Applied Machine Learning
Assignment 2

Sai Vishnu Kanisetty SXK175300

Objective of this exercise:
Experimentation of classification algorithms (SVM, Decision Tree and Boosting) on two problems:

- Predict whether the articles published by Mashable will be high shared or low shared
- Predict the Success of Bank Telemarketing

Index: This exercise has been implemented in both Python (Data Cleaning & SVM) and R (Decision Tree & Ensemble). Different topics discussed in this article are:

1. Definition for different algorithms used and their hyper parameters
2. Data Understanding for both the datasets
3. Results of experimentation for Mashable problem
4. Results of experimentation for Telemarketing
5. Summary using both the datasets
6. Next Step to improve the model accuracy

Step 1: Definition for different algorithms used and their hyper parameters

For both the datasets, Accuracy has been considered as the metric to be maximized.

Support Vector Machines: Without using cross validation, RBF kernel with cost = 1 and gamma =1 gives a training accuracy of 100%, whereas the test accuracy is only ~50%, which calls a need for cross validation during model selection. The approach and results without cross validation is not mentioned in the report but the code is available in python notebook.

- Three kernels have been used in this exercise: Linear kernel, Radial Bias Functional Kernel and Sigmoid kernel
- Kernel function and hyperparameter
    - **Linear Kernel:** Definition: $\langle x, x' \rangle$
        - Hyperparameters:
            - Cost: Penalty parameter C of the error term. Range of the values experimented with [0.1,1,10,100]
    - **Radial Bias Functional kernel:** Definition: $\exp(-\gamma \|x - x'\|^2)$
        - Hyperparameters:
            - Cost: Range of the values experimented with [0.1,1,10,100]
            - Gamma: Range of the values experimented with [0.1,1,10,100]
    - **Sigmoid:** Definition: $\tanh(\gamma \langle x, x' \rangle + r)$
        - Hyperparameters:
            - Cost: Range of the values experimented with [0.1,1,10,100]
            - Gamma: Range of the values experimented with [0.1,1,10,100]
            - Coef0: Range of the values experimented with [0.0,0.5]
- The algorithm which gives the best three-fold cross validation accuracy among all the above experiments has been considered as the best solution of SVM

- Learning curves of test and train accuracy by varying the number of observations has been used to understand the parameters facing issues of bias and variance

Decision Tree:

- Owing to the outliers present in the data, Classification and Regression using the towing as the split criteria, which can handle the outliers better than ID3 and C5.0 [1], has been used in this exercise.
- Hyperparameters for pruning:
    - Cp: Complexity parameter. Any split that does not decrease the overall lack of fit by a factor of CP is not attempted. Range of the values experimented with [0.05,0.01,0.005,0.001,0.0005,0.0001]
- Tree is fully grown using a very low Cp value and the train and test accuracy are identified
- Pruning is performed as long as the test data accuracy of the pruned tree is less then using an unpruned tree. Pruning helps to identify the Cp value (number of leaves) which gives the highest test accuracy.

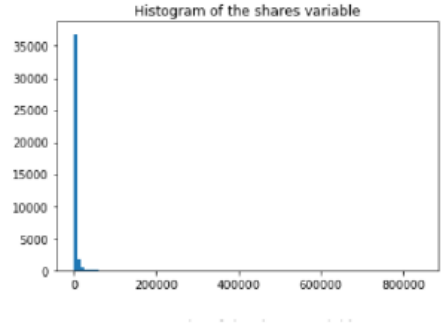
Ensemble:

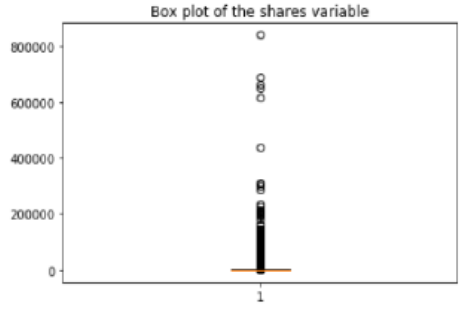
- Adaboost with CART as a weak learner, with accuracy always more than 50% is used.
- Three-fold cross validation with "number of learners" as a hyper parameter is used to identify the hyperparameter value which gives the highest accuracy. Range of number of learners experimented with: [5,10,25,50,75,100,150,200]
- Using the hyperparameter (number of learners) which gives the highest accuracy, pruning is performed over a range of Cp values. Range of Cp values experimented with [0.01,0.005,0.001,0.0005,0.0001]

Step 2: Data Understanding for both the datasets

Mashable Dataset:

Data has 39,644 records of 61 variables and there are no missing values. Two non-predictive features (URL and Timedelta) have been dropped from further analysis. All the numerical predictor features have been normalized. Some information about the shares variable: Mean being more than the median and having values above the 2 standard deviations (box plot) shows that the data is positively skewed

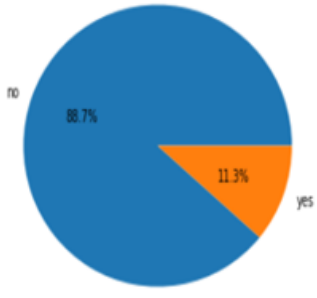




Median of the shares (1400) has been considered cutoff for the classification. The split of the articles to high shares (20,082) and low shares (19,562) will be ~50% because of considering median as the split point. Data has been divided into 70% train and 30% test in both the datasets.

Telemarketing Data:

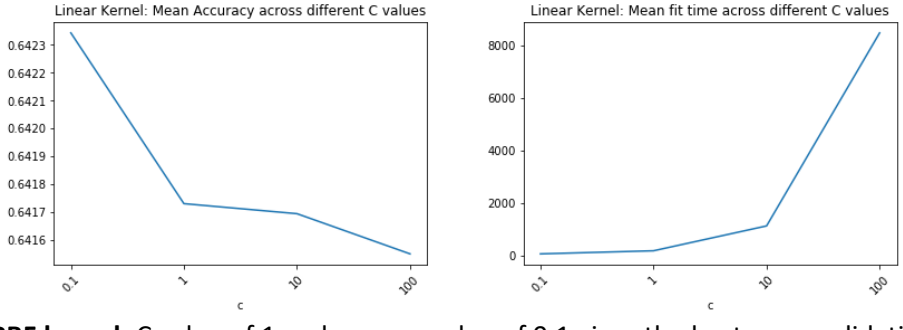
Data published in UCI machine learning has been used [2]. Data has 20 variables that are likely to describe whether a customer has subscribed to the product in telemarketing or not. All the categorical variables have been converted using one-hot encoding. All the numerical variables have been normalized. Success rate of telemarketing is 11.3%. Data has been divided into 70% train and 30% test in both the datasets. This dataset is interesting because of the class imbalance present in predictor variable compared to Mashable dataset.



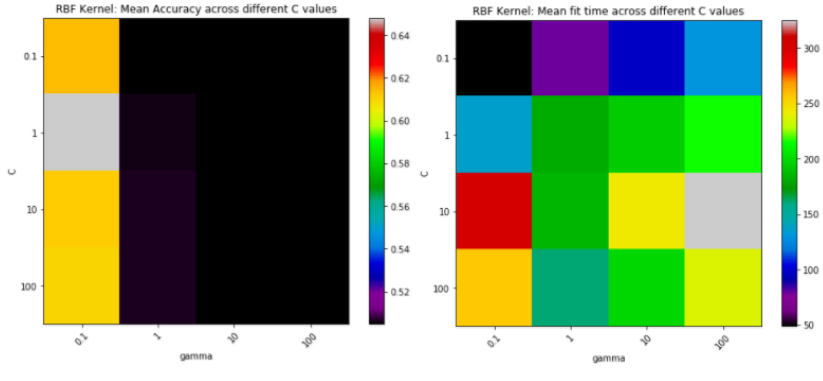
Step 3: Results of experimentation for Mashable problem

SVM

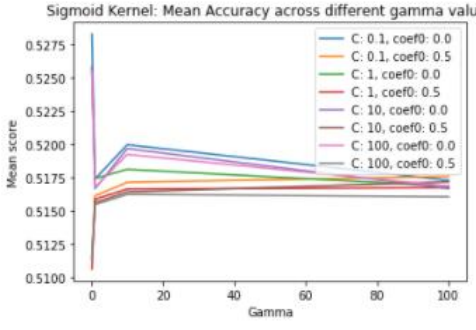
1. Results of three fold cross validation across hyperparameters mentioned in step 2. Each of kernels have two plots, by varying the hyperparameters over accuracy and time taken.
    a. **Linear kernel:** C of 0.1 gives the best cross validation accuracy of 64.2% and it takes the least time compared to other hyperparameters.

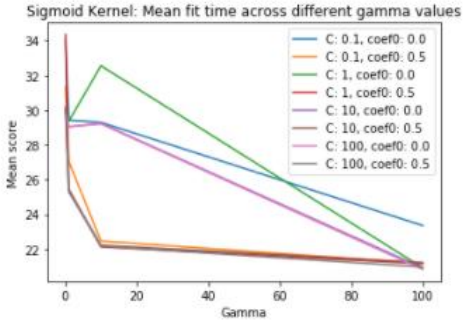


    b. RBF kernel: C value of 1 and gamma value of 0.1 gives the best cross validation accuracy of 64.8%. For the range of gamma [1,10,100], the cross validation accuracy is ~50%.



c. **Sigmoid Kernel:** C value of 0.1, coef0 value of 0.0 and gamma of 0.1 gives the highest cross validation accuracy of 52.8%.

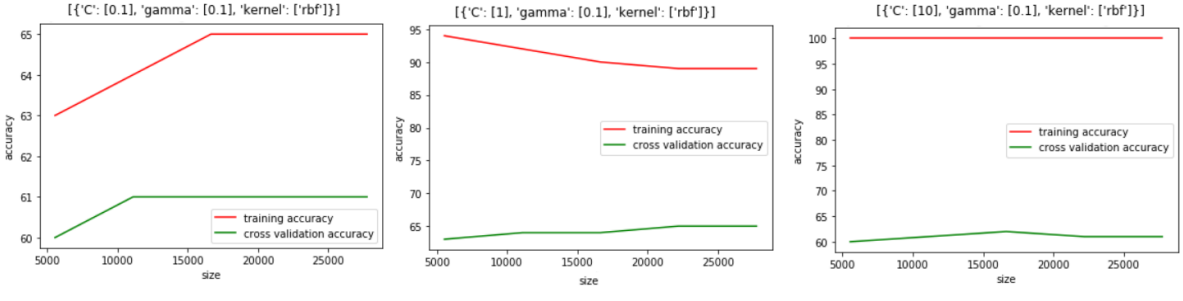




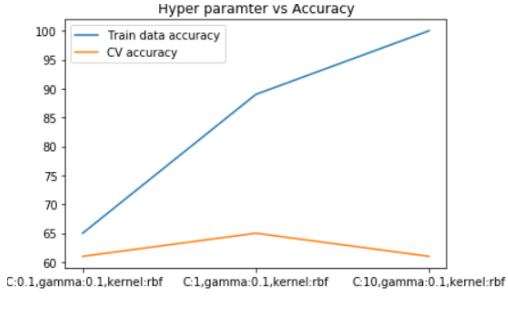
2. Maximum cross validation accuracy has been observed in RBF (c 1 and gamma 0.1). Accuracy for the test and train data using this kernel is:

```
Accuracy for train 88.7495495495   Accuracy for test 64.0575079872
confusion matrix for train          confusion matrix for test
[[12796 1904]                       [[3866 2073]
[ 1218 11832]]                      [2202 3753]]
```

3. A difference in accuracy of 24 points has been observed between train and test data. Varying the dataset size across different hyperparameters and evaluating the training(E_{in}) and cross validation(E_{out}) accuracy to check for bias and variance problem.



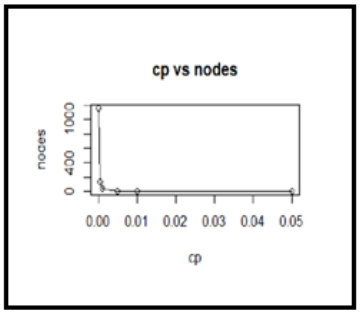
First kernel represented above suffers from bias, third kernel represented above suffers from variance problem. The kernel identified from the experimentation has less bias and less variance. A plot of hyperparameters vs Test and train accuracy shows the same. As we increase the complexity of the kernel, test accuracy initially increases and decreases post that, whereas the train accuracy increases.

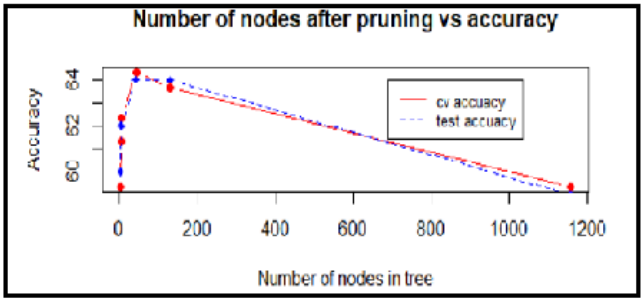


Decision Tree

Using a very low cp value (0.0001) and no pruning has resulted in a training accuracy of 83% and test accuracy of 59%.

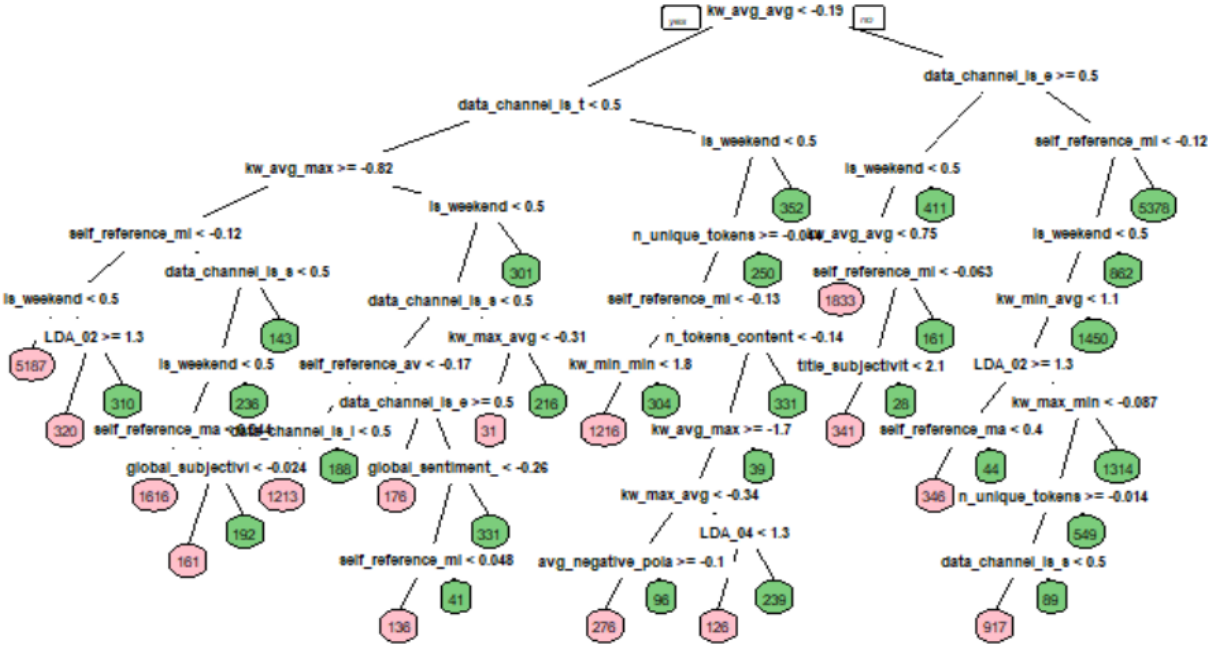
Pruning: As the cp value decreases, we allow tress to have more nodes. Using pruning the highest test accuracy (64%) and cross validation accuracy (64.3%) have been identified with a cp value of 0.001, which resulted in a tree with 41 nodes. Using cp higher than 0.05 or less number of nodes than 2 is not recommended because it results in test accuracy less than 59%, which is the test accuracy using the whole tree without pruning. Using less number of tress leads to high bias and more number of trees leads to high variance.





Plot of decision tree which resulted in the highest test accuracy:

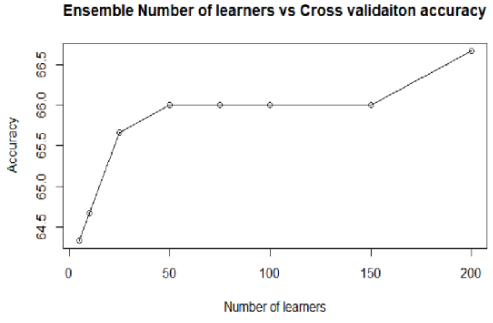
pink for low shares and green for high shares

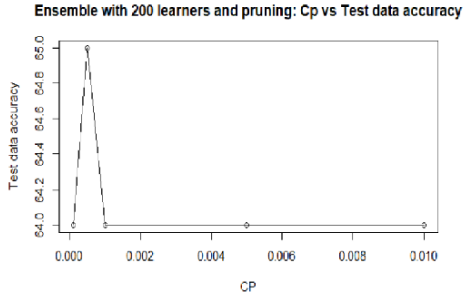


Ensemble

Left side of the image below: By varying the number of learners, cross validation accuracy of 66.5% has been achieved using 200 learners.

Right side of the image below: By pruning the tree built using 200 learners, it is observed that using a cp value of 0.0005 gives the highest test accuracy of 65%. Improvement over decision trees has been observed.





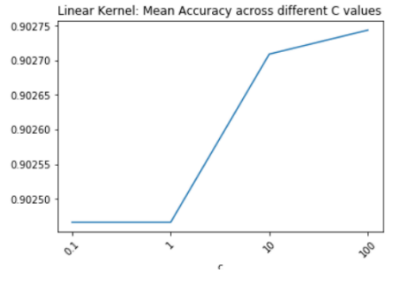
Summary:

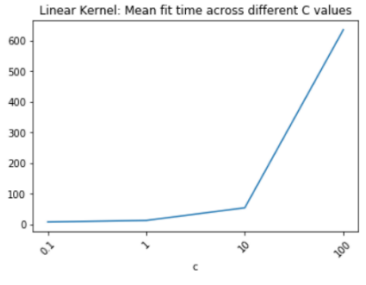
Among SVM, Decision Trees and Ensemble, highest test accuracy of 65% has been observed using an ensemble of 200 learners and using cp value of 0.0005.

Step 4: Results of experimentation for Telemarketing problem

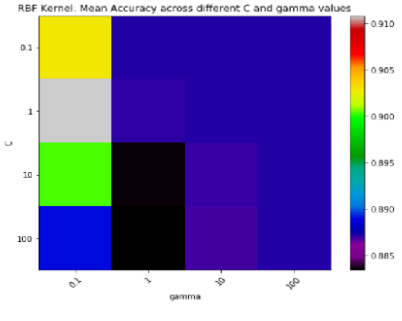
SVM

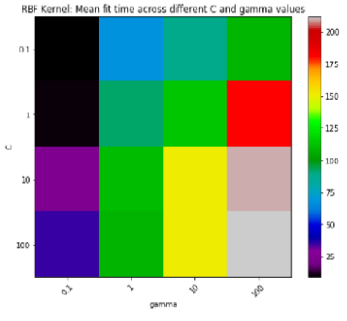
1. Results of three fold cross validation across hyperparameters mentioned in step 2. Each of kernels have two plots, by varying the hyperparameters over accuracy and time taken.
    a. **Linear kernel:** C of 100 gives the best cross validation accuracy of 90.2% and it takes the highest time compared to other hyperparameters.



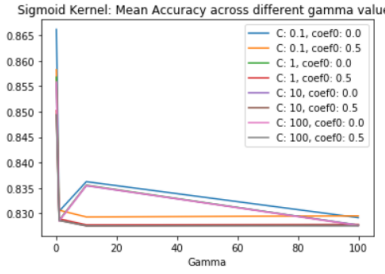


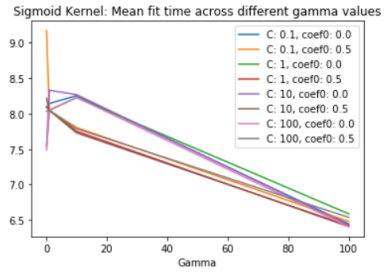
b. RBF kernel: C value of 1 and gamma value of 0.1 gives the best cross validation accuracy of 91.08%. For the range of gamma [1,10,100], the cross validation accuracy is ~88%.





c. Sigmoid Kernel: C value of 0.1, coef0 value of 0.0 and gamma of 0.1 gives the highest cross validation accuracy of 86.6%.

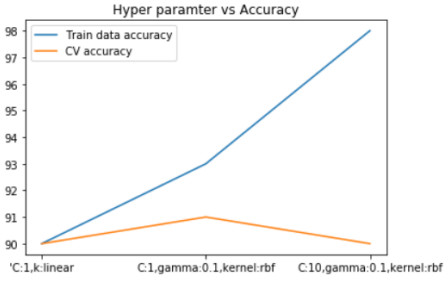




2. Maximum cross validation accuracy has been observed in RBF (c 1 and gamma 0.1). Accuracy for the test and train data using this kernel is:

```
accuracy for train 93.2641947903  accuracy for test 91.2195516711
confusion matrix for train        confusion matrix for test
 [[25206  1568]                    [[10674   791]
 [  374  1683]]                    [  294   598]]
```

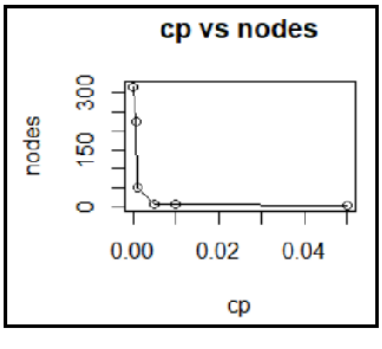
3. A difference in accuracy of ~2 points has been observed between train and test data. As we incre ase the complexity of the kernel, test accuracy initially increases and decreases post that, where as the train accuracy increases.

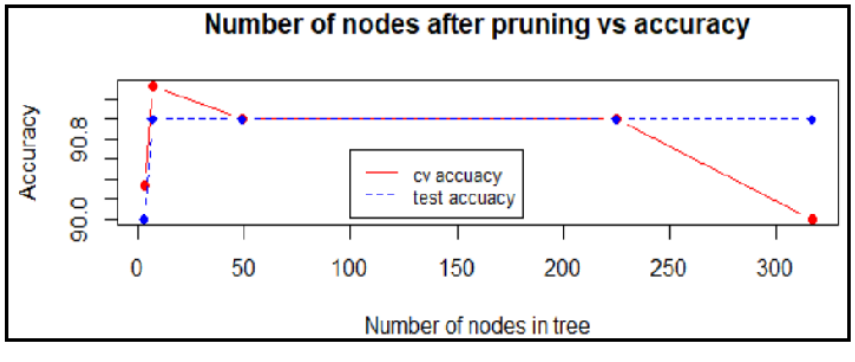


Decision Tree

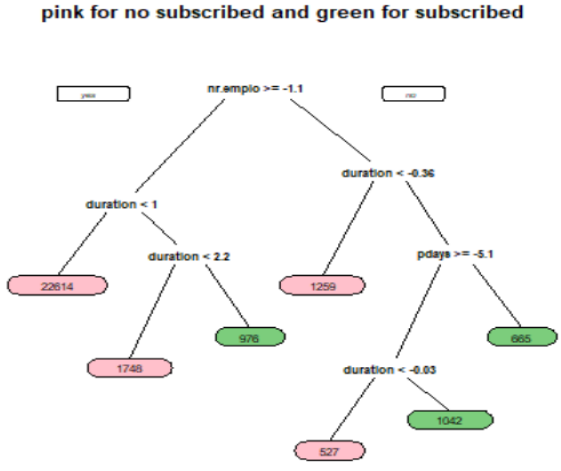
Using a very low cp value (0.0001) and no pruning has resulted in a training accuracy of 94% and test acc uracy of 91% .

Pruning: As the cp value decreases, we allow tree to have more nodes. Using pruning, the highest test ac curacy (91%) and cross validation accuracy (91.3%) has been identified with a cp value of 0.01, which res ulted in a tree with 7 nodes. Pruning has not improved the test set accuracy. Using cp higher than 0.01 o r less number of nodes than 7 is not recommended because it results in test accuracy less than 91%, whi ch is the test accuracy of the entire tree without pruning.





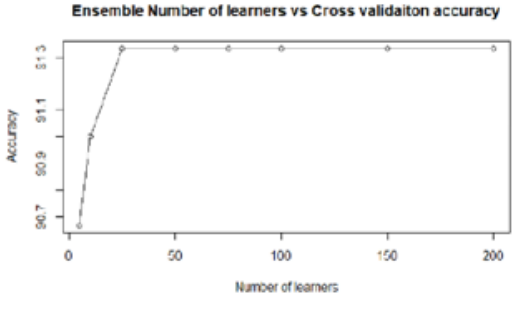
Plot of decision tree which resulted in the highest test accuracy:

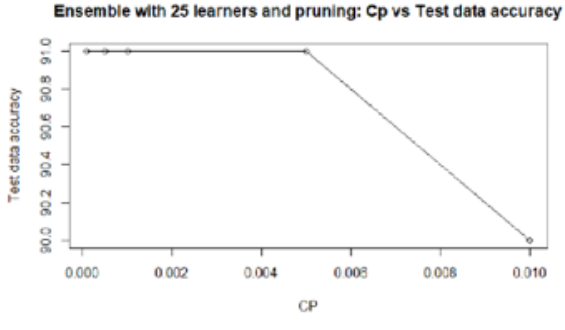


Ensemble

Left side of the image below: By varying the number of learners, cross validation accuracy of 91.3% has been achieved using 25 learners and the accuracy doesn't change post that.

Right side of the image below: By pruning the tree built using 25 learners, it is observed that using a cp value of [0.005,0.001,0.0005,0.0001] gives the test accuracy of 91%. No improvement over decision trees has bee n observed.





Summary:

SVM, Decision Trees and Ensemble result in approximately the same test accuracy (91%). Using decision tress is recommended in this scenario because of less time and more interpretability.

Step 5: Summary using both datasets

1. Among SVM kernels, in both the datasets, RBF kernel has given the highest accuracy
2. Accuracy using boosting is greater than or equal to decision trees
3. Experimentation with boosting consumes a lot of time when compared to other methods

Step 6: Next Step to improve the model accuracy

- Using polynomial kernel in SVM
- Trying other ensemble techniques like boosting, bagging

Reference:

1. http://saiconference.com/Downloads/SpecialIssueNo10/Paper_3-A_comparative_study_of_decision_tree_ID3_and_C4.5.pdf
2. https://archive.ics.uci.edu/ml/datasets/Bank+Marketing