STATISTICAL DATA MINING PROJECT

PREDICT EMPLOYEE ATTRITION

Neha Bairathi

Ankit Singh

Nikhil Surve

BUSINESS DOMAIN

- HR analytics
  - Correlates people data with business data.
  - Helps managers and leadership to understand patterns in HR data.
  - Provides deeper insight into factors affecting business performance.
  - Supports better decision making for improving business processes.

USEFULNESS IN BUSINESSES

- HR analytics gaining popularity in recent times.
- Several organizations adopting HR analytics for multiple HR processes.
- IT giants like Microsoft, Google, IBM involved.
- Helps increase effectiveness of different HR processes related to recruitment, employee engagement and turnover.

EXAMPLE - IBM HR ANALYTICS

- Use employee data to make better workforce decisions and increase operational performance.
- Functions:
  - Forecast workforce requirements and determine how to fill open positions.
  - Identify factors for greater employee satisfaction and productivity.
  - Discover reasons for employee attrition and identify high-value employees at risk of leaving.
  - Establish effective training and career development initiatives.
- IBM Kenexa solutions for HR analytics
  - Creates visualizations that uncover patterns in data based on HR questions.
  - Identifies drivers of an event and recommends alternative paths.

EXAMPLE - HR ANALYTICS BY GOOGLE

- Project Oxygen
  - Objective: To evaluate whether manager quality has an impact on performance.
  - Data collection using past performance appraisals, employee surveys and interviews.
  - Findings: Good management makes a difference and managers generally have a set of 8 key attributes.
  - Attributes involve being a good coach that helps in the employee's career development or being a good communicator with a clear vision and strategy for the team over having good technical skills.
  - Based on such insights, Google developed a new management training program around these skills.

EXAMPLE - HR ANALYTICS BY GOOGLE

- The PiLab
  - People and Innovation Lab
  - Conducted various experiments to determine most effective approaches for managing people and productive environment.
  - Google' free food program: A small attempt to create a workplace where employees don't want to leave. Also, improved the employee health by reducing the calorie intake using scientific experiments.

NEED FOR PREDICTING ATTRITION

- Losing a valuable employee in between a project might cause delay in project completion.
- Losing a key team member might hamper the team dynamics affecting all future projects.
- Finding the right replacement for the lost employee is a cumbersome task.
- Additional cost might be incurred in training the new employee.
- These unwanted situations could be avoided by predicting employee attrition and providing incentives for staying to employees with a chance of leaving.

PROBLEM STATEMENT

- Whenever a well-trained and well-adapted employee leaves the organization, it creates a vacuum.
- The project aims at identifying factors affecting employee attrition like salary hikes, growth opportunities, work environment, business travel opportunities, superior – subordinate relationship, recognition and appreciation, years since last promotion etc.
- These factors would then be used to predict employee attrition.
- This prediction would help in retaining valuable employees by providing incentives.

COMPLICATION

- Employee attrition could be attributed to a wide variety of factors ranging from personal preferences to work environment to monetary goals.
- Identifying the most relevant factors is a difficult task.
- Trade-off between building a simple model and making better predictions by including more independent variables.
- Find the right model which has a low false negative rate since classifying an employee at risk of leaving as not at risk is critical.

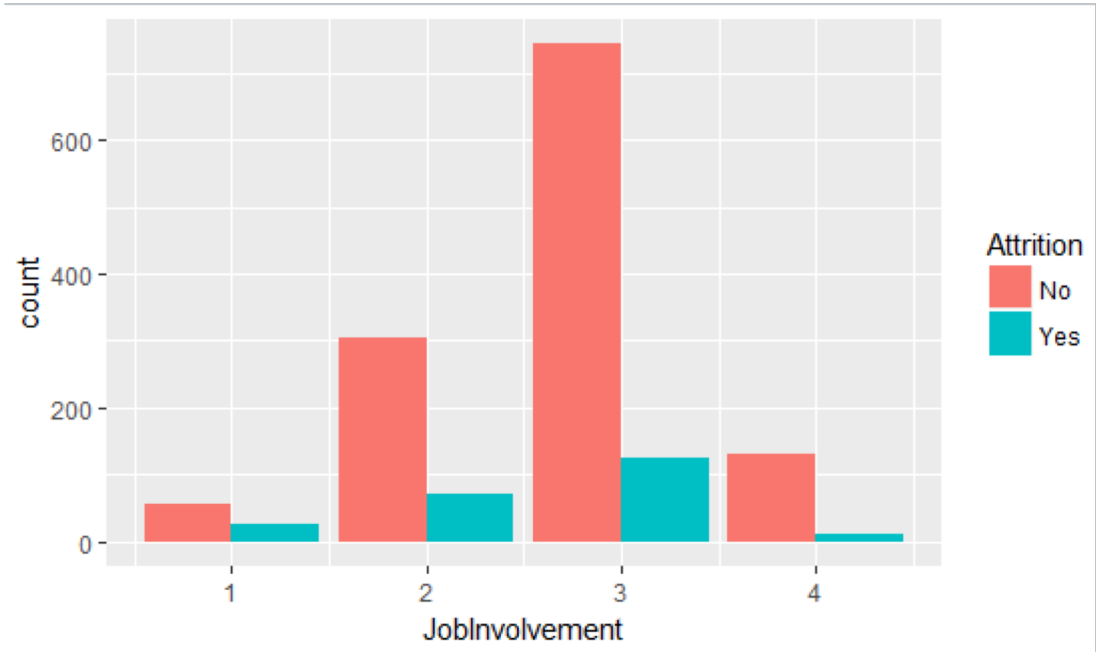
TURNING POINT

- Work with IBM employee attrition data set having 35 different attributes that could influence attrition.
- Explore the data and relationship between attrition and different attributes to identify the most relevant predictors.
- Run different classification models using the identified predictors and compare results to identify the best one.

EXPLORATORY DATA ANALYSIS - 1

- Attrition rate against job involvement

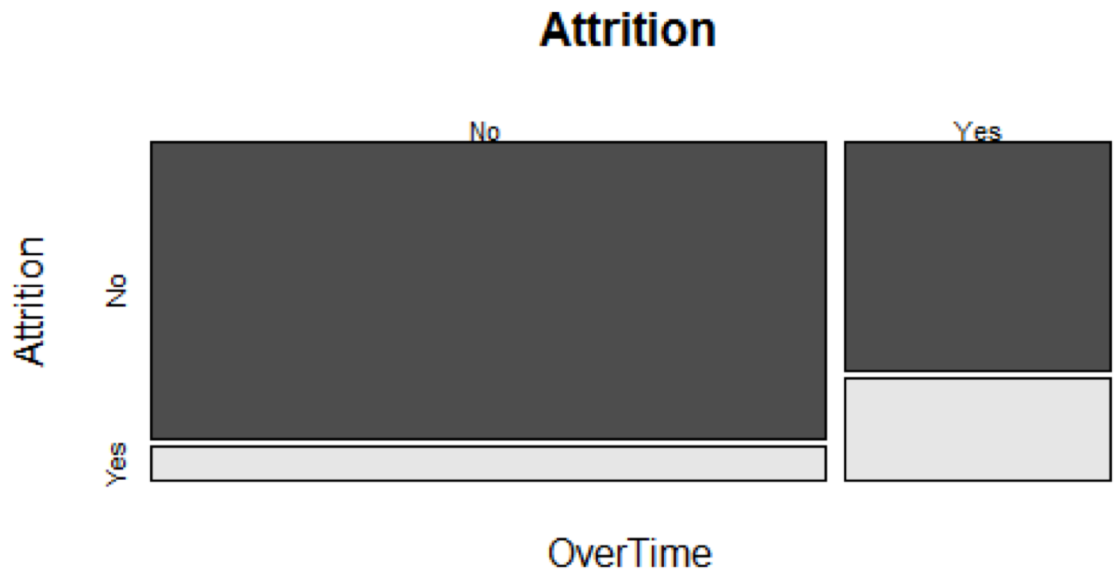
```
> tapply(Attrition$AttritionInt, Attrition$JobInvolvement, mean)
        1         2         3         4
0.33734940 0.18933333 0.14400922 0.09027778
```



EXPLORATORY DATA ANALYSIS - 2

- Attrition rate against overtime

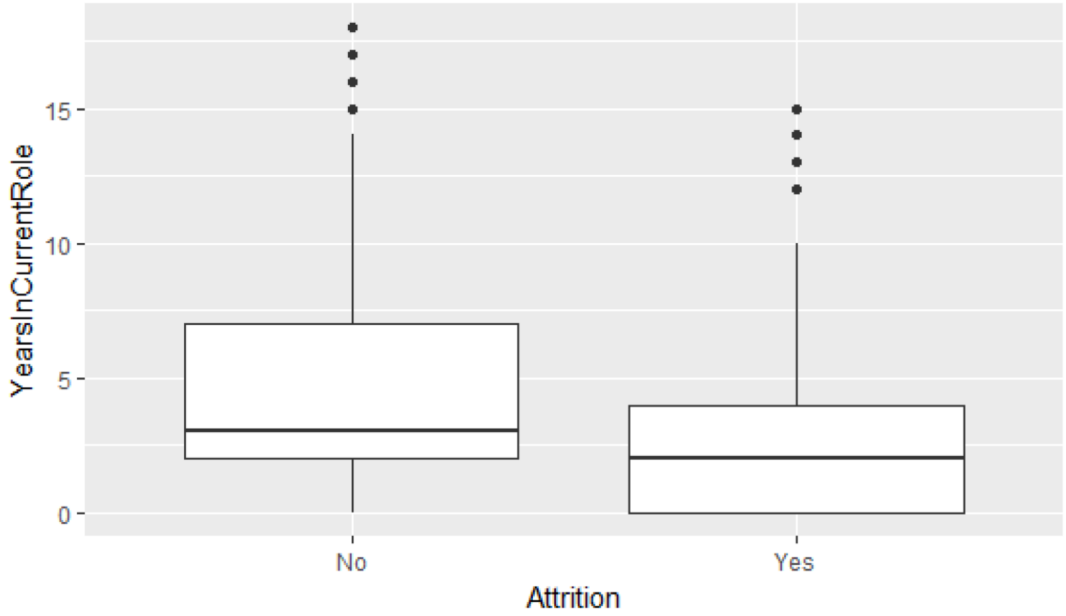
```
> tapply(Attrition$AttritionInt, Attrition$OverTime, mean)
       No       Yes
0.1043643 0.3052885
```



EXPLORATORY DATA ANALYSIS - 3

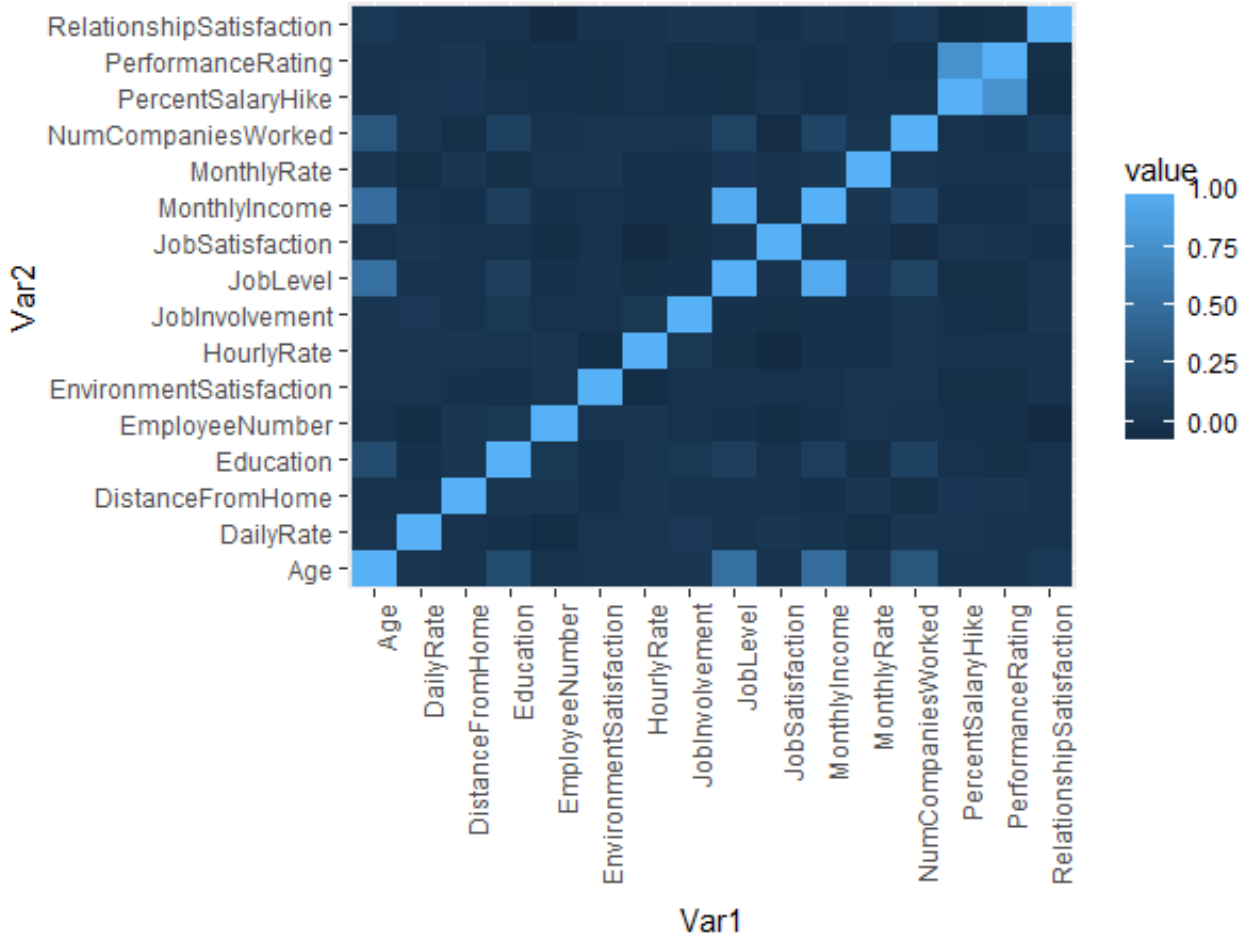
- Attrition rate against years in current role

```
> tapply(Attrition$AttritionInt, Attrition$YearsInCurrentRole, mean)
         0          1          2          3          4          5          6          7

0.29918033 0.19298246 0.18279570 0.11851852 0.14423077 0.02777778 0.05405405 0.13963964

         8          9         10         11         12         13         14         15

0.07865169 0.08955224 0.06896552 0.00000000 0.10000000 0.07142857 0.09090909 0.25000000

        16         17         18
0.00000000 0.00000000 0.00000000
```



EXPLORATORY DATA ANALYSIS - 4

- Correlation Matrix for all variables



LOGISTIC REGRESSION

- Accuracy of 85.94%

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6510  -0.5683  -0.3707  -0.2216   3.2495

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                2.97448    0.57299   5.191 2.09e-07 ***
EnvironmentSatisfaction   -0.44396    0.08675  -5.118 3.10e-07 ***
YearsInCurrentRole        -0.10804    0.03105  -3.480 0.000502 ***
TravelFrequentlyYes        0.74991    0.22152   3.385 0.000711 ***
JobInvolvement            -0.64422    0.12752  -5.052 4.38e-07 ***
OverTimeYes                1.59479    0.19506   8.176 2.93e-16 ***
Age                       -0.05966    0.01144  -5.214 1.85e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 915.90  on 1028  degrees of freedom
Residual deviance: 733.49  on 1022  degrees of freedom
AIC: 747.49

Number of Fisher Scoring iterations: 5
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  363   9
       Yes  53  16

               Accuracy : 0.8594
                 95% CI : (0.8234, 0.8905)
    No Information Rate : 0.9433
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2805
 Mcnemar's Test P-Value : 4.734e-08

            Sensitivity : 0.64000
            Specificity : 0.87260
         Pos Pred Value : 0.23188
         Neg Pred Value : 0.97581
             Prevalence : 0.05669
         Detection Rate : 0.03628
   Detection Prevalence : 0.15646
      Balanced Accuracy : 0.75630

       'Positive' Class : Yes
```

NAÏVE BAYES

- Accuracy of 85.49%

```
          Length Class      Mode
apriori   2      table      numeric
tables    6      -none-     list
levels    2      -none-     character
call      3      -none-     call
x         6      data.frame list
usekernel 1      -none-     logical
varnames  6      -none-     character
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  366   6
       Yes  58  11

               Accuracy : 0.8549
                 95% CI : (0.8185, 0.8864)
    No Information Rate : 0.9615
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2067
 Mcnemar's Test P-Value : 1.83e-10

            Sensitivity : 0.64706
            Specificity : 0.86321
         Pos Pred Value : 0.15942
         Neg Pred Value : 0.98387
             Prevalence : 0.03855
         Detection Rate : 0.02494
   Detection Prevalence : 0.15646
      Balanced Accuracy : 0.75513

       'Positive' Class : Yes
```

SUPPORT VECTOR MACHINE

- Accuracy of 84.81%

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.1428571

Number of Support Vectors:  379

 ( 213 166 )


Number of Classes:  2

Levels:
 No Yes
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  370   2
       Yes  65   4

               Accuracy : 0.8481
                 95% CI : (0.8111, 0.8803)
    No Information Rate : 0.9864
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0837
 Mcnemar's Test P-Value : 3.605e-14

            Sensitivity : 0.66667
            Specificity : 0.85057
         Pos Pred Value : 0.05797
         Neg Pred Value : 0.99462
             Prevalence : 0.01361
         Detection Rate : 0.00907
   Detection Prevalence : 0.15646
      Balanced Accuracy : 0.75862

       'Positive' Class : Yes
```

INFERENCE

- Logistic Regression provided us with the most accurate predictions.

- The 3 most important factors affecting the Attrition rate.
  - Overtime
  - Job Involvement
  - Environment Satisfaction

- Also, professional early in their career tend to switch the company at a higher rate as compared to those who have been working for a long time.

THANK YOU