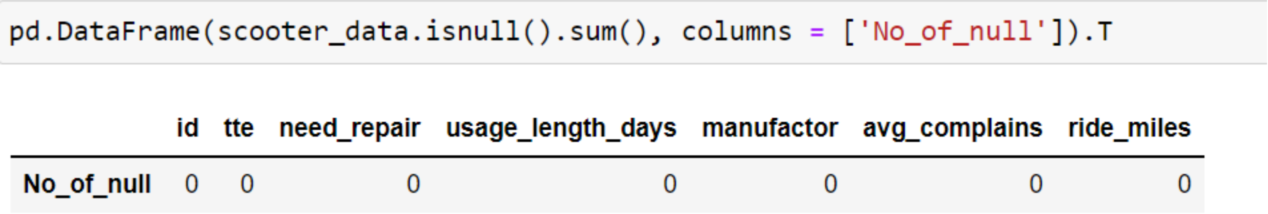
The project use serval regression models for the analysis of the right-censored survival data. The Kaplan-Meier estimator and log-rank test are more useful for the dataset with categorical variables while Random Survival Forest works for both categorical and quantitative predictor variables.

While Kaplan-Meier method estimates survival function according to a variable, Cox proportional-hazards model or Random Survival Forest algorithm for multivariate survival model. In this project, the performance of both estimators is almost the same (Random Survival Forest perform a little).

# 6. Reference

| *Variables* | Explanation | Data type |
|---|---|---|
| *Id* | Identification number of scooter | Int |
| *Tte* | Time to event | float |
| *Need_repair* | Event which is True if the scooter was repaired, False otherwise. | Boolean |
| *Usage_length_days* | Number of days the scooters have been used. | Float |
| *manufactor* | Name of manufacturer | Object |
| *Avg_complains* | Average number of complains in the last three months | Float |
| *Ride_miles* | Accumulated riding miles of the scooter | Float |

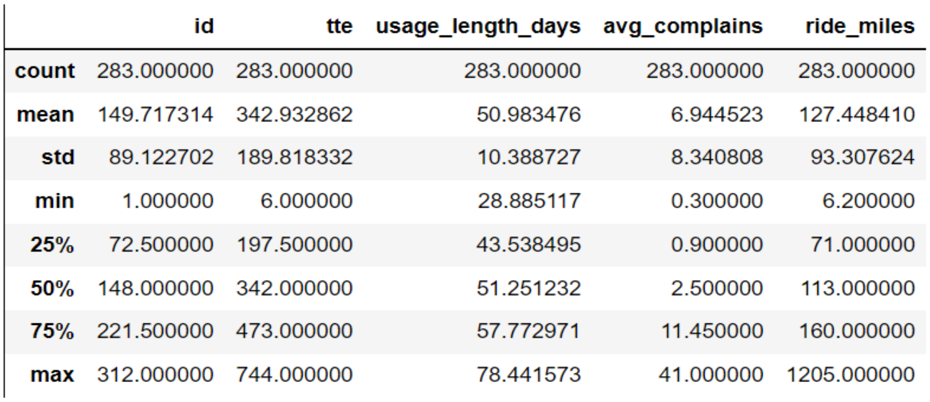
Table 1: Data set and explanation



```
pd.DataFrame(scooter_data.isnull().sum(), columns = ['No_of_null']).T
```

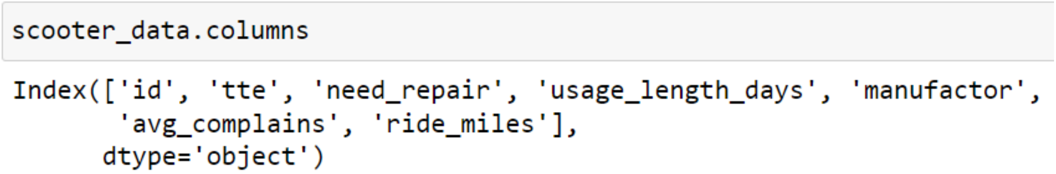
| | id | tte | need_repair | usage_length_days | manufactor | avg_complains | ride_miles |
|---|---|---|---|---|---|---|---|
| **No_of_null** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure 1: Check missing values



| | id | tte | usage_length_days | avg_complains | ride_miles |
|---|---|---|---|---|---|
| **count** | 283.000000 | 283.000000 | 283.000000 | 283.000000 | 283.000000 |
| **mean** | 149.717314 | 342.932862 | 50.983476 | 6.944523 | 127.448410 |
| **std** | 89.122702 | 189.818332 | 10.388727 | 8.340808 | 93.307624 |
| **min** | 1.000000 | 6.000000 | 28.885117 | 0.300000 | 6.200000 |
| **25%** | 72.500000 | 197.500000 | 43.538495 | 0.900000 | 71.000000 |
| **50%** | 148.000000 | 342.000000 | 51.251232 | 2.500000 | 113.000000 |
| **75%** | 221.500000 | 473.000000 | 57.772971 | 11.450000 | 160.000000 |
| **max** | 312.000000 | 744.000000 | 78.441573 | 41.000000 | 1205.000000 |

Figure 2: Statistical description



```
scooter_data.columns
```

```
Index(['id', 'tte', 'need_repair', 'usage_length_days', 'manufactor',
       'avg_complains', 'ride_miles'],
      dtype='object')
```

Figure 3: Columns

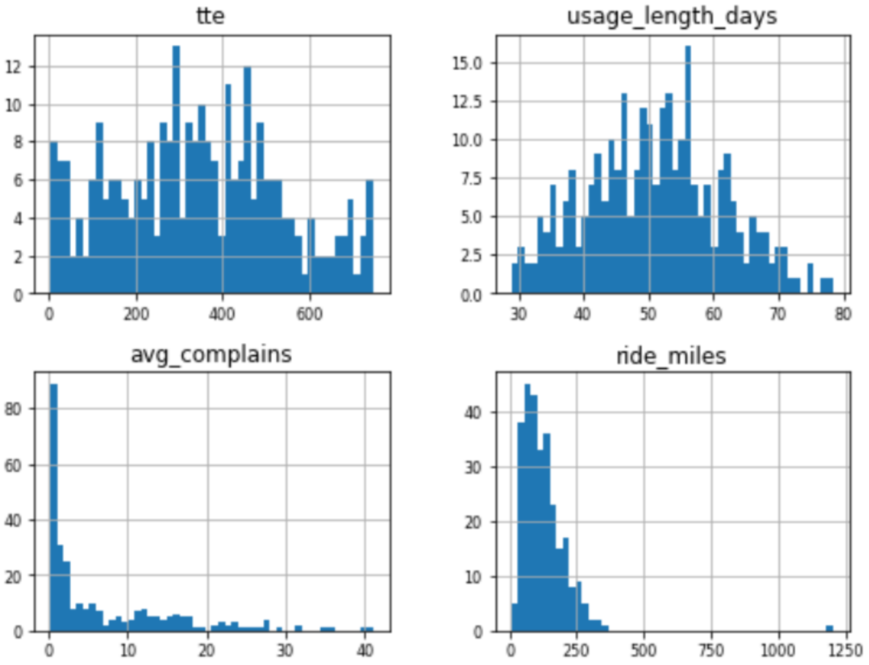


Figure 4: Numeric data distribution

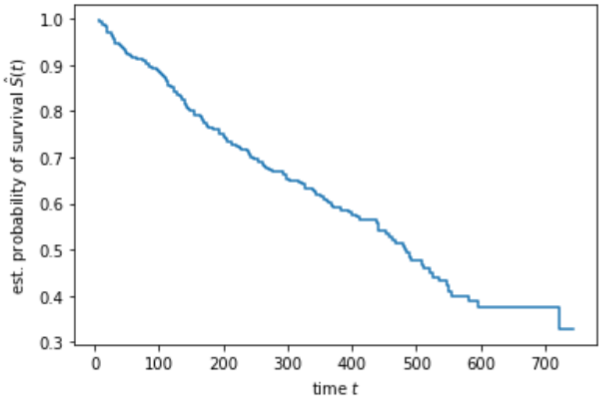


Figure 5: Survival analysis with Kaplan-Merier

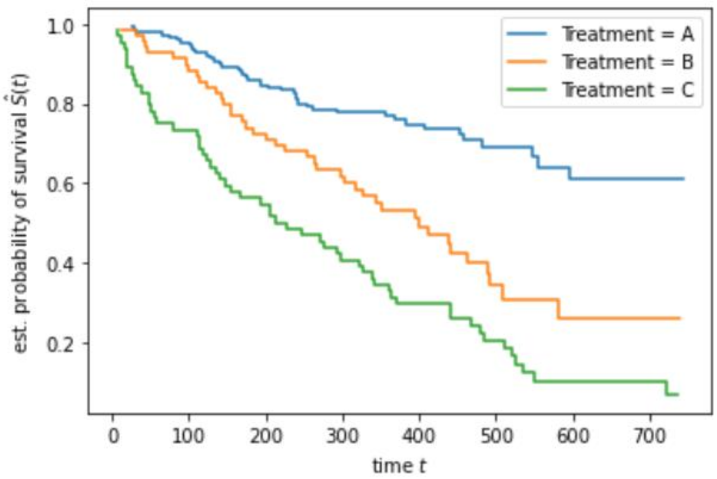
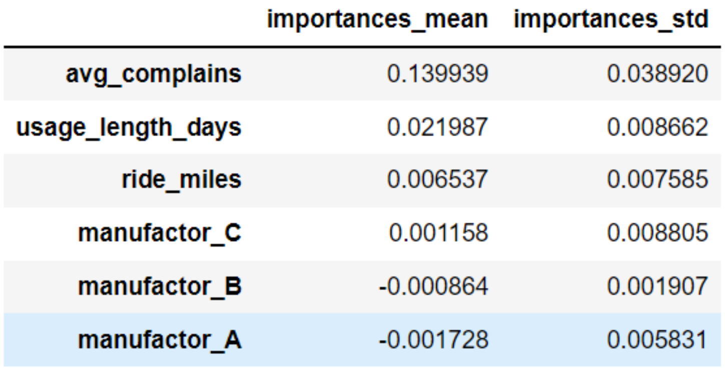
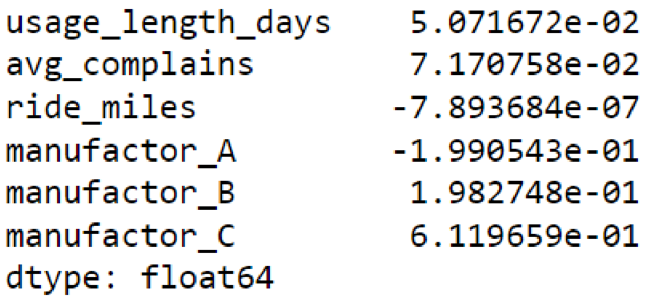


Figure 6: Survival rate by manufacture



| | importances_mean | importances_std |
|---|---|---|
| **avg_complains** | 0.139939 | 0.038920 |
| **usage_length_days** | 0.021987 | 0.008662 |
| **ride_miles** | 0.006537 | 0.007585 |
| **manufactor_C** | 0.001158 | 0.008805 |
| **manufactor_B** | -0.000864 | 0.001907 |
| **manufactor_A** | -0.001728 | 0.005831 |

Table 2: Feature importance with permutation_importance function



```
usage_length_days     5.071672e-02
avg_complains         7.170758e-02
ride_miles           -7.893684e-07
manufactor_A         -1.990543e-01
manufactor_B          1.982748e-01
manufactor_C          6.119659e-01
dtype: float64
```

Table 3: Cox model coefficient

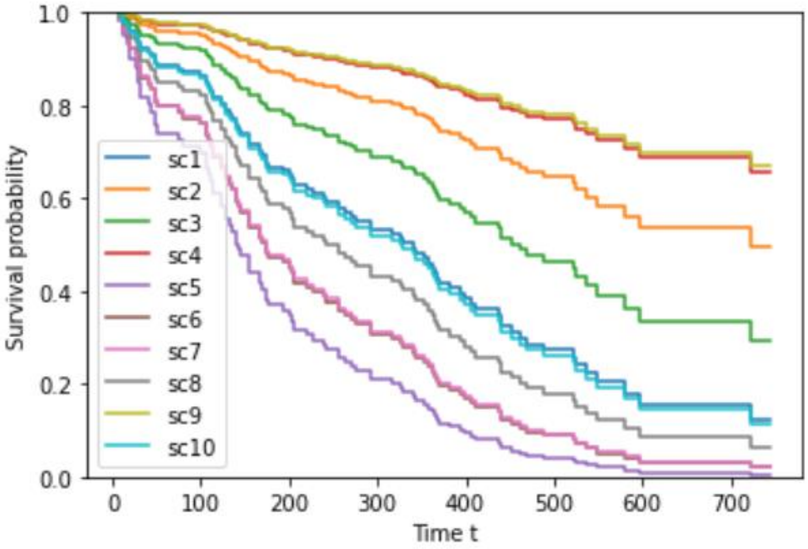


Figure 7: survival function with Cox model