HR Analytics Project

**Problem Definition :**

**Understanding the Attrition in HR Analytics and predict Attrition.** 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

**Data Analysis :**

Hr analytics data is classification problem statement. This data predict Attrition. Data have object and integer data type present . data have no null and duplicates value present. Data have 1470 rows and 35 columns.

Features name : Label name :

**Attrition**

**Age**

**BusinessTravel**

**DailyRate**

**Department**

**DistanceFromHome**

**Education**

**EducationField**

**EmployeeCount**

**EmployeeNumber**

**EnvironmentSatisfaction**

**Gender**

**HourlyRate**

**JobInvolvement**

**JobLevel**

**JobRole**

**JobSatisfaction**

**MaritalStatus**

**MonthlyIncome**

**MonthlyRate**

**NumCompaniesWorked**

**Over18**

**OverTime**

**PercentSalaryHike**

**PerformanceRating**

**RelationshipSatisfaction**

**StandardHours**

**StockOptionLevel**

**TotalWorkingYears**

**TrainingTimesLastYear**

**WorkLifeBalance**

**YearsAtCompany**

**YearsInCurrentRole**

**YearsSinceLastPromotion**

**YearsWithCurrManager**

**Exploratory Data Analysis (EDA) :**

First import python liebraries and import data using read\_csv() function.

info() : check information of data. Data have no null value present and check data types. Object and integer data type present .

isnull().sum() : check null values present in features. No null values present in data.

Duplicated().sum() : check duplicates values in data. No duplicate value present in data.

Value\_counts() : check whitespace and unique value count. No whitespace present and check unique value count. Drop three feature over18,employeenumber and employeecount because not make sence and showing multicollinearity problem.

Sperate object and numerical data for convert object data to integer data type.

Convert object to integer data type usning LabelEncoder. Then we check statastices usning describe() method.

Seprate continuous feature for check statustics.

Describe() : check data statastics, check mean, standard deviation , min , median, max, outlier and data skewness. Monthly income, total working year , year at company, year in current role, years sinces last promotion and year curr manager have outlier and skewness present

Skew() : check skewness value if skewness value close to 0 then no skewness present and if skewness value more than 0 then like 1 then skewness present

Boxplot() : usning graph we check outler. Outler present in Monthly income, total working year , year at company, year in current role, years sinces last promotion and year curr manager.

Distplot() : uning graph we check data is normal distributed or not. Age, daily rates, distance form home, hourly rate, monthly rate, and percent salary hike have normal distributed data and reamaning continuous data is not normal distributed.

Barplot() : check relation with label usning barplot how attrition will aftected. Also use sctterplot(),stripplot() and regplot()

Regplot(): age, daily rate, education, environment satisfaction, jobinvolment, joblevel , job satisfaction, monthly income, relationship satisfaction, stock opion level, total working year, training times last year, work life balance, year at company, year at current role and year with manager this feature is correlated with label when feature level is increase then attrition is decreases. example age feature when age is increase then attrition is less. Over time, department and marital status feature is when feature level is increase then attrition is increase. example when overtime feature is increase then attrition is increase. Reaming feature is nutral not increase and not decrease.

Standard hour feature is delect because this feature not netural , possitve and negative its nothing contribute to label.

We keep outlier in normalized form because most of feature present more outlier when we delect outlier then data loss occure that why we keep outlier using power transform method.

Powertransfomer() : we keep outliers in normalized form, also check outler using boxplot . year at company present outler then we delect outlier using inter quantile range method

We seprate feature and label for check relation in feature and label.

Corrwith(): Check relationship feature and label usning corrwith method. Performancerating , hourly rate and daily rate is less contribute with label. Department , distance from home, educationfeild, gender, jobrole , marital status, monthly rate, numcompanies worked and overtime is positive relation with label and remaining is negative relation with label.

dailyrate , hourlyrate, performancerating and joblevel is less contribute with label also check with heatmap.

Heatmap() : check multicollinearity problem and relationship using heatmap.

Year at company feature collinearity with year in current role.

Year at company feature collinearity with year with curr manager.

Also we check multicollinearity problem using variance inflation favtor score.

Job level feature collinearity with monthly income.

Hourly rate, performacerating and business travel is less contribute to label.

Variance\_infation\_factor (): Before use variance\_inflation\_factor() method standardized the data usning standardscaler(). We standardized only features data. Use standardscaler to standardized the features data and convert in dataframe. Than use variance\_inflation\_factor in features to check multicollinearity problem. Consider vif score =5. if VIF score >5 then multicollinearity problem and VIF < 5 then no multicolinearty problem.

Job level, monthly income and year at company features have more than 5 score then multicollinearity issue is there.

Delect job level and year at company feature for remove multicollinearity problem.

Also delect features which is less contribute with label dailyrates, hourly rates, monthy rates and performance rating. Again check vif score to confirm for mutlicolinearity problem remove or not.

Check data balanced or not usning countplot(). Data is imblanced, balanced data usning SMOTE method over sampling method. After balanced data build models.

Check underfitting and overfiltting usning cross validation techniques. After check select best model. The best model is AdaBoostclassifier(). Increase the accuracy using hyperparameter tuning.

Accuracy increase one percent.

Save the model and load the model

**Pre-processing Pipeline :**

Import Libraries

Import dataset

Check nulls value

Check duplicates value

Check whitespace and unique value

Check unique value

Seprate object and numric data

Convert object data to integers data

Check statastics

Check outlier and treat outlier

Check multicollinearity problem and remove multicollinearity problem

Check data balanced

Builds models

Check underfitting and overfitting

Select best model

Hyperparameter tuning

Save the model

Load the model

**Building Machine Learning Models :**

* LogisticRegression model:
  + - * + Accuracy is 80%
        + Type 1 erorr is 62
        + Type 2 erorr is 11.
* KneighborsClassfier model :
  + - * + Accuracy is 76%.
        + Type 1 erorr is 75.
        + Type 2 erorr is 12.
* BaggingClassifier model :
  + - * + Accuracy is 86%.
        + Type 1 erorr is 24.
        + Type 2 erorr is 26.
* RandomforestClassifier model :
  + - * + Accuracy is 88%.
        + Type 1 erorr is 12.
        + Type 2 erorr is 31.
* AdaBoostClassifier model :
  + - * + Accuracy is 86%.
        + Type 1 erorr is 28.
        + Type 2 erorr is 21.
* GradientBootingClassifier model :
  + - * + Accuracy is 89%.
        + Type 1 erorr is 10.
        + Type 2 erorr is 29.
* DecisionTreeClassifier model :
  + - * + Accuracy is 76%.
        + Type 1 erorr is 59.
        + Type 2 erorr is 27.
* SVC model :
  + - * + Accuracy is 86%.
        + Type 1 erorr is 25.
        + Type 2 erorr is 24.

Check underfitting or overfitting using cross validation technique

* LogisticRegression model :
  + - * + Cross validation mean score is 76%
        + ( Accuracy score – cross validation score )
        + 80% - 76% = 4%
* KneighborsClassfier model :
  + - * + Cross validation mean score is 83%
        + ( Accuracy score – cross validation score )
        + 76% - 83% = -7%
* BaggingClassifier model :
  + - * + Cross validation mean score is 86%
        + ( Accuracy score – cross validation score )
        + 86% - 90% = -4%
* RandomforestClassifier :
  + - * + Cross validation mean score is 88%
        + ( Accuracy score – cross validation score )
        + 88% - 93% = -5%
* AdaBoostClassifier :
  + - * + Cross validation mean score is 87%
        + ( Accuracy score – cross validation score )
        + 86% - 87% = -1%
* GradientBootingClassifier :
  + - * + Cross validation mean score is 90%
        + ( Accuracy score – cross validation score )
        + 89% - 90% = -1%
* DecisionTreeClassifier :
  + - * + Cross validation mean score is 83%
        + ( Accuracy score – cross validation score )
        + 76% - 83% = -7%
* SVC model :
  + - * + Cross validation mean score is 90%
        + ( Accuracy score – cross validation score )
        + 86% - 90% = -4%

The best model is AdaBoostClassifier and GradientBoostingClassifier . we select only one model. Select AdaBoostClassifier and increase the accuracy usning hyperparameter tuning.

Applying hyperparameter tuning increase one percent accuracy 87%.

And reduced type 1 erorr is 0 , type 2 erorr is 47.

Model build sucsessfuly. Lastly save the model with filename HR dumb method from pickle librabrie.

**Concluding Remarks :**

Increase the attrition because employee overtime more and monthly income is less , percent hike salary is less that’s why not statisfied and he will leave company. Also which employee home is long distance from company then he will be leave.

Reduced the attrition then reduced employee overtime, increase monthly income, increase percent hike and better job environment.