Final Project

Matthew Baker, Don Padmaperuma, Subhalaxmi Rout, Erinda Budo

2020-12-14

Abstract

HR Analytics finds out the people-related trends in the data and helps the HR Department take the appropriate steps to keep the organization running smoothly and profitably. Attrition is a corporate setup is one of the complex challenges that the people managers and the HRs personnel have to deal with it.

In this research assignment, we investigated data on employee attrition of a company. This is a fictional data set created by IBM data scientists.

We have collected this dataset from Kaggle, using the below link: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

Methodology

We obtained the data set from Kaggle.com using this link: [https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset]. The fictional data set was originally created by IBM data scientists to uncover the facts that lead to employee attrition and explore important question like what are the important factors influence attrition among employees. Also the original dataset can be accessed from the Link- https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/. Then it was saved in our group github repository as a .csv file for the convenience of the analysis purposes. The Attrition dataset had 1470 observations with 35 variables. Out of those 35 there exists the targen variable Attrition with possible outcomes "Yes" and "No". With our experiment results We will do the analysis based on Gender, Education, Income, Working Environment, and lastly, build a predictive model to determine whether an employee is going to quit or not.

Experimantation and Results

Data Preparation

Checking for missing values and removing non value attributes

```
##                     Age               Attrition          BusinessTravel
##                       0                       0                       0
##               DailyRate              Department        DistanceFromHome
##                       0                       0                       0
##               Education           EducationField            EmployeeCount
##                       0                       0                       0
##          EmployeeNumber  EnvironmentSatisfaction                  Gender
##                       0                       0                       0
##              HourlyRate          JobInvolvement                JobLevel
##                       0                       0                       0
```

```
##                   JobRole           JobSatisfaction            MaritalStatus
##                         0                         0                        0
##            MonthlyIncome               MonthlyRate       NumCompaniesWorked
##                         0                         0                        0
##                   Over18                  OverTime        PercentSalaryHike
##                         0                         0                        0
##        PerformanceRating RelationshipSatisfaction            StandardHours
##                         0                         0                        0
##         StockOptionLevel         TotalWorkingYears    TrainingTimesLastYear
##                         0                         0                        0
##          WorkLifeBalance             YearsAtCompany       YearsInCurrentRole
##                         0                         0                        0
##  YearsSinceLastPromotion      YearsWithCurrManager
##                         0                         0
```

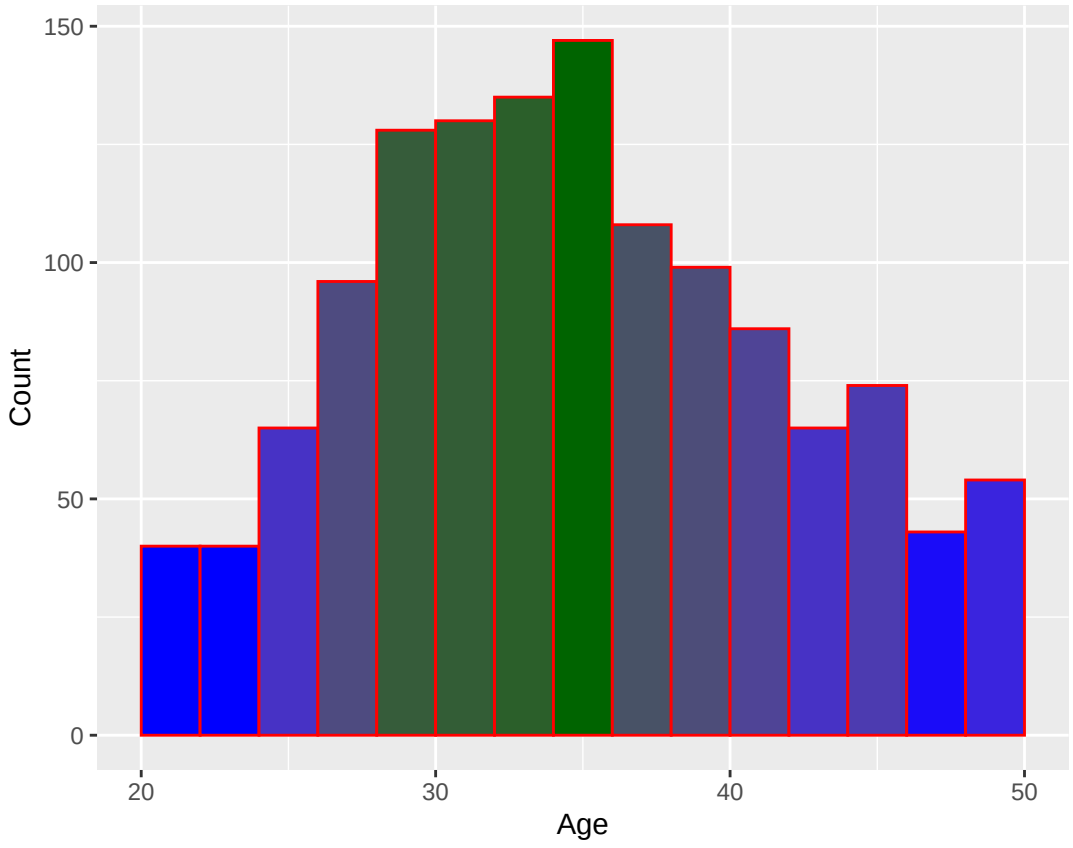
```
## Data Set has  1470  Rows and  31  Columns
```

Fortunately no missing data or duplicate data.

Also, some of the attributes that are categorical are represented as integers in the dataset. We need to change them to categorical.

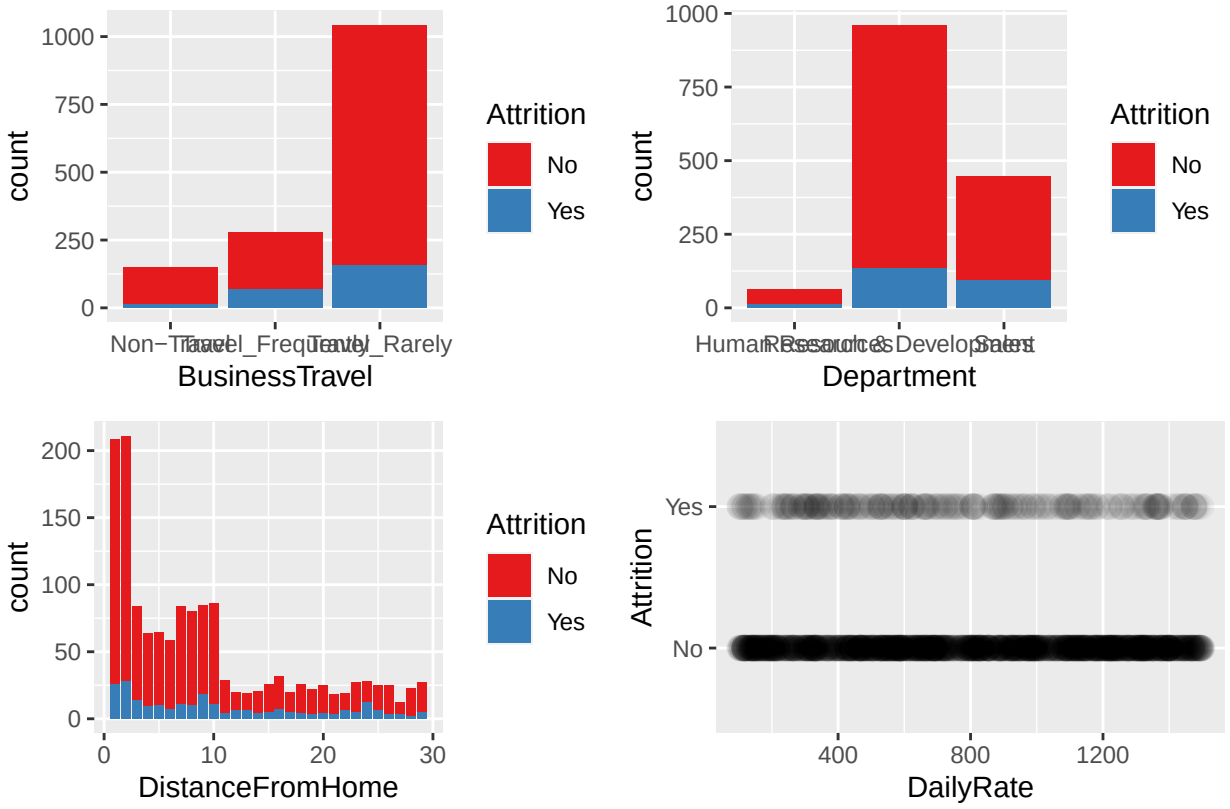
Visualization

In this section, we can visualize the infuance of each variable on Attrition of the organization.



Age plot and Fig 1

Fig 1



1. Age: We see that majority of employees leaving the org are around 30 Years (year 28-36). Average age is between 30 to 40.
2. Business Travel: Among people who leave, most travel frequently or rarely.
3. Department: Among people attrited employees from HR dept. are less.It is because of low proportion of HR in the organization(Fig 1).
4. Distance From Home: Contrary to normal assumptions, a mojority of employees who have left the organization are near to the Office.
5. Daily Rate: We are not able to see any distinguishable feature here(Fig 1).

Fig 2

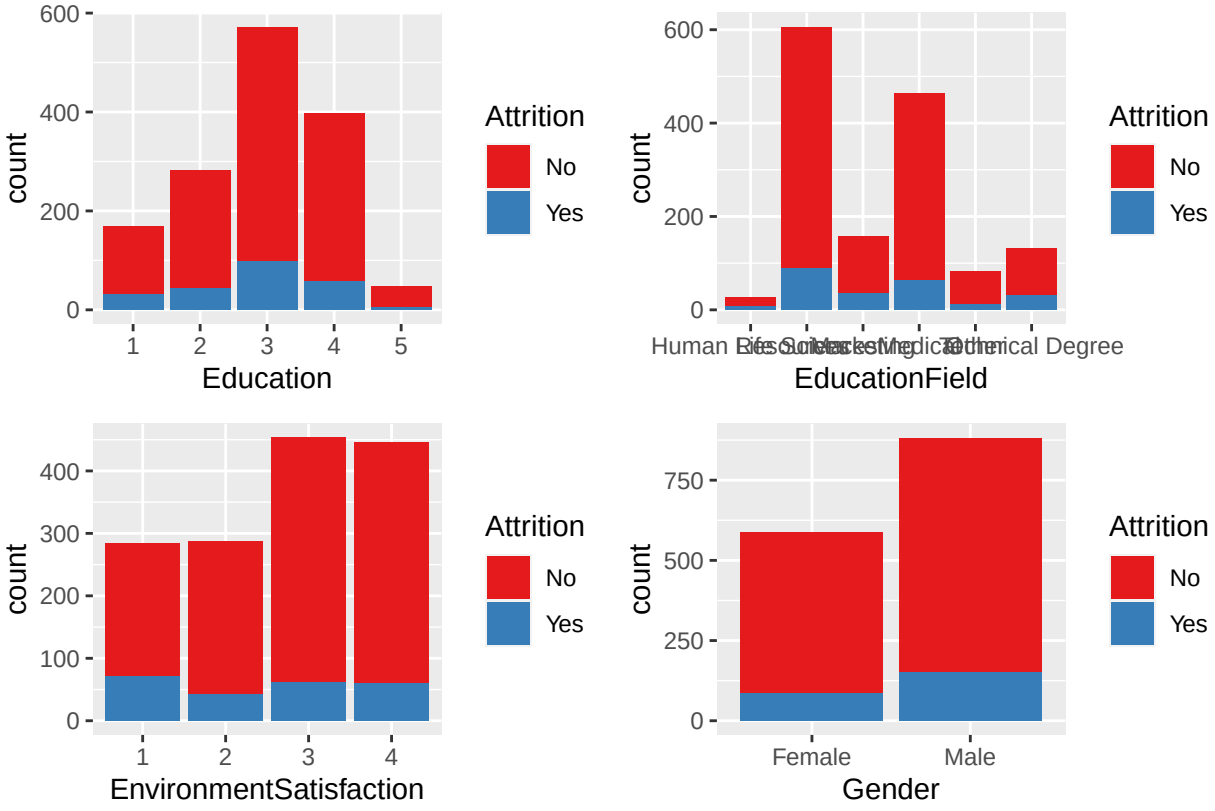


Fig 2

6. Education: From the data we know that, 1-'Below College', 2-'College', 3-'Bachelor', 4-'Master' 5 'Doctor'. Looking at the plot we see that very few Doctors attrite. May be because of less number. Based on the data most of the employees have Bachelors degree level education.
7. Education Field: On lines of the trend in Departments, a minority of HR educated employees leave and it is majorly because of low proportion of the HR in the organization.
8. Employee Count : It is an insignificant variable for us.
9. Employee Number: It is also an insignificant variable for us.
10. Environment Satisfaction: Ratings stand for: 1-'Low', 2-'Medium', 3-'High', 4-'Very High'. We don't see any distinguishable feature(Fig 2).

Fig 3

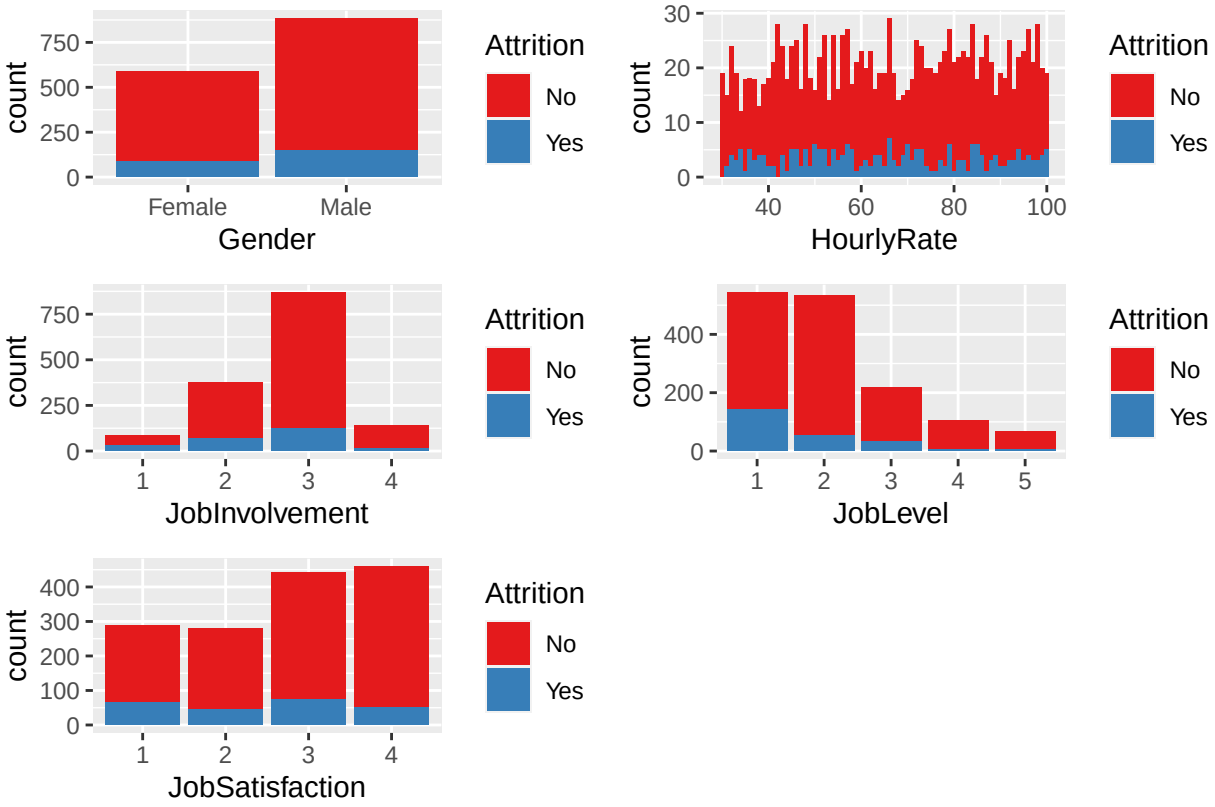


Fig 3

11. Gender: Majority of separated employees are Male and the reason might be because around 61% of employees in the dataset are Male.
12. HourlyRate : There seems to be no straightforward relation with the Daily Rate of the employees.
13. Job Involvement: Ratings stand for 1-'Low', 2-'Medium', 3-'High', 4-'Very High'. Majority of employees who leave are either very highly involved or least involved in their Jobs.
14. JobLevel: Job Level increases the number of people quitting decreases.
15. Job Satisfaction: As per data 1-'Low', 2-'Medium', 3-'High', 4-'Very High'. We see higher attrition levels among lower Job Satisfaction levels.

Fig 4

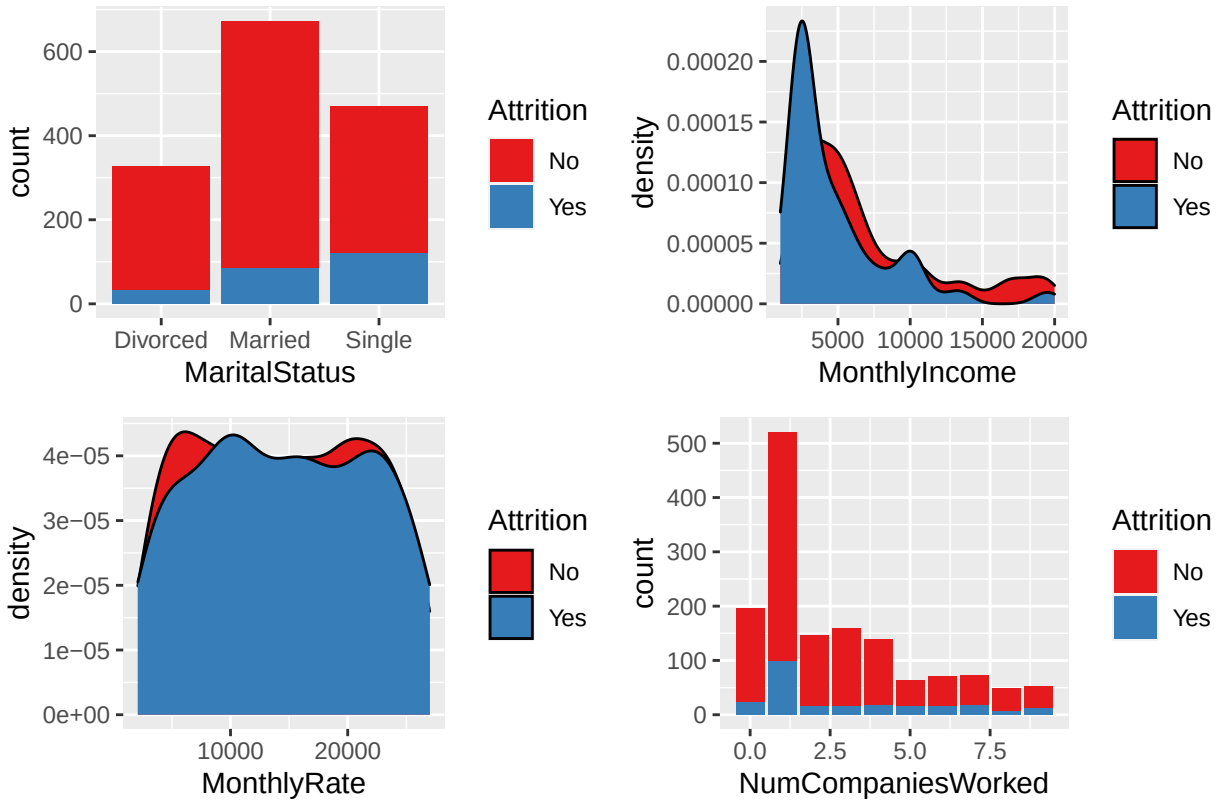


Fig 4

16. Marital Status:Attrition is on higher side for Single and lowest for Divorced employees. Most employees are married.
17. Monthly Income: We see higher levels of attrition among the lower segment of monthly income. If looked at in isolation, might be due to dissatisfaction of income.Higher number of employees earn less.
18. Monthly Rate: We don't see any inferable trend from this. Also no straightforwad relation with Monthly Income.
19. Number of Companies Worked: We see a clear indication that many people who have worked only in One company before quit a lot.

Fig 5

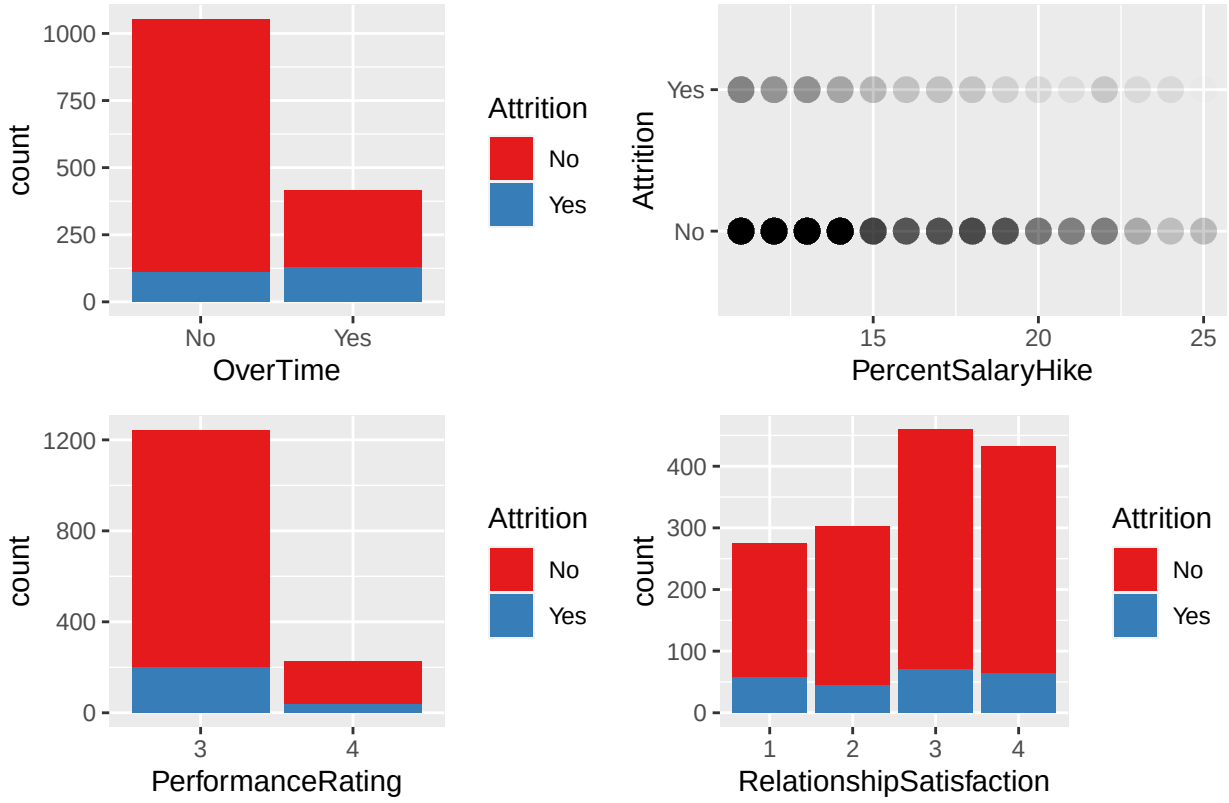


Fig 5

20. Over Time: Larger Proportion of Overtime Employees are quitting. 21. Percent Salary Hike: We see that people with less than 15% hike have more chances to leave. 22. Performance Rating: 1-'Low', 2-'Good', 3-'Excellent', 4-'Outstanding'. We see that we have employees of only 3 and 4 ratings. Lesser proportion of 4 raters quit. 23. Relationship Satisfaction: 1-'Low', 2-'Medium', 3-'High', 4-'Very High'. Higher number of people with 3 or more rating are quitiing. There are considerable amount of low and medium relationship satisfaction in this organization.

Fig 6

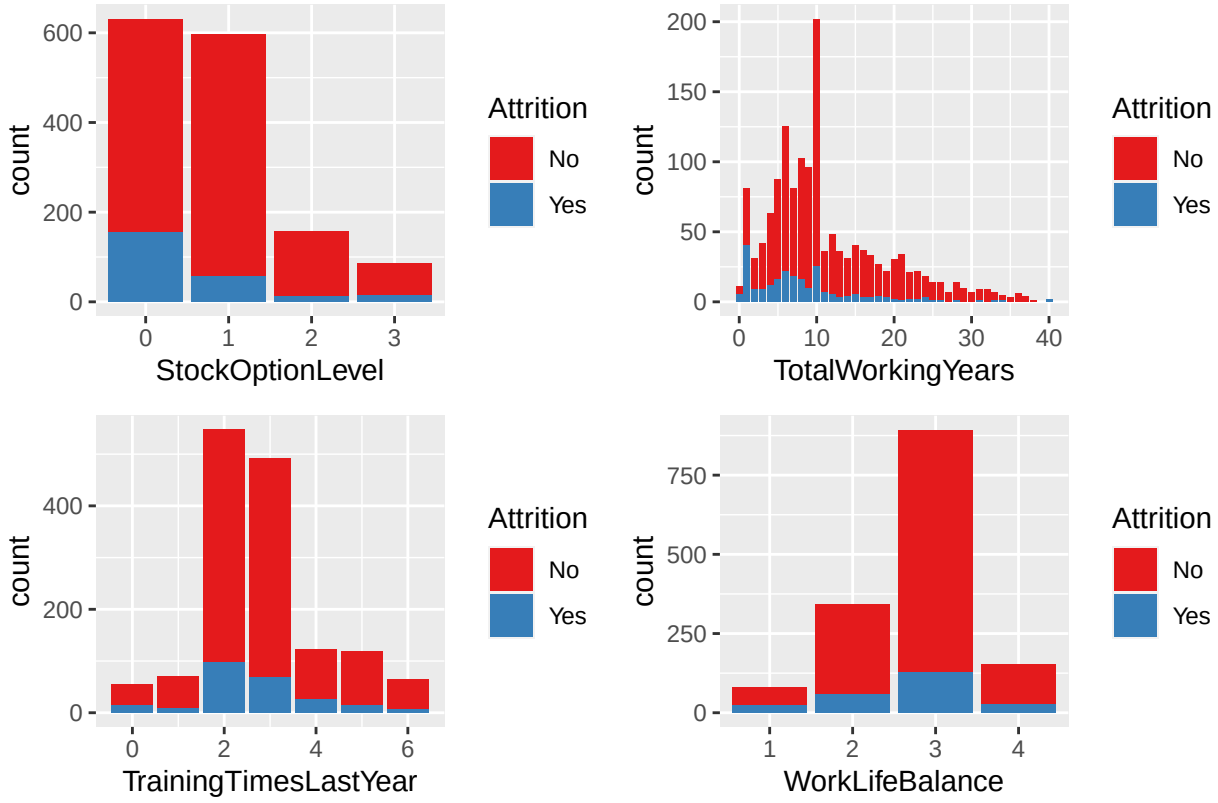


Fig 6

24. Stock Option Level: Larger proportions of levels 1 & 2 tend to quit more. 25. Total Working Years: We see larger proportions of people with 1 year of experiences quitting the organization also in bracket of 1-10 Years. Higher the number of experience you have, you tend to stay in the job. 26. Traning Times Last Year: This indicates the no of training interventions the employee has attended. People who have been trained 2-4 times is an area of concern. 27. Work Life Balance:Ratings as per Metadata is 1 'Bad' 2 'Good' 3 'Better' 4 'Best'. As expected larger proportion of 1 rating quit, but absolute number wise 3 is on higher side.

Fig 7

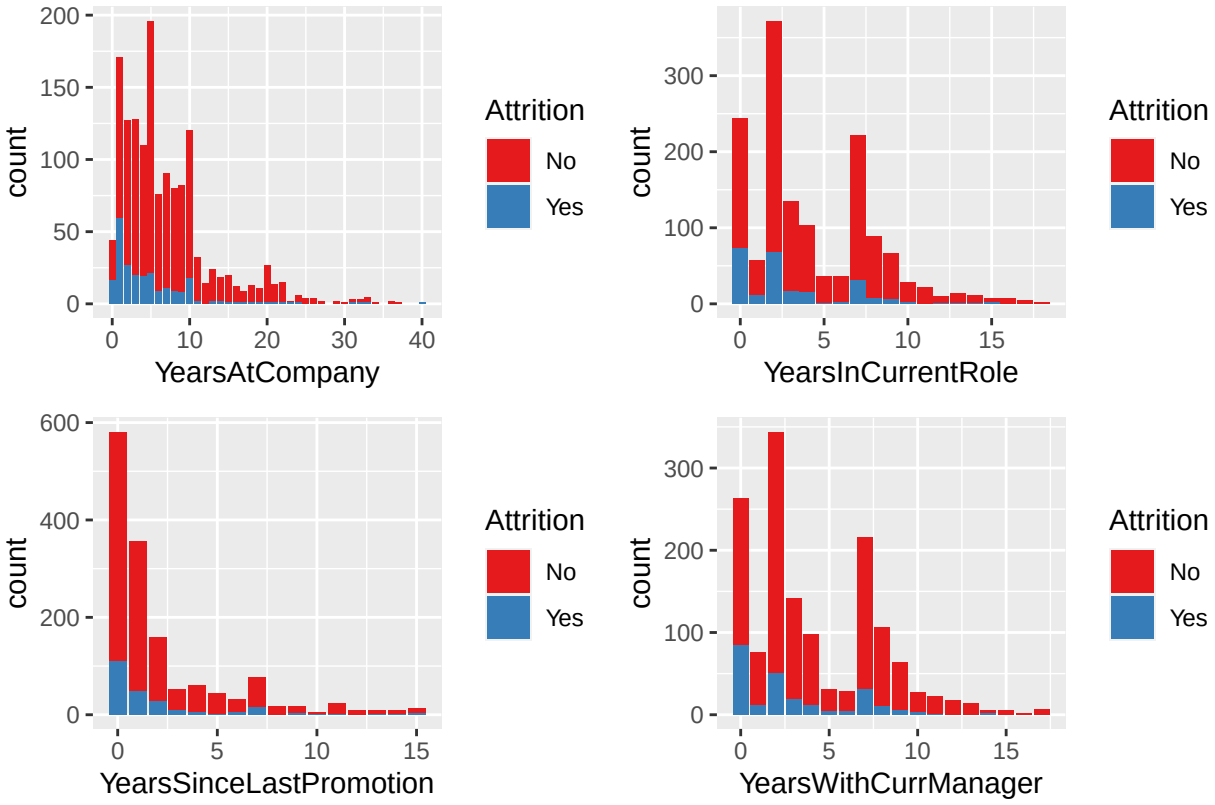


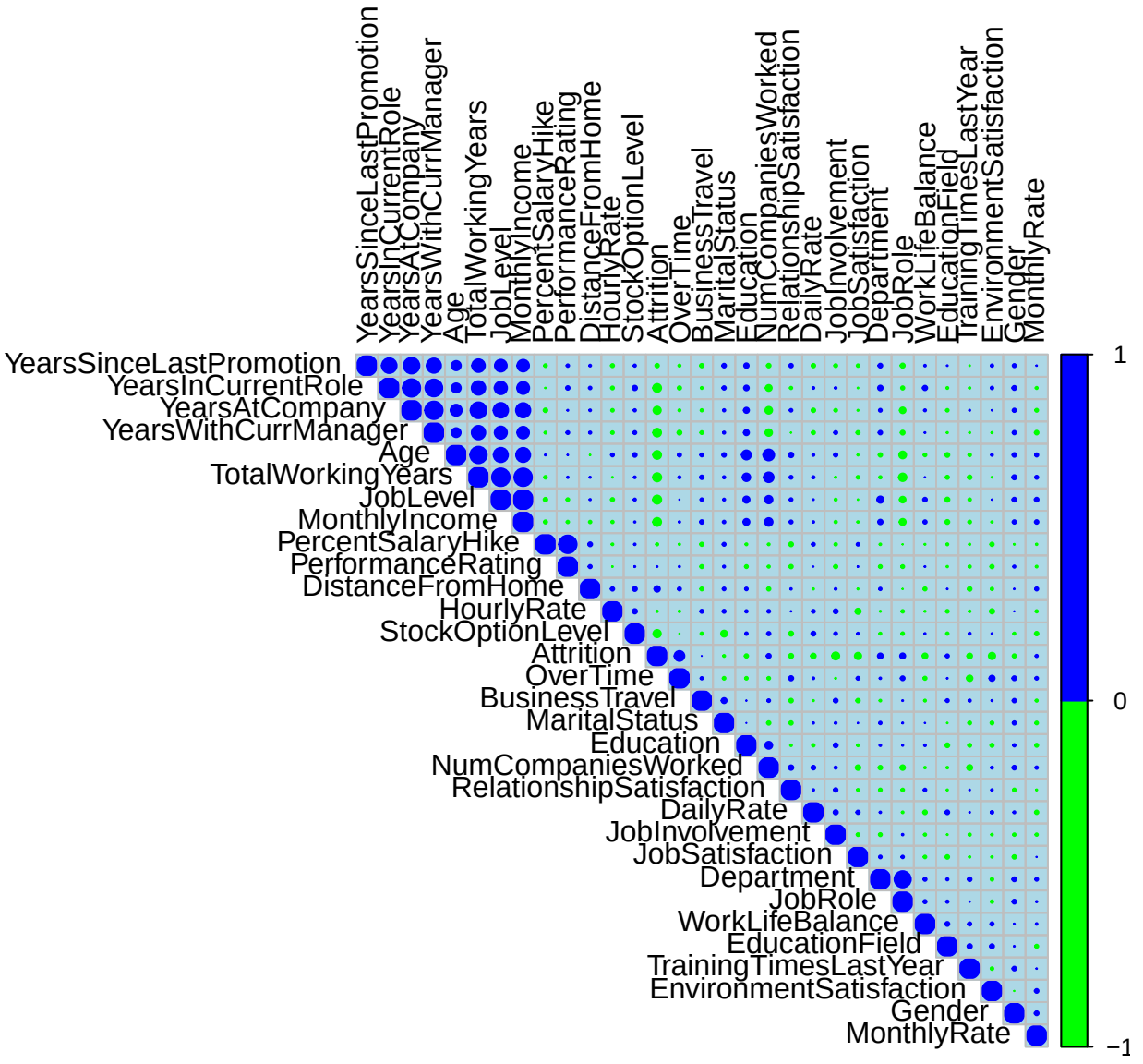
Fig 7

28. Years at Company: Larger proportion of new comers are quitting the organization. Which sidelines the recruitment efforts of the organization.
29. Years In Current Role: Plot shows a larger proportion with just 0 years quitting. May be a role change is a trigger for Quitting.
30. Years Since Last Promotion: Larger proportion of people who have been promoted recently have quit the organization.
31. Years With Current Manager: As expected a new Manager is a big cause for quitting.

Correlation Below plot shows correlated variables, such as with Attrition overtime is positively correlated however MonthlyIncome negatively co-related.

To get all numerical data we will apply below changes on some attributes.

- BUSINESS TRAVEL (1=Non Travel, 2=Travel Frequently, 3=Tavel Rarely)
- DEPARTMENT (1=Human Resources, 2=Research & Development, 3=Sales)
- EDUCATION FIELD (1=Human Resources, 2=LIFE SCIENCES, 3=MARKETING, 4=MEDICAL SCIENCES, 5=OTHERS, 6= TEHCNICAL DEGREE)
- GENDER (2=FEMALE, 1=MALE)
- JOB ROLE (1=Healthcare Representative, 2=Human Resources, 3=Laboratory Technician, 4=MANAGER, 5= Manufacturing Director, 6= REASEARCH DIRECTOR, 7= RESEARCH SCIENTIST, 8=SALES EXECUTIEVE, 9= SALES REPRESENTATIVE)
- MARITAL STATUS (1=DIVORCED, 2=SINGLE, 3=MARRIED)
- OVERTIME (1=NO, 2=YES)



Data Preparation

Lets split the data in to 2 parts i.e train and test. Train contains 75% of data and test contains 25% of data.

Train data has 1103 rows and 31 coumns. Test data has 367 rows and 31 columns.

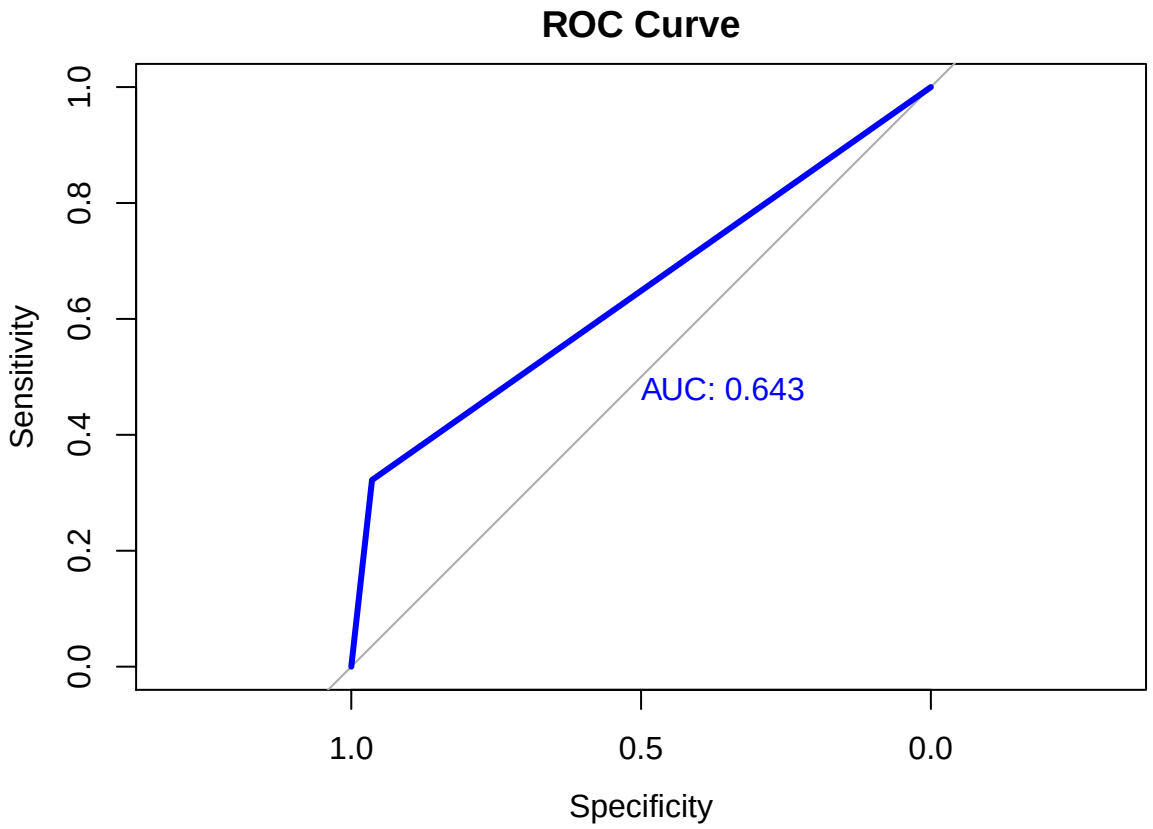
Model Building

Model1 - Logistic Regression with all features We will apply logistic regression model with all variables.

```
##
## Call:
## glm(formula = Attrition ~ ., family = binomial(link = "logit"),
##     data = hr_train)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6585  -0.5170  -0.2989  -0.1321   3.3787
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               4.395e+00  1.305e+00   3.367 0.000760 ***
## Age                      -2.264e-02  1.489e-02  -1.521 0.128256
## BusinessTravel           -6.818e-02  1.537e-01  -0.444 0.657362
## DailyRate                -3.012e-04  2.379e-04  -1.266 0.205565
## Department                8.920e-01  2.918e-01   3.057 0.002238 **
## DistanceFromHome          3.966e-02  1.177e-02   3.370 0.000753 ***
## Education                 1.373e-02  9.590e-02   0.143 0.886148
## EducationField            2.128e-02  7.720e-02   0.276 0.782778
## EnvironmentSatisfaction  -3.777e-01  9.111e-02  -4.145 3.40e-05 ***
## Gender                   -2.565e-01  2.013e-01  -1.274 0.202691
## HourlyRate                7.443e-04  4.895e-03   0.152 0.879147
## JobInvolvement           -5.386e-01  1.375e-01  -3.918 8.93e-05 ***
## JobLevel                 -4.571e-01  3.216e-01  -1.421 0.155217
## JobRole                  -4.418e-02  5.802e-02  -0.761 0.446391
## JobSatisfaction          -3.977e-01  9.000e-02  -4.419 9.90e-06 ***
## MaritalStatus             5.600e-02  1.323e-01   0.423 0.672140
## MonthlyIncome            -1.109e-05  7.622e-05  -0.146 0.884312
## MonthlyRate               7.638e-06  1.375e-05   0.555 0.578572
## NumCompaniesWorked        1.666e-01  4.166e-02   3.998 6.39e-05 ***
## OverTime                  2.008e+00  2.112e-01   9.509  < 2e-16 ***
## PercentSalaryHike        -2.365e-02  4.243e-02  -0.557 0.577224
## PerformanceRating        -1.268e-01  4.395e-01  -0.289 0.772909
## RelationshipSatisfaction -2.494e-01  9.215e-02  -2.707 0.006797 **
## StockOptionLevel         -6.377e-01  1.322e-01  -4.822 1.42e-06 ***
## TotalWorkingYears        -5.875e-02  3.124e-02  -1.881 0.060001 .
## TrainingTimesLastYear    -1.058e-01  8.094e-02  -1.307 0.191207
## WorkLifeBalance          -3.109e-01  1.321e-01  -2.354 0.018589 *
## YearsAtCompany            8.601e-02  4.176e-02   2.060 0.039418 *
## YearsInCurrentRole       -1.296e-01  4.949e-02  -2.619 0.008832 **
## YearsSinceLastPromotion   1.331e-01  4.557e-02   2.922 0.003480 **
## YearsWithCurrManager     -9.847e-02  4.952e-02  -1.989 0.046743 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 974.94  on 1102  degrees of freedom
## Residual deviance: 696.61  on 1072  degrees of freedom
## AIC: 758.61
##
## Number of Fisher Scoring iterations: 6
```

	Values
Accuracy	0.8610354
precision	0.8813056
F1-sensitivity	0.9642857
specificity	0.3220339
f1_score	0.9209302
AUC	0.6431598

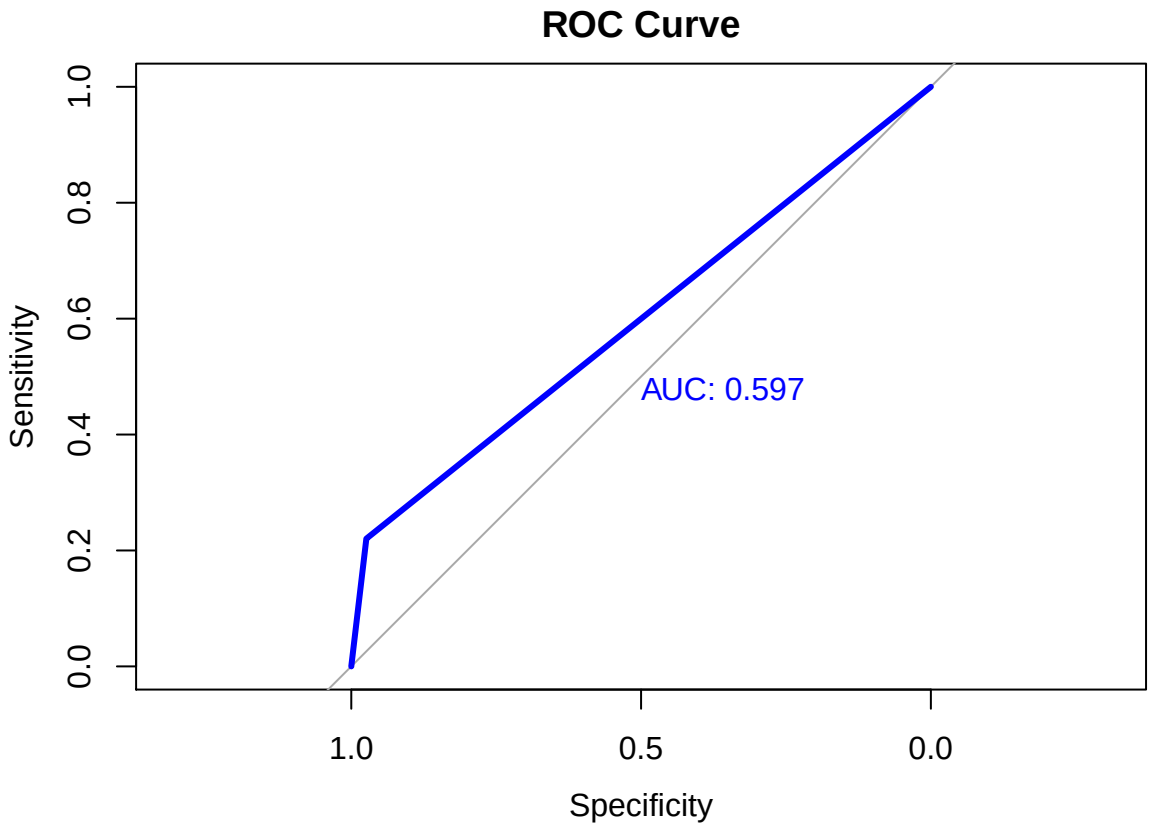


Model2 - Logistic Regression with significant features There are some insignificant variable present in model1 so this model we will remove those variables.

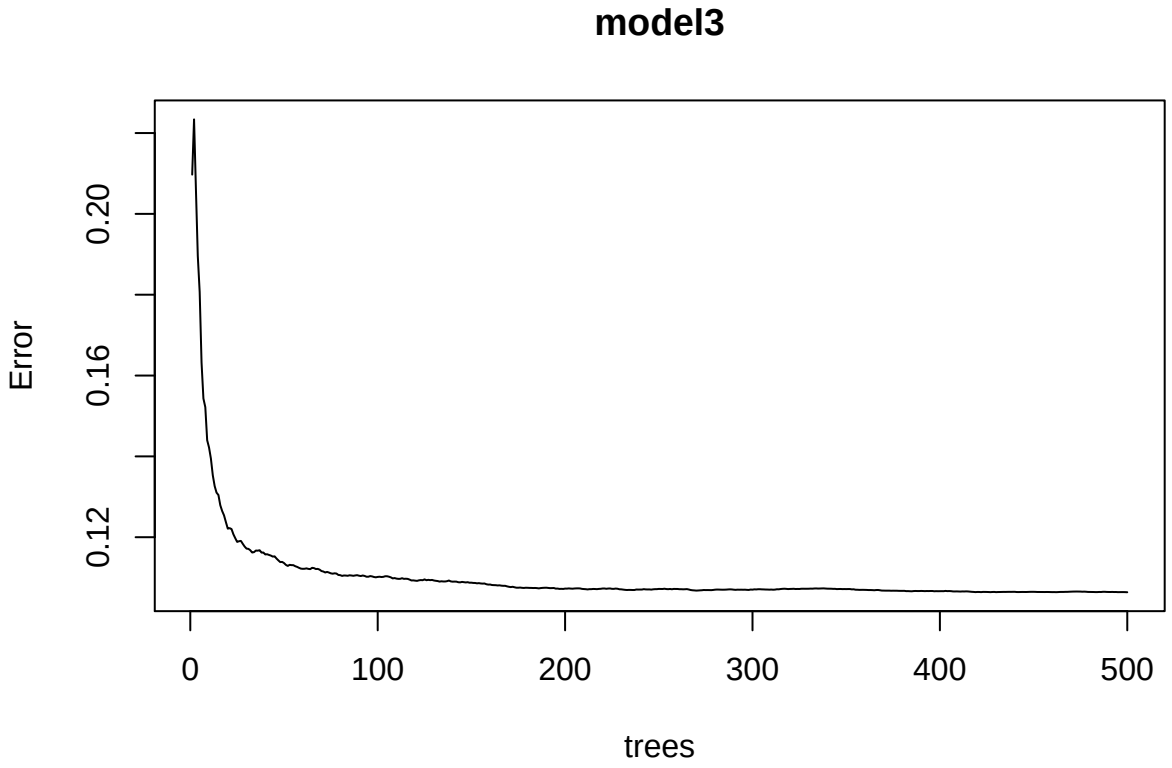
```
##
## Call:
## glm(formula = Attrition ~ . - BusinessTravel - Department - Education -
##     EducationField - Gender - HourlyRate - JobLevel - JobRole -
##     MaritalStatus - MonthlyIncome - MonthlyRate - PercentSalaryHike -
##     PerformanceRating - TotalWorkingYears - TrainingTimesLastYear -
##     YearsAtCompany, family = binomial(link = "logit"), data = hr_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7804  -0.5382  -0.3327  -0.1810   3.3091
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               5.3019394  0.8125469   6.525 6.80e-11 ***
```

```
## Age                       -0.0567969  0.0118972  -4.774 1.81e-06 ***
## DailyRate                 -0.0002773  0.0002311  -1.200 0.230158
## DistanceFromHome           0.0334165  0.0111773   2.990 0.002793 **
## EnvironmentSatisfaction   -0.3396858  0.0865671  -3.924 8.71e-05 ***
## JobInvolvement            -0.5414986  0.1325409  -4.086 4.40e-05 ***
## JobSatisfaction           -0.3469010  0.0855849  -4.053 5.05e-05 ***
## NumCompaniesWorked         0.1214537  0.0389367   3.119 0.001813 **
## OverTime                   1.8620636  0.1982625   9.392  < 2e-16 ***
## RelationshipSatisfaction -0.2302391  0.0886106  -2.598 0.009368 **
## StockOptionLevel          -0.6245823  0.1290149  -4.841 1.29e-06 ***
## WorkLifeBalance           -0.3009344  0.1267686  -2.374 0.017602 *
## YearsInCurrentRole        -0.1066898  0.0433740  -2.460 0.013903 *
## YearsSinceLastPromotion    0.1328820  0.0380678   3.491 0.000482 ***
## YearsWithCurrManager      -0.0825263  0.0431574  -1.912 0.055848 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 974.94  on 1102  degrees of freedom
## Residual deviance: 736.34  on 1088  degrees of freedom
## AIC: 766.34
##
## Number of Fisher Scoring iterations: 6
```

	Values
Accuracy	0.8528610
precision	0.8670520
F1-sensitivity	0.9740260
specificity	0.2203390
f1_score	0.9174312
AUC	0.5971825



Model3 - Random Forest Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

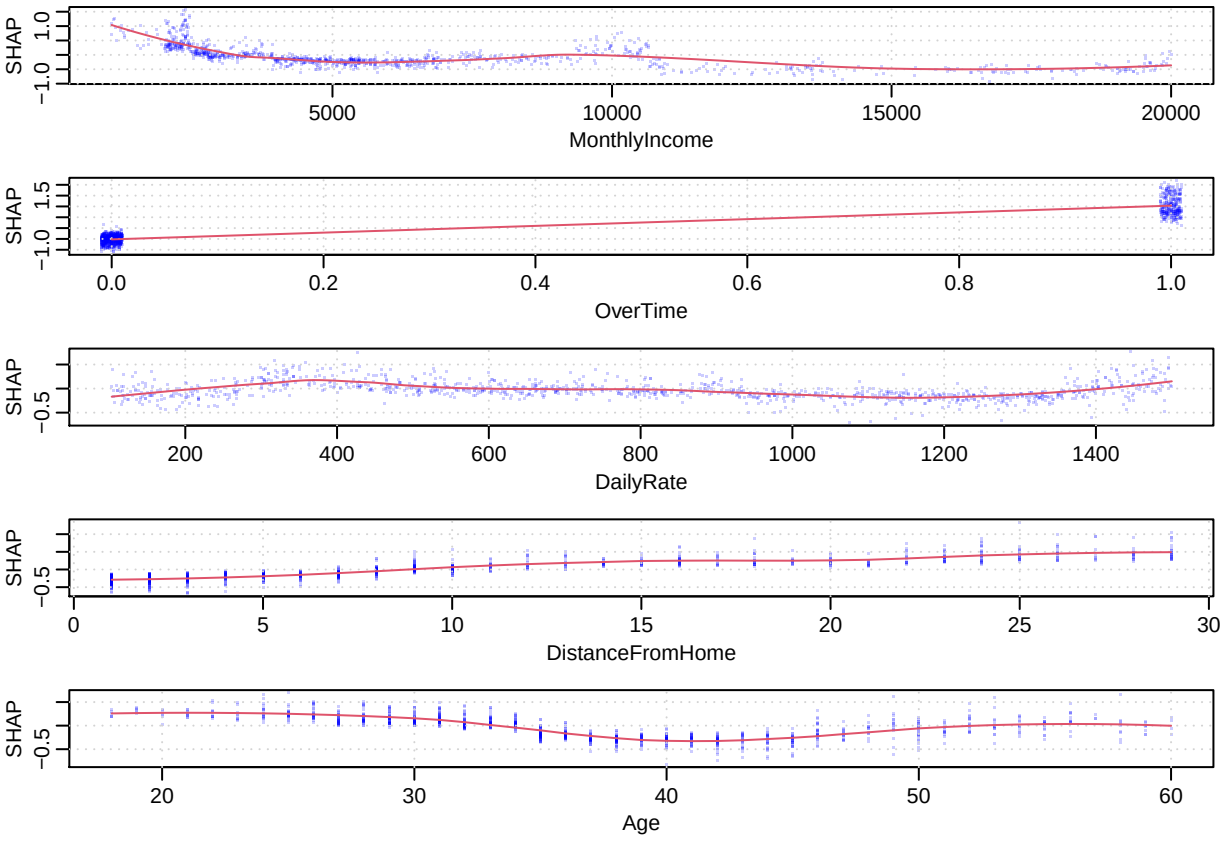


	Values
Accuracy	0.8474114
precision	0.8579545
F1-sensitivity	0.9805195
specificity	0.1525424
f1_score	0.9151515
AUC	0.5665309

Model4 - XGB Classification or regression technique that generates decision trees sequentially, where each tree focuses on correcting the previous tree model. The final output is a combination of the results from all trees.

```
## [02:50:44] WARNING: amalgamation/../src/learner.cc:516:
## Parameters: { set_seed } might not be used.
##
##   This may not be accurate due to some parameters are only used in language bindings but
##   passed down to XGBoost core.  Or some parameters are not used but slip through this
##   verification. Please open an issue if you find above cases.
##
##
## [1]  train-auc:0.845138
## [2]  train-auc:0.900112
## [3]  train-auc:0.937300
## [4]  train-auc:0.954677
## [5]  train-auc:0.964713
## [6]  train-auc:0.973820
## [7]  train-auc:0.988154
## [8]  train-auc:0.990990
```

```
## [9]  train-auc:0.994497
## [10] train-auc:0.996623
## [11] train-auc:0.998737
## [12] train-auc:0.998955
## [13] train-auc:0.999077
## [14] train-auc:0.999453
## [15] train-auc:0.999599
## [16] train-auc:0.999781
## [17] train-auc:0.999800
## [18] train-auc:0.999970
## [19] train-auc:1.000000
## [20] train-auc:1.000000
```

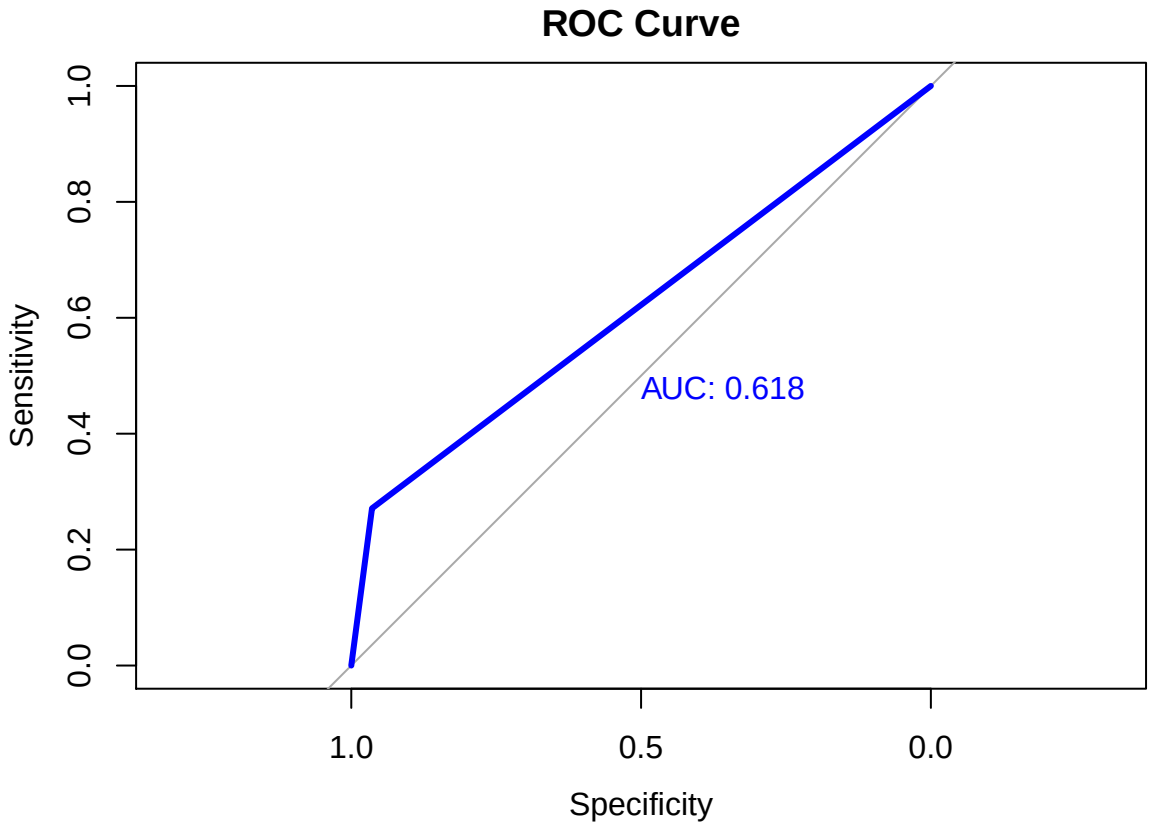


	Values
Accuracy	0.8528610
precision	0.8670520
F1-sensitivity	0.9740260
specificity	0.2203390
f1_score	0.9174312
AUC	0.5971825

Above `xgb.plot.shap` shows top 5 feature has more impact on attrition. MonthlyIncome graph clearly shows persion has high income tends less likely to leave the company than the person have low income.

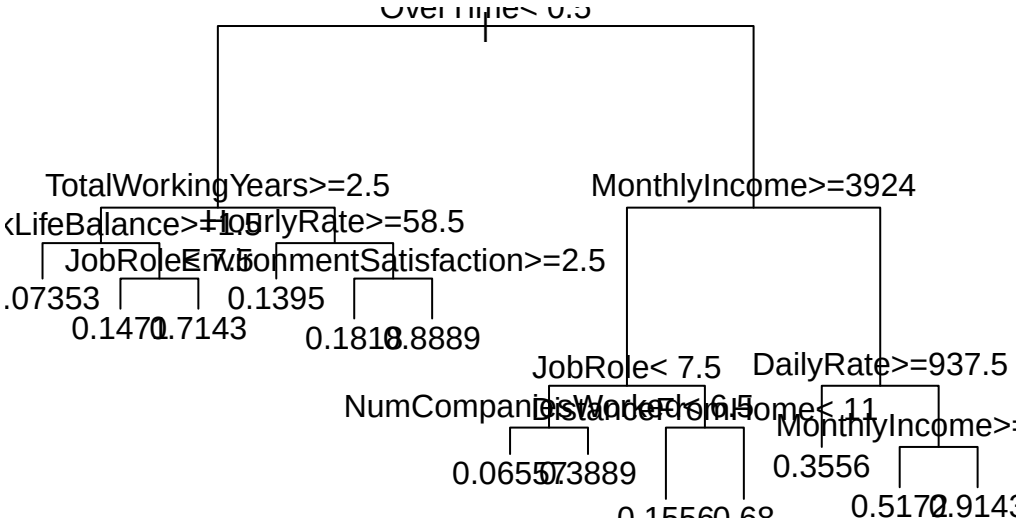
Model 5 - Support Vector Machines A technique that's typically used for classification but can be transformed to perform regression. It draws a division between classes that's as wise as possible

	Values
Accuracy	0.8528610
precision	0.8735294
F1-sensitivity	0.9642857
specificity	0.2711864
f1_score	0.9166667
AUC	0.6177361



Model 6 - Decesion Tree Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter.

```
model6 <- rpart(formula = Attrition ~., data=hr_train)
plot(model6)
text(model6)
```



	Values
Accuracy	0.8092643
precision	0.8584337
F1-sensitivity	0.9253247
specificity	0.2033898
f1_score	0.8906250
AUC	0.5643573

To see matrices from all models add the data in to a dataframe.

Show 10 entries Search:

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Accuracy	0.861035422343324	0.852861035422343	0.847411444141689	0.852861035422343	0.852861035422343	0.809264305177112
precision	0.881305637982196	0.867052023121387	0.857954545454545	0.867052023121387	0.873529411764706	0.858433734939759
F1-sensitivity	0.964285714285714	0.974025974025974	0.980519480519481	0.974025974025974	0.964285714285714	0.925324675324675
specificity	0.322033898305085	0.220338983050847	0.152542372881356	0.220338983050847	0.271186440677966	0.203389830508475
f1_score	0.92093023255814	0.917431192660551	0.915151515151515	0.917431192660551	0.916666666666667	0.890625
AUC	0.6431598062954	0.597182478538411	0.566530926700418	0.597182478538411	0.61773607748184	0.564357252916575

Showing 1 to 6 of 6 entries Previous 1 Next

Among all 6 models, Model 1 is doing well due to high accuracy and and high AUC.

Summary

From what we have seen so far, we have been able to come to the following conclusions:

- Employees stay if they have high income or more working years at company
- Employees who stay they gets promotion or they are satisfy with their ratings
- Employees who have stock options they more likely to stay
- Employees prefer to leave if they do overtime

The Logistic Regression models provides better result than any other models.

Reference :

XGBOOST - https://www.youtube.com/watch?v=frCu6eSI8R0

SVM - https://www.datacamp.com/community/tutorials/support-vector-machines-r

RF - https://towardsdatascience.com/random-forest-in-r-f66adf80ec9