### 

### 

### 

### 

### 

### 

### 

### 

### 

### **IBM HR ANALYTIC DATASET**

**ISDS-574**

### 

### 

### 

### **CONTENTS**

1. Executive Summary
2. Introduction

Background and Problem Description

Questions of interest

Dataset Description

Connection to literature

3. Data Exploration and Preprocessing

Summary Statistics

Data Visualizations

Data Cleaning

Correlation Analysis

Data Partitioning

4. Data Mining tasks

k-Nearest Neighbor

Logistic Regression

Classification tree

5. Results

6. Conclusion and Recommendations

7. Appendix and References

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### **EXECUTIVE SUMMARY**

Our project aims to use the fictional IBM HR Analytic dataset which contains employee attrition data. The focus is mainly on uncovering the factors that lead to employee attrition and to find out how much they affect our outcome. We performed a quick exploratory data analysis of some important variables and finally implemented data mining techniques (k-NN, Logistic Regression, Classification trees) to facilitate business decisions.

Conclusions of the analysis revealed that main general reason behind attrition in the company is most likely the effort-reward imbalance, especially for people who are working overtime. Our simple classification tree shows that apart from Overtime, monthly income is another deciding criterion for employee attrition. If employees are not well paid, they tend to leave the company for better opportunities. Some other factors like Age, frequent business travels, and job level also contribute in the employee attrition.

Based on our research, we have constructed a very simple pipeline of uncovering the factors affec ting employee attrition from some basic Exploratory Data Analysis (EDA) to implementing 3 different algorithms to build our model. That being said, there is a room for quite a lot of improvement. Class imbalance could be tackled by using more aggressive models like different variants of Gradient Boosting Machine (GBM) or by smoothing the dataset, since simple models seem to perform fairly only after making some adjustments (as discussed further in this report).

### **INTRODUCTION**

### **Background and Problem Description:**

### IBM is an American Multinational technology company with operations in over 170 countries. It manufactures and markets computer hardware, middleware, software, and offers hosting and consulting services. Employee attrition is a major cost to their organization, and predicting the turnover rate is one of the most pressing needs of HR. Some of these costs are tangible such as training expenses, the time invested in the employees etc. Others are intangible in the form of product ideas, customer relationships, etc. By accurately predicting employee attrition, IBM can take steps to keep their employees happy and less susceptible to competitive overtures as well as save on time and money to hire somebody else.

### The goal of our project is to use the fictional dataset provided by the IBM Research Scientists to uncover different factors that lead to higher employee attrition. We aim to build a model to get a better explanation of the critical features that are linked to attrition. The variables in this dataset include information about the employees’ demographics, evaluation about the company and the level of satisfaction with the company. Our team used different data mining techniques to find out how and which of these variables are the most significant ones.

### **Questions of interest:**

### Following are some questions of interest that we aim to answer by the end of the project:

### Does employee attrition depend on Monthly Income?

### Does the job satisfaction depend on the job level? If yes, do these factors influence the employee attrition rate?

### Do the employees associated with an organization since a longer time have a lower churn rate as compared to the ones that have been recently hired?

### Does a specific department need overtime which results in dissatisfaction among the employees?

### Which of the job roles has the highest probability of employee attrition?

### Finally, is there are relationship between an employee’s age and the attrition rate?

### **Dataset description:**

### The original dataset contains 1470 observations and 35 variables which are as follows:

### 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Field description** | **Variable type** | **String/****Numeric** | **Ordinal/****Nominal** |
| **Age** | **Age of employees in the company** | **Continuous** | **Numeric** |  |
| **Attrition** | **Target variable. 2 levels: Yes, No** | **Categorical** | **String** | **Nominal** |
| **BusinessTravel** | **Whether an employee travels frequently, rarely or doesn’t travel** | **Categorical** | **String** | **Nominal** |
| **DailyRate** | **Daily pay rate of an employee** | **Continuous** | **Numeric** |  |
| **Department** | **Department under which an employee works** | **Categorical** | **String** | **Nominal** |
| **DistanceFromHome** | **Distance of the work place from an employee’s home** | **Continuous** | **Numeric** |  |
| **Education** | **Education level of an employee. Value labels:****1- Below College 2- College 3- Bachelors 4- Masters 5- Doctor** | **Categorical** | **Numeric** | **Nominal** |
| **EducationField** | **Employee’s educational field like Life Sciences, Medical, etc.** | **Categorical** | **String** | **Nominal** |
| **EmployeeCount** | **Unary variable containing value 1** | **Unary** | **Numeric** |  |
| **EmployeeNumber** | **ID Number of an employee** | **Continuous** | **Numeric** |  |
| **EnvironmentSatisfaction** | **How satisfied an employee is with his work environment. Value labels: 1- Low 2- Medium 3- High 4- Very high** | **Categorical** | **Numeric** | **Ordinal** |
| **Gender** | **Gender of an employee: Male, Female** | **Categorical** | **String** | **Nominal** |
| **HourlyRate** | **Hourly pay rate of an employee** | **Continuous** | **Numeric** |  |
| **JobInvolvement** | **How heavily an employee is involved with his job. Value labels:****1- Low 2- Medium 3- High 4- Very High** | **Categorical** | **Numeric** | **Ordinal** |
| **JobLevel** | **Level of job of an employee: Value labels:****1- Low 2-Medium 3- High 4- Very high** | **Categorical** | **Numeric** | **Ordinal** |
| **JobRole** | **Current job role of an employee.** | **Categorical** | **String** | **Nominal** |
| **JobSatisfaction** | **How satisfied an employee is with his current job. Value labels:****1- Low 2- Medium 3- High 4- Very high** | **Categorical** | **Numeric** | **Ordinal** |
| **MaritalStatus** | **Whether an employee is married, single or divorced.** | **Categorical** | **String** | **Nominal** |
| **MonthlyIncome** | **Monthly income of an employee** | **Continuous** | **Numeric** |  |
| **MonthlyRate** | **Monthly pay rate of an employee.** | **Continuous** | **Numeric** |  |
| **NumCompaniesWorked** | **Total number of companies an employee has worked.** | **Continuous** | **Numeric** |  |
| **Over18** | **Whether an employee is over 18 or not.** | **Unary** | **String** |  |
| **OverTime** | **Whether an employee works overtime or not. 2 levels: Yes, No** | **Categorical** | **String** | **Nominal** |
| **PercentSalaryHike** | **Salary hike of an employee based on his performance** | **Continuous** | **Numeric** |  |
| **PerformanceRating** | **Rating given to each employee’s performance. Value labels:****1- Low 2- Good 3- Excellent 4- Outstanding** | **Categorical** | **Numeric** | **Ordinal** |
| **RelationshipSatisfaction** | **Satisfaction level of the customer-employee relationship.****Value labels:****1- Low 2- Medium 3- High 4- Very High** | **Categorical** | **Numeric** | **Ordinal** |
| **StandardHours** | **Standard hours an employee works (bi-weekly)** | **Unary** | **Numeric** |  |
| **StockOptionLevel** | **Level of stocks provided by the company: 0, 1, 2** | **Categorical** | **Numeric** | **Ordinal** |
| **TotalWorkingYears** | **Total number of years an employee has been working** | **Continuous** | **Numeric** |  |
| **TrainingTimesLastYear** | **The no. of trainings an employee has taken in the previous year** | **Continuous** | **Numeric** |  |
| **WorkLifeBalance** | **How well an employee can balance his work life. Value labels:****1- Bad 2- Good 3- Better 4- Best** | **Categorical** | **Numeric** | **Ordinal** |
| **YearsAtCompany** | **The total number of years an employee has spent with the company** | **Continuous** | **Numeric** |  |
| **YearsInCurrentRole** | **The number of years an employee has spent in his current job role** | **Continuous** | **Numeric** |  |
| **YearsSinceLastPromotion** | **The number of years since employee’s last promotion** | **Continuous** | **Numeric** |  |
| **YearsWithCurrManager** | **The number of years an employee has worked with his current manager** | **Continuous** | **Numeric** |  |

### 

### **Connection to the literature:**

### Before proceeding with the IBM fictional dataset and our approach, we referred to the below mentioned three models which were built on the same fictional dataset.

1. IBM Employee attrition by Numerics
2. Employee attrition by RF and GBM
3. IBM HR Analysis with 90.3% Accuracy and 89% Area under the curve

**IBM Employee attrition by Numerics**

The main focus of this model was to answer the following question.

What are common characteristics of employees lost in attrition compared to those who stay in IBM's fictional dataset?

Neither XL-Miner nor R-Programming was used to build this model or answer the question. Instead it used Python code to draw some of the important conclusions.

This model basically focused on representing the data through different kinds of visualizations like using point plots, box plots, kernel density diagrams, means, standard deviations, and z-tests.

Below mentioned are some of the characteristics or rather the conclusions drawn from the analysis.

* The fictional company on average loses staff that are 3 - 4 years younger than those who stay.
* Employees lost in attrition tend to have lower daily rates than those who stay.
* Employees lost in attrition tend to have longer commute distances than those who stay.
* Employees lost in attrition are less satisfied with their work environment on average than those who stay.
* Employees lost in attrition tend to be lower in job level than those who stay.
* Employees lost in attrition tend to have lower monthly average income on average than those who stay.
* Employees lost in attrition had poorer work-life balance on average than those who stay.
* Employees who are not paid for overtime are most likely to leave the Organization as compared to the ones who stayed, etc.

**Employee Attrition by RF and GBM**

This model unlike the above mentioned model used predictive modeling to predict the employee attrition on the fictional IBM dataset.

The steps involved in this model were:

1. Exploratory data analysis: In this technique they lookout at the feature distributions that how one feature is correlated to other.

-Data quality checks : to look for any null values.

-Correlation Matrix: how features are related to one another.

2. Categorical Encoding: To encode all the categorical features into dummy variables.

3. Finally they implemented the Machine learning models by splitting the data set into training and test datasets.

Random Forest method was used in order to predict the results.

(*Random Forest gathers a group (or ensemble) of decision trees and uses their combined predictive capabilities to obtain relatively strong predictive performance - "strong learner”*)

This model fetched 88% accuracy. And the two main features responsible for employee attrition came out to be Overtime effects and Marital status. At first glance this might seem to be a very good performing model. However when we think about how skewed our target variable where the distribution of yes and no's are 84% and 26%, therefore our model is only predicting slightly better than random guessing.

**IBM HR Analysis with 90.3% ACC and 89% AUC:**

The developers of this model used R-Programming to develop this prediction model.

The steps involved in building this model were:

Removing all previous variables -> Reading data -> Viewing the data -> Summarize the data with 1470 observations and 35 variables -> Check for missing values -> Check for duplicate records.

This model similar to our model used visualizations based on different attributes like age, Monthly income, marital status, etc. to predict the results.

It also used Random Forest method for prediction. The results fetched came out to be 90.3% accurate with 89% area under the curve.

This was the best out of the three models which we referred.

### 

### 

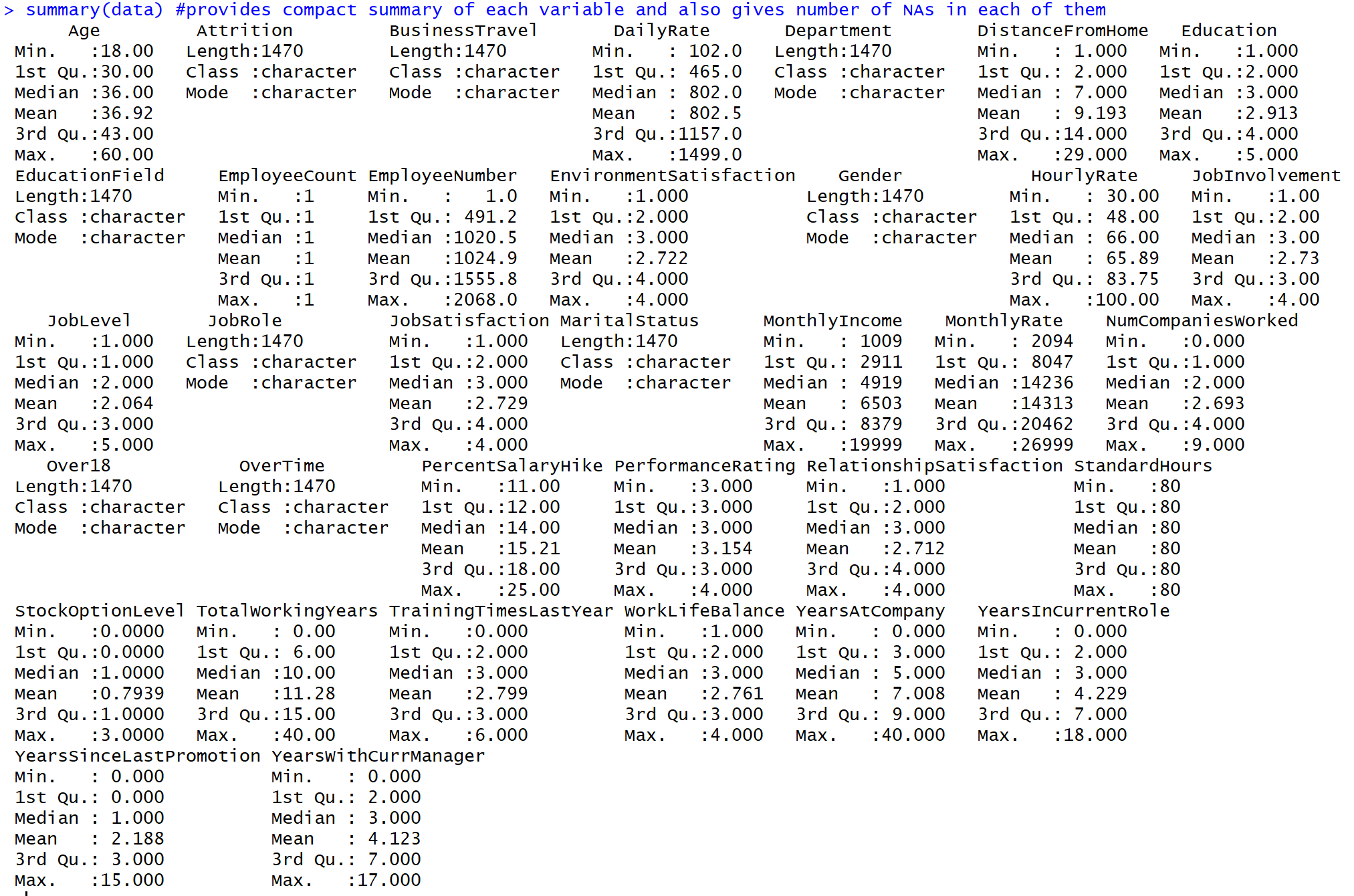
### 

### **DATA EXPLORATION AND PRE-PROCESSING**

### Before implementing data mining algorithms, it is important to understand the original data and transform the raw data into a processed form. The team used various methods to explore the data and convert it into the form ready for analysis.

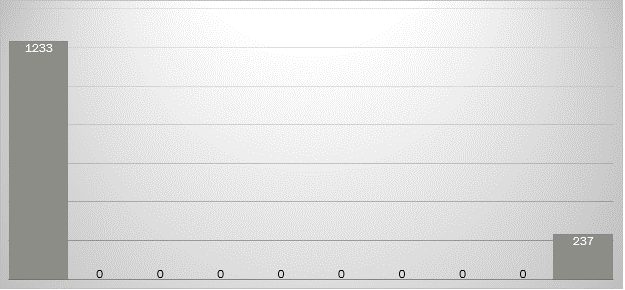
### **Summary Statistics**

### Descriptive Statistics were performed on all the variables. The little difference between mean and median of the numerical variables show that there is little or no skewness in these variables. It also revealed other characteristics of the variables like min, max, 1st and 3rd quartile, count etc.



**Data visualizations**

### Distribution of target variable “Attrition”: 237 out of the 1470 employees tend to leave the company which accounts to approximately 16% of the total number of employees. As is obvious from the histogram below, there is great imbalance in 2 classes of our target variable.



**Distribution of continuous variables against “Attrition”:**



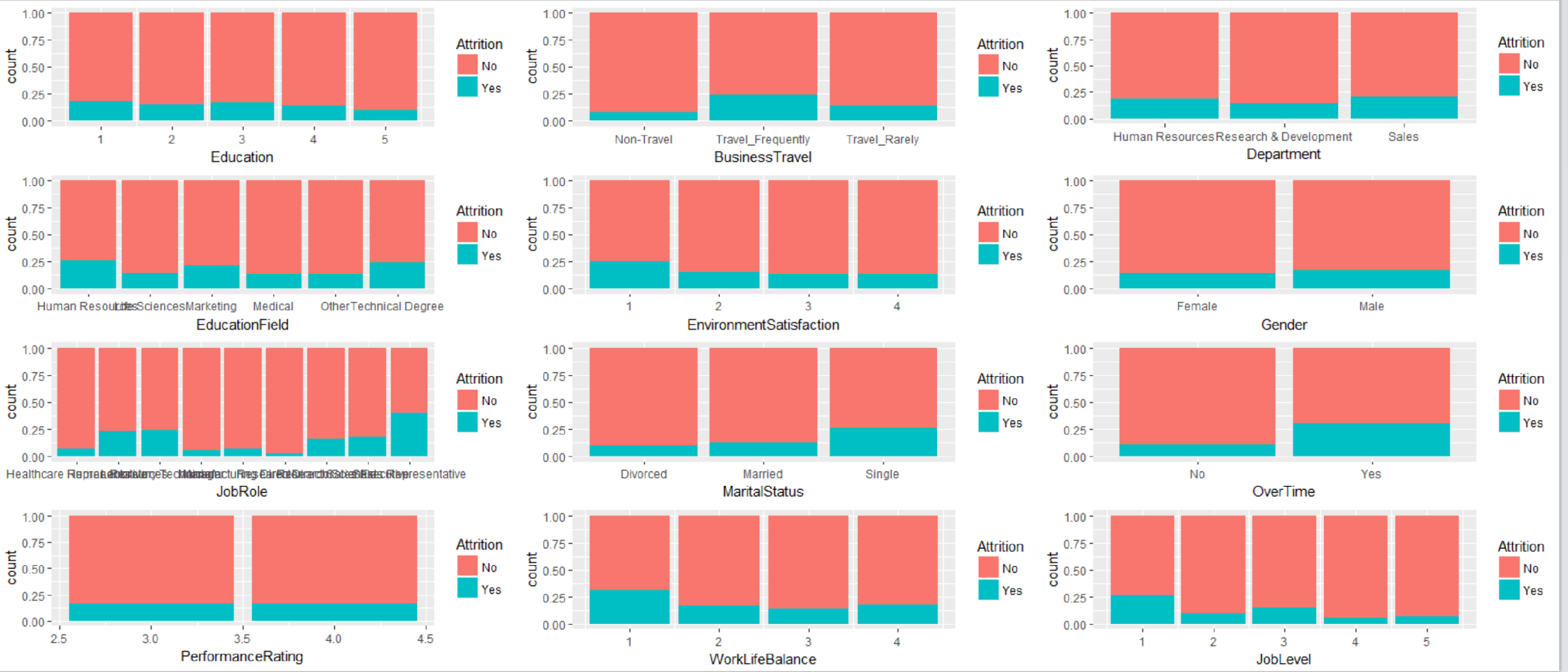
### The above plot matrix reveals some interesting relationships between our predictors and the target variable “Attrition”. Few of which are as follows:

### The attrition rate is higher for employees whose age ranges between 18-30. That makes sense since the older employees would just look for a secure job whereas the younger ones who must establish their careers tend to leave the company more often for better opportunities.

### As the employee’s distance from home to the company increases, the attrition rate increases too.

### As the total working years of an employee (Work experience) increases, the attrition rate seems to be decreasing.

### **Distribution of categorical variables against “Attrition”:**



Some interesting relationships from the above plot matrix are:

### The Attrition rate for different Education levels is almost similar, except for employees with PhD degree (their attrition rate is slightly less than the others).

### The employees who tend to travel frequently show higher rate of attrition as compared to the other 2 categories.

### Employees in Human Resources and Sales department have a slightly higher rate of attrition than R&D.

### Employee’s satisfaction rate with the company’s environment and his work-life balance is directly proportional to the attrition rate.

### Employees who are single tend to have a higher turnover rate.

### Employees who work Overtime tend to leave the company more than the others. (Effort-reward system of the company seems to be not great).

### **Data cleaning**

### **Detecting missing values:** Fortunately, our dataset had no missing values, so we went ahead with the next steps.

### **Detecting Outliers:** We plotted boxplots of various variables and arranged them in a descending order in order to find any outliers if present. Luckily, there were no outliers.

### **Removing insignificant variables:** The following variables contained Unary values and were therefore insignificant in contributing to our analysis, so we directly removed them.

### Over18

### StandardHours

### EmployeeCount

### **Transforming categorical variables into dummy variables:** After the removal of insignificant variables, our dataset had 32 variables out of which following 16 were categorical. They were converted into dummies for individual values.

### Department

### BusinessTravel

### Education

### EducationField

### EnvironmentSatisfaction

### Gender

### JobInvolvement

### JobRole

### JobLevel

### JobSatisfaction

### MaritalStatus

### OverTime

### PerformanceRating

### RelationshipSatisfaction

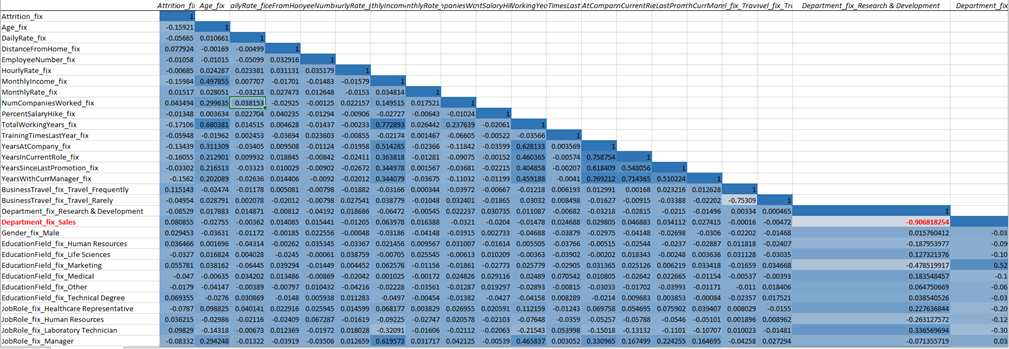
### WorkLifeBalance

### StockOptionLevel

### **Removing reference variables:** The category that had the highest frequency in each categorical variable was removed as a reference variable.

### **Correlation Analysis**

### The full correlation matrix was huge (a part of it is shown below). The group used a cutoff of 0.80 to remove variables which were highly correlated with each other. The 2 variables that showed highest correlation of -0.907 were Department\_Sales and Department\_Research & Development. We decided to remove both based on their correlation with each other and with the outcome variable “Attrition”. After their removal, the final dataset ready for further analysis had 62 variables.



### 

### **Data partitioning**

We did a partitioning of 75:25 to partition the dataset into training and validation sets. This partition will be used for all of the following data mining methodologies to measure the accuracy of each model.

**DATA MINING TASKS**

Since the primary goal of this project was to find out whether employee leaves the company or stays, the data mining goal was determined to be that of classification. We performed classification based k-NN, Logistic Regression, and Classification tree to execute classification. After implementing all three models, we compared them and decided to go with the model which was the most accurate one out of them all.

**Classification based k-NN:**

This algorithm is used for classification of the categorical outcome. To classify a new record, this method relies on finding similar records in the training data. These neighbors are then used to derive a classification for the new record. A new record is classified by a majority vote of its neighbors, with the new record being assigned to the class most common amongst its K nearest neighbors.

In our project, Attrition is the outcome variable which is categorical in nature with values “Yes/No”. Hence, this method would be appropriate to classify. We chose an optimal value of k=10 (reason discussed in next section) which was precise and reduced the overall noise.

**Logistic Regression:**

Logistic regression is similar to linear regression, except that it is used with a categorical response. We used Logistic Regression to reduce the number of predictors. As we need to classify the output variable, it is a classification problem and we use logistic regression. We used three methods of logistic regression - Forward selection, Backward Selection, and Stepwise selection. We built all these models on the training data and then tested it on the validation data. The predictors are related to the response Y via a nonlinear function called the logit.

Our output variable is Attrition and all other variables are input variables. We started with 0.5 cut-off but as we got the class 1 error too large, we reduced the cut-off value to 0.15. We compared all the three models and chose Forward selection as our best model as it gives low overall error and low class 1 error.

**Classification Tree:**

Classification tree method is used in our project since it is a recommended algorithm when the data mining task contains classifications of outcomes, and the goal is to produce set of rules that can be easily explained. A classification tree is built through recursive partitioning which is an iterative process of splitting data into partitions, and then splitting it up further on each of the branches. The objective is to obtain homogenous set of labels.

**RESULTS**

**Summary Table:** The following table summarizes the results from all the methods.

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter** | **k-NN** | **Logistic Regression** | **Classification Tree** |
| **1. Overall Error** | 28.8828 | **20.4359** | 24.9320 |
| **2. Class 1 Error** | 35.1351 | 29.7297 | 12.8834 |
| **3. Class 0 Error** | 27.3037 | 18.0887 | 27.0213 |
| **4. ROC AUC** | 0.7410 | 0.8312 | 0.8743 |

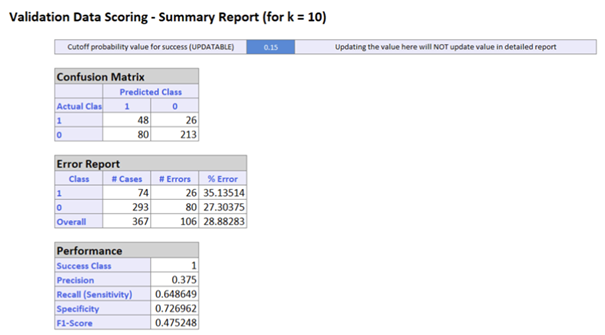
From the above result table, we can see that Classification Tree has highest ROC AUC as 87.43% and lowest Class 1 error as 12.8834 but has higher overall error when compared to Logistic Regression. Also, k-NN has higher overall error. Hence, we have selected **Logistic Regression - Forward Selection** as our final model.

**Additional Tables/ Figures and Interpretation of Results:**

**1. K Nearest Neighbors (kNN):**

* The output variable for our dataset is Attrition which is a categorical variable with the values of Yes and No. Hence, we used classification based kNN method for classifying our categorical outcome.

**Interpreting the results**



* **Value of k**

o Initially we took a range for k : 1 to 20

o The best k was at k=1 with a class 1 error of 75%. Since this error was quite high and

there was an imbalance between class 1 and class 0 error, we decided to try with

different values of k and got the best results for **k=10** as follows:

o Class 1 error : 35.13 %

o Class 0 error : 27.3%

o Overall error : 28.8%

* **Cutoff = 0.15**

o A cutoff less than 0.5 ends up classifying more records as 1’s and a cutoff greater

than 0.5 will end up classifying fewer records as 1’s.

o We decided to use a cutoff value less than 0.5 because our main objective was to

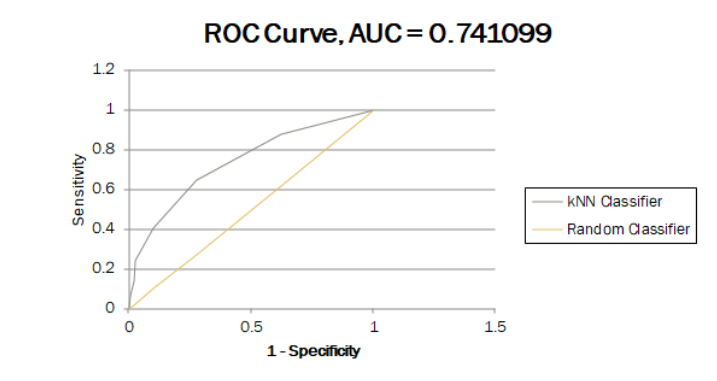
classify more class 1 records which comprised of the employees who left the

company (attrition = yes).

o We tried with different values of the cutoff (0.2, 0.25, 0,4) but class 1 error was quite high as compared to class 0.

o We got the best results for cutoff =0.15 and hence decided to go with it.

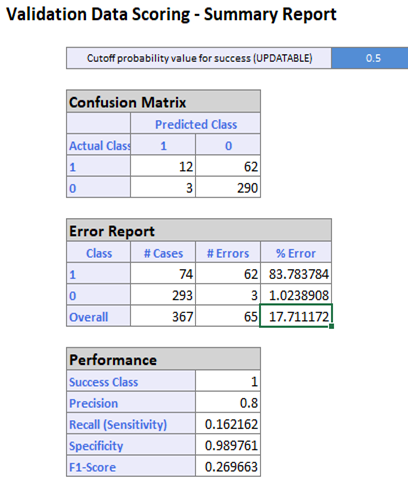
**Receiver Operating Characteristics (ROC)**

****

* Better performance is reflected for the curves where AUC is closer to 1.
* As depicted in the above ROC curve for kNN, the Area under curve (AUC)= 0.741099 which implies that it is a good model, but we get better results with logistic regression as described in the next section.

**2.** **Logistic Regression**:

With 0.5 as the cut-off -

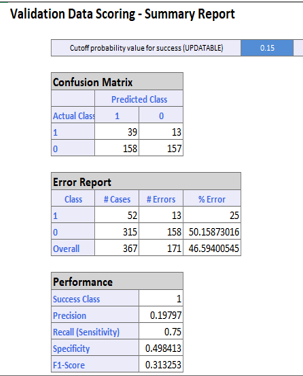
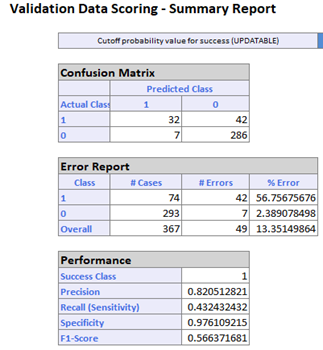


As we can see, the class 1 % Error is very high i.e. **83.7837**. Hence, we decided to reduce the cut-off value. A cutoff less than 0.5 ended up classifying more records as 1’s.

We tried different values of cut-off such as 0.4, 0.3, 0.25, 0.2. Finally, we achieved the best results with cutoff = **0.15**.

Similar to Linear Regression, we can also use variable selection methods to reduce the number of predictors.

Here, we have used Forward Selection, Backward Selection and Stepwise Selection methods to reduce the predictors.



Backward Selection Stepwise Selection

For both Backward and Stepwise selection methods, we used the training data set to build the model and tested it on the validation data.

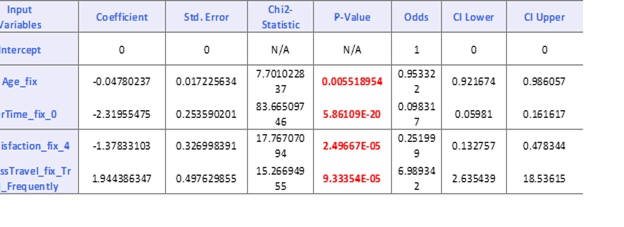
For Backward Selection,

Overall error is low = 13.3514. Class 1 error is **56.7567** and the class 0 error is 2.3890. We don’t go forward with this model because, the class 1 error is high compared to the class 0 error which is not desirable.

For Stepwise Selection,

Overall error= **46.5940.** Even though the class 1 error is low, the overall error is too high. Hence, we don’t proceed with this model.

Now, we consider Forward Selection method. Similar to previous methods, we used training data set to build the model and tested it on the validation data.



From the above table, we can see that the Predictors Age\_fix, OverTime\_fix\_0, JobSatisfaction\_fix\_4 and BusinessTravel\_fix\_Travel\_Frequently have the p-values less than 0.05. Hence, these variables are significant variables.

From this table, with the help of coefficients we can build a Logit function. This function is used to relate the predictors to the response Y i.e. Attrition in this case.

**Logit (Attrition) = 0.0056432 - 0.04780237\* Age\_fix - 2.31955475\* OverTime\_fix\_0 - 1.37833103\* JobSatisfaction\_fix\_4 + 1.944386347 \* BusinessTravel\_fix\_Travel\_Frequently**

Interpreting the Logit function:

1. For every unit increase in Age\_fix (Continuous variable), the Attrition will decrease by 0.04780 keeping other variables unchanged.
2. For every unit increase in OverTime\_fix\_0 (Binary variable), Attrition will decrease by 2.31955475 keeping other variables unchanged.
3. For every unit increase in JobSatisfaction\_fix\_4 (Binary variable), Attrition will decrease by 1.37833103 keeping other variables unchanged.
4. For every unit increase in BusinessTravel\_fix\_Travel\_Frequently (Binary variable), Attrition will increase by 1.944386347 keeping other variables unchanged.

*Binary predictor effects*:

OverTime\_fix\_0 and JobSatisfaction\_fix\_4 are associated with employees not leaving the jobs.

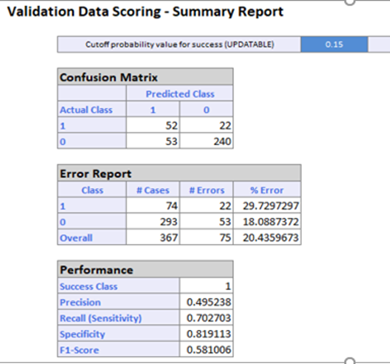
BusinessTravel\_fix\_Travel\_Frequently is associated with employees leaving the jobs.

*Continuous predictor effects*:

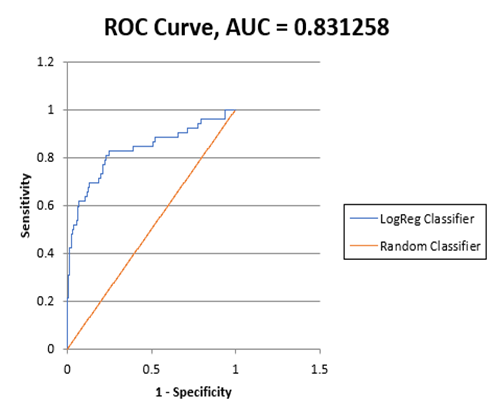
Age\_fix is associated with the employees not leaving their jobs.

The performance measures used to compare the three models are Confusion Matrix, Misclassification Error Rate from validation data and ROC.

As we can see in the summary report below, the overall error is **20.4359**. Also, the class 1 error = 29.7297 and the class 0 error = 18.0887 are low as well.



**Receiver Operating Characteristics (ROC)**



ROC Curve is a more popular method for plotting the two measures is through ROC (receiver operating characteristic) curves. ROC helps determine if the model does a good job of identifying members of a particular class. It plots the pairs {sensitivity, 1-specificity} as the cut-off value increases from 0 and 1. The sensitivity of a classifier is its ability to detect the important class members correctly. Here, Area Under Curve (AUC) is **0.831258**. As AUC is closer to 1, we can say that our model is a good model.

When we compared the three models, **Forward Selection** method gives us the best results i.e. low overall error, low class 1 error, low class 0 error.

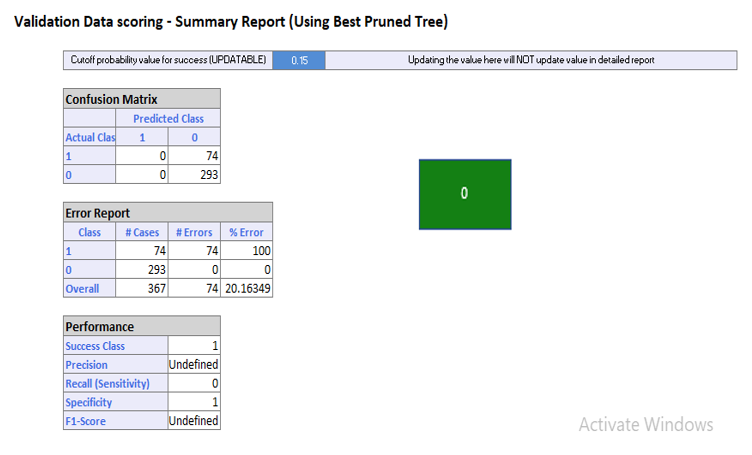
**3. Classification Tree:**

The cutoff for all the trees is considered as 0.15.

The data was split as 75-25

**Best Pruned Tree:**

The tree is one standard deviation above the minimum error tree.

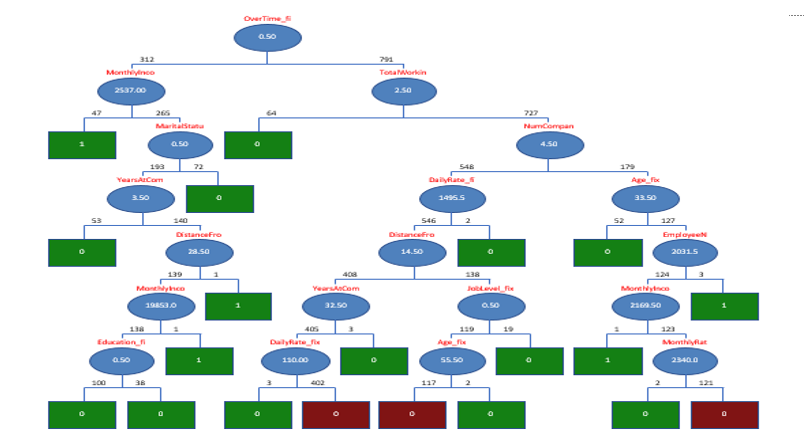


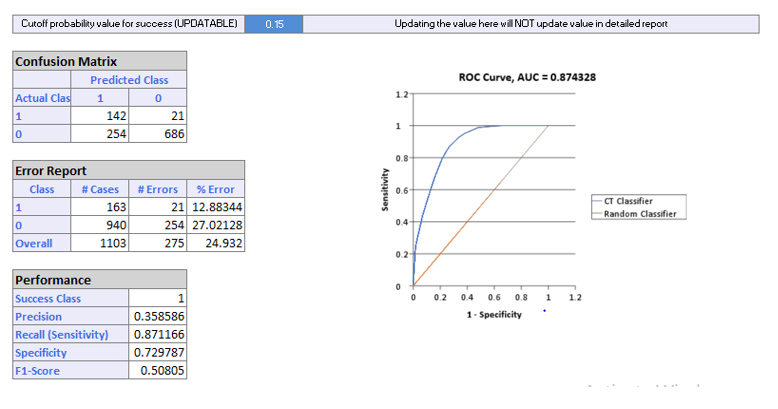
As it can be verified in the output, the best pruned tree came out to be zero in our model.

As shown in the confusion matrix, all the values of 1 are misclassified as zeros. Therefore, there is 100 % error in class one. We do not want this in the model as we are more interested in finding out the class 1 outcome. The overall error is 20.16%.

**Full grown tree:**

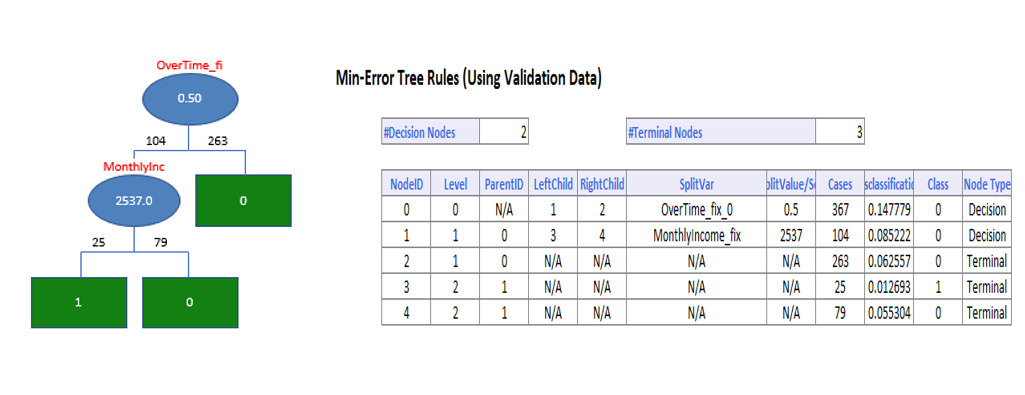
As shown below, the error for class 1 for full grown tree is 12.88%. Whereas, the overall error is 24.93%. The problem with full grown tree is that it considers all the features. This may lead to overfitting of the data. To avoid these problems, we consider taking the minimum error tree.





**Minimum Error Tree:**

This tree was proved to be the best for our model. As seen from the tree below, there are two nodes for this tree: Overtime and Monthly Income. 263 workers did not do overtime at all and hence were classified as Class 0, whereas 104 workers did overtime. These 104 workers were further classified based on the monthly income. 79 workers had their monthly income greater than 2537 and were satisfied with the job. On the other hand, 25 workers were not satisfied with the job and were found positive in contributing to the attrition rate.



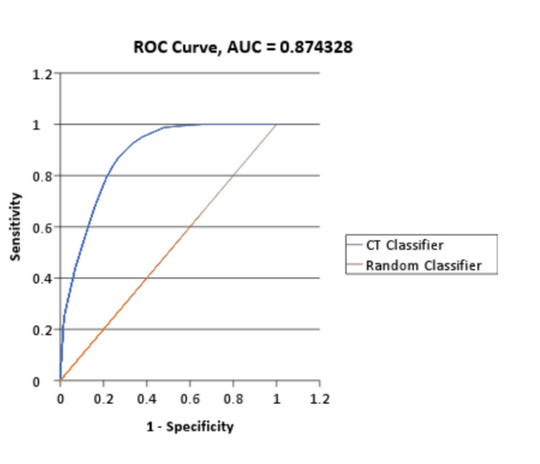
**Interpretation based on the minimum error tree:**

Concluding based on the results, the most important factors leading to the attrition rate were overtime and Monthly income.

Working overtime does affect the satisfaction derived from any job. No worker will be happy if he/she is given continuous overtime shifts.

Monthly income is also an obvious factor in deciding the outcome (Attrition rate). If the monthly income is low, the workers will be tempted to leave the company for better opportunities.

**Receiver Operating Characteristics (ROC)**

****

As can be seen from above plot, the AUC of Classification Tree comes out to be 0.8743 which is approximately equal to 87%. Hence, this model proves to be quite good.

**CONCLUSION**

The goal for this dataset was to find out what are the characteristics of the employees that are most likely to leave their jobs.

After due data exploration we uncovered the patterns and relationships between predictors. Initially the cut off was set at 0.50 but that had very high class one error then after trying various values the cutoff of 0.15 was found to be satisfactory. We implemented three learning methods namely k-NN, Logistic Regression and Classification tree. Of the three the overall percentage error and class 1 error was lowest for Forward selection of Logistic Regression hence we decide to go with that method. Having a method with low class 1 error was important for us as we wanted to accurately classify the 1’s in the dataset.

The main factors that affected the attrition rate of IBM employees are Age, Overtime, Job satisfaction and Travel frequency.

**Age:** It can be concluded that as the age of the employee increases he is less likely to leave his job.

**Overtime:** As the overtime per employee is high it is evident that employees tend to leave their jobs.

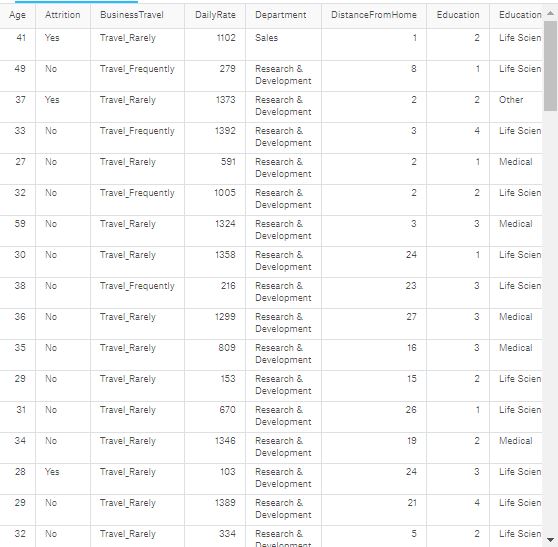
**Travel Frequency:** The attrition rate increases as the employee’s travel frequency increases.

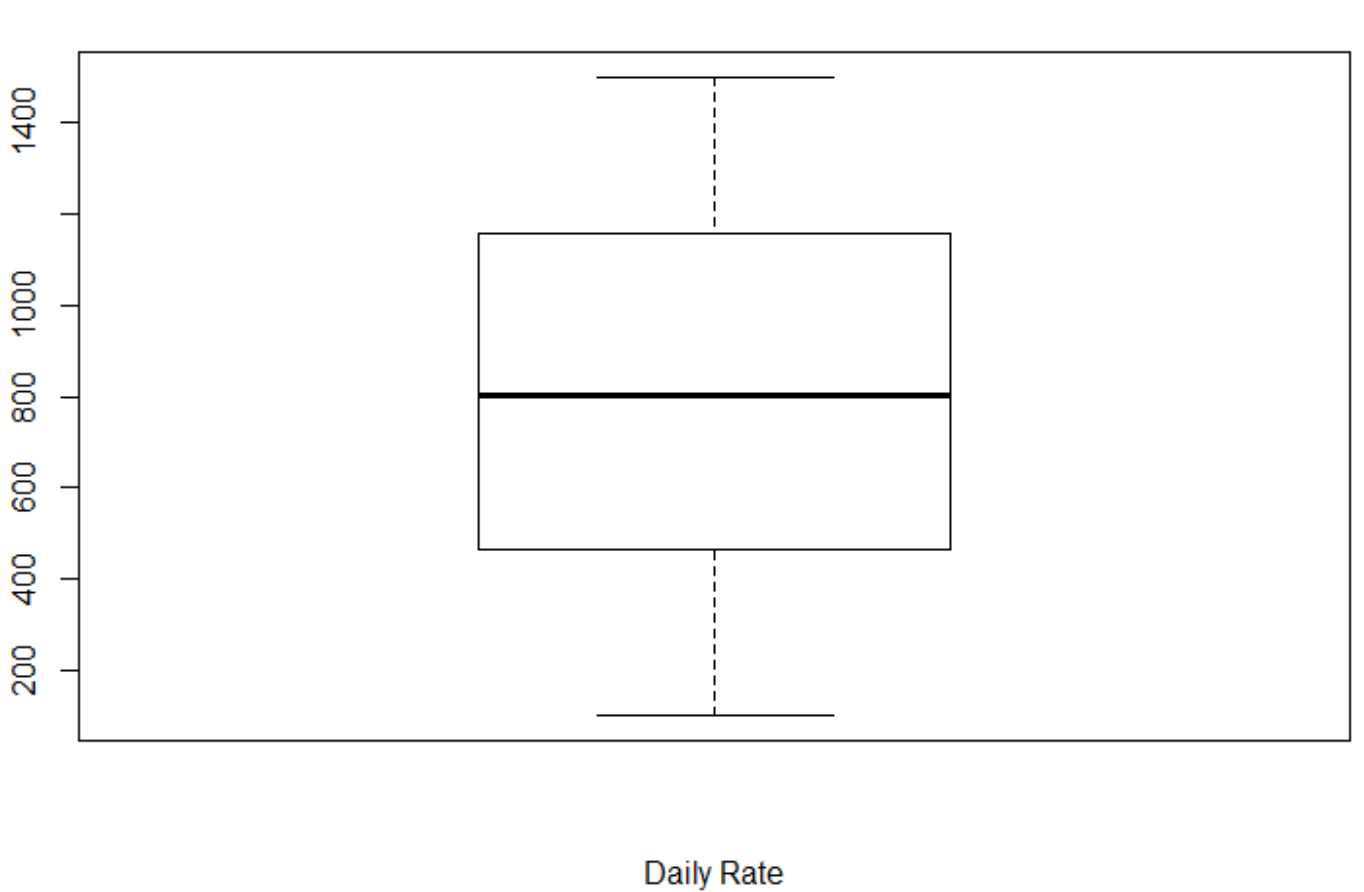
**Job satisfaction:** It is evident from the dataset that the as job satisfaction declines employee’s are more likely to look for a change in their jobs.

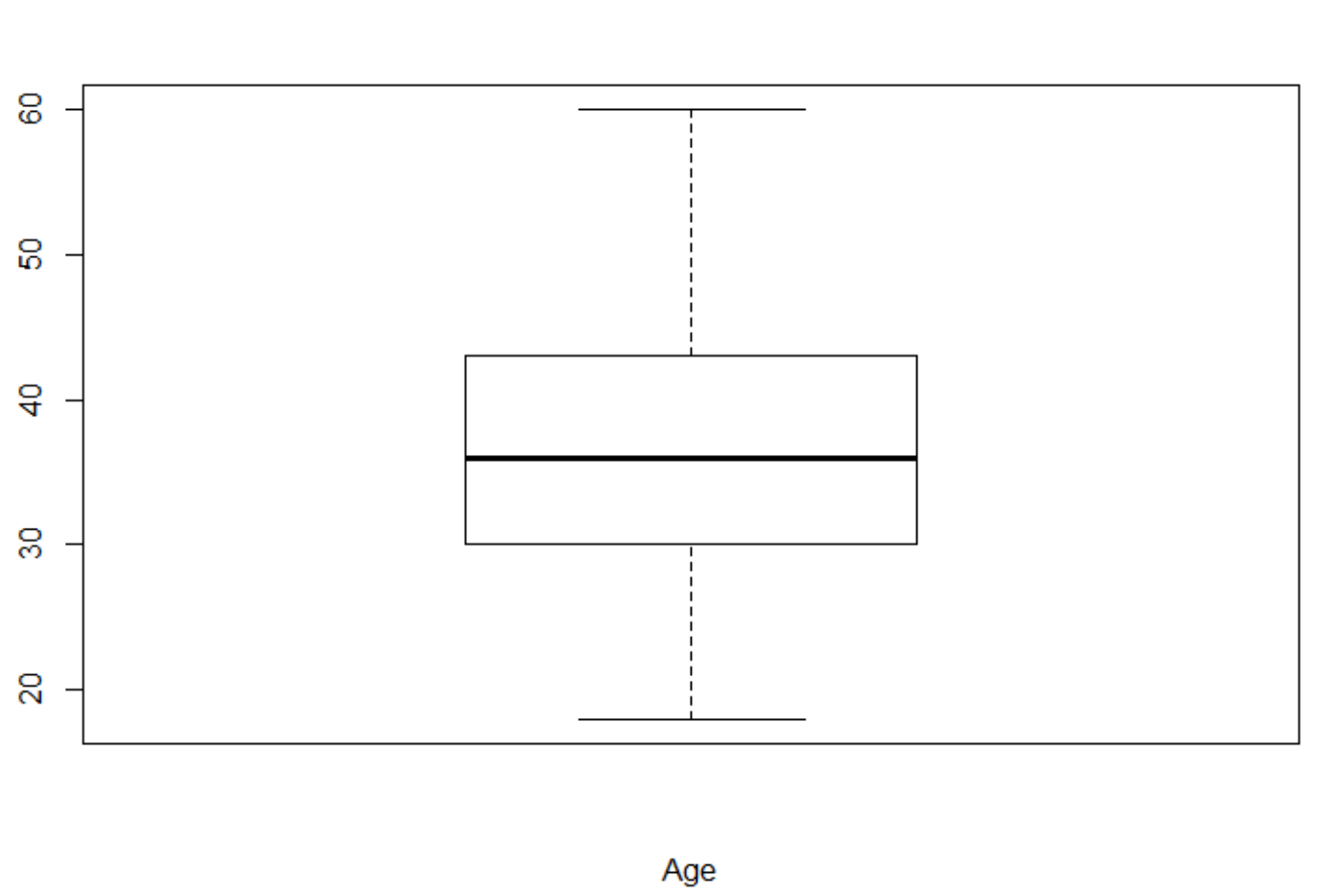
**Recommendations** for IBM would be that they should focus on their young employees as the attrition rate in this particular age group is the highest. IBM should consider limiting the overtime per employee as that can disrupt the work life balance of an employee making him leave his job. Travel frequency also should be limited or should be scheduled only when necessary to retain their employees.

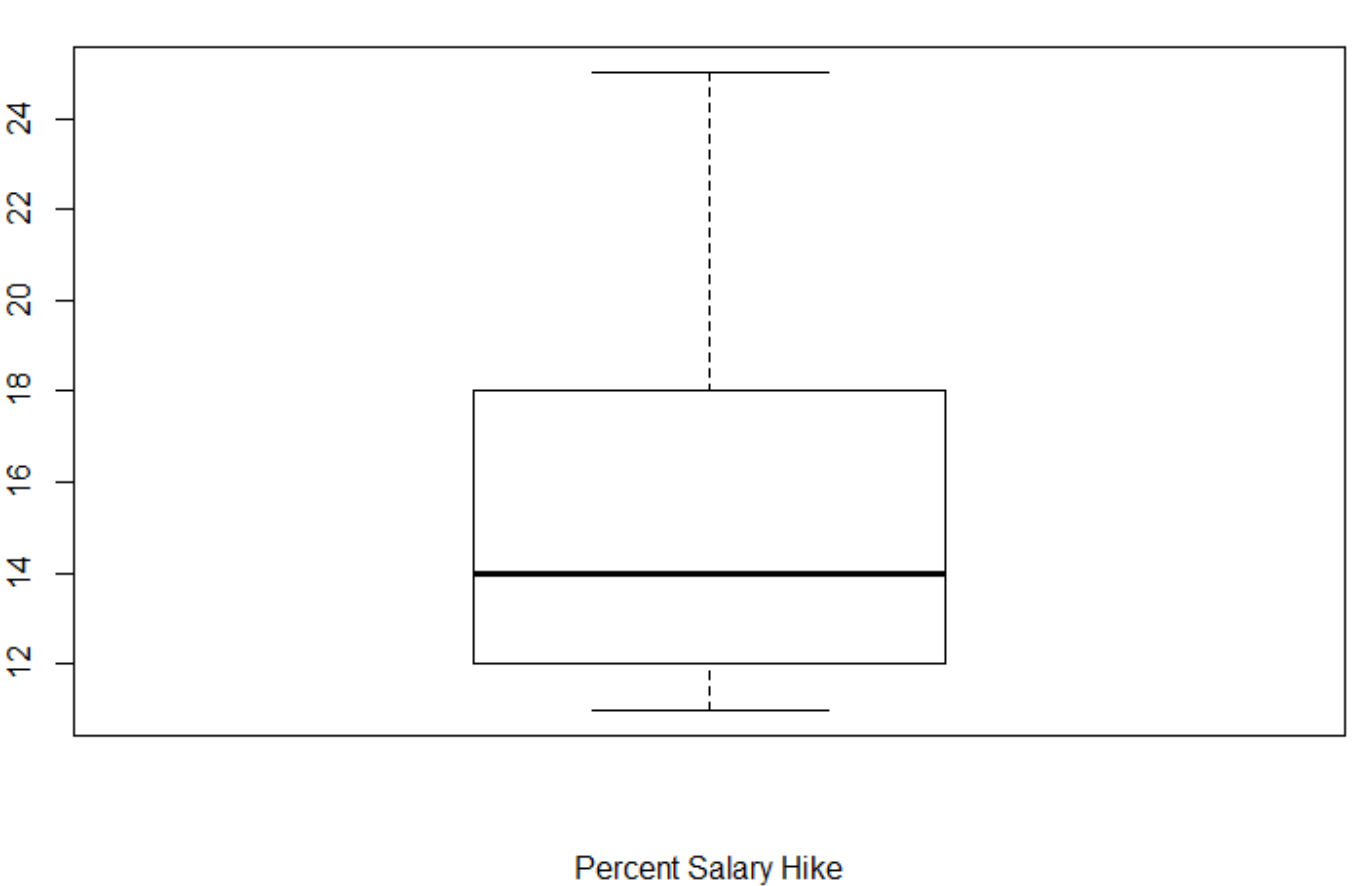
**APPENDIX**

**Appendix A: Portion of original dataset:**



**Appendix B: Some of the boxplots for detecting outliers**





**REFERENCES**

*Shmueli, Galit, Patel, Nitin, & Bruce, Peter (2010)*. Data Mining for Business Intelligence. New

Jersey: John Wiley & Sons, Inc.