Assignment 2

Applied Machine Learning

Namrata Srivastava – NXS190007

The scope of this project is to implement support vector machines, decision trees and boosting on two datasets and compare the performances and accuracy of each function with the other using various parameters.

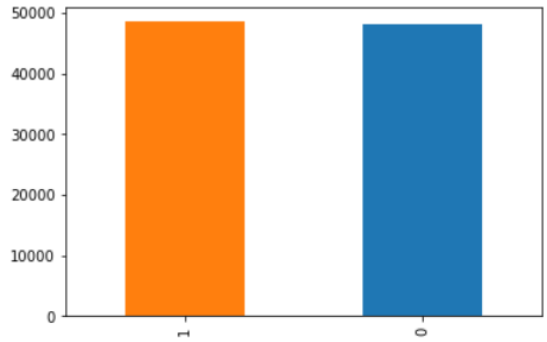
SGEMM CPU Kernel Performance

The dataset is about running time of multiplication of two 2048 matrices consisting 14 parameters. It was downloaded from UCI machine learning repository. There are no missing values, hence data imputation was not required. The dependent variable is "***Run***" which is the average of four runs of the dataset. Binary classification of the dataset has been done by thresholding the same with the target variable's median value.

- ***Median value of the target variable = Run = 69.407***

Hence, target variable's values greater than the median value were classified as 1 and all others were classified as 0.

```
1    48542
0    48098
Name: Run, dtype: int64
```



- 40% of the dataset was randomly selected to perform the support vector machines in order to improve its performance.
- Class 1 – 121255 values of the target variable greater than median value.
- Class 0 – 120345 values of the target variable less than median value.

SUPPORT VECTOR MACHINES

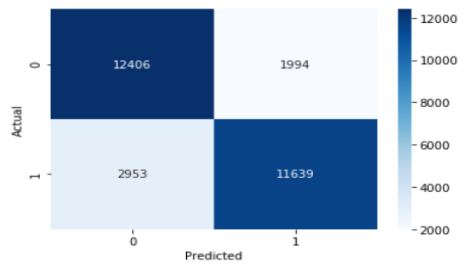
Support vector machine is a supervised machine learning algorithm which is used to classify the datasets. It basically creates a hyperplane which helps in distinguishes classes. There are different types of Kernel present in SVM which helps in improving the accuracy of the classification. Although the kernel productivity is slow, but since classification accuracy is high, these are used as classification algorithms.

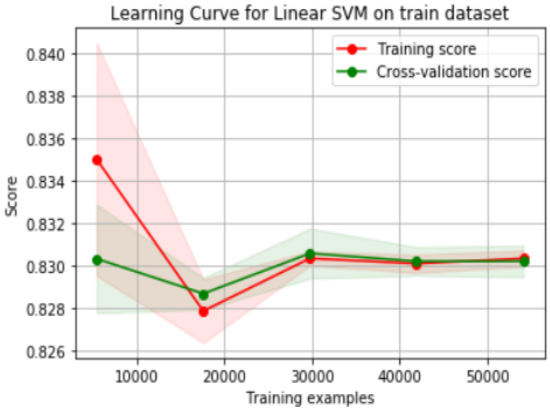
A) LINEAR KERNEL: -

```
0.8247298494242693
              precision    recall  f1-score   support

           0       0.81      0.86      0.83     14400
           1       0.85      0.80      0.82     14592

    accuracy                           0.83     28992
   macro avg       0.83      0.83      0.83     28992
weighted avg       0.83      0.83      0.83     28992
```



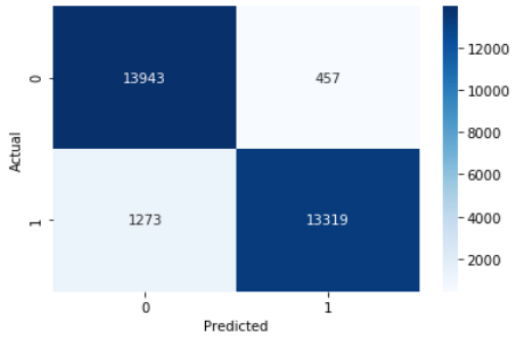


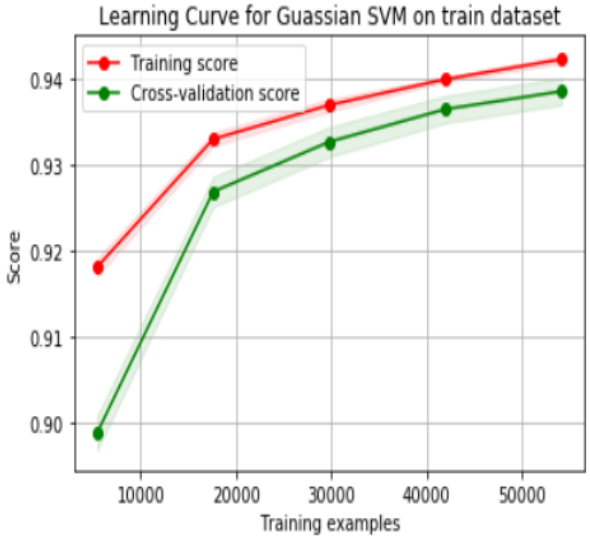
- Accuracy of the linear model is 83% which means 83% of the sample data are linearly separable. Training SVM with Linear kernel is faster than other kernels. Optimisation parameters, C, is 1. Number of observations correctly classified as 1 is 11639 observations.

- Cross-validation is used to avoid overfitting in a model. In cross-validation, a fixed number of folds (or partitions) of the data is created and analysis is done on each fold, and then average the overall estimate.

B) GUASSIAN KERNEL: -

	precision	recall	f1-score	support
0	0.92	0.97	0.94	14400
1	0.97	0.91	0.94	14592
accuracy			0.94	28992
macro avg	0.94	0.94	0.94	28992
weighted avg	0.94	0.94	0.94	28992

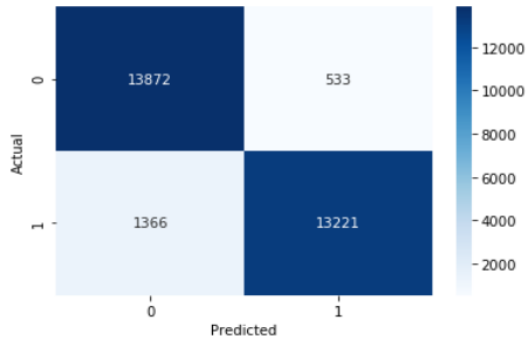


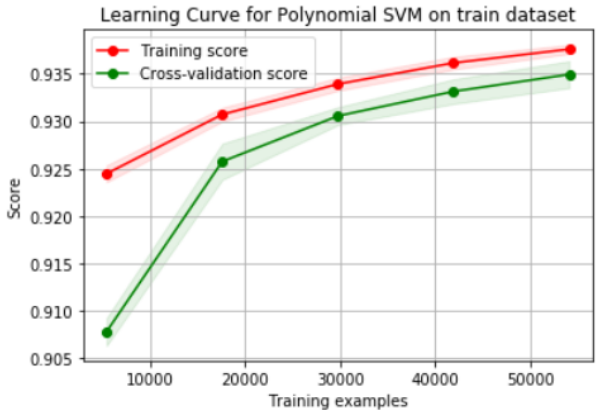


- Gaussian kernel is a kernel with the shape of a Gaussian (normal distribution) curve. Accuracy of the gaussian model is 94.03%. Optimisation parameters, C, is 1. Number of observations correctly classified as 1 is 13319 observations.
- It overcomes the drawback of linear kernel by showing an improved performance of identifying the class 1 features and thus increasing accuracy to approximately 94%.

C) POLYNOMIAL KERNEL: -

	precision	recall	f1-score	support
0	0.91	0.96	0.94	14405
1	0.96	0.91	0.93	14587
accuracy			0.93	28992
macro avg	0.94	0.93	0.93	28992
weighted avg	0.94	0.93	0.93	28992





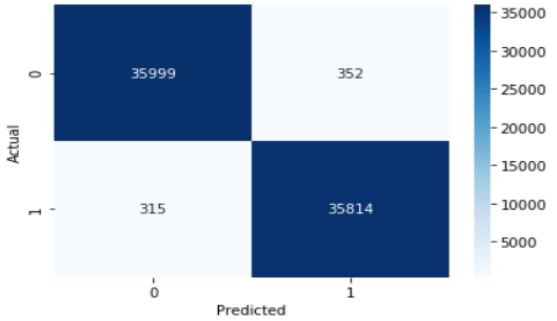
- Accuracy of the polynomial model is 93.6%. Optimisation parameters, C, is 1. Number of observations correctly classified as 1 is 13236 observations.
- It overcomes the drawback of linear kernel by showing an improved performance of identifying the class 1 features and thus increasing accuracy to approximately 93.6%

DECISION TREES

Decision tree classifier is used to split the data multiple times according to certain cut-off values in the features. It is an unpruned tree which can be large. Pruning on decision trees is done by giving the max features, criteria, random state and leaf nodes to control memory consumption, complexity and size of the trees.

A) DECISION TREE: -

	precision	recall	f1-score	support
0	0.99	0.99	0.99	36351
1	0.99	0.99	0.99	36129
accuracy			0.99	72480
macro avg	0.99	0.99	0.99	72480
weighted avg	0.99	0.99	0.99	72480

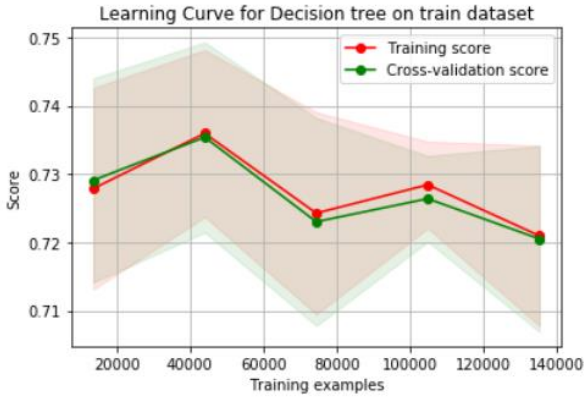


Accuracy of decision tree without pruning is 99%.

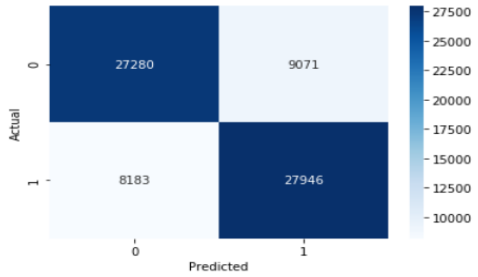
PRUNING: Both Gini and impurity are measures of impurity of a node. A node having multiple classes is impure whereas a node having only one class is pure. Entropy signifies the disorder in the classification of the dataset. Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen.

B) DECISION TREE WITH GINI IMPURITY: -

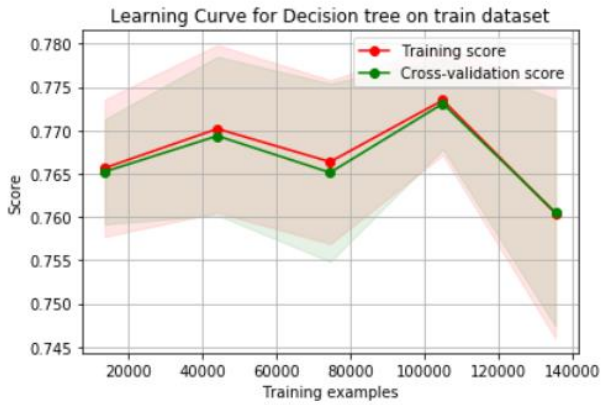
Accuracy of decision tree with Gini impurity is 76%.



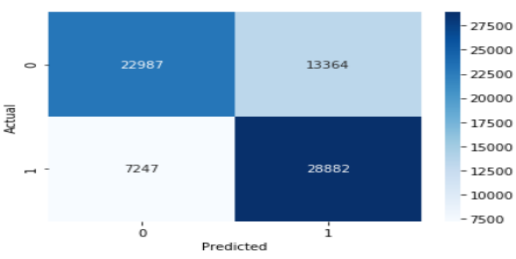
	precision	recall	f1-score	support
0	0.77	0.75	0.76	36351
1	0.75	0.77	0.76	36129
accuracy			0.76	72480
macro avg	0.76	0.76	0.76	72480
weighted avg	0.76	0.76	0.76	72480



C) DECISION TREE WITH ENTROPY: -



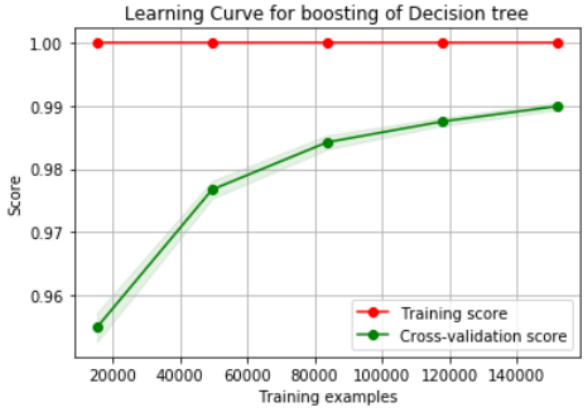
	precision	recall	f1-score	support
0	0.76	0.63	0.69	36351
1	0.68	0.80	0.74	36129
accuracy			0.72	72480
macro avg	0.72	0.72	0.71	72480
weighted avg	0.72	0.72	0.71	72480



Accuracy of decision tree with Gini impurity is 72%.

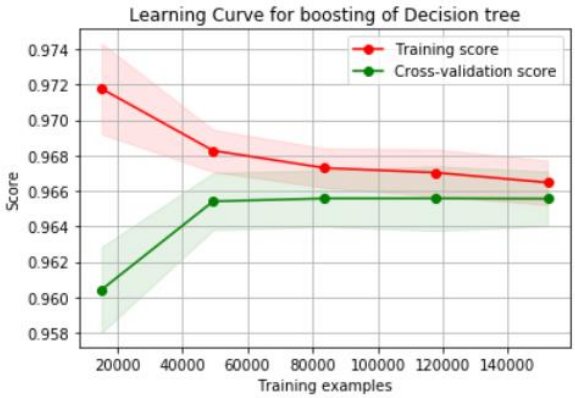
BOOSTING

Boosting with Decision Tree



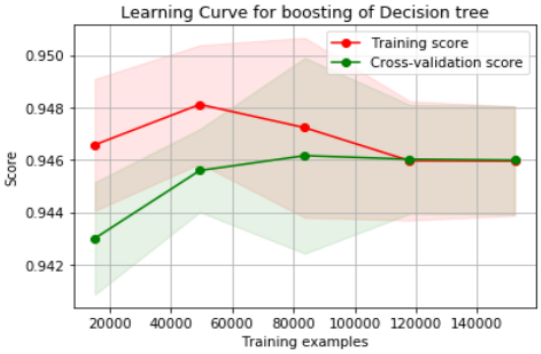
As the training sample size increases, bias variance trade-off decreases between the training and cross validation. But the model is not able to represent the underlying relationship. It is a poor fit.

Boosting with Pruning – Gini Impurity



Bias-variance trade-off decreases between the training and cross validation indicating that there exists a relationship and is a good fit to the model.

Boosting with Pruning – Entropy

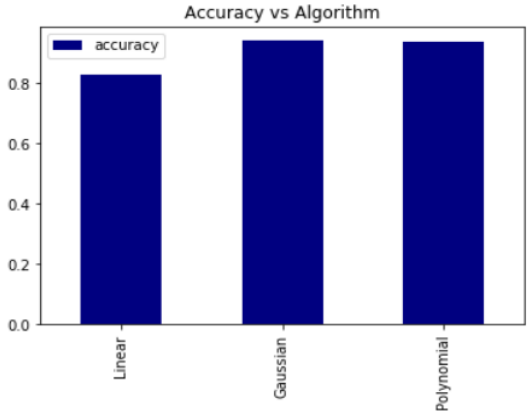


Bias-variance trade-off decreases and merges indicating that it is between the training and cross validation indicating that there exists a relationship and is a good fit to the model.

DISCUSSION

Comparison between Kernel Accuracies:

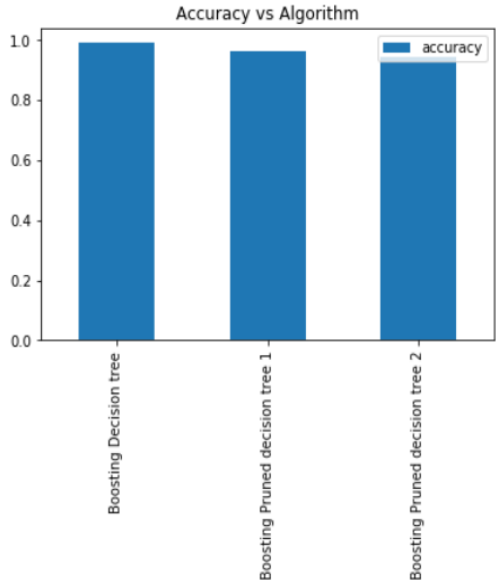
	accuracy
Linear	0.828194
Gaussian	0.940639
Polynomial	0.934499



The highest accuracy is given by gaussian kernel with an accuracy percentage of 94.06%.

Comparison between Decision Trees with and without Pruning

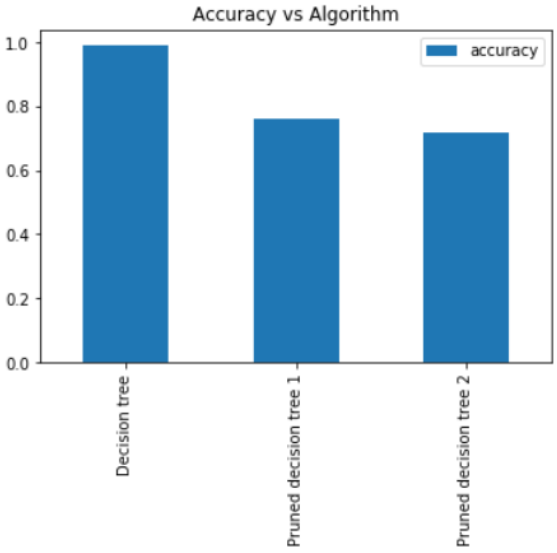
	accuracy
Boosting Decision tree	0.990977
Boosting Pruned decision tree 1	0.964100
Boosting Pruned decision tree 2	0.944178



The highest accuracy of 99% is in case of decision trees where no pruning is done and the decision tree is made with the whole dataset.

Comparison between Boosting trees with and without Pruning

	accuracy
Decision tree	0.990797
Pruned decision tree 1	0.761948
Pruned decision tree 2	0.715632



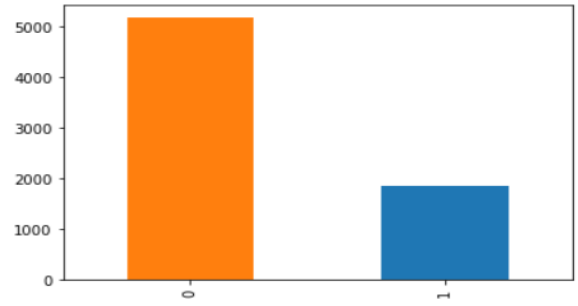
The highest accuracy of 99% in boosting is in case of decision trees where no pruning is done and the decision tree is made with the whole dataset.

SECOND DATASET – TELCO CHURN RATE

This dataset has been selected from Kaggle website, which shows the churn rate of customers for Telco companies. The data set includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

```
0    5174
1    1869
Name: Churn, dtype: int64
```

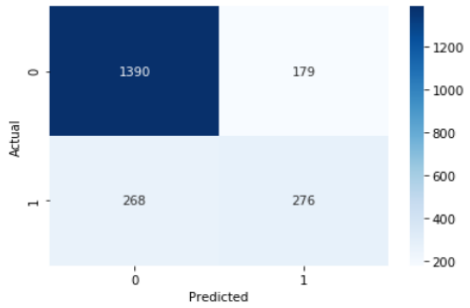


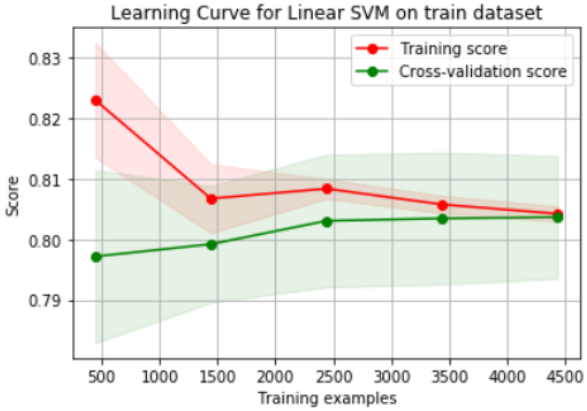
Through this dataset we can understand possible customer data and customer retention programs and predict behaviours to retain customers.

Target Variable is: Churn

LINEAR KERNEL

	precision	recall	f1-score	support
0	0.84	0.89	0.86	1569
1	0.61	0.51	0.55	544
accuracy			0.79	2113
macro avg	0.72	0.70	0.71	2113
weighted avg	0.78	0.79	0.78	2113

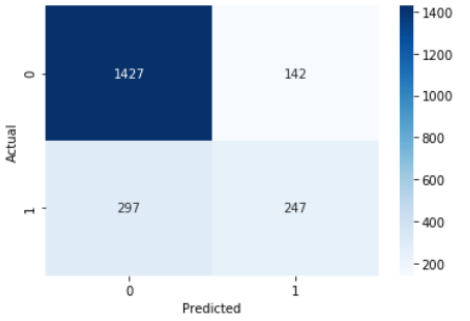


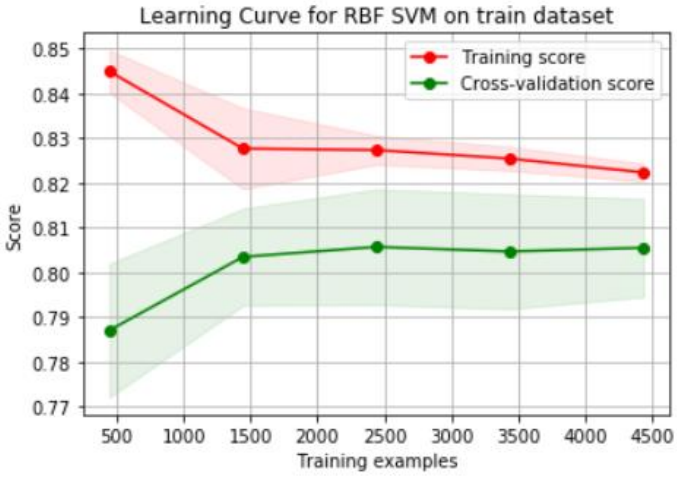


- Accuracy of the linear model is 79% which means 79% of the sample data are linearly separable. Training SVM with Linear kernel is faster than other kernels. Optimisation parameters, C, is 1. Number of observations correctly classified as 1 is 1390 observations.
- Cross-validation is used to avoid overfitting in a model. In cross-validation, a fixed number of folds (or partitions) of the data is created and analysis is done on each fold, and then average the overall estimate.

RBF KERNEL

	precision	recall	f1-score	support
0	0.83	0.91	0.87	1569
1	0.63	0.45	0.53	544
accuracy			0.79	2113
macro avg	0.73	0.68	0.70	2113
weighted avg	0.78	0.79	0.78	2113



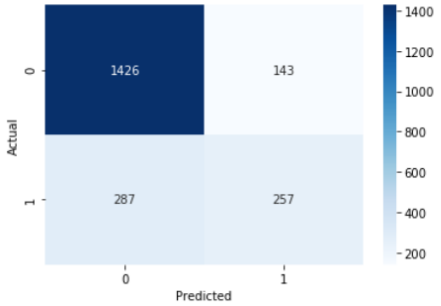


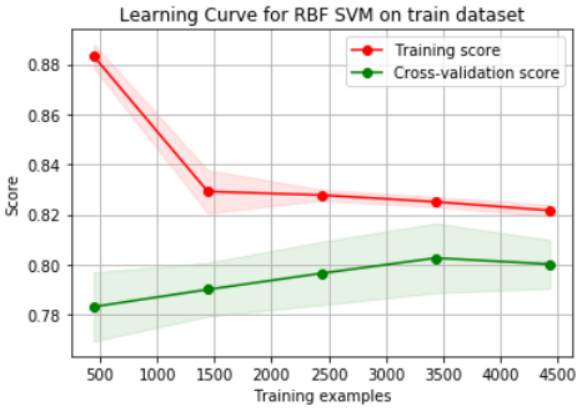
- Gaussian kernel is a kernel with the shape of a Gaussian (normal distribution) curve. Accuracy of the gaussian model is 79%. Optimisation parameters, C, is 1. Number of observations correctly classified as 1 is 1427 observations.

POLYNOMIAL KERNEL

- Accuracy of the polynomial model is 80%. Optimisation parameters, C, is 1. Number of observations correctly classified as 1 is 1426 observations.

	precision	recall	f1-score	support
0	0.83	0.91	0.87	1569
1	0.64	0.47	0.54	544
accuracy			0.80	2113
macro avg	0.74	0.69	0.71	2113
weighted avg	0.78	0.80	0.79	2113

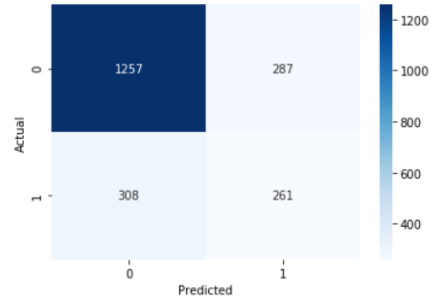


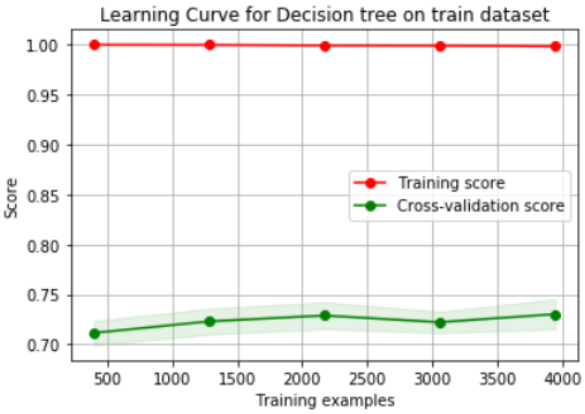


DECISION TREES

A) DECISION TREE: -

	precision	recall	f1-score	support
0	0.80	0.81	0.81	1544
1	0.48	0.46	0.47	569
accuracy			0.72	2113
macro avg	0.64	0.64	0.64	2113
weighted avg	0.72	0.72	0.72	2113

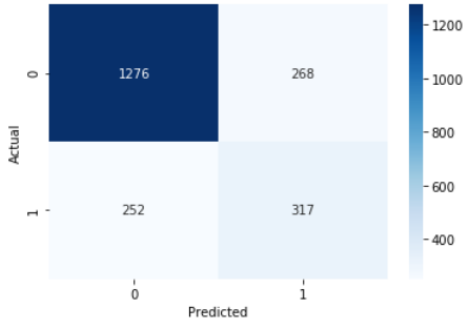


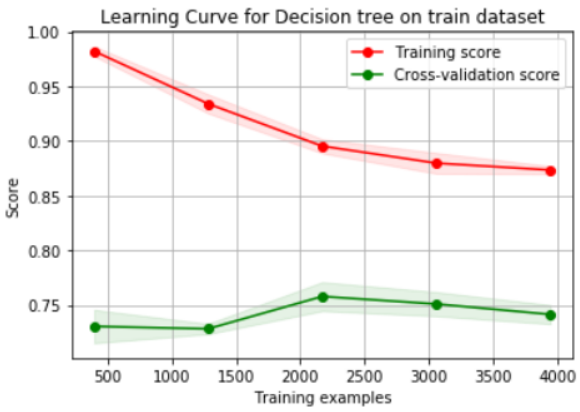


- Accuracy for Decision Tree is 72%.
- Learning curve indicates a poor fir in the model with high bias variance trade-off.

B) DECISION TREE WITH GINI IMPURITY: -

	precision	recall	f1-score	support
0	0.84	0.83	0.83	1544
1	0.54	0.56	0.55	569
accuracy			0.75	2113
macro avg	0.69	0.69	0.69	2113
weighted avg	0.76	0.75	0.75	2113

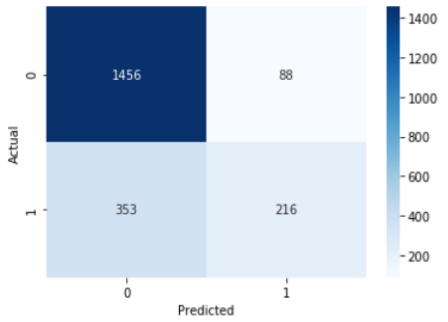


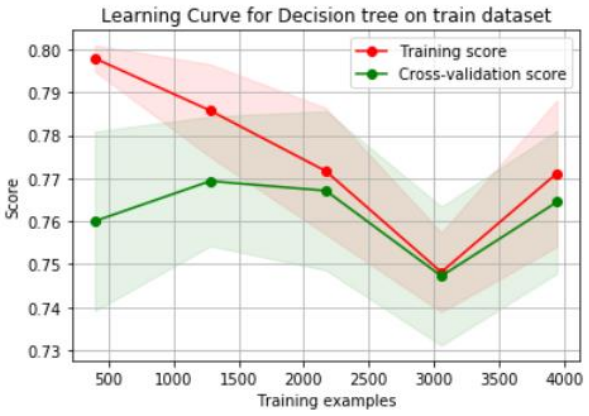


- Accuracy for Decision Tree after pruning with Gini Impurity is 75%.
- Learning curve indicates a poor fir in the model with high bias variance trade-off.

C) DECISION TREE WITH ENTROPY: -

	precision	recall	f1-score	support
0	0.80	0.94	0.87	1544
1	0.71	0.38	0.49	569
accuracy			0.79	2113
macro avg	0.76	0.66	0.68	2113
weighted avg	0.78	0.79	0.77	2113

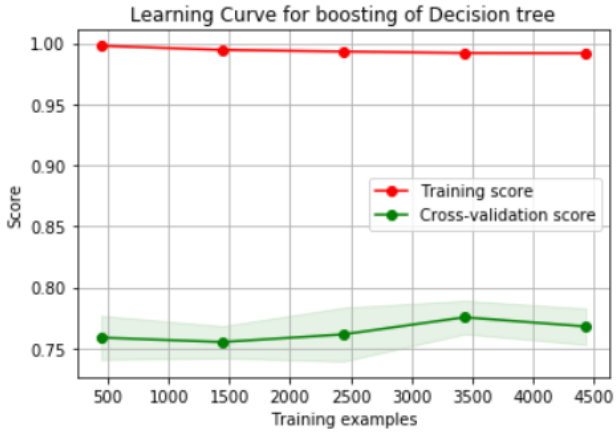




- Accuracy for Decision Tree after pruning with Entropy is 79%.
- Learning curve indicates a poor fir in the model with high bias variance trade-off. More flexibility

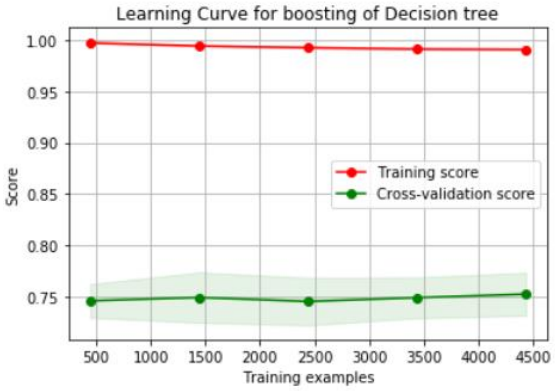
BOOSTING

Boosting with Decision Tree

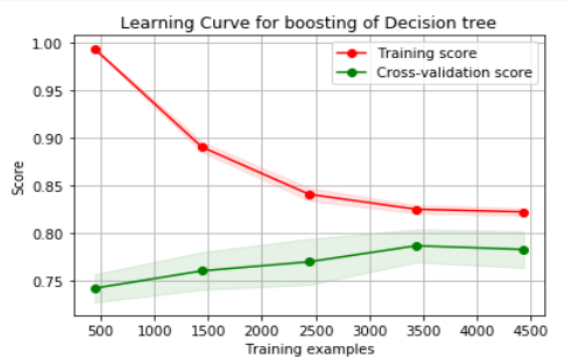


As the training sample size increases, bias variance trade-off is not decreasing between the training and cross validation. The trade-off started improving but could no move further.The model is not able to represent the underlying relationship. It is a poor fit model. Adding more samples of data might improve the model.

Boosting with Pruning – Gini Impurity



As the training sample size increases, bias variance trade-off is not decreasing between the training and cross validation. The model is not able to represent the underlying relationship. It is a poor fit model. More variables and number of observations might help us in getting useful result.

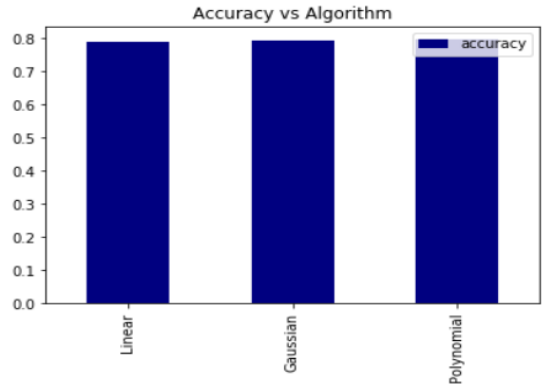


Boosting with Pruning – Entropy

This model is said to be underfitting and has a high bias. By adding more features, we can increase the flexibility of the model.

DISCUSSION

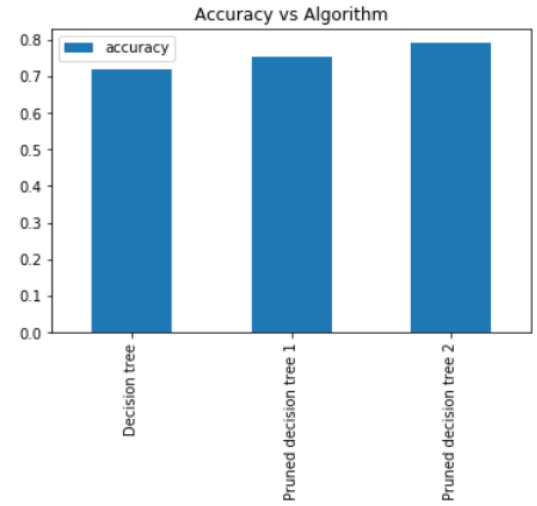
	accuracy
Linear	0.788452
Gaussian	0.792239
Polynomial	0.796498



The highest accuracy is given by gaussian kernel with an accuracy percentage of 94.06%.

Comparison between Decision Trees with and without Pruning

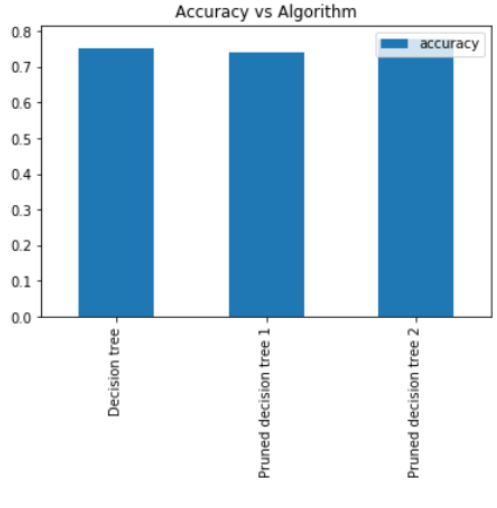
	accuracy
Decision tree	0.718410
Pruned decision tree 1	0.753904
Pruned decision tree 2	0.791292



The highest accuracy of 79% is in case of decision trees when criteria is entropy.

Comparison between Boosting trees with and without Pruning

	accuracy
Decision tree	0.752485
Pruned decision tree 1	0.740653
Pruned decision tree 2	0.778514



The highest accuracy of 79% is in case of decision trees when criteria is entropy.