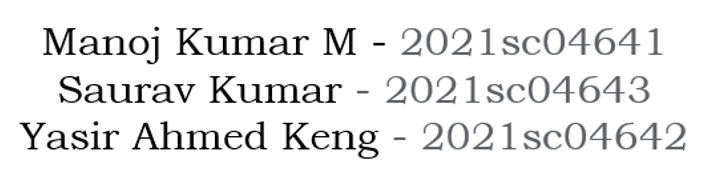
# Group-139 DM Assignment -1 (ATTRITION)



# Analyzing and Building models for Predicting Attrition.

1. Data Science Proposal to classify the Attrition of an employee.

1.Basic Understanding is to observe and predict if the employee will leave the company or not. There are several features provided for thousands of staffs and with the help of these features like age, working hours, salary hike etc we have to come to a conclusion. Predicting this will lead to hiring better employees as well as keeping an eye on the employee performance so that attrition could be reduced to acceptable levels.

2. Exploratory Data Analysis (EDA) & Preprocessing

* The dataset is provided to us in csv format. It has 32 columns (features ) and 1470 rows (employees).
* Several features are of continuous type, others are categorical.
* **Missing Data:** The dataset contains two columns with all values as NULL (Date\_of\_termination & Unnamed: 32), when imported to Jupyter Notebook. These columns are hence removed.
* The **Target Variable** - Attrition contains only two types of values Yes/No.   
   No 1233  
   Yes 237  
  This type of data is called **unbalanced** data where there are only few data whose target variable value is yes and more with no.
* Performed the **Univariate Analysis** and noted the results along with all plots (mentioned in the Jupyter notebook). BoxPlot revealed some **outliers** (dots away from the box lines) but on scrutiny found they are not real outliers.   
  Used **Matplotlib** pyplot and **seaborn boxplot** for continuous features. **Seaborn** **countplot**– for categorical features.
* Performed **Bi-Variate Analysis**. BoxPlot for continuous features and Crosstab for categorical features. Observed several interesting correlations as mentioned in the Jupyter notebook. Also performed Multivarite analysis of some correlated values.
* **Correlation Matrix**: Convert Categorical data to numeric data using **LabelEncoder** from SKLearn and plot using **Seaborn Heatmap.** The higher value of correlation is indicated by darker blue color. We are focusing on more than 60%, for now as follows:  
  JobLevel vs Monthly Income has high correlation.   
  JobLevel vs Total Working years has high correlation.   
  Year with company and years with current manager.   
  Total working years and age.  
  **Feature Engineering:** Could not find much features to exclude or to create a new feature.
* Performed **Normalization** - Scaling the dataset, using SKLearn - **Standard Scaler,** as the data were not uniformly distributed

3. Build a model to classify the Attrition of an employee.

This is a classification problem – predicting attrition possibility of an employee. Initially we divided the whole dataset into two parts using SKLearn library using **train\_test\_split**, and scale it.

* Training dataset – 80%
* Testing dataset – 20%

80/20 is considered generally a good ratio.  
The main purpose of this is to evaluate the trained model with the unseen data for better and correct evaluation.   
  
We have chosen this sklearn.linear\_model – Logistic Regression as baseline model.

**Model 1 - Logistic Regression (Baseline Model)**

The accuracy for the model is approximately 86% but looking at the poor recall for category 1 it seems like this is not the appropriate model for the data.

**Model 2: Choosing a better model for the data lets try using a decision tree classifier.**

The accuracy has drastically dropped by using Tree based model that is a decision tree classifier.

**Model 3: Try RandomForestClassifier of sklearn.ensemble**

The accuracy for the RFC model is less than the LR model.

**Model 4: Boosting algorithms for the data set**

The accuracy score has improved slighly by using a boosting algorithm that is XGBOOST CLF. ACC\_SCORE=87.414%

**Model 5: ADABOOST CLASSIFIER With hyperparameter tuning.**

The accuracy for ADA Boost Classifier is found better than others. Fine tuning is performed to get best results.

**Model 6: Deep Learning approach using Neural Networks.**

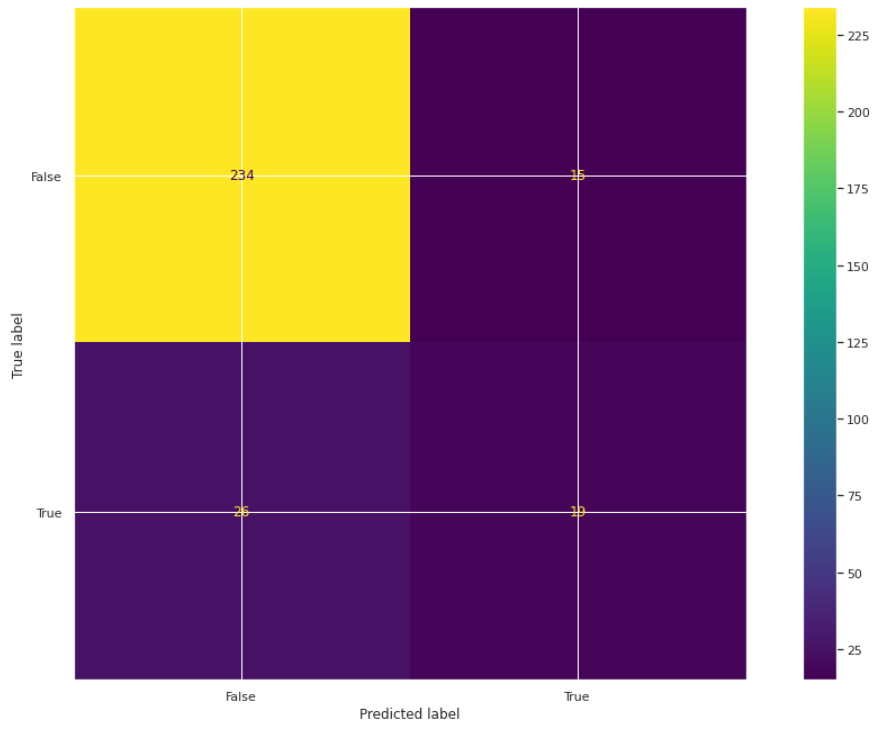
Tried DL using the Neural Network (Keras), still found the above ADABoost to be better.

3. Selecting the best Classifier

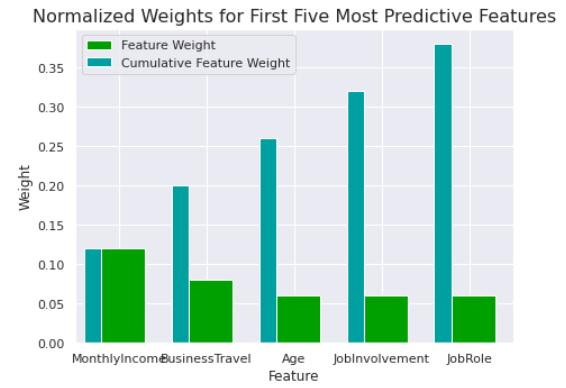
**The ADABoost model was able to classify with better results than other models.** For Evaluation of the model we have used Accuracy and Recall.

(The below mentioned scores are of test dataset only)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Recall of 0** | **Recall of 1** | **Accuracy** |
| Logistic Regression | 0.98 | 0.24 | 0.87 |
| Decision Tree Classifier | 0.84 | 0.33 | 0.77 |
| Random Forest | 0.98 | 0.18 | 0.86 |
| XGBoost | 0.97 | 0.36 | 0.87 |
| ADABoost | 0.94 | 0.42 | 0.86 |
| Neural Network – Deep Learning | 0.94 | 0.22 | 0.83 |

4. Confusion Matrix of the final selected classifier (ADABoost)  
  


5. Top five features used by the model for classification.



**Summary:**

We verified the dataset and performed necessary corrections. Splitted the data into Train dataset and Test dataset. Built several data models based on the EDA. Tested the dataset and found the ADABoost model was able to classify and predict the Attrition with better results than other models, as per the above comparison chart(3). We found some features were effecting the Attrition much more than others, as mentioned in above chart (5).