IBM HR Analytics Employee Attrition & Performance, Goals and Machine Learning

Capstone Project by Sneha Rani

Introduction

Attrition is the normal life cycle of employment. Attrition refers to employees who leave their jobs due to normal circumstances. In other words, employees are leaving not because they have a problem with the company or their jobs – it's a matter of life unfolding.

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.

It's expensive, non-productive and frustrating. It is not caused by one single factor rather it involves multiple factors which need to be defined and analysed further to determine the impact which leads to the high attrition rate.

Employee attrition, a big cause for concern for firms, ranges between 15 per cent and 20 per cent. It has been known to exist all long. However, with technology changing rapidly and manpower costs increasing, attrition is high and hurts badly. This can lead the company to huge monetary losses by these innovative and valuable employees. Someone well said

"You don't build a business. You build people, and people build the business"

Companies that maintain a healthy organization and culture are always a good sign of future prosperity. 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. These predictive insights give managers the opportunity to take corrective steps to build and preserve their successful business.

Problem

Our client is IBM and they want to know the factors that lead to employee attrition. A major problem in high employee attrition is its cost to an organization. The client would like to know the contribution of varied factors and analysis to determine which factors contribute the most to employee attrition. In order to address this problem I need to predict the employee attrition based on the listed factors and derive the top factors that lead to high attrition.

Approach

1) I will build machine learning models which will predict the attrition based on the features.

2) I will derive the relationship between the feature score and attrition.

3) Create actionable suggestions for improving the likelihood of retention.

Data Pre-processing

We are going to use the data present in the source link:-

https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

The data was downloaded from Kaggle. It is pretty straightforward and clear. Each row represents an employee; each column contains employee attributes. My dataset consists of 35 features and consist of (Int and Object) data type which include:

Data Dictionary

Column Name	Data Types	Column Description
Age	Int64	Age of each employee.
Attrition	Object	Attribute 'yes' and 'no' indicating employee left or stayed in the company.
BusinessTravel	Object	Travel frequency of the employee.
DailyRate	Int64	Daily wage rate of the employee.
Department	Object	List of different department where the employee works.
DistanceFromHome	Int64	Distance from home to work location of each employee.
Education	Int64	Level of education completed.
EduactionField	Object	Area of education.
EmployeeCount	Int64	Count of each employee.
EmployeeNumber	Int64	Employee Id number
EnvironmentSatisfaction	Int64	Satisfaction level of employee in the environment.
Gender	Object	Male/Female employee.
HourlyRate	Int64	Hourly wage rate of each employee.
JobInvolvement	Int64	Level of involvement in the job.
JobLevel	Int64	Categorical level of job of each employee.
JobRole	Object	Designated job role.
JobSatisfaction	Int64	Categorical job satisfaction level.
MaritalStatus	Object	Single/Married/Divorced
MonthlyIncome	Int64	Income of each month
MonthlyRate	Int64	Monthly wage rate
NumCompaniesWorked	Int64	Number of previous company employee had worked.
Over18	Object	Employee meets over 18 criteria or not.
OverTime	Object	Indicates overtime of each employee.

PercentSalaryHike	Int64	Salary hike percentage.
PerformanceRating	Int64	Categorical rating of performance.
RelationshipSatisfaction	Int64	Level of relationship satisfaction between employee and company.
StandardHours	Int64	Standard working hours.
StockOptionLevel	Int64	Stock option level given to each employee
TotalWorkingYears	Int64	Total work experience.
TrainingTimesLastYear	Int64	Time spent by each employee during training.
WorkLifeBalance	Int64	Balance level of each employee in their work life.
YearsAtCompany	Int64	Number of years in same company.
YearsInCurrentRole	Int64	Number of years in current role.
YearsSinceLastPromotion	Int64	Number of years since last promoted.
YearsWithCurrManager	Int64	Number of years with current manager.

Tools Used

Pandas: Loading the data, data wrangling and manipulation.

Scikitlearn: Libraries for classifiers, Model evaluation, Metrics, Cross validation, Feature Importance

MatplotLib and Seaborn: Data Visualization

Data Cleaning / Data Wrangling

Employee attrition dataset was downloaded from the data source link in the .csv file format and then it was imported on the Jupyter notebook using pandas. As the dataset was precleaned, there was no missing/null values in any of the columns. Dataset was then checked to know data types, shape and definition of each column in order to filter the column. Certain columns like DailyRate, EmployeeCount, EmployeeNumber, JobLevel, MonthlyRate, Over18, PerformanceRating, StandardHours, StockOptionLevel, TrainingTimeLastYear were found to be insignificant and was removed from the final dataset for further analysis.

Exploratory Analysis

- Certain initial findings were then performed for my dataset which included interesting questions such as:
- Which age-group people contribute maximum attrition?
- What is the count of married people and unmarried people attrition rate? Are married people more prone to attrition?
- What is the count of people working OverTime and YearsInCurrentRole? How working overtime (or not), and the years in role relate to employee attrition?

- What is the count of attrition of each department on the basis of RelationshipSatisfaction? Does satisfaction level has any impact on employees leaving these department?
- Do JobSatisfaction and JobRole impact gradual loss of employees? Are these two features have a common pattern?

To start off, age was grouped into four categories as young adults (15-24 Yrs), mid age adults (25-40 Yrs), mid to old adults (41-54 Yrs) and old age adults (55-64 Yrs) and graph was plotted which showed maximum contribution of different age- group people based on the attrition(Fig 1).

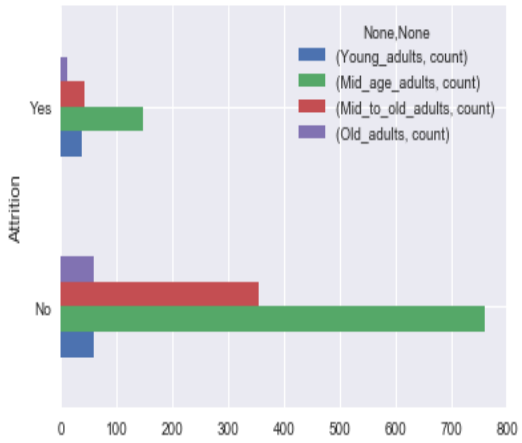


Fig 1

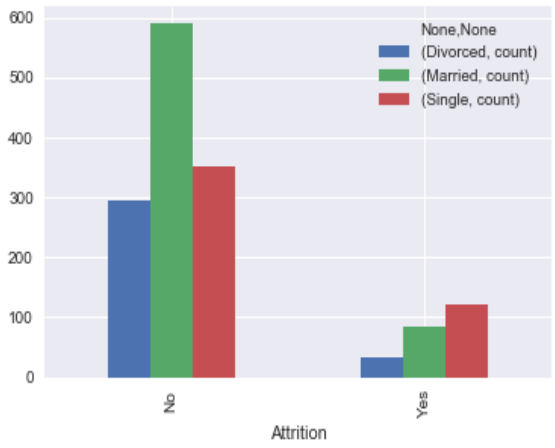


Fig 2

My second analysis was based on the marital status and was surveyed that the people who were 'Single' contributed more towards attrition (Fig 2). Next analysis gave the information about the people who had worked overtime under same current role left the company most (Fig 3). Next thing figured out was research and development department contributed maximum to the attrition and relationship satisfaction level had no special role to be played (Fig 4). Another important initial finding on job role and job satisfaction showed that employees under role of sales executive, research scientist, laboratory technician left the company most having low job satisfaction (Fig 5).

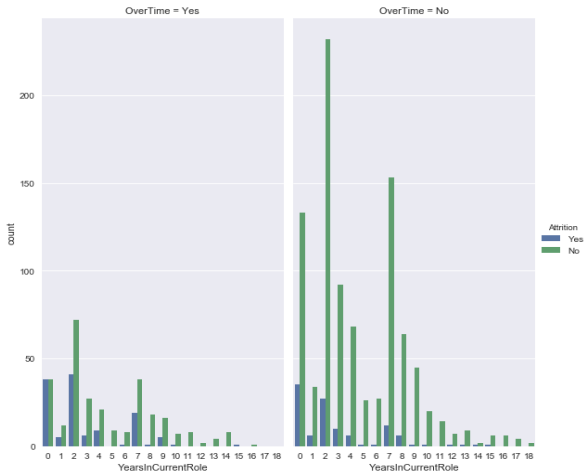


Fig 3

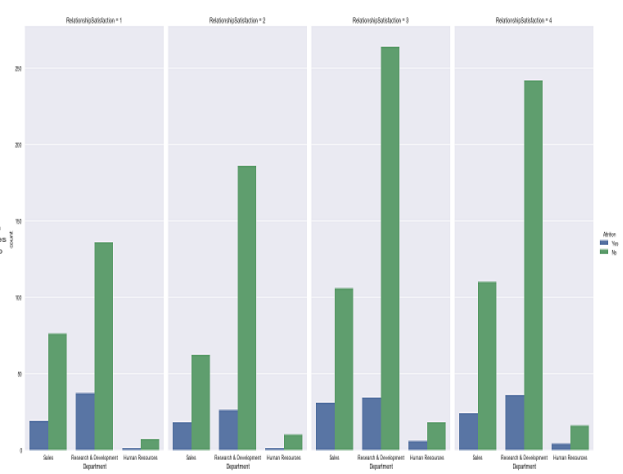


Fig 4

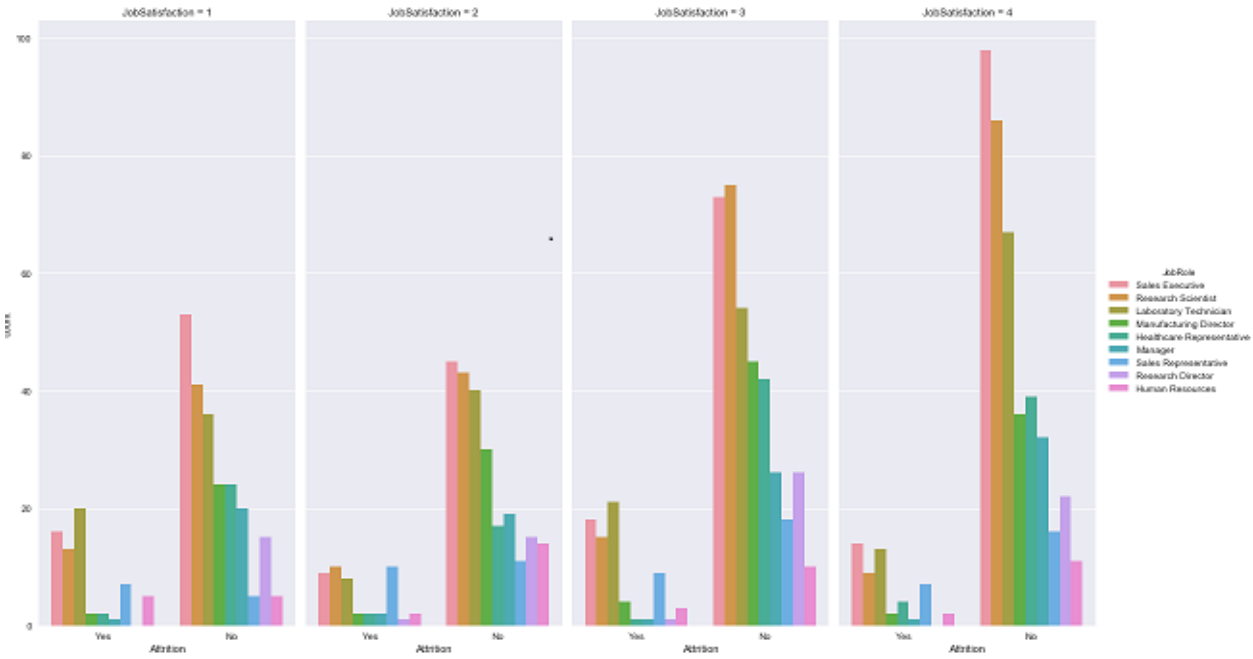
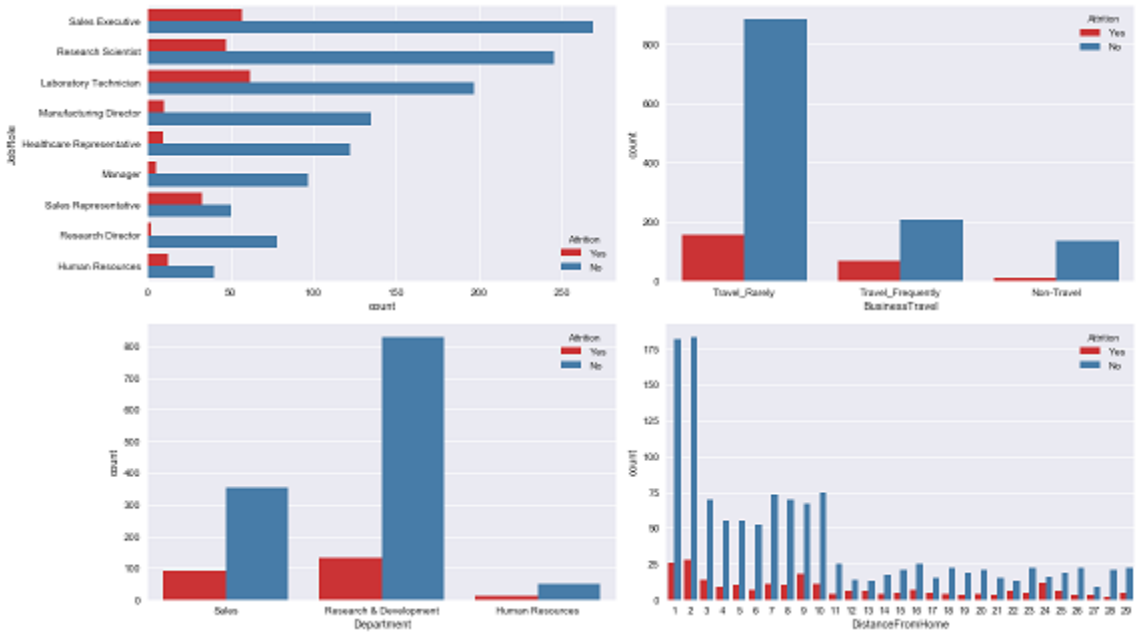
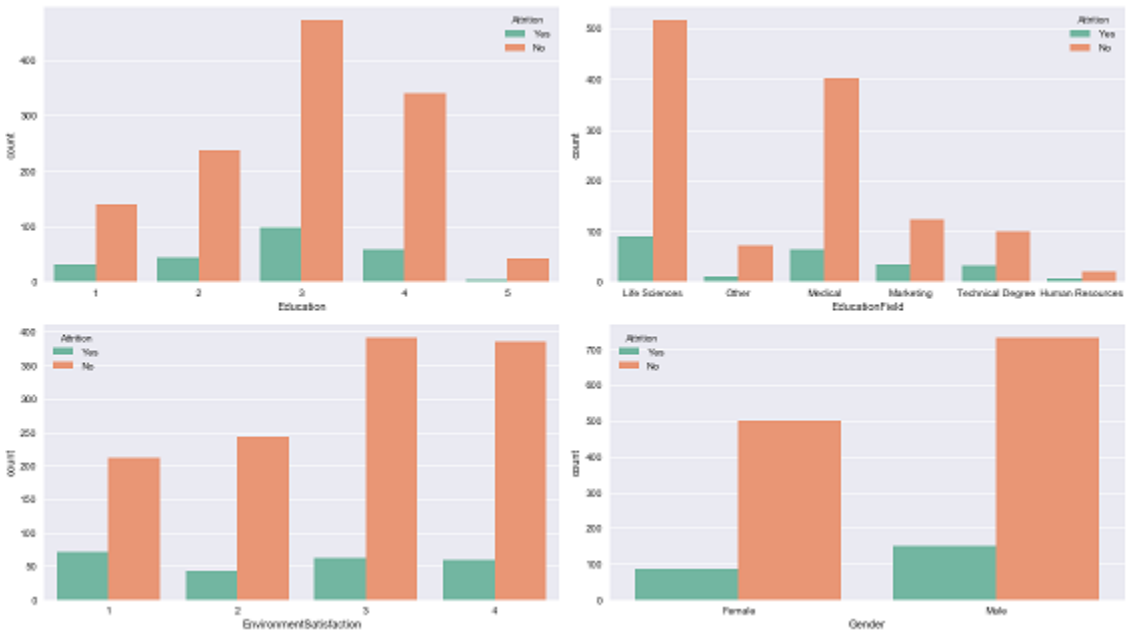
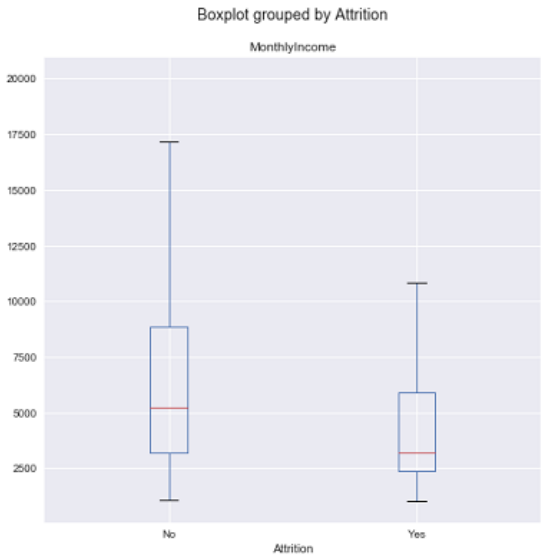


Fig 5

Very importantly, to know each variable influence on attrition of the organization, each variable was independently plotted against attrition with some sort of an interpretation. Some of the plot like factor plot, bar plot, box plot were used for most of the visualization which depicts the relationship on attrition out of which few were plotted as shown below:







As I was getting closer with the analysis done earlier on the age-group, some interesting facts revealed out which made me to go for further investigation. Few, statistics were calculated using statistical inferential for finding the relationship between two variables ‘Age’ and ‘Attrition’. In order to find the dependency between these two variable, Chi-Square Contingency Test was performed which revealed that the potential dependency between those two variables were statistically significant. Next Chi-Contingency test was performed for deciding independence between ‘Gender’ and ‘Attrition’. Even though male and female contributed to the attrition rate, Chi-Square Contingency Test of Independence showed high p-value which synced with the null hypothesis that these variables were independent.

Modeling the Data

Now knowing the fact that how each variable behaved on attrition, it was time to prepare my first base line model using logistic regression analysis. Let's stop for a while here. It was now time to take a deep breath think logically what basically I have done and what I need to do better prediction.

We all are aware about the real time known facts that attrition brings decreased productivity. People leave causing others to work harder which contributes to more attrition, in turn contributes to increased cost and lower revenue. This forces additional cost and austerity measures. There could be various ways to address this problem but to correctly understand the problem and think logically in better approach or applying best technique gives better way to move ahead and find best solution by choosing best model. What I think is that it depends on us, how well we understand these real time problems and address them correctly.

Before moving forward to build the best model and come out with any type of model prediction, these were certain question that had come up was 'Do really I have sufficient data to apply machine learning?' 'Is my model ready to be applied for future prediction based on attrition whether employee will leave the company or not? This was one of my experience that I faced and the challenges that had come while working on my dataset.

After figuring out my initial findings and having laid a foundation in Inferential Statistics and Hypothesis Testing, it's was time to see those ideas in action by implementing the best model and applying Machine Learning Algorithms based on the dataset. Basically we needed to first understand what was meant by Machine Learning and it's actually an art and science of giving computers the ability to learn to make decisions from data without being explicitly programmed.

As I am dealing with IBM Employee Attrition dataset, my goal is to know the contribution of varied factors and predict which factors contribute the most to employee attrition, seeing all these qualities I was now able to identify where my dataset falls under? Let me tell something about supervised and unsupervised learning before stepping ahead for those who might not be aware about it.

With Supervised learning, we have clearly labelled dependent and independent variables. The dependent variable (target) is known. For example, for my dataset 'Attrition' is the target variable. If we do have a clearly labelled *y* variable, we are performing supervised learning because the computer is learning from our clearly labelled dataset. It is learning the relationship between our *X* variables and our *y* variables. Supervised learning can be broken down into regression and classification problems based on continuous and categorical data of target variable. With unsupervised learning, we do not have a clear dependent variable. Based on my dataset and target variable 'y', I performed classification task.

Before preparing my model, I had created dummy variables for categorical type for easy interpretation. Now, the final data was ready to be modelled with first column as 'Attrition' in order to define the parameter with 'X' (features) and 'y' (target). It is denoted by Matrix 'X' and vector 'y' in terms of classification.

When we talk about fitting the model we would like to ensure two things - we have found the best model (in terms of model parameters) and next the model is highly likely to generalize i.e. perform well on unseen data. I tried building logistic regression base line model by using 'L1' and 'L2' regularization. The data was first split into a training and test (hold-out) set. Data was then trained on training set and tested for accuracy on testing set. After seeing the classification result it was analysed that even though the accuracy was good around 80% for both L1 and L2 regularization – the recall, precision and f1 score were not that good for the training and testing datasets which meant that the larger class (Class 0) i.e. Not-Attrition is over-influencing the model which meant that there was class imbalance which is most common problem associated while dealing with classification problems.

Here the number of data points belonging to the minority class (in our case, "Attrition") was far smaller than the number of the data points belonging to the majority class ("No Attrition").

Tuning the Model

In order to find the resulting model ‘M’ in my case, I used Grid Search to tune my model using best hyper parameter and corresponding penalty. In Logistic Regression, the most important parameter to tune is the regularization parameter `C`. It is very important, because we need to make sure that our model is general and it works beyond our data sets. In other words it should ideally work on data it has never seen. Still classification result was more or less same as was for base line model and the recall, precision and f1 score were not good. My next step was trying different algorithms like Random Forest classifier, CART, Ada Boost classifier and finding the highest recall score.

Here we need to notice that the high recall score means when employees left the company, model predicted correctly most of the time. Therefore, a good recall score is important to avoid unpredicted loss of the employees in the company.

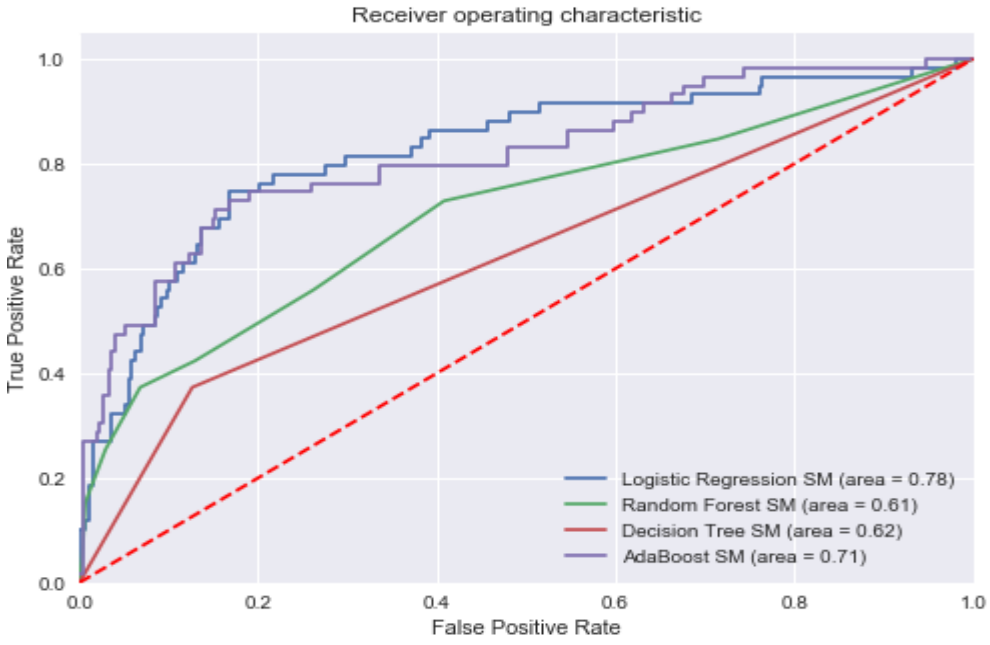
From the above ROC Graph, we found that the Ada Boost Model had the highest area under curve value of 0.70 and was more close to the top left corner which is one of the criteria for a good model.

Feature Importance

I tried computing the relative importance of each attribute by finding highest positive and negative coefficient by absolute value. This could be further used to inform a feature selection process. From the feature importance analysis report, I found that the features like OverTime_Yes, JobInvolvement_Low with the high positive coefficients (towards attrition) made some sense and were somewhat correlated to the things which influence employees to leave the company in short duration of time. On the other hands, the features like OverTime_No, JobSatisfaction_Very High, JobInvolvement_Very High with high (absolute value) negative coefficients (towards non-attrition) made sense too and were somewhat correlated to the things which influence employees to stay at the company for the longer duration of time.

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) and 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).

In short after analysing all classification reports, Ada Boost Classifier model with recall score of 0.46 showed best model among all. Also, the training performance was better than test performance which indicated that model is neither over-fitting nor erroneous. But to combat **Imbalanced Classes**, I tried **Synthetic Samples (SMOTE)** by randomly sampling the attributes from instances in the minority class as majority class data had over influenced the model. I repeated training of all the above models on the new balanced dataset and also created ROC curve comparing each of these models.



Logistic Regression Model ‘M1’ with tuning parameter C=10 was having higher recall score and so was concluded the best model for this dataset.

Conclusion

Based on all the above work, I developed a machine learning model which fairly predicts the employee attrition in the company based on all the features. This work also handled a real time issue of data imbalance and developed a model countering it. I was also able to logically build the top five features that may impact employee attrition and top five features that influences employee retention.

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.