STA 545/EAS 506
Statistical Data Mining I

Analysis of IBM HR Employee Attrition and Performance Data

Group 7

Prasanna Krishna Reddy Jeedipally (50441716)
Bharath reddy Madi (50441662)
Divya Sharvani Kandukuri (50442906)

Dataset

IBM HR Analytics Employee Attrition & Performance

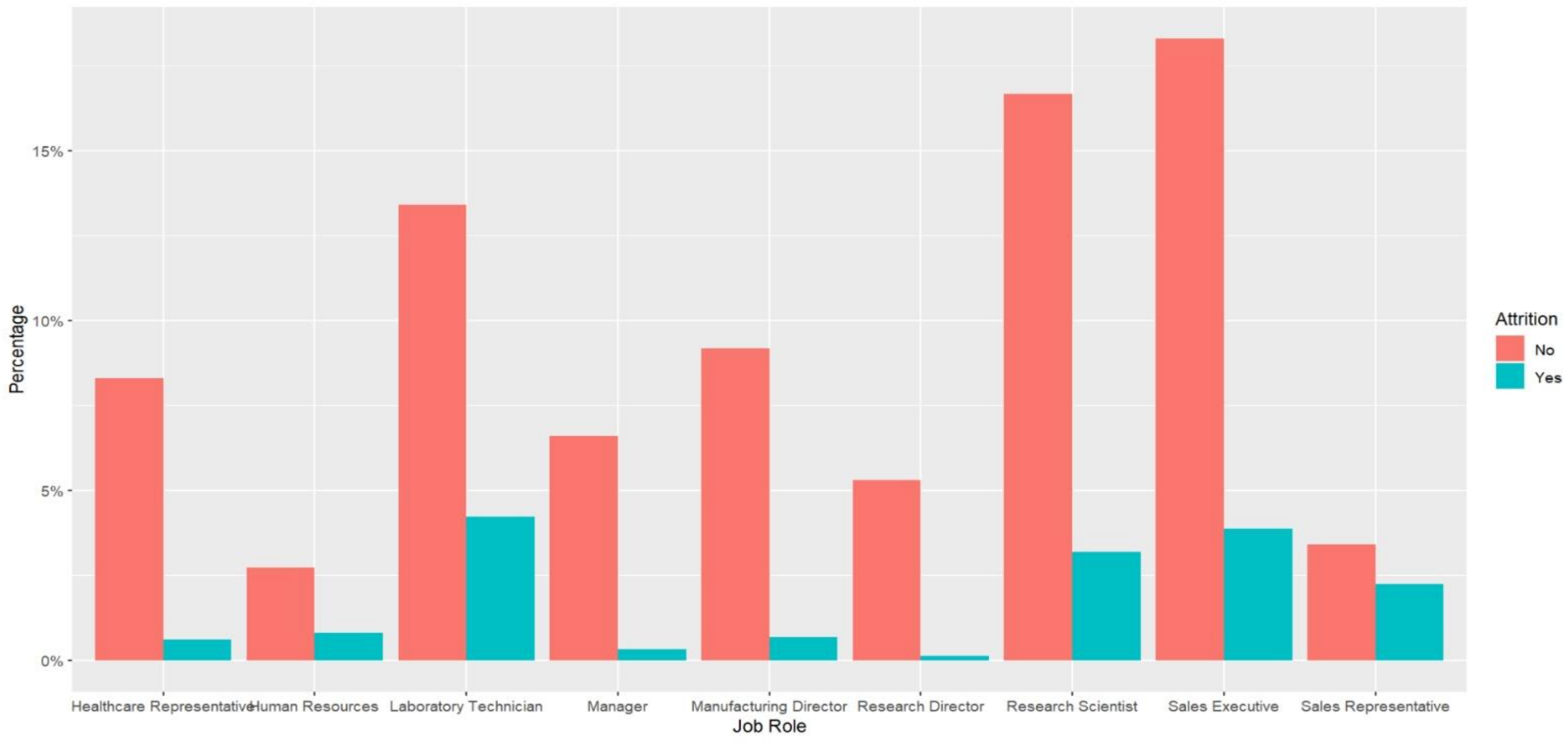
Discover the factors that lead to employee attrition

```
## spc_tbl_ [1,470 x 35] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Age                     : num [1:1470] 41 49 37 33 27 32 59 30 38 36 ...
##  $ Attrition               : chr [1:1470] "Yes" "No" "Yes" "No" ...
##  $ BusinessTravel          : chr [1:1470] "Travel_Rarely" "Travel_Frequently" "Travel_Rarely" "Travel
##  $ DailyRate               : num [1:1470] 1102 279 1373 1392 591 ...
##  $ Department              : chr [1:1470] "Sales" "Research & Development" "Research & Development" "
##  $ DistanceFromHome        : num [1:1470] 1 8 2 3 2 2 3 24 23 27 ...
##  $ Education               : num [1:1470] 2 1 2 4 1 2 3 1 3 3 ...
##  $ EducationField          : chr [1:1470] "Life Sciences" "Life Sciences" "Other" "Life Sciences" ..
##  $ EmployeeCount           : num [1:1470] 1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber          : num [1:1470] 1 2 4 5 7 8 10 11 12 13 ...
##  $ EnvironmentSatisfaction : num [1:1470] 2 3 4 4 1 4 3 4 4 3 ...
##  $ Gender                  : chr [1:1470] "Female" "Male" "Male" "Female" ...
##  $ HourlyRate              : num [1:1470] 94 61 92 56 40 79 81 67 44 94 ...
##  $ JobInvolvement          : num [1:1470] 3 2 2 3 3 3 4 3 2 3 ...
##  $ JobLevel                : num [1:1470] 2 2 1 1 1 1 1 1 3 2 ...
##  $ JobRole                 : chr [1:1470] "Sales Executive" "Research Scientist" "Laboratory Technic
##  $ JobSatisfaction         : num [1:1470] 4 2 3 3 2 4 1 3 3 3 ...
##  $ MaritalStatus           : chr [1:1470] "Single" "Married" "Single" "Married" ...
##  $ MonthlyIncome           : num [1:1470] 5993 5130 2090 2909 3468 ...
##  $ MonthlyRate             : num [1:1470] 19479 24907 2396 23159 16632 ...
##  $ NumCompaniesWorked      : num [1:1470] 8 1 6 1 9 0 4 1 0 6 ...
##  $ Over18                  : chr [1:1470] "Y" "Y" "Y" "Y" ...
##  $ OverTime                : chr [1:1470] "Yes" "No" "Yes" "Yes" ...
##  $ PercentSalaryHike       : num [1:1470] 11 23 15 11 12 13 20 22 21 13 ...
```

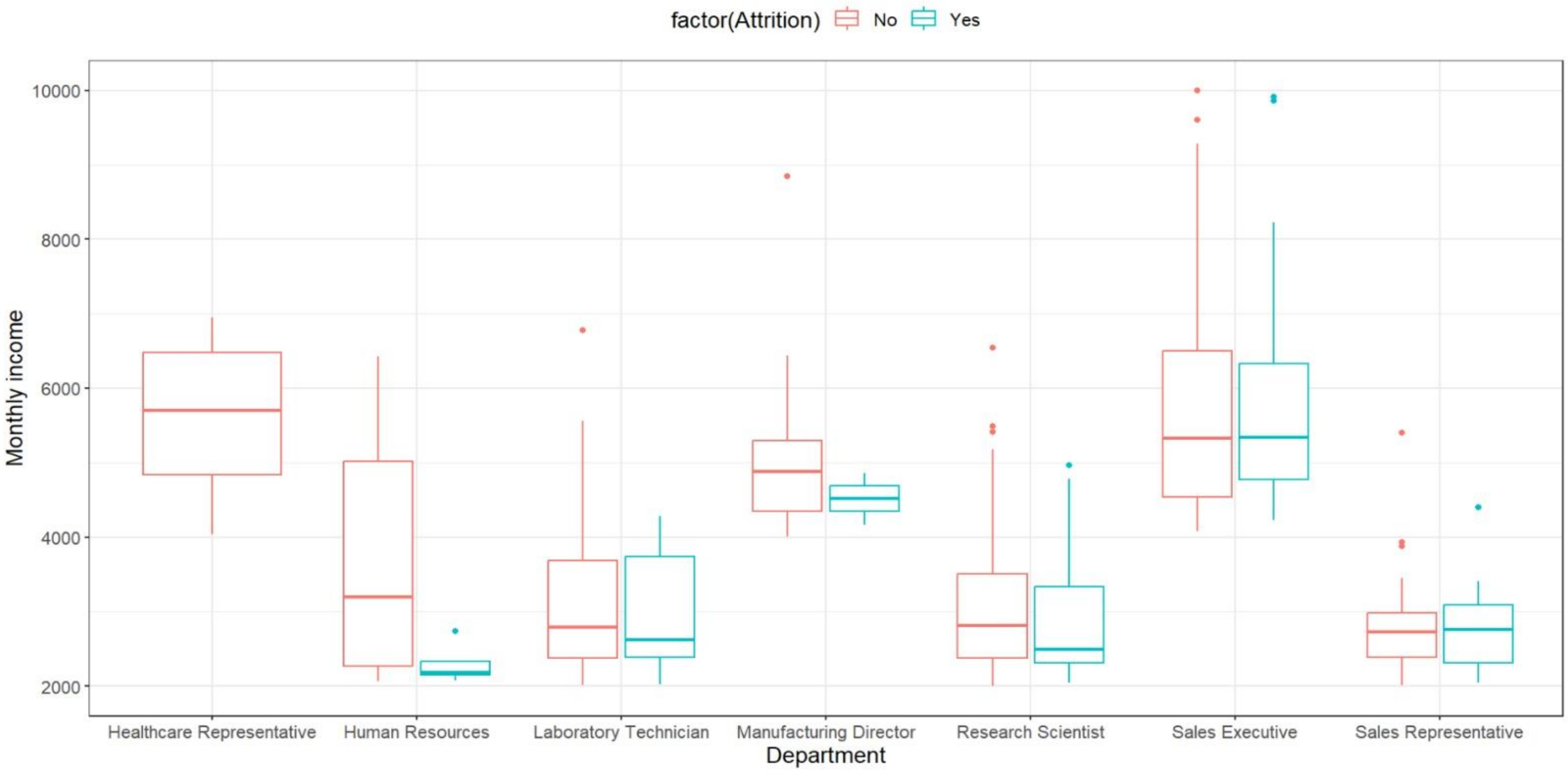
4

```
##  $ PerformanceRating       : num [1:1470] 3 4 3 3 3 3 4 4 4 3 ...
##  $ RelationshipSatisfaction: num [1:1470] 1 4 2 3 4 3 1 2 2 2 ...
##  $ StandardHours           : num [1:1470] 80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : num [1:1470] 0 1 0 0 1 0 3 1 0 2 ...
##  $ TotalWorkingYears       : num [1:1470] 8 10 7 8 6 8 12 1 10 17 ...
##  $ TrainingTimesLastYear   : num [1:1470] 0 3 3 3 3 2 3 2 2 3 ...
##  $ WorkLifeBalance         : num [1:1470] 1 3 3 3 3 2 2 3 3 2 ...
##  $ YearsAtCompany          : num [1:1470] 6 10 0 8 2 7 1 1 9 7 ...
##  $ YearsInCurrentRole      : num [1:1470] 4 7 0 7 2 7 0 0 7 7 ...
##  $ YearsSinceLastPromotion : num [1:1470] 0 1 0 3 2 3 0 0 1 7 ...
##  $ YearsWithCurrManager    : num [1:1470] 5 7 0 0 2 6 0 0 8 7 ...
```

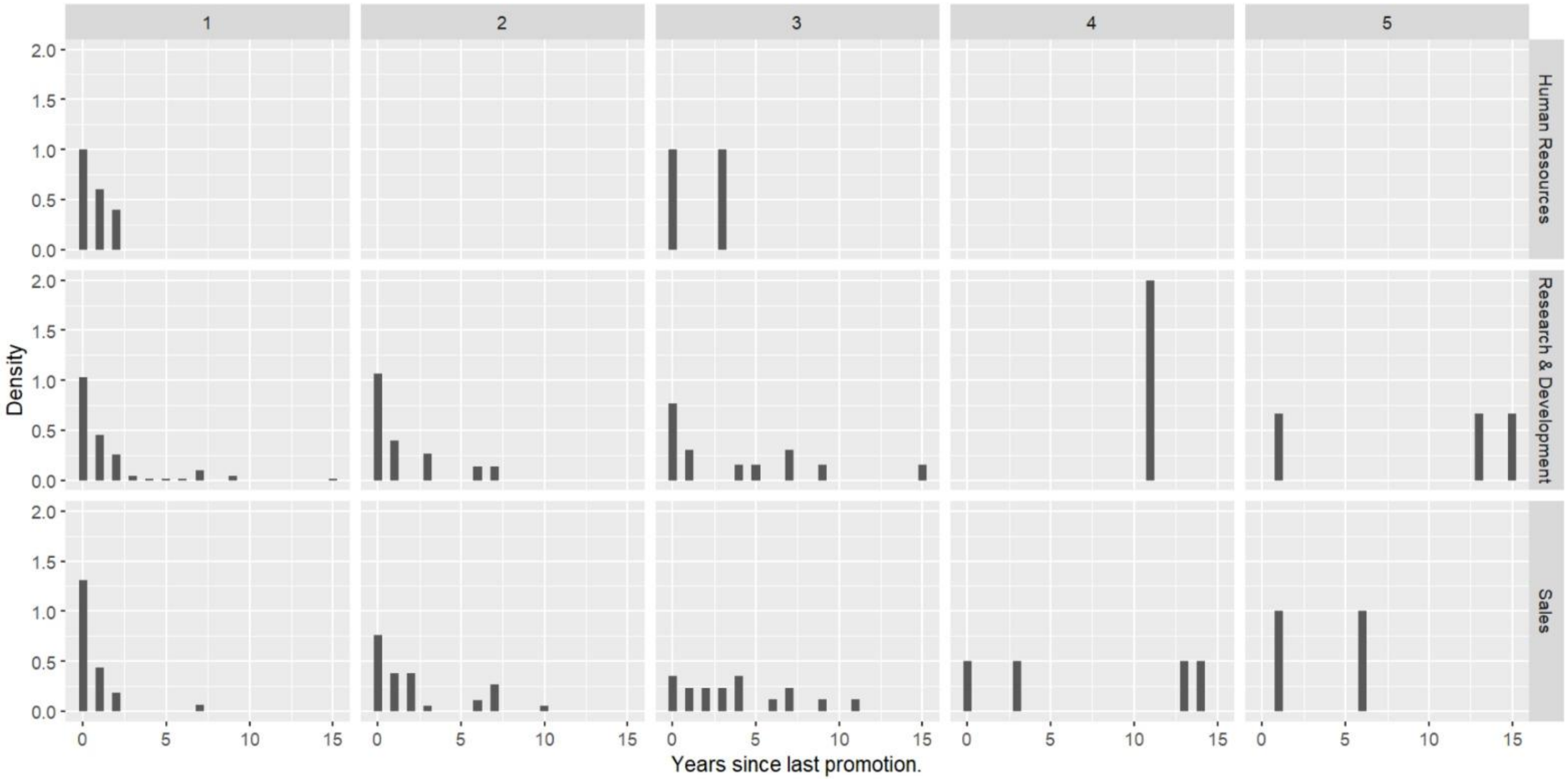
Bar chart to display Division of employees by job roles



Box plot to visualize division of employee's based on monthly income

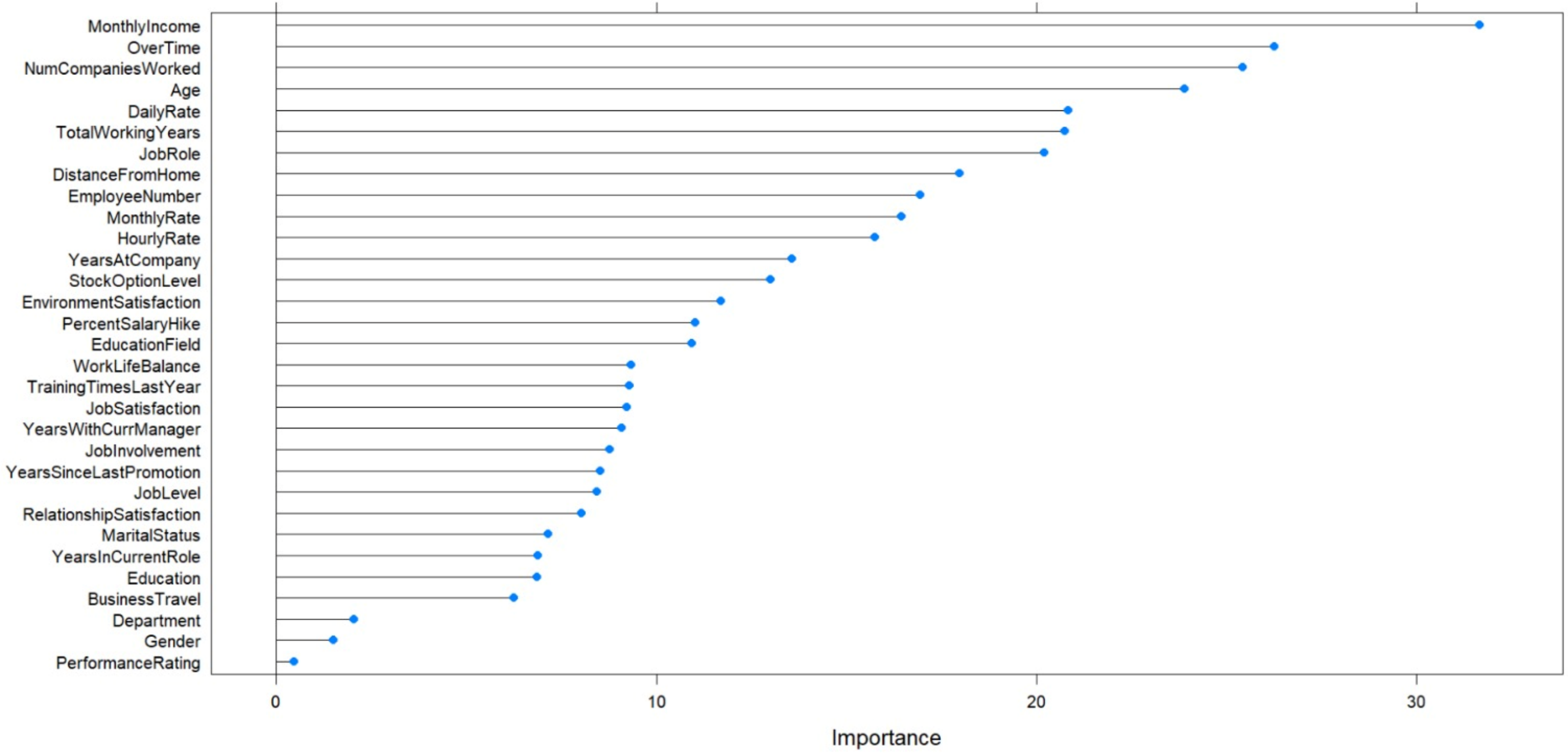


Department wise division of employees by years since last promotion



Feature Selection

Variable importance ranking using random forest



Rose - Random Over-Sampling Examples

- Functions to deal with binary classification problems in the presence of imbalanced classes. Synthetic balanced samples are generated according to ROSE.

Resampling the data

Data imbalance is handled by using ROSE

Before balancing the data

```
table(df_train$Attrition)
```

```
##
##   No Yes
## 864 166
```

After balancing the data

```
df_train %<>% as.data.frame()
#ROSE(admit~., data = train, N = 500, seed=111)$data
df_train <- ROSE(Attrition ~ .,
                 data=df_train,
                  N=1030,
                  seed=111)$data
```

```
table(df_train$Attrition)
```

```
##
##   No Yes
## 507 523
```

Supervised Learning Models

SVM using radial kernel

Support vector machines are a famous and a very strong classification technique which does not use any sort of probabilistic model like any other classifier but simply generates hyperplanes or simply putting lines, to separate and classify the data in some feature space into different regions.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  286  22
       Yes  83  49

               Accuracy : 0.7614
                 95% CI : (0.7187, 0.8005)
    No Information Rate : 0.8386
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3454

 Mcnemar's Test P-Value : 4.759e-09

            Sensitivity : 0.6901
            Specificity : 0.7751
         Pos Pred Value : 0.3712
         Neg Pred Value : 0.9286
             Prevalence : 0.1614
         Detection Rate : 0.1114
   Detection Prevalence : 0.3000
      Balanced Accuracy : 0.7326

       'Positive' Class : Yes
```

Random forest model

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction committee is more accurate than that of any individual tree.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  291  23
       Yes  78  48

               Accuracy : 0.7705
                 95% CI : (0.7283, 0.809)
    No Information Rate : 0.8386
    P-Value [Acc > NIR] : 0.9999

                  Kappa : 0.354

 Mcnemar's Test P-Value : 7.735e-08

            Sensitivity : 0.6761
            Specificity : 0.7886
         Pos Pred Value : 0.3810
         Neg Pred Value : 0.9268
             Prevalence : 0.1614
         Detection Rate : 0.1091
   Detection Prevalence : 0.2864
      Balanced Accuracy : 0.7323

       'Positive' Class : Yes
```

Boosted Logistic Regression

Boosting the logistic regression model is a way to convert a set of weak learners to a strong model. The weak learners specialize on different subsets of data. The subsequent models will do the classification task on the misclassified data. The final model can be a weighted sum of your weak models. With boosting, you can get better results since it can reduce bias as well as variance.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  268  20
       Yes 101  51

               Accuracy : 0.725
                 95% CI : (0.6807, 0.7662)
    No Information Rate : 0.8386
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3044

 Mcnemar's Test P-Value : 3.523e-13

            Sensitivity : 0.7183
            Specificity : 0.7263
         Pos Pred Value : 0.3355
         Neg Pred Value : 0.9306
             Prevalence : 0.1614
         Detection Rate : 0.1159
   Detection Prevalence : 0.3455
      Balanced Accuracy : 0.7223

       'Positive' Class : Yes
```

Stack of Models

Model Stacking is a way to improve model predictions by combining the outputs of multiple models that we modelled bove and running them through as another machine learning model called a meta-learner. This model has given the highest accuracy amongst the other three models. This is a kind of ensembling technique.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  308  24
       Yes  61  47

               Accuracy : 0.8068
                 95% CI : (0.7668, 0.8427)
    No Information Rate : 0.8386
    P-Value [Acc > NIR] : 0.9675

                  Kappa : 0.4103

 Mcnemar's Test P-Value : 9.432e-05

            Sensitivity : 0.6620
            Specificity : 0.8347
         Pos Pred Value : 0.4352
         Neg Pred Value : 0.9277
             Prevalence : 0.1614
         Detection Rate : 0.1068
   Detection Prevalence : 0.2455
      Balanced Accuracy : 0.7483

       'Positive' Class : Yes
```

	Models <chr>	Accuracy <dbl>	Recall <dbl>	Precision <dbl>	Time <dbl>
1	SVM RBF	0.76136...	0.69014...	0.37121...	25.29
2	Random Forest	0.77045...	0.67605...	0.38095...	405...
3	Stacking	0.80681...	0.66197...	0.43518...	97.28
4	Boosted Logistic Regression	0.72500...	0.71830...	0.33552...	15.66

4 rows

Evaluation

Conclusion

From all the above models trained the stack of models has given the highest accuracy and this model can used for prediction purposes in future work.

THANK YOU!