Table of Contents

Executive Summary

A dataset pertaining to the HR department of an organization was used to make insights in identifying the key drivers of attrition and monthly salary. This dataset helped solve problems in classifying the employees who were going to leave the organization versus the others and predicting the monthly income drawn by employees of the same company. Among the 35 variables of the dataset, key variables had to be picked by variable selection methods for model building. The list of variables picked up by the forward selection and backward selection for MLR and the rpart functions for CART gave almost same variables which proved that the selection mechanisms were similar for the different modelling techniques. Three classification models were built using Logistic Regression, K-Nearest Neighbors and Classification tree to best solve the issue of attrition. Three prediction models were built using Multi-linear regression, K-Nearest Neighbors and Regression tree to predict the values of monthly income based on explanatory variables. These models built were evaluated based on the accuracy measures and recommendations were made on the key driving factors in the best fitted model. It was interesting to find that the variation of accuracy of the three methods for the classification and prediction problem varied by a very small margin. The Root Mean Squared Error (RMSE) was used as a constant metric to evaluate the prediction model of monthly income. The Misclassification Error Rate was used as a constant metric to evaluate the classification model built using the three techniques. The Multiple Linear Regression Model for prediction and the Logistic Regression model for classification gave us the best accuracy measures.

Introduction

The foundation of data mining and the concepts of model building learnt in this Data Mining class laid the base for solving an analytics problem. The entire process of identifying the right dataset, cleaning the data, interpreting the data, splitting the data, building models, evaluating models to identify best performance helped us greatly in completing this project.

Background

Initiating the project, we were looking to apply the model building concepts learnt in a niche space where analytics was not predominantly used. To our surprise we came across an HR department dataset pertaining to the employees of a company. This dataset originally had 35 variables and 1470 records as shown in Table1. There were several categorical and continuous variables that helped us derive on solving two problems with respect to this dataset.

Problem Description

The 2 questions that we intended to solve through this project are given below:

1. Why do employees leave the firm? How do we classify employees who would leave?
2. How do we predict the monthly salary earned by an employee?

Description of Variables

	Variable Name	Varable Description	Data Type / Variable Type
1	Age	Age of Employee	Integer
2	Attrition	Currently Working or Left	Categorical
3	BusinessTravel	How often does he travel?	Categorical
4	DailyRate	Daily Pay Scale	Integer
5	Department	Department of Employment	Categorical
6	DistanceFromHome	Distance from Home	Integer
7	Education	Education Qualifications	Integer
8	EducationField	Field of Education	Categorical
9	EmployeeCount	Internal Value	Integer
10	EmployeeNumber	Unique ID	Integer
11	EnvironmentSatisfaction	Satisfaction Rating of Employee	Integer
12	Gender	Male/Female	Categorical
13	HourlyRate	Hourly Pay	Integer
14	JobInvolvement	Rating on Job Involvement	Integer
15	JobLevel	Seniority of Role	Integer
16	JobRole	Designation	Categorical
17	JobSatisfaction	Rating on Employee Satisfaction	Categorical
18	MaritalStatus	Married/ Unmarried	Categorical
19	MonthlyIncome	Take home monthly income	Integer
20	MonthlyRate	Monthly Pay	Integer
21	NumCompaniesWorked	Number of previously worked companies	Integer
22	Over18	Is age over 18?	Categorical
23	OverTime	Does he work overtime?	Categorical
24	PercentSalaryHike	Salary Hike	Integer
25	PerformanceRating	Rating received in Evaluation	Integer
26	RelationshipSatisfaction	Satisfaction of Relationship	Integer
27	StandardHours	Number of work hours	Integer
28	StockOptionLevel	Category of stock levels received	Integer
29	TotalWorkingYears	Years of Experience	Integer
30	TrainingTimesLastYear	Count of Trainings attended	Integer
31	WorkLifeBalance	Ratings on Work Life Balance	Integer
32	YearsAtCompany	Years at this company	Integer
33	YearsInCurrentRole	Years at this role	Integer
34	YearsSinceLastPromotion	Years since last promotion	Integer
35	YearsWithCurrManager	Years with the current manager	Integer

Table 1:- Variable Description

Descriptive Statistics

Descriptive statistics was performed on all the variables. Statistics for the outcome variables monthly income and attrition is listed in the table below. Charts of data analysis for the variables is attached in the subsequent analysis section.

MonthlyIncome	
Mean	6484.278545
Standard Error	126.802849
Median	4883
Mode	2342
Standard Deviation	4701.975483
Sample Variance	22108573.44
Kurtosis	1.030395106
Skewness	1.378174962
Range	18990
Minimum	1009
Maximum	19999
Sum	8915883
Count	1375

Attrition	
No of Yes	226
No of No	1149
Percentage of Yes	16.4%
Percentage of No	83.6%

Table 2. Descriptive Statistics

Data Preprocessing

This section contains a discussion on the purposes and methods used for reduction of the data. The first step in this data mining project is to refine the raw data as shown in (Fig-1) by removing the N/A variables or irrelevant variables. The dataset contains missing variables and these will have to be handled using an appropriate imputation method.

In our project, we are working on HR Dataset which contains 1470 rows and 35 columns. After carefully reviewing the whole dataset, and implementing it in R Studio we came to a conclusion that columns YearsAtCompany and RelationshipSatisfaction have missing data more than 30% of the data. We plotted a heat map for the raw dataset and the results is shown below in Fig1.

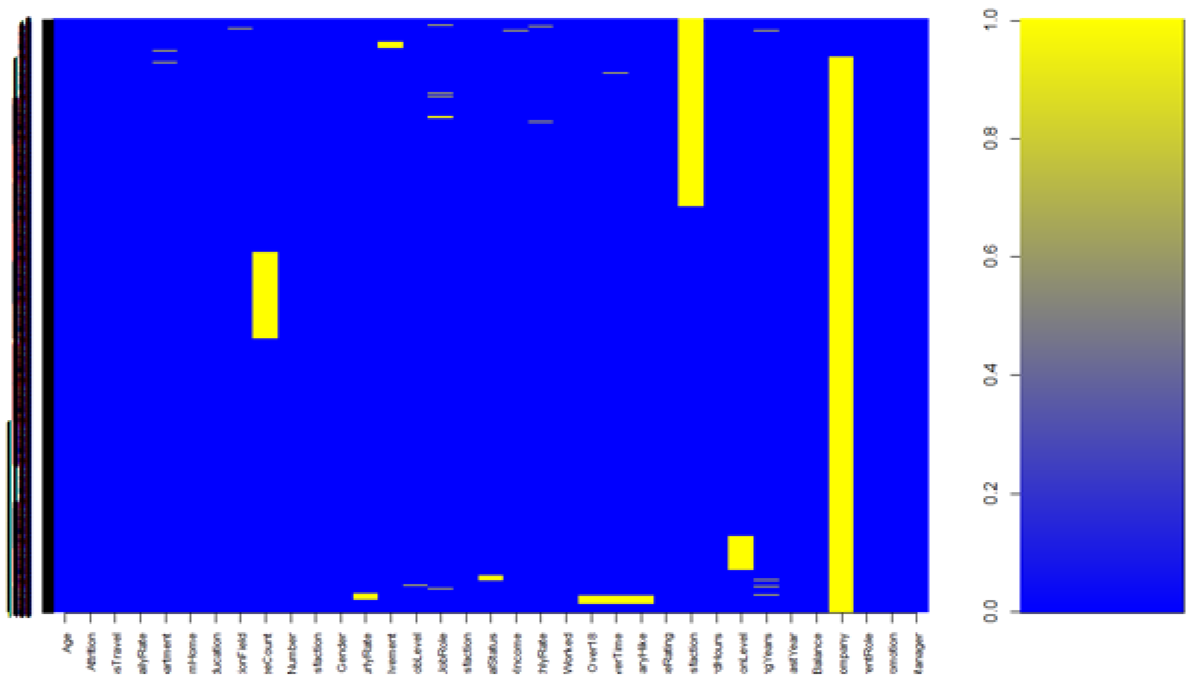

Fig 1: Raw Dataset

After removal of YearsAtCompany and RelationshipSatisfaction, the data has 1470 rows and 33 columns. The visualization of the data is shown below.

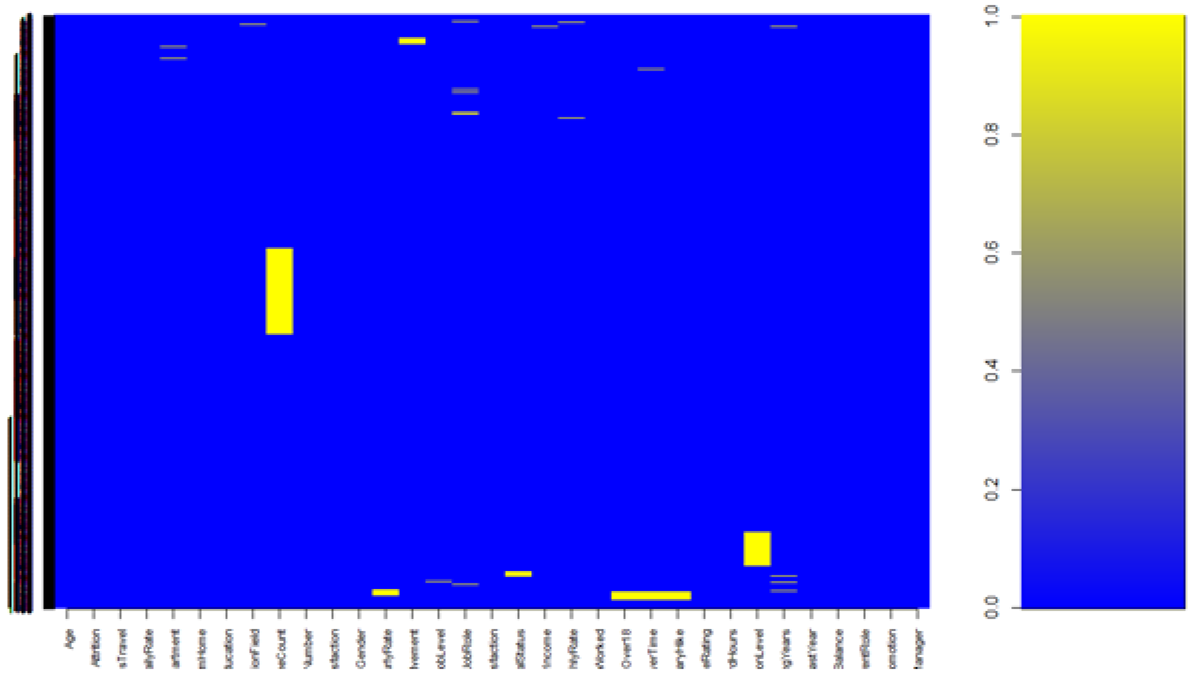

Fig 2: Removing Columns having more than 30% missing value

After carefully reviewing the new dataset , it was noticed column name EmployeeCount contains all the value as 1, so we did the median imputation for that column8 and filled out the missing value by the median of the whole column. We also found out another column i.e. StockOptionLevel where we could fill out the missing value by doing median imputation on the column. So, plotting the heat map for the dataset after performing mean imputation the data set is as illustrated in Fig.3.

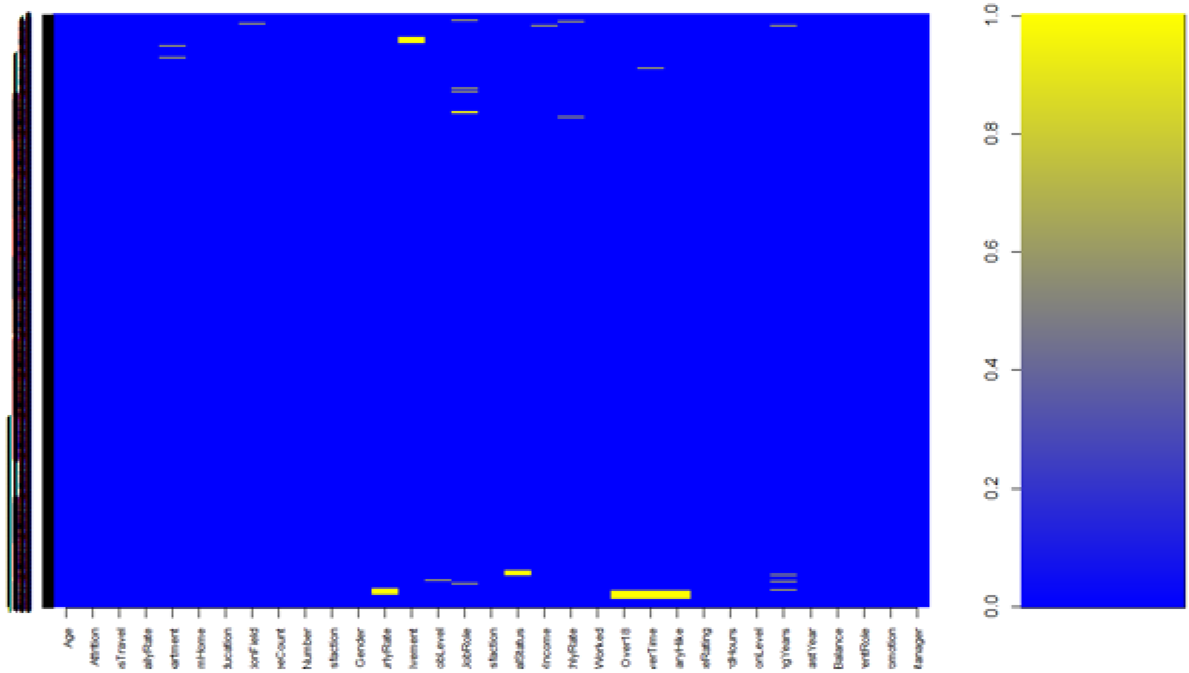

Fig 3: Median Imputation Method

After performing mean imputation we are done with reducing the columns in the dataset. So, now we will move forward in reducing the rows. After going through the rows in the dataset, we could not find any relevant pattern where we could apply any imputation methods. So, we deleted the rows which had missing value in the dataset. The heatmap of the cleaned dataset after removing the rows that contains missing values is as shown in Fig 4.

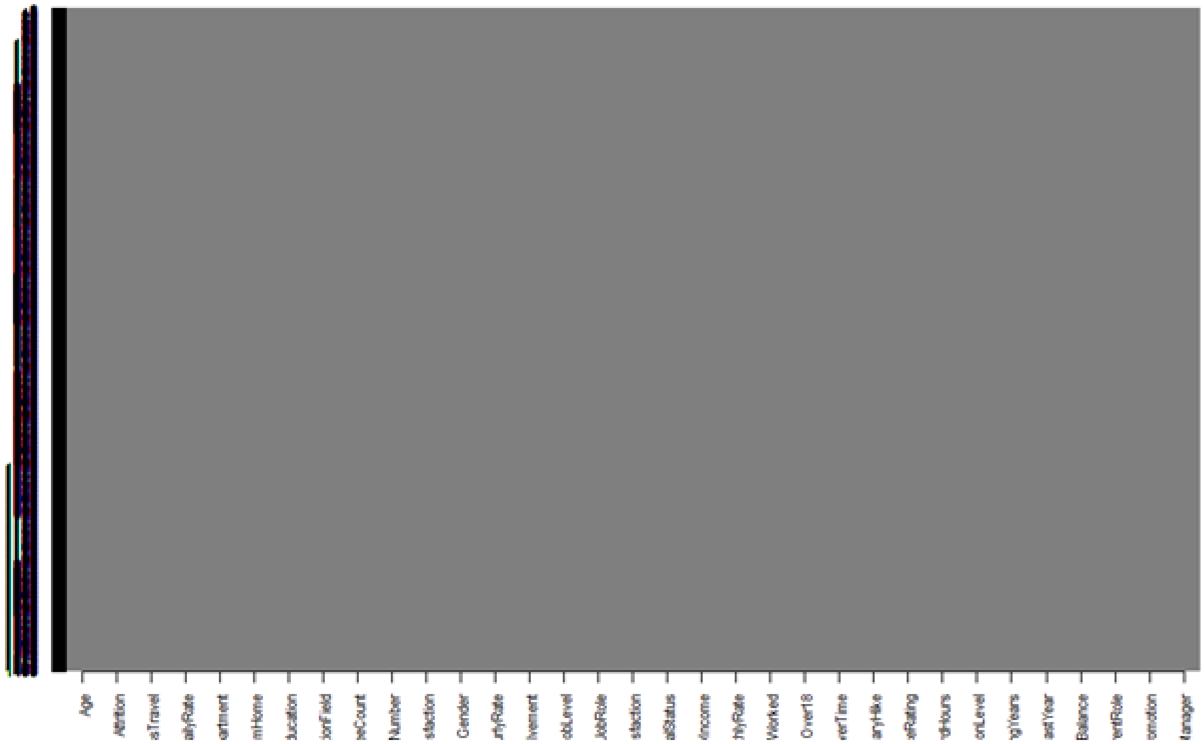

Fig 4: Clean Dataset

We can clearly see from the above figure-4 that there are no missing values in the dataset. But the main question is can we proceed with this datasets for the data mining task? The answer is “NO”. We need to create dummy variables and create categorical data into numeric data where needed.

For prediction purpose, we need to convert character values into numeric values so we assigned 1 to Yes and 0 to No in the Attrition column. Similarly in Gender column, we assigned 1 to Male and 0 to Female and for OverTime column, we assigned 1 to Yes and 0 to No. Since some predictive models require the use of dummy variables, the categorical variables were converted into dummies for individual variables. We created dummy variables for columns MaritalStatus, Department and JobRole. After doing all the steps for data reduction we are left with 1375 rows and 48 columns to proceed upon prediction and classification.

Sampling and Partitioning

The original dataset includes 1375 records. To perform any prediction and classification tasks we need the data partition into training data and validation data. In our project, we divided the data into 60-40%. Training data i.e. 60% consists of 825 rows and 48 columns while validation data i.e. 40% consists of 550 rows and 48 columns.

Correlation Plot

The correlation coefficient is a way to determine how one variable tend to change when other does. The sign of the correlation coefficient indicates the direction of the association. Positive sign indicates a strong relationship while as negative sign indicates a weak relationship. In our project, we have used “ggcorrplot” library for better visualization purpose. It can be seen from the correlation figure5 that our data in the datasets are not highly correlated and we can proceed further with the dataset.

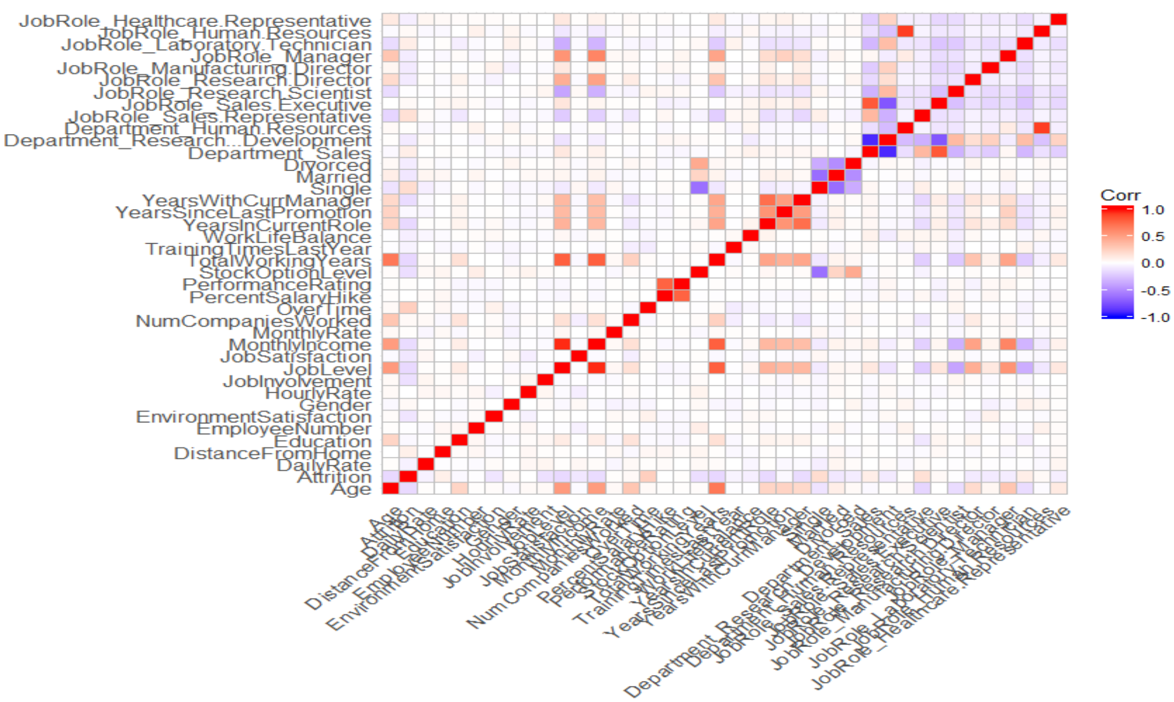

Fig 5: Correlation plot

Exploratory Data Analysis:

Distribution of continuous variable

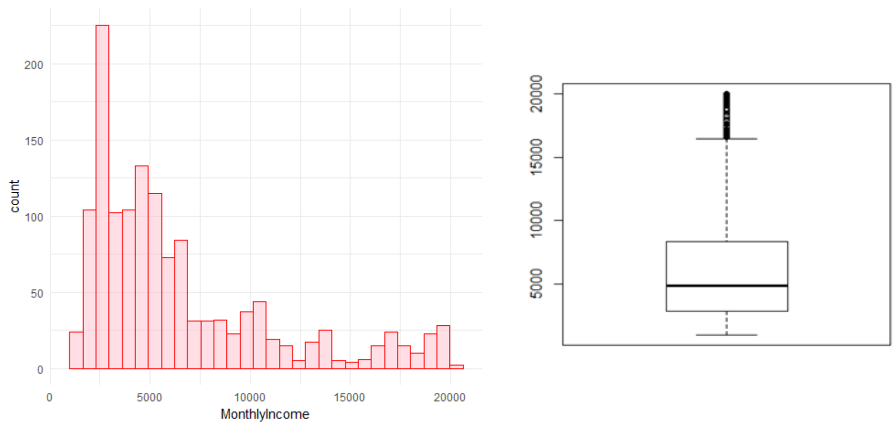

Fig 6: Histogram and boxplot for Monthly Income

The left figure is a histogram for monthly income of Employees. For example, for the 250 count the monthly income is 3000. The right figure is a box plot where the median Income is 5000 and there are outliers of employees above 16,000 salary. A classification model is built for the monthly income.

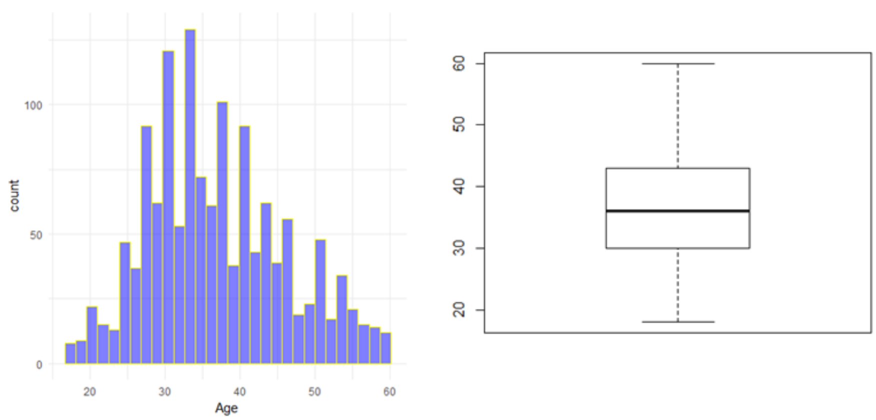

Fig 7: Histogram and boxplot for Age

For the next example, the left figure shows a histogram for age and count and on the right side is the box plot for where the median age is 35 and there are no outliers.

Distribution of Categorical Variables:

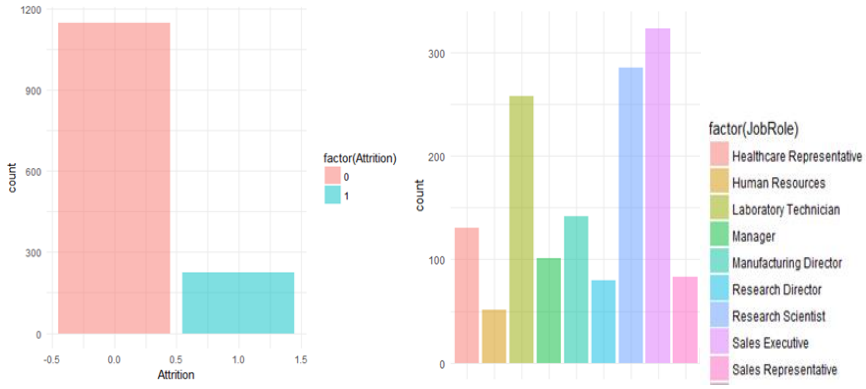

Fig 8: Barplot for Attrition and JobRole

Two categorical variables are described which are Attrition and Job Role. The left Bar Plot s for attrition and it clearly shows that the 0 factor is greater than 1 factor. 0 Factor means the employees who stayed and 1 Factor means the employees who left. The results showed that people tend to stay at their jobs rather than changing them frequently. A classification model was prepared by us just to determine why this was happening.

The right figure shows a barplot for Job Roles and attached are different roles applicable and their count.

Methods Employed:

Prediction of Monthly Income	**Classification of Attrition**
Multiple Linear Regression	Logistic Regression
k-NN	k-NN
Regression Tree	Classification Tree

Model Building

Prediction Problem: Prediction of Monthly Income

1. **Multiple Linear Regression (MLR)**

Regression equation is the mathematical formula is applied to the explanatory variables to best predict the dependent variable. Regression analysis is often used for prediction of a variable and thus answers the why question. In this particular case prediction of monthly Income from the explanatory variables is the question. A typical expression representing the elements of an Ordinary Least Squares Regression (OLS) is illustrated below in Fig 9.

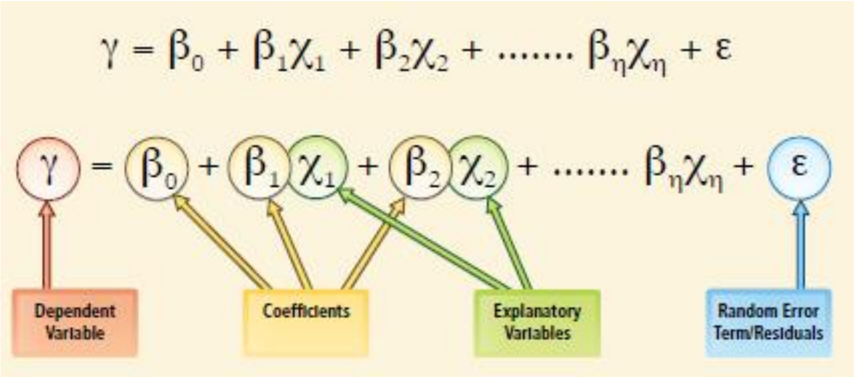

Fig 9: Elements of an OLS Regression (Scott L,ESRI)

Variable Selection

Selection of significant explanatory variables of prior to building the best regression model is necessary. Variable selection is performed on training data whereas model accuracy is tested on validation data.The four popular variable selection methods often implemented in statistical tools are :

1. Forward Selection
2. Backward Selection
3. Stepwise selection

4. Best subset selection

The first two methods forward and backward selection were simulated in R programming language utilizing RStudio for this particular project. Package “MASS” facilitates variable selection methods via function ‘step’. Snapshots of the the models along with significant variables for respective variable selection methods are illustrated below . Algorithm iterates based on Akaike information criterion (AIC) with the model with lowest AIC selected at convergence.

Forward Selection:

```
Call:
lm(formula = MonthlyIncome ~ JobLevel + JobRole_Manager + JobRole_Research.Director +
    TotalWorkingYears + JobRole_Laboratory.Technician + JobRole_Sales.Representative +
    JobRole_Research.Scientist + JobInvolvement + YearsWithCurrManager +
    Department_Human.Resources + Age + YearsSinceLastPromotion +
    HourlyRate + PerformanceRating + Divorced, data = train)


Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   1602.145    488.396   3.280  0.00108 ** 
JobLevel                      2622.620     91.971  28.516  < 2e-16 ***
JobRole_Manager               4301.011    210.262  20.455  < 2e-16 ***
JobRole_Research.Director     4089.666    218.886  18.684  < 2e-16 ***
TotalWorkingYears               66.169     10.338   6.401 2.62e-10 ***
JobRole_Laboratory.Technician -928.022    141.310  -6.567 9.16e-11 ***
JobRole_Sales.Representative  -981.774    196.687  -4.992 7.34e-07 ***
JobRole_Research.Scientist    -645.593    139.672  -4.622 4.42e-06 ***
JobInvolvement                -151.010     55.717  -2.710  0.00686 ** 
YearsWithCurrManager           -46.545     13.913  -3.345  0.00086 ***
Department_Human.Resources    -485.031    200.088  -2.424  0.01557 *  
Age                            -11.411      6.190  -1.843  0.06565 .  
YearsSinceLastPromotion         28.483     14.577   1.954  0.05105 .  
HourlyRate                       3.711      1.954   1.900  0.05783 .  
PerformanceRating             -200.352    111.032  -1.804  0.07153 .  
Divorced                      -152.021     97.450  -1.560  0.11915    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1139 on 809 degrees of freedom
Multiple R-squared:  0.9456,    Adjusted R-squared:  0.9446 
F-statistic: 937.7 on 15 and 809 DF,  p-value: < 2.2e-16
```

Backward Selection:

```
Call:
lm(formula = MonthlyIncome ~ Age + HourlyRate + JobInvolvement +
    JobLevel + PerformanceRating + TotalWorkingYears + YearsSinceLastPromotion +
    YearsWithCurrManager + Department_Sales + Department_Research...Development +
    JobRole_Sales.Representative + JobRole_Research.Scientist +
    JobRole_Research.Director + JobRole_Manufacturing.Director +
    JobRole_Manager + JobRole_Laboratory.Technician, data = train)


Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       1098.130    503.614   2.180  0.02951 *
Age                                -11.842      6.197  -1.911  0.05636 .
HourlyRate                           3.498      1.954   1.790  0.07389 .
JobInvolvement                    -147.655     55.801  -2.646  0.00830 **
JobLevel                          2629.325     91.919  28.605  < 2e-16 ***
PerformanceRating                 -200.406    111.062  -1.804  0.07153 .
TotalWorkingYears                   66.447     10.354   6.417 2.36e-10 ***
YearsSinceLastPromotion             23.868     14.732   1.620  0.10559
YearsWithCurrManager               -44.264     13.966  -3.169  0.00159 **
Department_Sales                   473.715    207.905   2.279  0.02296 *
Department_Research...Development  645.150    222.193   2.904  0.00379 **
JobRole_Sales.Representative      -959.757    202.717  -4.734 2.59e-06 ***
JobRole_Research.Scientist        -805.966    175.595  -4.590 5.14e-06 ***
JobRole_Research.Director         3895.887    243.459  16.002  < 2e-16 ***
JobRole_Manufacturing.Director    -329.429    178.403  -1.847  0.06518 .
JobRole_Manager                   4206.976    215.199  19.549  < 2e-16 ***
JobRole_Laboratory.Technician    -1096.129    177.136  -6.188 9.67e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1139 on 808 degrees of freedom
Multiple R-squared:  0.9457,    Adjusted R-squared:  0.9446
F-statistic: 879.2 on 16 and 808 DF,  p-value: < 2.2e-16
```

Results

Both the models are highly accurate as the adjusted R2 is about 0.9446. RMSE values on validation data by forward selection is 1111.548 and backward selection is 1113.593. Thus, forward selection model was chosen for the prediction of monthly income.

Interpretation of explanatory variables:

1. Job level, Job role , Job level , years with current manager and total working years in the company are the significant explanatory variables with p-value lesser than 0.05 at 95% confidence interval.
2. p-value of the independent variables noted above < 0.05, Hence, we can reject null hypothesis. Also p-value of intercept in forward selection method is < 0.05.
3. H_0: Population slope coefficient ≠ 0, i.e the slope of the trendline is not equal to zero in the aforementioned variables.
4. Higher job level such as job role of manager or research director is associated with higher monthly income.
5. Job role of sales representative and laboratory technician indicate lower monthly salaries.

Graph of predicted values and actual values of monthly income follow a linear trend without any outliers as depicted in the figure below. Thus, the prediction model built on monthly income is accurate.

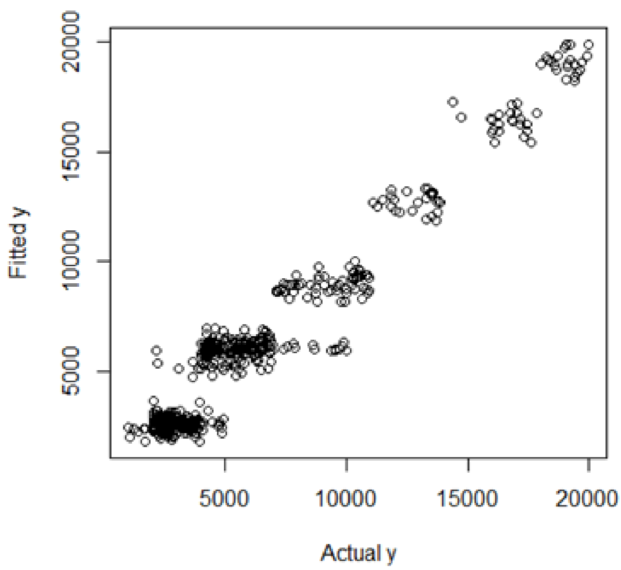

Fig:10 Actual y values vs predicted y values

Assumptions on Multiple Linear Regression Model:

i) Normality of Residuals: Residuals of the regression model is normally distributed for both forward and backward selection methods as illustrated in the figure below.

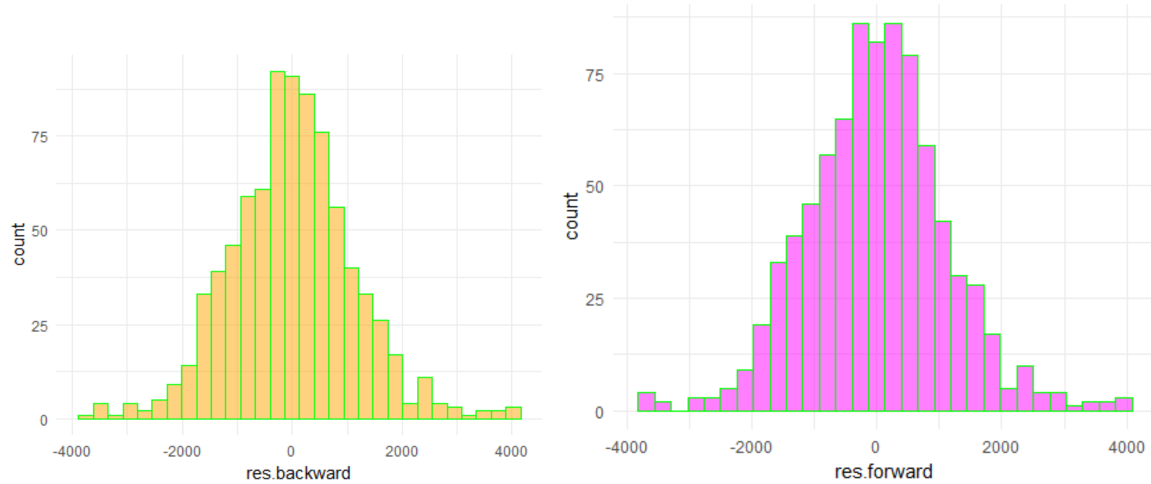

Fig 11: Histogram of residuals

ii) Linearity of Continuous Variables: Scatter plot of dependent and independent variables is linearly distributed for continuous variables.

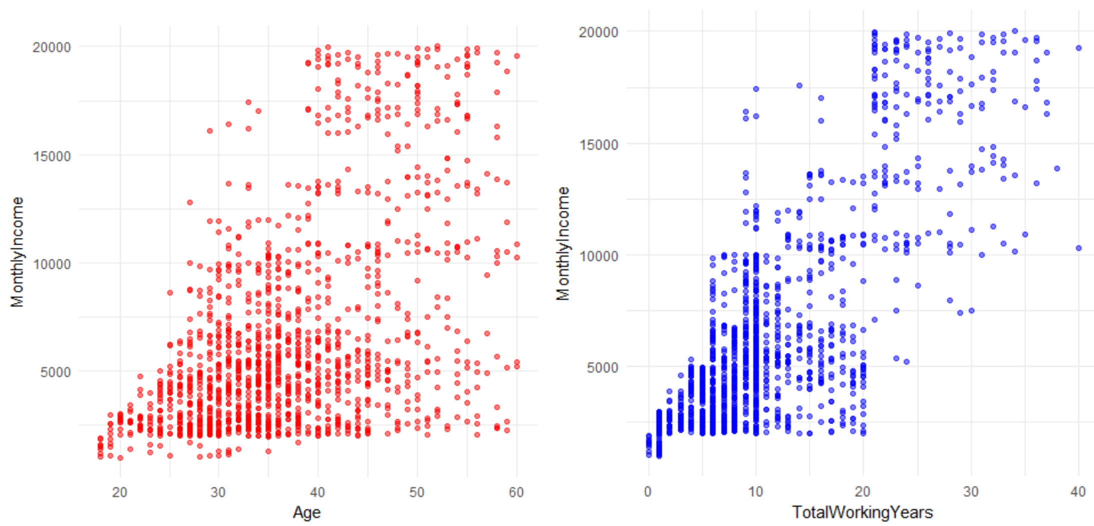

Fig 12: Scatter plot of continuous variables

iii) Linearity of Categorical Variables:

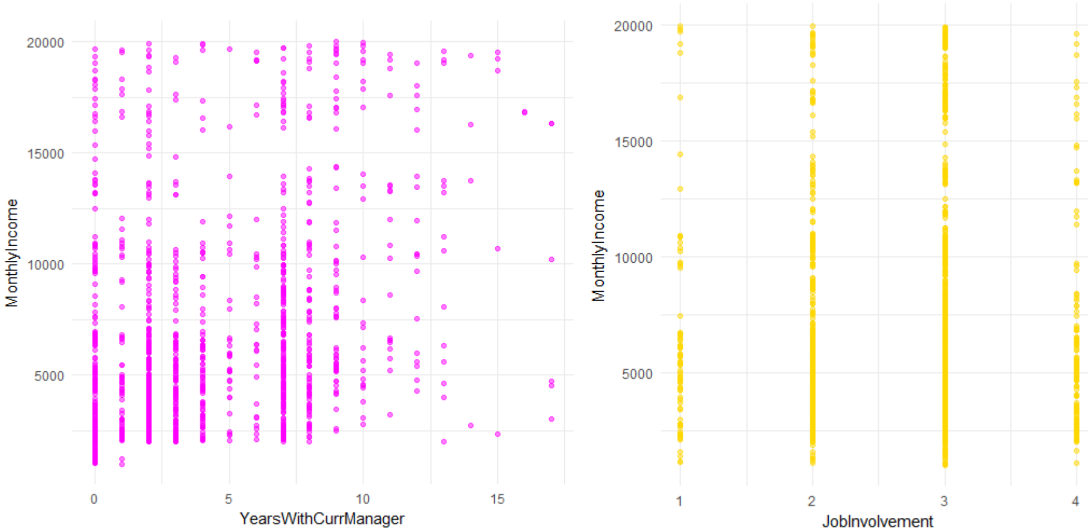

Fig:13 Scatter plot of categorical variables

iv) Independency: Residuals are evenly scattered on both the sides of the axis

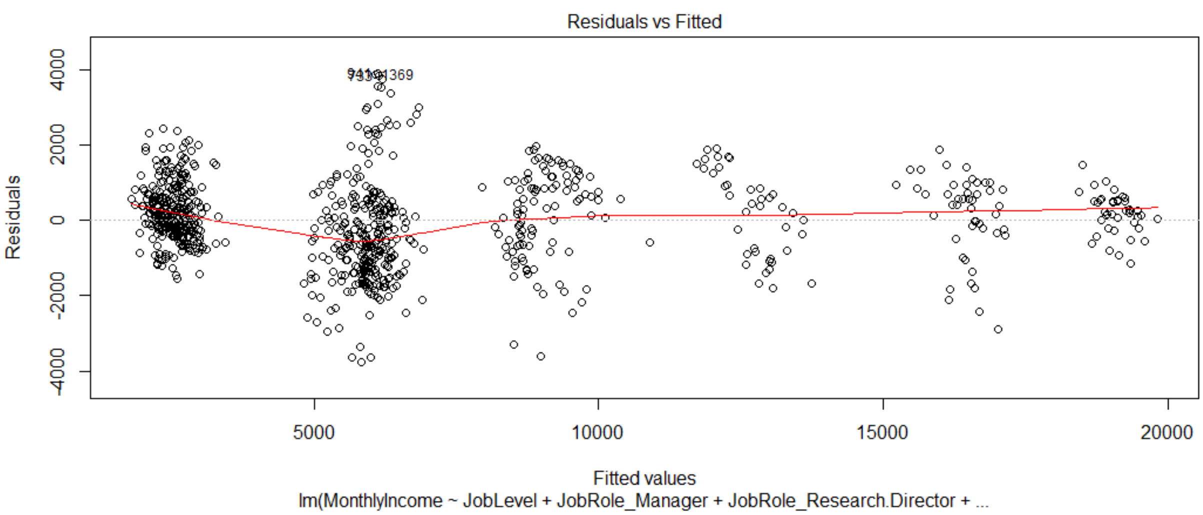

Fig: 14 Plot of fitted values and residuals

k-Nearest Neighbours Method: Prediction

k-nearest neighbors algorithm (k-NN) is a data driven method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression.

In *k-NN regression*, the output is the monthly income for the object. This will predict the monthly income using kNN.

k-NN Function for Regression

```
knn.bestk_reg = function(my_train, my_test, my_ytrain, validation_income, k.max = 20) {

  #each_rmse = rep(NA, k.max)
  each_rmse <- as.data.frame(matrix(0, ncol = 4, nrow = k.max))
  colnames(each_rmse) <- c("K", "RMSE", "MSE", "MAD")
  for (i in 1:k.max){

    knn_reg_obj_val = knn.reg(train = my_train,
                              test = my_test,
                              y = my_ytrain, k = i)

    each_rmse[i,1] <- i
    each_rmse[i,2] <- rmse(validation_income, knn_reg_obj_val$pred)
    each_rmse[i,3] <- mse(validation_income, knn_reg_obj_val$pred)
    each_rmse[i,4] <- mad(validation_income, median(knn_reg_obj_val$pred))

  }
  return(each_rmse)
}
```

To explain this in detail, first of all the package which we implemented was 'FNN'. This package consists of function knn.reg for executing regression. This function will give us RMSE result for each of the corresponding k-values.

The following is the K-Chart which is shown below. There are RMSE, MSE, MAD for the corresponding k-values. Lowest RMSE computed by the program was for k=16. However, as the difference between the RMSE values after k=9 is very less. Hence, we choosed best k=9 which saved lot of iterations, and corresponding RMSE is 1766.928

K chart

	K	RMSE	MSE	MAD
1	1	2407.705	5797042	2813.975
2	2	2034.873	4140709	3058.604
3	3	1926.339	3710781	3256.037
4	4	1861.340	3464588	3495.229
5	5	1852.911	3433279	3548.900
6	6	1839.826	3384961	3555.028
7	7	1798.283	3233821	3756.167
8	8	1795.167	3222625	3752.461
9	**9**	**1766.928**	**3122036**	**3949.564**
10	10	1774.138	3147567	4017.846
11	11	1777.530	3159612	4089.685
12	12	1788.979	3200446	4073.691
13	13	1775.530	3152508	4005.415
14	14	1757.700	3089510	4009.056
15	15	1761.706	3103608	4032.326
16	16	1750.305	3063568	4154.523
17	17	1755.585	3082080	4162.312
18	18	1754.447	3078084	4217.256
19	19	1771.898	3139621	4155.650
20	20	1777.401	3159155	4164.030

Model Evaluation

The k-NN model built using k-NN gives us the best k value of 9 with an RMSE of 1766.928. The value is shown in the K Chart. 9 being an odd number will help serve as the best k for this model.

CART: Regression Tree

The structure of the regression tree below explains the division of factors that influence the monthly income variable to the maximum extent. The variables that play an important role in altering the monthly income of employees are TotalWorkingYears, JobLevel, JobRole_Research Director. It is quite natural to note from the categorical nature of the variables that the employees having an experience of less than 20.5 years draw a lesser pay when compared to those above this segment. The JobLevel is also a natural indication of the 5 level of employees drawing proportional salaries with the level 1 representing the base level of salaries and 5 representing the highest level of salaries. The color of the leaf nodes with respect to intensity of the blue color signifies the magnitude of the salary values. The darker shades of blue represent higher salaries whereas the lighter shades represent the relatively lower values. The percentage values represented in each leaf node of the tree represents the percentage of training data belonging to the model falling in each category of leaf node.

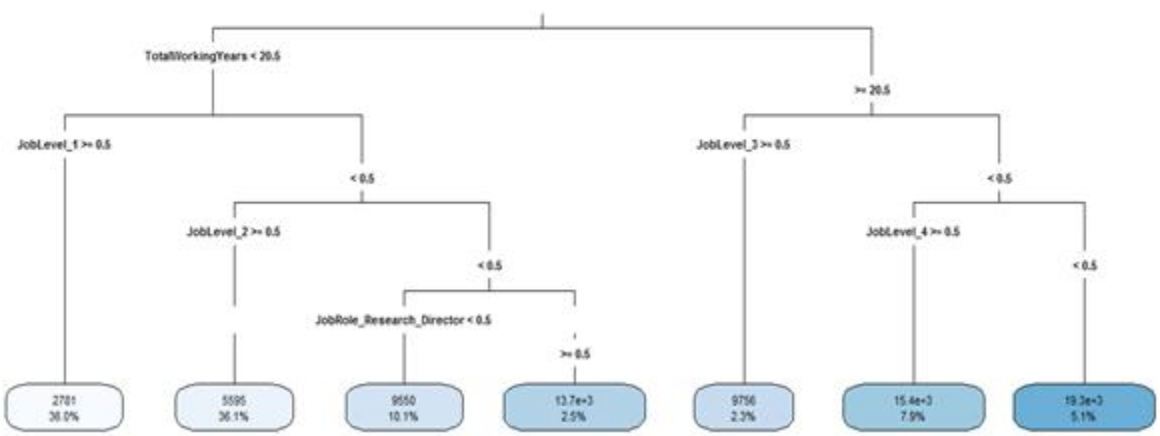

Fig: 15 Regression Tree

Structural Schema of Regression Tree

```
n= 844

node), split, n, deviance, yval
      * denotes terminal node

 1) root 844 20289910000  6751.953
   2) TotalWorkingYears< 20.5 715  6142445000  5106.239
     4) JobLevel_1>=0.5 304   180944200  2781.092 *
     5) JobLevel_1< 0.5 411  3102340000  6826.056
      10) JobLevel_2>=0.5 305   705220000  5595.043 *
      11) JobLevel_2< 0.5 106   605023600 10368.120
        22) JobRole_Research_Director< 0.5 85   245988000  9549.647 *
        23) JobRole_Research_Director>=0.5 21    71615620 13681.000 *
   3) TotalWorkingYears>=20.5 129  1477743000 15873.540
     6) JobLevel_3>=0.5 19    42880780  9755.842 *
     7) JobLevel_3< 0.5 110   600937200 16930.240
      14) JobLevel_4>=0.5 67   208061300 15435.460 *
      15) JobLevel_4< 0.5 43     9919053 19259.300 *
```

Model Building, Evaluation and Accuracy

The Regression tree model was built using the “rpart” function in R programming where all the variables were given and the function by itself picked the best variable parameters to predict the monthly income value. To evaluate the prediction models built by the three methods were decided to compare the Root Mean Square Error (RMSE) values and the value for CART for the prediction of monthly income turned out to be at 1168.528. This means that the model can predict the income of an employee based on the other variables with an accuracy range of plus or minus 1168 USD.

2. Classification Problem: Classification based on Attrition

Logistic Regression

If the question is to predict a binary variable also called classification problem then logistic regression is the preferred choice. Logistic Regression is used to predict the probability that a given example belongs to the “1” class versus the probability that it belongs to the “0” class. In this particular case the variable of interest is Attrition with “1” class associated with leaving and “0” class with staying at the company. The curve is constructed using the natural logarithm of the “odds” of the target variable and function used is sigmoid or logistic function as depicted in the figure below.

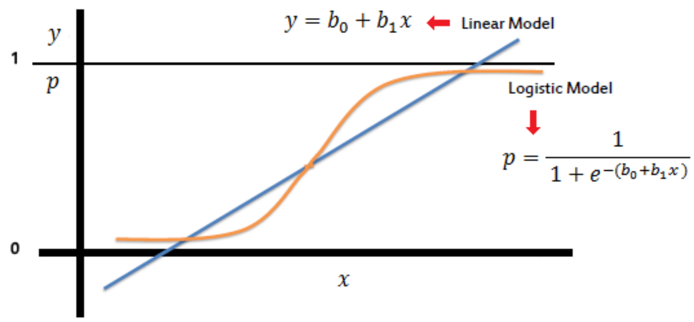

Fig: 16 Logistic Regression Curve (Saedsayad.com)

Variable Selection:

Similar to selection of variables in Multiple Linear Regression (MLR) forward and backward selection methods were utilized. Snapshots of the the models along with significant variables for respective variable selection methods are illustrated below.

Forward Selection:

```
glm(formula = Attrition ~ OverTime + JobLevel + Single + JobSatisfaction +
    JobInvolvement + EnvironmentSatisfaction + JobRole_Sales.Representative +
    DistanceFromHome + WorkLifeBalance + JobRole_Laboratory.Technician +
    YearsInCurrentRole + YearsSinceLastPromotion + TotalWorkingYears +
    NumCompaniesWorked, family = "binomial", data = train)
```

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                    2.18044    0.80972   2.693 0.007085 ** 
OverTime                       2.01881    0.24431   8.263  < 2e-16 ***
JobLevel                      -0.33530    0.20170  -1.662 0.096440 .  
Single                         0.98626    0.22928   4.302 1.70e-05 ***
JobSatisfaction               -0.50195    0.10644  -4.716 2.41e-06 ***
JobInvolvement                -0.57551    0.15021  -3.831 0.000127 ***
EnvironmentSatisfaction       -0.33735    0.10305  -3.274 0.001062 ** 
JobRole_Sales.Representative   1.50297    0.39873   3.769 0.000164 ***
DistanceFromHome               0.03976    0.01359   2.926 0.003428 ** 
WorkLifeBalance               -0.29497    0.15280  -1.930 0.053552 .  
JobRole_Laboratory.Technician  0.64019    0.29668   2.158 0.030941 *  
YearsInCurrentRole            -0.11102    0.04924  -2.255 0.024158 *  
YearsSinceLastPromotion        0.16659    0.04851   3.434 0.000595 ***
TotalWorkingYears             -0.06356    0.02940  -2.162 0.030608 *  
NumCompaniesWorked             0.08825    0.04926   1.791 0.073228 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Backward Selection:

```
Call:
glm(formula = Attrition ~ DistanceFromHome + EnvironmentSatisfaction +
    JobInvolvement + JobSatisfaction + NumCompaniesWorked + OverTime +
    TotalWorkingYears + WorkLifeBalance + YearsInCurrentRole +
    YearsSinceLastPromotion + Single + Married + Department_Sales +
    JobRole_Sales.Representative + JobRole_Research.Scientist +
    JobRole_Laboratory.Technician + JobRole_Human.Resources,
    family = "binomial", data = train)
```

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                    0.77004    0.88405   0.871 0.383735    
DistanceFromHome               0.03891    0.01365   2.851 0.004358 ** 
EnvironmentSatisfaction       -0.32890    0.10387  -3.166 0.001543 ** 
JobInvolvement                -0.58096    0.15240  -3.812 0.000138 ***
JobSatisfaction               -0.53469    0.10889  -4.910 9.09e-07 ***
NumCompaniesWorked             0.09764    0.04982   1.960 0.050032 .  
OverTime                       2.03978    0.24734   8.247  < 2e-16 ***
TotalWorkingYears             -0.07752    0.02415  -3.210 0.001327 ** 
WorkLifeBalance               -0.29481    0.15369  -1.918 0.055079 .  
YearsInCurrentRole            -0.11519    0.04931  -2.336 0.019495 *  
YearsSinceLastPromotion        0.17155    0.04925   3.483 0.000496 ***
Single                         1.30187    0.33605   3.874 0.000107 ***
Married                        0.47733    0.32997   1.447 0.148008    
Department_Sales               0.58643    0.37984   1.544 0.122615    
JobRole_Sales.Representative   1.74411    0.45226   3.856 0.000115 ***
JobRole_Research.Scientist     0.91404    0.40162   2.276 0.022852 *  
JobRole_Laboratory.Technician  1.49028    0.41092   3.627 0.000287 ***
JobRole_Human.Resources        1.35153    0.61062   2.213 0.026872 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Results

Forward Selection

	Predicted 0	Predicted 1
Actual 0	443	22

Actual 1	56	29

Table: 3 Misclassification error forward selection

Misclassification Error: 14.1818%

Backward Selection

	Predicted 0	Predicted 1
Actual 0	439	26
Actual 1	56	29

Table:4 Misclassification error backward selection

Misclassification Error: 14.9090%

Misclassification error is almost similar for both the models and model from forward selection method was chosen.

Interpretation of explanatory variables:

1. Working over time , being single, job roles of sales representative and laboratory technician are associated with higher probability of leaving the company.
2. Employees with more involvement in the job , satisfied with their job and environment and having a proper work life balance stay at the company.

k-Nearest Neighbours for Classification

In k-NN classification method, the output is attrition. An outcome is classified by a majority vote of its neighbors, with the outcome being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). In general odd number of k is assigned to execute the algorithm, so that there is an outcome.

Classification for k-NN

	Predicted

Actual	**0**	**1**
0	**465**	**15**
1	**85**	**68**

Table:5 Classification for kNN

Misclassification Error
= (Errors)/(Total Records in Validation Set)

= (85+15)/(533)= 0.150632

Results against Validation Data

Misclassification Rate - 0.150632
Accuracy Rate - 84.93%
Sensitivity - 0.3411765
Specificity - 0.9526882

CART: Classification Tree

The structure of the classification tree in the figure 17. below gives a view of how the factors influencing the employee's decision to leave the company correlates with the attrition of the employees. In this classification model, we see the variables such as TotalWorkingYears, Overtime, Single, NumCompaniesWorked, EmployeeNum and WorkLifeBalance are the factors that greatly influence the attrition rate. In this case again taking a closer look at the leaf node indicates that more the number of leaf nodes the better where can use many variables in classifying the data and obtaining better accuracy measures. This model was very good for the fact that engaged a total of 14 variables in building the tree model.

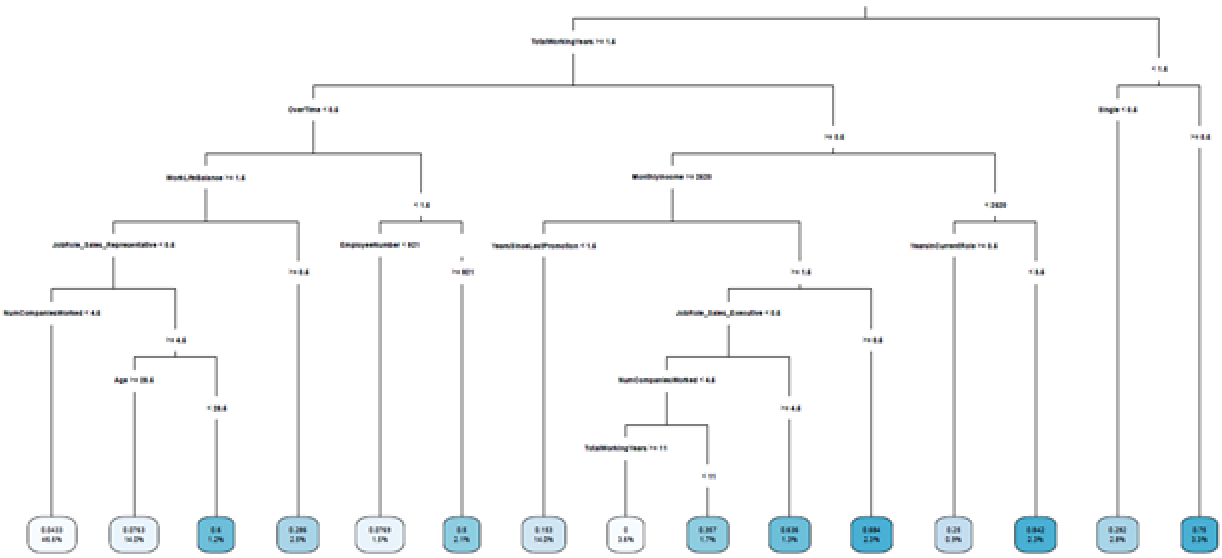

Fig:17 Classification Tree

Structural Schema of Classification Tree

```
node), split, n, deviance, yval
      * denotes terminal node

  1) root 844 114.7618000 0.16232230
    2) TotalWorkingYears>=1.5 792  93.9987400 0.13762630
      4) OverTime< 0.5 573  43.9790600 0.08376963
        8) WorkLifeBalance>=1.5 542  35.3357900 0.07011070
         16) JobRole_Sales_Representative< 0.5 521  30.0345500 0.06142035

           32) NumCompaniesWorked< 4.5 393  16.2646300 0.04325700 *
           33) NumCompaniesWorked>=4.5 128  13.2421900 0.11718750
             66) Age>=28.5 118   8.3135590 0.07627119 *
             67) Age< 28.5 10   2.4000000 0.60000000 *
         17) JobRole_Sales_Representative>=0.5 21   4.2857140 0.28571430 *
        9) WorkLifeBalance< 1.5 31   6.7741940 0.32258060
         18) EmployeeNumber< 921 13   0.9230769 0.07692308 *
         19) EmployeeNumber>=921 18   4.5000000 0.50000000 *
      5) OverTime>=0.5 219  44.0091300 0.27853880
       10) MonthlyIncome>=2520 192  33.3697900 0.22395830
         20) YearsSinceLastPromotion< 1.5 118  15.2542400 0.15254240 *
         21) YearsSinceLastPromotion>=1.5 74  16.5540500 0.33783780
           42) JobRole_Sales_Executive< 0.5 55   9.3818180 0.21818180
             84) NumCompaniesWorked< 4.5 44   4.4318180 0.11363640
              168) TotalWorkingYears>=11 30   0.0000000 0.00000000 *
              169) TotalWorkingYears< 11 14   3.2142860 0.35714290 *
             85) NumCompaniesWorked>=4.5 11   2.5454550 0.63636360 *
           43) JobRole_Sales_Executive>=0.5 19   4.1052630 0.68421050 *
       11) MonthlyIncome< 2520 27   6.0000000 0.66666670
         22) YearsInCurrentRole>=3.5 8   1.5000000 0.25000000 *
         23) YearsInCurrentRole< 3.5 19   2.5263160 0.84210530 *
    3) TotalWorkingYears< 1.5 52  12.9230800 0.53846150
      6) Single< 0.5 24   4.9583330 0.29166670 *
      7) Single>=0.5 28   5.2500000 0.75000000 *
```

Model Building, Evaluation and Accuracy

The classification tree was built using the rpart function and the visualization plot of the tree was made using the rpart.plot function. This tree also picked the variables by itself after assessing the correlation factors and came up with the model involving 14 variables in it. The classification models are evaluated in this project using the confusion matrix and Misclassification Error rate.

Confusion Matrix

	0	1
0	409	33
1	62	27

Table:6 Confusion Matrix

Misclassification Error Rate

= (Incorrect Predictions/Total Data in Validation Data Set) = (33+62/531) = 17.89%

Conclusion

Model Comparison

	Prediction (Monthly Income) RMSE	**Classification (Attrition) Misclassification Error Rate**
MLR & LR	**1111.584**	**0.1418182**
K-NN	**1790.405**	**0.150632**
CART	**1168.528**	**0.1789077**

Table:7 Model Comparison

For prediction of monthly income Multiple Linear Regression, k-NN and Regression trees were evaluated. Similarly, for classification of attrition Logistic Regression, k-NN and classification trees analysis was performed.RMSE values and Misclassification Error Rates of the respective models were compared. RMSE values of Multiple Linear Regression **1111.584** and Misclassification Error Rates **14.18182** of Logistic Regression were lowest. Thus, they were considered the best models for prediction of monthly income and classification of attrition in present scenario.

Best Model Interpretation

Prediction of Monthly Income: Higher job level such as job role of manager or research director is associated with higher monthly income.Job role of sales representative and laboratory technician indicate lower monthly salaries.

Classification of Attrition: Working over time , being single, job roles of sales representative and laboratory technician are associated with higher probability of leaving the company.Employees with more involvement in the job , satisfied with their job and environment and having a proper work life balance stay at the company.

Business Decisions

- ❖ The RMSE indicates that the prediction of Monthly Income of the employee can be done with an approximation of plus or minus 1111.584 USD.

- ❖ The probability of an employee leaving the firm can be predicted with an accuracy of 14.18%.

Learnings from this project:

It was difficult to perform the project in R, and specifically we had some issues in choosing the package for the kNN method. After finding FNN package also it was hard to figure out the function used to execute the kNN regression method.
Lastly we came to know many different things about the HR company datasets. We also found various dependencies used to find whether a person will stay in the company (0) or leave the company (1).
Every results we got were compared with the XLminer output and we observed that there wasn't much difference in the results. If we had more time we could have learned more R functions and packages and used them in our project to get better results. Thus, at the end we can say that R was a bit challenging, while XLMiner was comparatively easy.

References

(n.d). Retrieved from https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/

Scott, L. (n.d.). An introduction to using regression analysis with spatial data. *ESRI*. Retrieved from http://www.esri.com/news/arcuser/0309/files/why.pdf

(n.d.). Retrieved from http://www.saedsayad.com/logistic_regression.htm

(n.d.). Retrieved from https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-1