**Predicting the Employee Attrition from IBM Data Set**

**Project Overview**

Employee Attrition is one of the most studied and important issues to organizations because of its adverse impact on workplace productivity and long-term growth strategies. It is expensive for organizations since they incur significant costs both directly and indirectly. When an employee performs voluntary attrition the critical information of the organization is carried along. Higher attrition rate is recognized as being a major issue for HR. In this study, we aim to determine the key factors that lead an employee to perform attrition.  

The data was collected from

<https://www.kaggle.com/HRAnalyticRepository/employee-attrition-data>

**Metrics**

A common metric for binary classifier is the accuracy. Accuracy takes into account the true positive and true negative value with equal weight.

**Accuracy** = (True Positive + True Negative) / Size of the dataset

**F-1 Accuracy:** This is a mean of Precision and Recall that gives a better measure of the incorrectly classified cases than the Accuracy

**Precision:** It is implied as the measure of the correctly identified positive cases from all the predicted positive cases

**Recall**: It is implied as the measure of the correctly identified positive cases from all the actual positive cases

These metrics are used when evaluating the classifiers.

**Analysis and Exploratory visualization**

**Representation of training data is provided below**

**Age (Real numbers)**

|  |  |  |
| --- | --- | --- |
| **Distinct Value** | 43 |  |
| **Distinct (%)** | 3.8 |
| **Missing** | 0 |
| **Minimum** | 18 |
| **Maximum** | 60 |
| **Zeros** | 0 |

**Attrition (Boolean Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 2 |  |
| **Distinct (%)** | 0.2 |
| **Missing** | 0 |

[**Business Travel**](file:///C:\Users\Hp\Analysis.html#pp_var_1371214274023623635) **(Categorical Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 3 |  |
| **Distinct (%)** | 0.3 |
| **Missing** | 0 |

**Daily Rate (Real Number)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 766 |  |
| **Distinct (%)** | 67.3% |
| **Missing** | 0 |
| **Maximum** | 102 |
| **Minimum** | 1499 |

**Department (Categorical Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 3 |  |
| **Distict(%)** | 0.3% |
| **Missing** | 0 |

**Distance from Home (Real Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 29 |  |
| **Distinct (%)** | 2.5% |
| **Missing** | 0 |
| **Maximum** | 29 |
| **Minimum** | 1 |

**Education (Categorical Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 5 |  |
| **Distinct (%)** | 0.4% |
| **Missing** | 0 |

**Education Field (Categorical Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 6 |  |
| **Distinct (%)** | 0.5% |
| **Missing** | 0 |

**Employee Number (Real Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 1139 |  |
| **Distinct (%)** | 100% |
| **Missing** | 0 |
| **Minimum** | 1 |
| **Maximum** | 1606 |

**Environment Satisfaction (Categorical Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 4 |  |
| **Distinct (%)** | 0.4% |
| **Missing** | 0 |

**Gender (Categorical Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 2 |  |
| **Distinct (%)** | 0.2% |
| **Missing** | 0 |

**Hourly Rate (Real Number)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 71 |  |
| **Distinct (%)** | 6.2% |
| **Missing** | 0 |
| **Minimum** | 30 |
| **Maximum** | 100 |

**Job Involvement (Categorical Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 4 |  |
| **Distinct (%)** | 0.4% |
| **Missing** | 0 |

**Performance Rating (Categorical Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 2 |  |
| **Distinct (%)** | 0.2% |
| **Missing** | 0 |

**Job Level (Categorical Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 5 |  |
| **Distinct (%)** | 0.4% |
| **Missing** | 0 |

**Job Role (Categorical Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 9 |  |
| **Distinct (%)** | 0.8% |
| **Missing** | 0 |

**Job Satisfaction (Categorical Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 4 |  |
| **Distinct (%)** | 0.4% |
| **Missing** | 0 |

**Marital Status (Categorical Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 3 |  |
| **Distinct (%)** | 0.3% |
| **Missing** | 0 |

**Monthly Income (Real Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 1057 |  |
| **Distinct (%)** | 92.8% |
| **Missing** | 0 |
| **Minimum** | 1009 |
| **Maximum** | 19999 |

**Number of Companies Worked (Real Data)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 10 |  |
| **Distinct (%)** | 0.9% |
| **Missing** | 0 |
| **Minimum** | 0 |
| **Maximum** | 9 |

**Over Time (Boolean)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 2 |  |
| **Distinct (%)** | 0.2% |
| **Missing** | 0 |

**Performance Salary Hike**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 15 |  |
| **Distinct (%)** | 1.3% |
| **Missing** | 0 |
| **Minimum** | 11 |
| **Maximum** | 25 |

**Total working years (Real numbers)**

|  |  |  |
| --- | --- | --- |
| **Distinct** | 40 |  |
| **Distinct (%)** | 3.5% |
| **Missing** | 0 |
| **Minimum** | 0 |
| **Maximum** | 40 |

**Data Preprocessing:**

* Data set is divided into a train set and a test set. 1139 data are used as a train set and 331 data are used as a test set to test the prediction

**Removing constant valued features:**

* We removed the data that have only have a specific constant value (Distinct value) that is the feature **Employee Count, Over18** and **Standard Hours**

**Feature Selection**

Since after encoding the categorical data, we have approximately 50 features, so we are trying to select the features that have the most impact on predicting the result of attrition. Among all of the other selection procedure we choose the lasso since it penalizes strongly.

Lasso finds these the following features that are not important while predicting in the logistic regression.

* Job Level
* In Business Travel: The Travel Rarely
* In Department: Human Resource and Research & Development
* All options of Educational Field
* All of the options for Job Role excel Laboratory Technician
* Marital Status other than single

**Logistic Regression**

Accuracy is 88.03%

Precision: 41.67%

Recall: 71.14%

We have implemented LDA, QDA, KNN, Decision Tree, Support Vector Machine, Random Forest with only the preprocessed data and the accuracy table is given below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Decision Support Tree** | **KNN** | **Support Vector Machine** | **Random Forrest** | **LDA** | **QDA** |
| 76.6 | 81.65 | 84.04 | 87.1 | 87.1 | 37.1 |