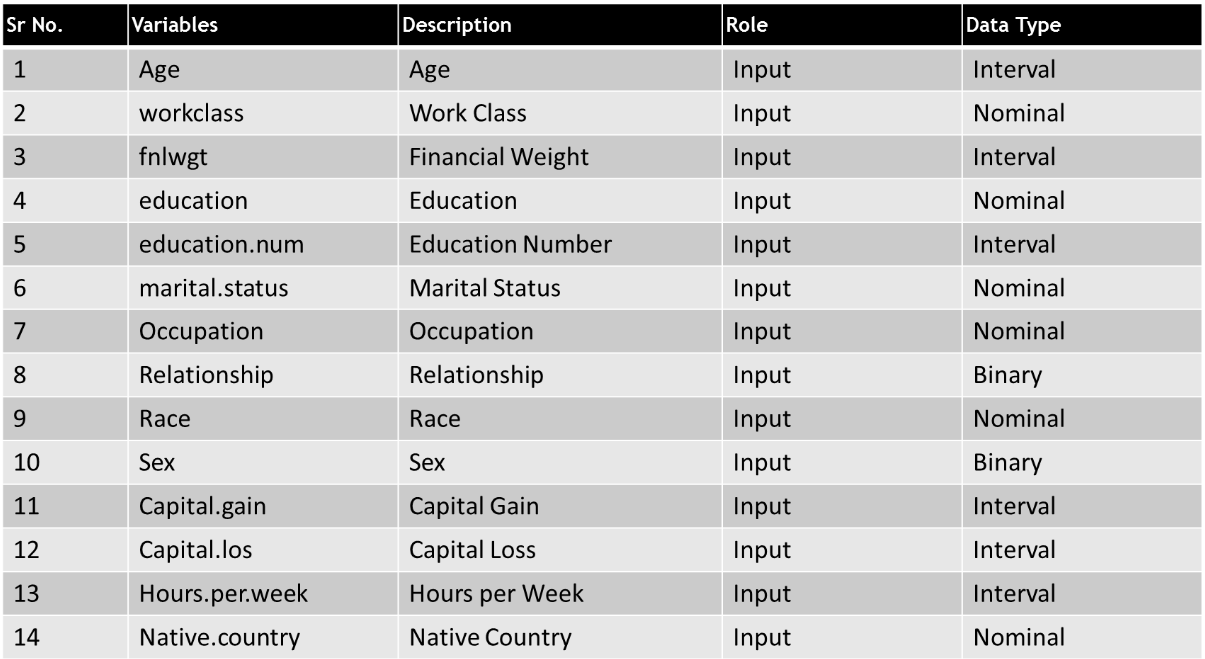
Objective: Learn Support Vector Machine, Decision Tree, and Boosting algorithm by implementing these algorithms against 2 different classification problems.

Classification Problem:

Adult Census: The dataset mainly contains census data in terms of age, education, relationship, race, work class, sex, hours per week and the annual income as a target with <50k or >50k classification labels. The prediction task is to determine whether a person makes over $50K a year. Source: https://www.kaggle.com/uciml/adult-census-income

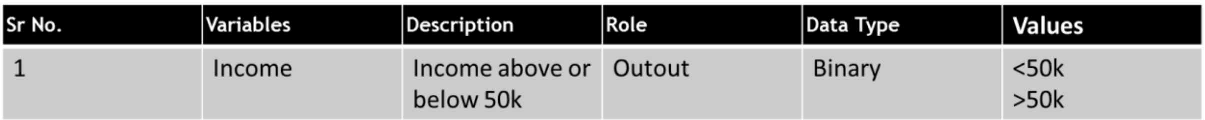
Census dataset gives a lot of valuable information which can be useful for many businesses. We will focus on analyzing the impact of diverse factors on annual income. Predict whether a person has income less than or more than 50k. Identify the most probable factors for a person's income to be more than 50k.This will help companies in target marketing by planning better strategies for a specific audience for the product they are marketing.

Input Variable:



Sr No.	Variables	Description	Role	Data Type
1	Age	Age	Input	Interval
2	workclass	Work Class	Input	Nominal
3	fnlwgt	Financial Weight	Input	Interval
4	education	Education	Input	Nominal
5	education.num	Education Number	Input	Interval
6	marital.status	Marital Status	Input	Nominal
7	Occupation	Occupation	Input	Nominal
8	Relationship	Relationship	Input	Binary
9	Race	Race	Input	Nominal
10	Sex	Sex	Input	Binary
11	Capital.gain	Capital Gain	Input	Interval
12	Capital.los	Capital Loss	Input	Interval
13	Hours.per.week	Hours per Week	Input	Interval
14	Native.country	Native Country	Input	Nominal

Output variable



Sr No.	Variables	Description	Role	Data Type	Values
1	Income	Income above or below 50k	Outout	Binary	<50k >50k

Some observational insights:

- There are number of people in the USA having salary <=50k
- If you work for more hour's week, you will have salary >50k
- The ratio of number people having salary <=50k to the number of people having salary >50k is minimum for Japan and maximum for the USA.

2: Human Resources Analytics:

The dataset has different parameters as shown in the table that tells us why the employees leave early. We will analyze these parameters to predict which employee can leave next.

The dataset is obtained from https://www.kaggle.com/ludobenistant/hr-analytics website

Companies spend a lot of resources on employee development for the growth of the business. if attrition rate is higher, then it will adversely impact companies image. Hence it very interesting analyze the reasons behind why the employee left the company and then prepare a model to predict whose chances are more to leave. This information can be used to reduce attrition rate.

Feature	Description	Role	Data Type
satisfaction_level	satisfaction level of employee in proportion	Input	continuous
last_evaluation	last evaluation of employee in proportion	Input	continuous
number_project	number of the project of employee	Input	Discreet
average_montly_hours	average monthly hours of employee	Input	continuous
time_spend_company	time spend in the company by an employee	Input	Discreet
Work_accident	Work accident of employee	Input	Discreet
promotion_last_5years	promoted in last 5 years or not	Input	Discreet
Department	Department of employee	Input	Discreet
salary	Salary	Input	continuous
left	Employee left or not	Target	Discreet

Data Preparation:

The Adult census dataset contains missing values and the class imbalance. Missing values are omitted and the dataset is operated using SMOTE function to resolve the class imbalance. The earlier ratio was 75:25 between target classification labels. After SMOTE the ratio is 52:48. Income as factored for categorical classification. The other categorical input parameters are also converted into a numerical equivalent.

HR dataset doesn't have any missing values but it contains class imbalance. The ratio of employee left to not left is 77:23. After SMOTE operation, the ratio is 53:47.

The newly generated rows contain decimal values even for pure integers such as age and other categorical values so they are rounded to integer values for both the dataset.

Once both datasets are cleaned, they are divided into train and test sets with 70:30 ratio respectively.

Target prediction for the Adult dataset and HR dataset using 3 different Algorithms which are:

1. Support Vector Machine
2. Decision Tree
3. Decision Tree with Boosting

Support Vector Machine:

The e1071 package is used to define the model using SVM. The models are prepared using 3 different kernels which are linear, polynomial, and sigmoid and the accuracy is calculated for all models for both datasets against test dataset.

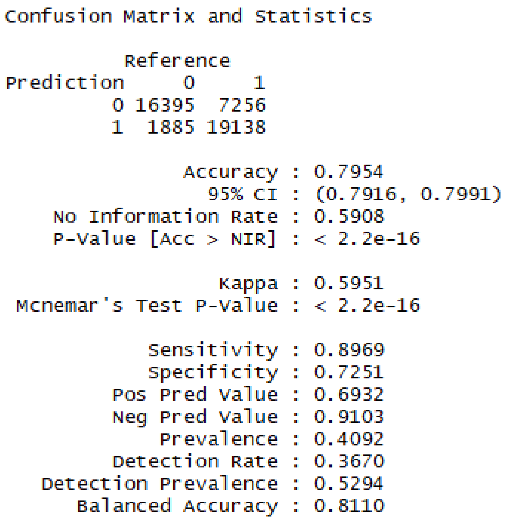
The model accuracy is as follows:

Algorithm	Dataset	Kernel	Accuracy	AUC for Test
Support Vector Machine	Adult Census	Linear	0.749	0.792
		Polynomial	0.795	
		Sigmoid	0.641	
	HR	Linear	0.776	0.942
		Polynomial	0.943	
		Sigmoid	0.573	

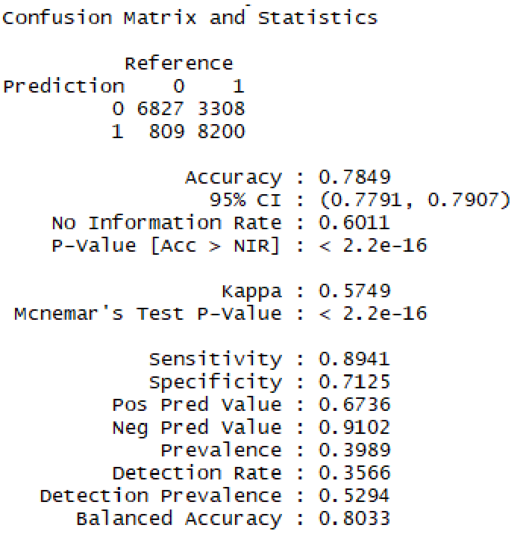
Models are performing well under Polynomial kernel so we can go ahead with this kernel.

For Adult Census with kernel Polynomial

Train Test



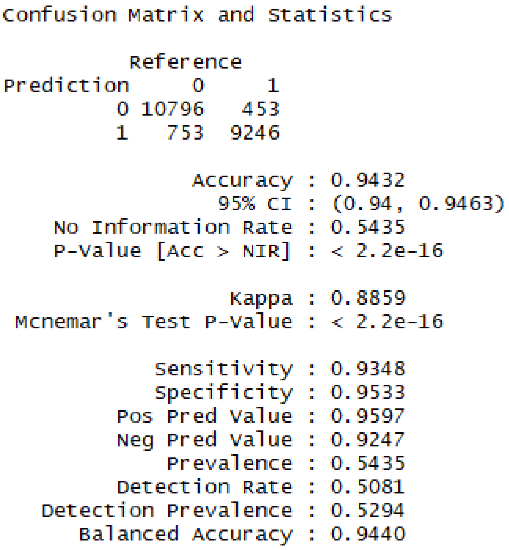
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 16395  7256
         1  1885 19138

               Accuracy : 0.7954
                 95% CI : (0.7916, 0.7991)
    No Information Rate : 0.5908
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5951
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8969
            Specificity : 0.7251
         Pos Pred Value : 0.6932
         Neg Pred Value : 0.9103
             Prevalence : 0.4092
         Detection Rate : 0.3670
   Detection Prevalence : 0.5294
      Balanced Accuracy : 0.8110
```



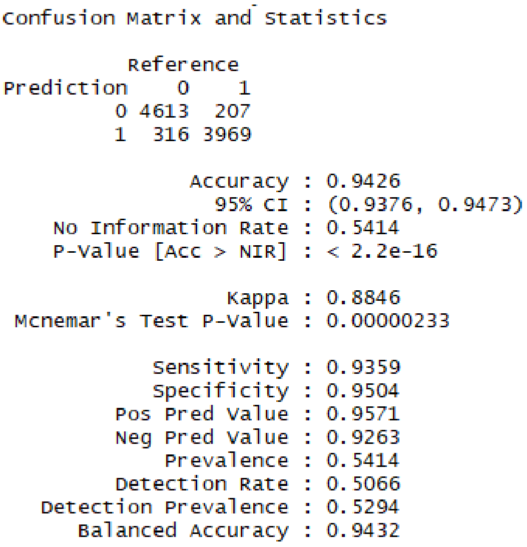
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6827 3308
         1  809 8200

               Accuracy : 0.7849
                 95% CI : (0.7791, 0.7907)
    No Information Rate : 0.6011
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5749
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8941
            Specificity : 0.7125
         Pos Pred Value : 0.6736
         Neg Pred Value : 0.9102
             Prevalence : 0.3989
         Detection Rate : 0.3566
   Detection Prevalence : 0.5294
      Balanced Accuracy : 0.8033
```

For HR with Kernel Polynomial

Train Test



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 10796   453
         1   753  9246

               Accuracy : 0.9432
                 95% CI : (0.94, 0.9463)
    No Information Rate : 0.5435
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8859
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9348
            Specificity : 0.9533
         Pos Pred Value : 0.9597
         Neg Pred Value : 0.9247
             Prevalence : 0.5435
         Detection Rate : 0.5081
   Detection Prevalence : 0.5294
      Balanced Accuracy : 0.9440
```



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4613  207
         1  316 3969

               Accuracy : 0.9426
                 95% CI : (0.9376, 0.9473)
    No Information Rate : 0.5414
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8846
 Mcnemar's Test P-Value : 0.00000233

            Sensitivity : 0.9359
            Specificity : 0.9504
         Pos Pred Value : 0.9571
         Neg Pred Value : 0.9263
             Prevalence : 0.5414
         Detection Rate : 0.5066
   Detection Prevalence : 0.5294
      Balanced Accuracy : 0.9432
```

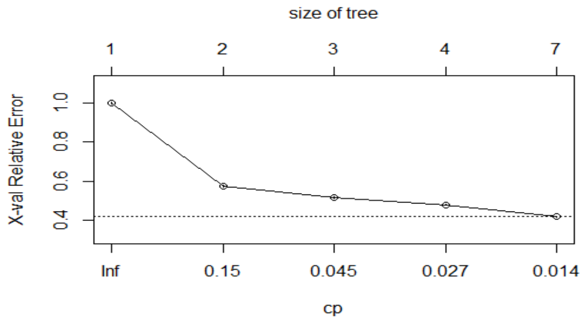
The confusion Matrix for Adult Census indicates TP and TN values as 16k and 19k which are good but FP is higher in both train and test set. The positive prediction value is lower (69% for train). To improve model performance, maybe we can experiment by adding more training sample for positive case. For HR, even though TP and TN are good but FN is higher than FP which is not good for Type 2 error.

Decision Tree

The decision tree models are generated using rpart and rattle package.

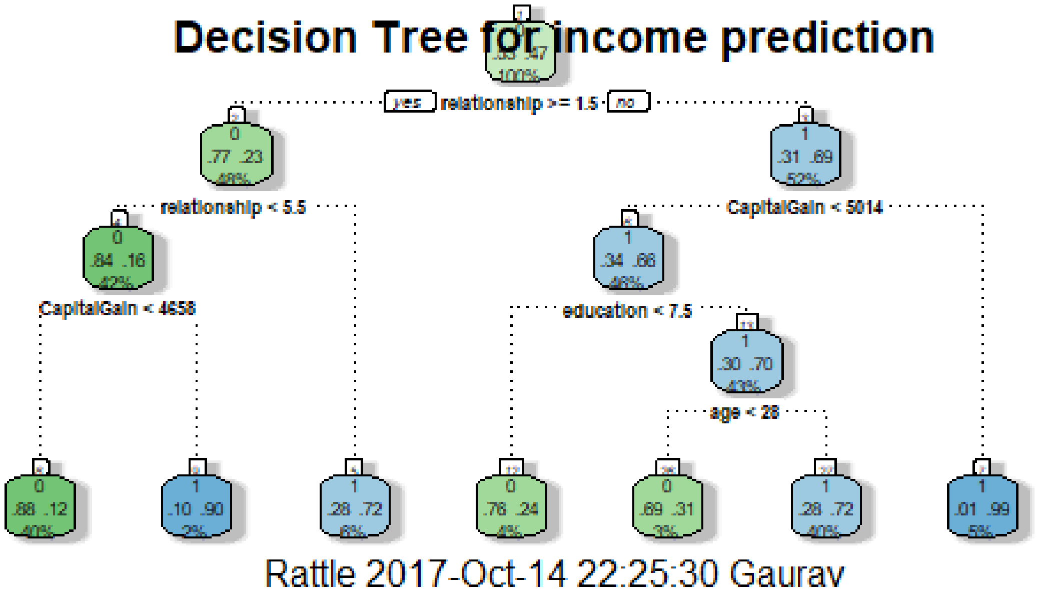
Adult Census Dataset: The models are generated using rpart function with the class method and with Gini and information split.

Accuracy of model with split Gini: 0.7952361
Accuracy of model with split information: 0.7974822
Based on the accuracy **Information** split is considered for a model generation with Tree pruning.
CP plot obtained for tree generated with information split gives tree depth for different CP values which are used to prune the tree. After pruning the tree for different values 0.011, 0.001 and the available lowest CP which is selected at runtime. I found it as 0.01.

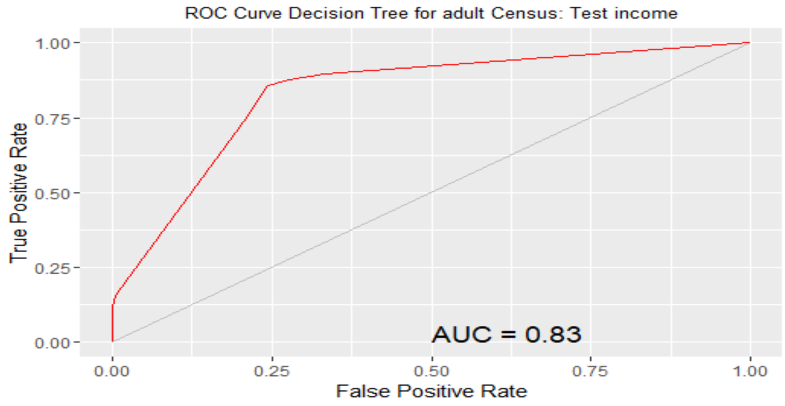


CP	Tree size	Accuracy
0.011	6	0.797
0.001	<7	0.797
0.01	7	0.797

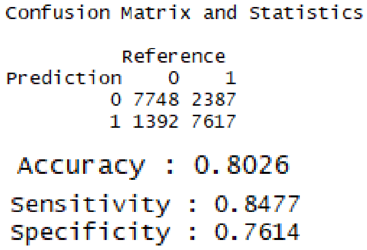
For different tree depths, there is no impact on the accuracy of the current model. Relative error also increases with reduction in the tree depth. Reducing the complexity will further increase the tree depths but can lead to overfitting the model.



The decision tree mainly decides classification using relationship, Capital Gain, education, and age. The Entropy is higher for a relationship.



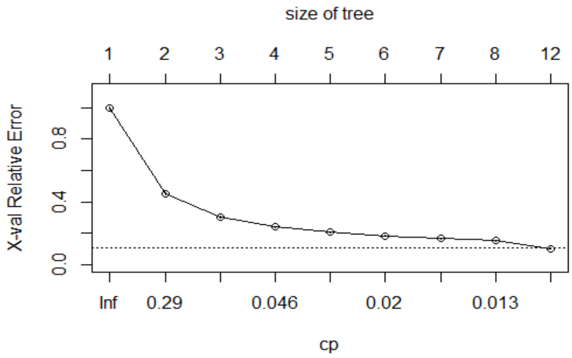
The model generated and tested with AUC 0.83 and other details are as follows:



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7748 2387
         1 1392 7617

 Accuracy : 0.8026
Sensitivity : 0.8477
Specificity : 0.7614
```

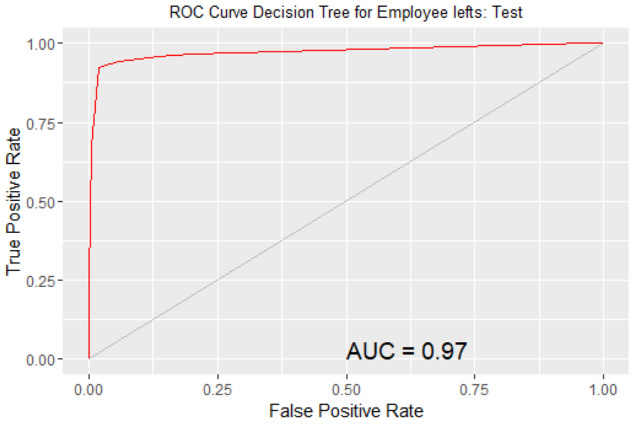
<u>HR Analytics Dataset</u>: The models are generated using rpart function with the class method and with Gini and information split.

Accuracy of model with split Gini: 0.9529929
Accuracy of model with split information: 0.9470621
<u>Based on the accuracy **Gini** split is considered for a model generation with Tree pruning.</u>
CP plot obtained for tree generated with information split gives tree depth for different Complexity (CP) values which are used to prune the tree. After pruning the tree for different values 0.011, 0.001 and the available lowest CP which is selected at runtime. I found it as 0.01.

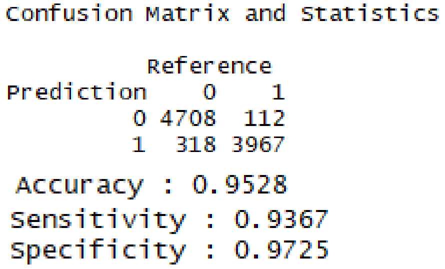


Complexity CP	Tree Depth	Accuracy
0.028	2	0.904
0.015	7	0.922
0.01	12	0.952

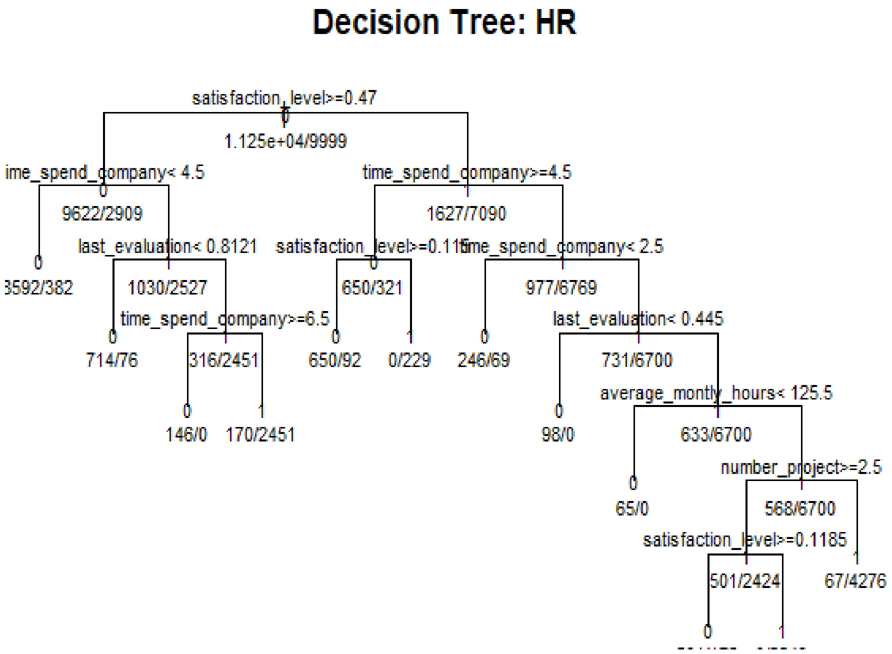
The complexity factor restricts the depth of tree. When its zero the tree will grow to its max depth but that will increase overfitting chances. Based on the accuracy values from the above table, the model is regenerated with Gini split and CP value 0.01.



The Receiver operating characteristics curve for test dataset with AUC 0.97 and other details are as follows:



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4708  112
         1  318 3967

 Accuracy : 0.9528
Sensitivity : 0.9367
Specificity : 0.9725
```



The decision Tree plot is mainly constructed using satisfaction_level, time_spend_company, last_evaluation, average_montly_hours, and number_project.

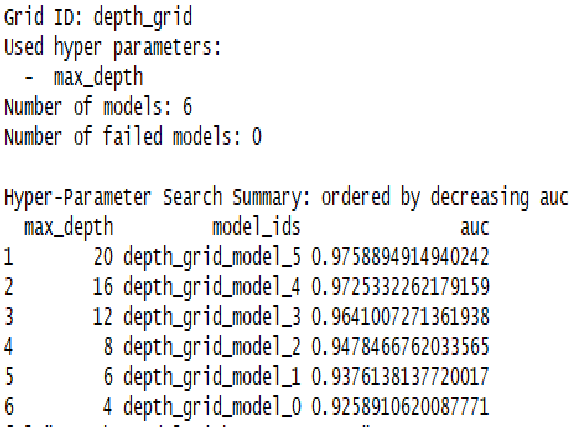
satisfaction level is very important for employee retention. The feature like department, work accident and promotion are not considered by the model. Even salary is not considered which is bit doubtful.

I have been working for more than 3 years and based on my experience, I can say that these parameters considered by models are important.

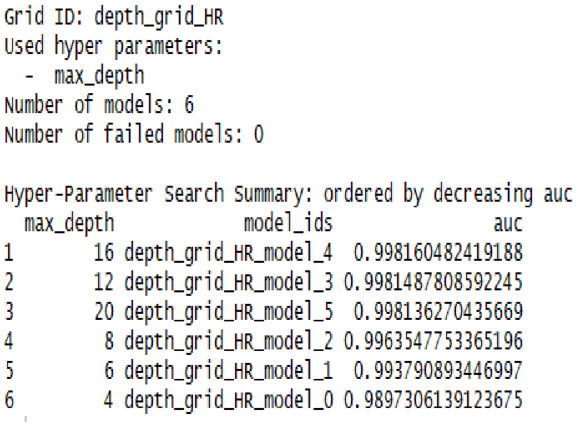
Boosting:

Model is generated using Decision Tree with Boosting using h2o package which consists gradient boosting algorithm. Both the datasets newly generated files after SMOTE operation are loaded again using h2o file loader. The datasets are divided into train, valid and test set. To understand model performance with pruning, the model is generated multiple times with hyper_params which are set to the list of max depths values such as 4,6,8,12,16,20 at the learning rate of 0.05 and ntrees = 10000.

For Adult Census & HR Dataset, model performances are as follows:



```
Grid ID: depth_grid
Used hyper parameters:
  -  max_depth
Number of models: 6
Number of failed models: 0

Hyper-Parameter Search Summary: ordered by decreasing auc
  max_depth          model_ids                auc
1        20 depth_grid_model_5 0.9758894914940242
2        16 depth_grid_model_4 0.9725332262179159
3        12 depth_grid_model_3 0.9641007271361938
4         8 depth_grid_model_2 0.9478466762033565
5         6 depth_grid_model_1 0.9376138137720017
6         4 depth_grid_model_0 0.9258910620087771
```



```
Grid ID: depth_grid_HR
Used hyper parameters:
  -  max_depth
Number of models: 6
Number of failed models: 0

Hyper-Parameter Search Summary: ordered by decreasing auc
  max_depth            model_ids                 auc
1        16 depth_grid_HR_model_4  0.998160482419188
2        12 depth_grid_HR_model_3 0.9981487808592245
3        20 depth_grid_HR_model_5  0.998136270435669
4         8 depth_grid_HR_model_2 0.9963547753365196
5         6 depth_grid_HR_model_1  0.993790893446997
6         4 depth_grid_HR_model_0 0.9897306139123675
```

The model with tree depth 20 has achieved a maximum area under the curve for Adult census whereas tree with depth 16 achieved maximum AUC for HR dataset. Thus, we cannot directly state that tree depth is directly proportional to AUC. It varies with the type of dataset.

This models max AUC are selected and tested against Test datasets.

Model	Dataset	Type	Accuracy	AUC	MSE
GBM	Adult Census	Train	0.986369	0.9991773	0.01906724
		Valid	0.924393	0.951779	0.05867855
		Test	0.925740	0.9779776	0.0566165
	HR	Train	0.999289	0.9999874	0.0020739
		Valid	0.986730	0.9981605	0.01162889
		Test	0.989791	0.9983714	0.009540592

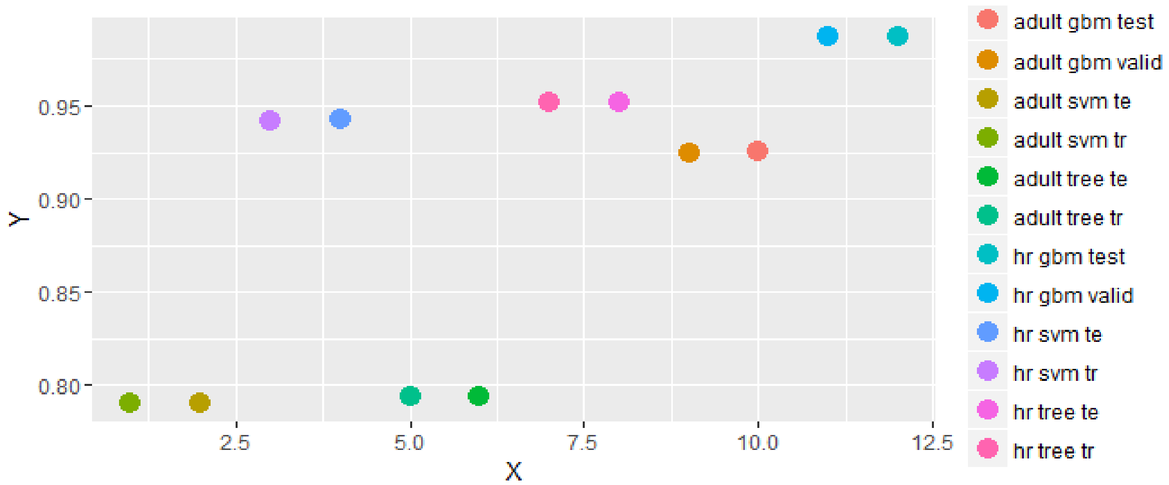
From the above tabular information, both the datasets performance are good for boosting algorithm.

Model Comparison:

After data cleansing, data preparations, and data partitions into training, testing and validation operations, different models are prepared, validated and tested using 3 different algorithms SVM, Decision Tree and Decision Tree with Boosting.

It can be observed that we may not get better accuracy based models just using only one algorithm.

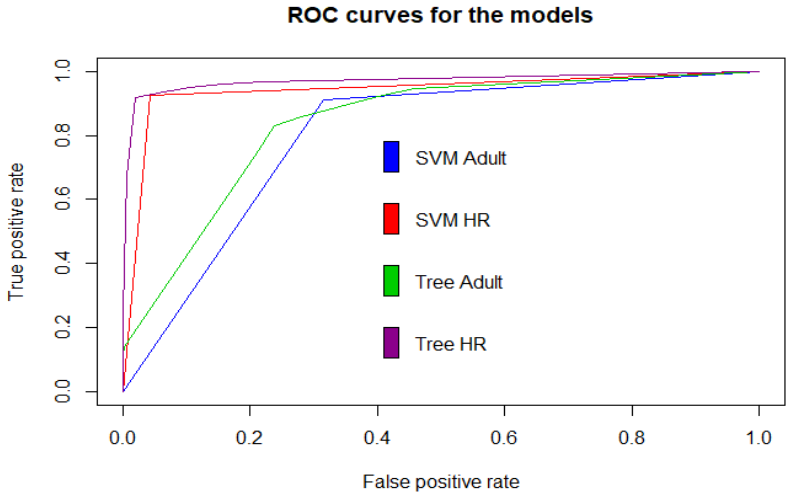
Model Accuracy: I have also captured the accuracy values for different models.



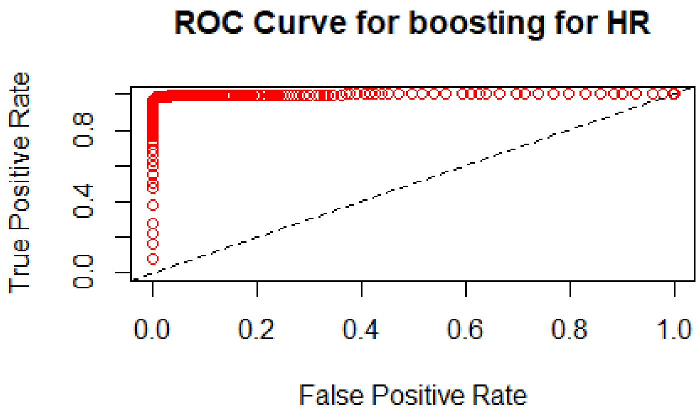
HR Dataset: Maximum model Accuracy is achieved using gbm with h20 package followed by decision tree and then SVM in the end.

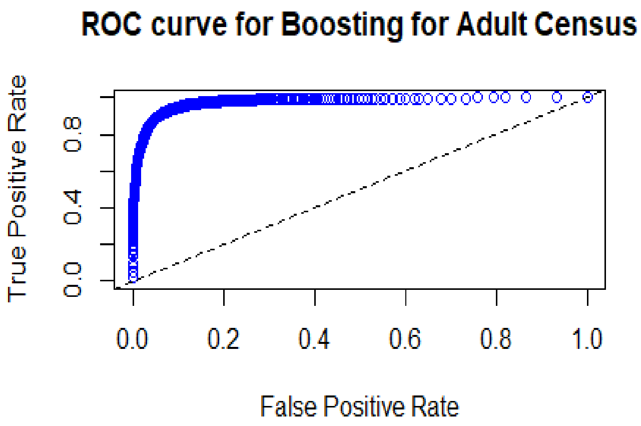
Adult Census Dataset: Maximum model accuracy is achieved using GBM algorithm followed by decision tree and then support vector machines.

For boosting, the accuracy for HR is close to approx. 98% whereas for the Adult census, the accuracy for SVM and decision tree is approx. 78%. The maximum accuracy for Adult Census is approx. 93%.



ROC curves resemble with the accuracy value plot of the models. AUC for all ROC curves is higher for GBM followed by Decision Tree and then SVM.





Overall, the boosting algorithm performs well as compare to decision tree and support vector machines. My dataset doesn’t have cases like cancer dataset where specificity is more important than accuracy. Here, I can reply on model accuracy, AUC to determine which model is doing well. With Boosting, your model does well on hard classification examples too which makes it better.

SVM algorithm results are not satisfactory in both the datasets. SVM performance can be increased by tuning the model against different cost and gamma factor. The kernel selection was done on an experimental basis. For polynomial kernel with default degree 3, The performance was better compared to other kernels. The model performance can be tested varying the degree until it doesn’t overfit.

To understand the impact of pruning of model, I chose tree depth parameter in boosting and tried to find AUC against every depth. In case of Decision Tree, I chose different complexity values to prune the tree. If the depth is too low then the model will not generalize well and it will underfit which make it bias.

The cross-validation helps to overcome the error factor in the training dataset based on observations in validation so that we don’t have to rely completely on Test dataset for model behavior.

Please refer R console for accuracy, roc, variable importance and other printed outputs.