Kumanan G

PGP - DSBA Online March’20

Capstone Project Notes 2

Model Building and Tuning

Choosing Metric

Before start building machine learning models it is important to decide what is the metric that we need to compare the different models on. In our business problem, we are finding or predicting which customer is going to buy the tourism package. The Buyers are categorised as 1 and non-buyers are categorised as 0. Recall score is chosen as the metric to compare. Recall is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labelled as belonging to the positive class but should have been). Recall Value ranges from 0 to 1.

Recall = $TP/(TP + FN)$

TP - True Positive

FN - False Negative

Understanding Confusion Matrix

A Confusion Matrix is generated for every model to understand the number of the correct and wrong classifications. A confusion matrix has four elements namely

1. True positive - Class 1 being correctly classified
2. True negative - Class 0 being correctly classified
3. False positive - Class 0 wrongly classified as class 1 (A non buyer being predicted as Buyer. It is also known as Type 2 error)
4. False Negative - Class 1 wrongly classified as class 0 (A potential customer predicted as non-buyer known as Type 1 error)

Actual class **Predicted class**	**1**	**0**
1	**TP**	FP
0	FN	**TN**

Since we are predicting our potential customers to buy the travel package, we cannot afford to lose our customer wrongly or making the type 1 error. So a model which has low False Negatives will be an ideal model.

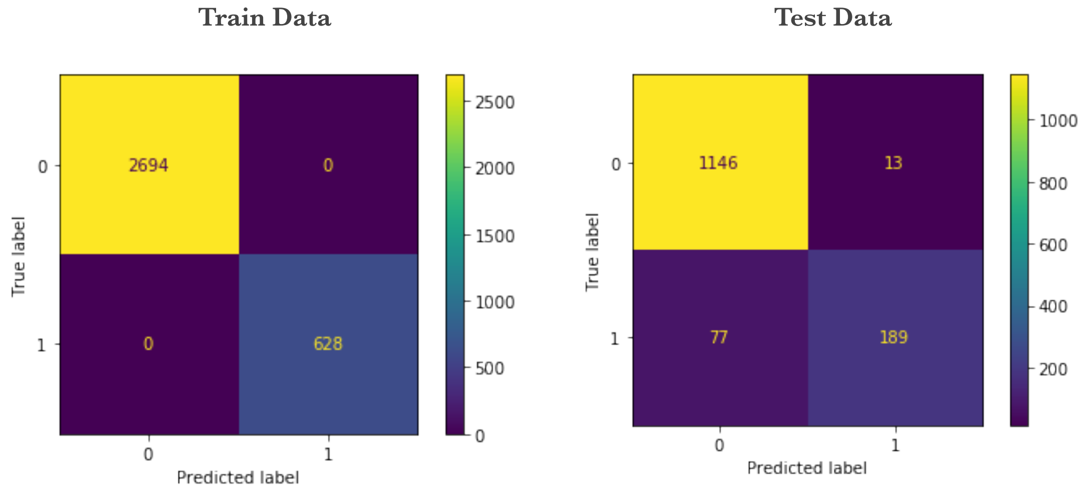
Various Models and Interpretation

Random Forest Classifier

Random Forest model is one of the simplest and powerful classifier model available from sklearn library. Random Forest is based on several decision trees. It is also one of the ensemble models. It got trained very well and predicted all classes right with the training data. But the same model did not perform well with test data. This model classifies 77 customers as non-buyers. The Recall value shows a clear overfitted model.

Training Recall - 1.0

Test Recall.   - 0.71


Train Data

Test Data

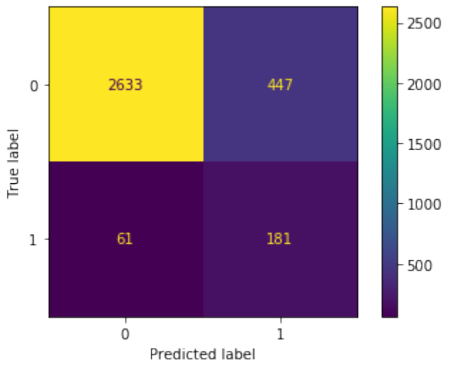
Logistic Regression

In contrast with its name (regression) its actually a classification model. The model available from sklearn's linear model library. Logit model's recall value for training is low compared with Random Forest and Decision Trees. But the model performed consistently with Test Data also.

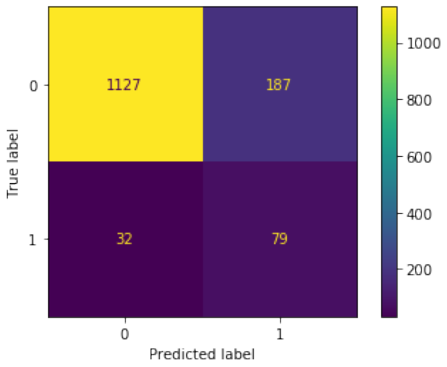
Recall on Train Data - 0.74

Recall on Test Data   - 0.71

Train Data



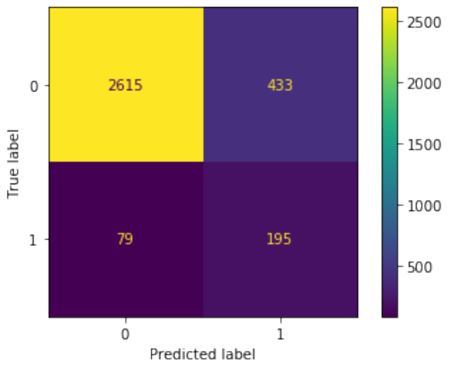
Test Data

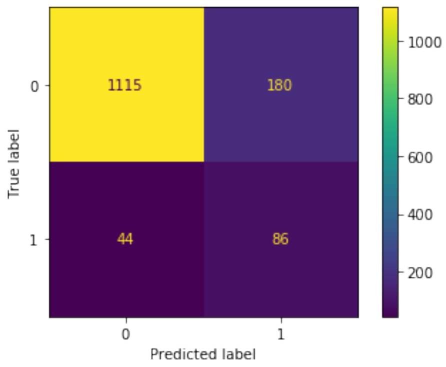


Discriminant Analysis

Linear Discriminant Analysis is another distance based algorithm available from sklearn's Discriminant analysis library. It uses distances between the data points to classify the customers between buyer and non-buyer. This model did not perform well on the test data.

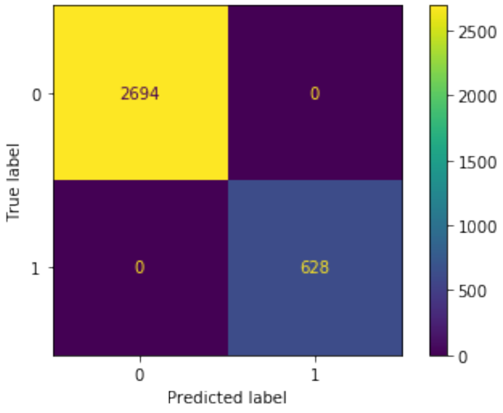
Train Recall - 0.71 Test Recall - 0.66

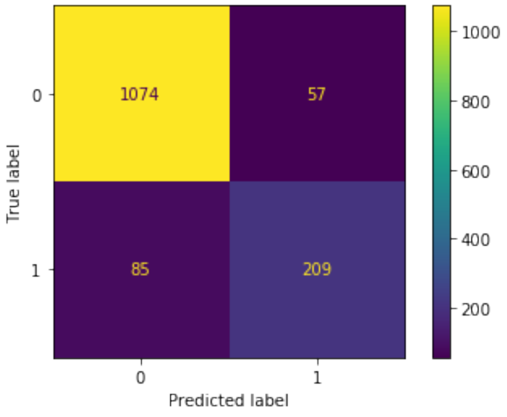




Decision Tree Classifier

Decision Tree model is also known as CART model. Decision Trees split based on gini index. It is a basic also a powerful tool in classification models. This model trained very well like random forest did. But it's a overfitted model since the test recall value is 0.71 while training recall is 1.





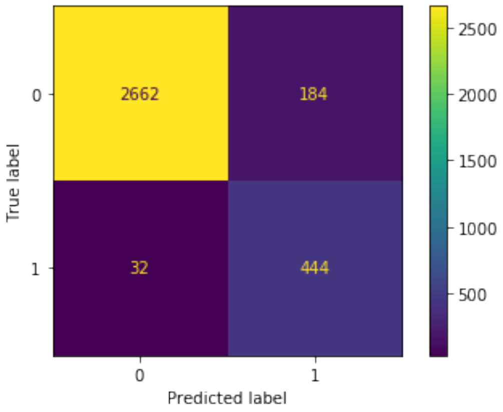
KNN Classifier

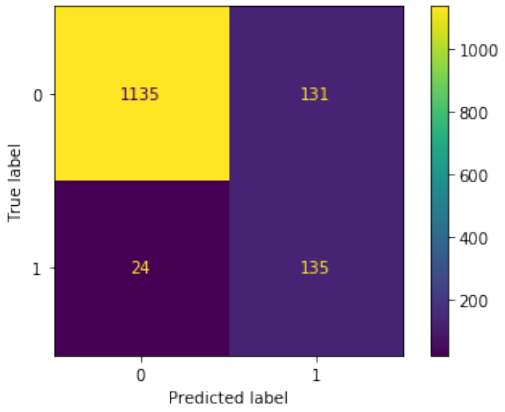
K- Nearest Neighbour classifier is one of the distance based model. It classifies the data based on the majority class of it's neighbouring points. This model gives the best recall value on test data than any other ML models.

Train Recall - 0.93

Test Recall - 0.85

This model is slightly overfitting based on the recall scores. But it's better than decision trees and random forest models.



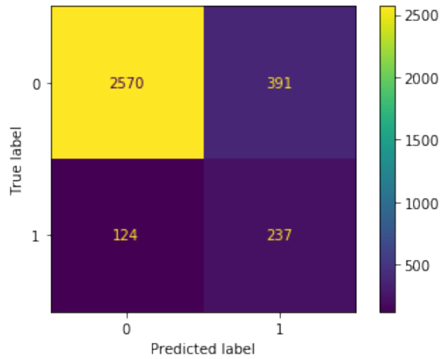


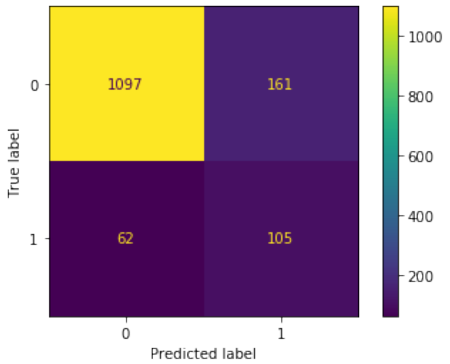
Naive-Bayes Model

It is a probabilistic model based on Bayes theorem. It assumes all the variables are independent. Naive-Bayes performed poorly on both training and test data.

Train Recall - 0.65

Test Recall - 0.62





Performance after Class Balancing

The performance of various models are checked after the minority class is balanced using SMOTE technique. Originally the dataset has 18% of buyers and 82% of non-buyers. So it is probable that models may be biased towards the non-buyers. So the minority class needs to be balanced to be equal with non-buyer's class. Logistic Regression, LDA and Naive Bayes models were rerun using SMOTE data points. But this didn't improve the performance of the models. These models trained slightly better than models with original data points. But on test data these models predicted very poorly.

Logistic Regression Train Recall	- 0.73	Test Recall	- 0.39
LDA Train Recall	- 0.73	Test Recall	- 0.40
Naive - Bayes	- 0.70	Test Recall	- 0.35

So Class Balancing did not work at well with our dataset. It is best to proceed with original datapoint.

Ensemble Modelling

Ensemble modelling is one of the effective tools to get better classification results. It combines various models to get the best result. Random Forest is one of the ensemble models which combines different decision trees to get trained and predicts the classes.

Voting Classifier is another ensemble technique which aggregates the prediction of each classifier and predict the class that gets the most votes. In voting classifier, we used Random Forest, LDA and Logistic Regression classifiers. But this ensemble technique is also not useful for our dataset. Voting classifier's test recall value is 0.40 which is very less than original Logistic Regression or LDA's scores.

Model Tuning

The Recall scores can still be improved by using tuning the various models. GridSearch is one of the technique for model tuning. GridSearch is available from sklearn's model selection library. Model Tuning involves various hyper parameters and value of those needs to be changed based on the output. This involves trial and error method for choosing the right hyper parameter value.

For example Random Forest's hyper parameters are

n_estimators - number of trees to be built

max_depth.   - depth of each trees

max_features. - maximum number of features to used in the training model

We tuned Random Forest, LDA and logistic Regression models. But only LDA performed well on both train and test data. While RF and Logistic models performed very poorly.

RF test recall -   0.28

Logistic test recall - 0.29

LDA test recall -   0.66

Interpretation of Models

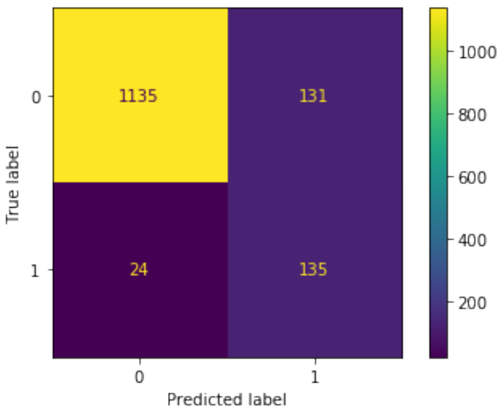
From the above various model built, it is clear that K-NN and Logistic regression works well both on train and test data. Even though Decision trees and random forest have same recall scores as logistic model, they are overfitting. So in our case CART, Random forest are unreliable even though they have decent scores.

The scores of all models built is presented in below table

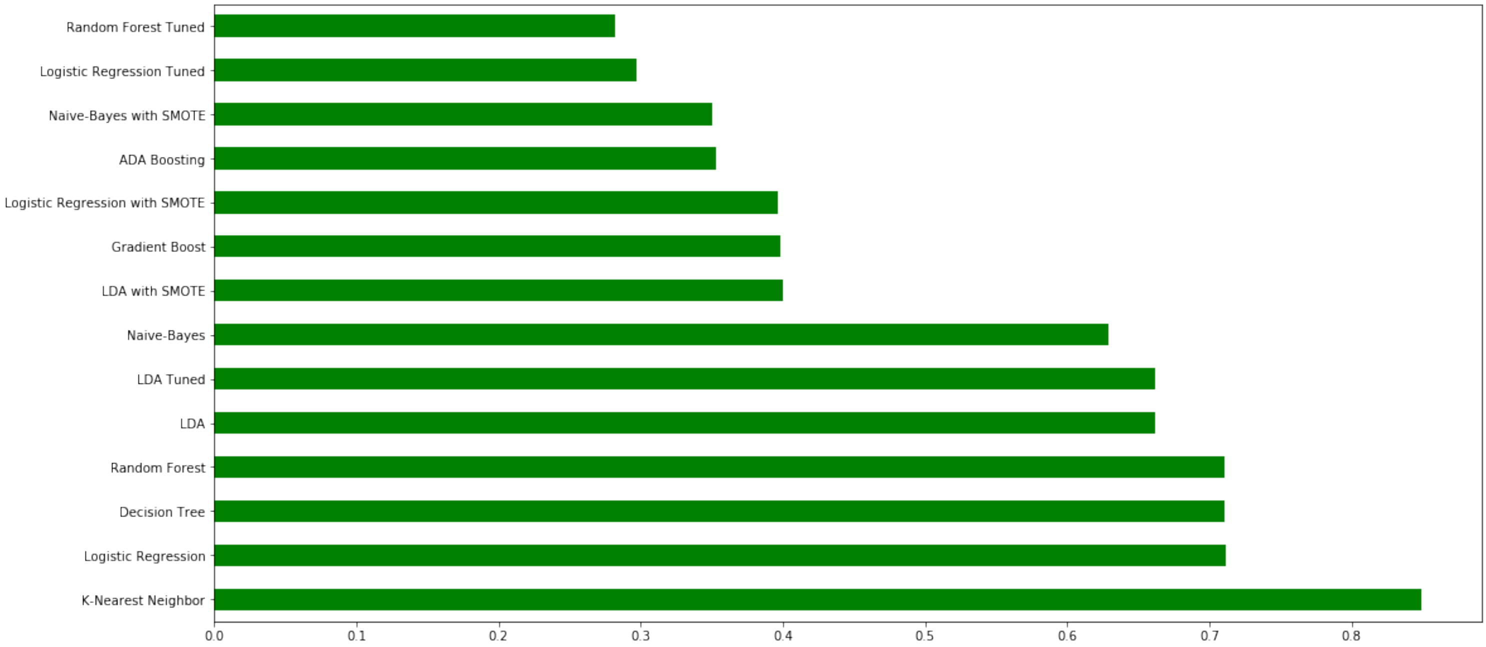
	Train_Recall	Test_Recall	Train_roc_auc	Test roc_auc
K-Nearest Neighbor	0.932773	0.849057	0.982376	0.930041
Logistic Regression	0.747934	0.711712	0.783230	0.800249
Decision Tree	1.000000	0.710884	1.000000	0.856188
Random Forest	1.000000	0.710526	1.000000	0.973110
LDA	0.711679	0.661538	0.782087	0.800285
LDA Tuned	0.711679	0.661538	0.782087	0.800285
Naive-Bayes	0.656510	0.628743	0.779640	0.789798
LDA with SMOTE	0.735283	0.400411	0.796368	0.798867
Gradient Boost	0.452229	0.398496	0.914803	0.875435
ADA Boosting	0.452229	0.398496	0.847234	0.815086
Logistic Regression with SMOTE	0.730916	0.396378	0.796675	0.798819
Naive-Bayes with SMOTE	0.702293	0.350534	0.789159	0.766359
Logistic Regression Tuned	0.288217	0.296992	0.783236	0.800226
Random Forest Tuned	0.332803	0.281955	0.879450	0.859293

KNN :

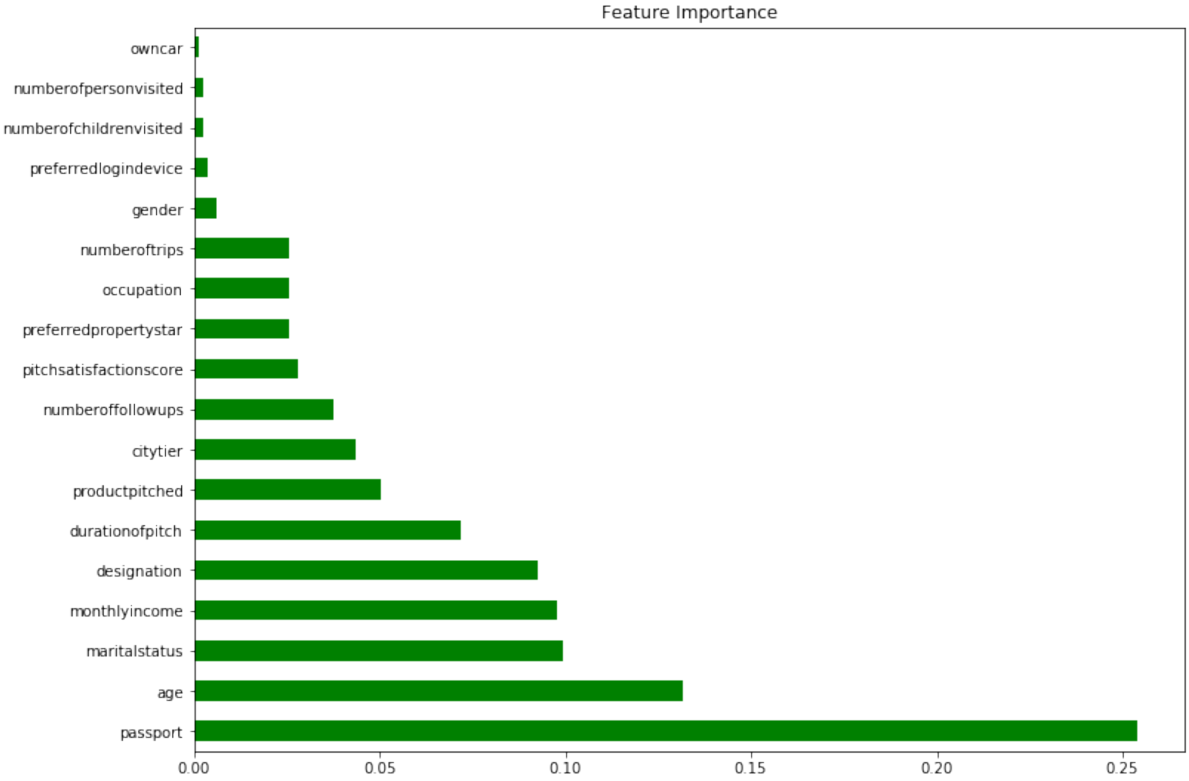
On test data KNN model only predicted 24 customers as non-buyers. By using this model the company loses 24 customers where as other models will make the company lose more customers. So it is best to use K-NN model for further study.



The below shows the Test Recall scores of all models built. KNN model stands out of all models.



The below graph shows the important variables which predicts the buyer. Customer having Passport is an important factor for predicting the buyer.



Thank you !!!