1. Motivation

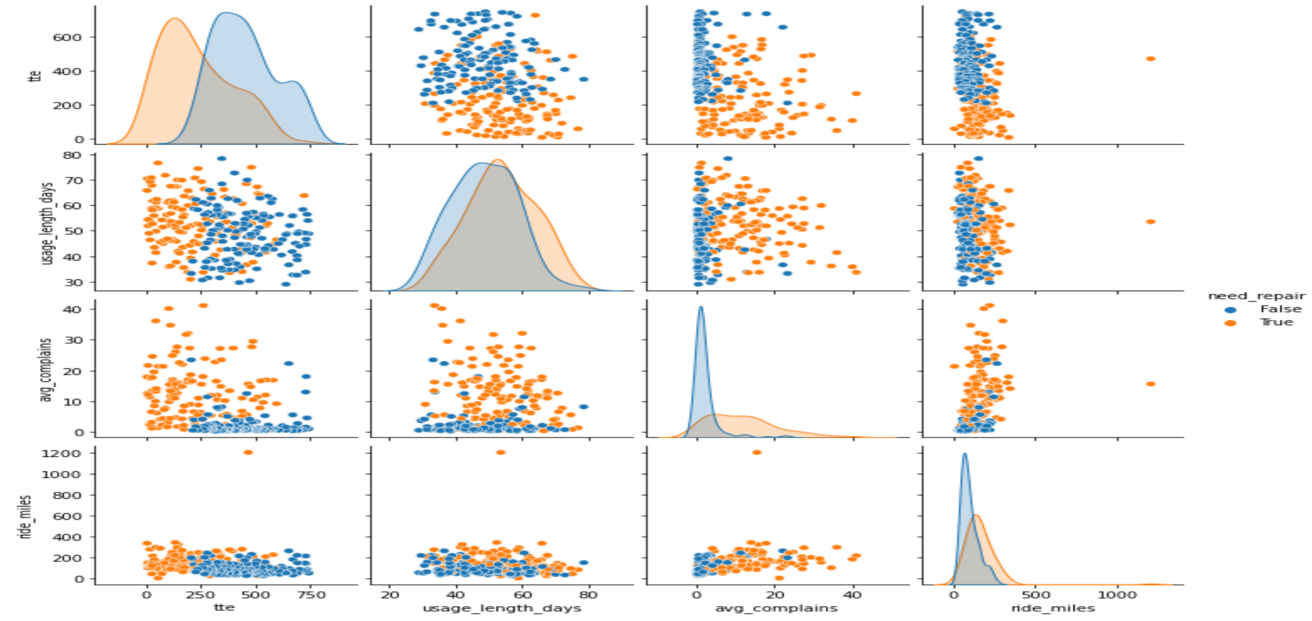
To evaluate the current scooter business and enhance its business profit, the cost is an important factor that is considered as when the cost is lower, the business would either have greater profit or have a better competitive advantage by lowering the price. The analysis is to investigate how current scooter business is, to know what factors affect the proportion of scooters that do not need to be repaired after certain time periods, find out the model for prediction of that proportion, and detect the factors that are common in the scooters having high proportion of being in good condition after some period of times, hence, making data- driven decision to reduce the cost for the business.

The structure of this report is organized as follows. First is the motivation, second is the data and exploratory analysis, third is the survival analysis, fourth is the model and fifth is the conclusion.

2. The data and exploratory analysis.

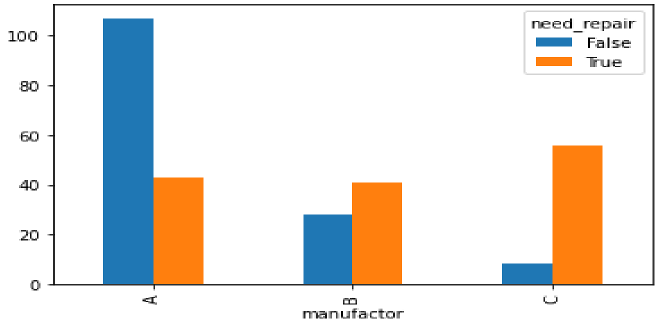
The data ‘cleaned better scooter.csv’ has 283 rows and 7 columns including ‘id’ column. The id column is removed in the beginning as it is not informative for the analysis. The remaining columns are: ‘tte’(time to event), ‘need_repair’, ‘usage_length_days’, ‘manufactor’, ‘avg_complains’, and ‘ride_miles’. The ‘need_repair’ columns contain boolean(True/False) data, ‘manufactor’ contains categorical data, and other columns contain continuous data.

Overall, the data in the continuous types are almost normally distributed respective to the ‘need_repair’ category.



The column ‘tte’ has a min value of 6 and max of 744. The min and max of ‘usage_length_days’ are respectively 28.88 and 78.44. The ‘avg_complains’ column ranges from 0.3 to 41. Meanwhile, the ‘ride_miles’ ranges from 6.2 to 1205.

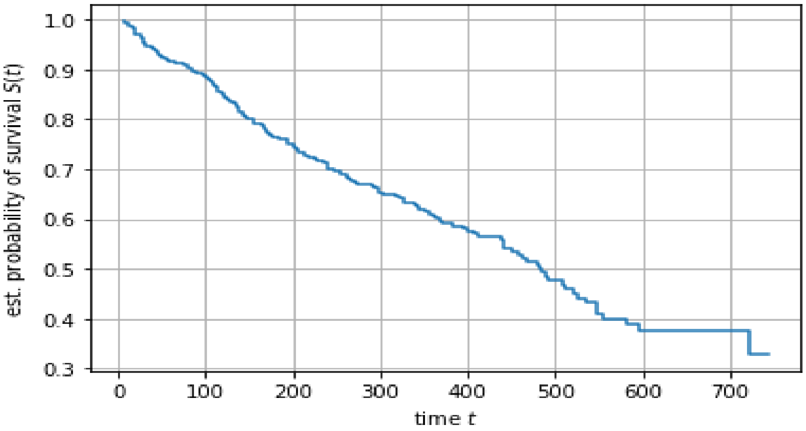
Regarding the ‘manufactor’ column, manufacturer A has the highest number of scooters that do not need repair while manufacturer C has the highest number of scooters that need repair.



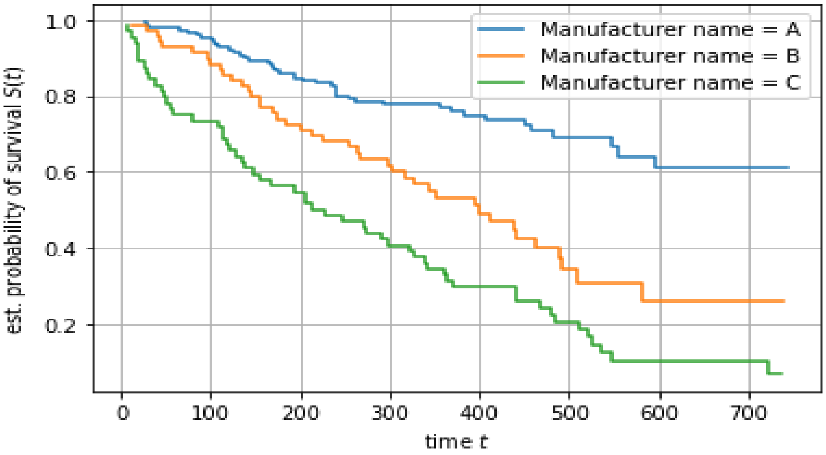
3. Survival Analysis

a) Survival analysis with Kaplan-Meier estimator

Below is the Kaplan-Meier plot for the 'tte' and 'need_repair' columns. In general, the estimated curve is neither late nor early. Overall, 50% of subjects survive to nearly 480 days. It means that after about 480 days, about the proportion of scooters that do not need to be repaired is 50%.



b) Survival analysis with Kaplan-Meier estimator for different manufacturer



The above graph shows the rate of scooters that do not need to be repaired until a different number of days. Each curve represents the scooters of different manufacturers. Overall, the curve of manufacturer A is the latest while the curve of manufacturer C is the earliest among 3 manufacturers. From the above graph, the scooters of manufacturer A are highly recommended because the curve of this manufacturer drops slowest among 3 manufacturers, or the proportion of scooters that do not need to be repaired of manufacturer A remains highest among all the durations given. For example, at a certain time of 300 days, about 80% scooters of manufacturer A do not need to be repaired, while only about 60% scooters of manufacturer B and about 40% scooters of manufacturer C do not need to be repaired.

c) Log- rank test

To further evaluate statistically the difference in the proportion of scooters that do not need to be repaired presented in b part, log-rank test is taken into account. The p-value of the log-rank test is yielded with 5.503522413332893e-14, which is less than 0.01(the threshold value). So, it can be concluded that the difference in probability of survival, or the probability of a scooter that does not need to be repaired is significantly different among 3 manufacturers.

4. Models

There are 2 models that are built to predict the time of survival (days until the time that scooters need to be repaired). The models used are CoxPHSurvival and Random Survival Forests.

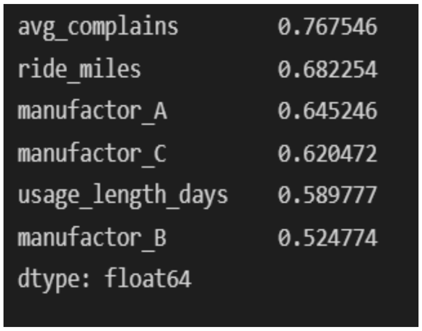
a) Model evaluation

To evaluate the models, the concordance index(c-index) score is taken from both models. The c-index is the evaluation of rank-order statistics for predictions against true outcomes. The higher the c-index is, the more predicted rank-orders are correct.

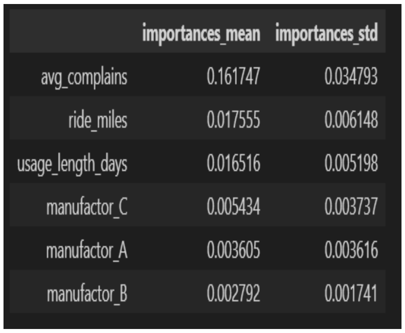
The c-index score of CoxPHSurvival for test data is about 0.796 and that score of Random Survival Forests Analysis is about 0.772. Because the c- index of CoxPHSurvival model is higher than Random Survival Forests model, CoxPHSurvival is considered the better model for prediction.

b) Feature evaluation

For both CoxPHSurvival and Random Survival Forests models, ‘avg_complains’ and ‘ride_miles’ are the most and second most important features with highest computed important scores. However, the feature importance is different in both models from the third most important position. The order of feature importance from the third most important position for CoxPHSurvival is ‘manufactor_A’, and continues with ‘manufactor_C’, usage_length_days, and ‘manufactor_B’, meanwhile, that order is different in Random Survival Forests as it starts from usage_length_days, ‘manufactor_C’, ‘manufactor_A’, and ends by ‘manufactor_B’. Below are the order of feature importance and important scores of the CoxPHSurvival(picture on the left) and Random Survival Forests(picture on the right).

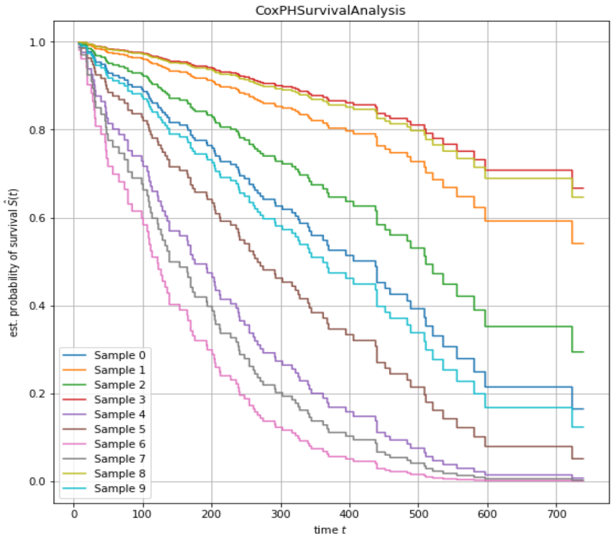


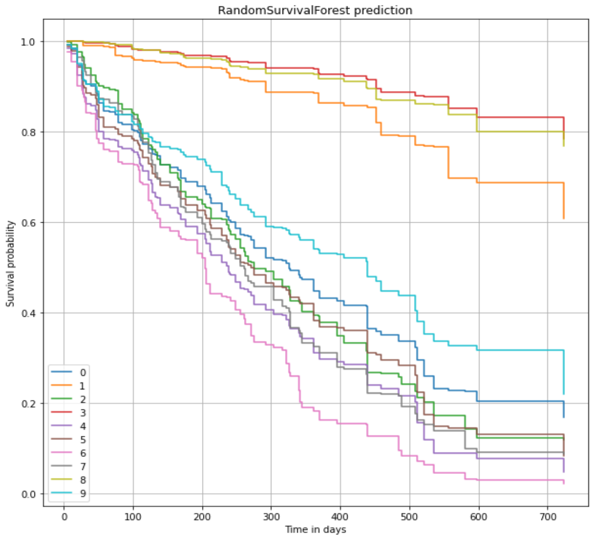
```
avg_complains        0.767546
ride_miles           0.682254
manufactor_A         0.645246
manufactor_C         0.620472
usage_length_days    0.589777
manufactor_B         0.524774
dtype: float64
```



	importances_mean	importances_std
avg_complains	0.161747	0.034793
ride_miles	0.017555	0.006148
usage_length_days	0.016516	0.005198
manufactor_C	0.005434	0.003737
manufactor_A	0.003605	0.003616
manufactor_B	0.002792	0.001741

c) Predict the samples





The file ‘ten_scooters.csv’ is used as data for prediction.

From both models, there are no stark differences in the survival curves respectively to each sample. However, it is worth noting that samples 1,3, and 8 have the latest survival curves, which mean that those samples have highest probability of survival as time goes by. For CoxPHSurvival model, the probability that the scooter does not need to be repaired after 600 days is about 70% for sample 3 and 8, about 60% for sample 1, and about 0% for sample 4,6 and 7. For Random Survival Forest, the probability that the scooter does not need to be repaired after 600 days is about 85% for sample 3 and 8, about 35% for sample 9 and about 5% for sample 6. The survival curves in Random Survival Forest prediction are later than those of CoxPHSurvival prediction.

Sample 1,3, and 8 are further investigated. Those scooters have some characteristics in common. Those scooters are the products of manufacturers A, have lower than 2 in average complaints, ride miles below 100 and usage days lower than 56.

5. Conclusion

Briefly, manufacturer A is highly recommended as a good supplier as the majority of its scooters do not need to be repaired as time goes by. Next, CoxPHSurvival is highly recommended as a potential model for prediction as it has a good c-index. Meanwhile, the most important features for survival function prediction are average complaints and ride miles. Also, some common characteristics of scooters having high survival rates (probability that the scooter does not need to be repaired as time goes by) are few complaints, have small ride miles and usage days values, and be the product of manufacturer A.