SUMMER INTERNSHIP-2020 – DATASCIENCE

EMPLOYEE ATTRITION ANALYSIS AND PREDICTION

INTRODUCTION:

This study aims at finding out the factors that leads to employee attrition and in predicting the employee attrition using some Machine Learning Algorithms.

PROCEDURE:

1. DATA PREPARATION

The given Employee Attrition dataset consists of no any missing values hence no need of formatting the dataset.

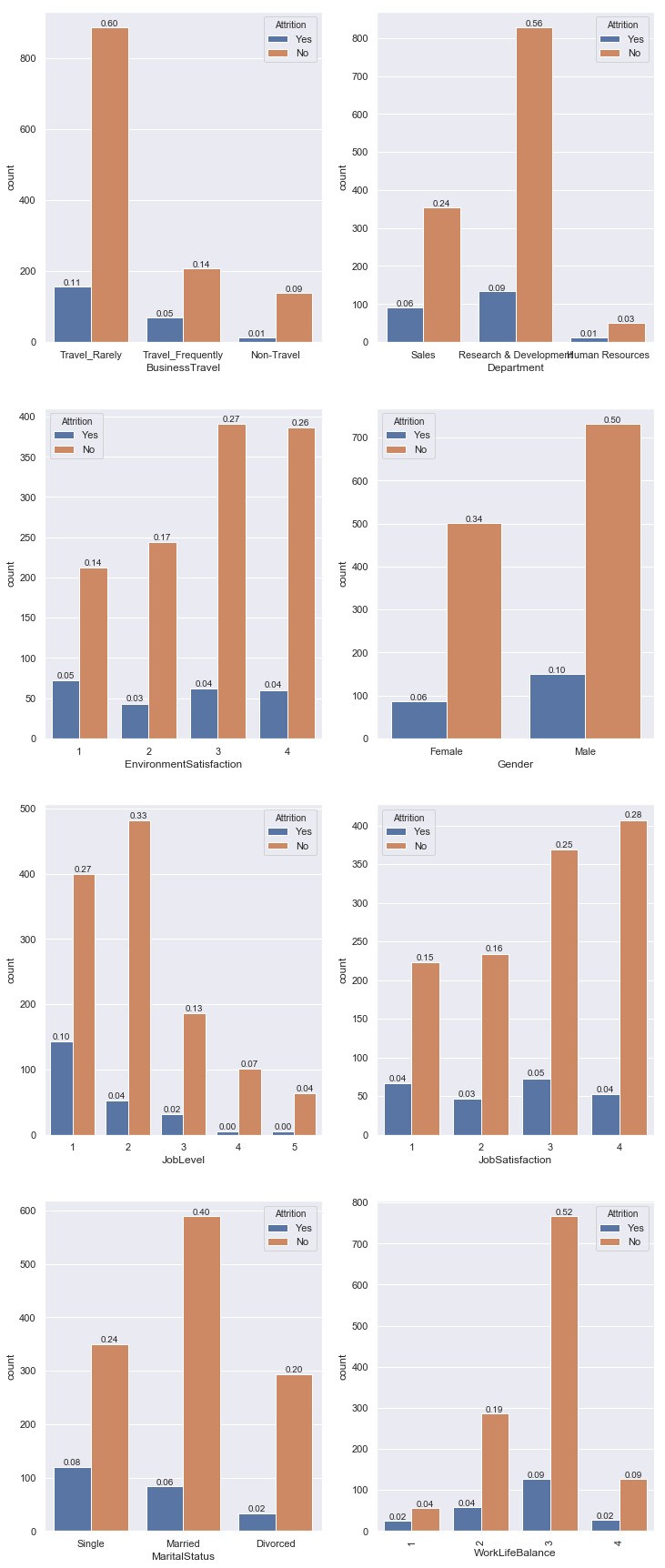
1. DATA ANALYSIS

By getting the unique values of each columns it has be seen that, 'EmployeeCount', 'Over18', 'StandardHours' have only one unique value and 'EmployeeNumber' has 1470 unique values. These features are not useful hence, these columns have been dropped.

1. DATA VISUALIZATION

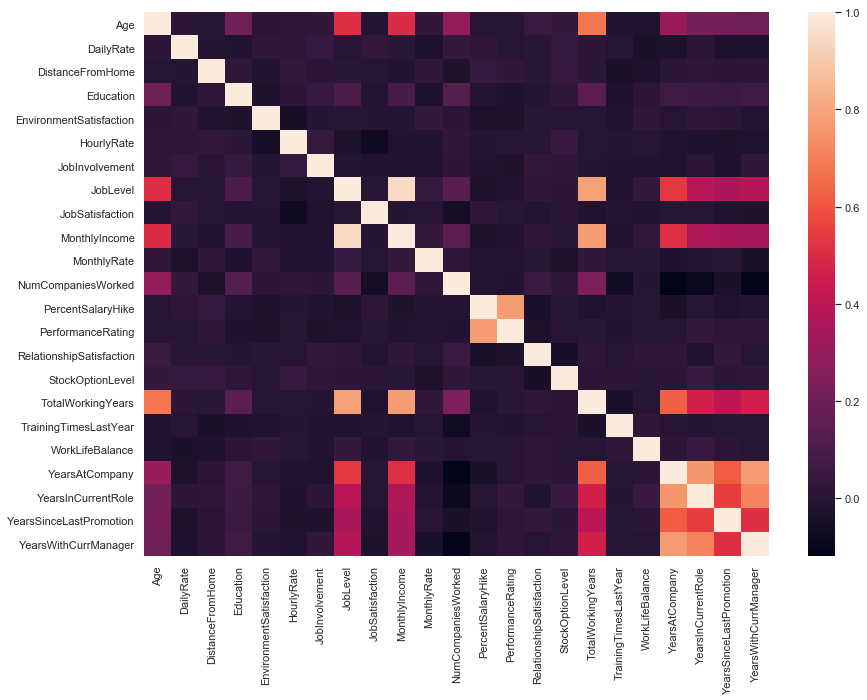
Multiple count plots have been used for visualization of data.

The following visualizations have been made:



1. The attrition rate is high for Frequently business travelers when compared with others
2. Environment Satisfaction Level1 has a high attrition rate
3. Single attrition rate is high in Marital Status
4. Male attrition rate is high when compared with the female
5. Attrition rates are high in these attributes in Sales Department
6. Jobsatisfaction Level1 has a high attrition rate
7. FEATURE SELECTION

Correlation heatmap has been used for feature selection for the model



The highly correlated variables are dropped and poorly correlated variables are taken for fitting the model

1. MODELLING THE DATA

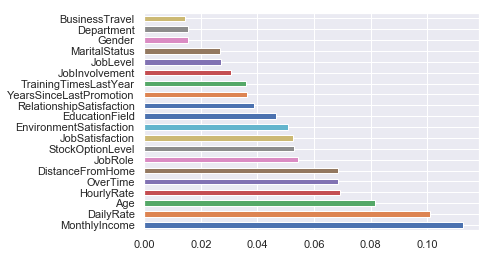
Since the dataset consists of some categorical data it has been converted into numeric data.

Three classification models Logistic Regression, Decision tree and Random Forest have been used in this study in order to find the suitable model for the dataset.

The accuracy level for LOGISTIC REGRESSION: 0.8695652173913043

The accuracy level for DECISION TREE: 0.7418478260869565

The accuracy level for RANDOM FOREST: 0.8435374149659864



Random Forest provides the importance of features in the model. The features that contributes the most in attrition are Monthly Income, Daily Rate, Age, Hourly Rate and Over Time.

CONCLUSION:

Though Logistic Regression shows higher accuracy than Random Forest, **Random Forest** is the suitable algorithm as the dataset contains both categorical and numeric data. The model gave accuracy of 0.84 which is quite good. The random forest works quite well even with the default parameters. It also doesn’t over fit easily because of its randomness feature.