Analysis Report

Data Preprocessing:

Step1:After exploring data, I found that there are multiple sensor attributes with many null values in it almost every feature had null value.

Step2:I replaced null value with mean of the every feature column of the data frame.after replacing null value, I have normalized each and every feature columns by Min max scaling as every feature was scaled differently which would have produced bias while modeling.

Step:3 I repeated same procedure for training and testing data.

Step4:As there are multiple attributes which would have taken lot of running time therefore I have applied pca where the no of features space was reduced to 11 column covering 95% variance.

Step5:As there are only few positive class in dataset I have applied oversampling to adjust class distribution of dataset.

Step6: I have applied same process to testing data as well.

Data Features:

After preprocessing training data had Rows=118000, column=11 and separate class label.

I have further split into training and validation set

Train set rows=94400

Validation set rows=23600

After preprocessing testing data had Rows=16000, column=11 and separate class label.

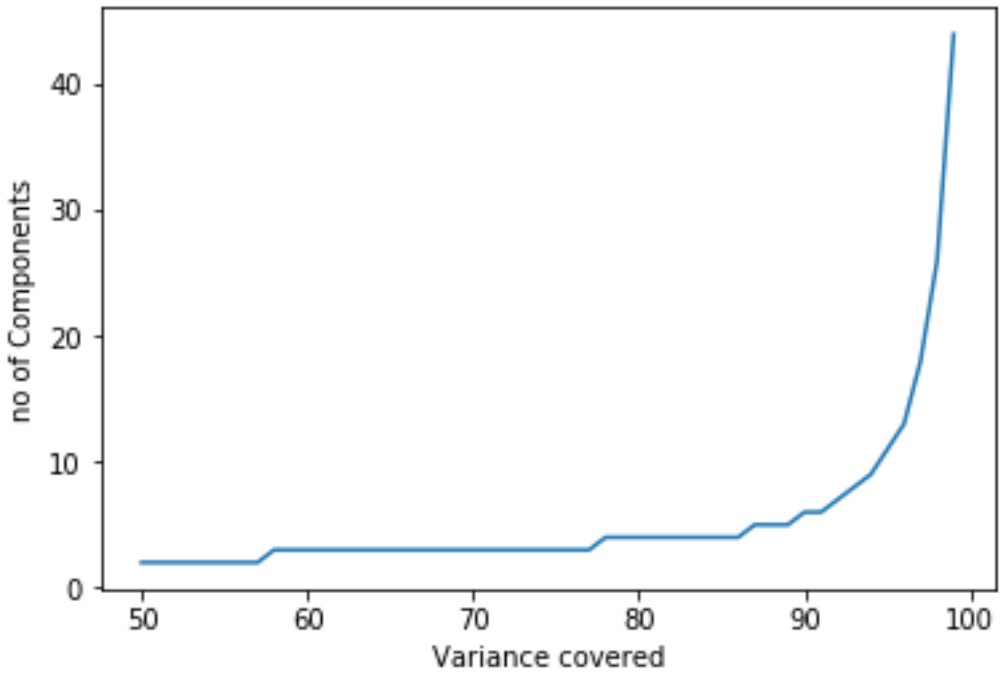
Outlier Detection:

For outlier detection I have selected all the features and saw that hich feature had unique value less than 50 .

Applying statistic technique and removing rows having outliers to original train data frame would lead me to very few rows remaining in training set as many of features have outliers where scaling was not done

In order to solve this problem I have applied scaling technique to original Data which lead to removal of dataframe.

PCA Analysis:

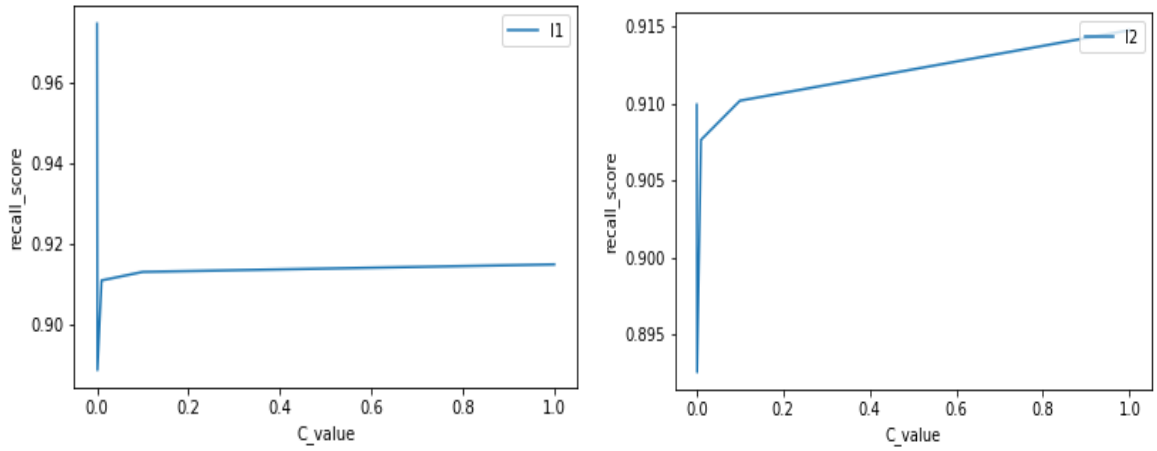


As you can see that 95% variance can be covered with 11 components.

Model Fitting:

LogisticRegression:

I have slected logistic regression because of sigmoid function which gives conditional probability for the particular class . This type of problem where logistic regression works better because of probability of instances belonging to that class.

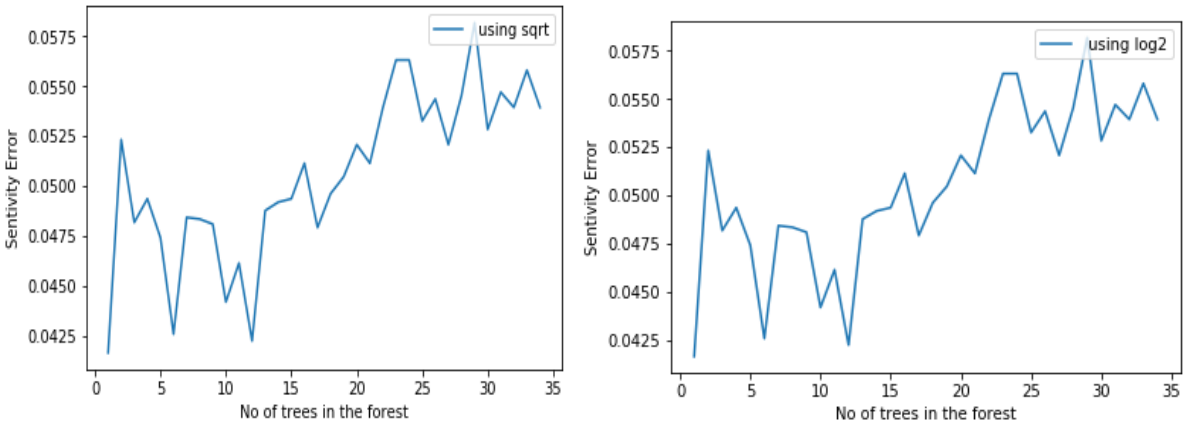


I have plot these graph for selecting feature for l1 and l2 penalty for regularization.

I have selected optimum feature based on above plot and then applied to test data.

Random_Forest:

RF works better in this type of classification problem because it requires no input feature preparartion and it implicitly select the features. Therefore I have tried no of trees in random forest and selected optimum no of trees based on the sensitivity error



Then I applied final model to test data

SVM:

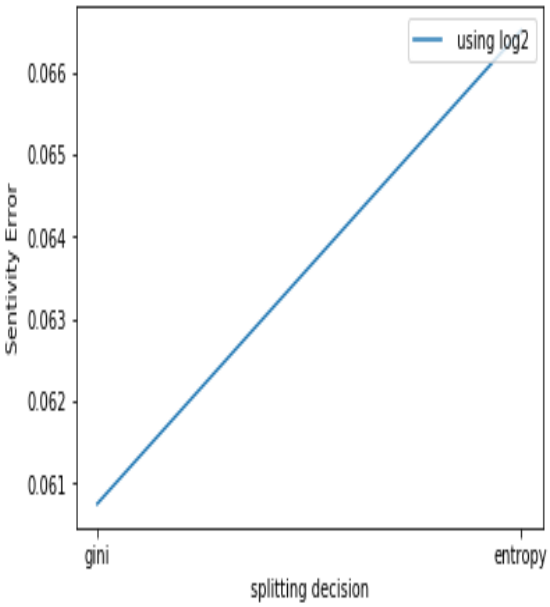
Support vector machine works better in this type of problem because of the feature space location it splits the data into two classes. Only disadvantage it general tends to overfit on training data therefore giving poor performance on test data.

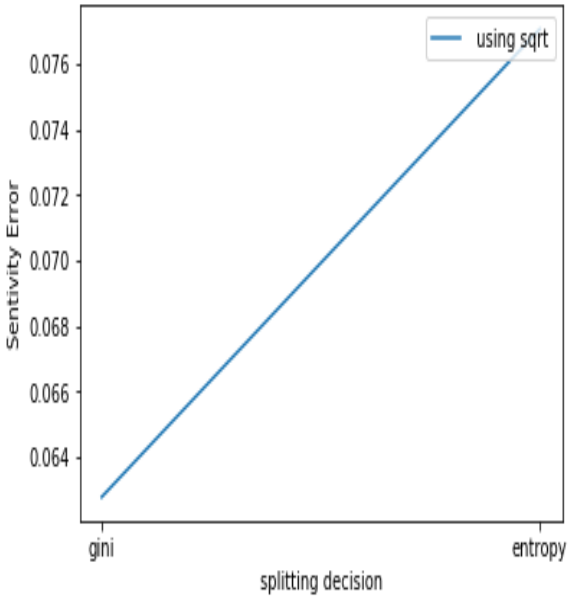
I have use sigmoid kernel and 0.01 gamma value. It takes lot of running time on training data as it calculate distance from classifier.

Decision Tree:

C4.5 works better because of similar reason of random forest but only with single tree therefore it might produce more error than random forest.

I have selected splitting criteria point based on the error obtained in following graph.





Results:

These are the result obtained after applying all the model with accuracy on validation set to test set to predict the classes

Algorithm	Accuracy	Sensitivity
Logistic regression	0.9583125	0.9173333333333333
SVM	0.9719375	0.44542772861356933
Decision_Tree	0.974375	0.44107744107744107
Random_Forest	0.90925	0.936

Comparison between different algorithms:

As you can see that even SVM has highest accuracy but it has lowest sensitivity and same is case with decision tree. Among all these we can see that Random forest outperforms every other model because it has highest sensitivity which we are aiming to predict. Even though having little less accuracy but sensitivity needs to be more as we are targeting to predict positive class which can save money for truck company when they can know truck failure due to APS components.

Random forest doesn't have any data distribution requirement plus it is very versatile flexible and can learn the features and split it according to information loss. All these qualities with no of decision trees formed in random forest make it more robust.

Therefore I would select Random Forest over other models in this scenario.