IBM HR Analytics: Employee Attrition

SPRINGBOARD DSC CAPSTONE PROJECT 1 BY SNEHA RANI

Introduction

- *Someone well said :*

  *“**You don't build a business. You build people, and people build the business.**”*

What do we mean by Attrition?

Attrition refers to employees who leave their jobs due to normal circumstances. So, it is basically a normal life cycle of employment.

In mathematical term, attrition rate is defined as the number of employees who leave a company during a specified time period divided by the average total number of employees over that same time period.

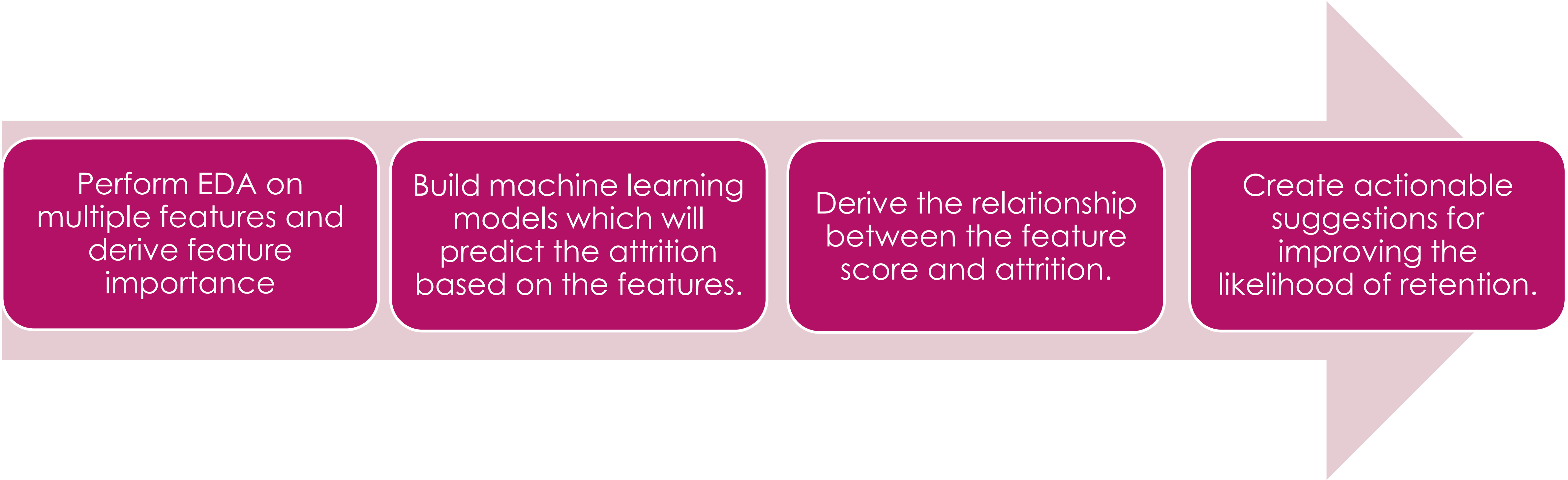
Problem and Client

- Client : IBM

- Client would like to know the factors that lead to employee attrition.

- In other terms, recognizing and understanding what factors that were associated with employee attrition will allow companies and individuals to limit this from happening and may even increase employee productivity and growth.

Approach



Perform EDA on multiple features and derive feature importance

Build machine learning models which will predict the attrition based on the features.

Derive the relationship between the feature score and attrition.

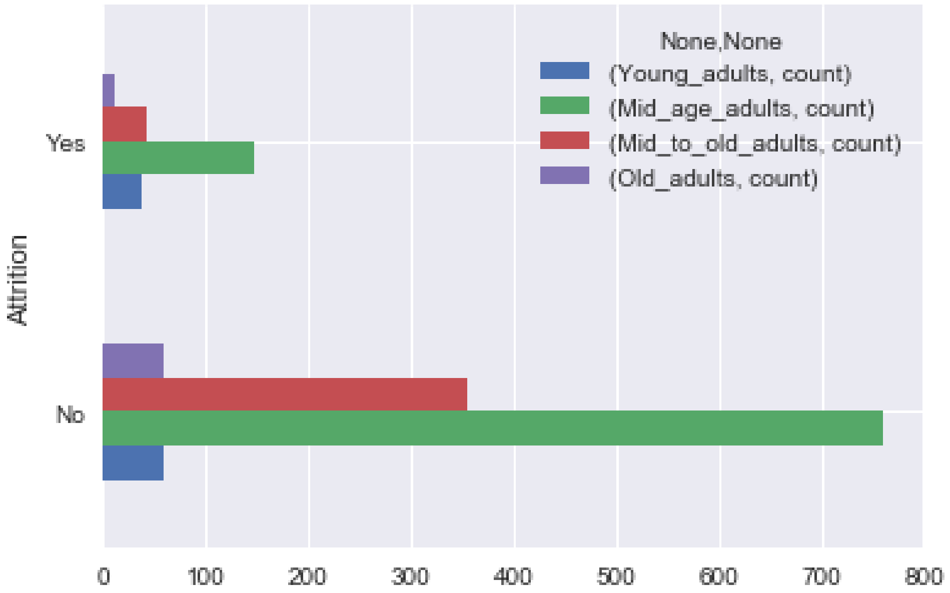
Create actionable suggestions for improving the likelihood of retention.

Tools Used

- **Pandas:** Loading the data, data wrangling and manipulation.
- **Scikitlearn:** Libraries for classifiers, Model evaluation, Metrics, Cross validation, Feature Importance
- **Imbalance-Learn:** SMOTE
- **MatplotLib and Seaborn:** Data Visualization

Exploratory Analysis

- ❑ **Which age-group people contribute maximum attrition?**



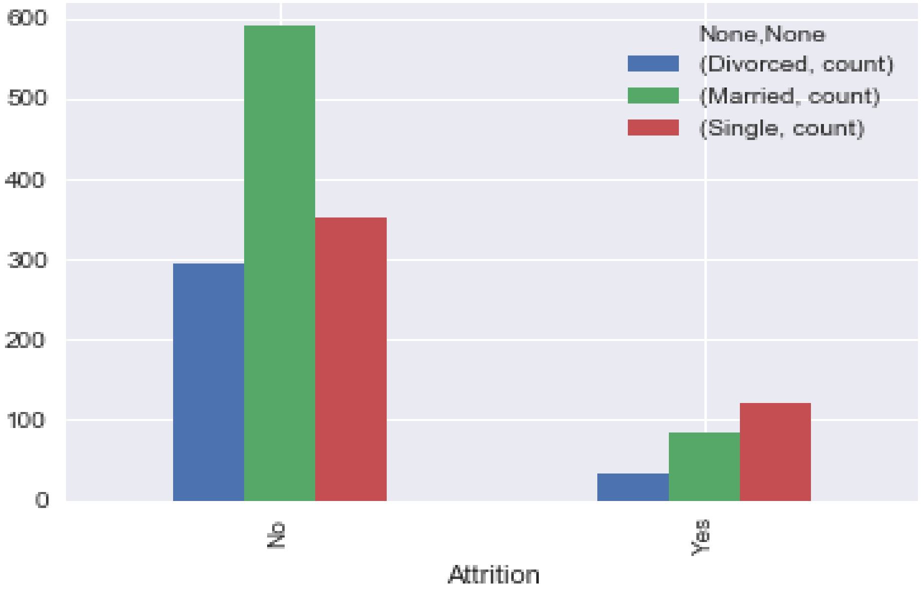
Age was grouped into four categories :

Young adults (15-24 Yrs)
Mid age adults (25-40 Yrs)
Mid to old adults (41-54 Yrs)
Old age adults (55-64 Yrs)

- Shows maximum contribution of different age-group people based on the attrition.

Continued…

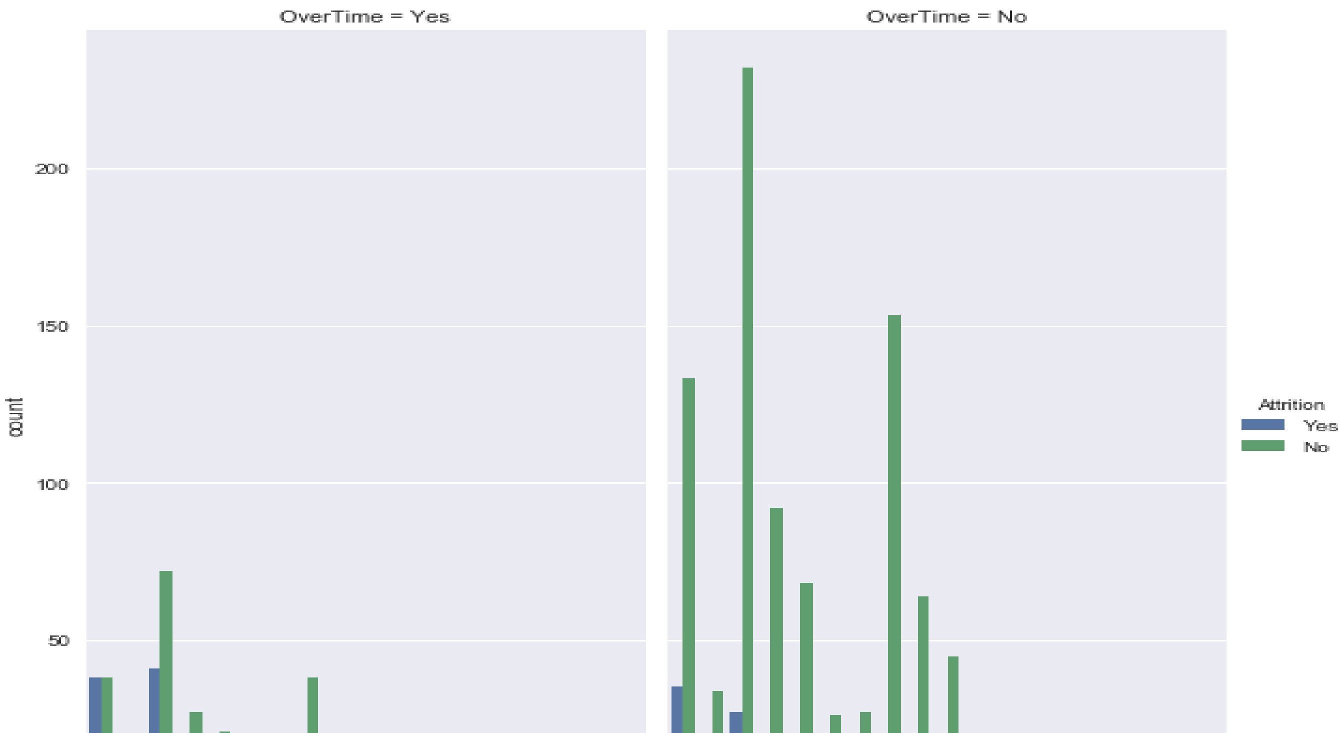
- ❑ **What is the count of married people and unmarried people attrition rate? Are married people more prone to attrition?**



- Analysis was based on the marital status.
- People who were ‘Single’ contributed more towards attrition.

Continued...

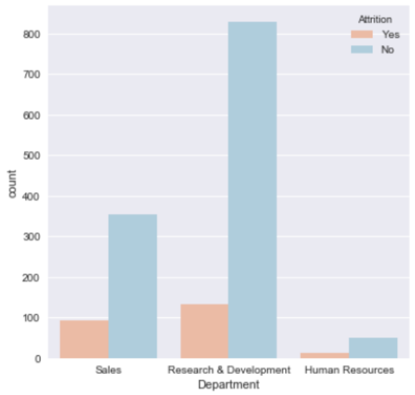
❑ **What is the count of people working OverTime and YearsInCurrentRole? How working overtime (or not), and the years in role relate to employee attrition?**

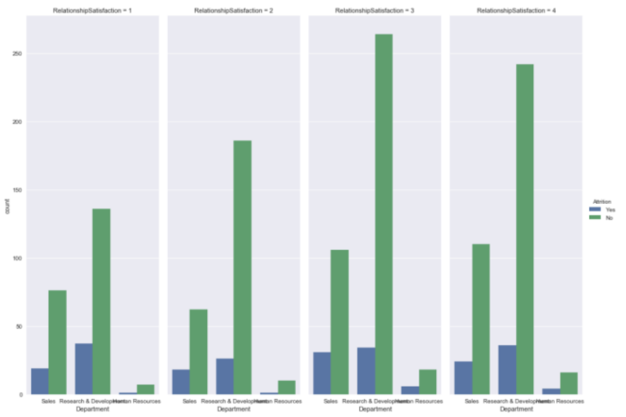


- It shows that people who had worked overtime under same current role left the company most.

Continued…

- ❑ **What is the count of attrition of each department on the basis of RelationshipSatisfaction? Does satisfaction level has any impact on employees leaving these departments?**

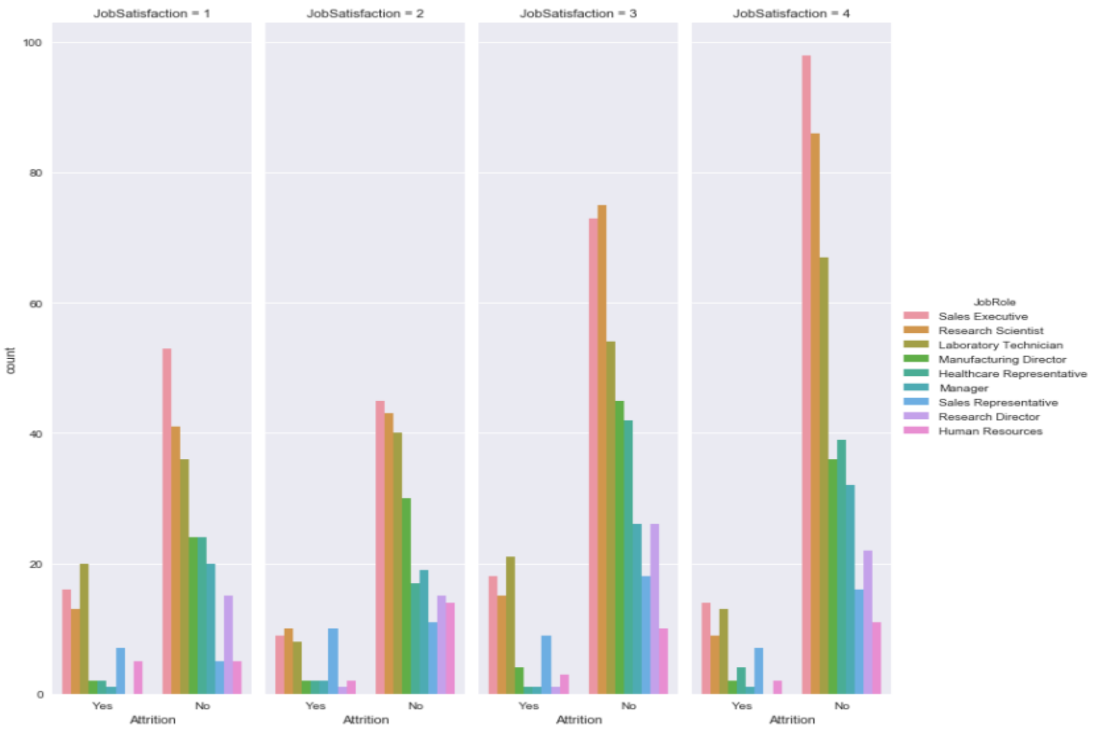




- It shows that research and development field has highest attrition count.
- Plot doesn't show any variance on attrition based on the relationship satisfaction feature.
- All four different levels (Low, Medium, High, Very High) has the similar number of attrition.

Continued...

- **Do JobSatisfaction and JobRole impact gradual loss of employees? Are these two features have a common pattern?**



- Employees in Sales Executive, Research Scientist, Laboratory Technician leave the company most having low and high (not very high) job satisfaction level.
- Sales Representative employees show a similar count of attrition under all job satisfaction levels.

Applying Inferential Statistics for my further findings

- ❑ **Is Age statistically impacting the Attrition Rate?**

- ❑ **Do male or female of different age groups impact differently to the employee attrition?**

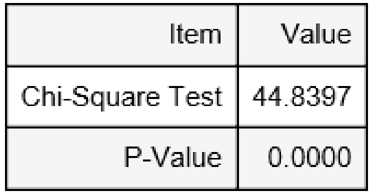
Since I am dealing with categorical features, an optimal test is Chi-square contingency test.

Chi-square Contingency test

- **Null hypothesis (H0):** *The 2 categorical variables are independent (there is no relationship between the variables)*
- **Alternate hypothesis (H1):** *The 2 categorical variables are dependent (there is a relationship between the variables)*

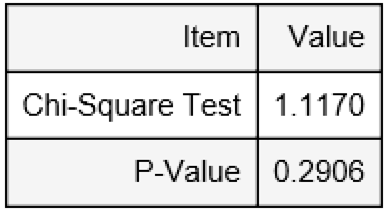
Results of my further findings using Chi-square test

- *Chi-square test shows 'Age_Group' and 'Attrition' are not statistically independent*
    - *Test has p-value of 0 (below the threshold value of 0.05) and so we rejected the null hypothesis,*



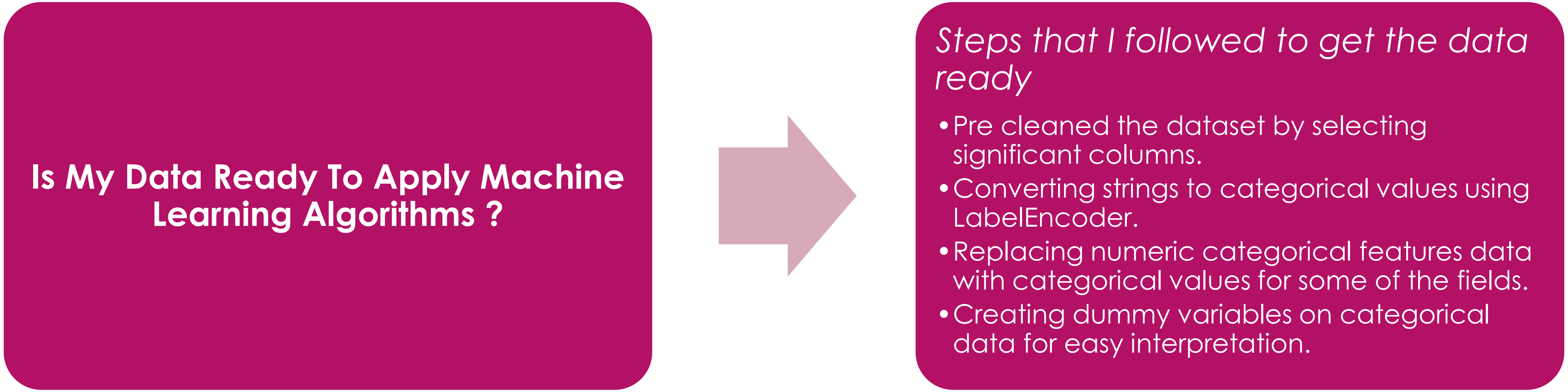
Item	Value
Chi-Square Test	44.8397
P-Value	0.0000

- *Chi-Square test shows 'Gender(Male or female)' and Attrition may be statistically independent.*
    - *Test has p-value of 0.2906(which is significantly higher than threshold value of 0.05) and so I cannot reject the null hypothesis.*



Item	Value
Chi-Square Test	1.1170
P-Value	0.2906

Modelling the Data



Is My Data Ready To Apply Machine Learning Algorithms ?

Steps that I followed to get the data ready

- Pre cleaned the dataset by selecting significant columns.
- Converting strings to categorical values using LabelEncoder.
- Replacing numeric categorical features data with categorical values for some of the fields.
- Creating dummy variables on categorical data for easy interpretation.

Build baseline model using logistic regression

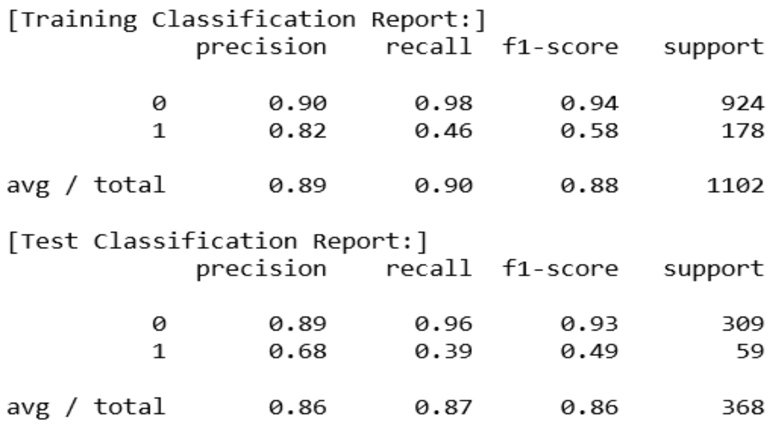
Approach:

- When fitting models, I wanted to ensure two things:
  - ✓ Finding the best model (in terms of model parameters).
  - ✓ The model is highly likely to generalize i.e. perform well on unseen. data.

- I first built logistic regression base line model using 'L2' and 'L1' regularization by :
  - ✓ Splitting the data into a training and test (hold-out) set.
  - ✓ Train on the training set, and test for accuracy on the testing set.

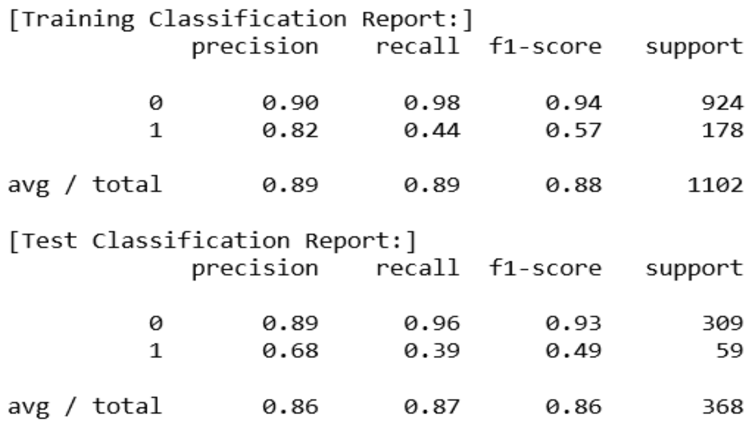
Classification Results

With L2 regularization, accuracy score was 0.8723 for test set and 0.8956 for train set.



```
[Training Classification Report:]
             precision    recall  f1-score   support

          0       0.90      0.98      0.94       924
          1       0.82      0.46      0.58       178

avg / total       0.89      0.90      0.88      1102

[Test Classification Report:]
             precision    recall  f1-score   support

          0       0.89      0.96      0.93       309
          1       0.68      0.39      0.49        59

avg / total       0.86      0.87      0.86       368
```

With L1 regularization, accuracy score was 0.8723 for test set and 0.8956 for train set.



```
[Training Classification Report:]
             precision    recall  f1-score   support

          0       0.90      0.98      0.94       924
          1       0.82      0.44      0.57       178

avg / total       0.89      0.89      0.88      1102

[Test Classification Report:]
             precision    recall  f1-score   support

          0       0.89      0.96      0.93       309
          1       0.68      0.39      0.49        59

avg / total       0.86      0.87      0.86       368
```

Based on classification report on Test set, it shows that larger Class(Class 0 or Not-Attrition class) is over influencing the model.

Hyper parameter optimization using Grid Search

- ❑ Comparing the recall value, 'L2' had slightly higher recall value than 'L1' in train data classification report.
- ❑ I used Grid Search to tune the model using best hyper parameter and corresponding penalty.

Result Summary:

- ✓ I created Resulting model 'M' which identify the best combination of **penalty** = '**L2**' (default) and the value of **C** = '**1**'.
- ✓ Classification report was more or less same as was for base line model.
- ✓ The recall, precision and f1 score were still not good.

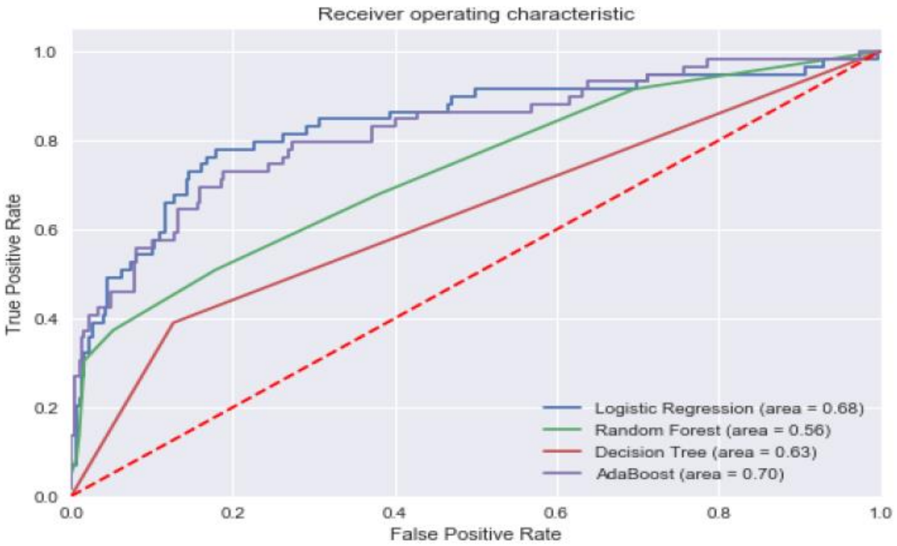
Next Step

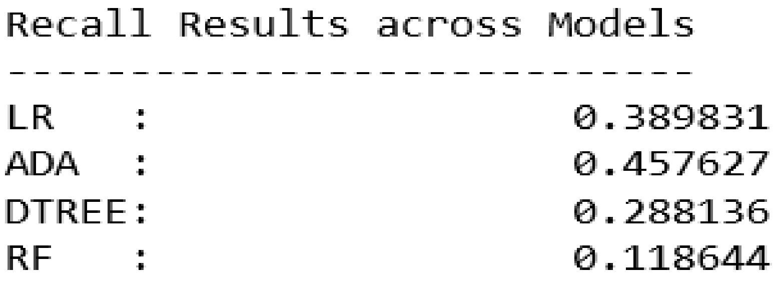
- ❑ Trying other algorithms:

- ✓ **Random Forest Classifier**
- ✓ **Decision Tree Classifier (CART)**
- ✓ **AdaBoost Classifier**

- ❑ Finding the highest recall score.

ROC Graph and Recall Score





```
Recall Results across Models
-------------------------------
LR    :                    0.389831
ADA   :                    0.457627
DTREE:                     0.288136
RF    :                    0.118644
```

From the ROC Graph,

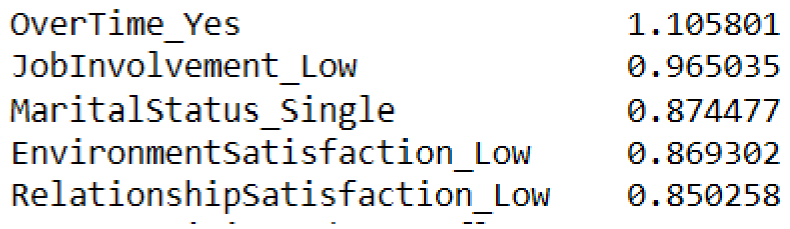
Ada Boost Model had the highest area under curve value of 0.70 and was more close to the top left corner which is one of the criteria for a good model.

- ❑ Results across different models shows that the AdaBoost has the highest recall score of 0.46.
- ❑ A good recall score is important to avoid unpredicted loss of the employees in the company.

Feature Importance

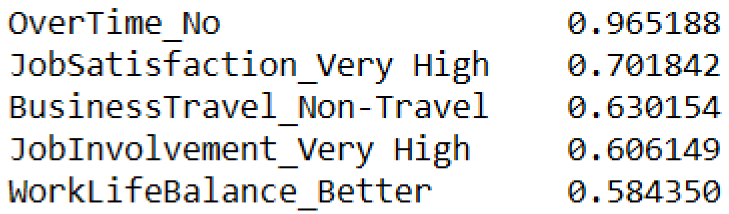
- Found highest positive and negative coefficient by absolute value.

Top 5 positive coefficients



```
OverTime_Yes                    1.105801
JobInvolvement_Low              0.965035
MaritalStatus_Single            0.874477
EnvironmentSatisfaction_Low     0.869302
RelationshipSatisfaction_Low    0.850258
```

Top 5 positive coefficients



```
OverTime_No                   0.965188
JobSatisfaction_Very High     0.701842
BusinessTravel_Non-Travel     0.630154
JobInvolvement_Very High      0.606149
WorkLifeBalance_Better        0.584350
```

- ✓ The one with the negative coefficient having the highest absolute weighted value will contribute the most important feature as belonging to class 0 (employees who have not left the company).
- ✓ The one with the positive coefficient having the highest absolute weighted value will contribute the most important feature as belonging to class 1 (employees who have left the company).

Feature importance analysis report

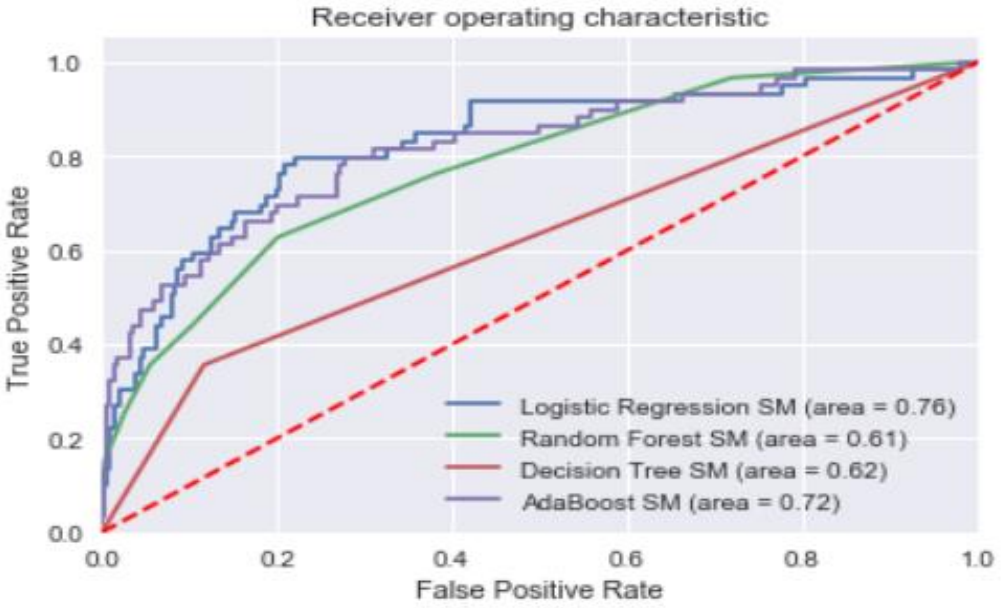
- ✓ Features like OverTime_Yes, JobInvolvement_Low with the high positive coefficients (towards attrition) makes some sense and are somewhat correlated to the things which influence employees to leave the company in short duration of time.
- ✓ Features like OverTime_No, JobSatisfaction_Very High, JobInvolvement_Very High with high (absolute value) negative coefficients (towards non-attrition) makes sense too and are somewhat correlated to the things which influence employees to stay at the company for the longer duration of time.

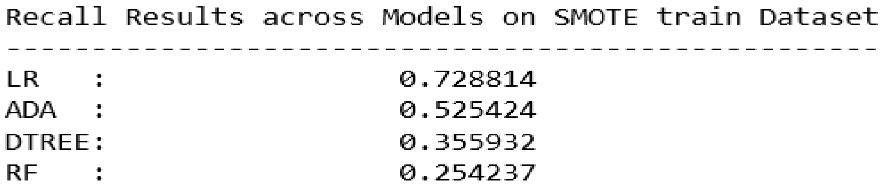
Summary Report for Imbalanced Dataset

- *AdaBoost Classifier model with recall score of 0.46 is one of the best model among all. Also, the training performance is better than test performance which does indicate that model is neither over-fitting nor erroneous.*
- *Logistic Regression Model 'M' also has same accuracy but the AdaBoost Model has better area under curve and Recall score so I have considered AdaBoost Classifier as the best model for this imbalanced Dataset.*
- *Class 1 (minority class) recall and F1 score are not so good when compared to Class 0 (majority) recall and F1 score. This means that Class 0 (majority class) data has over-influenced the model.*
- *So, to combat **Imbalanced Classes**, I used **Synthetic Samples (SMOTE)** by randomly sampling the attributes from instances in the minority class.*

Summary Report for Balanced Dataset using SMOTE

ROC Graph using Different Models using SMOTE

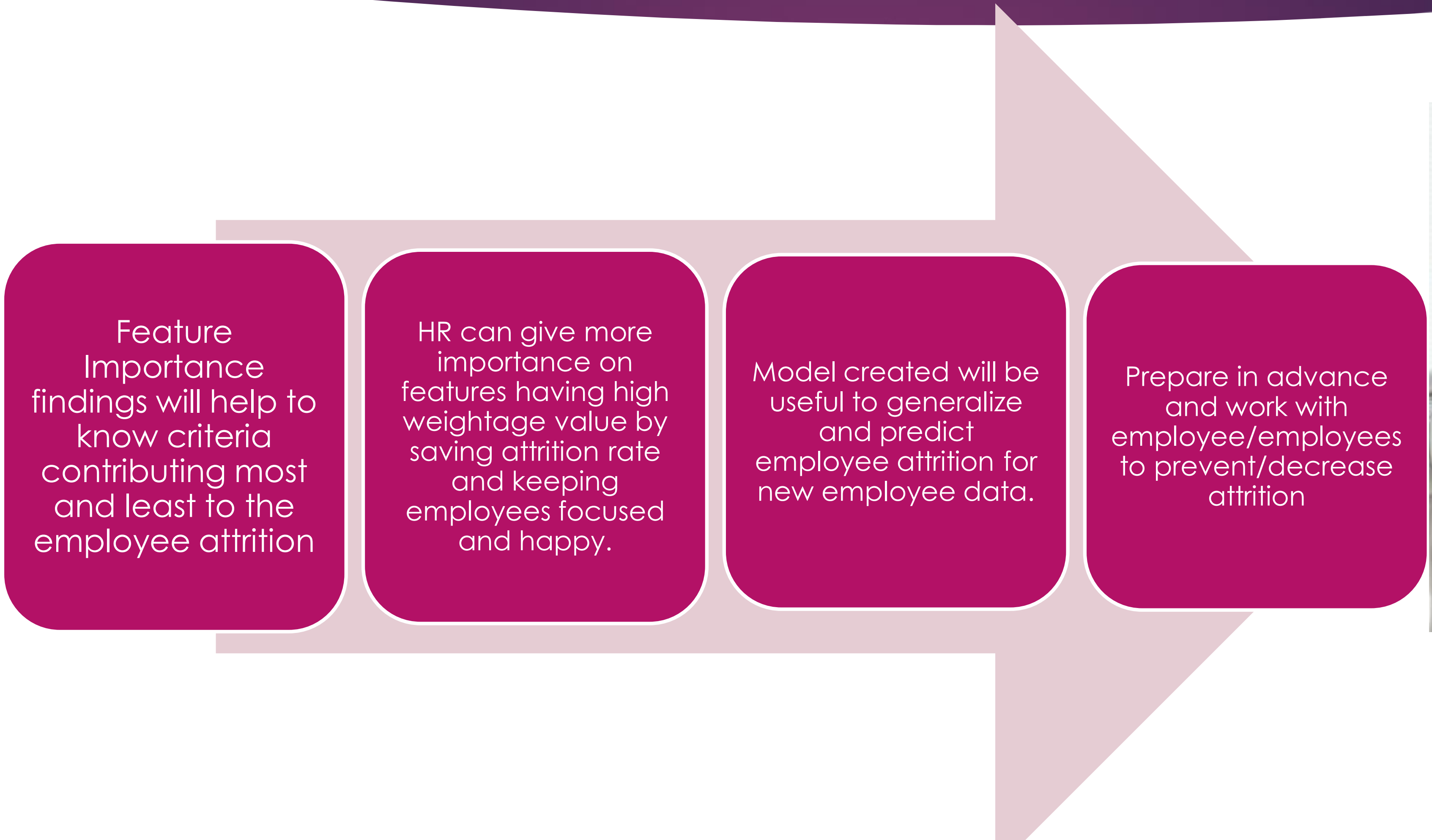




```
Recall Results across Models on SMOTE train Dataset
---------------------------------------------------
LR    :                   0.728814
ADA   :                   0.525424
DTREE:                    0.355932
RF    :                   0.254237
```

- From the above work done based on recall score, I see that best model (AdaBoost) for imbalanced dataset has changed to Logistic Regression model with tuning parameter 'C=10'.
- AdaBoost Model which was the best model on **Imbalanced Dataset** is second best model as it has lower recall value when applied on **Balanced Dataset**.

Recommendations for the Client



Feature Importance findings will help to know criteria contributing most and least to the employee attrition

HR can give more importance on features having high weightage value by saving attrition rate and keeping employees focused and happy.

Model created will be useful to generalize and predict employee attrition for new employee data.

Prepare in advance and work with employee/employees to prevent/decrease attrition

Further Research

- To build a better model on any dataset with imbalanced class as this one requires different resampling using other techniques and more analysis trying different machine learning algorithms.
- My further research includes hyper-parameter tuning for all other algorithms that I used for my project work excluding Logistic Regression

THANK
YOU!