Employee Attrition Analysis And Prediction

Course Project — DS-203: Programming for Data Science
Prof. Amit Sethi & Prof. Manjesh Hanawal

Avinash Gupta
20D110005
20d110005@iitb.ac.in

Meera Upasana
190110046
190110046@iitb.ac.in

Gautham Mallya
200110041
200110041@iitb.ac.in

Prakhar Patel
180020065
180020065@iitb.ac.in

Abstract—**Employee attrition is the process by which employees leave the workforce.There comes a time when an employee wants to leave a company – for personal or professional reasons. Employee retention is crucial for any company's success and the attrition rate is the metric that provides insight into how well the company is retaining their employees. By measuring and analyzing the attrition rate, we can identify the problems the firm needs to solve to ensure employee satisfaction as well as the firm's success. According to OMAM's report lower hike in salary and more percentage of variable pay leads to more attrition.**

We have used various classification models like Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbour to predict attrition of an employee based on multiple features like monthly income, overtime, distance from home, total working years etc.

The features that influence the attrition the most are overtime followed by marital status and the best performing model to predict the same are Logistic Regression and K-Nearest Neighbour.

I. Introduction

Companies invest a lot of human effort and financial resources in recruiting employees and training them according to their strategic needs. Therefore, the employees are a real investment for organizations. When an employee leaves the company, the organization is losing a valuable employee and the resources invested in recruiting and selecting those employees and training them for their roles. Consequently, the organization must continuously recruit, train, and prepare new staff to fill vacancies. It is in the company's interest to control and bring down the employee attrition rate: attrition is when an employee resigns or retires from a company. It is the group of satisfied, motivated, and loyal employees that form the core of an organization and impact the overall productivity of an organization. Companies aim to retain such employees due to the value they add to the company and their role in the firm's sustained growth. Job dissatisfaction data is a powerful indicator of whether an employee plans to resign and maybe switch companies.

In our report, we analyze the reasons or motivations that push an employee to leave the company and hence the adoption of timely appropriate countermeasures by the HR department. Starting from the dataset, we identify the main factors related to the employee's attrition and propose a classification based on the statistical evaluation of the data. The application of classification algorithms can help the HR department by allowing the company to adopt staff management support tools. We have prepared classification models based on a dataset by IBM analytics, which includes 35 features and about 1500 samples.

II. Background

According to OMAM consultants' report, companies of different sectors like automobile, banking, e-commerce, insurance, IT, etc., had attrition rates ranging between 7-19%in 2019. Their report suggests that the Insurance sector had the highest attrition rate at 19%, followed by the Retail Banking sectors, both at 18%, while the Automobile sector had the lowest attrition rate at 7%. The report also shows that sectors with higher variable pay and lower salary hike had more attrition rates. But, the report only shows the dependence of attrition rate on these two features. In our Problem Statement, we plan to take this one step ahead and try to find out which features or reasons the attrition rate of a company might depend upon. Our feature set includes various features like the employee's department, job satisfaction, monthly income, salary hike percentage, job role, daily rate, etc. We aim to find top features that contribute the most towards attrition.

III. Methodology

- Collect the employee data set, which consists of employee observations
- Apply data cleaning techniques to prepare the data set, like checking for null entries, eliminating highly correlated features
- Perform a descriptive analysis of data to detect the key factors and trends that contribute to attrition

- Elaborate the data set for the training and testing phase and try several classification algorithms to process it
- Based on the results collected, compare many performance metrics of machine learning models and conclude which model gives the most accurate results for the given problem and propose an HR support solution that implements the classification model

IV. Experiments and Results

A. Data Description

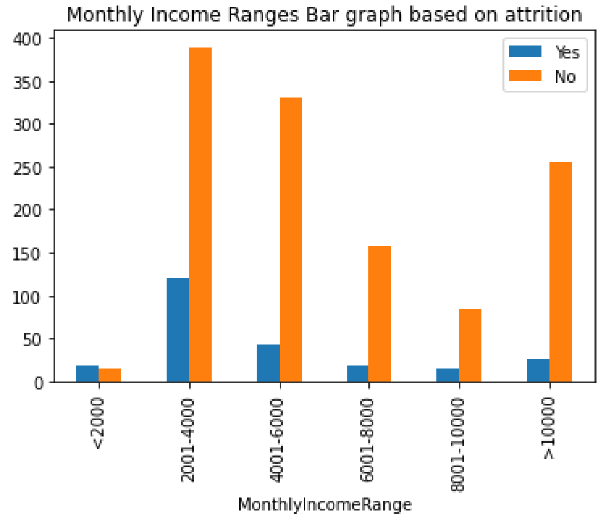
The data set used has following features:

- Age
- Monthly income
- Attrition
- Monthly rate
- Business travel
- Number of previous employers
- Daily rate
- Over 18
- Department
- Overtime
- Distance from home
- Per cent salary hike
- Education
- Performance rating
- Education field
- Relations satisfaction
- Employee count
- Standard hours
- Employee number
- Stock option level
- Environment satisfaction
- Total working years
- Gender
- Training times last year
- Hourly rate
- Work-life balance
- Job involvement
- Years with company
- Job level
- Years in current role
- Job role
- Years since last promotion
- Job satisfaction
- Years with current manager
- Marital status

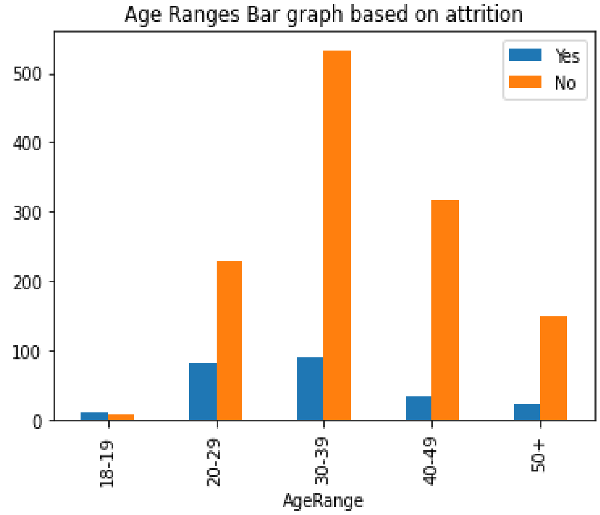
B. Exploratory Data Analysis

We started with the missing value treatment, and then removal of highly correlated features for model building. The features that have good correlation between them(mod>0.75) are following:
Monthly Income, Performance Rating, Total Working Years, Years in Current Role and Years with Current Manager.

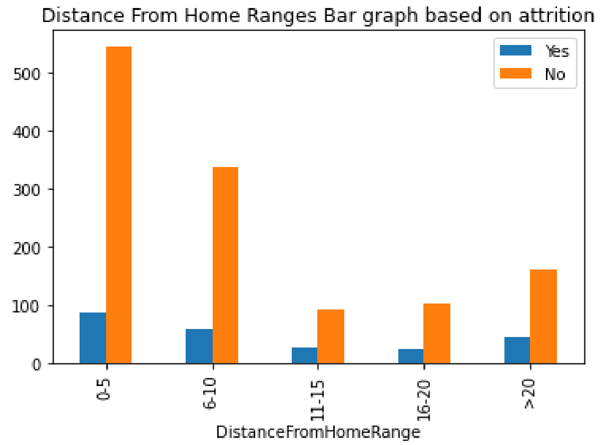
The descriptive analysis of data set characteristics was conducted by relating each feature to the target variable “Attrition”. Following are plots for the same:



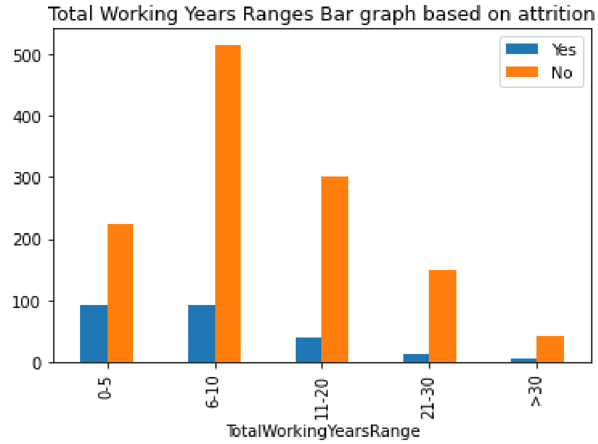
Attrition is relatively higher among lower income brackets, possibly because of employees looking for a higher salary at other firms.



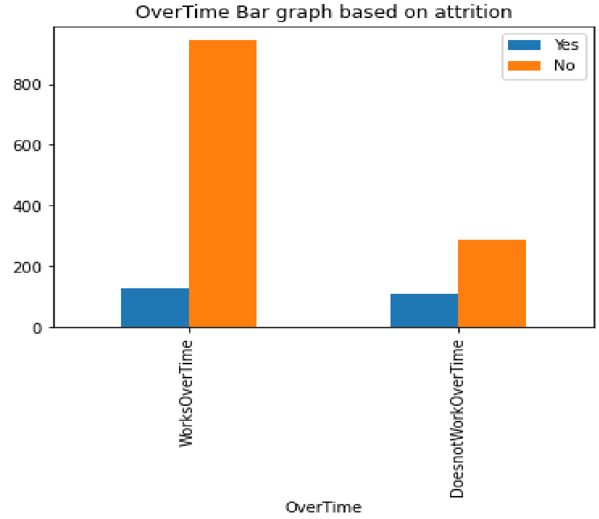
Young employees show higher tendency to leave the firm, chasing better career prospects and opportunities, as they are at a stage where they can afford to take risks. The older employees prefer job security.



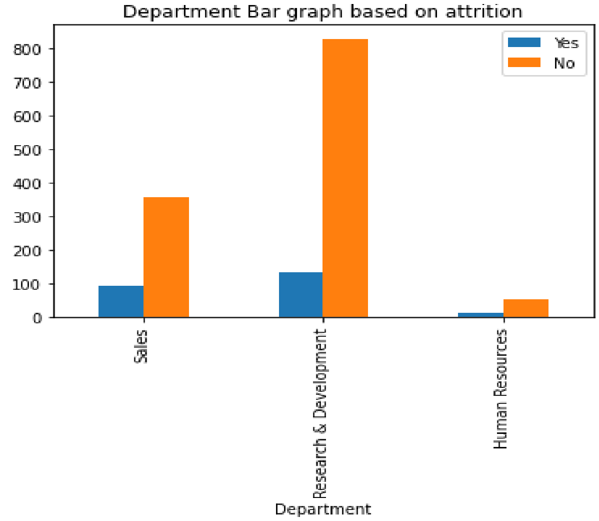
Distance from Home factor does not give many concrete inferences about attrition



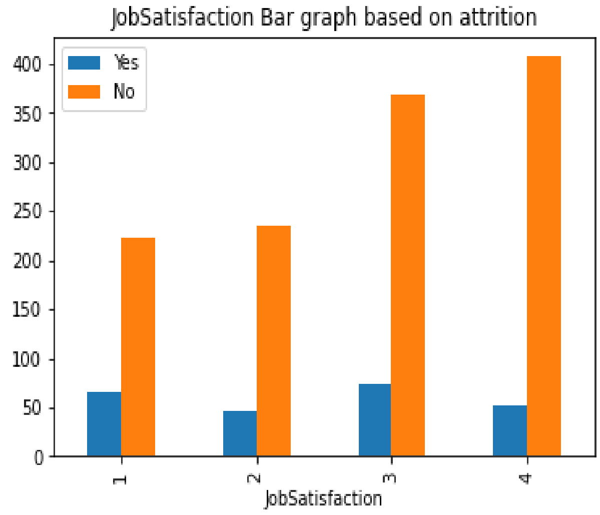
The graph reveals that employees in the early stages of their career are more likely to switch companies than experienced employees.



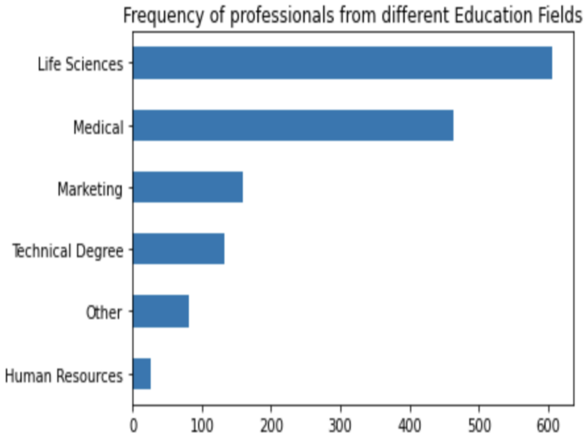
Among those working overtime, only a relatively small employees have shown tendency to leave the company, however, a significant number of those who do not, have left the company.



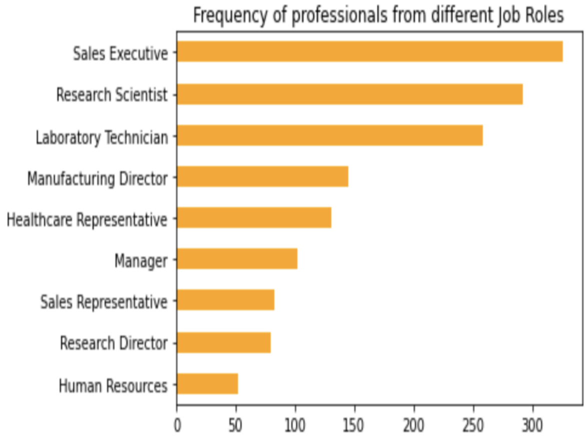
The R&D Department and the HR department see higher retention in employee number than the Sales department, possibly because of the difference in competitiveness in the different sectors



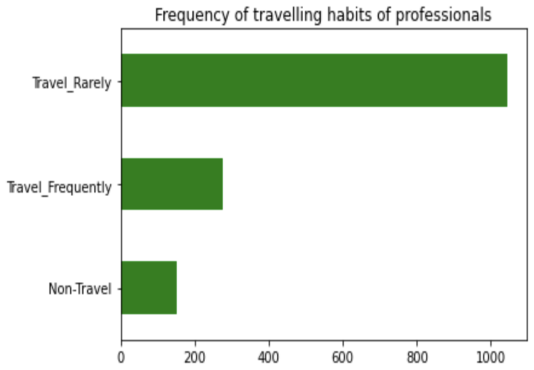
The number of employees from different Education Fields is shown in the following graph:



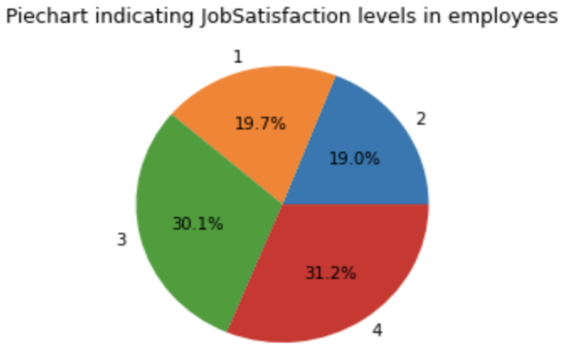
A majority of the employees in the company have majored in Life Sciences and Medical Fields. The number of employees in different Job Roles is shown in the following graph:



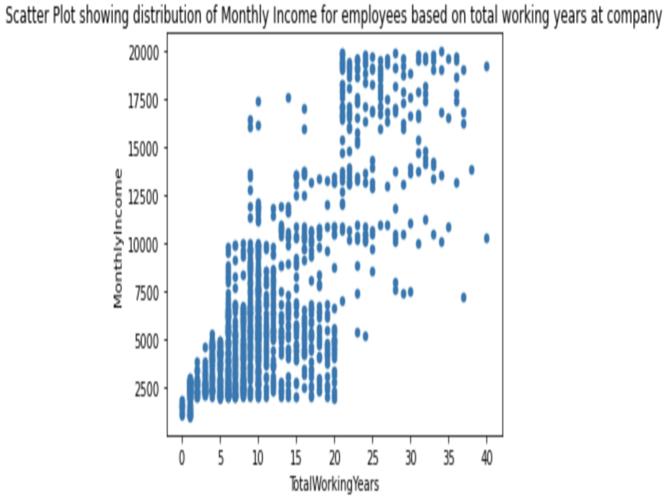
Employees in the company are required to travel according to the needs of their roles. The information regarding Travel Habits is captured in the next graph:



From the dataset, we can infer the distribution of job satisfaction of employees classified into four categories as follows:



For an idea about how much employees earn with increasing years of service we have made a scatter plot:



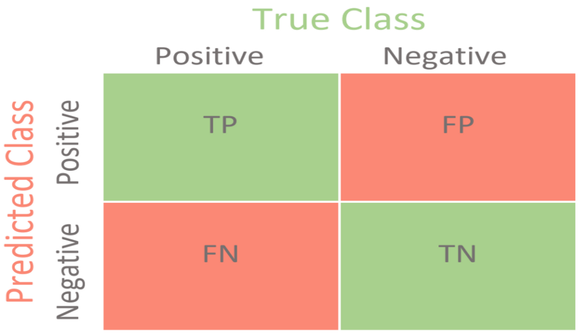
The plot is typical of any company: Density of points is higher where employees with less experience and the corresponding monthly income is lower. Sparsity is observed among employees with great amount of experience and they occupy the higher end of the income spectrum.

C. Models Used

1) **Logistic Regression** is probably the most commonly used supervised learning classification method and the results obtained here from the Logistic Classifier have been used as a benchmark for more complex models. We have used the saga solver, an extension of the stochastic average gradient descent approach, which uses a random sample of previous gradient values and allows for L1 regularization. A 4-fold cross validation procedure has been used to improve the estimated performance of the machine learning model and reduce the standard error associated with the results.
2) **The Random Forest Classifier** consists of an ensemble of decision trees (DT) where each DT is a branching structure that represents a set of rules, distinguishing values in a hierarchical form. Depth of each DT and the number of such DTs to be used in our ensemble were the hyperparameters which were selected after running a gridsearch on the training set for each ensemble.
3) **KNN** is a classification algorithm. It uses the k value and distance metric(Euclidean distance) to measure the distance of new points to nearest neighbors.The smaller the k value the greater the noise with the data; however, we can smooth this out by increasing the value of k.K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
4) **Decision Trees** (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

Confusion Matrix: A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if there are an unequal number of observations in each class or if there are more than two classes in one's dataset. Calculating a confusion matrix can give you a better idea of what one's classification model is getting right and what types of errors it is making.

Below is the process for calculating a Confusion Matrix. We need a test dataset with expected outcome values. Make a prediction for each row in your test dataset. From the expected outcomes and predictions count: The number of correct predictions for each class. The number of incorrect predictions for each class, organized by the class that was predicted. These numbers are then organized into a table, or a matrix as follows:



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

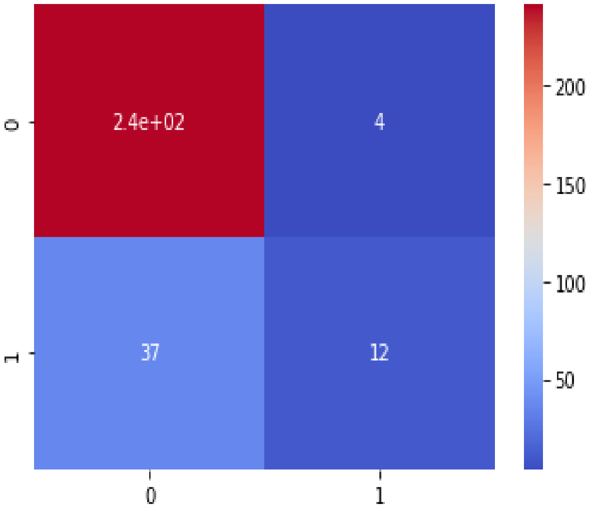
Expected down the side: Each row of the matrix corresponds to a predicted class. Predicted across the top: Each column of the matrix corresponds to an actual class. The counts of correct and incorrect classification are then filled into the table.

The total number of correct predictions for a class go into the expected row for that class value and the predicted column for that class value.

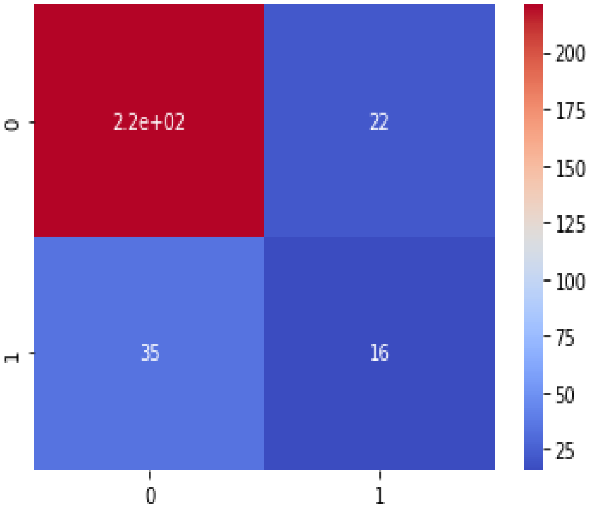
In the same way, the total number of incorrect predictions for a class go into the expected row for that class value and the predicted column for that class value.

Following are the confusion matrices for Logistic regression, Decision Tree and Random Forest respectively, and a plot of accuracy score vs number of neighbors for the KNN Model.

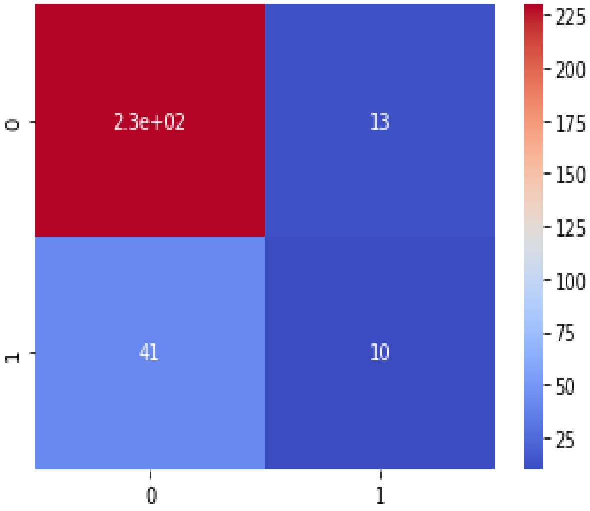
- Logistic Regression



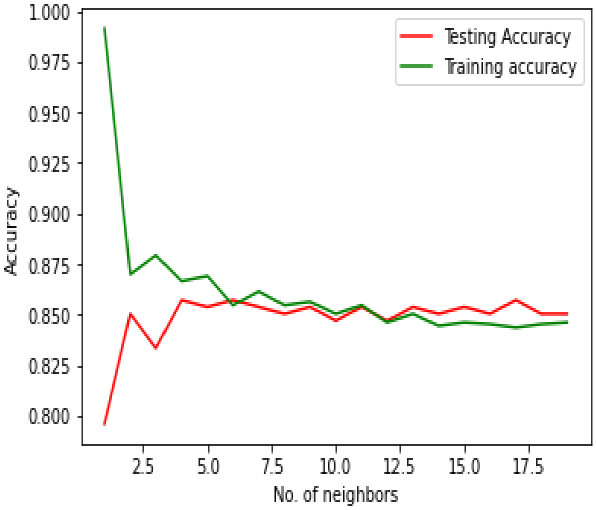
- Decision Tree



- Random Forest



- K-Nearest Neighbours



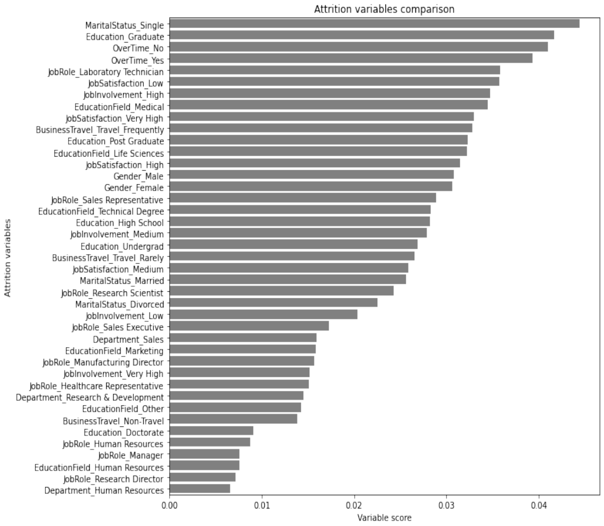
These trained models are then tested against the test data.

D. Model Scores

The different scores for each model have been tabulated as follows:

Model Implemented	Accuracy Score	Precision Score	Recall	F1 score
Logistic Regression	0.8605	0.75	0.2449	0.8295
Decision Tree	0.8601	0.4211	0.3137	0.7945
Random Forest	0.8163	0.4348	0.1961	0.7866
K-Nearest Neighbor	0.8571	1.0	0.1064	0.8050

V. Learning & Conclusions



With respect to the results we came up with the following conclusions:

1.From variable score plot we observe **overtime** variable has maximum influence on attrition followed by **marital status**
2.Single employees have more attrition rate than married employees
3.Graduate employees have greater attrition rate as compared to post-graduates

This information is of utmost importance to firms as it will help them in introducing specific reforms considering the above analysis and retain valuable employees and thus boost firm's growth.

We also conclude that according to accuracy scores every model performs similar with marginal differences in predicting whether an employee will leave the firm or not. By numerical values of F1 scores, **Logistic Regression** and **K-Nearest Neighbour** models are best for predicting.

VI. Contribution of Team Members

Meera Upasana - Exploratory data analysis using data visualization & data cleaning techniques
Avinash Gupta - Employee Attrition prediction models & model score analysis
Gautham Mallya - Machine learning model optimization & report documentation
Prakhar Patel - Project methodology, analysis & report documentation

References

- IBM HR Analytics Employee Attrition Dataset
- Predicting Employee Attrition Using ML Techniques
- OMAM-Increment-Trend-India-2019
- Employee Satisfaction Index
- Confusion Matrices in Machine Learning