**IBM HR Analytics Employee Attrition Modeling.**

**DESCRIPTION**

IBM is an American MNC operating in around 170 countries with major business vertical as computing, software, and hardware. Attrition is a major risk to service-providing organizations where trained and experienced people are the assets of the company. The organization would like to identify the factors which influence the attrition of employees.

**Data Dictionary:**

**Age**: Age of employee

**Attrition**: Employee attrition status

**Department**: Department of work

**DistanceFromHome**

**Education**: 1-Below College; 2- College; 3-Bachelor; 4-Master; 5-Doctor;

**EducationField**

**EnvironmentSatisfaction**: 1-Low; 2-Medium; 3-High; 4-Very High;

**JobSatisfaction**: 1-Low; 2-Medium; 3-High; 4-Very High;

**MaritalStatus**

**MonthlyIncome**

**NumCompaniesWorked**: Number of companies worked prior to IBM

**WorkLifeBalance**: 1-Bad; 2-Good; 3-Better; 4-Best;

**YearsAtCompany**: Current years of service in IBM

**Analysis Task:**

* Import attrition dataset and import libraries such as pandas, matplotlib.pyplot, numpy, and seaborn.
* Exploratory data analysis.
  + - * Find the age distribution of employees in IBM
      * Explore attrition by age
      * Explore data for Left employees
      * Find out the distribution of employees by the education field
      * Give a bar chart for the number of married and unmarried employees
* Build up a logistic regression model to predict which employees are likely to attrite.

**Assessment:**

(Note: For code refer the attached notebook file in PDF format.)

**Import libraries and dataset:**

* Import libraries such as pandas, matplotlib.pyplot, numpy, and seaborn.
* Import the dataset in to ‘dataframe’ using ‘pd.read\_csv()’ function.

**Exploratory data analysis:**

* View the first five observations using head ().
* View the column names using ‘dataframe.columns.values’
* View the concise summary of the dataframe using dataframe.info ().
* View the Shape of the dataframe using ‘dataframe.shape’.
  + The dataframe has 1470 rows and 13 columns.
* The Attrition dataset has 1470 observations with 13 variables. Out of the 13 variables, there exists one target variable 'Attrition' with possible outcomes Yes and No. The other 12 variables are independent variables.

**Age distribution of Employees:**

* Plot a histogram for Age distribution of Employees.
* Use the ‘dataframe['Age'].hist(bins=70)’.

**Explore attrition by age:**

* Using plt.scatter(dataframe.Attrition,dataframe.Age, alpha=0.1), plot a scatter plot to explore the accumulation of attrition by age.
* The Scatter plot shows most of the attritions are centered at the age of late 20’s and early 30’s of employees.

**Explore data for Left employees:**

* The dataset is well organized with no missing values. This is a dataset of Binary Classification Problem, so the Distribution of instances among the 2 classes is visualized using a bar plot.
* The plot shows the distribution of Attrition. Out of the total of 1470 observations 1233 is No, whereas 237 is Yes. We will treat this imbalance after splitting the data into Training and Test Set.

**Find out the distribution of employees by the education field:**

* Using below code, ‘dataframe.EducationField.value\_counts().plot(kind='barh',color='g',alpha=.65)’, plot the distribution of employees by the education field.
* Using below code,

‘pd.crosstab(dataframe["EducationField"],dataframe["Attrition"]).plot(kind="barh",stacked=True)’, plot the Attrition by Education Field.

* Life Sciences and Medical fields have the highest number of employees and highest number of attrition rate. The percentage of employees who have attrition against those who have been retained seems to be approximately same in all the education fields.

**Give a bar chart for the number of married and unmarried employees:**

* Explore the data for marital status by plotting a bar chart using ‘dataframe.MaritalStatus.value\_counts().plot(kind='bar',alpha=.5)’
* Find the Attrition by Marital Status using , ‘pd.crosstab(dataframe["MaritalStatus"],dataframe["Attrition"]).plot(kind="bar",stacked=False)’.
* From the plot, we can see t**he highest attrition is well correlated to ‘Single’ followed by ‘Married’ & ‘Divorced.**

**Build up a logistic regression model to predict which employees are likely to attrite:**

* Perform the data preprocessing.
* Replace the Attrition ‘Yes’ and ‘No’ to ‘’1’ and ‘0’ respectively.
* In EducationField column, replace

'Life Sciences' as ‘1’,

'Medical' as ‘2’,

‘Marketing' as ‘3’,

'Other' as ‘4’,

'Technical Degree' as ‘5’,

'Human Resources' as ‘6’.

* In Department Column, replace

'Research & Development' as ‘1’,

'Sales' as ‘2’,

'Human Resources' as ‘3’.

* Change the data type of all columns to ‘int64’.

**Train and Test the data:**

* Import the logistic regression model from sklearn library.
* Split the dataset into testing and training data.
* Fit the training data into the logistic regression model.
* Find the training accuracy score using ‘model2.score(X\_train,y\_train)’.
* The training Accuracy of 84.15% is achieved by the model.
* Find the Validation accuracy score using ‘model2.score(X\_test,y\_test)’.
* The validation accuracy of 84.35% is achieved by the model.

**Build and Evaluate Model:**

* Generate the Classification Report of the model using,

‘metrics.classification\_report(y\_test, predicted)’.

* Print the precision, recall and f1-score.

**Conclusion:**

* According to the Performance Analysis, it can be concluded that the Machine Learning Predictive Model has been successful in effectively classifying 84.35% unknown (Validation Set) examples correctly and has shown quite descent statistical figures for different performance metrics.