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

**Problem Definition**

Every company have a HR Department (Human Resource) Department. Here **HR analytics** means companies hire new employees. Company want to give training newly hired employees. For effectiveness in their performance, company also want to give training for their existing employees. This is highly effort and challenge for company. HR analytics is the field of analytics area for applying analytics processes to the HR Department of an organization in the hope of improving the performance of employees. HR analytics provide insight to each process by gathering data and using to make relevant decisions about how to improve these processes.

**Attrition in HR**

High attrition in employees affect organization in many ways. HR professionals plays a major role in designing company compensation programs, work culture and motivation system that helps organization retain employees.

We mention that attrition affects companies in many ways. High attrition in employee gives costs to an organisation. Company wants to hire new employees, wants to give training for employees this will affect lose in company’s costs and times. Company loses an experienced employee. A new worker does not be a perfect in first days in work. They have errors and issues while working because they have no experience. This is all company affect attrition in employees. Now let’s check how HR analytics help in employee attrition by building machine learning model. As usual collect data, analyse data and predict data using machine learning model.

As we want predict the employee attrition or not, we want build a classification model.

**Data Analysis**

We collect some raw data, now we want to analyse the raw data. Dataset have some features and a target variable. In this process we want to clean data by treating missing values, encoding etc.. Dataset is small, around 1470 rows were present in dataset. When dataset have low data, accuracy of dataset is also low it is because more data gives more accuracy if machine have more data to learn it gives more accuracy. So let’s analyse the data in dataset for our project.

Dataset have integer, categorical data type features. As our target variable is categorical we want to encode those features and target variable for machine to understand. When we check the missing values it is not showing any missing values. When we look at the statistical summary of dataset, maximum age of the employee shows 60 and minimum age of employee is 18. Maximum 40 years an employee work in company. Each year employee got trained at least one time. These are some statistical summary of features of integer data type. Now we can look on dependent variable, most of the employees are not attrition, using count plot we can visualize the count of dependent variable. And also we can observe the imbalance of the dependent variable we want to balance it.

Now we can check how other features impact in dependent variable. As our dependent variable is categorical we can compare with categorical features. When we compare with one of the feature ‘gender’ most of the employees were male also most of male were attrition. We can conclude that most of male employees are attrition. Another feature is ‘Job role’ lot of job roles have employees, most of number of employees have the ‘Sales executive’ job role and also most of them were attrition in compare to other job roles. When we check the marital status of employees, most of the employees were get married. But the employees were not married have high attrition rate. While divorced employees have low attrition. So a married and divorced employees have low attrition rate. When we go through the employees ‘Over time’ jobs most of them were not working overtime, as compare to overtime workers the attrition rate of employees were low in employees who working for overtime. So we can conclude that giving employees overtime work results high attrition rate. No let’s compare the age of employees. When we compare age with attrition most of the attrition rate were in age of 28 to 35. The employees age 28 to 35 were have high attrition rate.

Now we can compare the dependent variable with the numerical features. We already compare age with target variable. We can compare daily rate with attrition rate. Attrition and daily rate are categorical and integer data type. To compare numerical and categorical data we can use pandas **Pivot Table.** By comparing using pivot table we can observe employees have high daily rate have low attrition rate. Comparing with distance from home feature, increase in distance from home results high attrition rate. We can conclude that daily travelling to company from home results high attrition rate, Employees gets tired and become weak. Comparing with environment satisfaction, good environment satisfaction gives low attrition rate, it gives good mental ability for employees and results low attrition rate. Hourly rate feature gives equal in attrition rate. It means having high hourly rate gives same number of attrition rate. When employee gets high rate in job involvement it gives high attrition rate. Here we can conclude that giving more jobs to employee results high attrition rate. Comparing job level of employee with attrition rate, high job level results low attrition rate. We can conclude that high job level gives low number of attrition rate. When an employee get satisfaction in his job it gives low attrition rate. So satisfaction in job is a major factor in attrition rate. When we come to monthly income, high monthly income results low attrition rate. So giving employees high salary gives low attrition rate. Daily rate and monthly rate are same. Monthly rate is calculated using daily rate, when monthly rate increases attrition rate also increases in our dataset. Comparing number of companies worked by employees with attrition rate. Here, higher number of companies worked results high attrition rate. Percent salary hike does not gives a best results in attrition rate. That means it gives equal number of attrition. And also performance rating also gives equal number of attrition rate. Satisfaction in relationship gives low attrition rate. Here relationship means relation with co-workers, satisfaction with colleagues gives low attrition rate. Now compare the stock level of company which with employees, higher number of stock level gives low attrition rate. Employees who have worked more years gives low attrition rate. So keeping employees more years results low attrition rate. So always keep the employees by giving benefits, it helps to low attrition rate. Always train employees to low the attrition rate. Training gives employees make perfect. Also gives promotion for employees to results low attrition rate. So we compare features with target variable.

**Correlation**

Now let’s look how features where correlated to each other. We can visually check the correlation using heat map.

Job level and total working years are correlated. It makes sense that, total working years or we can say that experience of employee gives higher job level. It means high number of working years gives high job level. Also these are related to age, when experience increases age also increases and job level also increases.

Years at company, years with current manager, Years since last promotion, Years with current role these features are correlated to age. This make sense, when the Years increases age also increases. And also these features are correlated to job level. Experience increases job level also increases. Total working years also correlated to the above mentioned features

So these features are needed for building machine learning model.

**Data Cleaning**

In dataset, there is no missing values. So we don’t need to treat it. But there is some features which is not impact to target variable and not help for prediction, so we can drop it from the dataset. Those features are ‘Over18’, ‘Employee Number’, ‘Standard Hours’, ‘employee Count’. Here, all employees age is 18+ so we don’t need over 18 feature. And employee number will not affect, it is unique number for all employees. All Employees standard hours of working are same, the count employee does not need for prediction. So we drop columns which not needed.

Then we encoded data having categorical variables using label encoder. Scaled features using any scaling technique. Balanced target variable using smote.

**EDA Concluding Remarks**

EDA (Exploratory Data Analysis) is one of the most important part of machine learning. We have compare every features with dependent variable of dataset. Some features affects the target variable. Before building machine learning models EDA says about the data, gives a detailed view of data.

In HR analytics, we check how the data are related to each other, how other features affect in attrition of employee. In above data analysis, I can conclude and in my point of view I understand that for low attrition rate, organization try to give training to employees and make a good working environment for them and try to keep the employees giving benefits, promotion to them. When a company gives these basic features for employees, organization can low rate of attrition.

**Pre-processing Pipeline**

There are many pre-processing techniques. Here we use pre-processing techniques for encoding and feature scaling. Using label encoder we encode categorical features. Feature scaling is one of the most important pre-processing technique while building machine learning. It helps machine for better understanding features and helps to gives best accuracy score. There are many encoding techniques label encoder, ordinal encoder, one hot encoding. But here we use label encoding technique for categorical data. Here we balanced the target variable using smote (Over Sampling).

**Building Machine Learning Models**

Before building machine learning model, we want separate our target variable and features. And wants to split model into train and test model. Here I separated features and target and split model into train and test model. 25% of data were used for testing and rest of 75% of data used for training. Now we can check the best random states. We can check at which random state has best training accuracy and testing accuracy. We want to select best random state which has equal number accuracies in test and train. This is used for avoiding under fitting and over fitting. By selecting best random state, apply it on model by fitting train data and predicting test data. After this print accuracy score, classification report gives f1 score, precision, recall and accuracy score of the model. To select best most model I applied four models here to check which one is best. Which model gives good Accuracy. Logistic regression, Decision tree classifier, Bernoulli NB, KNeighbors classifier. Among these also ensemble techniques like random forest and ada boost classifiers. These model gives e a good accuracy score as well. For selecting final model, we want to compare accuracy score and cross validation score of our models. Among these models which have equal CV score and accuracy score, we can select that model as our final model. So by checking it I got ada boost classifier as my final model. Reason I have mentioned early, this model have equal CV score and accuracy score.

**Hyper Parameter Tuning**

Tuning a model gives best parameters for our model and which parameters gives best accuracy. There are two types of hyper parameter tuning **GridSearchCV, RandomSearchCV.** Grid search CV is better than Random search CV. So here I choose grid search CV for tuning. Gives parameters of our final model **Adaboost classifier.** And gives CV score equals to 8. After this, fitted train data for grid search. After that select best parameters. We can apply the best parameters for adaboost and fit the train data and predict the test data. After giving best parameters, get the accuracy of 92% which is good.

AUC ROC Curve

Higher the Area Under the curve it is the better performance of the model. Here when I check the AUC of my final model it gives higher area under the curve.so my model is having best performance.

After all these steps we want to save our model in object file using joblib or pickle.

**Concluding Remarks**

As we get the accuracy score of my model is 92%. It gives predicted and original value is same.