**Project 2 - Implement Support Vector Machines (SVM), Decision trees, Boosting on Classification problems**

**Dataset – Employee Attrition**

This is a fictional data set created by IBM data scientists. The dataset contains variables such as monthly income, overtime hours, tenure etc. We have to develop a model to understand whether the employee will the organization or not.

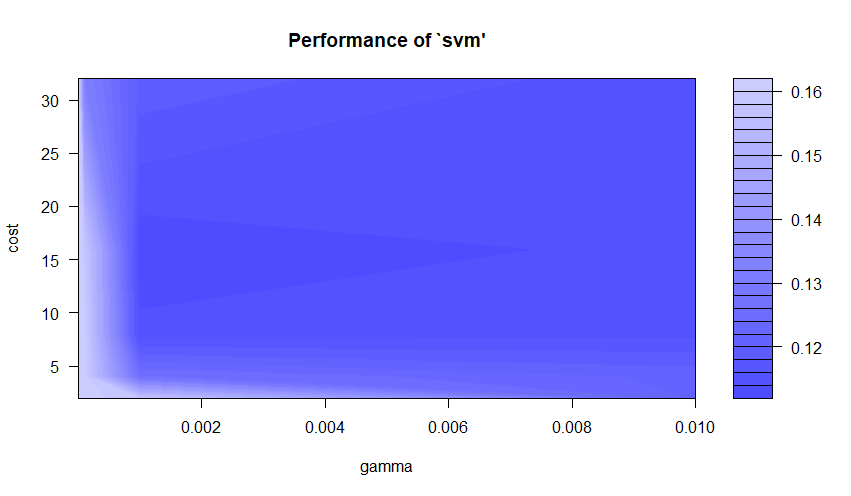
This problem is interesting as it helps us understand what are the important variables that makes an employee leave an organization.

We shall build 3 types of models – SVM, Decision Tree, Gradient Boosting.

**Support Vector Machines**

Using Support vector machines, we aim to classify employees utilizing 3 different kernels – Radial, Linear, Polynomial.

We first tune the SVM model based on different gamma and cost values using a grid.



We see from the above graph that the best tuning hyperparameters for SVM are gamma = 0.001, cost = 16 for a performance: 0.1126214

We shall use the above Hyperparameters to build our SVM models for different kernels.

**Radial Kernel**

Using the Radial Kernel, we get the following Confusion Matrix

actual

fitted No Yes

No 353 44

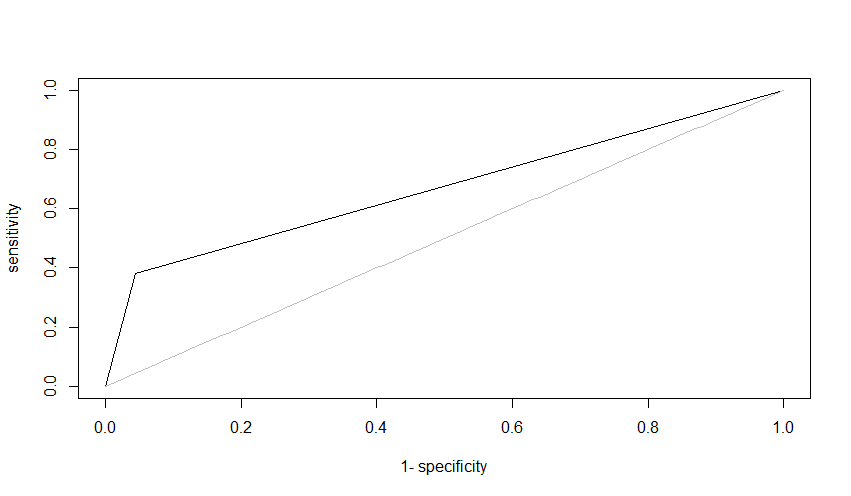
Yes 16 27

Accuracy is 86.36%

True Positive Rate - When employee leaves, and how often our model predicts that: 62.97%

Specificity - When employee stays, and how often our model predicts that: 88.91%

Next, we build the ROC curve to find the Area under the curve metric.



AUC = 0.6684606

**Linear Kernel**

Using the Linear Kernel, we get the following Confusion Matrix:

actual

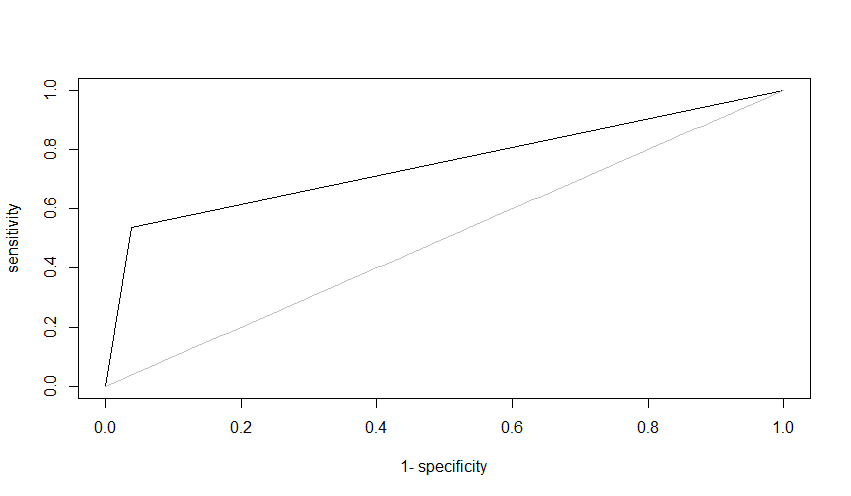
fitted No Yes

No 355 33

Yes 14 38

Accuracy is 89.31%

True Positive Rate - When employee leaves, and how often our model predicts that: 73.07%

Specificity - When employee stays, and how often our model predicts that: 91.49%

AUC = 0.7486354

**Polynomial Kernel**

Using the Polynomial Kernel, we get the following Confusion Matrix

actual

fitted No Yes

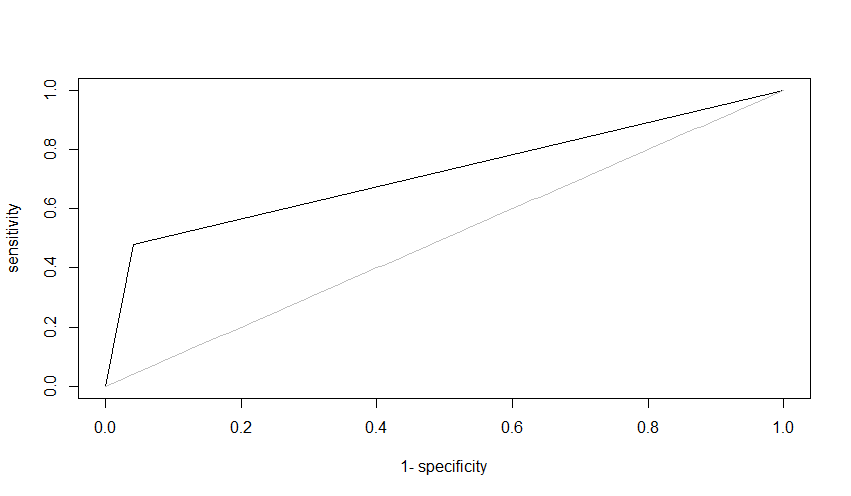
No 354 37

Yes 15 34

Accuracy is 88.18%

True Positive Rate - When employee leaves, and how often our model predicts that: 69.38%

Specificity - When employee stays, and how often our model predicts that: 90.53%

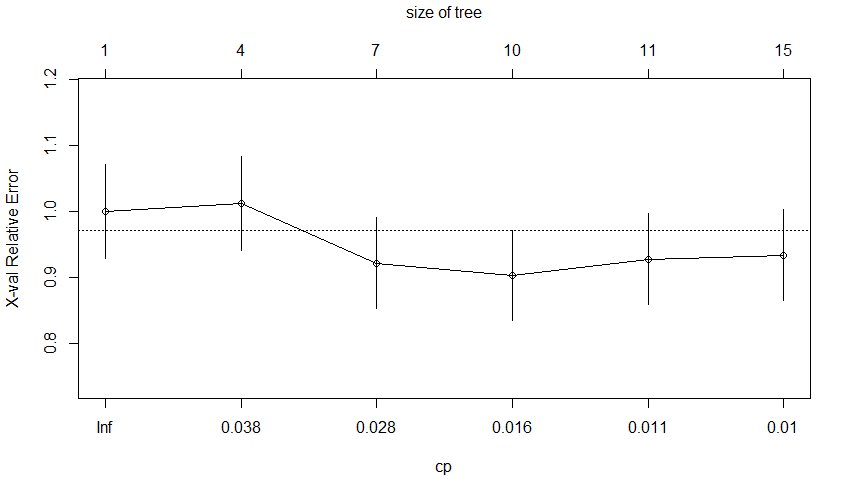


AUC = 0.7191114

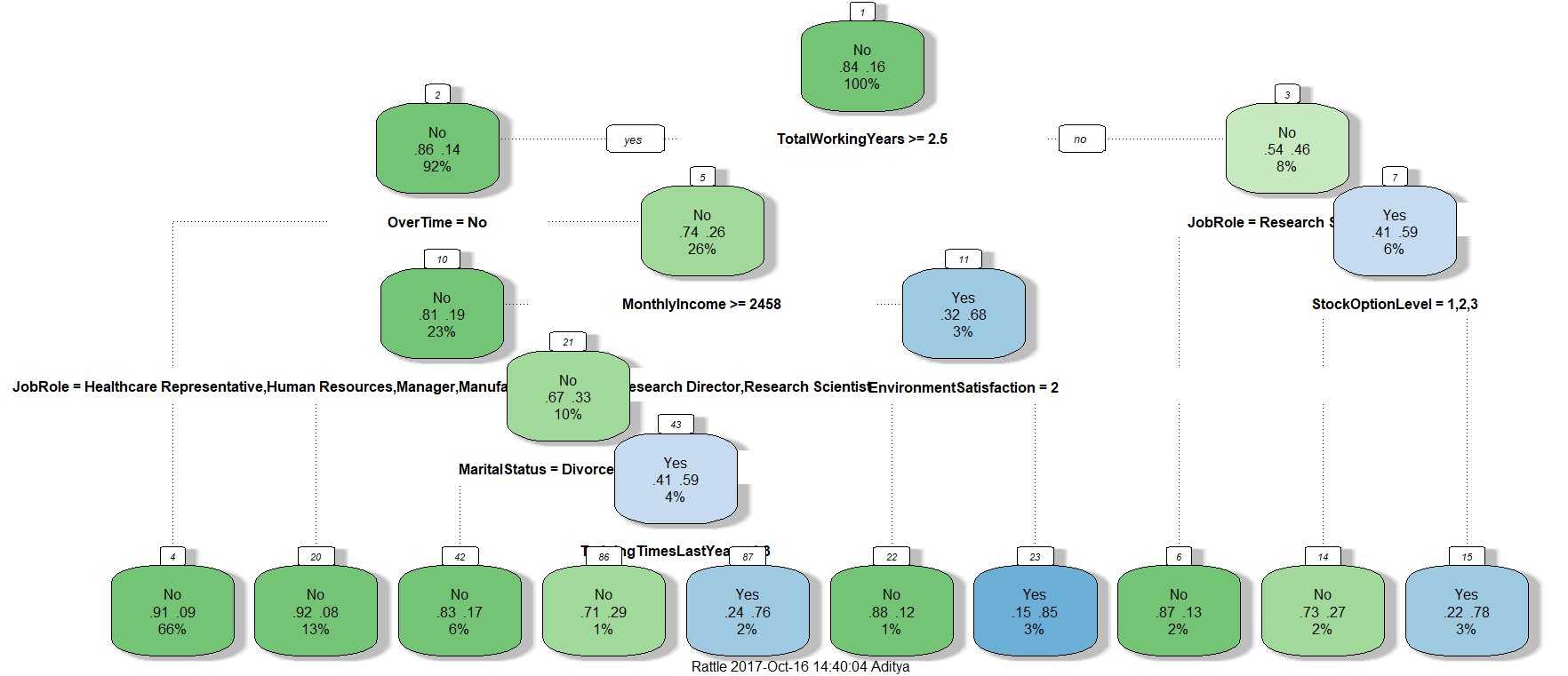
Linear Kernel performs the best among all the kernels with Accuracy of 89.31 and AUC = 0.7486354

**Decision Tree**

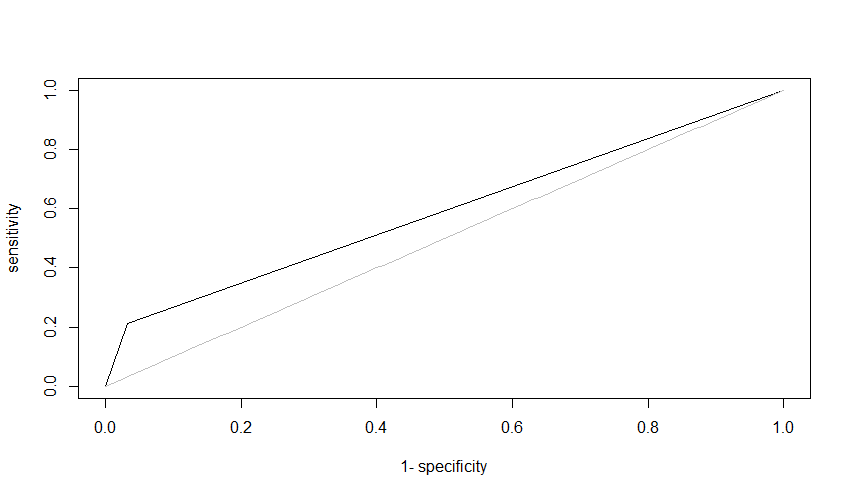
Next, we use a Decision tree to build our model.



We use pruning and check for what size of the tree the cross-validation error is the least. Using the size of the tree with lowest cross validation error we build our decision tree.



Main variables affecting employee attrition: TotalWorkingYears, JobRole, Overtime, MonthlyIncome & MaritalStatus.



actual

fitted No Yes

No 357 56

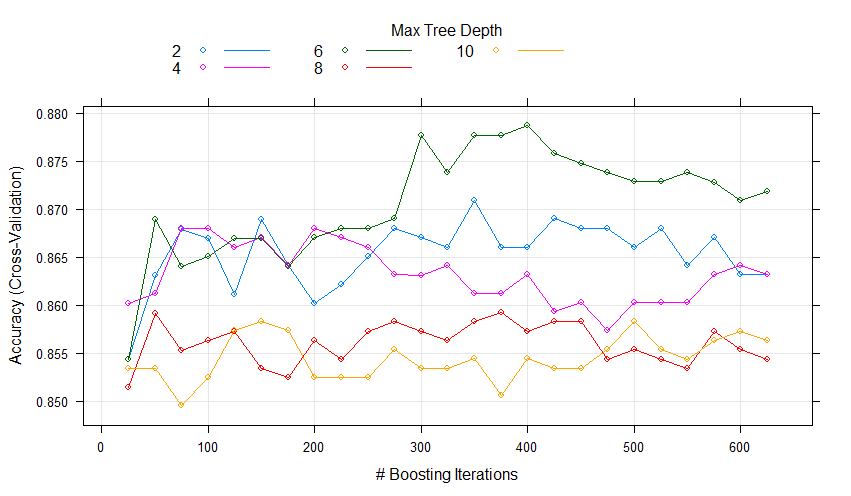
Yes 12 15

Accuracy = 372/440 = 84.54%

AUC = 0.5893736

**Boosting**

Next, we utilize Boosting to build our model. We build a gbm grid having all our hyperparameter values for tree depth and Boosting iterations.



We plot the cross-validation error for different tree depths and boosting iterations to find the best parameters.

We then use the fitted model to get predictions and find the below top variables affecting Attrition.

**var**  **rel.inf**

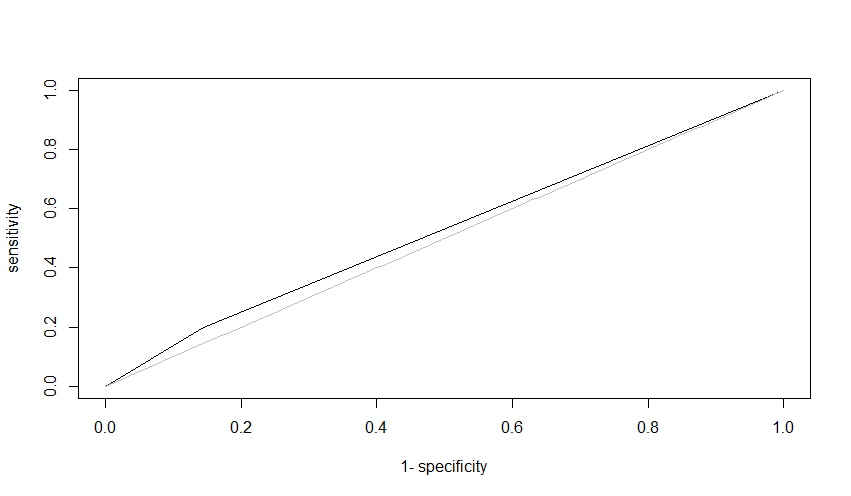
JobRole 8.9938366

MonthlyIncome 8.6570092

Age 6.1900461

DailyRate 5.2741509

OverTime 5.1781483



AUC = 0.6497767

actual

fitted 0 1

0 360 48

1 9 23

Accuracy = 383/440 = 87.04%

**Comparison of the three Models**

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | AUC |
| SVM | 89.31% | 0.7486354 |
| Decision Tree | 84.54% | 0.5893736 |
| Boosting | 87.04% | 0.6497767 |

SVM provides the best result as it has the highest accuracy as well highest AUC metric