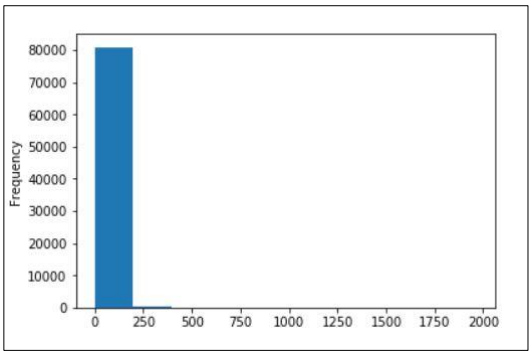
Problem Statement: It is desired to classify the Facebook Comment volume into high and low based on a chosen threshold. The parameters chosen are Page likes, Page Checkins, The total number of comments before selected base date/time, The number of comments in last 24 hours, relative to base date/time, The number of comments in last 48 to last 24 hours relative to base date/time, The number of comments in the first 24 hours after the publication of post, Base time, Post length, Post Share Count, H Local and Post Published Weekday. In this problem, accuracy is chosen to be the metric of interest as it is essential to be accurate in classifying posts as high and low volume as these are then usually set as bench marks for online marketing.

About Facebook Data Set:
As most of our data is skewed towards 0 comments, it would be advisable to use our understanding of the business case to threshold the data. For the scope of this project however, we threshold the data at 15 comments.



Frequency Distribution of Target

	Target	TargetClass
count	81312.000000	81312.000000
mean	7.190611	0.085436
std	36.049374	0.279532
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	3.000000	0.000000
max	1966.000000	1.000000

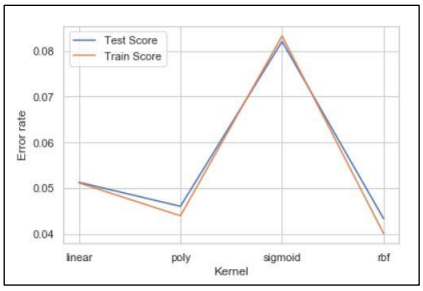
Statistical description of Target

In line with industry best practices and the theory available around choosing a reliable model for our data, it is decided to use 3-fold cross validation to choose the model.

Benefits:

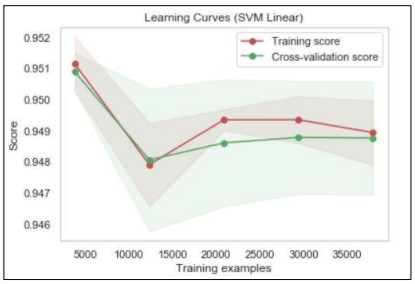
- All the examples in the dataset are eventually used for both training and testing. Finally, the true error is estimated as the average error rate
- 3 folds ensure low computing costs

SVM:
To get a general feel of how our data yields to various classification methods, we perform a *Zero One loss analysis* of linear, polynomial, sigmoid and radial basis function(rbf) kernels without cross validation.



We see that the sigmoid function has extremely high error rates and hence we decide to perform further analysis on linear, polynomial and rbf kernels.

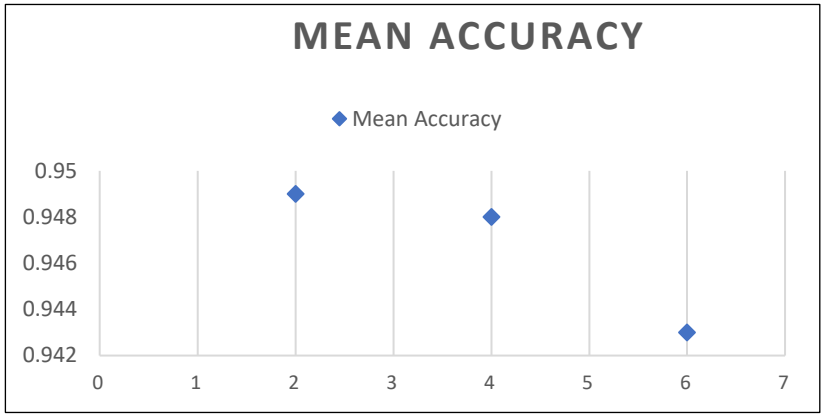
a. Linear Kernel:



On performing exhaustive grid search on the hyper parameter C between values of 0.001 and 10, we find that for C =10, the model performs the best under in the linear kernel of SVM. we see that the train and cross validation errors decrease and converge to comparable levels in the range 0.9475 to 0.95. The average accuracy is 94.8%.

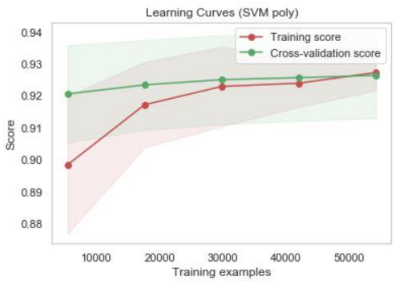
b. **Polynomial Kernel:**

Degree of the polynomial dictates the flexibility of the model. We experiment with degree 2, 4 and 6.

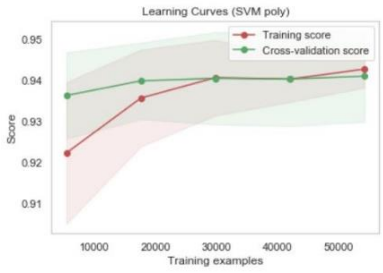


We calculate the average accuracy for the 3 folds in each of the models. The results obtained are as plotted. Thus we expect the highest accuracy to be between 2 and 4 degrees.

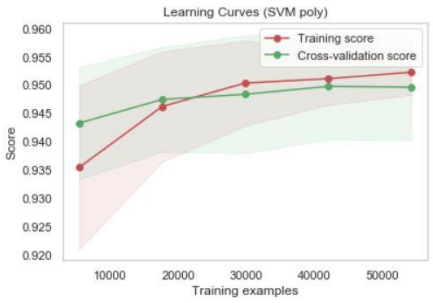
The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. Thus, we experiment with the values of 0.001, 0.05, 0.1 and 1. The accuracy v/s sample size curves obtained are as below:



Poly kernel, Degree 2, C 0.001

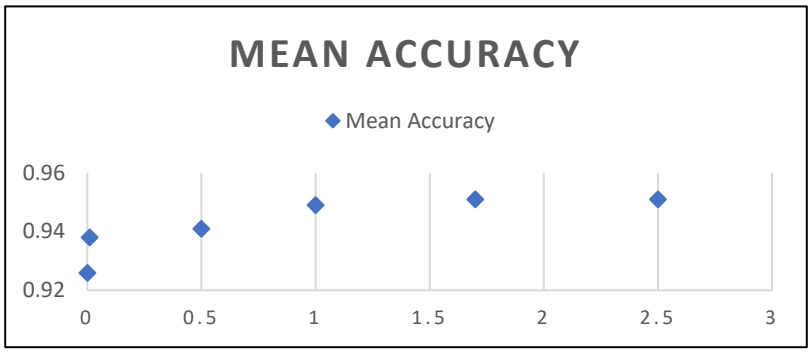


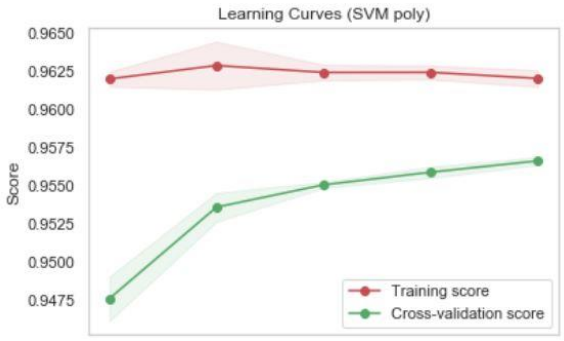
Poly kernel, Degree 2, C 0.1



Poly kernel, Degree 2, C 1

On plotting the average accuracy for the k folds for each of these c values at degree =2, we get the graph as plotted:
The accuracy increases with C. Thus, we expect higher values of C to give us our best model.

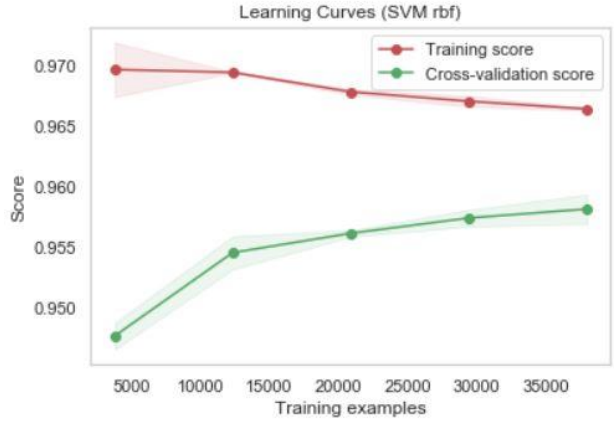




On experimenting with different Degrees and C values, we choose a model with Degree 3 and C 10 as it provides an average k fold validates accuracy of 95.6%.

The learning curve is as plotted.

c. **Radial Basis Function:**



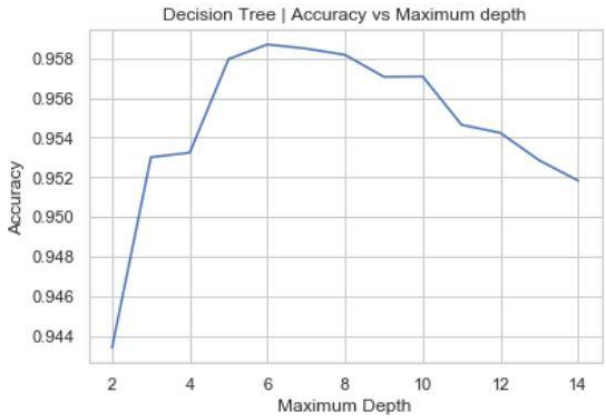
On performing exhaustive grid search on the hyper parameter C and gamma values between values of 0.001 and 10 and 0.001 and 1 respectively, we find that for C =10 and gamma =0.1, the model performs the best under in the radial basis function kernel of SVM. The average accuracy is 95.8%.

SUMMARY OF SVM:
Thus, we can say that all kernels give comparable validation performance when tuned.

SVM Kernel	Average Score for k folds
Linear, c=10	94.8%
Polynomial, deg=3, c=10	95.6%
Rbf C=10, gamma=0.1	95.8%

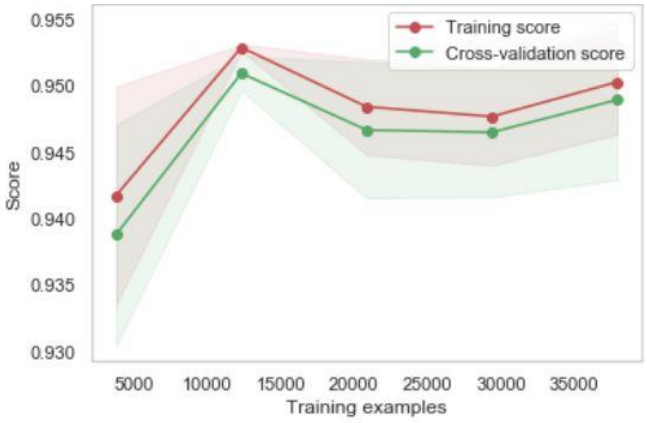
Decision Trees:

Decision trees are widely used and the most intuitive for classification problems. Here we experiment with depth and number of leaves of the tree



On plotting the accuracy against various depths at which the tree was pruned, we observe that our most accurate version of the model was achieved at depth of 6. It dwindles after that due to over fitting.
Thus we pick depth of 6 for further experimentation.

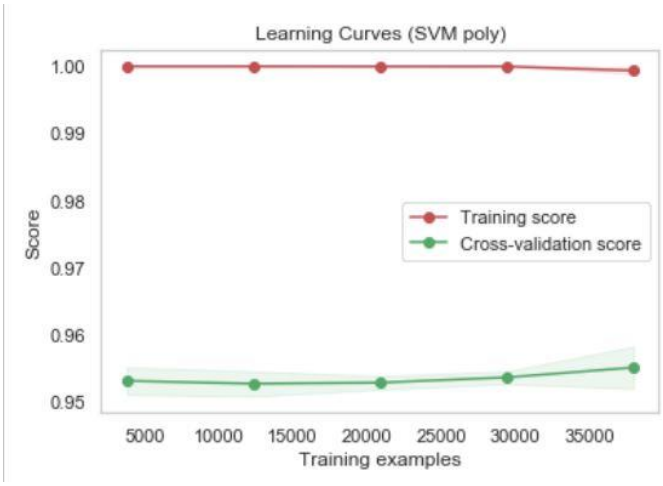
On experimenting with various hyperparameters like depth of tree, number of leaves number of splits etc, we find the model with the following hyperparameters to give us best accuracy: max_depth=2, max_leaf_nodes = 5, min_samples_leaf=5, min_samples_split=2



Based on our experimentation, we have picked the version of our decision tree having maximum depth of 6 and maximum leaves 8. The average accuracy of this is model with k fold validation is 94.9%.

Boosting:

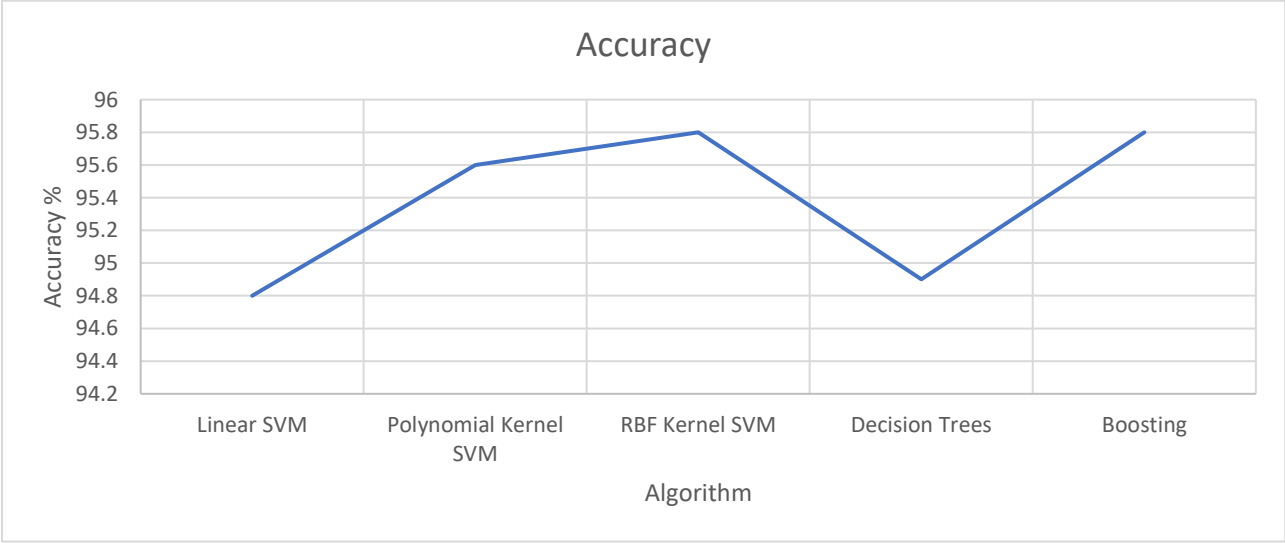
Using ensemble methods, we can increase the accuracy of our model by fine tuning the hyper parameters. We chose to work with AdaBoost Algorithm for this dataset. On passing decision trees to the model and fine tuning the learning rate and tree depth, we obtain the following learning curve:



On computation of the accuracy for We see that the validation error for AdaBoost Algorithm is around 94.8%.

As expected, boosted decision trees perform better than decision trees.

Accuracy Performance:



As in this problem, prediction of Facebook comment volume will help decide the right time, day and other such features of online marketing, the accuracy plays an important role here in predicting a post to be high comment or low comment volume. Thus, we can pick boosted trees as they give high accuracy and ensure good classification.

Bank Marketing Data:

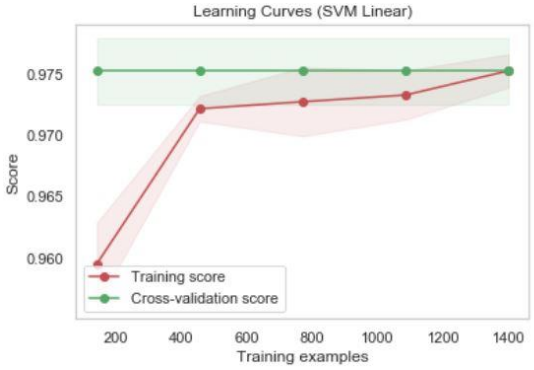
The data is sourced from UCI Machine Learning Repository. It has data around the efficiency of the marketing campaign launched by the bank to acquire more term deposits. This is an interesting dataset as it provides an opportunity to work on predictive marketing analytics and could be used to fine tune targeted marketing in the future. As for the scope of this project, the focus is on correctly classifying the unseen test data with good accuracy.

Input variables:
Age (numeric)
Default: has credit in default? (categorical: 'No', 'Yes', 'Unknown')
Duration-last contact duration, in seconds (numeric)

Output variable (desired target):
y - has the client subscribed a term deposit? (binary: 'Yes', 'No')

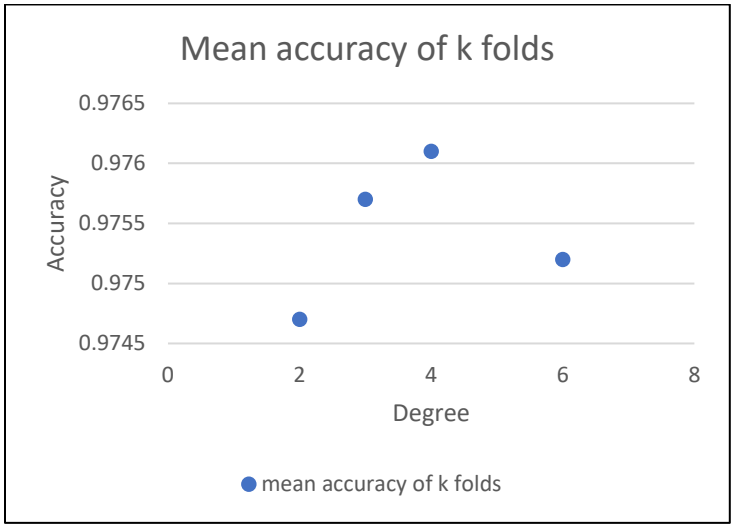
SVM:

a) Linear Kernel:



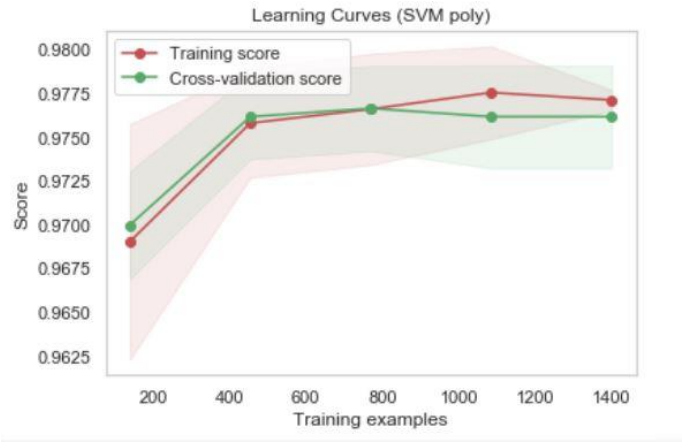
On performing GridSearch with Linear Kernel and experimenting with various values of C, the best value of C is chosen as 0.001 giving an average accuracy of 97.5 with 3 folds of cross validation.

b) Polynomial Kernel:

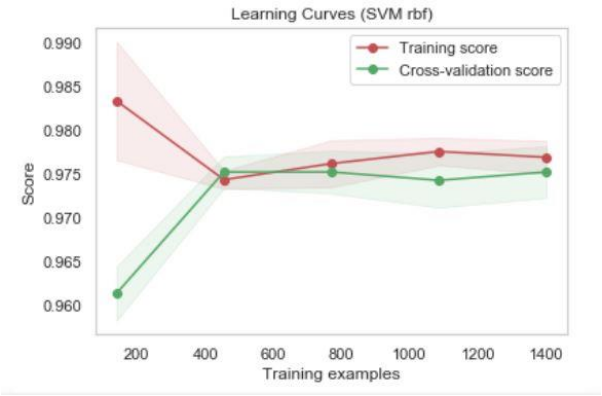


We plot the average accuracy of the k folds to find that a polynomial kernel of degree 4 provides maximum validation accuracy.

On performing experimentation with various values of C, we find that soft bound of 1 provides an accuracy of 97.6 for degree 4. The learning curve is as plotted below.

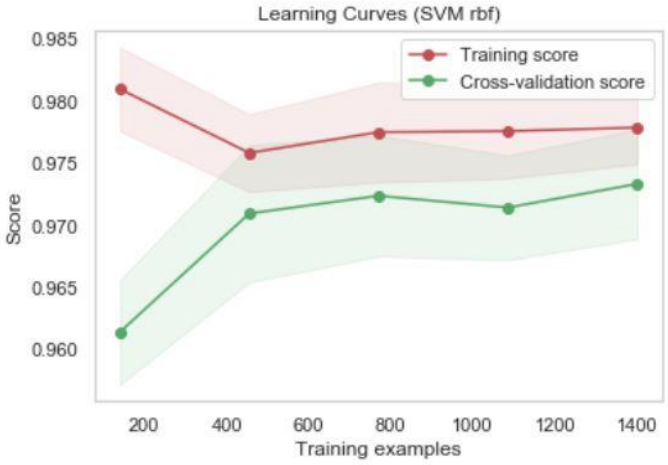


c) RBF Kernel:



On performing experimentation with various values of C and gamma, we find that soft bound of 10 and gamma of 0.1 provides an accuracy of 97.57. The learning curve is as plotted.

Decision Trees:



On performing experimentation with various values of criterion, maximum depth, maximum leaf nodes min samples leaf and min sample split, we chose model with the plotted learning curve. The accuracy for this model is 96.9

Boosted Trees:

On performing experimentation with various values of depth, for boosted decision trees. We find the model with depth 2 provides highest classification accuracy of 97.2%

Summary:

Algorithm	Accuracy (%)	Error (%)
SVM-Linear	97.5	2.5
SVM-Polynomial	97.6	2.4
SVM-RBF	97.5	2.4
Decision Trees	96.9	3.1
Boosted Decision Trees	97.2	2.8

As in this business case, it is important to accurately classify the if the targeted customer will invest in the deposit with the bank, as it forms the basis of further marketing strategies and helps banks plan their resources accordingly. Thus, we choose accuracy as our prime metric to judge the performance of each algorithm. In this case, all the chosen algorithms give good accuracy however, the polynomial kernel gives the best accuracy. Thus, we choose it as our selected model. Plotted below is the confusion matrix for polynomial kernel of SVM.

	Predicted (1)	Predicted (0)
Actual (1)	**871**	**4**
Actual (0)	**24**	**1**