Predicting failures using machine learning

Tong Chun Ho, Department of Mathematics, HKUST
Lai Cheuk Man, Department of Mathematics, HKUST
Wong Ngo Cheung, Department of Mathematics, HKUST

Introduction

We predicted the failure by analyzing the dataset “OW = 2, PW = 1” provided in https://www.kaggle.com/c/nexperia-predictive-maintenance

We utilized classification tree and random forest to find the important factors. Then we build serval classification models by logistic regression, lda, qda, knn and svm. Finally, we do the qualification of uncertainty by bagging/ bootstrap.

Data Preprocessing

Missing Value Imputation

In this dataset, missing values are represented by blank values. Therefore, we changed the data into character first and the values were shown in NA. As the missing data in label column are small in size compared to the sample size, we decided to omit these data. Then, we assigned the value of “True” with 1 and 0 for “False”. For the value in count, mean and standard deviation, we replaced the NA with 0.

Categorical Factors

We changed the date column into several categorical factors. For one of the factors, it was ranged in a specific month. Taking the first datum which has a date of 2015 -05- 31 as an example, we classified it as 1505 which means 1 if the date is in 2015-05. For the last group of the categorical factors, we deleted it as we set it as a reference group.

Splitting training data and test data

We separated the data of “train.csv” into 70% of training data and remaining 30% of test data.

Modelling

Regression tree

We performed a regression tree with all the factors except the status and found out that it stopped splitting at the node of 2. The terminal node is time14.count. The error rate is 0.08112724.

True/ Predicted	0	1
0	3260	253
1	0	0

Random Forest

We grew 1000 trees with the number of factors n = sqrt(115). The result turned out a misclassification rate 0.0001314406. By exploring the importance of the forest, we found out that time6, time8, time14, time15 and time17 have a significant mean decrease gini.

True/ Predicted	0	1
0	3260	253
1	0	0

Factor	Mean Decrease Gini
time6.count	36.3246983
time6.mean	38.2850650
time6.sd	39.6447376
time8.count	26.7994366
time8.mean	37.9965521
time8.sd	37.5176300
time14.count	37.0643670

time14.mean	42.1997027
time14.sd	40.4898047
time15.count	33.3691650
time15.mean	35.7438105
time15.sd	35.8490630
time17.count	32.7452149
time17.mean	35.7306395
time17.sd	31.9369231

Logistic Regression

We built three models where the first model consisted of the factors mentioned above, time14 for the second model and only time14.count for the third model. The error rates of first two were a bit lower. We decided to use the count mean and s.d. Of time14 for the next model construction.

Model 1:

True/ Predicted	0	1
0	3259	253
1	1	0

Model 2:

True/ Predicted	0	1
0	3259	253
1	1	0

Model 3:

True/ Predicted	0	1
0	3258	253
1	2	0

Linear Discriminant Analysis

The error rate was similar to those of the previous models.

True/ Predicted	0	1
0	3256	251
1	4	2

Quadratic Discriminant Analysis

The error rate increased.

True/ Predicted	0	1
0	3177	253
1	2	0

K-nearest Neighbors

We constructed three models with different value of k. The error rate of the third model was the lowest.

Model 1 (k = 1)

True/ Predicted	0	1
0	3026	231
1	234	22

Model 2 (k = 5)

True/ Predicted	0	1
0	3237	252
1	23	1

Model 3 (k = 10)

True/ Predicted	0	1
0	3258	253
1	2	0

Support Vector Machine

We perform svm with linear, radial and polynomial with cost = 0.01. The radial svm took gamma as 1 and the polynomial considered the degree as 1. The error rates were the same.

Model 1 (linear)

True/ Predicted	0	1
0	3260	253
1	0	0

Model 2 (radial)

True/ Predicted	0	1
0	3260	253
1	0	0

Model 3 (polynomial)

True/ Predicted	0	1
0	3260	253
1	0	0

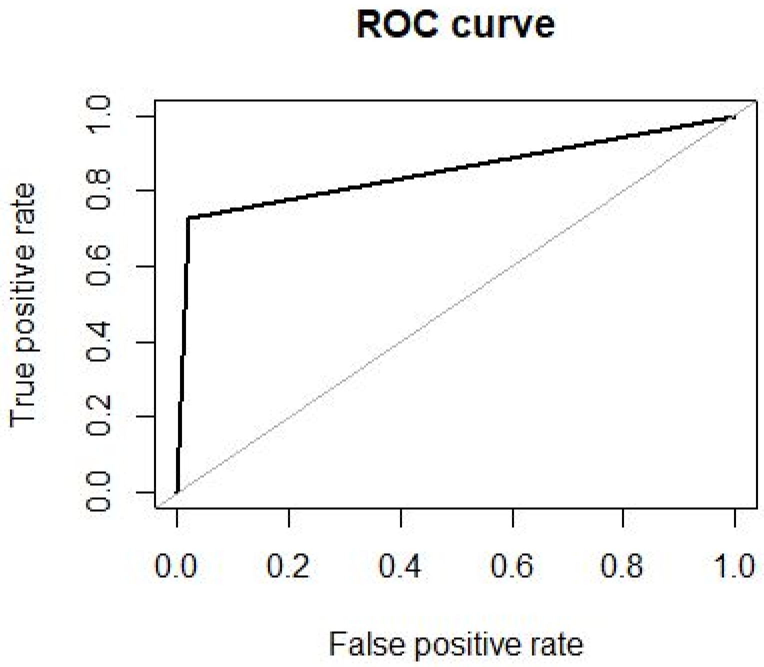
Sensitivity

From the aspect of sensitivity, knn performed the best.

Method	Error Rate
Classification Tree	0
Random Forest	0
Logistic Regression	0
Linear Discriminant Analysis	0.007905138
Quadratic Discriminant Analysis	0
K-nearest Neighbors	0.0869565
Support Vector Machine	0

Result

Area under AOC: 0.854



Analysis

In the models produced by classification tree and random forest, we noticed that the date of the error happened does not affect the failures.

Second, the models constructed by logistic regression demonstrated that the 14th time interval is much important to the others. The model with 14the time interval and other intervals had a similar error rate as that with only the 14th time interval.

Third, in the confusion matrices, the sensitivity is very low. It showed that the models perform badly when predicting the value of “true”.

Conclusion

The error rates of model built by different approaches were similar but the sensitivity of KNN with k = 1 was the lowest. Therefore, we decided to use the KNN model.

Due to the low value of k, it tends to do well by randomly sampling feature rather than training. Therefore, we decided not to do the bagging or bootstrap.

To examine the model performance, we used another dataset “OW = 4, PW = 1” as a test dataset. The area under ROC is 0.854. The shape and the area showed that the model did quite well in this dataset.

Contribution

Data Preprocessing

- Lai Cheuk Man

Model

- Tong Chun Ho
- Wong Ngo Cheung