**HR ANALYTICS – ATTRITION**

**USING ML**

By Prachi Parmar

**PROBLEM DEFINATION**

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.

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.

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 INFORMATION**

**FEATURES:**

**Age**: Ranges from 18 to 60 Years.

**BusinessTravel**: Travel Rarely, Frequently, Non-Travel.

**DailyRate**: Ranges from 102 to 1499.

**Department**: HR, Sales, Research & Development.

**DistanceFromHome**: Ranges from 1 to 29.

**Education**: 5 categories (1,2,3,4,5) based on no of year of higher education.

**EducationField**: HR, Technical Degree, Medical, Marketing, Life Sciences, Others.

**EmployeeCount**: 1, count of employee is 1 for all rows.

**EmployeeNumber**: Unique key for every employee.

**EnvironmentSatisfaction**: Categories in (1,2,3,4) as of level of satisfaction.

**Gender**: Male/ Female

**HourlyRate**: Ranges from 30 to 100.

**JobInvolvement**: Categories in (1,2,3,4) as of level of involvement.

**JobLevel**: Categories in (1,2,3,4,5) as of level of Job in the Company.

**JobRole**: Sales Executive, Research Scientist, Laboratory Technician, Manufacturing Director, Healthcare Representative, Manager, Sales Representative, Research Director, Human Resources

**JobSatisfaction**: Categories in (1,2,3,4) as of level of satisfaction.

**MaritalStatus**: Married, Single, Divorced

**MonthlyIncome**: Ranges from 1009 to 19999

**MonthlyRate**: Ranges from 2094 to 26999

**NumberCompaniesWorked**: Ranges from 0 to 9.

**Over18**: All rows have Yes value.

**OverTime**: Yes/ No

**PercentSalaryHike**: Ranges from 11 to 25 %

**PerformanceRating**: Categories of 3,4 rating.

**RelationshipSatisfaction**: Categories in (1,2,3,4) as of level of satisfaction.

**StandardHours**: 80 for all rows.

**StockOptioLevel**: Categories of 0,1,2,3 levels.

**TotalWorkingYears**: Ranges from 1 to 40.

**TrainingTimesLastYear**: Categories (1,2,3,4,5,6)

**WorkLifeBalance**: Categories 1,2,3,4

**YearsAtCompany**: Ranges from 0 to 40.

**YearsInCurrentRole**: Ranges from 0 to 18.

**YearsSinceLastPromotion**: Ranges from 0 to 15

**YearsWithCurrentManager** : Ranges from 0 to 17

**TARGET VARIABLE**: *ATTRITION* *(YES/ NO)*

**REMARKS**

* 24 Categorical Feature
* 9 Continuous Features
* 1 Target Variable having 2 target classes Yes/ No
* 1 Unique Key, Employee Number
* Size of Data Set is 1470 rows.

**DATA ANALYSIS**

* Categorical Features Like Over 18 (Yes), Standard Hours (80) and Employee Count (1) has same value for all rows, therefore these features can be dropped.
* Employee Number is unique for all rows, like a primary key or ID Value, this feature can be dropped as this feature is to uniquely identify each row and has no relation with the target Variable.
* Among the Continuous features, Monthly income and Monthly Rate has highest spread of data and have highest standard deviation.
* No missing values are present in the Data.
* Imbalanced Classes of the Target Variable which consist of 83.8 % of No Attrition Data and 16.1% of Yes Attrition Target Class. Oversampling needed to balance the imbalance class as under-sampling will lead to loss of information because imbalance is high.
* Data types of the Features of int, float and object types.
* Objects types Data are converted to numeric Labels with Label Encoder.
* Before Removing Outliers, Skewness observed in Continuous Features Monthly Income.
* Outliers Observed in Monthly Income.
* Outlier Treatment done with ZSCORE Method which gave Data Loss of 5.6% which is considerable.
* After Outliers Removal, Monthly Income does not have skewness.
* Skewness is still present in the categorical features which can be ignored.
* Scaling the Data with Min Max Scaler.

**EDA REMARKS**

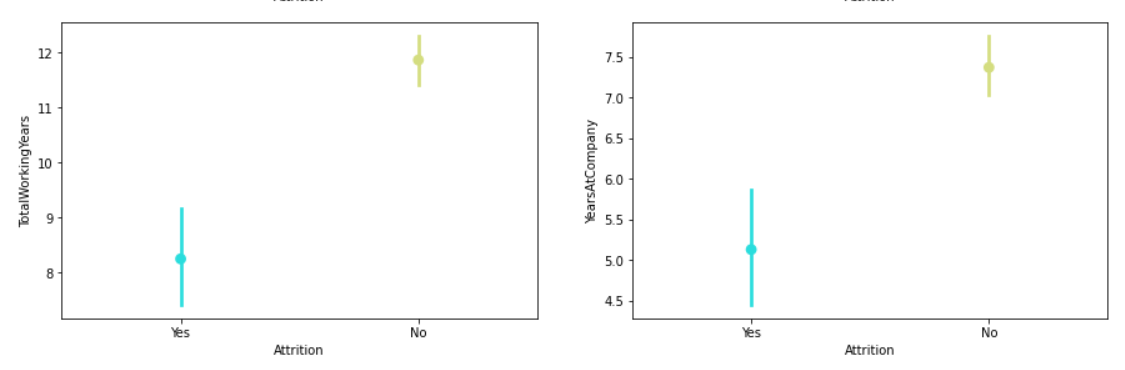
* Visualization of Data involve univariant analysis, for categorical data with count plots, Bivariant Analysis between features, Features and Target variable with many different visualization techniques. Multivariant Analysis involve correlation table between all the features (predictors and target variable) and pair plots.
* Imbalanced Classes observed in Target Classes with 84:16 ratio with No:Yes Attrition Labels.

**FEATURE INFORMATION**

* 70% of employees travel Rarely for Business, 18% employees travel frequently for Business, and 10% of employee do not travel for Business.
* 65% of employees are in Research and Development Department, 30% are in Sales Department and 4% are in HR Department of the Company.
* Highest no of employees has done their education in life Sciences about 41% of employees, 31% in Medical.
* Company has gender ratio of 60%:40% with male and female.
* Top Job Roles in the companies are 22% as Sales Executive,19% as Research Scientist, 17% Laboratory Technician.
* Around 37% of employees belong of Job level 1,2 Each.
* Count of employee with Job Involvement 3 is highest with 59% of employee count.
* Most of the employees (35% of employees) has number of companies worked as 1.
* 85% of employees have performance rating 3. Others have 4.
* Only 19% of the Employees do Over Time in the company.
* 45% of the employees are married, 31% of the employees are unmarried and 22% of the employees are divorced.
* 75% of employees belong have been recently promoted within 0 to 2 Years.
* 55% of employees belong to salary increment bracket of 11-14%
* Only 17% of employees have more than 20% salary increment. Rest belong to the bracket of 15-20% of salary increment.

**TARGET VRIABLE RELATIONSHIP WITH FEATURES:**

* Average Value of Age, Daily Rate, Monthly Income, Years at company and total Years With Company is lower for the employees who belong in yes category of Attrition.



* As you can observe, Total working hours and Years at Company have Higher Average Value for No Attrition.
* Employee having large distance from home has greater probability to be in Attrition.
* In the Attrition Yes Category, Highest count of Business Travel type is for travel\_rarely and lowest count for non-travel.
* Low probability of HR to belong in Yes Attrition Category.
* Attrition count of Females are less and high for No attrition. Males have high probability for Attrition.
* Employees do belong to No Overtime have high probability to not belong to Attrition.
* For High Job Levels, probability of No attrition increases as compared to low job levels.
* for Higher value of years\_in\_current\_role, Attrition probability for No increases. Attrition Yes majorly belongs to low values of years\_in\_curernt\_role.
* Attrition majorly belongs in low values of years\_with\_manager.
* Multi collinearity Exist between the features: Total Working Years and jobLevel, Total Working Years and Age, Monthly Income and StockOption Level
* Correlation of Attrition was moderate with YearsWithCurrManager, YearsInCurrentRole, YearsAtCompany, TotalWorkingYears, Stock Level Option, Overtime, Monthly Income,

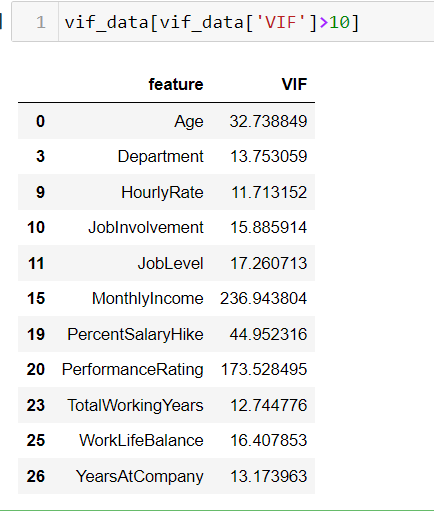
Marital Status, JobSatisfaction, JobInvolvement, Age.

**PRE-PROCESSING PIPELINE**

* Pre-processing pipeline involving Data Cleaning, Outlier detection, Outlier treatment, Scaling of Data, Feature selection or feature reduction with PCA .
* Outliers detected in Monthly Income Feature, treated with Zscore method .
* After Removal of Outliers, 6.2% data loss observed.
* No missing data to handle.
* Multi collinearity observed, After applying VIF factor analysis for multicollinearity

Considering the VIF threshold to be 10, there are few features with high VIF, Look at Below image.

**OBSERVATION**: High VIF value suggest high correlation between features.

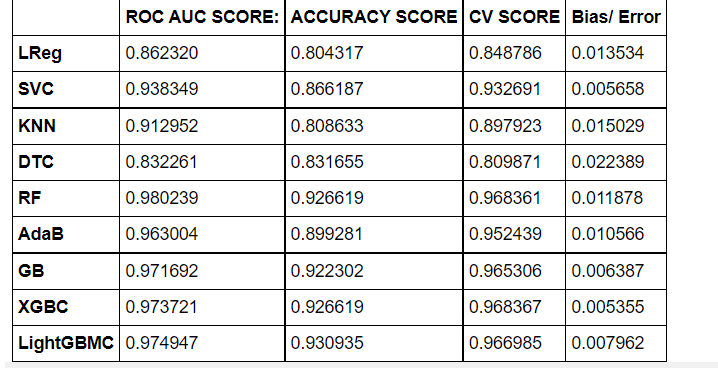


Features were removed after VIF analysis , features removed were: ['Age', 'JobLevel', 'PerformanceRating', 'MonthlyIncome', 'WorkLifeBalance' ] to reduce Multicollinearity.

* Dataset scaled with Min Max Scaler.
* Skewness threshold considering +/- 0.55, Skewness removed in continuous features after outlier removal.
* PCA used for finding how many features give more than 98% information. We observe 24 Features gives upto 98% information about the Target Class, as noise component in the data is less, concluding no need for PCA reduction because of less noise component.

**BUILDING MACHINE LEARNING MODELS**

* As we have Imbalanced classes of Target Label, performed oversampling to avoid biased accuracy for minority or majority target class.
* Over Sampling Technique used was SMOTE, It adds data of the minority class to have balanced set of target classes data.
* Evaluation Metric Chosen is : ROC AUC Score.
* Finding Best Random state. that is 168.
* Splitting the data into train set and test set with 70:30 ratio with sklearn.model\_selection.train\_test\_split.
* For Training Score, Cross validation method is used. Cross validation fits the models in n number time with different set of sample, so as to avoid biased results towards majority target class.
* Algorithms used for modelling is Logistic Regression, SVC, KNN, Decision Tree Classifier, En-sembling techniques.
* Ensemble Bagging and boosting models used are Random Forest, Ada Boost Model, Gradient Boost Model, XGBoost Model, Light GB Model.
* Training scores are calculated with cross validation with cv = 5 and scoring = ’roc\_auc’ as ROC\_AUC score is our evaluation metrics for the model.
* Please refer below to see the accuracy score, roc auc score, training score and the error term which is the difference between the training score and validation score (ROC AUC METRIC)



* Observation from the above table, all the ensembling techniques perform better than the simple models.
* As we can see, all the models have less error, suggesting the models the not underfitting or overfitting.
* The least error metric is in XGBoost Model.
* XGBoost Classifier is taken for hyper parameter tuning.
* Hyper Parameter Tuning is performed with Randomized Search Method with max\_iter = 70.

**HYPER PARAMETER TUNING OF XGBOOST**

* Parameters Chose for Hyper Parameter tuning are:

params = {

'n\_estimators': [50,100,150,200],

'max\_depth': [5,10,15,20,25],

'learning\_rate': [0.1,.5,.2,0.01],

'gamma': [.1,.5,1,2,5],

'subsample': [1,.8,.75,.6],

'colsample\_bytree':[1,.8,.75,.6],

'min\_child\_weight': [1,3,5,2] }

* Fitting the model to RandomizedSearchCV

boost = XGBClassifier()

grid = RandomizedSearchCV(boost,params,cv=5,n\_iter=70,scoring='roc\_auc',verbose=2)

* Tuning the XGB Model with above parameter we get tuned model having best parameters.
* Best Estimators after Randomized Search:

{'subsample': 0.6,

'n\_estimators': 150,

'min\_child\_weight': 1,

'max\_depth': 25,

'learning\_rate': 0.5,

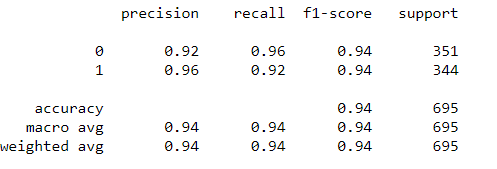
'gamma': 0.1,

'colsample\_bytree': 0.75}

**EVALUATION OF TUNED MODEL:**

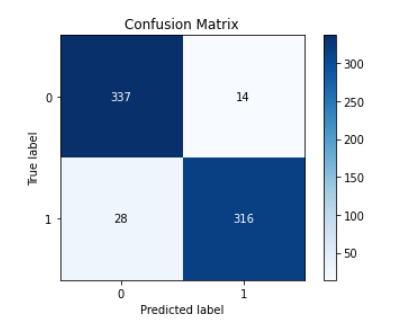
* + Obtaining A Accuracy Score : 0.**9395**
  + Obtaining A ROC AUC SCORE : 0.**9776**
  + Obtaining A Training Score with Cross Validation : 0.**9721**

**CLASSIFICATION REPORT**

****

No Biased results toward majority class, both target class has good recall and precision score as you can see in the classification report with a F1 Score of .94%. Suggesting, Results are not biased for a target class.

**CONFUSION MATRIX**

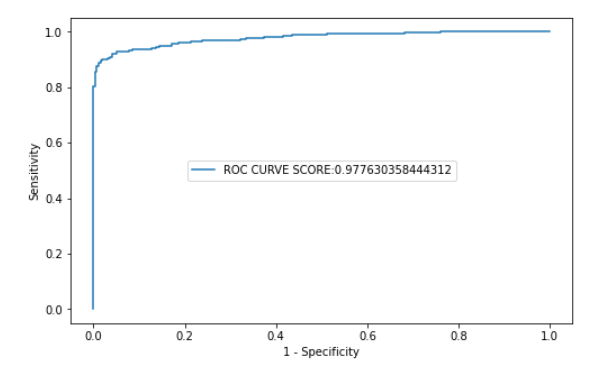
****

Confusion matrix tells us the number of test prediction correctly predicted and the prediction which were not correctly predicted.

AS we can observe, we have high No of Correctly predicted labels, that is True positive and true negative and only 6-7% of the data are predicted wrongly , that is False Positive and False Negative.

**ROC AUC CURVE : EVALUATION METRIC**

Below is the Roc Curve and its auc\_score is 0.9776 suggesting that 97% correct prediction of positive class.

****

**SAVING AND LOADING THE MODEL:**

* After modelling the Data, we save the final build model with joblib, look at the code snippet below:

**import** **joblib**

joblib.dump(model,'model\_svc\_attrition.pk')

* Loading the Model for future prediction of HR Attrition.

joblib.load('model\_svc\_attrition.pk')

OUTPUT

XGBClassifier(base\_score=0.5, booster='gbtree', colsample\_bylevel=1, colsample\_bynode=1, colsample\_bytree=0.75, gamma=0.01, gpu\_id=-1, importance\_type='gain', interaction\_constraints='', learning\_rate=0.05, max\_delta\_step=0, max\_depth=25, min\_child\_weight=1, missing=nan, monotone\_constraints='()', n\_estimators=150, n\_jobs=8, num\_parallel\_tree=1, random\_state=0, reg\_alpha=0, reg\_lambda=1, scale\_pos\_weight=1, subsample=0.8, tree\_method='exact', validate\_parameters=1, verbosity=None)

**UNDERFITTING/ OVERFITTING CHECK:**

* As training score and testing score (roc auc score) difference is very or both scores are very near to each other, with a training and testing score of 97%. We can rule out over fitting and underfitting.

**CONCLUDING REMARKS**

* XGBoost Model is builds with an Accuracy score of 94%.
* High F1-Score suggests prediction of Both majority and minority classes are done, absence of biased results. F1 Score achieved was 94%
* Evaluation metrics choose to evaluate the model was ROC\_AUC score which is 97.7% of our model, suggesting that 97.7% correct prediction of Positve clas (Label 1 Class).
* Saved the model with Job lib.
* Training score metrics was Cross Validation Score.
* Multicollinearity existed in the model which was analysed with Variance Inflation Factor and manually selected the features for modelling of data.
* After Feature Selection, we chose 25 Features out of 34 Features for modelling.
* According to PCA, 22 Features were able to give upto 98% information about the Data, Noise component is less.