BITS Pilani
Pilani | Dubai | Goa | Hyderabad
Work Integrated Learning Programmes

Attrition Analysis

Group Number 74 – Santhi Maadhaven, Sabarish V

Data Mining

M.Tech Data Science and Engineering

10-Jan-2022

Overview

- **Objective**

Employee Attrition refers to % of employee who leaves the organization. Attrition of employees is the issue faced by many companies across the world, where valuable and experienced employees leave the company on a regular basis. Its loss to an organization particularly in case of service industry where client prefer to interact with familiar people. They will lose business in case there is more attrition as confidence and trust will be lost. Not only in terms of money but also knowledge base & experience over time will not be increased and having new joiners in the team prone to make errors till they gain knowledge. This is a major problem and our main objective of this exercise is to analyze & develop model that can help to predict attrition.

- **Methodology**

For this problem, we will use (C) Cross (I) Industry (S) Standard (P) Process for (D) Data (M) mining (CRISP-DM) Business model and steps that we will follow are understanding of problem, data preprocessing, data analysis, model training & validation, model predictions and Visualization of results

Accuracy of the prediction depends on the data and model to be used. Aim of this study is to focus on these two parameters to maximize the accuracy of the predictive model. Employee attrition problem is a binary classification problem and due to simplicity & interpretability we will choose Decision Tree and Logistic Regression. We need dataset of employee with details around left or not which will help to analyze the reasons why employees left (Descriptive Analytics), predict who will leave (predictive Analytics) and what steps can be taken for retaining existing employees (Prescriptive Analytics).

Methodology

- **Logistic Regression**

Logistic regression is a popular classification algorithm used to assign observations to a discrete set of classes. It's a predictive analysis algorithm and based on the concept of probability. It transforms output using sigmoid function to return a probability value. It's much easier to train, implement, efficient, fewer assumptions of variables and decent degree of accuracy

- **Decision Tree**

Decision Tree is used for both classification and regression problems and mostly preferred for classification problems. Its tree structured classifier where internal nodes represent the features of a dataset, branches represent the decision rules and each lead node represents the outcome

Dataset & Understanding of Data

- **Size of the Dataset**

It have got 1470 data values and 33 features

- **Variable Type**

Attributes	Data Type	Category	Type	Description
Attrition	Character	Binary	Target	Describes whether employee left the company or not with 'Yes' or 'No' value
Age	Numeric	Continuous	Predictor	Describes the age of the employee with numeric value
BusinessTravel	Character	Categorical	Predictor	Describes whether employee travel for business with three values 'Non-Travel', 'Travel_Rarely' , 'Travel_Frequently'
DailyRate	Numeric	Continuous	Predictor	Describes daily rate of the employees with numeric value
Department	Character	Categorical	Predictor	Describes which department that employee belongs to
DistanceFromHome	Numeric	Continuous	Predictor	Describes how much distance is workplace from home
Education	Numeric	Ordinal	Predictor	Describes education level of the employee with values from 1 to 5

EducationField	Character	Categorical	Predictor	Describes which field of education that employee have studied
EmployeeNumber	Numeric	Nominal	Predictor	Describes employee identify number
EnvironmentSatisfaction	Numeric	Ordinal	Predictor	Describes environment satisfaction level of employee from 1 to 4(‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’)
Gender	Character	Binary	Predictor	Describes gender of the employee
HourlyRate	Numeric	Continuous	Predictor	Describes hourly rate of the employee
JobInvolvement	Numeric	Ordinal	Predictor	Describes job involvement of the employee from 1 to 4
JobLevel	Numeric	Ordinal	Predictor	Describes job level of the employee from 1 to 5
JobRole	Character	Categorical	Predictor	Describes job role of the employee
JobSatisfaction	Numeric	Ordinal	Predictor	Describes job satisfaction of the employee from 1 to 4(‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’)
MaritalStatus	Character	Nominal	Predictor	Describes marital status of the employee
MonthlyIncome	Numeric	Continuous	Predictor	Describes monthly income of the employee with numeric value
MonthlyRate	Numeric	Continuous	Predictor	Describes monthly rate of the employee with numeric value
NumCompaniesWorked	Numeric	Discrete	Predictor	Describes no. of companies employee worked
Over18	Character	Binary	Predictor	Describes whether employee age is over 18 years
OverTime	Character	Binary	Predictor	Describes whether employee worked overtime
PercentSalaryHike	Numeric	Continuous	Predictor	Describes % of hike employee have received
PerformanceRating	Numeric	Ordinal	Predictor	Describes the performance rating of the employee with values 3 to 4
RelationshipSatisfaction	Numeric	Ordinal	Predictor	Describes relationship satisfaction of employee with value range from 1 to 4(‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’)
StockOptionLevel	Numeric	Discrete	Predictor	Describes stock option level which employee holds with values from 0 to 4
TotalWorkingYears	Numeric	Discrete	Predictor	Describes total years of employee experience
TrainingTimesLastYear	Numeric	Discrete	Predictor	Describes on times of training employee received last year
WorkLifeBalance	Numeric	Ordinal	Predictor	Describes work life balance of the employee with range from 1 to 4 (1 ‘Bad’ 2 ‘Good’ 3 ‘Better’ 4 ‘Best’)
YearsAtCompany	Numeric	Discrete	Predictor	Describes years of employment at the current company
YearsInCurrentRole	Numeric	Discrete	Predictor	Describes number of years at current role of the employee
YearsSinceLastPromotion	Numeric	Discrete	Predictor	Describes number of years since employee last promoted
YearsWithCurrManager	Numeric	Discrete	Predictor	Describes number of years with current manager

- **Data Distribution and handling imbalanced data**

  Data Distribution:

  Out of total employees – 1470, resource who stay is 1214 (No) & resource who left is 235 (Yes) per data set given for target variable ‘Attrition’. Dataset seems to be imbalanced as only 16% denotes attrition and fewer samples to predict model. However, it’s easily predictable that more employees will stay than leave.

  For imbalanced data sets, precision & recall will be the right metrics and not accuracy. Accuracy is not the right metric for imbalanced datasets because model can learn to predict majority class most of the time.

Recall (True Positive Rate) will help us in answering the below
- Does model successfully find % of employees leaving?

Precision (Measure of Exactness) will help us in answering the below
- How many will leave company that model predicts?

If we focus on maximizing precision, we are trying to minimize "over-shooting”. If model predicts that someone will leave, we take action to make them stay. If they actually weren’t going to leave, then company will waste their resources to keep someone who would stay regardless.

With recall, we try to minimize "under-shooting" . Higher the recall and fewer false negatives. If model doesn’t identify employee leave correctly, then company will have made no effort at all to save employee. We will use Recall as the primary metric for model evaluation and comparison. We will not compromise on the precision, we will take account of F1 which combines both recall and precision. Higher the f1 score, the better the prediction.

Feature Wrangling

- **Missing Value** :

While traversing through dataset, it is found that ‘DistanceFromHome’,’ Education’,’ Hourly Rate’,’ MonthlyRate’,’ MonthlyIncome’ have ‘NAN’ values and handled it through dropna() function in Pandas Python. Approach taken for missing value is to remove tuples from the dataset.

Post removal of tuples, dataset we will be dealing with is 1449 . Dataset is validated for null values as well post tuple removal for “NaN” and conversion of data type

- **Irrelevant Value :**
  - ‘Over18’ is having single value and will not help in prediction so this have to be removed
  - ‘EmployeID’ is just ID number and doest help in prediction so this is removed
- **Aggregate Value :**

Use numerical binning concepts; aggregation of data is done for better analysis & prediction
  - ‘Age’ is categorized into '<29', '30-39', '40-49','50-59','60-69'
  - ‘MonthlyIncome’ is categorized into ‘Low’,’Medium’,’High’

- **EDA outcomes and discussion**

Age & Monthly Income:

Majority of employees falls in age group of 20-40. Attrition rate also seems to be higher in that group. Mid-level employees 30-40 leaving organization more followed by 20-30 age groups. One of the contributing factors for leaving organization is due to low income. Attrition at senior level > 40 seems to be very less.

MonthlyIncome from 0 to 6666 are high and attrition happens with employees whose salary falls in that range. Also while doing the analysis with Age, it is found that employee at mid senior level (30-40) getting low monthly income (< 6666) which is clearly seems to be one of the factors.

Job Related Factors:

Considering JobInvolvement rating category as 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’, majority of the employees involved high/very high at job. Attrition rate is there across irrespective of job involvement.

Employee with Job level 1, 2 are more in the dataset and have high attrition rate as well. Job Role such as 'Laboratory technician’, ‘Sales Executive' tend to leave company more followed by 'Research Scientist’. Clearly R&D department is high and attrition also high with job roles 'Laboratory technician','ResearchScientist'

Education Related Factors:

Employees who studied Life science & medical education field are more in the dataset. Attrition also seems to be high in those fields

Travel Related Factors:

Those who travel very rarely is high in the dataset and attrition rate is high in that category as well

Satisfaction & Wellbeing:

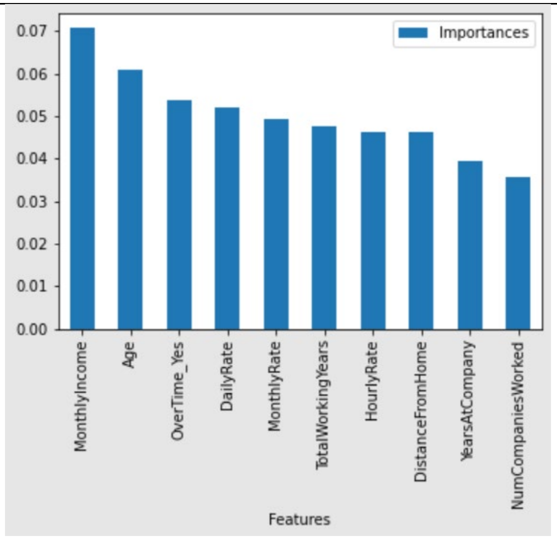
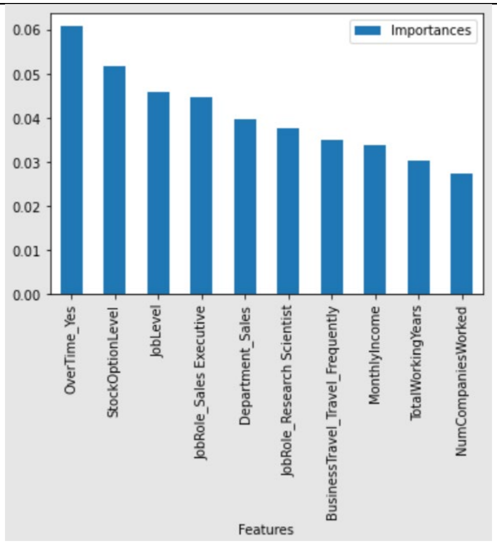
Majority of the employees have high job, relationship & environment satisfaction. Employees have rated work life balance as better. While comparing with the attrition rate, attrition is spread across irrespective of satisfaction level & better work life balance. Those who work overtime also tends to leave more than who doesn't but majority of employees don't do overtime.

Other Factors:

- Male employees are more in the dataset and attrition rate also high.
- Employees with married status is more in the whole dataset but attrition is high with single status followed by married

Feature Engineering & Selection

Categorical values are given dummy values and then embedded into final data model. So final Data set size used for modeling & prediction is (1449, 45)
Two techniques used for feature selection are

Random Forest Classifier	**XgBoost**
This is a Bagging Algorithm which aggregates a specified number of decision trees. These decrease impurities (Gini Impurity) over all trees & improve purity of then node. At start of the trees, greatest decrease in impurity will be at start of the trees and nodes with the least decrease in impurity occur at the end of trees. This will help in selecting feature of most importance.	Gradient boosting automatically provides estimates of feature importance from a trained predictive model. After boosted trees are constructed, it is easy to retrieve importance scores for each attributes.
	
Feature removed Based on the feature engineering techniques used for importance , below features	**Feature removed** Based on the feature engineering techniques used for importance , below features are removed

are removed from the final data set which is of very low value. 'JobRole_Human Resources', 'JobRole_Laboratory Technician', 'JobRole_Manager', 'JobRole_Manufacturing Director','JobRole_Research Director', 'JobRole_Research scientist' ,'JobRole_Sales Executive', 'JobRole_Sales Representative','EducationField_Life Sciences', 'EducationField _Marketing' ,'EducationField_Medical', 'EducationField_Other', 'EducationField_Technical Degree','PerformanceRating', 'Department _ Research & Development' , 'Departmen t _ Sales' ,'BusinessTravel _Travel_Rarely', 'Gender_ Male','MaritalStatus_Married'	from the final data set which is of very low value. 'PerformanceRating','JobRole_Human Resources', 'JobRole_Research Director','JobRole_Manufacturing Director', 'JobRole_Manager'

Modeling Results

- **Logistic Regression**

Using RandomForest technique

Training Data : (1159, 25) & Test Data : (290, 25)

```
Accuracy_score training data set: 0.8766177739430544
Accuracy_score test data set    : 0.8655172413793103

Classification Report:
              precision    recall  f1-score   support

           0       0.86      1.00      0.93       243
           1       0.90      0.19      0.32        47
    accuracy                           0.87       290
   macro avg       0.88      0.59      0.62       290
weighted avg       0.87      0.87      0.83       290

Confusion Matrix: [[242   1]
                   [ 38   9]]
```

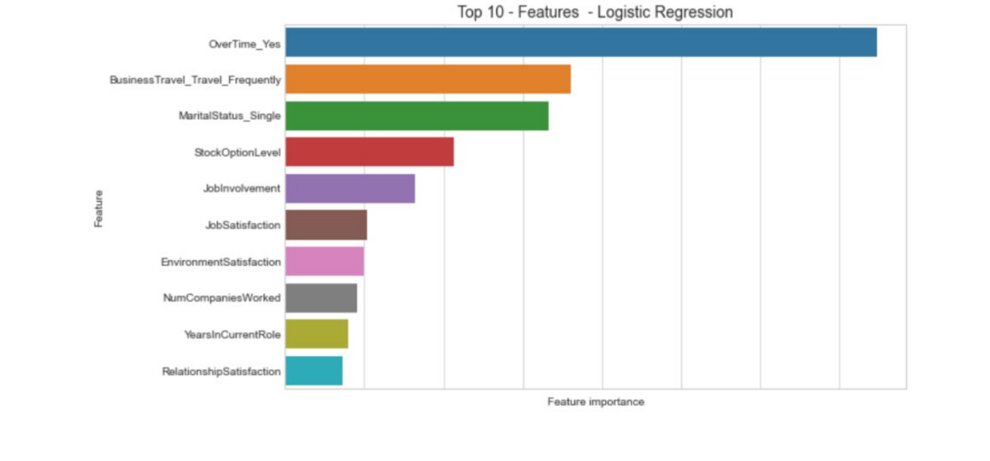
Using XGBoost technique

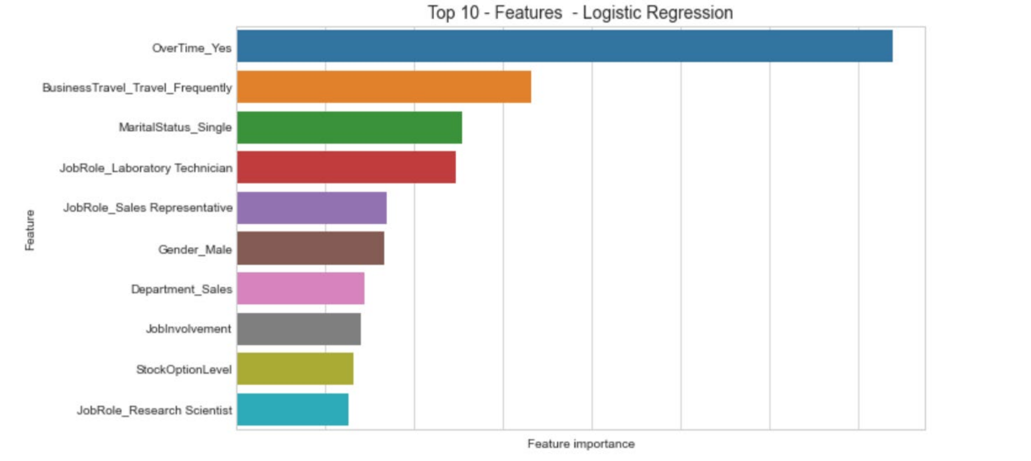
```
Training Data : (1159, 39) & Test Data : (290, 39)
Accuracy_score training data set: 0.8852459016393442
Accuracy_score test data set    : 0.8827586206896552
Classification Report:
                   precision    recall  f1-score   support

           0       0.89      0.99      0.93       243
           1       0.84      0.34      0.48        47

    accuracy                           0.88       290
   macro avg       0.86      0.66      0.71       290
weighted avg       0.88      0.88      0.86       290

Confusion Matrix: [[240   3]
                    [ 31  16]]
```





- **Decision Tree**

Using RandomForest technique

```
Training Data : (1159, 25) & Test Data : (290, 25)
Accuracy score training data set: 1.0
Accuracy Score test data set    : 0.7862068965517242
Classification Report:
              precision    recall  f1-score   support
           0       0.87      0.88      0.87       243
           1       0.33      0.30      0.31        47

    accuracy                           0.79       290
   macro avg       0.60      0.59      0.59       290
weighted avg       0.78      0.79      0.78       290

Confusion Matrix: [[214  29]
                   [ 33  14]]
```

Using XGBoost technique

```
Training Data : (1159, 39)& Test Data : (290, 39)
Accuracy_score training data set: 1.0
Accuracy Score test data set    : 0.7931034482758621
Classification Report:
          precision    recall  f1-score   support

           0       0.89      0.86      0.87       243
           1       0.38      0.43      0.40        47

    accuracy                           0.79       290
   macro avg       0.63      0.64      0.64       290
weighted avg       0.80      0.79      0.80       290

Confusion Matrix: [[210  33]
                   [27 20]]
```

BITS Pilani
Pilani | Dubai | Goa | Hyderabad
Work Integrated Learning Programmes

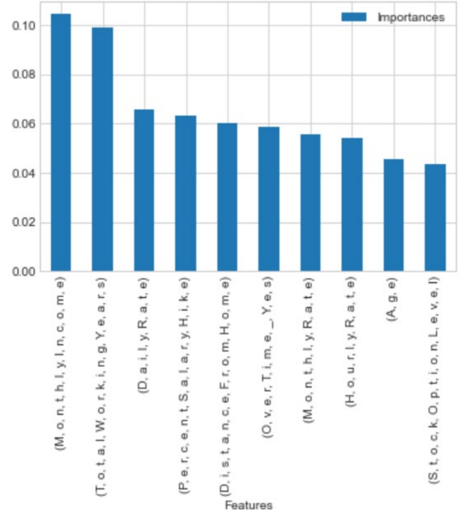
Attrition Analysis

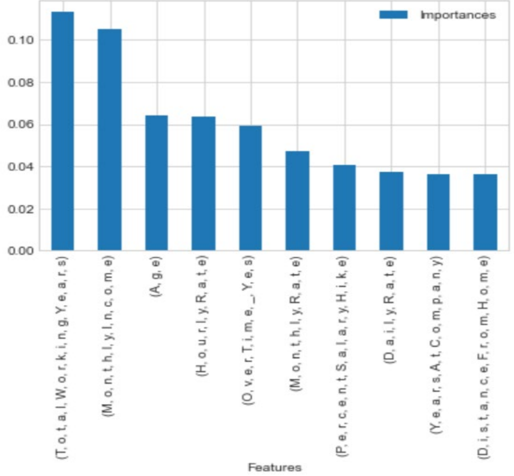
Group Number 74 – Santhi Maadhaven, Sabarish V

Data Mining

M.Tech Data Science and Engineering

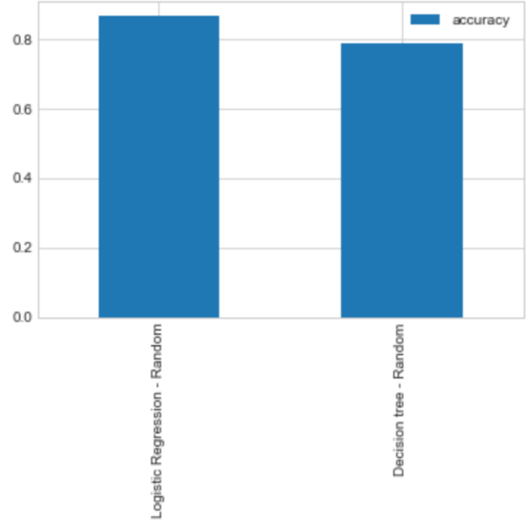
10-Jan-2022

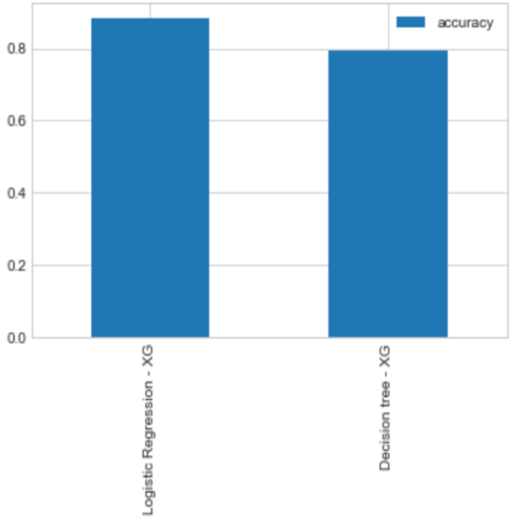




Conclusion

```
'Logistic Regression - Random': 0.8655172413793103, 'Decision tree - Random': 0.7862068965517242
'Logistic Regression - XG': 0.8827586206896552, 'Decision tree - XG': 0.7931034482758621
```





- Based on the above analysis , it can be concluded that
    - Logistic Regression prediction is better when compared to Decision tree as accuracy rate seems to be higher
    - Looking at confusion matrix, it can be found that there are 20% True Positive & Negative for Decision Tree but for logistic regression its 12%
    - Based on the logistic regression prediction, below are suggestions can be shared to organization
        - Don’t make your employees work over time
        - Engage with your employees more to improve satisfaction levels & understand concerns at job level
        - Have regular feedback mechanism for performance so that job involvement can be more
        - Have regular connect with employees who travel frequently to identify pain points/concerns

-