**HR Analytics Project- Understanding the Attrition in HR Classification**

**Problem Definition:** Every year a lot of companies hire a number of employees. The companies invest time and money in training those employees, not just this but there are training programs within the companies for their existing employees as well. The aim of these programs is to increase the effectiveness of their employees. But where HR Analytics fit in this? and is it just about improving the performance of employees?

**HR Analytics**

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead, it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes.

**Attrition in HR**

Attrition in human resources refers to the gradual loss of employees overtime. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees.

How does Attrition affect companies? and how does HR Analytics help in analyzing attrition? We will discuss the first question here and for the second question, we will write the code and try to understand the process step by step.

**Attrition affecting Companies**

A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

**Data Sources:The training data for this project are available here:**<https://github.com/dsrscientist/IBM_HR_Attrition_Rate_Analytics>

**About the Dataset: We will understand all features from the dataset.**

A screenshot of a computer

Description automatically generated with medium confidence

HR-Employee-Attrition dataset having 1470 rows and 35 features. Where **Attrition** is the resultant feature

Features names are as follow.

A screenshot of a computer

Description automatically generated with medium confidence

Dataset contains Categorical and Numericle type data.

Dataset contains following Categorical data

|  |  |
| --- | --- |
| **Attrition** | Employee Leaving The Company (0=No, 1=Yes) |
| **Business Travel** | (1=No Travel, 2=Travel Frequently, 3=Tavel Rarely) |
| **Department** | (1=Hr, 2=R&D, 3=Sales) |
| **Education Field** | (1=Hr, 2=Life Sciences, 3=Marketing, 4=Medical Sciences, 5=Others, 6= Tehcnical) |
| **Gender** | (1=Female, 2=Male) |
| **Job Role** | (1=Hc Rep, 2=Hr, 3=Lab Technician, 4=Manager, 5= Managing Director, 6= Reasearch Director, 7= Research Scientist, 8=Sales Executieve, 9= Sales Representative) |
| **Marital Status** | (1=Divorced, 2=Married, 3=Single) |
| **Over 18** | (1=Yes, 2=No) |
| **Overtime** | (1=No, 2=Yes) |

Dataset contains following Numericle data

|  |  |
| --- | --- |
| **Age** |  |
| **Daily Rate** | - Salary Level |
| **Distance From Home** | - The Distance From Work To Home |
| **Education** |  |
| **Employee Count** |  |
| **Employee Number** | - Employee Id |
| **Enviroment Satisfaction** | - Satisfaction With The Enviroment |
| **Hourly Rate** | - Hourly Salary |
| **Job Involvement** | - Job Involvement |
| **Job Level** | - Level Of Job |
| **Job Satisfaction** | - Satisfaction With The Job |
| **Monthly Income** | - Monthly Salary |
| **Monthy Rate** | - Monthy Rate |
| **Numcompanies Worked** | - No. Of Companies Worked At |
| **Percent Salary Hike** | - Percentage Increase In Salary |
| **Performance Rating** | - Erformance Rating |
| **Relations Satisfaction** | - Relations Satisfaction |
| **Standard Hours** | - Standard Hours |
| **Stock Options Level** | - Stock Options |
| **Total Working Years** | - Total Years Worked |
| **Training Times Last Year** | - Hours Spent Training |
| **Work Life Balance** | - Time Spent Bewtween Work And Outside |
| **Years At Company** | - Total Number Of Years At The Compnay |
| **Years In Current Role** | -Years In Current Role |
| **Years Since Last Promotion** | - Last Promotion |
| **Years With Current Manager** | - Years Spent With Current Manager |

The dataset contains the details of the employees who are working in an organization. The dataset contains both dependent and independent variables and also contains both categorical and numerical data. In this dataset "**Attrition**"" is our target variable which has two classes. So this is a "**Classification type**" problem in which we need to increase the attrition of the employees.

**Exploratory Data Analysis**

Above details features details we get the datatypes of features. This gives the information about the dataset which includes indexing type, column type, no-null values and memory usage.

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

There are two unique values in the label 'Yes' and 'No'. We can say Yes means the employess who are facing attrition and No means employees who are not facing any attrition.

Graphical user interface, text, application

Description automatically generated

This gives the list of values in the label. As we can see 237 employees facing attrition and 1233 employess are not facing any attrition. There is a data imbalancing issue in this dataset, will make it balance later Nunique methode, we an see number of unique contain present in each feature.

#### Text Description automatically generated with medium confidence

Since the columns EmployeeCount, Over18 and StandardHours have only 1 count so they can be dropped since they won't affect our model. Also EmployeeNumber is taken on the basis of unique ID of the employees which does not helps so we can drop this too.

#### **Detect the missing values**

The dataset is well organised with no missing values we can see with isnull().sum() function and with heatmap graph.

|  |  |
| --- | --- |
|  |  |

**We will use describe() method for calculating some statistical data like percentile, mean and std of the numerical values of the Series or DataFrame. As our dataset having both numeric and object series and also the DataFrame column sets of mixed data types.** Describe methode uses columns contain continuous type of data

## **A picture containing graphical user interface Description automatically generated**

The summary of this dataset looks perfect since there is no negative or invalid values present.

We can observe the following things.

* The counts of all the columns is same for all features, no missing values present in the data.
* The mean is more than the median(50%) in most of the columns which means they are skewed to right. The min age of the employee is 18 and max is 60 and most of the employees are in between 36.
* In few columns the median 50% is more than the mean which means they are skewes to left.
* Some of the columns have huge difference in mean and the standard deviation.

Correlation heatmap is graphical representation of correlation matrix representing correlation between different variables.

A screenshot of a computer

Description automatically generated with medium confidence

Using above Heat map corr() observation is Over18 is blank StandardHours, EmployeeCount, EmployeeNumber having all NAN, so better we will drop those columns.

Also in HR Analytics Project we can see Joblevel, MonthlyIncome, Age, TotalWorkingYears, YearsatCompany, YearsIncurrentRole, YearssinceaLastPromotion these are features are positively co realated with

Statisticle **Univariate Analysis** with distplot, from Univariate observation we can see that data dataset contains both continous and categoricle data. Data is not normally distributed very skewed distance\_from\_home, monthly\_income, number\_of\_compnies\_worked, Prcentage\_salary\_hike, job\_satisfaction. Age, Dailyrate, EmployeeNumber, HourlyRate, MonthlyRate are mostly normally distributed.

Diagram, shape

Description automatically generated

*Fig. Univariate Analysis with distplot*

**Bivariate Analysis** we can do analysis with atrition label.

Interpreting Relationship between Dependent Variable and Independent Variables

Timeline

Description automatically generated

*Fig. Bivariate Analysis with countplot*

**Multivariate Analysis** using pairplot.

* This pair plot gives the pairwise relation between the columns which is plotted on the basis of target variable "Attrition". Here we can observe the relation between the features and label.
* Some of the features have strong linear relationship and most of features are highly correlated with each other.
* Some of the features have outliers and skewness, will remove them later.

A picture containing shoji, building, tiled

Description automatically generated

*Fig. Multivariate Analysis with pairplot*

**Identifying the outliers**

We will check for are any outliers present with Boxplot. With boxplt we can see many columns are having outliers presentin MonthlyIncomeme, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager. We will remove these outliers using either Zscore or IQR mathod in the further steps

Text, calendar

Description automatically generated

*Fig. Identifying the outliers with Boxplot*

**Removing outliers by Zscore and IQR Methode**

Text

Description automatically generated

Dataloss of 5.6% with zscore.

Graphical user interface, text, application

Description automatically generated

Dataloss of 56% with IQR which is very high.

We removed outliers with dataloss of 5.6%, which is less than 6% using zscore.

Table

Description automatically generated with medium confidence

**yeo-johnson method**

Removed the skewness using yeo-johnson method.

Diagram, engineering drawing

Description automatically generated

The looks normal compare to the old data and the skewness is also reduced. Now our data is cleaned.

Visualizing the correlation between label and features using bar plot

Chart

Description automatically generated

Here we can notice the columns BusinessTravel and HourlyRate have very less corrrelation with the target. Lets drop those columns

# Model Preparation

## For model preparation we will Separate the features and label variables into x and y

# A picture containing icon Description automatically generated

# Standard scaler

## Scaled the data using standard scalarizaion method to overcome with the issue of data biasness.

## Text Description automatically generated

Encoding the categorical columns using label encoder

Table

Description automatically generated

Check the count of label feature.

Graphical user interface, text, application, email

Description automatically generated

Here we can notice that the data is not balanced. Let's use oversampling method to balance the data.

**Handle Imbalance Data** **with** **SMOTE**

## **To Handle Imbalance Data.**This two-class dataset is imbalanced (84% vs 16%). As a result, there is a possibility that the model built might be biased towards to the majority and over-represented class. After applying Synthetic Minority Oversampling Technique (SMOTE) to over-sample the minority class, some improvement in both F1-score & Recall can be observed. Class Imbalance issue present so we will apply SMOTE to fix this isssue

Graphical user interface, text, application, email

Description automatically generated

Now we can see y\_1 is balaced properly.

# Training and Testing Data

Separate data into a training set and a test set . This is a very standard approach in Machine Learning. The random\_state parameter is simply a seed for the algorithm to use (if we didn't specify one, it would create different training and test sets every time we run it) **Find for which state we are getting best accuracy with DecisionTreeClassifier. Below we can see at 122 random state we are getting best accuracy score 85%.**

**Now with 122 random state we done train test split for taining and testing data.**

**Graphical user interface, text, application

Description automatically generated**

**Building Machine Learning Models:**

**Prepare model with algorithms LogisticRegression, SVC, KNeighborsClassifier, RandomForestClassifier, DecisionTreeClassifier, GaussianNB, AdaBoostClassifier, ExtraTreesClassifier, XGBClassifier.**

**Text

Description automatically generated**

We can also observe the details of all algorithm with Acuracy score, confusion matrix (which contains True-Positive, False-Positive rate and False-Negative, True -Negative rate) and with classification report (which conatins precision, recall, f1-score, support). Details of ExtraTreesClassifier are as follow

**Table

Description automatically generated**

**Cross Validation Score**

**We will do cross validation** Cross-validation is usually the preferred method because it gives your model the opportunity to train on multiple train-test splits. This is tells us a better indication of how well the model will perform on unseen data. Following are the Cross Validation Score of all models.

**Table

Description automatically generated**

The **Area Under the Curve** (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The AUC is Higher, performance of the model is better at distinguishing between the positive and negative classes.

The AUC value lies **between 0.5 to 1** where 0.5 denotes a bad classifer and 1 denotes an excellent classifier. All models are giving AUC value more than 0.5, we need to find which is the best model among all.

**Text

Description automatically generated**

**Chart, line chart

Description automatically generated**

**Here with following comparision table of Test Score,Cross Validation Score,ROC AUC Score**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Model** | **Test Score** | **Cross Validation Score** | **ROC AUC Score** |
| **1** | LogisticRegression | 0.719828 | 0.849331 | 0.719345 |
| **2** | SVC | 0.693966 | 0.834896 | 0.691071 |
| **3** | KNeighborsClassifier | 0.771552 | 0.817606 | 0.776488 |
| **4** | RandomForestClassifier | 0.905172 | 0.850035 | 0.904911 |
| **5** | DecisionTreeClassifier | 0.793103 | 0.770005 | 0.794048 |
| **6** | GaussianNB | 0.773707 | 0.783674 | 0.776042 |
| **7** | AdaBoostClassifier | 0.840517 | 0.870221 | 0.842411 |
| **8** | ExtraTreesClassifier | 0.93319 | 0.854372 | 0.933036 |
| **9** | XGBClassifier | 0.907328 | 0.860136 | 0.907887 |

**We can observe following,**

LogisticRegression and SVC test score is less but cross validation score is incresed and diffeerence is more.

KneighborsClassifier and GaussianNB less difference in test score and in cross validation score

RandomForestClassifier, ExtraTreesClassifier, XGBClassifier having high test score and cross validation is less than test test score difference is less.

DecisionTreeClassifier cross validation score is less than test score and difference is also less.

So we can say than AdaBoostClassifier ,KNN, GaussianNB,DTC are having overall less differenece .

But ExtraTree giving best test score.

# Hyper parameter tuning

So we will do Hyper parameter tuning for SVC, KNN, DecisionTreeClassifier, ExtraTreesClassifier and AdaBoostClassifier, RandomForestClassifier and with Accuracy Score we will conclude which is best model for HR Analytics prediction project.

After Hyper Parameter Tuning, Random Forest is giving best acuracy score 91.0% and F1-score 91%

Extra Tree Classifier is giving best acuracy score 93.10% and F1-score 92.85%.

So we conclde that Extra Tree Classifier is the best model for HR Analytics prediction project. Following is the ROC Curves for comparision of all models after Hyper parameter tuning

*Fig. ROC Curves for comparision*

# Chart, line chart Description automatically generated

# Model Saving

Graphical user interface, text, application

Description automatically generated

# Conclusion::Best Model

# So we will choose Extra Trees classifier as best model

Graphical user interface

Description automatically generated with medium confidence

We can observe both original and predicted attrition values are same. Conclusion is **Extra Trees classifier as best model.**

**Thank you**