Analysis of APS Failure at Scania Trucks

Stella Lang

Data Preprocessing

The original training dataset provided contains 60000 observations with 171 attributes and 16000 observations with 171 attributes in testing dataset. First, I loaded the data into R and converted the label class to factor and the rest to numeric. Since the dataset contains lots of missing values, I removed columns with over 10% NAs in the training dataset, which reduces the number of attributes to 143. Then I replaced NAs with median corresponding to each column. Last, I transformed all attributes except class to a scale between 0 and 1, while retaining rank order and the relative size of separation between values. After scaling, I found that column 81 (attribute “cd_000”) returns NaNs for the whole column. Since all observations have the same value for attribute “cd_000”, I removed “cd_000” from the training dataset as well, repeating the same process for testing dataset. After data preprocessing, the training and testing datasets have 60000 and 160000 observations respectively with 142 attributes.

Models

In order to predict whether or not truck component failures are related to the Air Pressure System (APS), I applied the following three machine learning approaches on the training dataset: Random Forest (RF), Support Vector Machine (SVM) and Gradient Boosting Machine (GBM), and assessed each model’s performance based on the corresponding prediction accuracy of testing dataset. `train` function from caret package and repeated 10-fold cross-validation are used to choose the best tuning parameters. In addition, since the data is quite unbalanced, I applied undersampling method to adjust the class distribution. Consider that the dataset is fairly large, training models on the whole training dataset would be quite time-consuming. Therefore, I use 10000 training data to test all 16000 testing data.

Random Forest

Random Forest is an ensemble method in which we create a classifier by combining several independent base classifiers. The ensemble classifier then coalesces all predictions to a final prediction based on a majority vote. By averaging several trees, there is a significantly lower risk of overfitting. It overcomes the major drawback of Decision Tree which is highly biased to training dataset. In addition, RF doesn't have strict restrictions on data and is able to deal with unbalanced and missing data. I used default tuning grid for random forest in `train` function.

Support Vector Machine

SVM is intrinsically suited for two-class problems. It works on the principle of fitting a boundary to a region of points which are all alike. Once a boundary is established, most of the training

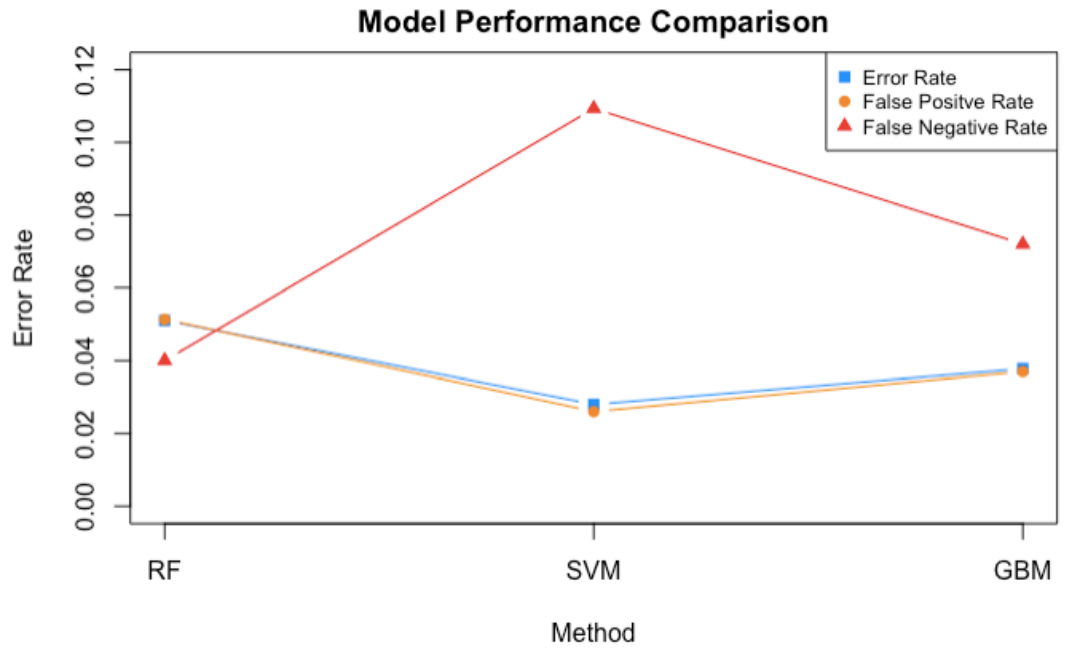
data is redundant. All it needs is a core set of points which can help identify and set the boundary. SVM is also computationally cheaper compared with logistic regression. For this dataset, I used linear kernel for SVM since the number of features is large. We may not need to map data to a higher dimensional space. In other word, the nonlinear kernel does not improve the performance. Using the linear kernel is good enough, and it only searches for the parameter C, which leads to shorter solving time.

Gradient Boosting Machine

Gradient Boosting Machine, unlike Random Forest, builds trees one at a time and uses each new tree to correct errors made by previous trees. It uses weighted averaging, which gives a more reliable prediction if overfitting is not an issue. I used default tuning grid for gradient boosting machine in train function.

Results

The plot below shows the general trends of how each model performs in terms of classification error rates, type I and type II error rates. From the plot, we can see that there is no significant difference in error rate and false positive rate among three models while the false negative rate of SVM differs greatly from RF.



The more detailed results for models implemented are shown below. In terms of classification error, the best result is achieved by SVM, which outperforms RF (improvement of 2.3%) and GBM (improvement of 1%). However, RF performs better than SVM and GBM in terms of total

cost. The cost for false negative is much higher than cost for false positive and RF has the lowest false negative rate which is 4%. Therefore, to minimize total cost, RF would be a better choice than SVM and GBM.

Method	Error Rate	FP	FP Rate	FN	FN Rate	Cost
RF	0.051	802	0.051	15	0.040	15520
SVM	0.028	407	0.026	41	0.109	24570
GBM	0.038	578	0.037	27	0.072	19280