Workflow

Different parts of all code have been set using "#### title ####". Use the hashtags in each Rscript to navigate and jump among sets.

In "Time_Features.R"

1. Used the in-time and out-time data to engineer several features. All features were combined together in HoursWorked dataset which was finally combined to the master dataset.
2. Merged the three data-sets (employee_survey, manager_survey, and HoursWorked) by the unique key (EmployeeID) which is common in all three datasets. This dataset is called "master".

In "EDA.R"

3. Detected and removed the basic anomalies if any
4. Treated x number of NA values which were present in five columns one variable at a time.
5. Used reasoning and Decision Trees to replace NAs

In "Univariate_Analysis.R"

6. Removed outliers, if any, during univariate analysis.
7. Categorized the variables into 7 categories and then conducted an in-depth analysis of all variables one by one mainly using Histograms. (Univariate Analysis)
8. Feature engineered another variable.

In "Bivariate_Analysis.R"

9. Bivariate Analysis was done using Histograms for categorical variables and Density Graphs for continuous variables for all features, against the target variable and insights were noted.

In "Model_Building"

10. Based on the outcome requirements, models were chosen and predictions were made. Their outcomes (accuracy, sensitivity, and specificity) was noted and compared. Before implementing each model, the data was prepared as per the requirements of each model

The models were implemented in the following order-

i. Random Forest
ii. Logistic Regression
    a. without SMOTE
    b. with SMOTE
iii. Decision Tree (CART model)
iv. K nearest neighbours
v. Naïve Bayes
vi. Support Vector Machine (SVM)

The tools used were R Software for most of it, MS Excel to create Table 1 and Table 2, and MS word for the report.

Variable data-types: The dataset consisted of a mix of categorical and continuous variables. All types of Nominal, Ordinal and Interval were present under categorical.

Target Variable:

Attrition is the target variable which has the output “Yes” or “No”.

There is slight Target Class Imbalance. Target Class Imbalance is when the majority class dominates the minority leading to very unequal distribution in the target class. This can lead to bias in prediction as the model will tend to lean towards the majority class while still managing to maintain high accuracy. Few ways to handle this is-

i. Oversampling: duplication of rows belonging to minority class.
ii. Under-sampling: Removing rows belonging to majority class.
iii. SMOTE or Synthetic Minority Oversampling Technique: Tends to strike a balance between (i) & (ii) by not losing too much data due to Under-sampling and by not unnecessarily bulking up the data by Oversampling a lot.

Most of the models give a biased prediction in such cases. Ensemble models such as Random Forest are efficient when dealing with imbalanced class problems.

We need to keep the class ratio maintained once we split the data in test and training for our model evaluation to avoid any further biases.

When using any of the above methods to handle target class imbalance, it should be applied on training set only and not the test/val set. After training the model, test/val set is used to evaluate the model. It is important for test/val set to have original data.

Variables inserted by Feature engineering are –

1. By in_time and out_time
   First, subtracted out-time from in-time to calculate number of hours worked each day. Columns with full NA values are public holidays such as Republic Day, May Day, Christmas and Gandhi Jayanti when all of the employees did not come to work. Remaining NA values are those when the employee took leave, therefore there is no in-time and out-time on those particular days for the employee. The featured generated are:
   a. Average_HoursWorked: Total hours worked / number of days worked
   b. LeaveTaken: sum of NA values row wise
   c. NumOvertimeDays: sum of days when employee has worked overtime
   d. Overtime: whether the employee worked overtime or not
2. Employee_survey and manager_survey
   a. EmployeeRating: Sum of all variables from employee_survey and manager_survey that measure employees' satisfaction and performance level. The employees are rated out of 20.

Division of variables into 7 sub-categories based on their characteristics

A. Employee Background
   1. Age
   2. Gender
   3. Marital Status
   4. Education
   5. Education Field
   6. Total Working Years
   7. Number of Companies Worked
B. Position and Experience

1. Department
2. Job Role
3. Job Level
4. Years at Company
5. Years with Current Manager

C. Payment/ Salary
    1. Monthly Income
    2. Percent Salary Hike
    3. Stock Option Level

D. Travel and Work Time
    1. Distance from Home
    2. Business Travel
    3. Overtime
    4. Number of Overtime Days
    5. Average Hours Worked
    6. Leave Taken

E. Employee Satisfaction
    1. Environment Satisfaction
    2. Job Satisfaction
    3. Work Life Balance

F. Employee Performance
    1. Job Involvement
    2. Performance Rating
    3. Employee Rating

G. Employee Development
    1. Years Since Last Promotion
    2. Training Times Last Year

False Positive vs False Negative in Attrition. Which is worse?

A **False Negative** means that the employee **will leave** the company but our model does not detect that and predicts that the employee will not leave.

A **False Positive** is the opposite, it means that the employee **will not leave** the company but the model predicts that the employee will leave.

Which one will prove to be more tragic for the company? As human resource is the most valuable asset in a company, False Negative (FN) is more problematic for the company (XYZ) because the company will not make any efforts to retain that employee as it thinks that he/she will not resign/leave. The company could have retained the employee if they identified him/her correctly and provided him with added benefits or addressed his woes, issues and complaints on time.

A False Positive will lead to the company spending extra time and resources making an employee stay in the company, who anyway was not going to leave in the first place. There will be brainstorming among HR managers about why he wants to leave and how can they retain him, while he/she is perfectly fine with the current situation.

Therefore, our aim is also to reduce the **False Negative Rate** which is calculated by:

1 – Sensitivity (True Positive Rate)

Where, **sensitivity** is calculated by: (True Positives / (True Positives + False Negatives))

Model Selection:

Multiple methods or algorithms were used to train the model. The model which gave the highest accuracy along with a balanced Sensitivity and Specificity was selected.

Requirements:

i. Handling imbalance in target class distribution: Ensemble models such as Random Forest is very efficient with handling target class imbalance due to Bootstrap-Aggregating (B-agging) method.

ii. Knowing the importance of variables is important: Logistic Regression, CART and RF models show us variable importance directly.

Keeping these requirements in mind **Random Forest** model was selected.

Random Forest model gave the highest accuracy of 98.37% on validation data with Sensitivity and Specificity being 98.19% and 98.40% respectively.

Other models such as Logistic Regression and Decision Tree (CART) model provide us with variable importance which is an important factor. To check the impact of target class imbalance in the dataset, Logistic Regression was performed twice, once with original data and the second time with sampling using SMOTE. It was observed that there was not much of a difference in both. However, LR w/o SMOTE had a batter balance in the True Positive Rate (Sensitivity) and True Negative Rate (Sensitivity). Therefore, for all other models oversampling or undersampling options have been eliminated.

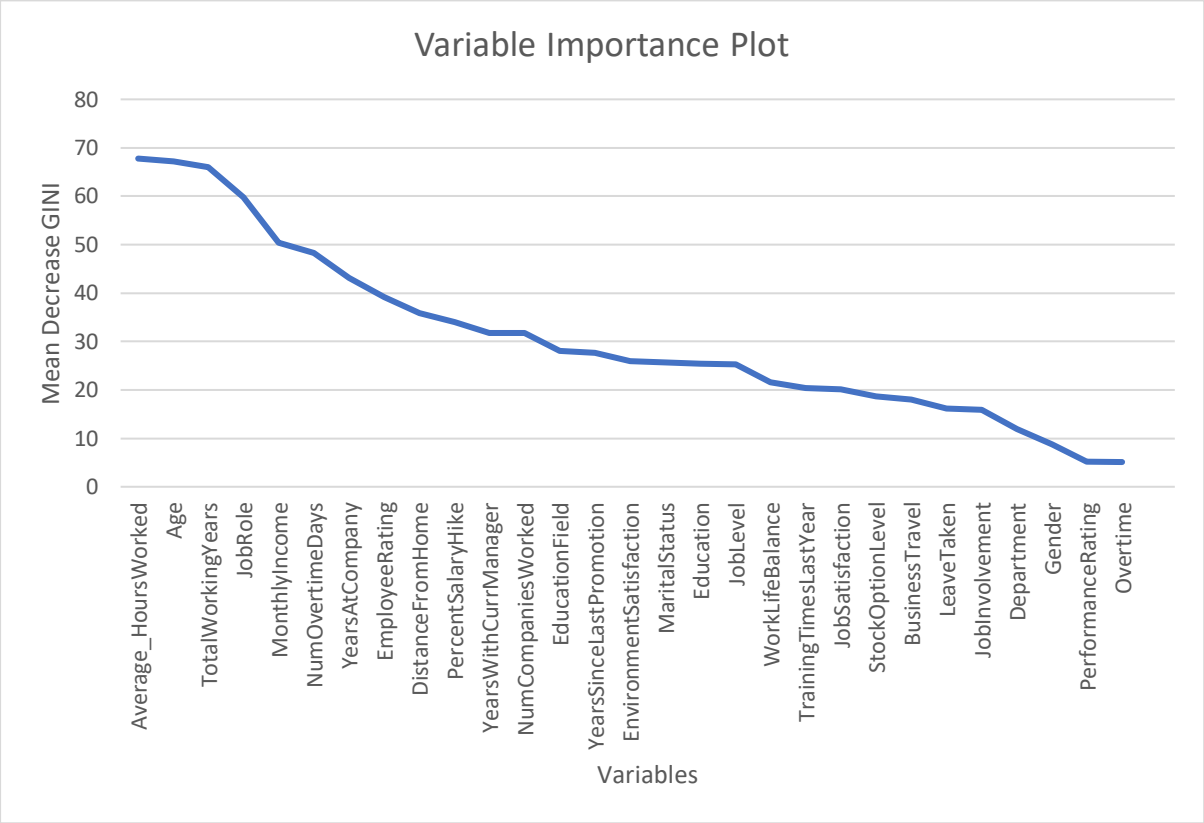
K nearest neighbours, Naive Bayes, and Support Vector Machines were other models which were tested. Out of these three, SVM with radial kernel gave the best prediction.

Table 1: Model Performance Comparison

	Accuracy	Sensitivity	Specificity	False Negative Rate	Rank
Random Forest	0.9837	0.982	0.984	0.018	1
Logistic Regression w/o SMOTE	0.8459	0.78	0.84	0.22	6
Logistic Regression with SMOTE	0.8531	0.75	0.86	0.25	7
Decision Tree	0.8513	0.89	0.85	0.11	3
k-NN	0.8477	0.916	0.847	0.084	5
Naïve Bayes	0.8486	0.823	0.848	0.177	4
Support Vector Machine	0.9012	0.897	0.902	0.103	2

Which of these variables is most important and needs to be addressed?

Table 2. Variable importance

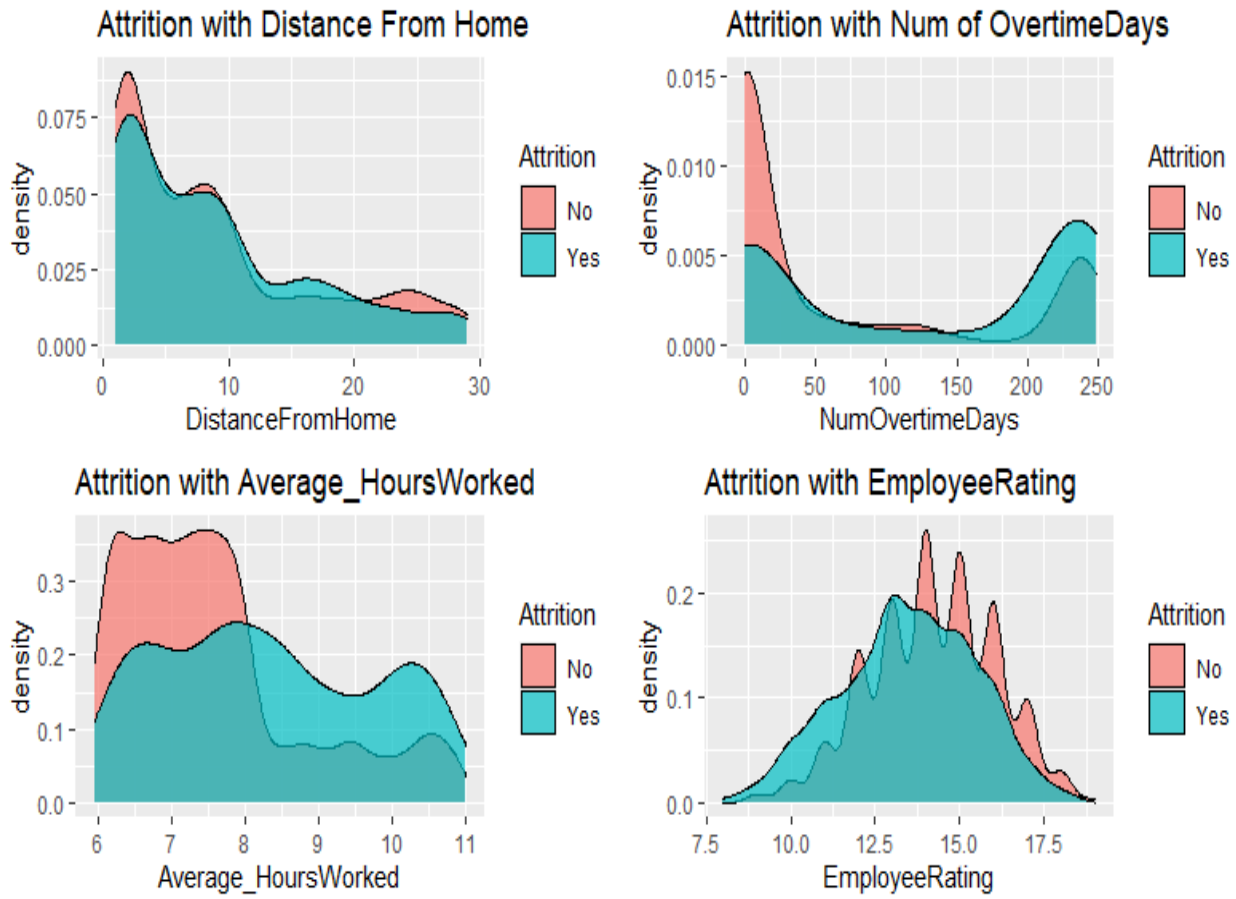


These are the important variables in accordance to the model which gave the best predictions (RF). The top 10 most important variables which need to be addressed and the category they belong to in order are-

1. Average_HoursWorked - Travel & Work Time
2. Age - Employee Background
3. TotalWorkingYears - Employee Background
4. JobRole - Position & Experience
5. MonthlyIncome – Payment & Salary
6. NumOvertimeDays - Travel and Work Time
7. YearsAtCompany - Position & Experience
8. EmployeeRating – Employee Performance
9. DistanceFromHome - Travel and Work Time
10. PercentSalaryHike - Payment & Salary

Let’s see their impact on Attrition group by group.

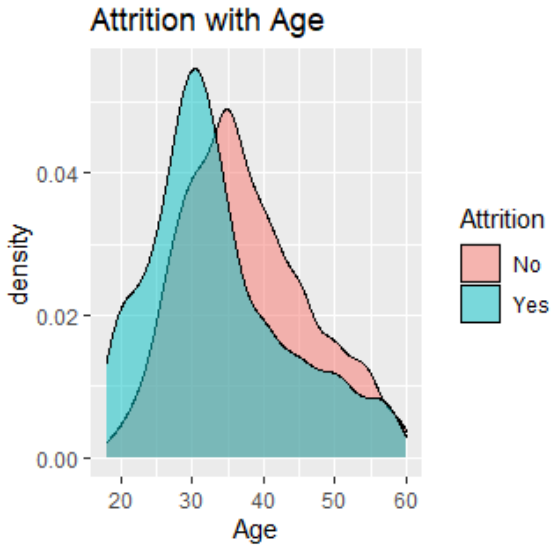
Travel & Work Time, and Employee Rating:

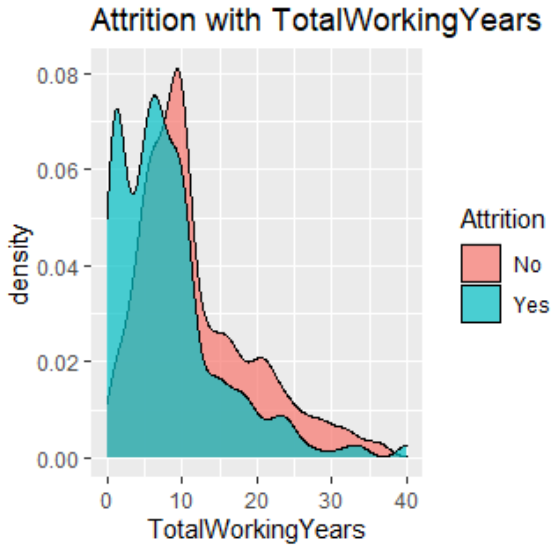


- Attrition rate sees an increase between after 10 kms but drops again after 21.
- Employees who have worked overtime more than 150 days have very high attrition rate.
- Attrition is very high for employees who work more than 8 hours a day
- Attrition is very high for employees whose rating is below 13. After 13 it decreases gradually.

Note: These illustrations were combined using “grid.arrange()” function from the package “gridExtra”. It has been removed from “Bivariate_Analysis.R” to prevent confusion and ensure smooth interpretability.

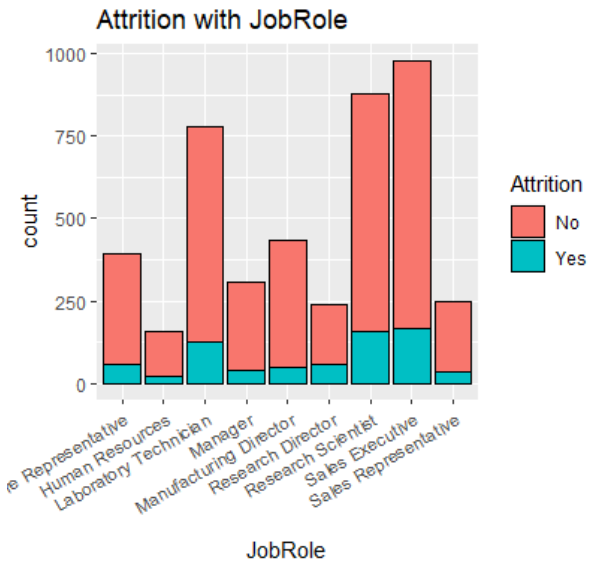
Employee Background:

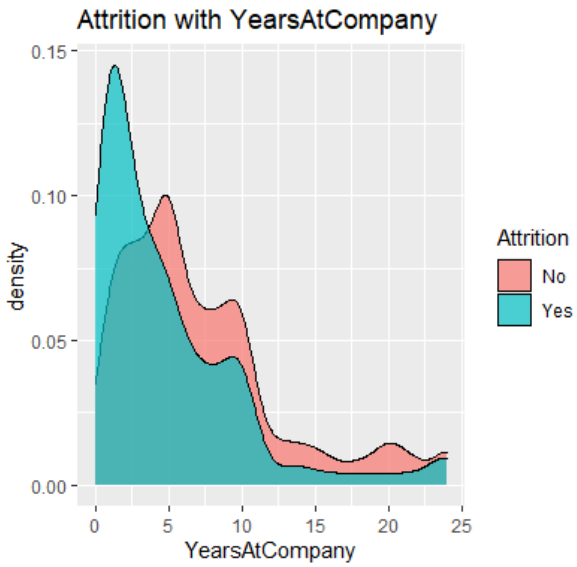




- Younger employees within 25-35 years have a higher attrition rate
- Higher Attrition for employees whose Total working years is less than 6-7. This is because it gets riskier with age to quit a job due to reasons such as family.

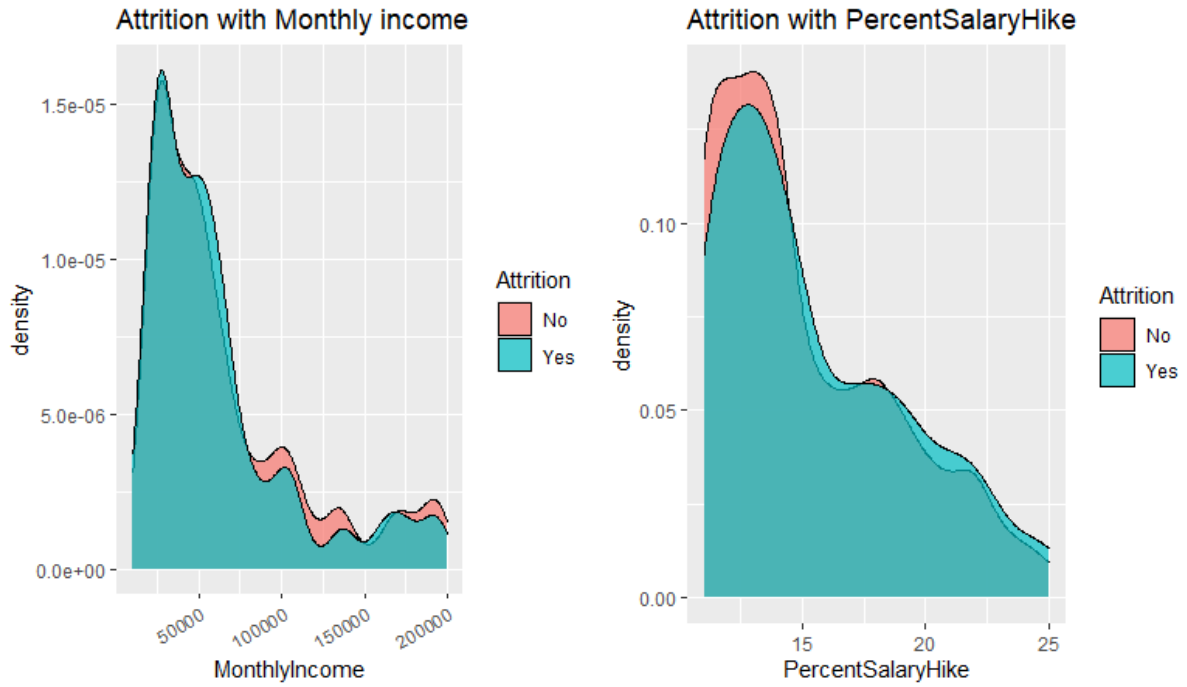
Position and Experience:





- Attrition is high for Research Scientist and Sales Executives
- Attrition is very high for employees who have been in the company for less than 2-3 years. For older employees the fear to join a new company, get used to the new environment and start from the bottom all over again, can be overwhelming. For new-comers it is usually not. Thus, they often leave.

Payment & Salary:



- Attrition is high for employees earning between 10,000 to 20,000 per month.
- Employees with low increase in salary every year have higher chances to leave.

What factors should the company focus on, in order to curb attrition? What changes should they make to their workplace, in order to get most of their employees to stay?

1) To begin with, the company should tackle the issue at its core. Employees can only leave the company if they have been hired by someone at some point thinking that he/she is perfect for the job: Hiring the right people. Spend more time in recruiting and selecting the right talent based on the quality of the candidate. Make sure that he/she is right for the role hired by increasing the depth of the selection process by including multiple Group Discussions, Technical Interviews and HR Interviews. Foe eg., it was observed that employees who have studied HR do not have job roles in the relevant field. Maybe a case of poor or neglected hiring. These cases will lead to increased attrition rate as such employees do not work in fields they specialize in, thus leading to low job satisfaction.

2) Research Scientist and Sales Executives have higher attrition rate. It was observed that these two job roles had the highest number of employees with very low job satisfaction. Increasing commission for Sales Executives is a well-known proven method of increasing employee retention for employees in the sales department.

3) The company can focus Providing extra incentives, perks and higher bonuses to newer and/or younger employees based on their performance. Providing better growth and promotion opportunities to such employees because younger employees are mostly in search of opportunities with better job roles and higher pay.

   Employees in their late twenties and early thirties usually get married and become parents. Such employees would appreciate and remember if they get additional fully paid family leaves at the time of occurrence of such events. Also, job security or company's contribution towards the employees' family's health insurance premiums will give the employee a sense of belongingness in the company and make him/her want to stay and work hard.

4) Monthly income and percentage increase in salary are very important factors. If an employee is highly satisfied with his/her job and has high performance, the company may still fail to retain them since their salary is very low or has been constant for several years in a row or the increase in salary per year is low/ not as desired.

5) Employees working overtime for more than 150 days have a very high attrition rate. Company can set a ceiling limit to the number of days an employee does overtime. For example, total number of permissible days to work overtime is 180 days. However, this can prove to be harmful for the company as it will lead to reduction in productivity.

Instead, company can increase the overtime payment after the employee has worked overtime for more than 180 days. For example-
Basic overtime rate: **1.5** x Daily Pay Rate x Overtime days worked
Overtime rate after employee has worked overtime for more than 180 days can be increased to 1.8 or 2 times: **1.8(**or **2)** x Daily Pay Rate x Overtime days worked

This will act as an incentive and employees will feel that their time and work given to the company is recognized and valued.
The same with Overtime Hours worked, this should be calculated on an hourly pay rate, and the cut off can be 10 hours.

6) As for low Employee Rating, if all the above factors are handled it will lead to an overall increase in satisfaction levels, which in turn, will curb attrition.

Conclusion:

Entire process of Analysis was carried out step by step. Appropriate models were selected based on the requirements and analysis was made. The core factors which resulted in such a high attrition rate was identified.

Now that Company XYZ has successfully identified the issue, proper and efficient steps should be taken to curb it. The results thus obtained should be used by the management to understand what changes they should make to their workplace, in order to get most of their employees to stay.