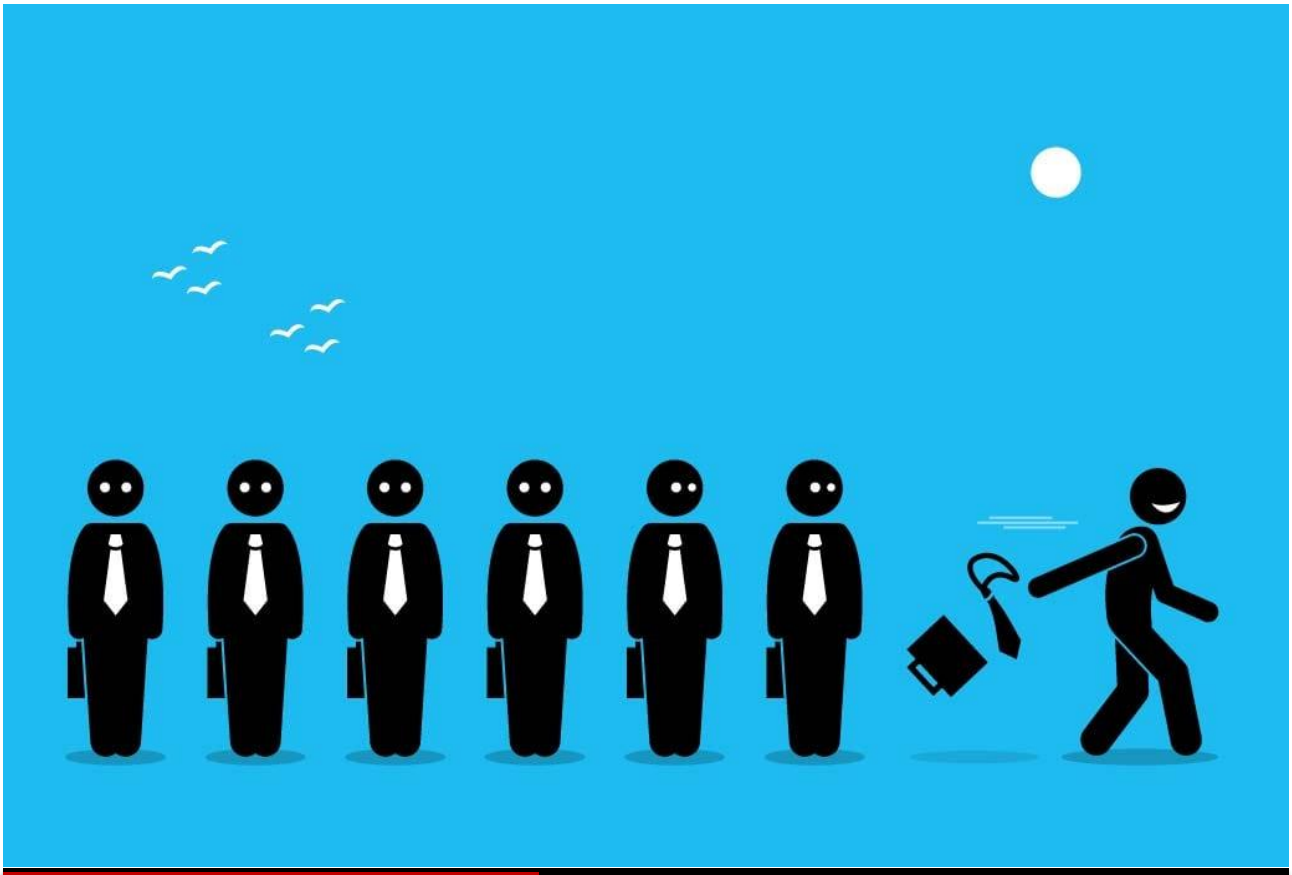
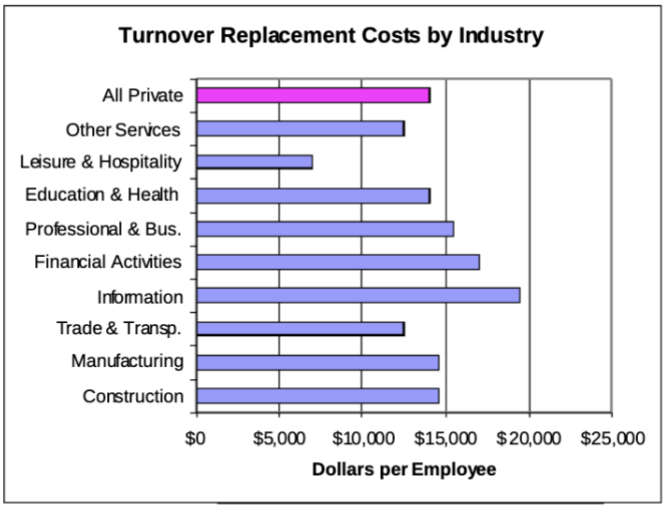


Employee Retention

Data Mining Final Project – Report

Prepared by Group 6 –
Harshit Mittal
Neelam Arya
Tawi Mankotia
Vishal Doshi

Executive Summary



Source: Employment Policy Foundation tabulation and analysis of Bureau of Labor Statistics, Employer Cost of Employee Compensation data.

Employee Attrition is expensive, often more than one would expect. In addition to replacement fees, the true cost of attrition involves costs such as productivity loss, workplace safety issues, and morale down spiral. It is the goal of this project to predict from sample data, the employees who are likely to leave and to uncover the causes of attrition. The findings and recommendations from this project **can save organizations hundreds of thousands of dollars yearly.**

Our proposal to tackle this problem is by first, using **2 classification models** (Support Vector Machines and Boosting Ensemble methods) to classify observations from a sample set of Employee Retention dataset into 2 classes. (i.e., whether they will leave the company or not). Second, to pit model performance of the 2 methods against each other and to pick the model that shows better relevant performance measures. (i.e., Specificity, Negative Prediction Value, and Accuracy). Finally, to use variable importance chart from the selected model, in combination with the frequency distribution of the target variable, to determine recommendations.

Upon performing the aforementioned tasks, we discover that **Support Vector Machines model** performs the best for our business case. It predicts attrition in the sample dataset with an **accuracy of 89%, Specificity of 97% and a Negative prediction Value of 90%**. The model tells us that the top 10 most important predictors for employee attrition are: Total working years, Overtime, Monthly income, Years at company, Years in current role, Job level, Years with current manager, Stock option level, Age and Job satisfaction score.

Introduction

Business Problem:

"To model and predict the factors that most greatly influence employee attrition."

Maintaining employee satisfaction and retaining them in the company is a challenge faced by companies since time immemorial. If an employee a company has invested heavily into leaves for "greener pastures", then this would result in company spending even more time and money to hire another resource, train them, bring them up to speed with company culture and the day to day workings. Leading to a huge monetary loss as well as a loss in efficiency, productivity and revenue.

Objectives:

- To determine which variables most greatly affect employee attrition using Support Vector Machines and Boosting Ensemble Learning Method.
- To Maximize Specificity, negative prediction rate and accuracy
- To model and predict which employee is likely to leave the company.

Motivation:

The great resignation that the market is experiencing right now, has only enhanced the importance of having a strong employee retention strategy in place. By using data analysis and data classification techniques, we want to firstly derive insights and secondly, use the same to the firm's advantage to reduce cost and increase productivity which has a direct impact on business revenue.

Analysis Methods:

Based on the target variable - "Attrition" and its classes, we decided to use supervised analysis methods to devise the model to predict attrition and identify the significant variables leading to attrition. Supervised methods are used when input and output variables are known and hence we used the following two methods:

- Support Vector Machine: It is a classification algorithm method used to distinctly identify two groups in a dataset by simply finding the maximum distance between the classes, which in our case are **YES** and **NO** for "Attrition" variable.
- Ensemble Method (boosting): It is a prediction model that predicts the target class for all the observations in the data. It runs classifiers in multiple iterations and accounts for misclassification and imposes more weights on previously misclassified records in the next iteration.

Data Exploration

Source:
The Employee Retention dataset is synthesized by IBM.

Dimensionality:
The dataset contains 1470 rows and 30 columns.

Data Quality:
There are no missing values in the dataset. There are also no outliers in the dataset. There is a class imbalance in the target variable that needs to be treated in order to increase prediction accuracy and reliability. The data is evenly distributed for all variable types, numerical and categorical (both nominal and ordinal).

Variables:

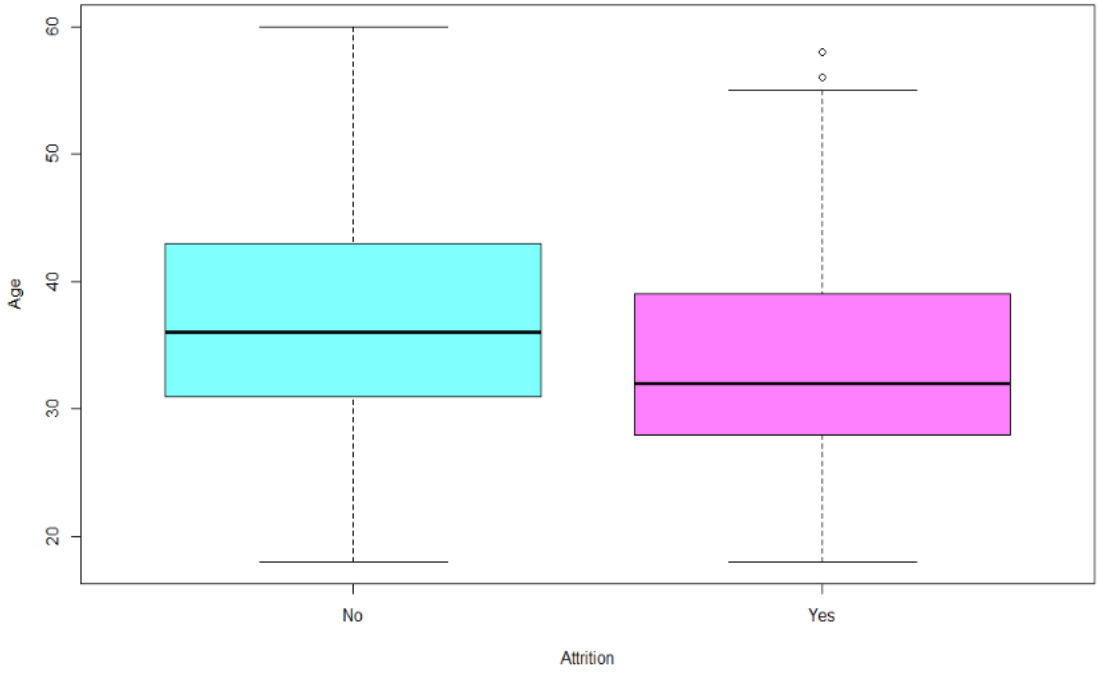
- Target variable: The Target variable is "Attrition", it is based on many variables in the dataset that contain information ranging from demographic, to organizational data. (E.g. Age, Monthly Income, Job Satisfaction score, Distance from home, Marital Status, etc.)

Variable Types		
Numerical	**Categorical**	
	Nominal	**Ordinal**
Age Distance from home Hourly Rate Monthly Income Number of Companies worked at Percent Salary Hike Total Working years Training Times Last Year Years at Company Years in Current Role Years Since Last Promotion Years with Current Manager	Department Gender Job Role Marital Status Over18 Overtime Attrition Education Field	Business Travel Environment Satisfaction Job Involvement Job Level Job Satisfaction Performance Rating Relationship Satisfaction Stock Option Level Work Life Balance Education

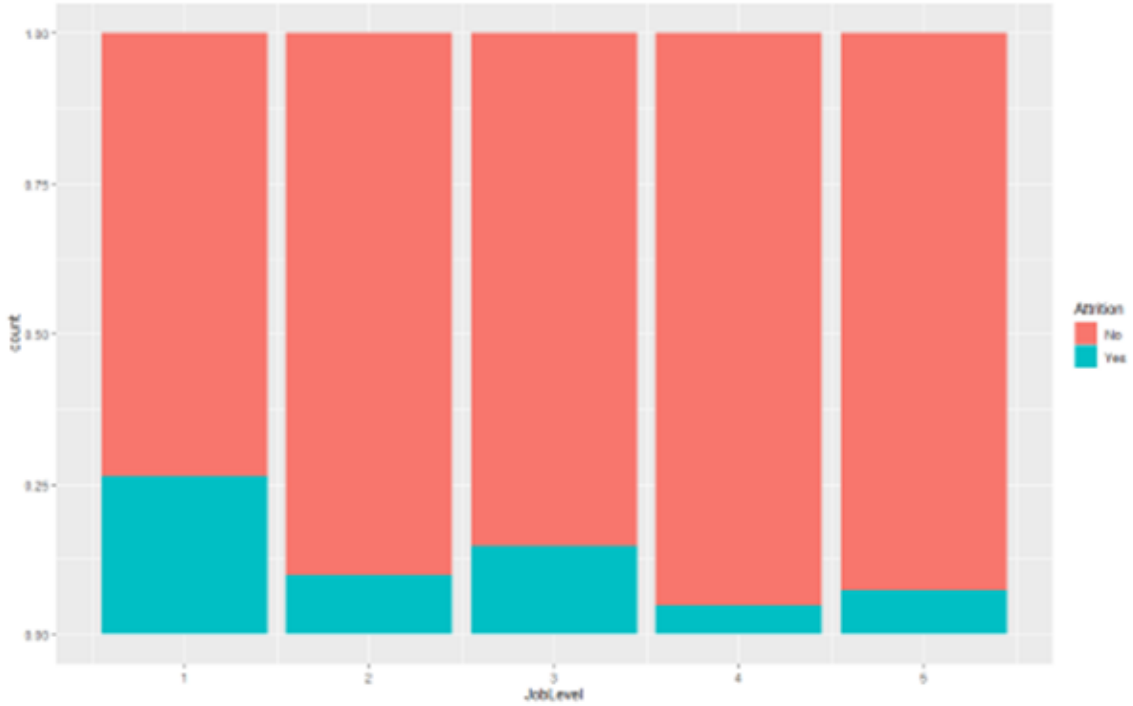
Exploratory Data Analysis

- As per the boxplot of distribution of Monthly Income over Attrition: People with a monthly salary higher than $5k are less likely to leave the company.
- As per the boxplot of distribution of Age over Attrition: Aged people seem less likely to leave the company.
- As per the boxplot of distribution of Number of Companies Worked at over Attrition: This has a similar distribution over both the classes in Attrition so we feel that this may not be a relevant variable for the prediction model.
- When we plot employee attrition rates with respect to the Job level, we notice that most attrition occurs among employees who are of the lowest and middle level of Jobs. This can be because they are new graduates looking to advance their careers quicker or those who don't feel like they are getting the recognition they deserve.

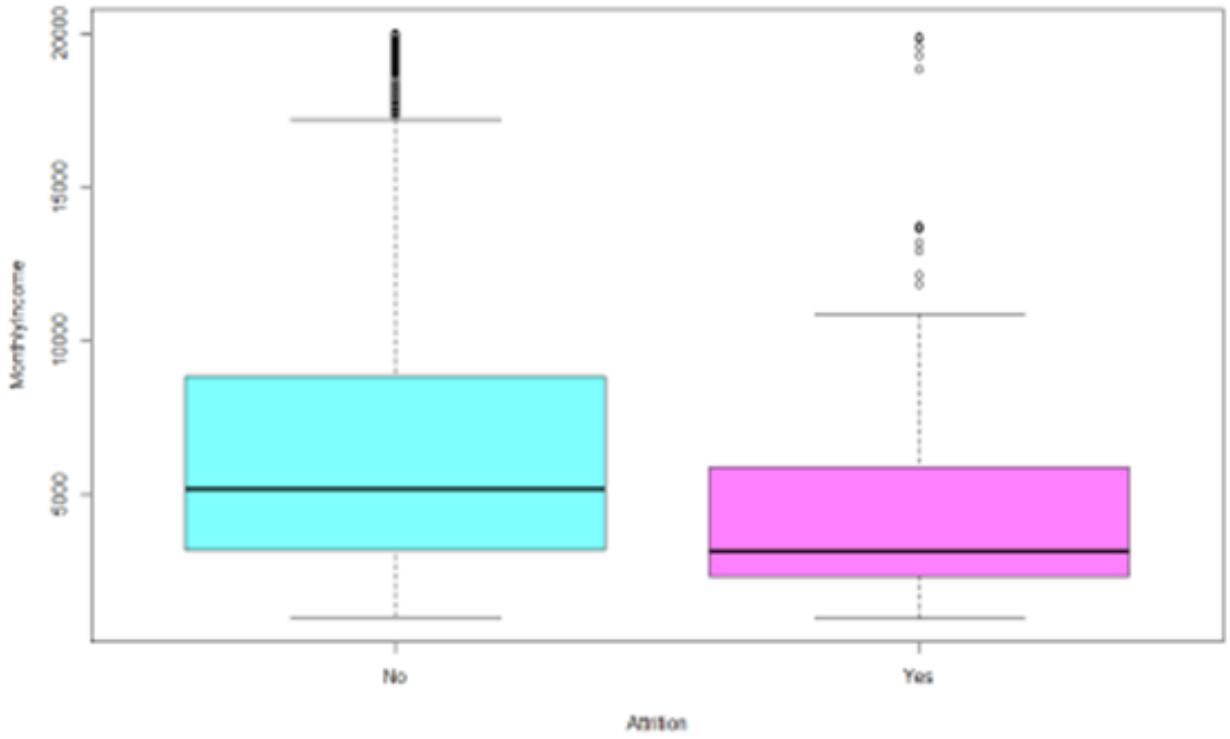
Data Visualizations:



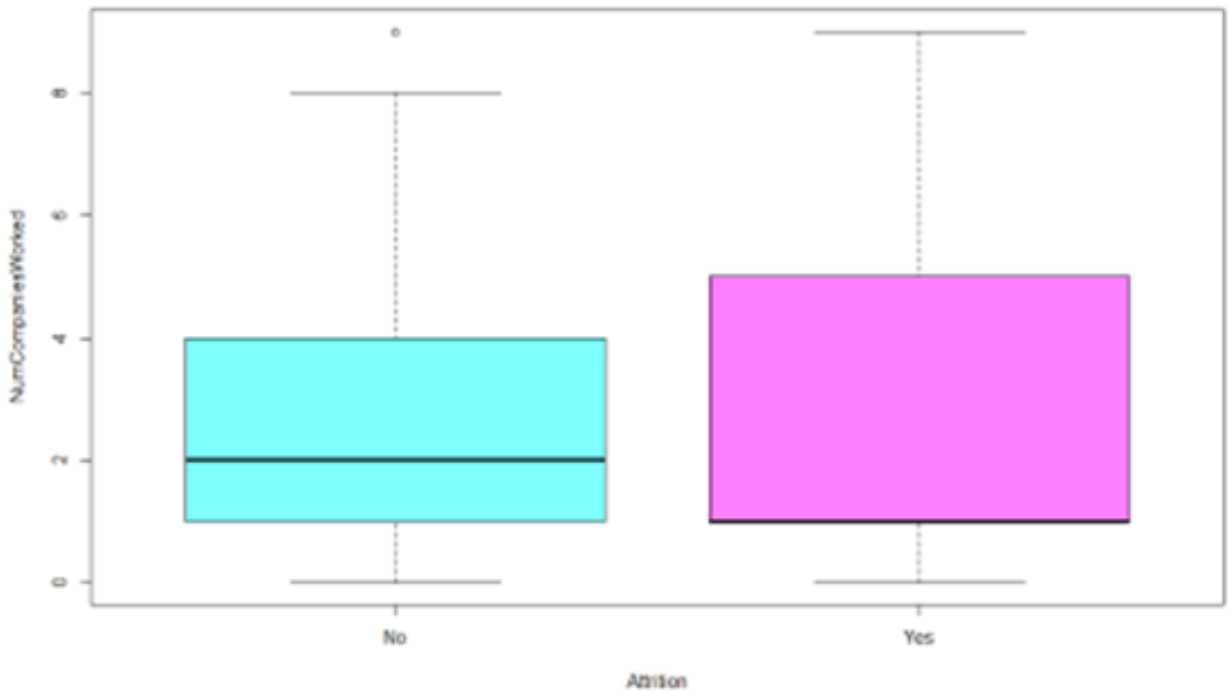
Graph 1 - Distribution of Age over Attrition



Graph 2 - Count of Attrition with respect to different Job Levels



Graph 3 - Distribution of monthly income with respect to Attrition



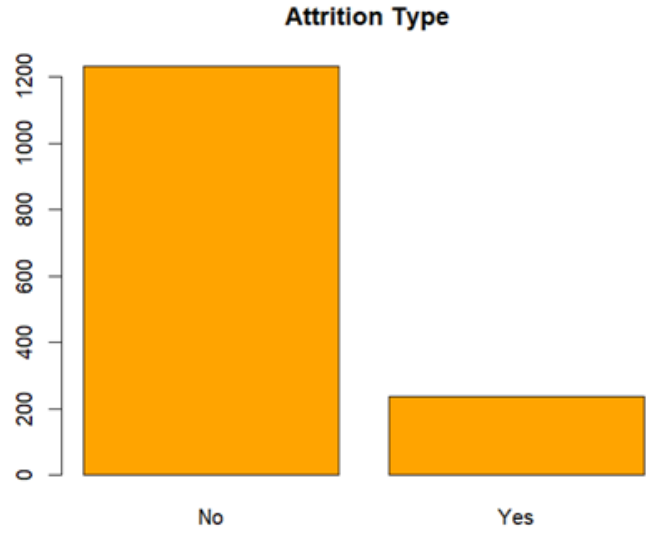
Graph 4 - Distribution of number of companies worked at with respect to Attrition

Data Pre-processing:

1. For **Support Vector Machine** analysis method:

    a. **Binarization -** Nominal variables are binarized so that we can more accurately predict the properties/impact of these variables on our analysis

    b. **Class Imbalance** – A huge class imbalance exists in data and this can impact the model's ability to correctly predict the Attrition rate. We will use weights method to treat class imbalance



c. **Weights** method used to treat class imbalance and assigned a higher weight to the minority class (Attrition = "Yes")
    i. No – 0.5962162
    ii. Yes – 3.09083146

d. **Trained** our model by using 75% of the data and the remaining 25% for **testing**
    i. Pre balancing data dimensionality - 1470 rows and 30 columns
    ii. Post balancing data dimensionality for Training set – 1103 obs. Of 46 variables
    iii. Post balancing data dimensionality for Testing set – 367 obs. Of 46 variables

2. For Ensemble (**Boosting**) analysis Method

a. **Missing Values** – No missing values were found

b. **Outlier Detection** – Checked for outliers using the Z-Score method and no outliers were detected

c. **Class Imbalance -** Boosting method takes care of the imbalance in the data, hence no external treatment is required

d. **Feature section** - Used wrapper method to remove irrelevant variables as the method is sensitive to irrelevant variables when more in number
    i. Irrelevant variables: Education, Hourly Rate, Over18, Percent Salary Hike, Performance Rating, Relationship Satisfaction Score, Training Times Last Year

e. Trained our model by using 75% of the data and the remaining 25% for testing
    i. Pre-balancing data dimensionality - 1470 rows and 30 columns
    ii. Post balancing data dimensionality for Training set – 1103 obs. Of 46 variables
    iii. Post balancing data dimensionality for Testing set – 367 obs. Of 46 variables

Analysis Results for Support Vector Machine

Test Model Performance:

A higher value of cost under hyperparameter tuning **(C = 4.63)** indicates that the model focuses on minimizing training errors.

```
Pre-processing: re-scaling to [0, 1] (64)
Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 882, 882, 883, 883, 882, 883, ...
Resampling results across tuning parameters:

  sigma        C           Accuracy   Kappa
  0.005083309  2.82987448  0.8694488  0.3255285
  0.007124801  0.20839469  0.8386261  0.0000000
  0.007313480  4.63409503  0.8727807  0.4213733
  0.008094202  0.11760278  0.8386261  0.0000000
  0.015936615  0.05935485  0.8386261  0.0000000

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.00731348 and C = 4.634095.
```

Goodness of fit:

```
> cbind(svm_train = SVM_trtune_conf$overall,
+       svm_test = SVM_tetune_conf$overall)
                    svm_train     svm_test
Accuracy       9.592022e-01 0.8882833787
Kappa          8.354610e-01 0.5004150204
AccuracyLower  9.457878e-01 0.8515061619
AccuracyUpper  9.700889e-01 0.9186275718
AccuracyNull   8.386219e-01 0.8392370572
AccuracyPValue 1.087843e-36 0.0048691991
McnemarPValue  8.025111e-08 0.0001781293
> # Class-Level
> cbind(svm_train = SVM_trtune_conf$byClass,
+       svm_test = SVM_tetune_conf$byClass)
                     svm_train   svm_test
Sensitivity          0.7696629 0.44067797
Specificity          0.9956757 0.97402597
Pos Pred Value       0.9716312 0.76470588
Neg Pred Value       0.9573805 0.90090090
Precision            0.9716312 0.76470588
Recall               0.7696629 0.44067797
F1                   0.8589342 0.55913978
Prevalence           0.1613781 0.16076294
Detection Rate       0.1242067 0.07084469
Detection Prevalence 0.1278332 0.09264305
Balanced Accuracy    0.8826693 0.70735197
```

- **Accurately** predicts **89%** of the unseen data.

- **Kappa Value of 50%** which is a moderate agreement of collected data with the variables. Which means that 50% of the time the model requires no additional variables to explain attrition in a particular observation. This rate can be improved by including more relevant variables in the data collection process or by including more observations for the training model.

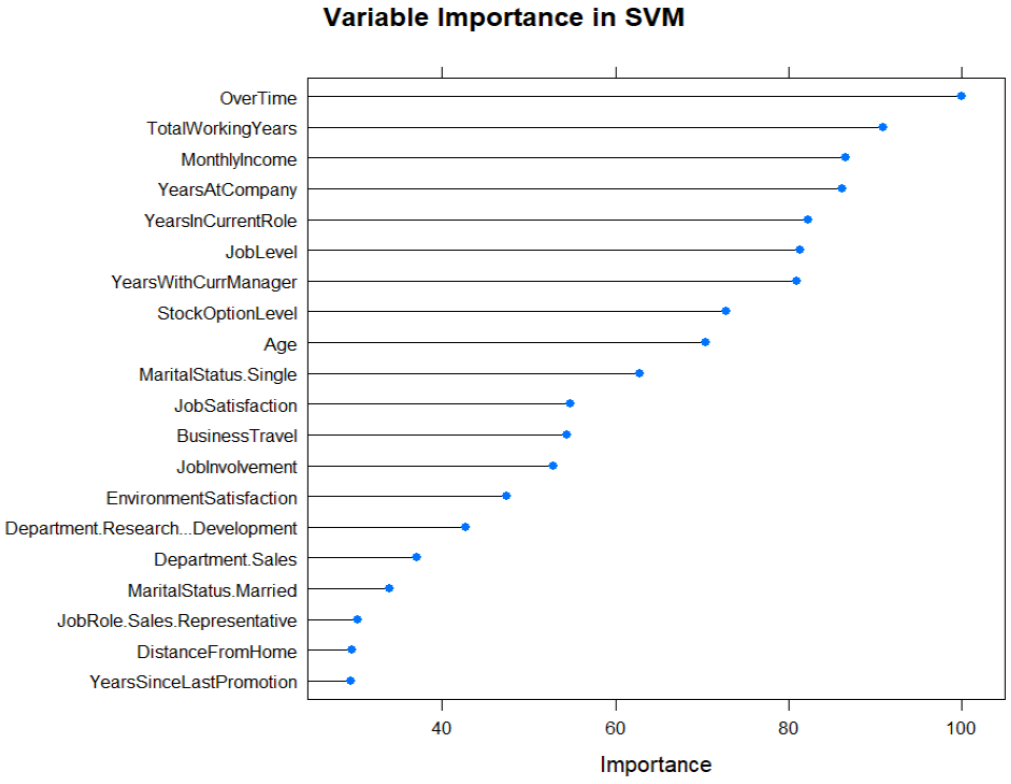
- It is a **balanced model**. It performs well on training as well as testing data. Close to 90% accuracy on both training and testing data.

- The model has high **Specificity (97%)**. This means that the model minimizes false negative rate to 3%.

- The probability of precisely predicting the **Positive Attrition rate (Yes) is 77%**. Which means that of all people predicted to leave the company, 77% of them actually will leave the company.

- The probability of precisely predicting the **Negative Attrition rate (No) is 90%.** Which means that of all people predicted to stay at the company, 10% of them actually will leave the company.

Variable Importance:



Model Validation for Business Case

- The model predicts attrition with a high Specificity, which will lead to a minimal opportunity cost of misclassifying employees who will leave.

- This coupled with a very high Negative Attrition prediction rate for the model, validates it for the business case.

Analysis Results for Ensemble (Boosting) Method

Test Model Performance and Goodness of fit:

```
                       bag           rf        boost
Accuracy       0.8474114441 8.610354e-01 8.664850e-01
Kappa          0.2882177738 2.838339e-01 3.825155e-01
AccuracyLower  0.8064711061 8.213654e-01 8.273538e-01
AccuracyUpper  0.8826317547 8.947507e-01 8.995669e-01
AccuracyNull   0.8392370572 8.392371e-01 8.392371e-01
AccuracyPValue 0.3665275738 1.426306e-01 8.614255e-02
McnemarPValue  0.0001064987 2.129708e-08 6.334248e-05
> ## By Class
> cbind(bag = ph[["bagging"]]$byClass,
+       rf = ph[["randforest"]]$byClass,
+       boost = ph[["boosting"]]$byClass)
                           bag         rf      boost
Sensitivity          0.27118644 0.22033898 0.33898305
Specificity          0.95779221 0.98376623 0.96753247
Pos Pred Value       0.55172414 0.72222222 0.66666667
Neg Pred Value       0.87278107 0.86819484 0.88427300
Precision            0.55172414 0.72222222 0.66666667
Recall               0.27118644 0.22033898 0.33898305
F1                   0.36363636 0.33766234 0.44943820
Prevalence           0.16076294 0.16076294 0.16076294
Detection Rate       0.04359673 0.03542234 0.05449591
Detection Prevalence 0.07901907 0.04904632 0.08174387
Balanced Accuracy    0.61448932 0.60205261 0.65325776
```

- **Accurately predicts 89%** of the unseen data

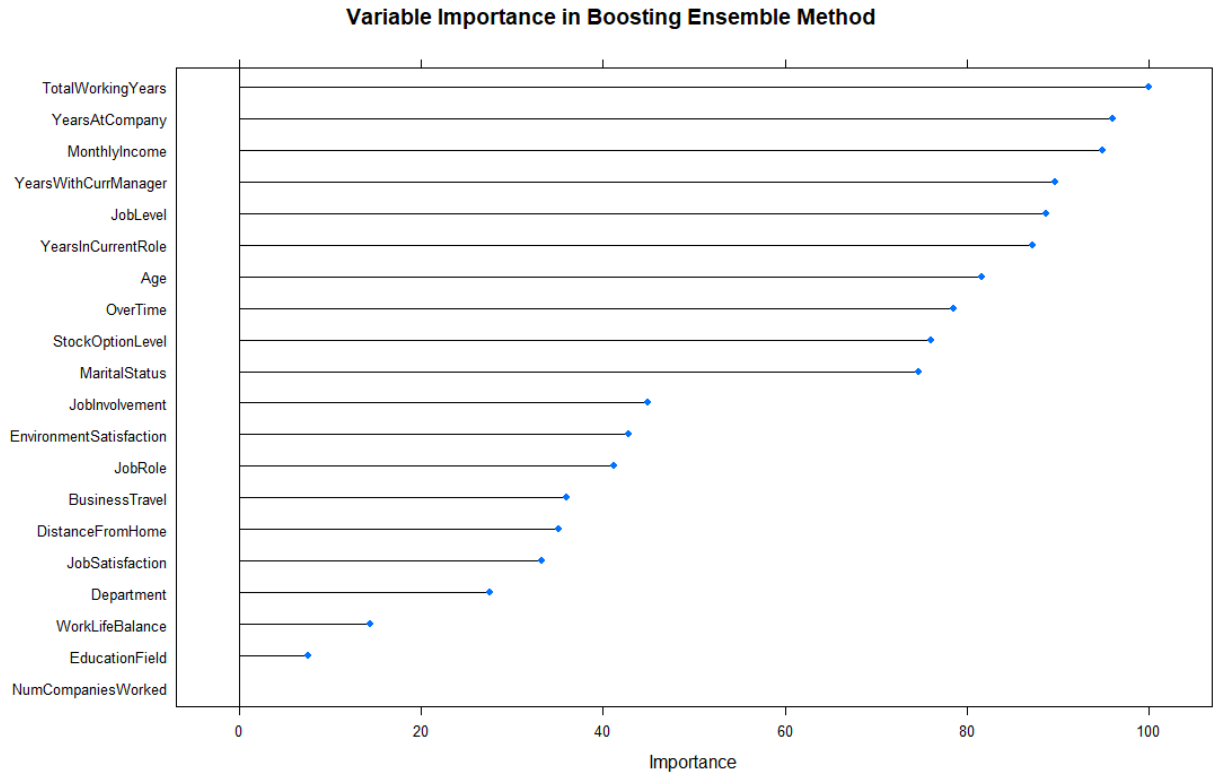
- **Kappa Value of 36.4%** which is a fair agreement of the collected data with the variables. Which means that 36.4% of the time the model requires no additional variables to explain attrition in a particular observation. This rate can be improved by including more relevant variables in the data collection process or by including more observations for the training model.

- The model has high **Specificity (96%)**. This means that the model minimizes false negative rate to 4%.

- The probability of precisely predicting the **Positive Attrition rate (Yes) is 66%.** Which means that of all people predicted to leave the company, 66% of them actually will leave the company.

- The probability of precisely predicting the **Negative Attrition rate (No) is 88%.** Which means that of all people predicted to stay at the company, 12% of them actually will leave the company.

Variable Importance:



Model Validation for Business Case - Boosting

- The model predicts attrition with a high Specificity, which will lead to a minimal opportunity cost of misclassifying employees who will leave.

- This coupled with a very high Negative Attrition prediction rate for the model, validates it for the business case.

Conclusion and Recommendations

Both SVM and Boosting methods give us models with high levels of accuracy, specificity, and positive/ negative prediction rates, which are the most relevant measures for our business case.

We recommend that the SVM model be chosen for this business problem as the performance measures are higher, albeit slightly when compared to Boosting.

Findings of the analysis

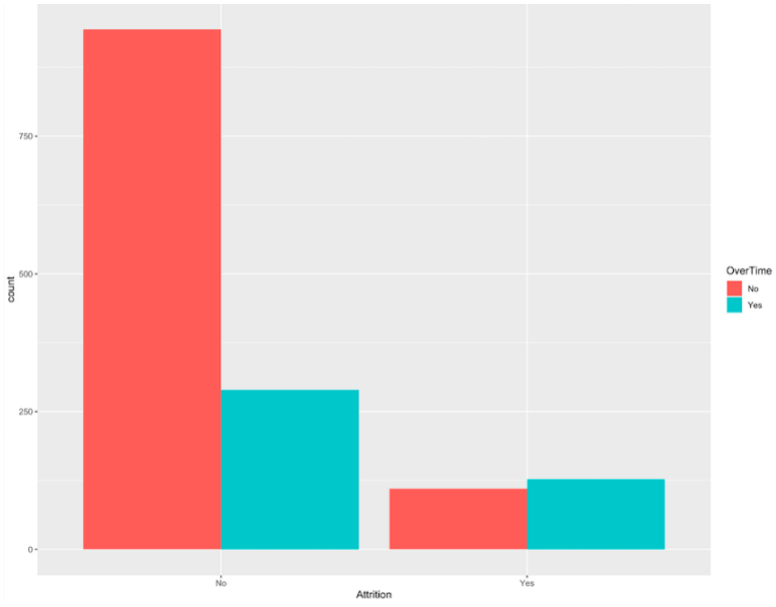
- The **SVM method predicts, with an 89% accuracy**, whether a given employee will leave the company based on the predictor variables.
- The method also gives us a **chart of the predictors and their importance to making the predictions**.

Implications of the findings for the business case:

The business can **use the information from the model**, i.e. the relative importance of the predictor to **make policy, and organizational decisions** on what will **minimize attrition rate** among its employees. The analysis addresses the business problem:

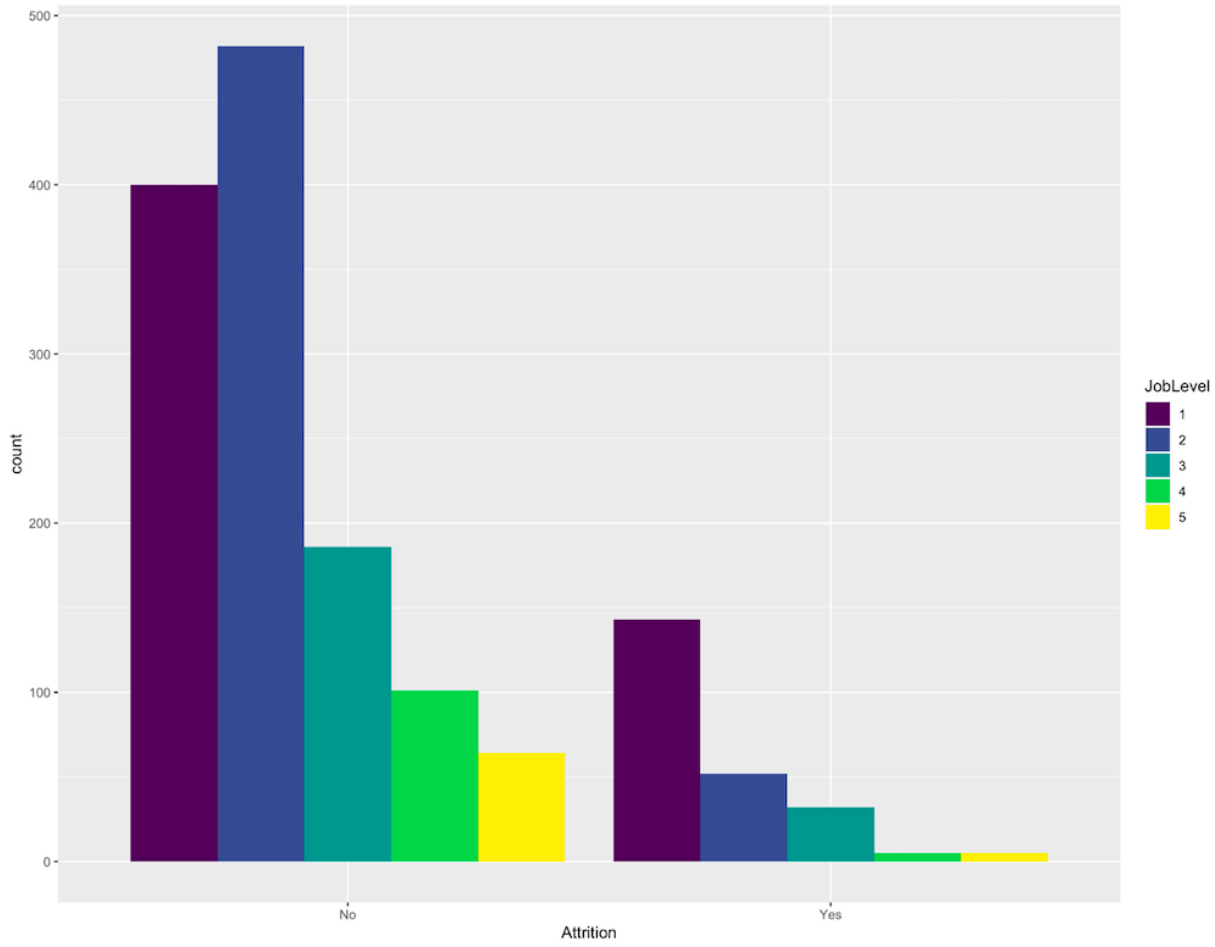
According to the results of the SVM model, we find the following 3 variables to be the most important Numerical, Nominal and Ordinal variables to determine employee Attrition. We make recommendations based on the distribution of these variables to understand what influences people to leave the company.

1. **Nominal - Overtime**



- Overtime is the most important nominal variable to predict attrition as per SVM model.
- Overtime can help a few employees take a bigger paycheck home, however, if done excessively it can lead to burnout.
- The management must strike a balance here.
- Overtime should be properly scheduled.

2. **Ordinal - Job Level**



- Job level is the most important ordinal variable to predict attrition as per SVM model
- To enhance job involvement among low job levels, the first approach to hire for a vacant position should be to hire internally
- This will encourage training and mentoring among low job levels and enhance job involvement score.

3. **Numerical - Monthly Income**

- Monthly income is the most important numerical variable to predict attrition as per SVM model
- Significant emphasis should be put on the other components of the compensation such as - bonus, allowance, benefits etc.
- Rewards and recognition for high performers