## Article on Evaluation Project 3 HR Analytics

## Problem Statement

##### 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 employee’s 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.

**Project pathway:**

* Firstly we imported the csv file in the pandas library using pd.read\_csv( ) method.
* The dataset contains 1470 rows and 35 columns.
* From the Observation of dataset we concluded the following:

1. In this HR dataset we have 1470 rows and 35 columns.
2. Non-null count is same for all Columns, so it seem that it contain No missing value. Still we need to perform Data integrity Check for null values in form of "-","NA" , any duplicate entry or error in Data.
3. Out of 35 we have 9 features with Object datatypes and rest are int64 types
4. Among all Numeric Variables 'Education','EnvironmentSatisfaction', 'JobInvolvement', 'JobSatisfaction', 'RelationshipSatisfaction', 'PerformanceRating', 'WorkLifeBalance' are ordinal variable. Unique range of all these ordinal Variable need to check.
5. Here We have Target Variable 'Attrition’

**Statistical Analysis :**

From the statistical Information we came to the following observations:

* Minimum Employee Age is 18 and Maximum age of employee 60.
* Average distance from home is 9.1 KM. It means that most of employee travel atleast 18 KM in day from home to office.
* On Average performance Rating of employees is 3.163 with min value 3.0. This Means that performance of most of employee is 'Good’. This implies that Attrition of Employee with 'Outstanding' or 5 rating need to investigate.
* 50% of Employees has worked at least 2 companies previously.
* For Monthly Income, Monthly Rate by looking at 50% and max column we can say outliers exist in this feature.
* By looking at Mean and Median we see that some of the features are skew in nature.
* For ordinal features statistical terminology of mean, median, std deviation doesnot make sense.
* StandardHours and EmployeeCount contain same value for all statistical parameter. It means they contain one unique value.

From the data visualization we came to following observations and conclusions:

**In Education columns:**

* More than 60 % employees educated at Masters & Bachelor. It interesting to find out in which department need this human resources.
* 30 % of Employees are highly educated which involves master and doctor degree.
* 39 % of Employees are graduate.
* Almost 19% Employees are educated upto college & 12% are below college.

**In department column:**

* 65.37% of Employees belong to Research & Development Department. Out of Total 961 Employee no of employee educated at Bachelors, Masters, Doctor are 379,255 and 30 respectively.
* Only 63 Employee work in HR department.

**In Education field:**

* 41.22 % Employee comes from Life science background followed by Medical profession with 31.56%.
* There are only 27 people with HR background and We know that 63 people work in HR Department from previous result. This implies that at least half employee working in HR department do not have HR background. This important as there is more probability of Employees Retention is when they are working in domain of interest or as per their education background. Dissatisfaction with want we doing can be seen as major reason of leaving job.
* Most of Employees with Technical degree are Bachelors.
* Most of Employees having Masters and Doctors belong to Life Science and Medical domain.
* R&D department almost everyone comes from profession or technical background except support staff. Factor like Salary Hike, travelling, overtime and Job level are things need to taken in consideration while analysing Attrition of this category.
* There are 159 Employee with Marketing background and all work in Sales Department.
* 50% Employees in sales department have background of Life sciences & Medical. So it will interesting to see attrition rate in these employees.

**In jobrole column:**

* There are 3 job role in HR Department, maximum of which are sales Executive with 446 Total Employees.
* Human Resources department has 2 Job role i.e. HR & Manager.
* There 6 different Job role in R&D department with total 961 employees and until now we know that all of them belong to their respective domain background.

**From attrition as per jobrole:**

* Percentage of attrition is high in Sales Representative, Laboratory Technician, Human Resources. This all job role comes at bottom in corporate hierarchy also Salary is comparatively less compare to other job role.
* Monthly Income, Job satisfaction, travelling are feature need to dive into for further insights in these job role.
* At the Top chart 62 Laboratory Technician has resign from job, followed by 57 sales executive and 47 Research Scientist.
* 16 % attrition rate for Research Scientist, which involve huge investment from company. Company not only loses employee but its knowledge base, expertise & Intellectual property rights in some cases.

**Observations from numeric features:**

* For Majority of people have spend 3 to 10 years at company.
* Most of people staying company up to 2 years after promotion.
* Majority of people are train 2-3 times in last year. If employees leaves job then it loss investment for company.
* Majority of people stay in same role for maximum 4 yrs.
* Majority of Employees have salary hike of 10 to 15%.

**Observations from age and attrition:**

1. The Attrition rate is minimum between the Age years of 34 and 35.
2. The Attrition rate is maximum between the Age years of 29 and 31.

**Preprocessing :**

1. **Outliers detections and removal**

There were some outliers Present in the columns "MonthlyIncome", "NumCompaniesWorked", "PerformanceRating", "StockOptionLevel", "TotalWorkingYears", "TrainingTimesLastYear", "YearsAtCompany", "YearsInCurrentRole", "YearsSinceLastPromotion", "YearsWithCurrManager", "Attrition".

Which we removed using zscore method.

1. **Skewness :**

We transformed the skewed data using power transform method

1. **Correlation:**

We checked for correlation and here are some observations from the correlation heatmap

* Age, JobLevel, MonthlyIncome is highly positively correlated with TotalWorkingYears.
* JobLevel is highly positively correlated with the MonthlyIncome.
* PercentSalaryHike is highly positively correlated with the column PerformanceRating.

1. **Balancing using Smote:**

As the target variable data was imbalanced we balanced it using Smote technique.

**Machine learning and model building:**

For machine learning we imported various machine learning classifiers like

* Logistic Regression
* Gaussian NB
* SVC
* Decision Tree Classifier
* KNeighbors Classifier
* Random Forest Classifer
* Extra Trees Classifier
* AdaBoost Classifier
* Gradient Boosting Classifier
* Bagging Classifier

Firstly, we found the best random state which was 73 at the 87% accuracy

Then we check the accuracy score for each machine learning classifier as the accuracy score was as follows:

* Accuracy Score of Logistics Regression was found to be 0.8718
* Accuracy Score of SVC was found to be 0.9137
* Accuracy Score of GaussianNB was found to be 0.8352
* Accuracy Score of DecisionTreeClassifier was found to be 0.7986
* Accuracy Score of KNeighborsClassifier was found to be 0.8313
* Accuracy score of RandomForestClassifier was found to be 0.9045
* Accuracy Score of AdaBoostClassifier was found to be 0.8562
* Accuracy Score of GradientBoostingClassifier was found to be 0.8888
* Accuracy Score of BaggingClassifier was found to be 0.8614

**Cross validation:**

Then we cross validate the above classification model using Cross\_val\_score. The result were as follows:

* Cross validation score of LogisticReggression was 0.6830 with the std deviation of 0.0349
* Cross validation score of SVC was 0.5945 with the std deviation of 0.0139
* Cross validation score of GaussianNB was 0.7366 with the std deviation of 0.0307
* Cross validation score of DecisionTreeClassifier was 0.8373 with std deviation of 0.0944
* Cross validation score of KNeighborsClassifier was 0.6960 with std deviation of 0.0143
* Cross validation score of RandomForestClassifeir was 0.9154 with the std deviation of 0.1077
* Cross validation score of AdaBoostClassifier was 0.8701 with the std deviation of 0.1304
* Cross validation score of GradientBoostingClaassifier was 0.8848 with the std deviation of 0.1543
* Cross validation score of BaggingClassifier was 0.8895 with the std deviation of 0.1198

So on the basis of maximum score in cross validation of Random Forest Classifier. We choose Random Forest classifier For further hyperparameter tuning.

**Hyperparameter Tuning:**

We tried the Hyper parameter tuning of Random Forest Classifier with the following parameter

{ 'bootstrap': [True], 'max\_depth': [5, 10,20,40,50, None],

'max\_features': ['auto', 'log2'],

'criterion':['gini','entropy'],

'n\_estimators': [5, 10, 15 ,25,50,100] }

The best parameters found to be:

{'bootstrap': True,

'criterion': 'entropy',

'max\_depth': 50,

'max\_features': 'log2',

'n\_estimators': 100}

**Final Model:**

So , we again Build the model using the best parameters. And after that the Accuracy score came to be 0.9084

**Saving the Model:**

After all that the final model is now saved using joblib by joblib.dump( ) method .

**So, this is how the Attrition of the employee is predicted**