EMPLOYEE RETENTION

FINAL PROJECT – STAT 642

Presented by Group 6:
Harshit Mittal
Neelam Arya
Tawi Mankotia
Vishal Doshi

Introduction

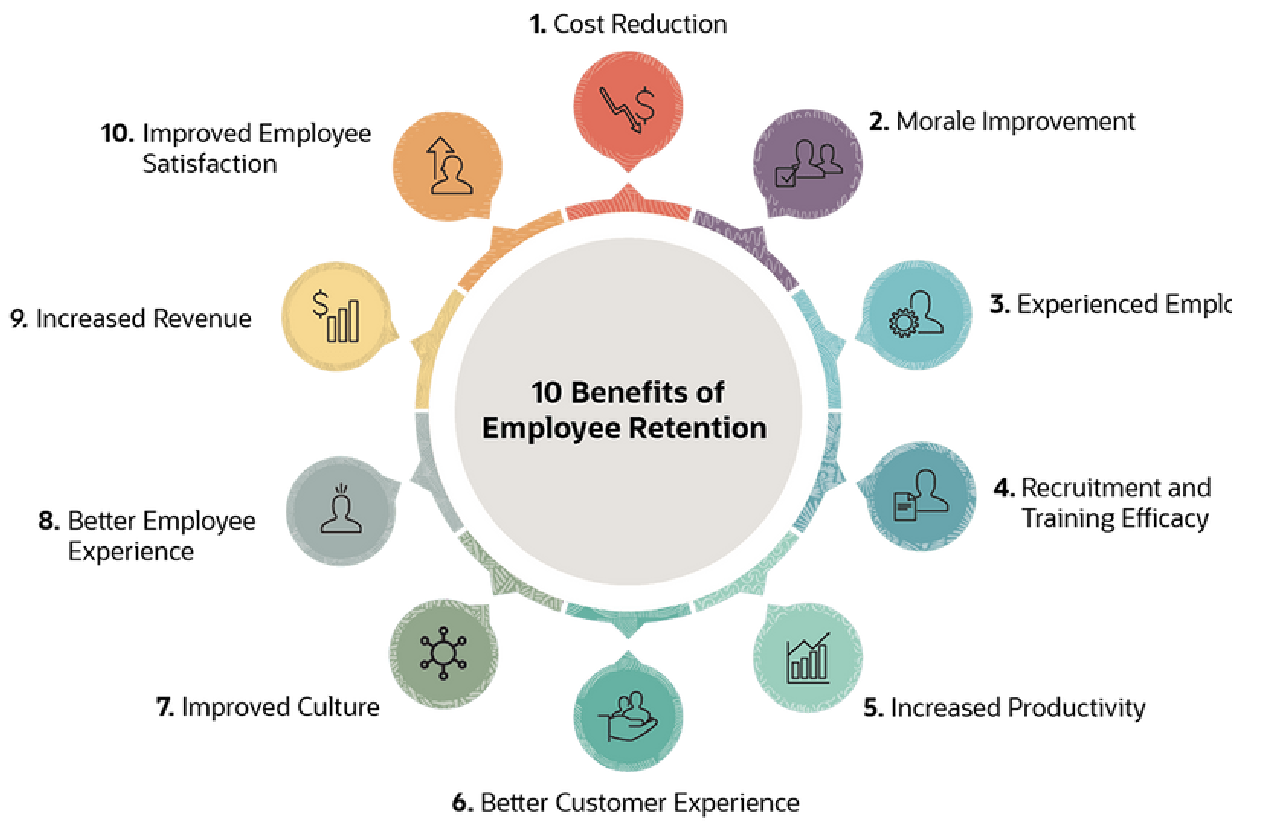
Business Objective

Maintaining employee satisfaction and retaining them is a challenge faced by companies since time immemorial

If an employee, into whom, a company has invested significant time and money leaves for "greener pastures", then this would mean that the company would have to spend additional time and money to hire another resource, train and bring them up to speed with the company culture and the day to day workings.

This will cause a monetary loss as well as a loss in efficiency, productivity, and revenue.

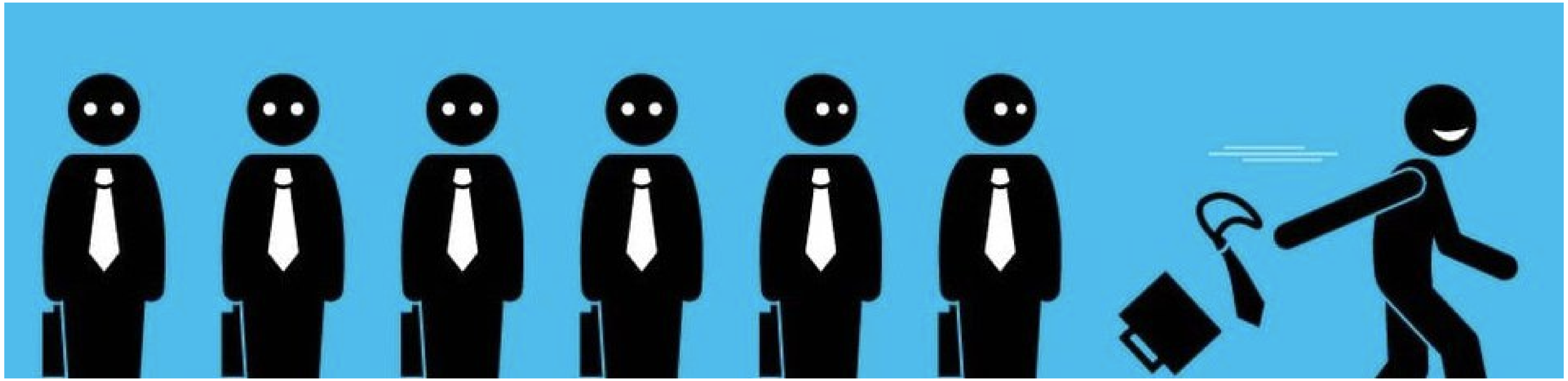
The objective is to minimize this loss by predicting the employees that will leave the company and identifying the factors that most greatly influence employee attrition.



Introduction

Goals:

- To determine which variables most greatly affect employee attrition using **supervised methods** - Support Vector Machines and Boosting Ensemble Learning Method.
- To model and predict which employee is likely to leave the company.



Introduction

Dataset Overview

1. The dataset is synthesized by IBM
2. Contains observations of 1470 employees and 30 variables.
3. 16.12% of the dataset contains observations of employees who have left the company.
4. The data is evenly distributed for all variable types, numerical and categorical (both nominal and ordinal)

Target Variable

Attrition is based on many variables in the dataset that contain information from demographic, to organizational data. (e.g. Age, Monthly Income, Job Satisfaction score, Distance from home, Marital Status, etc.)

Data Exploration

Dataset Quality

There are no missing values 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.

Variable Types

Numerical

1. Age
2. Distance From Home
3. Hourly Rate
4. Monthly Income
5. # Companies Worked at
6. Percent Salary Hike
7. Total Working Years
8. Training Times Last Year
9. Years At Company
10. Years In Current Role
11. Years Since Last Promotion
12. Years With Curr Manager

Categorical - Ordinal

1. Business Travel
2. Environment Satisfaction
3. Job Involvement
4. Job Level
5. Job Satisfaction
6. Performance Rating
7. Relationship Satisfaction
8. Stock Option Level
9. Work Life Balance
10. Education

Categorical - Nominal

1. Department
2. Gender
3. JobRole
4. Marital Status
5. Over Time
6. Education Field

General Methodology

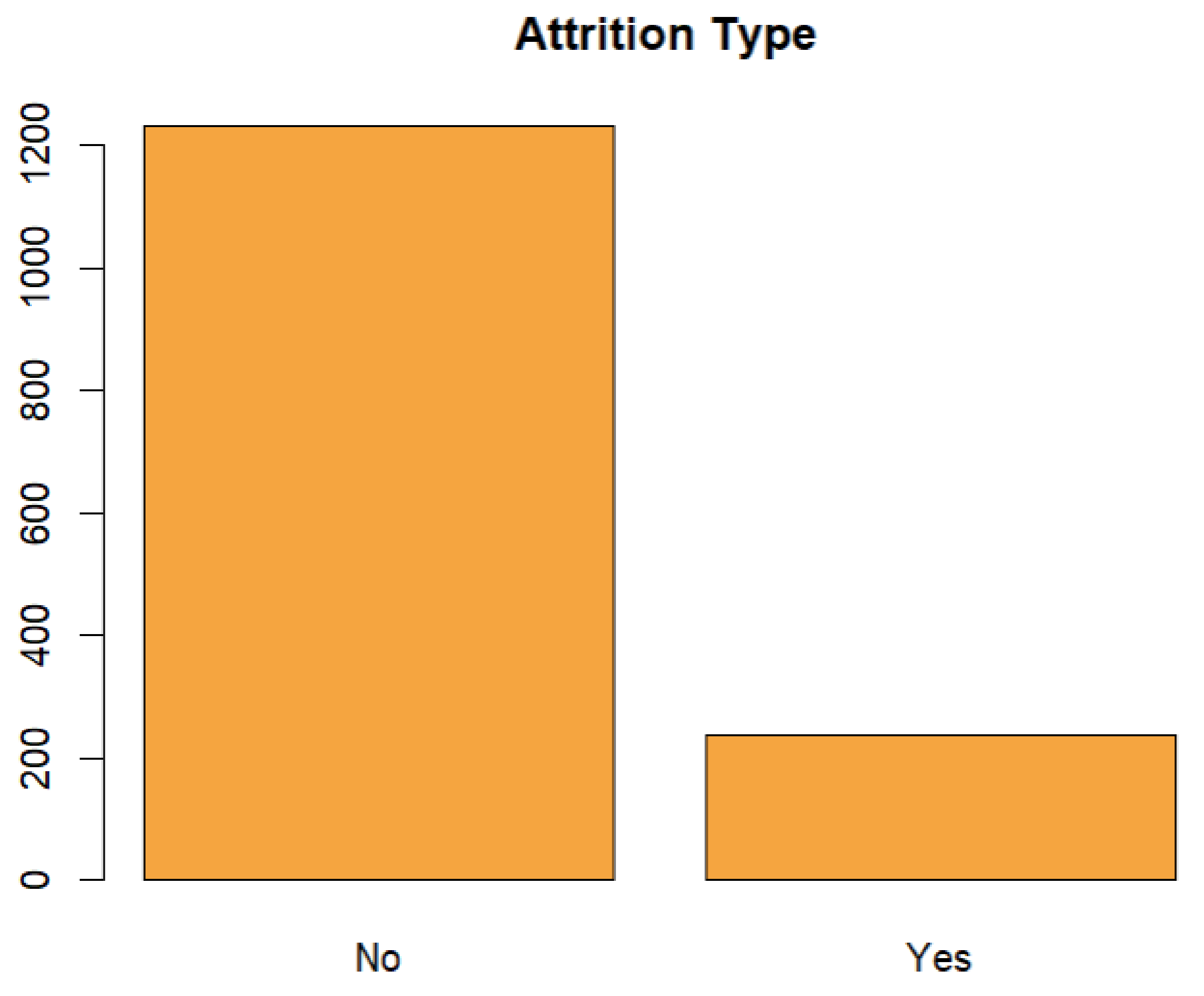
- Cleaning the data
- Exploratory Data Analysis
- Model Training
- Assessing Model Performance
- Validating the model

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.

Data Pre Processing

- Checking for **Missing Values** – No missing values
- **Outlier Detection** – No outliers. Checked using the Z-Score method
- **Class Imbalance** - Huge class imbalance exists in data and this can impact the model's ability to correctly predict the Attrition rate.
- **Weights** method used to treat class imbalance and assigned a higher weight to the minority class (Attrition = "Yes")
- Trained our model by using 75% of the data and the remaining 25% for testing



SVM - Findings

Test Model Performance

- A higher value of cost under **hyperparameter tuning** (**C = 4.63**) indicates that the model focuses on minimizing training errors
- **Accurately** predicts **89%** of the unseen data
- **Kappa Value** of **50%** which is a moderate agreement of collected data with the variables

Goodness of fit

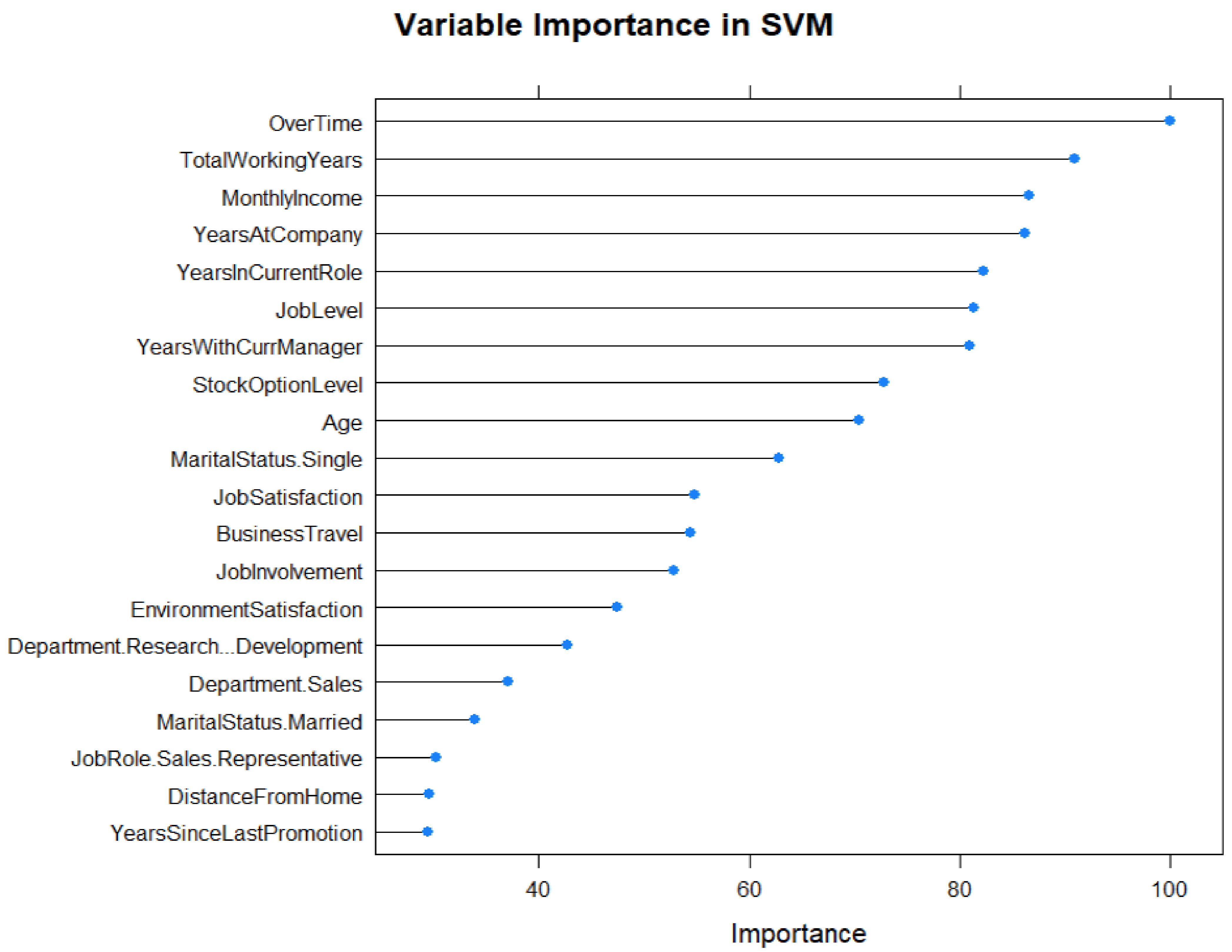
- It is a balanced model: Performs well on training as well as testing data. Close to **90% accuracy** on both training and testing data
- Has high **Specificity(97%)**. This means that the model correctly predicted employees who will not leave
- The probability of precisely predicting the **Positive Attrition rate(Yes) is 77%**
- The probability of precisely predicting the **Negative Attrition rate(No) is 90%**

Model Validation

- The model predicts attrition with a **high Specificity** which will lead to a minimal opportunity cost of misclassifying employees who will leave. This along with a very **high Negative Atttrition** prediction rate, validates our model for the business case

SVM - Variable Importance

As per Support Vector Method following are the 20 most **significant predictors** of Attrition:



Ensemble Method (Boosting)

- It is a prediction model that predicts the target class for all the observations in the data
- It is an iterative model that accounts for misclassification and imposes more weights on misclassified records in the next iteration

Data Pre Processing

- Checking for **Missing Values** – No missing values
- **Outlier Detection** – No outliers. Checked using the Z-Score method
- Class Imbalance - method takes care of the imbalance in the data, hence **no external treatment is required**
- Feature section - Used **wrapper method** to remove irrelevant variables
- Trained our model by using 75% of the data and the remaining 25% for testing

Ensemble Method - Findings

Test Model Performance

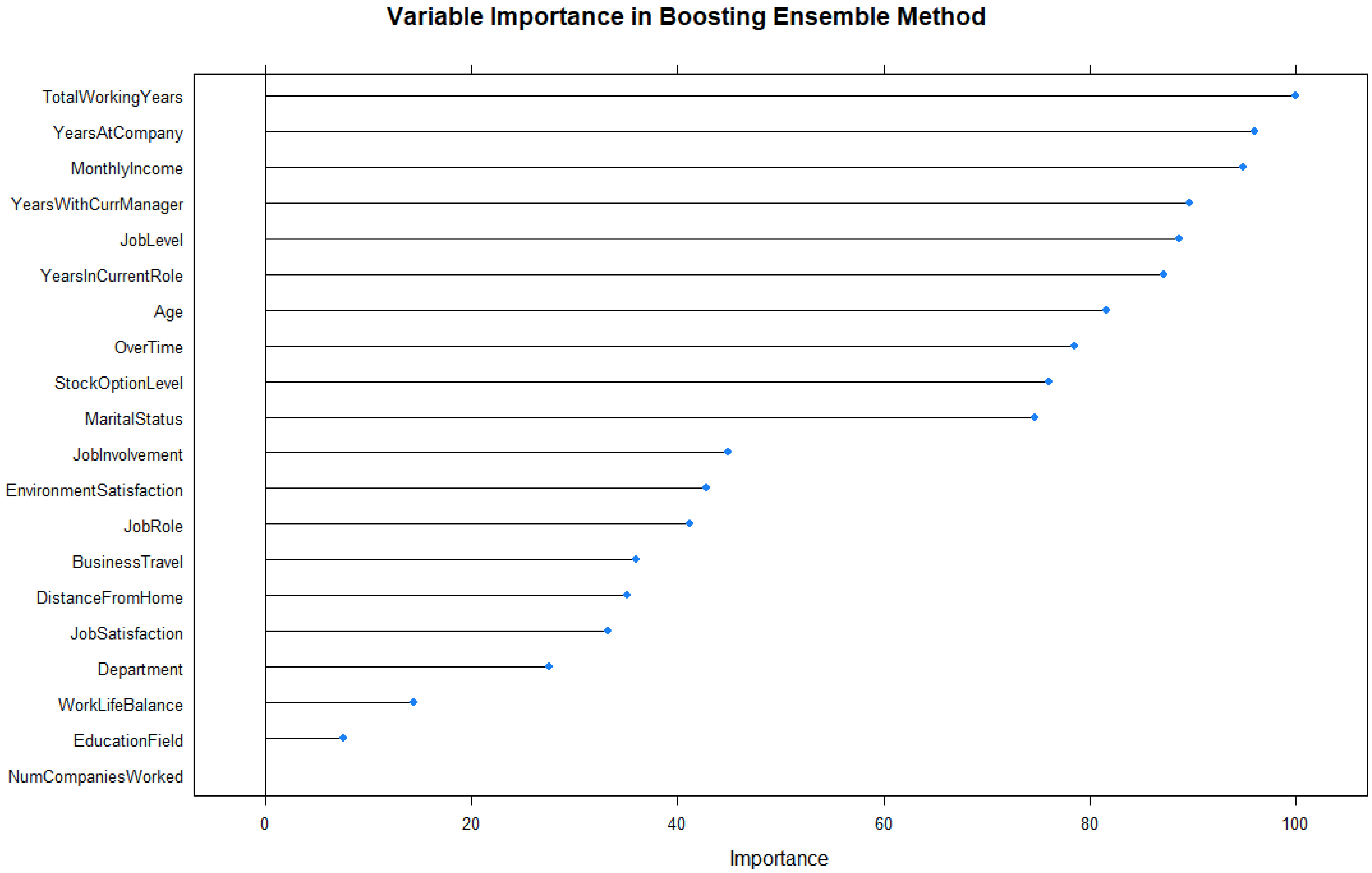
- **Accurately** predicts 89% of the unseen data
- **Kappa Value** of 36.4% which is a fair agreement of the collected data with the variables

Goodness of fit

- Has **high Specificity** (96%). This means that the model correctly predicts the employees who will not leave 96% of the time of all "No" predictions
- The probability of precisely predicting the **Positive Attrition rate** (Yes) is 66%
- The probability of precisely predicting the **Negative Attrition rate** (No) is 88%

Ensemble Method – Variable Importance

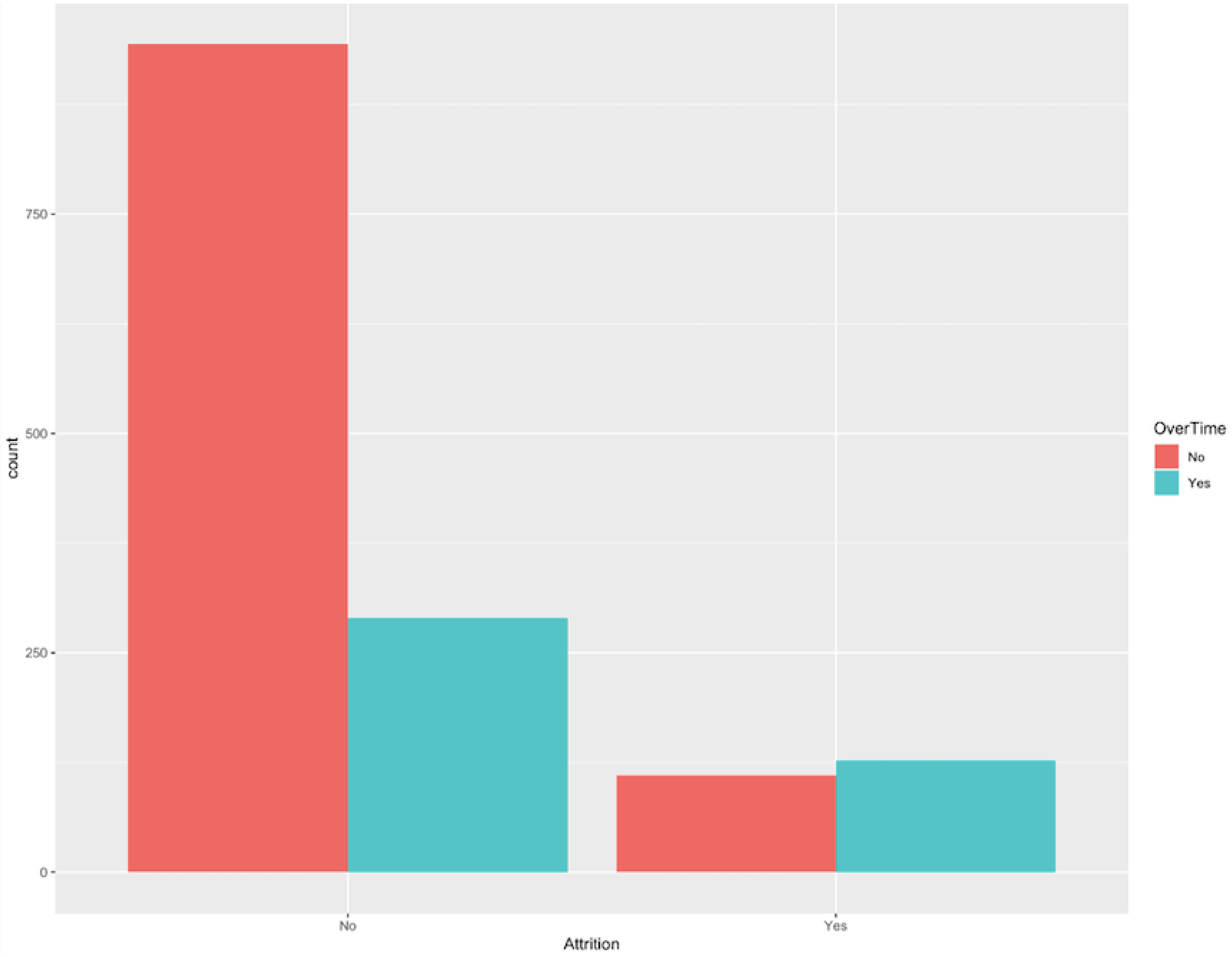
As per Boosting Ensemble method following are the 20 most **significant predictors** of Attrition:



Recommendation - Comparing Models

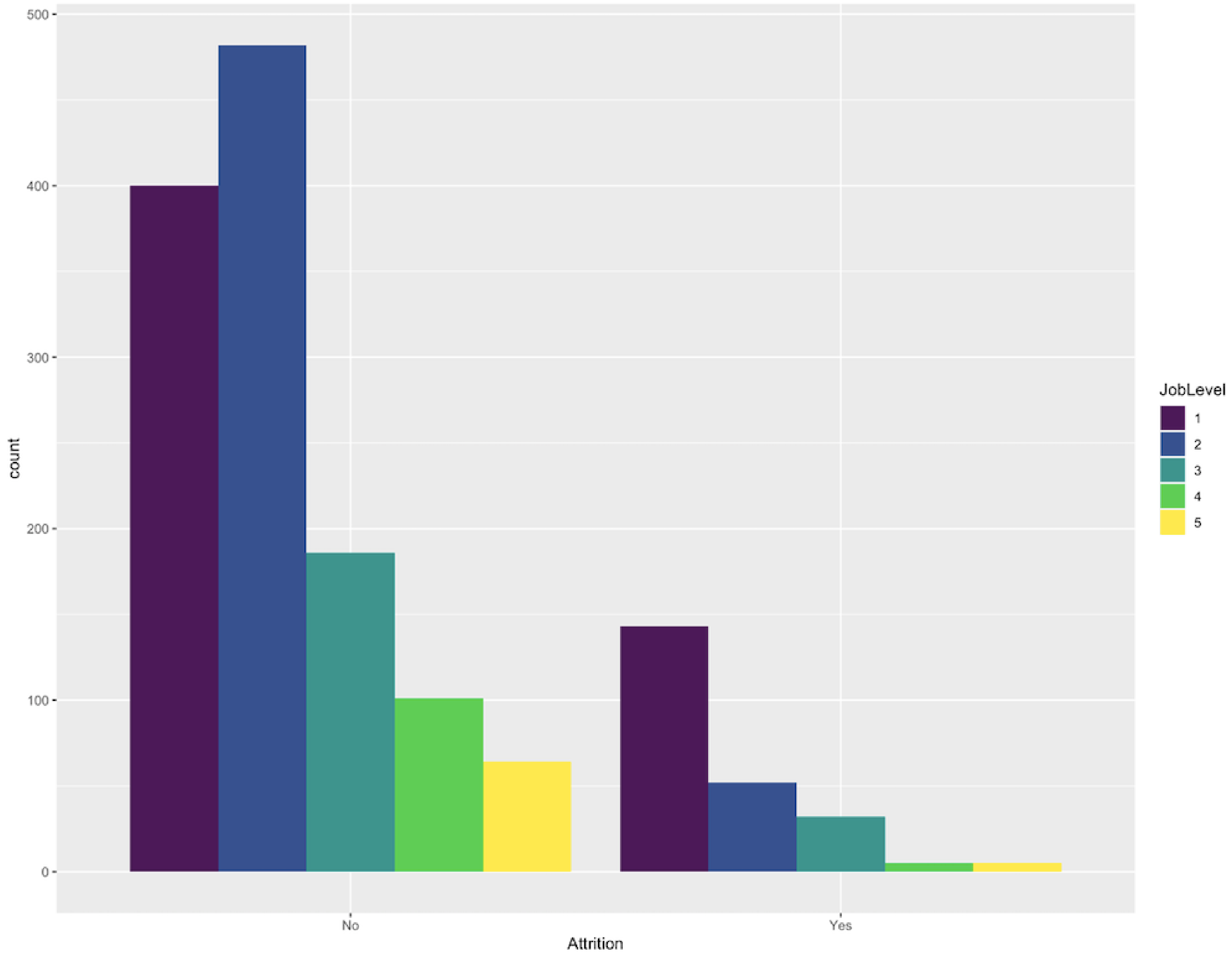
- Both SVM and Boosting methods give us models with high levels of accuracy, specificity, and positive/ negative prediction rates, which are the most relevant measures for our business case.
- We recommend that the SVM model be chosen for this business problem as the performance measures are higher, albeit slightly when compared to Boosting

Findings based on SVM



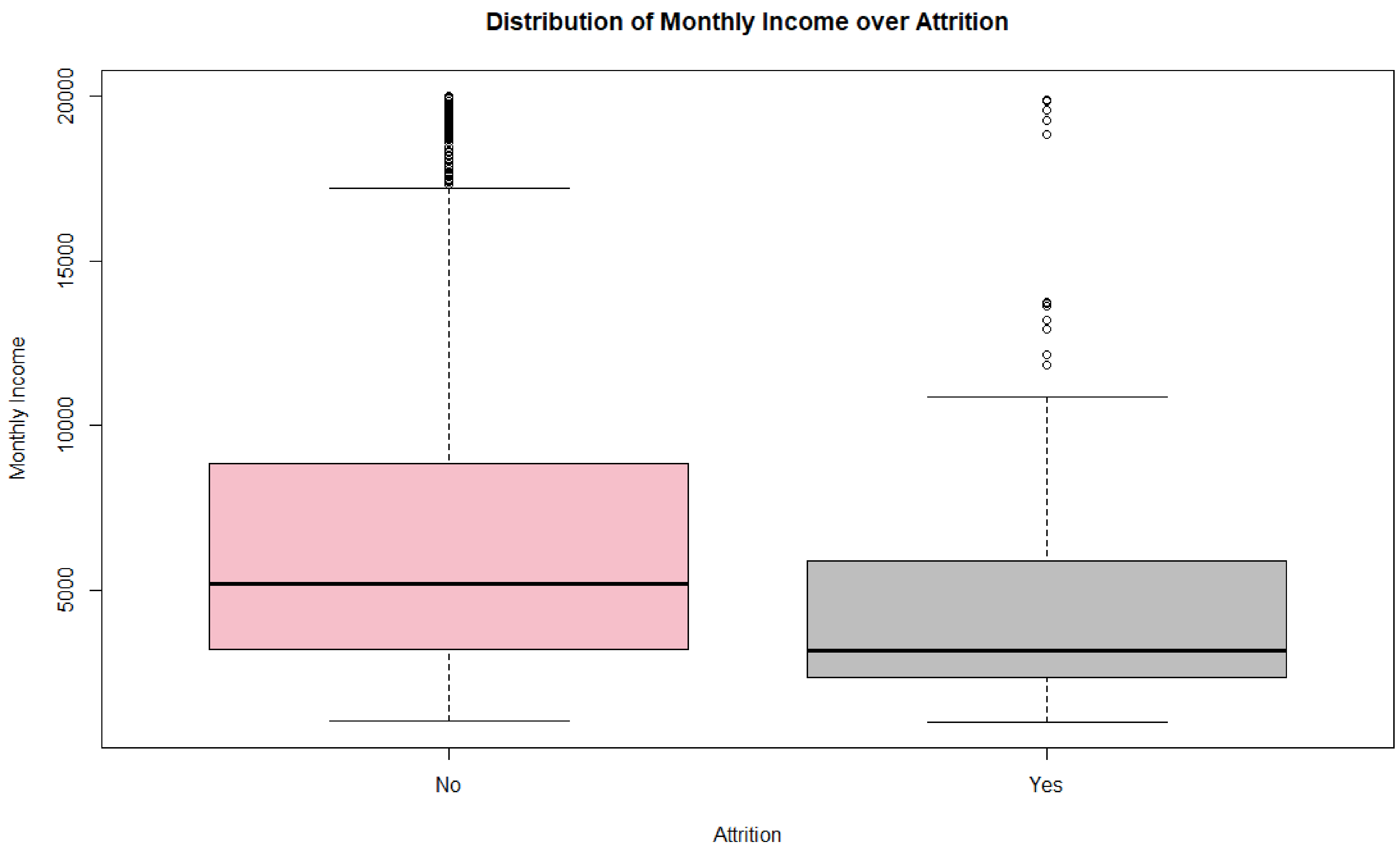
- Overtime is the most important nominal variable to predict attrition as per SVM model
- Overtime can help a few employees take a bigger paycheck home, however, if done excessively it can lead to burnout
- Must strike a balance here
- Overtime should be properly scheduled

Findings based on SVM



- 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

Findings based on SVM



- 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

Thank you!