**Name: Phat Nguyen**

# Survival Analysis

## Predicting survival rate of Scooters

## 1. Introduction and motivation

**Better Scooter Oy** - a limited liability company based in Finland ("**Company**" hereinafter) is offering a subscription-based service of renting electric scooters to customers in the Helsinki Metropolitan region. According to a recent article, 3 to 5 years is the average lifespan of an electric scooter for a private scooter with an average of daily use with one-time charging. However, the number is significantly lower for scooters offered by subscription-based businesses such as the Company, due to users using them less carefully than if they use their own vehicles.

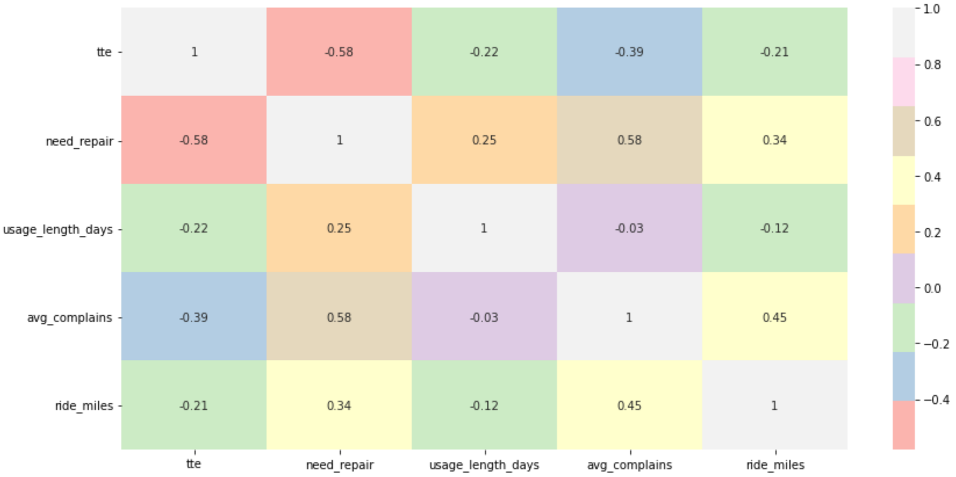
The company has tasked the author of the report to build a model to predict the survival rate of 10 scooters that the Company has in inventory in order to better predict the time the scooters need to be repaired.

## 2. Discussion of data and methods

The author received historical data from the Company in the format of a CSV file. The data has 283 rows (equal to 283 observations of 283 scooters) and 7 columns (equal to 7 data points of these scooters). There were no null values. The columns include:

(a) **id**: (int) Identification number of the scooter.
(b) **tte**: (float) Time to event.
(c) **need_repair**: (binary) Event, True means the scooter was repaired.
(d) **usage_length_days**: (float) Number of days when the scooters have been used.
(e) **manufactor**: (categorical) Name of the manufactors (A, B and C).
(f) **avg_complains**: (float) Average number of complaints in last three months.
(g) **ride_miles**: (float) Accumulated riding miles of the scooter.

The author decided to remove the column '**id**' due to the lack of explanatory values.



| | tte | need_repair | usage_length_days | avg_complains | ride_miles |
|---|---|---|---|---|---|
| tte | 1 | -0.58 | -0.22 | -0.39 | -0.21 |
| need_repair | -0.58 | 1 | 0.25 | 0.58 | 0.34 |
| usage_length_days | -0.22 | 0.25 | 1 | -0.03 | -0.12 |
| avg_complains | -0.39 | 0.58 | -0.03 | 1 | 0.45 |
| ride_miles | -0.21 | 0.34 | -0.12 | 0.45 | 1 |

**Name: Phat Nguyen**

**Figure 1:** correlation heatmap among numeric variables.

We quickly draw a correlation heatmap of the numeric variables.  Noticeably, we can see that there *might* be a correlation between ***avg_complains*** and ***need_repair*** (score=0.58).

We can also clearly see from the figures below that the number of scooters that needed to be repaired is similar to the number that did not. Additionally, scooters from manufacturers B and C seems to break down much more compared to those made by manufacturer A.

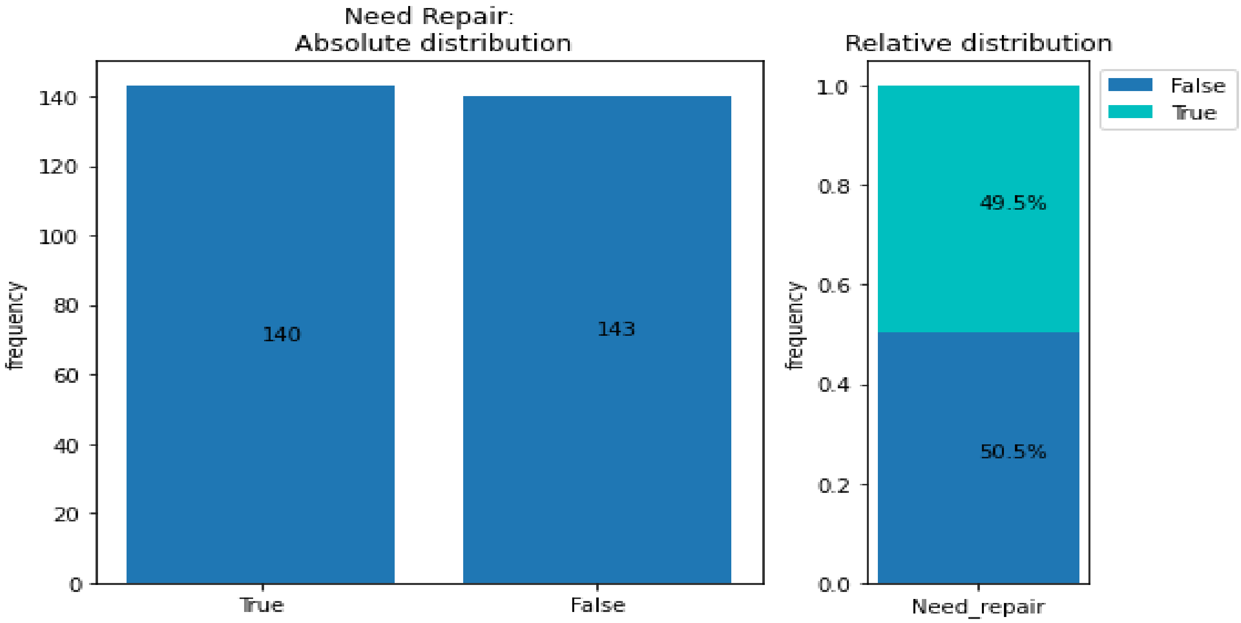


**Figure 2:** Absolute and Relative distribution of the number of scooters that needed repairing and those that did not.

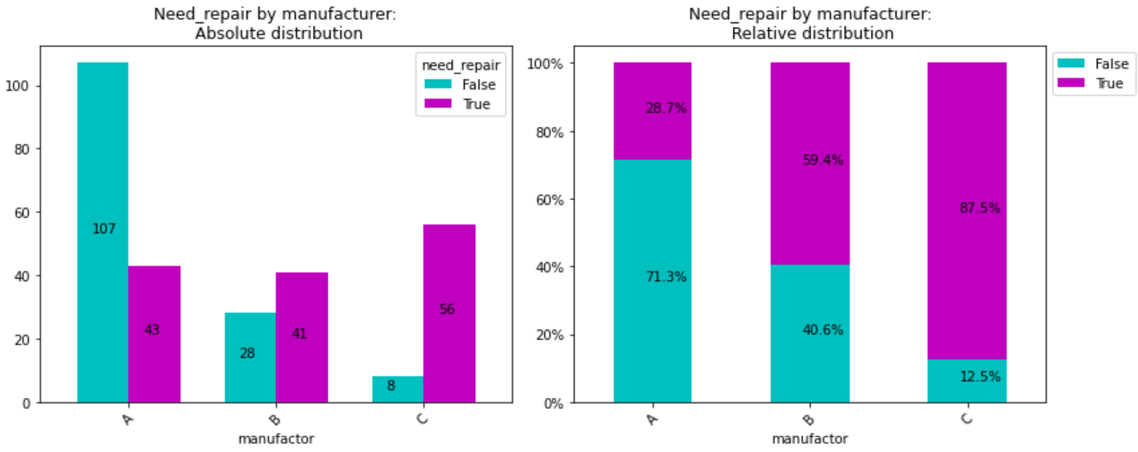


**Figure 3:**  Absolute and Relative distribution of the number of scooters from different manufactors

**Name: Phat Nguyen**

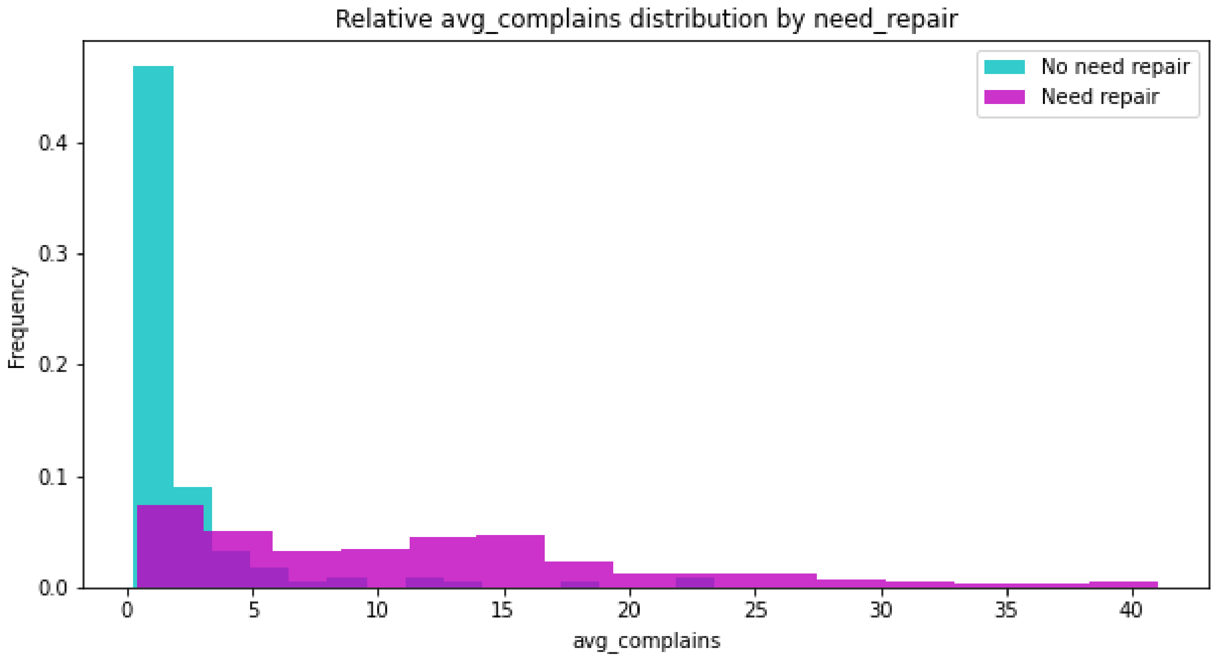


**Figure 4:** Relative distribution of the average number of complaints for scooters that needed repairing and those that did not.

We can see that generally, scooters that don't need repairing receive a smaller number of complaints. On the contrary, scooters that needed repairing receive more complaints.

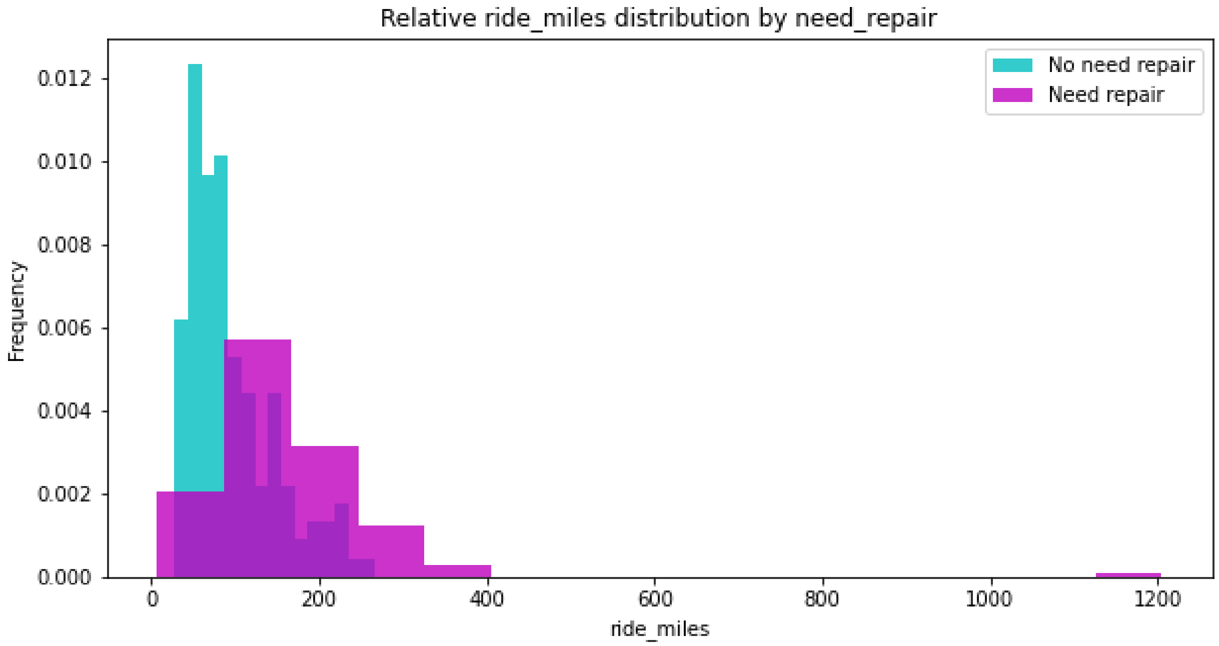


**Figure 5:** Relative distribution of the accumulated distance the scooters have ridden among those that needed repairing and those that did not.

We can see that generally, scooters that were used more (had more ride_miles) don't need to be repaired.

**Name: Phat Nguyen**

Using Kaplan Meier Estimator, a survival analysis is conducted with the historical data to build a survival analysis model to predict the survival rate of the scooters. Moreover, we look into the difference between scooters made by manufacturers A, B, and C.

## 3. Results

The result of Kaplan Meier Estimator is plotted in figure 6 on the right. Accordingly:

- At 300 days, the scooter has just over 65% of survival rate.
- At 500 days, the scooter has just under 50% of survival rate.

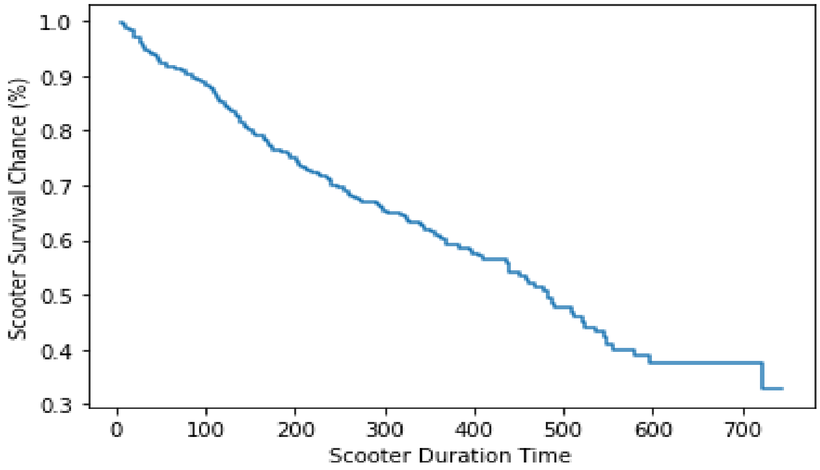


**Figure 6:** Kaplan Meier Survival Curve of all scooters.

Additionally, we can clearly see the difference in survival rate among scooters made by manufacturers A, B, or C. As shown in figure 7, scooters by manufacturer A have the best survival rate, followed by those made by manufacturer B, and scooters made by manufacturer C has the worst survival rate.

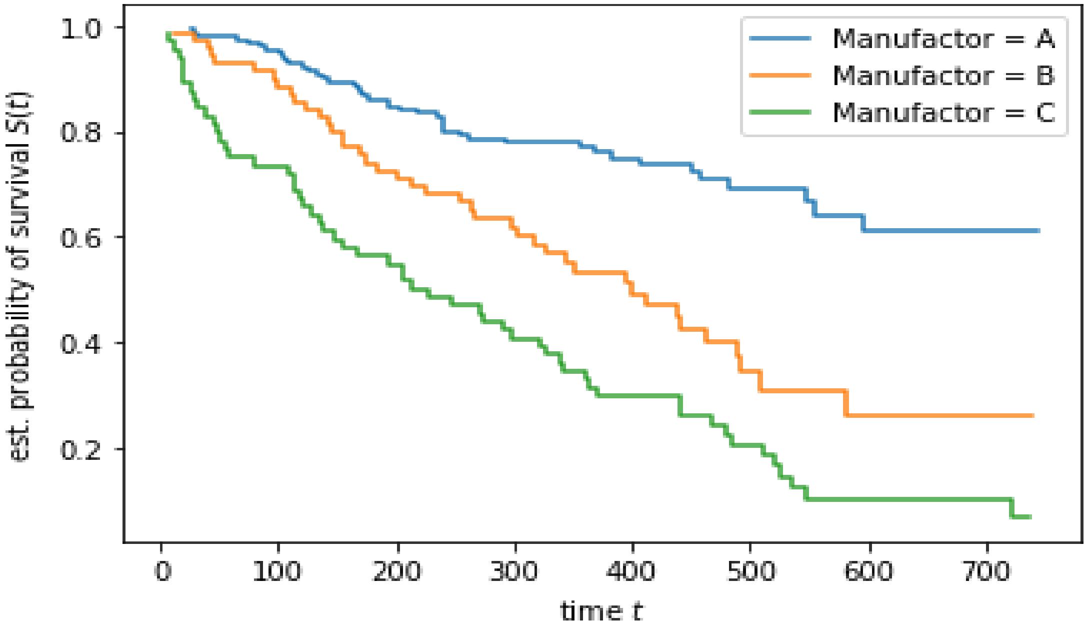


**Figure 7:** Kaplan Meier Survival Curve of scooters made by different manufacturers.

The log-rank test result shows that the ***p-value = 5.503522413332893e-14 < 0.001***. This means the null hypothesis (there is no difference between groups A, B, and C in terms of the distribution of time until the event occurs) is *rejected*. The test confirms our interpretation of figure 7 that there are differences in terms of survival rates between groups A, B, and C.

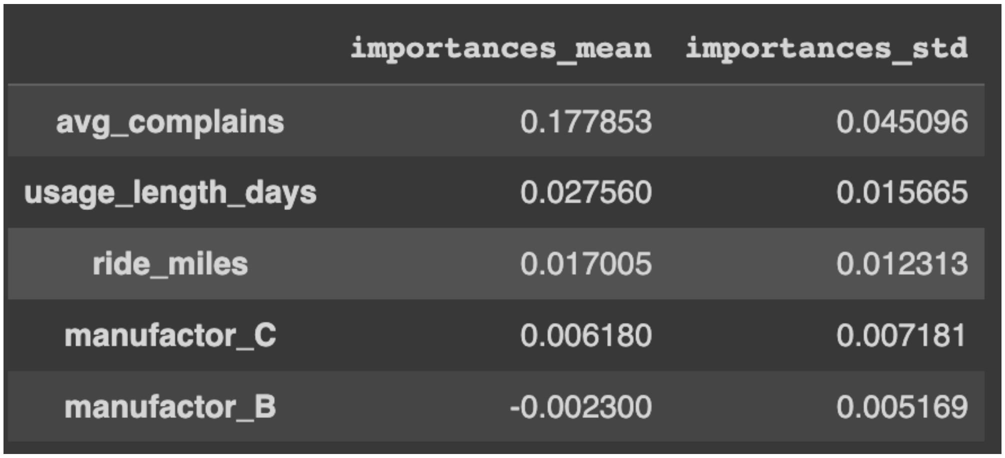
After using get_dummies method on categorical variable "***manufactor***"**,** the author split the data into training and testing sets with test_size=0.25 and random_state=2022. Next, the author also conducted **Cox Regression** and **Random Survival Forest** ("**RSF**" hereinafter) models. For the Random Survival Forest model, the parameters are:

- Test size: 25%
- Random state: 2022
- Number of trees: 1000
- Min samples split: 6
- Min samples leaf: 3

The c-index score are as follows:

- Cox Regression: 0.771
- RSF: 0.76

⇒ Both scores are higher than 0.7 which means the models are both strong, with Cox Regression being *slightly* better than Random Survival Forest.



| | importances_mean | importances_std |
|---|---|---|
| avg_complains | 0.177853 | 0.045096 |
| usage_length_days | 0.027560 | 0.015665 |
| ride_miles | 0.017005 | 0.012313 |
| manufactor_C | 0.006180 | 0.007181 |
| manufactor_B | -0.002300 | 0.005169 |

**Figure 8:** Estimate of feature importance by permutation.

Furthermore, the author analyzed the importance of different features of the random survival forest model. The result in figure 8 shows that on relative terms, the number of average complaints is the feature with more importance than others when it comes to predicting survival rates (highest importances_mean and rather low standard deviation). On the otherhand, other features (with low importances_mean and high standard deviation) have low effect on the survival rates.

**Name: Phat Nguyen**

## 4. Conclusions and discussions

Last but not least, we used the RSF model to predict the survival rate of the 10 scooters in inventory.

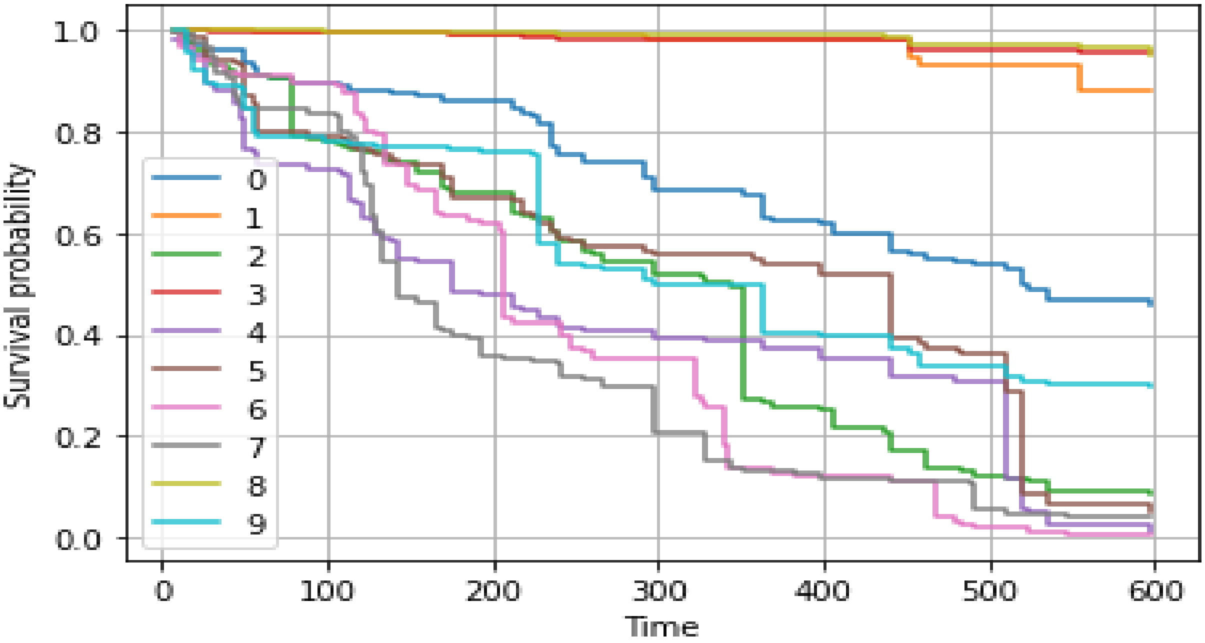


**Figure 9**: RSF Survival Curve of 10 scooters in the inventory.

The chart shows that scooters with ID 9, 4, and 2 will have excellent survival rates. For instance, even after 600 days, the survival rates of these scooters are still 90% and above. The number on the chart needs +1 since ID from 1 to 10 is numbered from 0 to 9.

On the other hand, scooters 1 and 10 have good survival rates. For instance, after 600 days, the survival rate is 50% and 30% respectively.

Last but not least, the remaining scooters have poor survival rates. For instance, after 600 days, the rates are under 10%.

Using the model and the chart, the Company can predict the survival rate of each scooter at a given time. Depending on the acceptable survival rates defined by the Company, the Company can prepare beforehand when certain scooters' survival rates will fall below, hence have corresponding measures to fix them before they break down.

Furthermore, the Company can also execute preventative maintenance measures to prolong the scooter's lifecycle, such as:

- Cleaning components
- Lubricate moving parts
- Replace battery
- etc.