CAPSTONE PROJECT: HR ANALYTICS

Table of Contents

Serial No.	Content	Page No.
1.1	Introduction, Data Report	3
1.2	Exploratory Data Analysis (EDA)	4 - 7
1.3	Insights from EDA	7 - 8
2.1	Model Building and Interpretation	9 - 20
2.2	Interpretation of the best model	20
2.3	Model Tuning	20 - 23
2.4	Final Interpretation of the best model	23
2.5	Conclusion and Summary	23 - 24

Introduction

- A fast growing start-up has been experiencing high attrition rates since the past three years. Instead of seeking reactive measures, they require predictive information with respect to the attrition of employees using the data they have available.
- The need of the project is to concretely arrive at the discerning factors that could potentially motivate an employee to leave the company. The company needs this study to aid them in solving their attrition problem in a proactive manner.
- This is a business opportunity for the company as it would help them to reduce the overall hiring and attrition costs which the start-up has been incurring, and forging proactive HR policies to retain employees on a long term basis.

Data Report

- The data available to us consists of all the relevant employee information corresponding to their respective rate of attrition (0 for no and 1 for yes). The data available has been recorded over a period of 3 years. No particular frequency is seen across the data as it is simply recorded for every employee the company has had since the past 3 years. The data is secondary in nature, made available by the company through their organizational records for each employee.
- Visual inspection of the data gives us **20 variables (in columns) and 5880 observations (in rows)**. A detailed summary in R gives us more precise information for each and every variable w.r.t. its mean, median, range etc. The description of each variable is available below-

Variable	Description
EmployeeID	Unique employee code
Attrition	Attrition flag
Age	Age of employee
TravelProfile	Status of travel in job profile
Department	Department of employee
HomeToWork	Distance between home to work
EducationField	Field of education of an employee
Gender	Gender of an employee
HourlnWeek	Work hours of an employee in a week
Involvement	Involvement of any employee in engagement activity organised by HR team. 5 highest | 1 Lowest
Designation	Employee designation
JobSatisfaction	Score of employee opinion survey. 5 highest | 1 Lowest
MaritalStatus	Marital status of employee
MonthlyIncome	Gross monthly income of employee
NumCompaniesWorked	Total number of company employee had worked in past
OverTime	Is employee is eligible to be paid for overtime
SalaryHikelastYear	Increment percent in last cycle
WorkExperience	Total year of work experience
LastPromotion	Year since last promotion
CurrentProfile	Year since in current profile

Variables	Type
EmployeeID	Numeric
Attrition	Numeric
Age	Numeric
TravelProfile	Character
Department	Character
HomeToWork	Numeric
EducationField	Character
Gender	Character
HourlnWeek	Numeric
Involvement	Numeric
Designation	Character
JobSatisfaction	Numeric
MaritalStatus	Character
MonthlyIncome	Numeric
NumCompaniesWorked	Numeric
OverTime	Numeric
SalaryHikelastYear	Numeric
WorkExperience	Numeric
LastPromotion	Numeric
CurrentProfile	Numeric

Exploratory Data Analysis

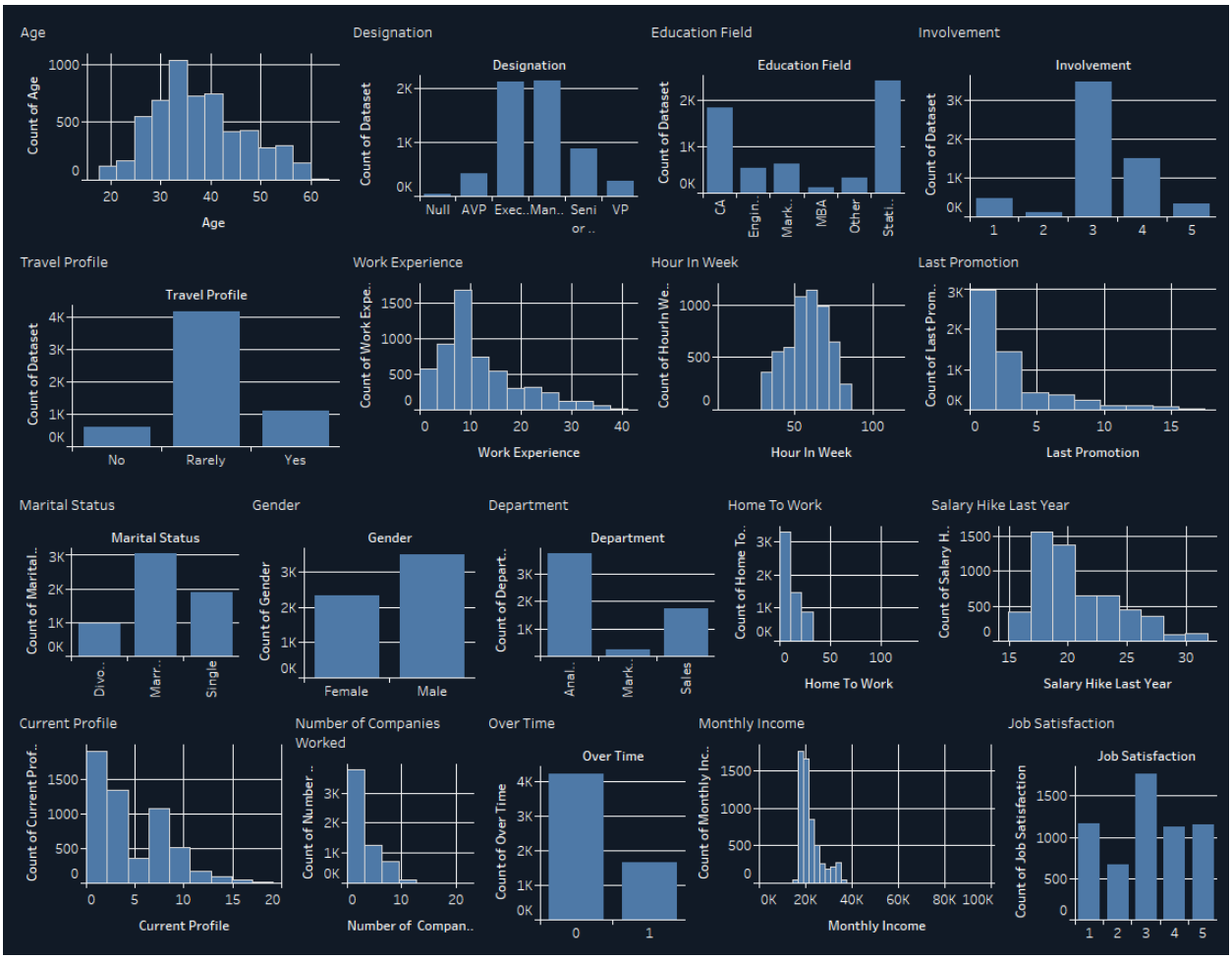
Variable Transformation

Out of the 20 variables, we will convert the following variables for further analysis.

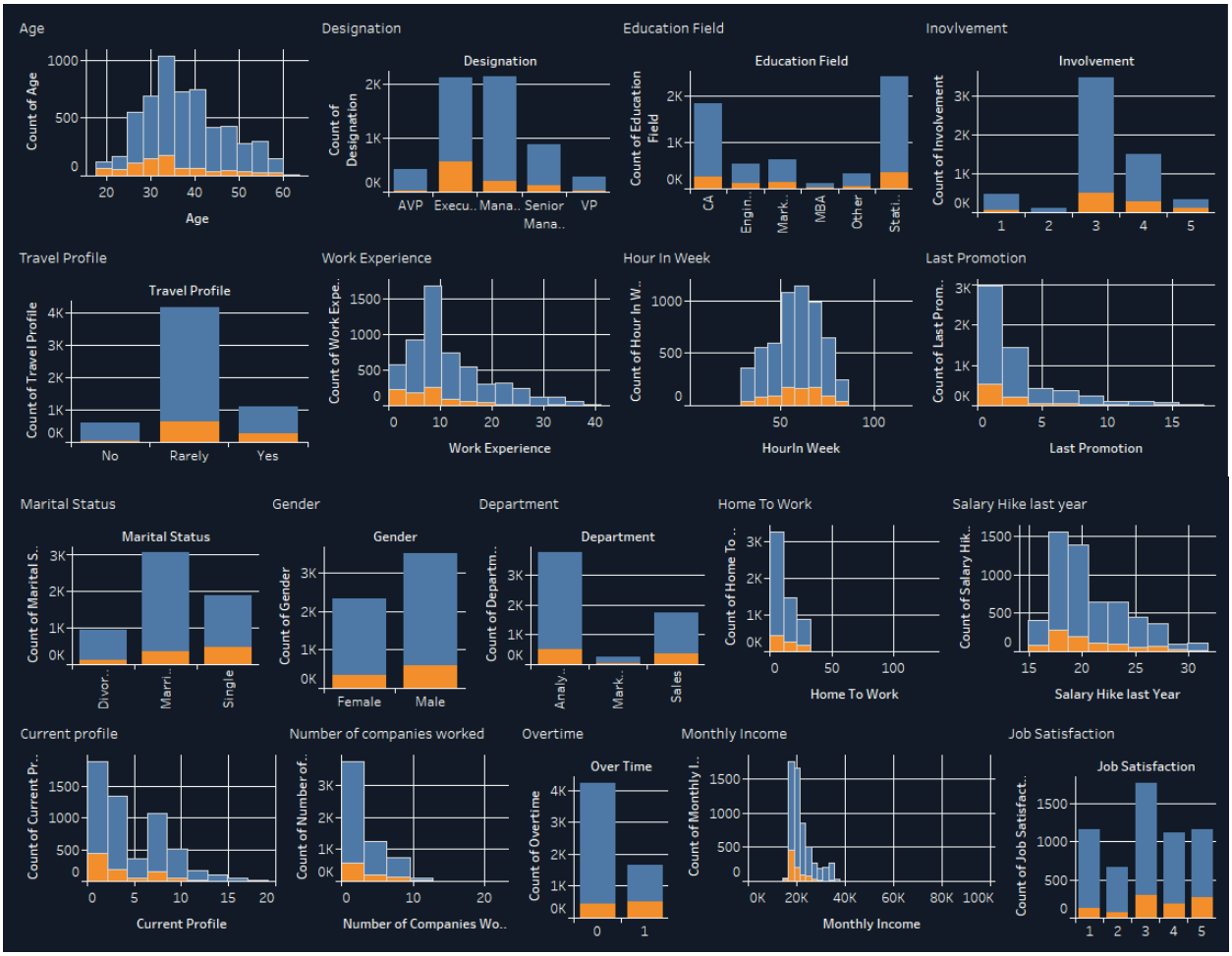
Variables	Type	Coversion to
Attrition	Numeric	Factor(with 2 levels:0,1)
TravelProfile	Character	Factor (with 3 levels: Yes, No, Rarely)
Department	Character	Factor (with 3 levels: Sales, Analytics, Marketing)
EducationField	Character	Factor (with 6 levels: CA, Engineer, Marketing Diploma, MBA, Statistics and Others)
Gender	Character	Factor (with 2 levels: Male, Female)
Involvement	Numeric	Ordered factor (with 5 levels where 1 is lowest and 5 is highest)
Designation	Character	Factor (with 5 levels: AVP, Executive, Manager, Senior Manager, VP)
JobSatisfaction	Numeric	Ordered factor (with 5 levels where 1 is lowest and 5 is highest)
MaritalStatus	Character	Factor (with 3 levels: Single, Married, Divorced)*
OverTime	Numeric	Factor (with 2 levels: 0,1)

* There were 4 distinct levels under *MaritalStatus* namely, 'Single', 'M', 'Married' and 'Divorced'. 'M' has been recoded as 'Married'. Similarly, for *Gender, 'F' has been recoded as 'Female'.*

Univariate Analysis



Bivariate Analysis

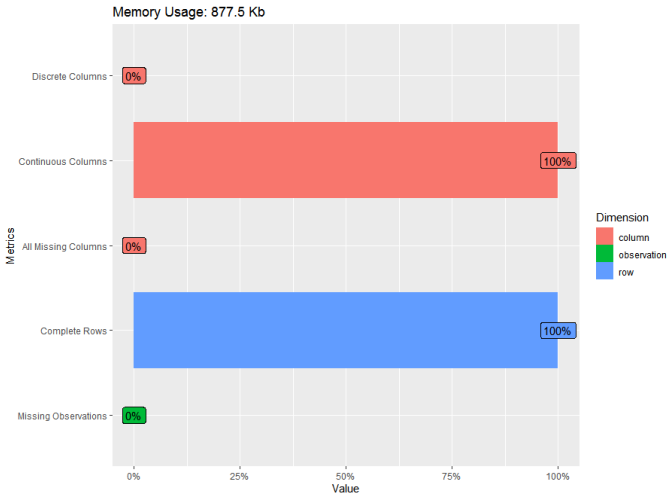


Removal of Unwanted Variables

- Variable *EmployeeID* is removed as it would not aid in any kind of analysis.

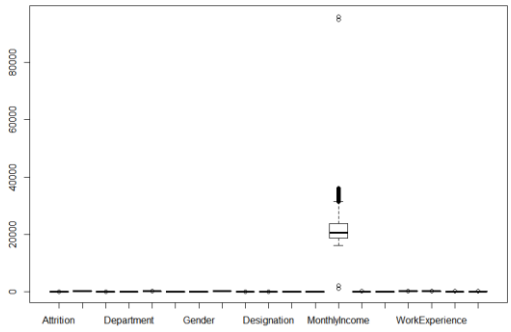
Missing Value Treatment

- Treatment of missing values is required for better insights and overall increase in performance of predictive models.
- Missing values are replaced with the median of the corresponding column using a loop function. After treatment, there should be no missing values.
- This treatment is generally advantageous given there is no loss of data, or skewing of data.

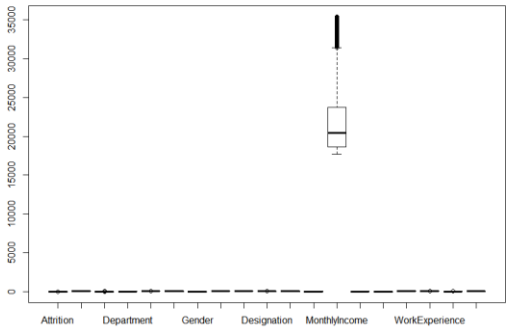


Outlier Treatment

- Outlier treatment is done using the squish function.
- The outliers in the dataset are treated by replacing the observations lesser than the value of the 1st percentile with value of the 1st percentile and the observations more than the 99th percentile with the value of the 99th percentile. The outlier treatment is done for every column in the dataset.
- This type of treatment is by far the most logical given it simply re-arranges the outliers identified by sorting them to their nearest quantile (here, the two extremes). Presence of outliers is normal and hence not removed as it will hamper the accuracy of the data.

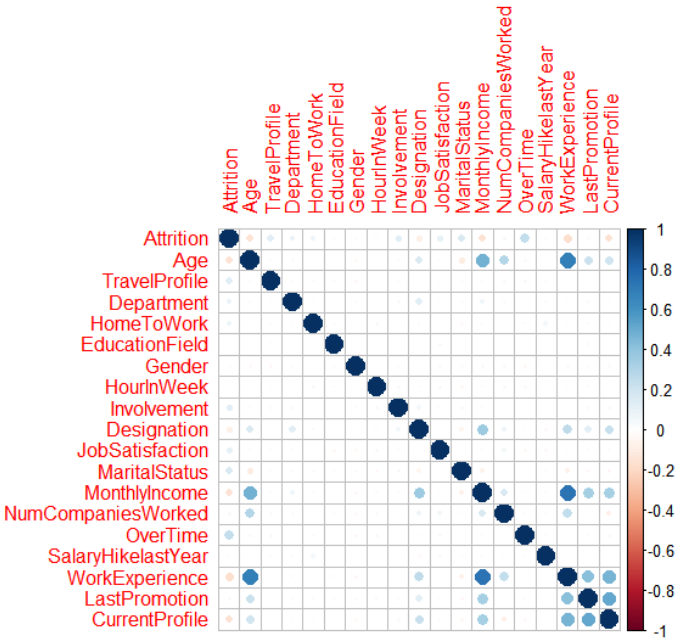


Outliers Before Treatment



Outliers After Treatment

Correlation Plot



Insights from Exploratory Data Analysis

- The data is highly imbalanced with the distribution of attrition approximately at 84% (for employees that didn't quit) and 16% (for employees that did quit). A highly imbalanced data will not give accurate results and hence **SMOTE (Synthetic Minority Oversampling Technique) is used to balance the data in 50-50 ratio.**
- Majority proportion of employees fall between the age groups of 25 to 40 and the average age bracket of employees who quit is roughly between 25 to 35 years.
- Most employees rarely travel and are seen to have a slightly higher attrition.
- A large proportion of employees are from Analytics and sales department-with Analytics recording the highest attrition rate closely followed by Sales, even though the number of employees in the sales department is less than half of the total .
- Majority proportions of employees have educational backgrounds in CA and Statistics and in turn also have a higher rate of attrition.
- Maximum employees contribute 50 to 70 working hours in a week and the employees within this range have a higher rate of attrition.
- Most employees score an average involvement score in terms of engagement activities and the ones who have an average or higher score are seen to have a slightly higher attrition rate.
- Most employees form the designations of Executives and Managers, and Executives are seen to have the highest rates of attrition.

- Most employees have recorded an average level of job satisfaction and those that recorded a higher level have a slightly higher rate of attrition.
- Over 3000 employees are married and Single employees are seen to have a higher attrition.
- Most employees with work experience roughly between 2 to 10 years have a recorded a higher attrition, although there is a significant presence of outliers in attrition rate w.r.t. work experience.
- Most employees were promoted within 0-5 years since last promotion and the highest rate of attrition is seen among those promoted within 0 to 4 years. While maximum numbers of employees have been in their current profiles between 0 to 10 years. And, higher rates of attrition are also observed within this bracket.
- Attrition rate is higher for employees with a monthly income of 15000 and 30000 per month, as most employees fall under this salary bracket.
- Most employees live within 10 kilometers of distance from work. Although higher rate of attrition are also observed within this group itself owing to its large sample.
- Employees who have worked with lesser companies i.e. between 0 to 5 companies have a higher rate of attrition because most employees fall under this range.
- Most employees are observed not doing overtime, less than half of the total employees are recorded doing overtime. A slightly higher rate of attrition is observed among employees who do overtime compared to those who do not.
- Majority of employees received a salary hike of 17 to 20 percent since last year and the employees receiving a hike within this bracket are seen to have a relatively higher attrition rate.
- High correlation is seen between:
    - Work Experience and Age
    - Monthly Income and Age
    - Monthly Income and Work Experience
    - Current Profile and Last Promotion
    - Current Profile and Work Experience
    - Last Promotion and Work Experience
    - Monthly Income and Current Profile
    - Monthly Income and Last Promotion
- No significant multicollinearity is observed between predictor variables.

Model Building and Interpretation

- In order to further explore the relationship between the response and the predictor variables, multiple models will be built and tested. This section will aim to understand which model is most suitable in determining the rate of attrition among the employees.
- This is a classifier problem and as such will be analyzed with the classification models on both train and test data.

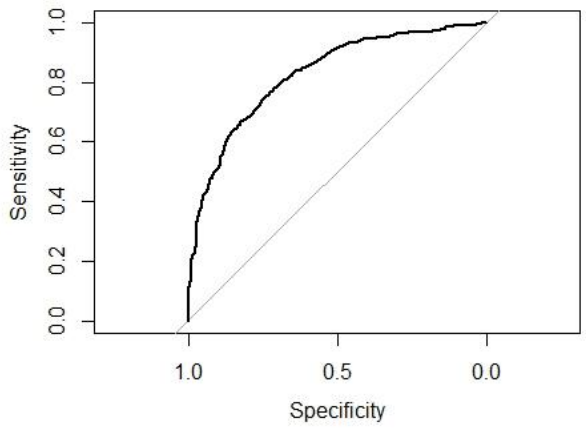
Model Validation

- Additionally, four metrics will be used across models to test the accuracy of the output, namely Accuracy, Sensitivity, Specificity and ROC (Received Operating Characteristic) and AUC (Area under the curve of ROC curve). The first three are obtained directly through the Confusion Matrix.
- **A confusion Matrix** is an error matrix that is displayed in a tabular form, showcasing the performance of an algorithm. It distributes True Negatives (TN), True Positives (TP) (Example: Correct number of predictions of the employees that attrite-this should be on the higher side), False Positives (FP) (Type I error) and False Negatives (FN) (Type II error). The lower these errors, the higher the accuracy of the model. In order to gauge this, we will look at the **Sensitivity (True Positive rate) and Specificity (True Negative rate)** of each confusion matrix. Higher distribution of both these metrics will yield the best results. **Accuracy** is obtained by a simple formula- **Accuracy = (TP + TN)/ (TP + TN + FP + FN)**
- **ROC curve** essentially displays the diagnostic ability of a classifier model against various thresholds. ROC curves helps to see how predictive models can distinguish between true positives and negatives. **AUC calculates** the area under the ROC curve and generally lies between 0 and 1. Higher the AUC, more accurate the model.
- **The most common classification models are selected to analyze the problem, including a few models based on multiple weak learner models; these are otherwise known as Ensemble models. Some of these ensemble models have been further tuned to give the best results.**

Model Building

1. Logistic Regression

- Logistic regression is widely known to be used when the Target variable is categorical, like for example *Attrition* is recorded as 0 and 1 as corresponding to 'employees that did not leave the company' and 'employees that left the company'.
- A simple model is formulated initially to gauge which variables contribute significantly towards attrition. After identifying the significant variables, the final model is built and their performance is checked through various metrics as mentioned above.
- The logistic regression model eliminated the variables *HourInWeek, Designation and WorkExperience* and yet provided with an ***overall accuracy of 84.75% with a Sensitivity of 91.21%, Specificity of 51.05% and AUC of 82.46%. (at the threshold of 0.7).***

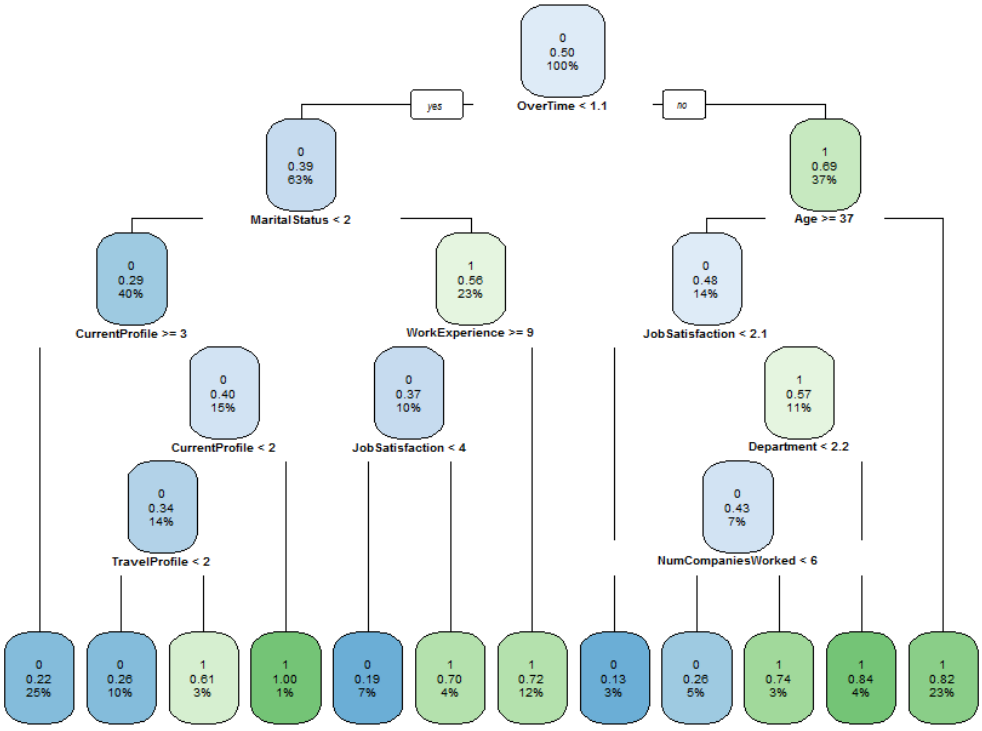


- According to the confusion matrix below, 145 out of 275 employees that attrite were correctly predicted, whereas 139 out of 1489 employees that did not attrite were incorrectly predicted to have done so.

	0	**1**
0	**1350**	**139**
1	**130**	**145**

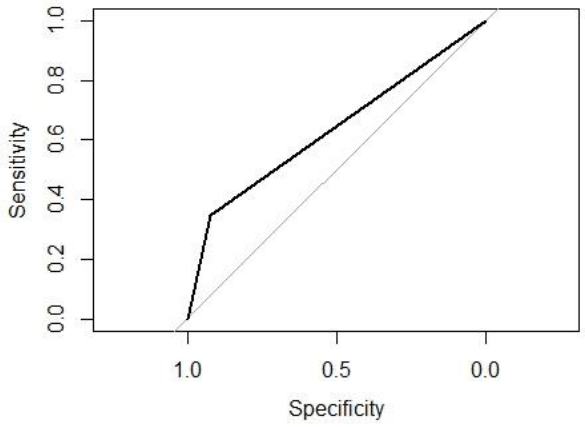
2. Classification and Regression Tree (CART) (Pre-Pruned)

- This is a decision tree model where each fork is split into predictor variables and each end node contains a prediction for the outcome variable.



Pre-Pruned CART Model

- The above Classification and regression tree shows us a very detailed model that is splitting observations multiple times. We note that, Overtime if less than or greater than 1.1 gives about 22 observations in total. Total number of splits in the tree is 11.
- Overtime is seen as the most important variable, followed by Age, Work Experience, Current Profile, Marital Status, Job Satisfaction, Monthly Income and so on.
- The pre- pruned CART Model provided an ***accuracy of 72.85 %, with Sensitivity at 73.04%, Specificity at 71.83% and AUC of 63.47%.***

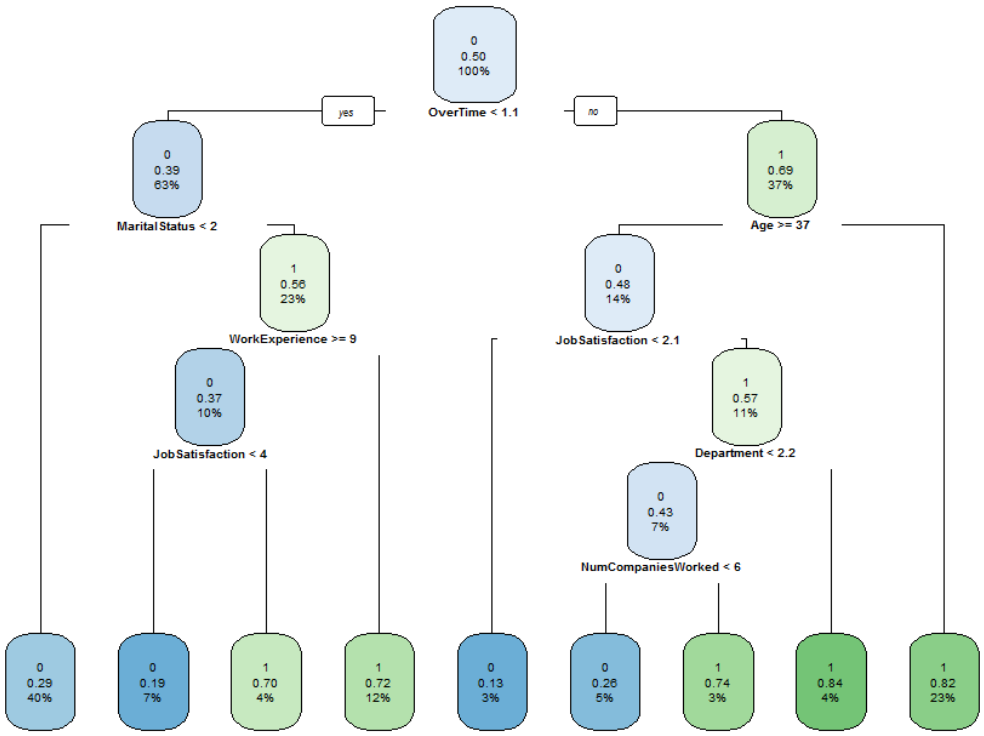


- According to the confusion matrix below, only 204 out of 603 employees that quit were correctly predicted and 80 out of 1161 employees that did not attrite were incorrectly predicted as so.

	0	1
0	1081	80
1	399	204

3. Classification and Regression Tree (CART) (Pruned)

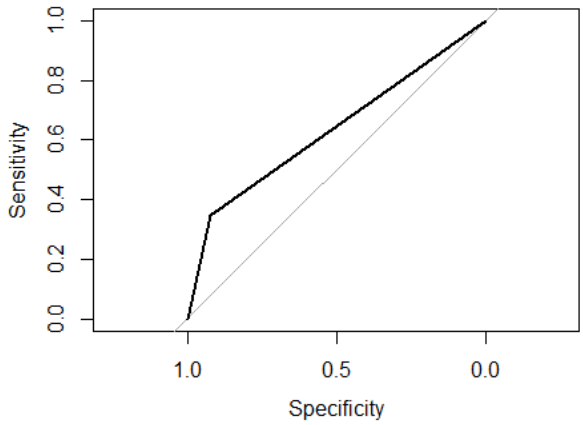
- Pruned model, as the name suggests involves cutting back the tree to avoid overfitting the model. The tree is pruned back slightly further than the minimum error. Here the control parameter (CP) is determined, plotted and the tree is pruned with the obtained CP value.



Pruned CART Model

-

- We note that, Overtime if less than or greater than 1.1 gives about 16 observations in total. Total number of splits is 8. Overtime is again seen as the most important variable followed by Age, Work Experience, Marital Status, Job Satisfaction, Monthly Income, Number of Companies Worked and so on.
- The pruned CART Model provided an ***accuracy of 74.15 %, with Sensitivity at 75.34%, Specificity at 67.96% and AUC of 63.52%. This is only slightly above the pre-pruned CART model.***

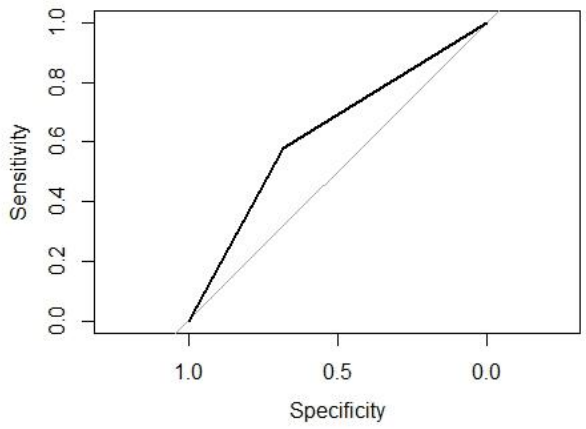


- According to the confusion matrix below, only 193 out of a total of 558 employees that quit were correctly predicted and 91 out of a total of a total of 1206 that didn't quit were incorrectly predicted to have to done so.

	0	**1**
0	**1115**	**91**
1	**365**	**193**

4. K-Nearest Neighbour

- KNN is a classifier which uses the Euclidean distance between a specified train and test samples of a data. It is not a mathematical formula. It essentially assumes that similar things exist in close proximity.
- The K-NN model provides an ***accuracy of 66.5%, with a Sensitivity of 68.11%, Specificity of 58.10% and AUC of 63.1%.***

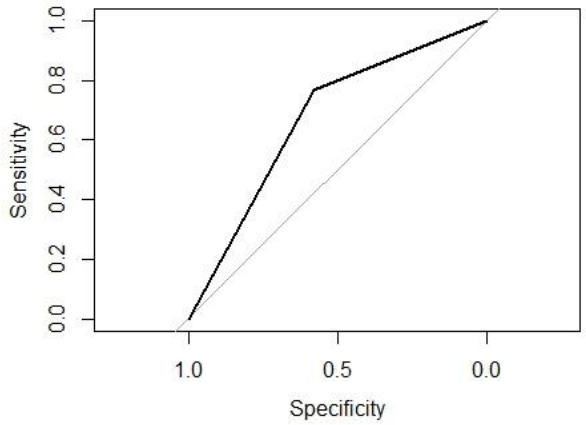


- It only managed to correctly identify 165 employees that left the company out of a total of 637, and incorrectly predicted 119 out of a total of 1127 that didn't attrite as having done so.

	0	**1**
0	**1008**	**119**
1	**472**	**165**

5. Naïve Bayes

- Naïve Bayes is a classifier that does not require us to specify a joint distribution between the conditional probabilities, which is essentially the naïve assumption. Naïve Bayes classifiers are a collection of classification algorithm based on Bayes' Theorem,
  P(A/B) = P(B/A) x P(A)/ P(B)
- The Prior probabilities are given as 50% and 50% for '0' i.e. employees do not attrite and '1' i.e. employees that attrite, respectively.
- Naïve Bayes requires us to specify the opposing conditional probabilities. All the predictor variables have a mean and standard deviation w.r.t employees that attrite and employees that do not attrite.
- Determining the posterior probabilities with the predict function and tabulating it against the test data, results in ***the accuracy rate of 56.07%, Sensitivity of 52.03%, Specificity of 77.11% and an AUC of 64.57%.***

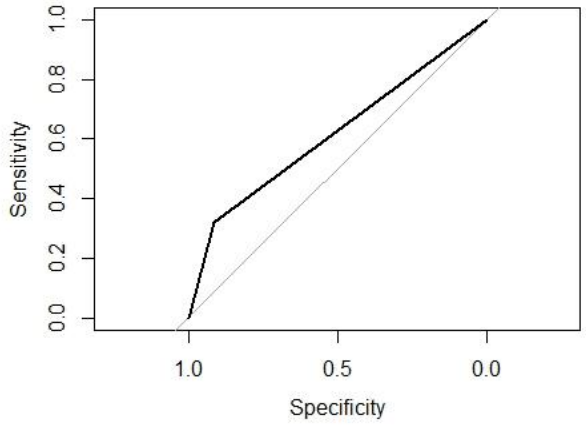


- According to the confusion matrix below, only 219 out of a total of 929 employees that quit were correctly predicted as such, and 65 out of a total 835 employees that didn't attrite were incorrectly predicted to have done so.

	0	**1**
0	**770**	**65**
1	**710**	**219**

6. Bagging

- Bagging, that often considers homogeneous weak learners, learns them independently from each other in parallel and combines them following some kind of deterministic averaging process.
- The ***accuracy rate through this ensemble method comes at 72.56% with Sensitivity of 74.39%, Specificity of 63.03% and AUC of 61.69%.***

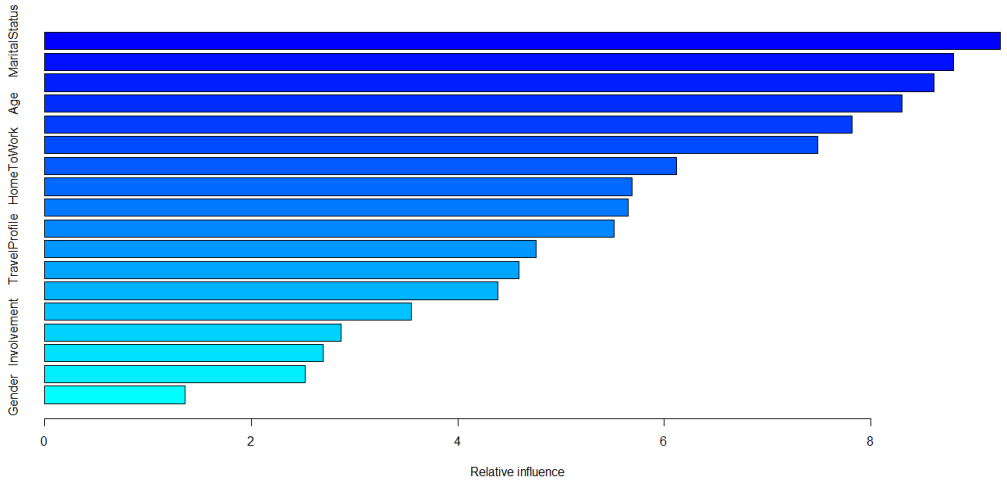


- 179 out of a total of 558 employees that quit were correctly predicted to have done so. Whereas, 105 out of a total of 1206 employees that didn't attrite were incorrectly predicted to have done so.

	0	**1**
0	**1101**	**105**
1	**379**	**179**

7. Gradient Boosting Machines

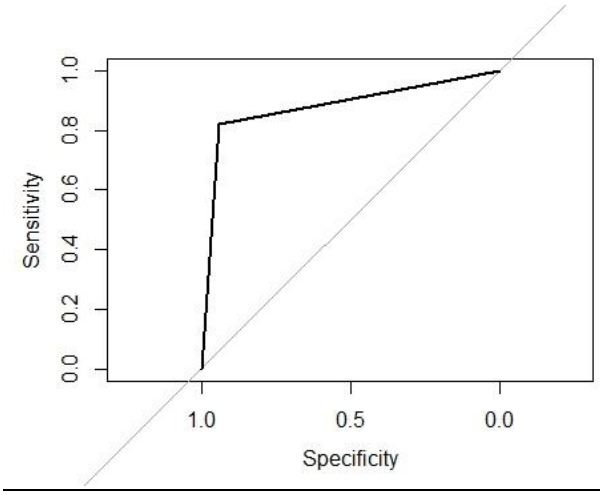
- **Boosting** often considers homogeneous weak learners, learns them sequentially in a very adaptive way (a base model depends on the previous ones) and combines them following a deterministic strategy.
- The gradient is used to minimize the loss function (error - difference between the actual values and predicted values). It is basically the partial derivative of the loss function, so it describes the steepness of our error function. In each round of training, the weak learner is built and its predicted values are compared to the actual values. The distance or difference between the prediction and reality represents the error rate of our model.



Variable	Relative Influence
MaritalStatus	9.25
MonthlyIncome	8.80
CurrentProfile	8.61
Age	8.30
Overtime	7.81

NumbCompaniesWorked	7.48
HomeToWork	6.12
JobSatisfaction	5.68
LastPromotion	5.65
WorkExperience	5.51
TravelProfile	4.76
SalaryHikelastYear	4.59
HourInWeek	4.39
Designation	3.55
Involvement	2.86
EducationField	2.69
Department	2.52
Gender	1.35

- Variable Importance is an important feature of GBM modeling. The variable importance table and plot shows the ranking of individual variables based on their relative influence, which is a measure indicating the relative importance of each variable in training the model. *As can be seen from the table and plot above, Marital Status, Monthly Income, Current Profile, Age are by far the most important variables in the gbm model.*
- The model has the ***accuracy rate of 92.12%, with a Sensitivity of 94.05%, Specificity of 82.04% and AUC of 88.05%.***

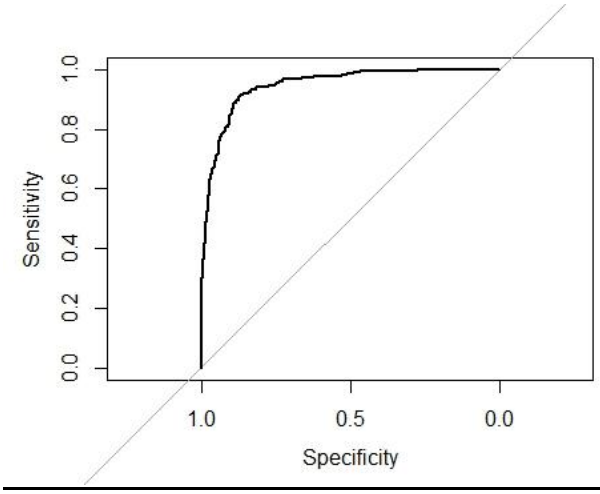


- The confusion matrix states that 233 out of 321 employees that quit have been correctly predicted. And, 51 out of a total of 1443 employees that didn't quit were incorrectly predicted to have done so.

	0	**1**
0	**1392**	**51**
1	**88**	**233**

8. eXtreme Gradient Boosting-XGBoost

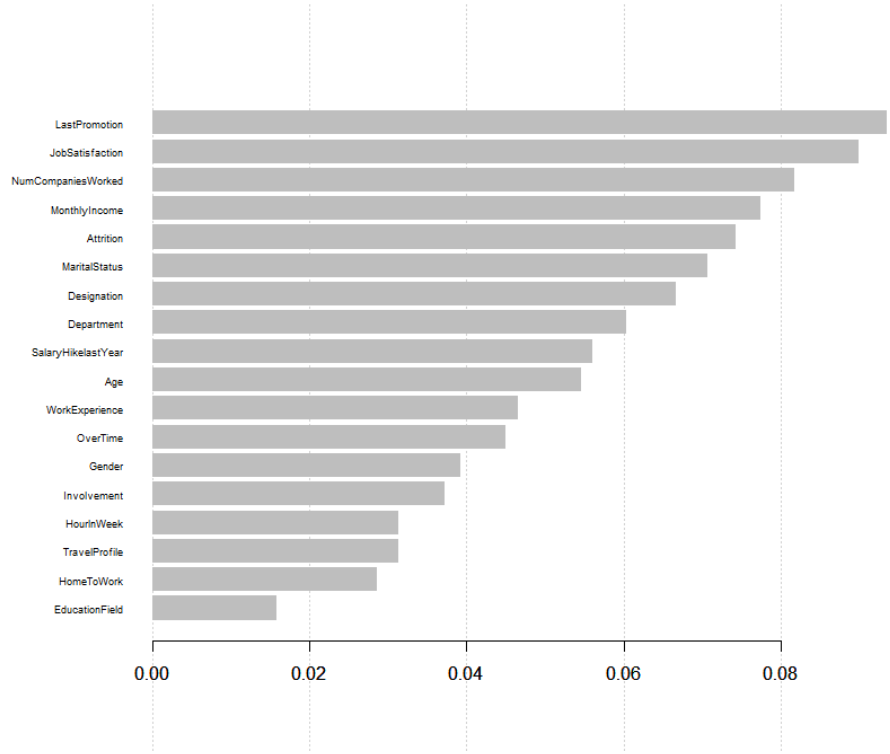
- XGBoost is a specific implementation of the Gradient Boosting method which delivers more accurate approximations by using the strengths of second order derivative of the loss function, L1 and L2 regularization and parallel computing.
- **The xgboost model delivers more accurate approximations in comparison to the gbm model.**
- It has an ***accuracy rate at 90.7%, with a Sensitivity of 95.56%, Specificity of 68.75% and AUC of 94.61%.***



- 220 out of a total of 320 employees that quit were correctly predicted. And, 64 out of a total of 1444 employees that didn't quit were incorrectly predicted to have done so.

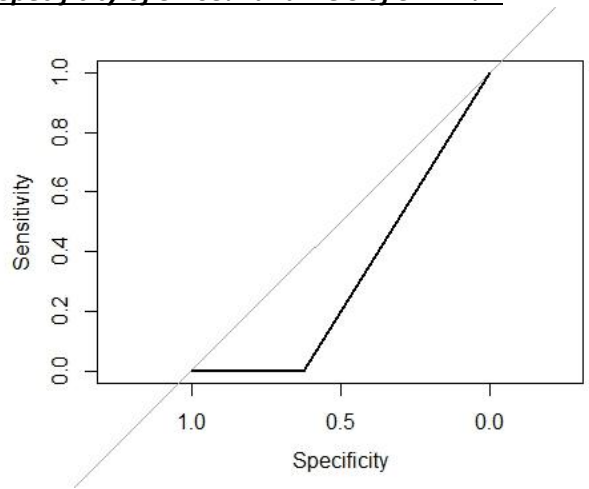
	0	**1**
0	**1380**	**64**
1	**100**	**220**

- Upon checking the variable importance for this model, the following is arrived at:



9. Random Forest

- The random forest is a supervised learning algorithm that randomly creates and merges multiple decision trees into one "forest." The goal is not to rely on a single learning model, but rather a collection of decision models to improve accuracy. The primary difference between this approach and the standard decision tree algorithms is that the root nodes feature splitting nodes are generated randomly.
- The confusion matrix for random forest model is unable to predict any number of employees that quit correctly, which essentially leads to a ***very low accuracy at 26.8%, with a Sensitivity of 0%, Specificity of 32.03% and AUC of 31.27.%.***



- There were no correct predictions made in this model, and 284 out of a total of 758 employees that didn't quit were incorrectly predicted as such.

	0	1
0	**474**	**284**
1	**1006**	**0**

Interpretation of the best model

Models	Performance Metrics			
	Accuracy	**Sensitivity**	**Specificity**	**AUC**
Logistic Regression	84.75%	91.21%	51.05%	82.46%
CART- pre-Pruned	72.85%	73.04%	71.83%	63.47%
CART- Pruned	74.15%	75.34%	67.96%	63.52%
K-NN	66.5%	68.11%	58.10%	63.1%
Naïve Bayes	56.07%	52.03%	77.11%	64.57%
Bagging	72.56%	74.39%	63.03%	61.69%
GBM	**92.12%**	**94.05%**	**82.04%**	**88.05%**
XGBoost	**90.7%**	**95.56%**	**68.75%**	**94.61%**
Random Forest	26.8%	32.03%	0	31.27%

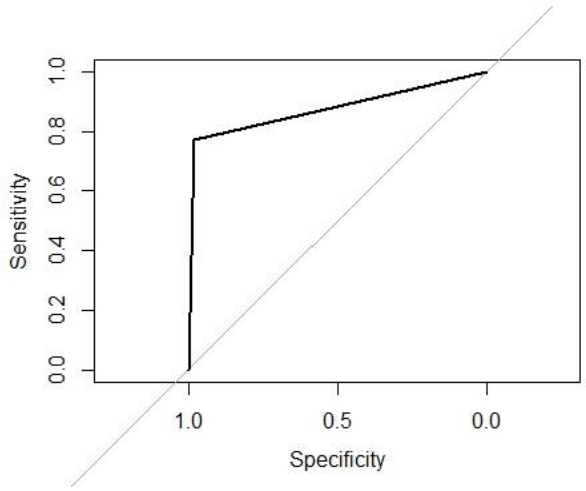
- Among the models above, Gradient Boosting Machines or **GBM had the highest accuracy rate at 92.12%** and AUC at 88.05%. However, **XGBoost had an accuracy rate of 90.7%** with AUC at 94.01%.
- However, there is more scope for XGBoost and Random Forest models with tuning.

Model Tuning

The following ensemble models are further tuned to get higher accuracy –

1. eXtreme Gradient Boosting

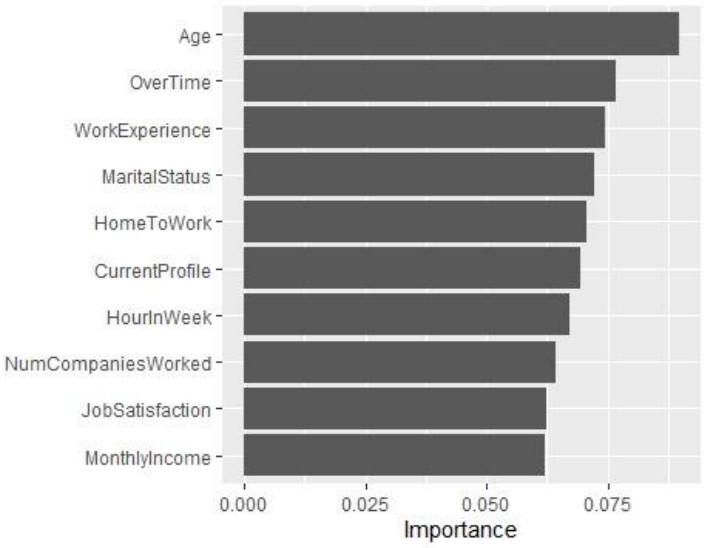
- Multiple tuning parameters are used to increase the accuracy of the model. After the model is tuned, the accuracy obtained is significantly higher than that of the default XGBoost model.
- ***This leads to a high accuracy rate of 94.1% - a significant boost from accuracy rate of 90.7% of our default XGBoost model. Sensitivity is at 94.93%, Specificity at 89.79%.***
- The AUC however falls to 87.6% from 94.6% which too is acceptable.



- The confusion matrix of the tuned XGBoost model correctly predicts 255 observations (employees that attrite) out of a total of 330. And, only 29 out of 1434 employees that didn't quit were incorrectly predicted.

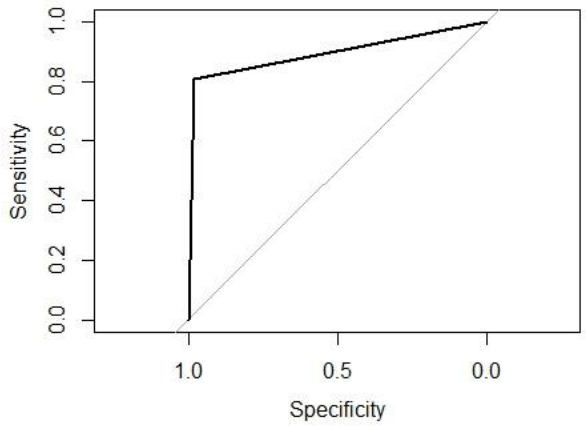
	0	**1**
0	**1405**	**29**
1	**75**	**255**

Top 10 variables contributing most significantly to the accuracy level above –



2. Random Forest

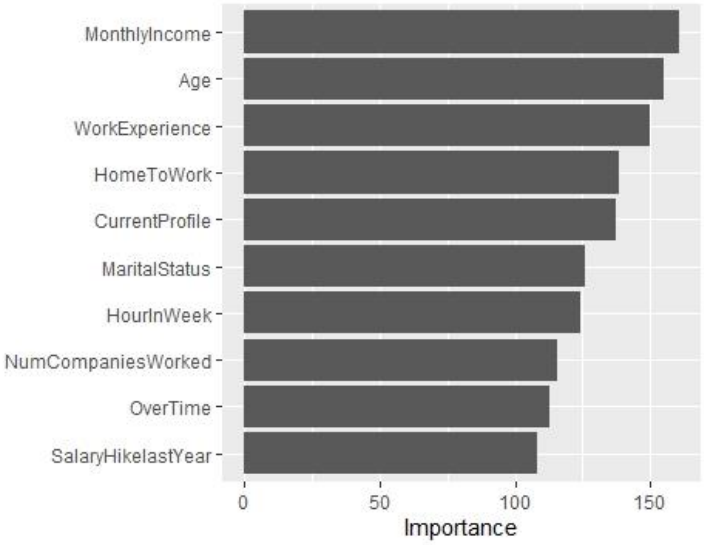
- The Random Forest model is tuned with multiple parameters and hyper-parameters until a high accuracy rate is arrived at. The confusion matrix of the tuned random forest model gives us an ***accuracy rate of 94.95%, with a Sensitivity of 90.85%, Specificity of 95.74% and AUC of 89.29%.***



- According to the confusion matrix below, 258 observations correctly predicted out of a total of 321(employees that attrite). Additionally, the proportion of incorrect predictions is on the lower side with only 26 observations incorrectly predicted out of 1443.

	0	**1**
0	**1417**	**26**
1	**63**	**258**

Top 10 variables contributing most significantly to the accuracy level above –



Final Interpretation of the best model

Models	Performance Metrics			
	Accuracy	**Sensitivity**	**Specificity**	**AUC**
Logistic Regression	84.75%	91.21%	51.05%	82.46%
CART- pre-Pruned	72.85%	73.04%	71.83%	63.47%
CART- Pruned	74.15%	75.34%	67.96%	63.52%
K-NN	66.5%	68.11%	58.10%	63.1%
Naïve Bayes	56.07%	52.03%	77.11%	64.57%
Bagging	72.56%	74.39%	63.03%	61.69%
GBM	92.12%	94.05%	82.04%	88.05%
XGBoost	90.7%	95.56%	68.75%	94.61%
XGBoost-tuned	94.1%	94.93%	89.79%	87.63%
Random Forest	26.8%	32.03%	0	31.27%
Random Forest-tuned	**94.95%**	**90.85%**	**95.74%**	**89.29%**

Conclusion and suggestions

- The variables mentioned above play a significant role in determining whether an employee will leave the company or not. **Random Forest can be seen to have the highest accuracy rate at almost 95% with a >90% sensitivity and specificity too.**

- An increase in Monthly income of competent employees (according to EDA, specifically employees within the salary bracket of 15K to 30K monthly) is likely to go a long way in retaining them. Changing the proportion of fixed vs variable component of the salary itself could prove to be beneficial.
- As employees within the age group of approximately 25 to 40 are more likely to quit, special efforts must be made in retaining that age group by devising succession planning, growth opportunity, inclusivity, cultural acceptance etc.
- Special incentives in terms of dynamic package, succession planning, benefits ought to be given to employees with work experience between 3 to 10 years, as it is likely that such employees may be on the look-out for packages/ benefits best in said industry.
- The issue of distance from home to work is solvable through travel allowances, pick-up/ drop-off options or the relatively new concept of 'work from home'.
- Employees who have been in their current profile for 2 to 8 years are seen to have a higher attrition rate. The HR department should focus on making sure that the employees do not feel stagnant or saturated in the company. An employee ought to feel challenged yet secure in their position at the company to stay motivated to give his/her best and as a result not feel the need to move.
- As a part of manpower planning efforts could be made towards hiring more settled/married employees who are seen to have the least amount of attrition as compared to single employees who despite being lower in number, have a high rate of attrition. This could reduce overall hiring/ attrition costs. Additionally, if feasible steps maybe taken to ensure retention of single employees to promote a diverse work environment. A healthy work environment would technically contain a mix of all kinds of gender, age groups, qualifications etc. Retention measures would differ for each such criterion and would have to be individually determined.
- It is observed that majority of employees commit approximately 40 to 80 working hours in a week, which roughly average to about 8 to 16 hours every day. Keeping this working hour range in mind, dedicated work timing could be prescribed with benefits and/or provisions for over-timing.
- Employees who have worked at fewer companies i.e., (1 to 10) are seen to have a higher rate of attrition. Such employees ought to be retained with competent salaries, benefits, etc. Additionally, hiring could be done by keeping a check of this and employees with more experience with multiple different companies could be selected.
- Lower the salary hike, higher the rate of attrition. Salary hikes should be performance based, and could be more dynamic for employees with potential in order to retain them. There may be provisions for additional salary hikes over and above the norm on the basis of an employee's appraisal report and overall performance.

<u>These predictor variables should be the key focus of the client whilst managing their rate of attrition. All of the above factors if put into a combined force is sure to provide results in ensuring long term retention of employees.</u>