IBM HR Analytics: Employee Attrition

Springboard DSC Capstone Project 1 by Sneha Rani

Introduction

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

In mathematical terms, 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 are associated with employee attrition would 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 and Client

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 attrition.

Approach

- I built machine learning models which will predict the attrition based on the features.
- I derived the relationship between the feature score and attrition.
- I offer 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 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 departments 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.

Scikit-learn: Libraries for Classifiers, Model evaluation, Metrics, Cross-Validation, Feature Importance

Imbalance-Learn: SMOTE

Data Visualization: MatplotLib, and Seaborn

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 determine data types, shape and definition of each column in order to filter the columns. 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).

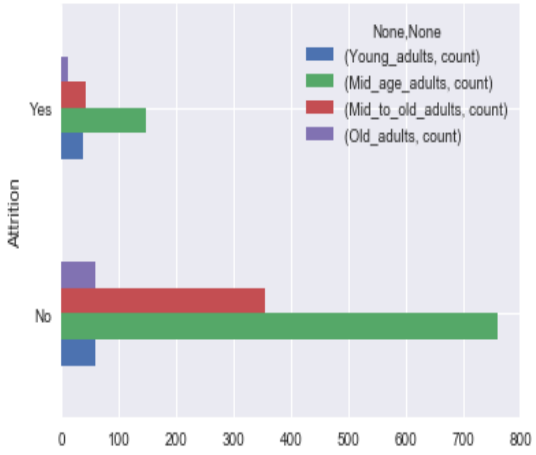

Figure 1

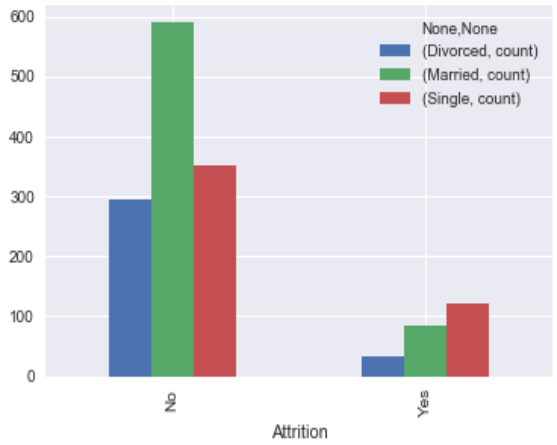

Figure 2

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 was research and development department contributing 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).

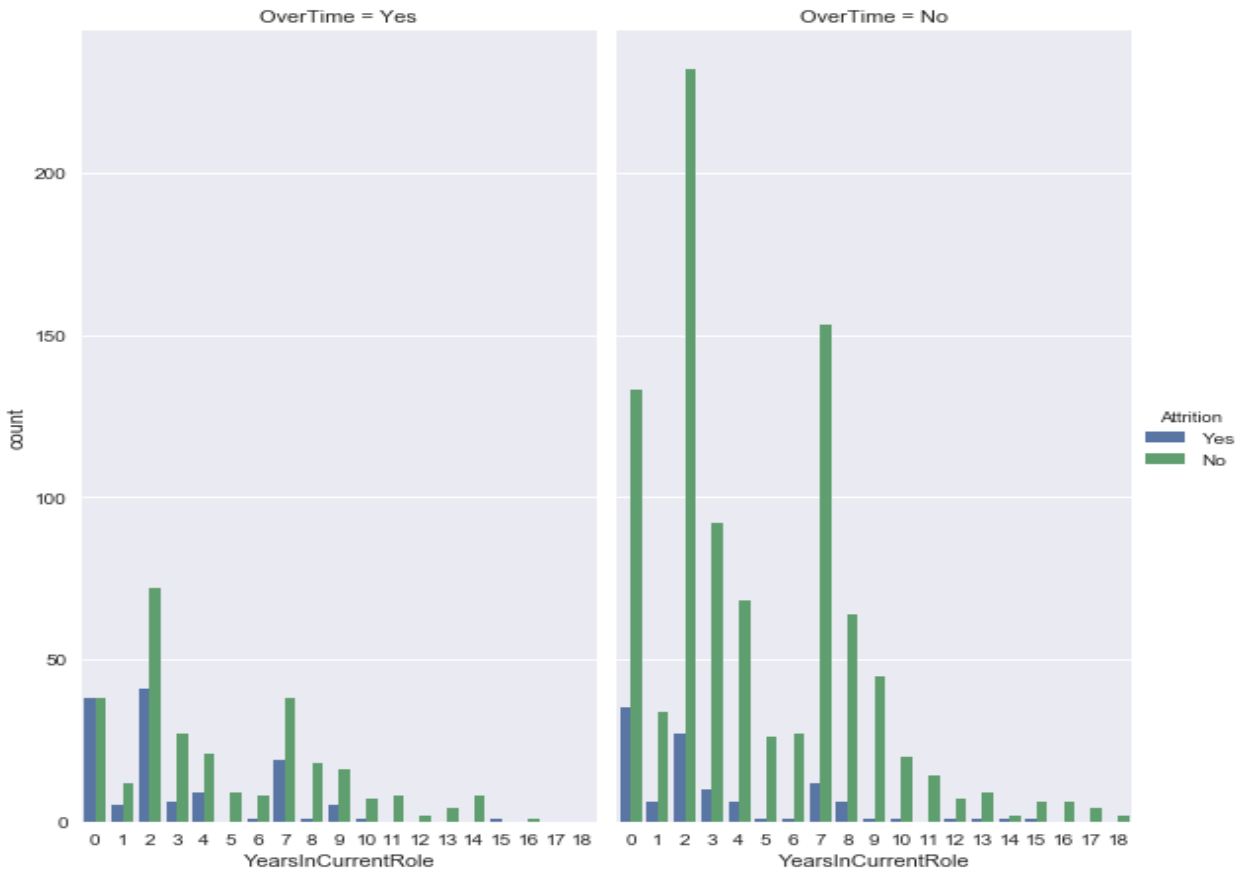
OverTime = Yes
OverTime = No
count
YearsInCurrentRole
YearsInCurrentRole
Attrition
Yes
No

Figure 3

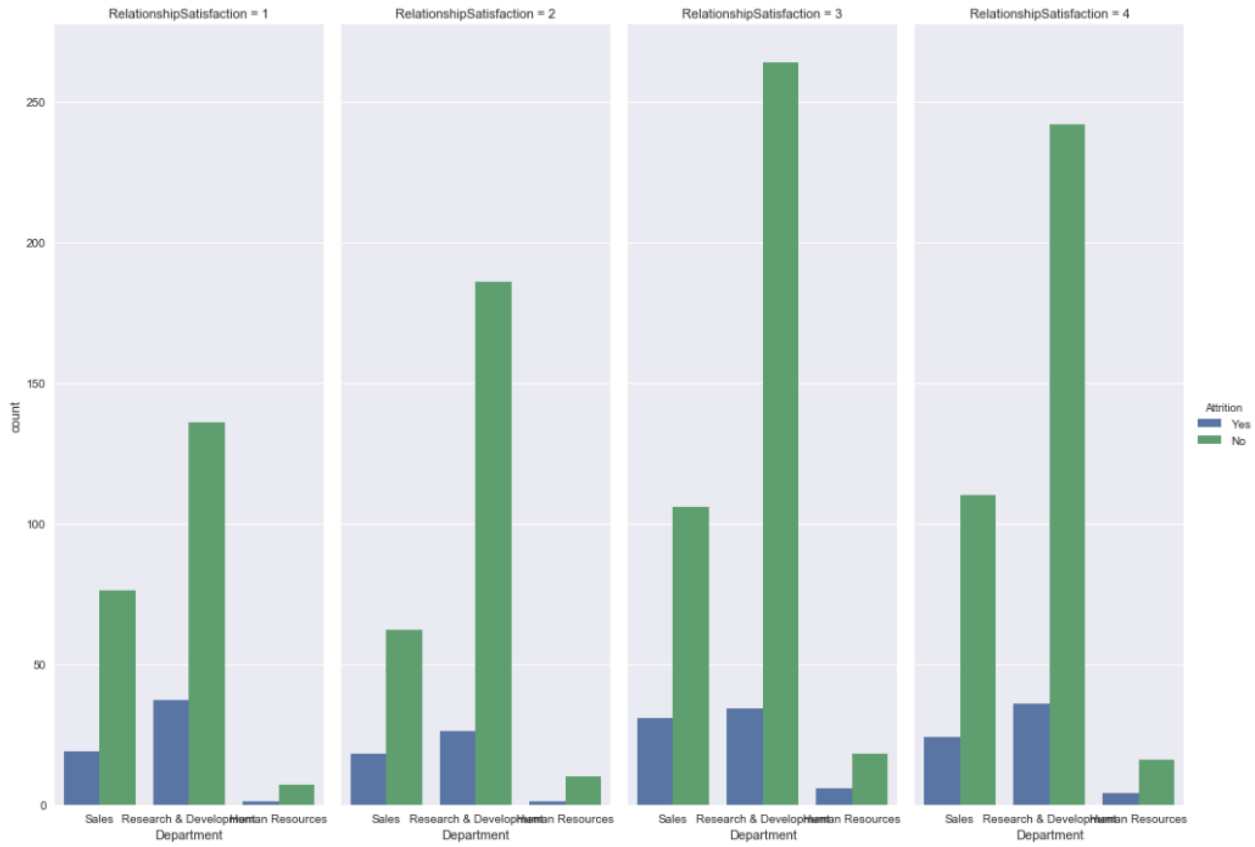
RelationshipSatisfaction = 1
RelationshipSatisfaction = 2
RelationshipSatisfaction = 3
RelationshipSatisfaction = 4
count
Sales
Research & Development
Human Resources
Department
Attrition
Yes
No

Figure 4

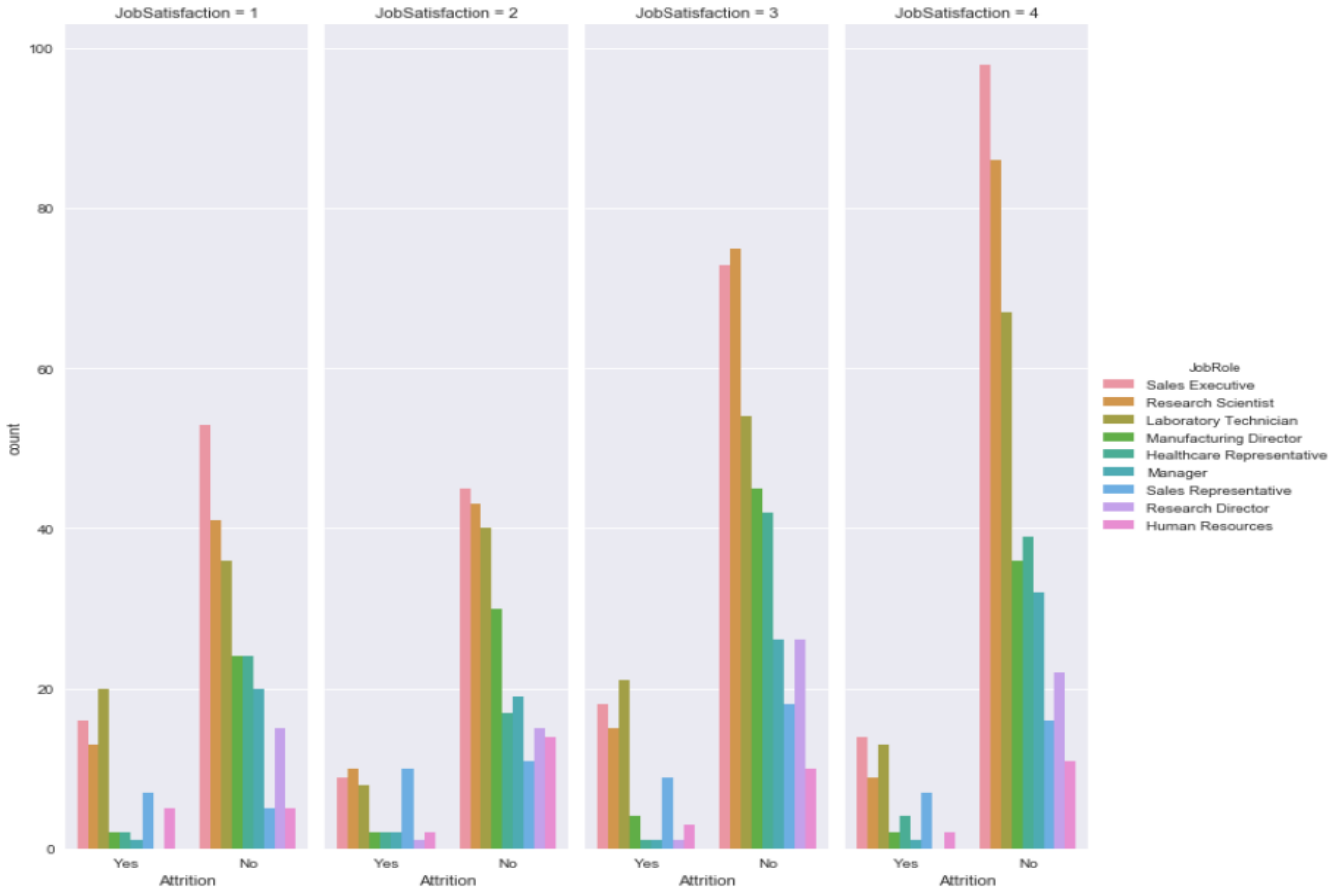

Figure 5

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 (Fig. 6, Fig. 7, and Fig. 8):

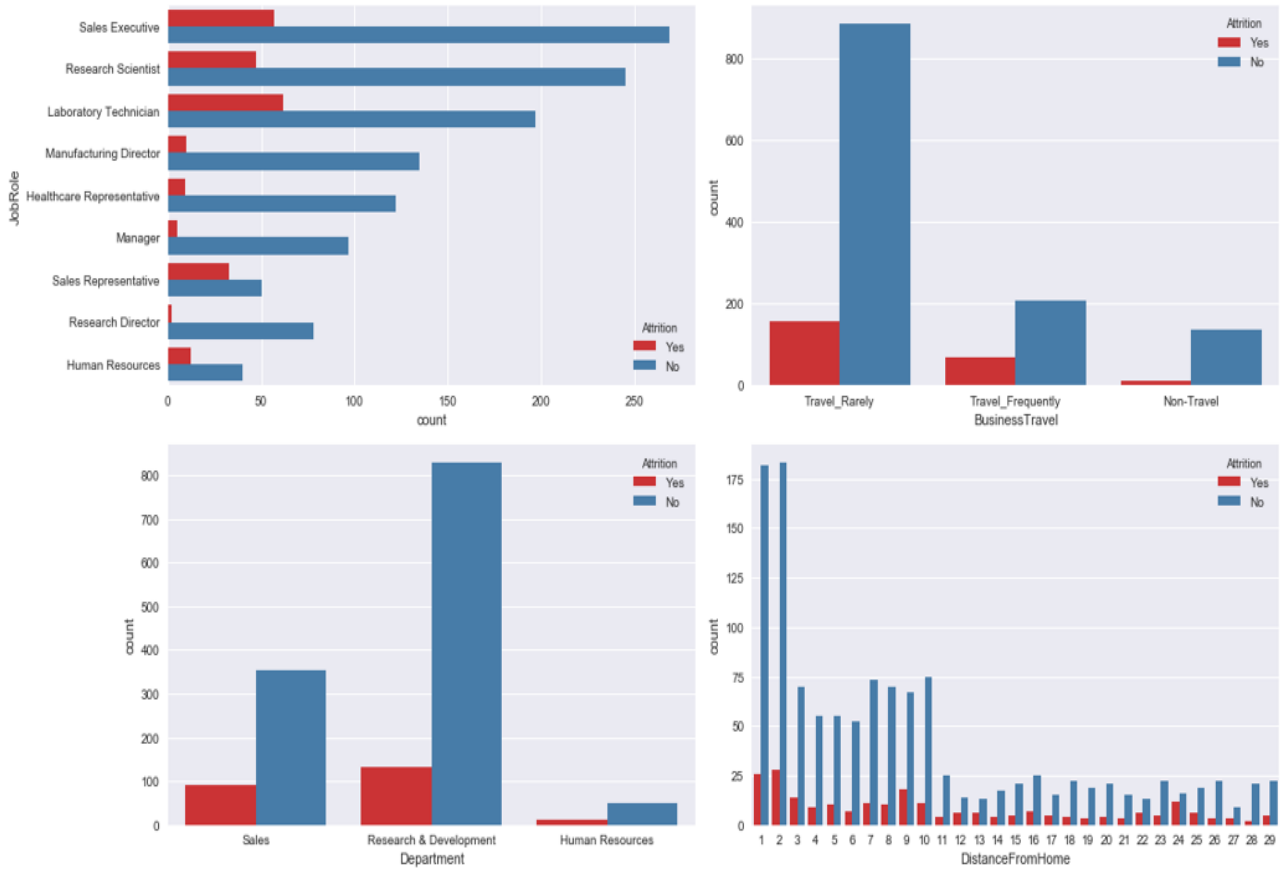

Figure 6

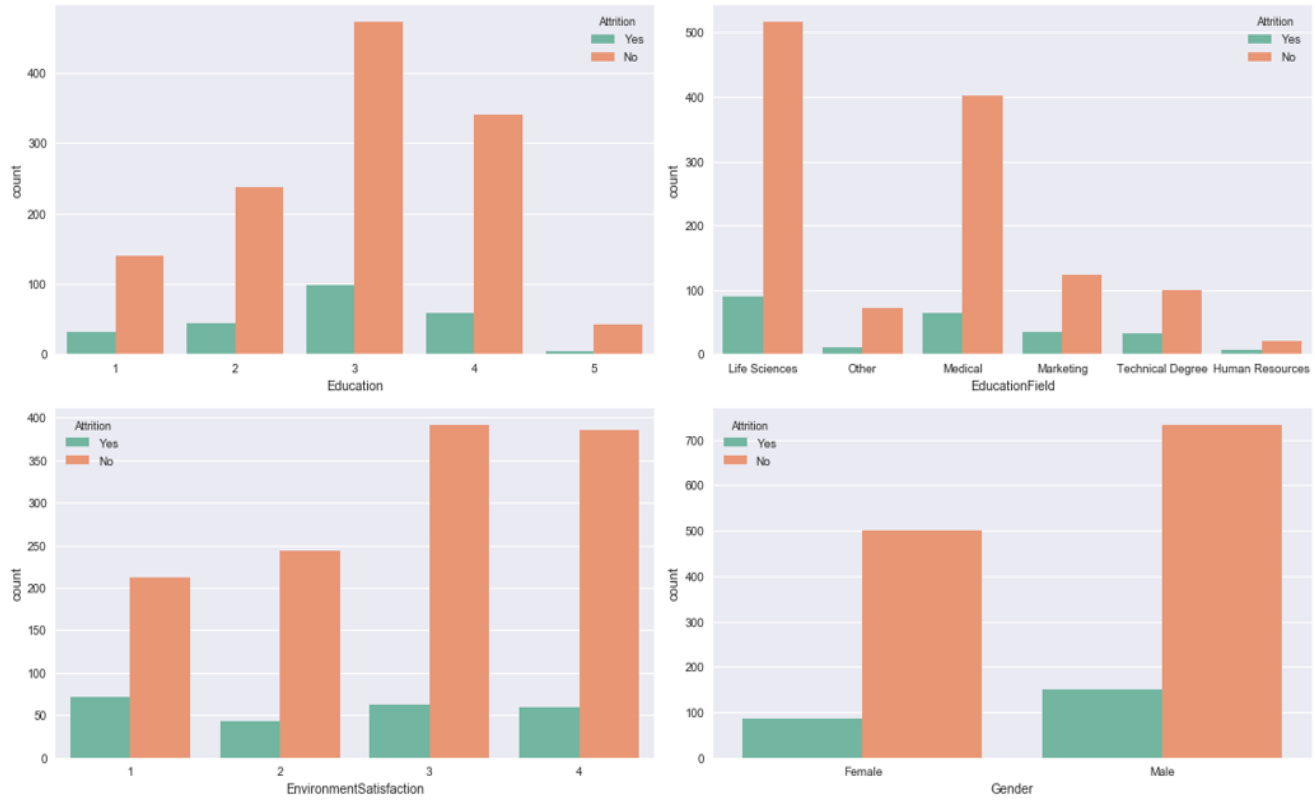

Figure 7

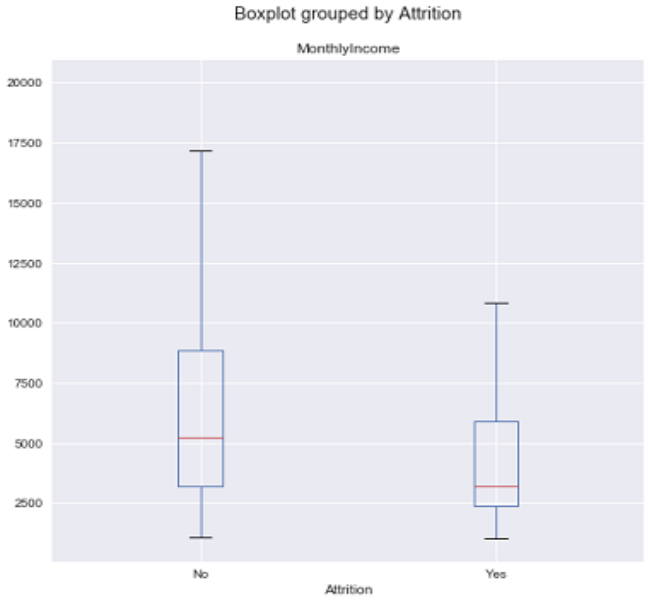

Figure 8

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 inferential statistics for finding the relationship between two variables 'Age' and 'Attrition'. In order to find the dependency between these two variables, 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 Test of Independence result shows high p-value(0.2906) which is higher than threshold p-value of 0.05, so we cannot reject the null hypothesis and makes me think not to be a useful feature to include in a predictive model, but I am not sure until I try all the variables in machine learning model.

Modelling the Data

Now knowing the fact that how each variable behaved on attrition, it was time to prepare my first baseline model using logistic regression. Before preparing my model, I had created dummy variables for categorical type for easy interpretation. I created dummy variables on BusinessTravel, Department, Education, EducationField, EnvironmentSatisfaction, Gender, JobInvolvement,JobRole, JobSatisfaction, MaritalStatus, OverTime, RelationshipSatisfaction, WorkLifeBalance. After creating dummy variables, final dataframe created includes 65 columns.

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. After the split, train data has 1102 samples where as test data has 368 samples.

Data was then trained on training set and tested for accuracy on testing set. With L2 regularization, accuracy score was 0.8723 and 0.8956 for testing and training set. With L1 regularization, accuracy score was 0.8723 and 0.8938 for testing and training 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(for L1 regularization value was 0.39, 0.49 and 0.68 respectively for Class 1) 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 Models

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 following algorithms: Random Forest classifier, CART, Ada Boost classifier and finding the highest recall score.

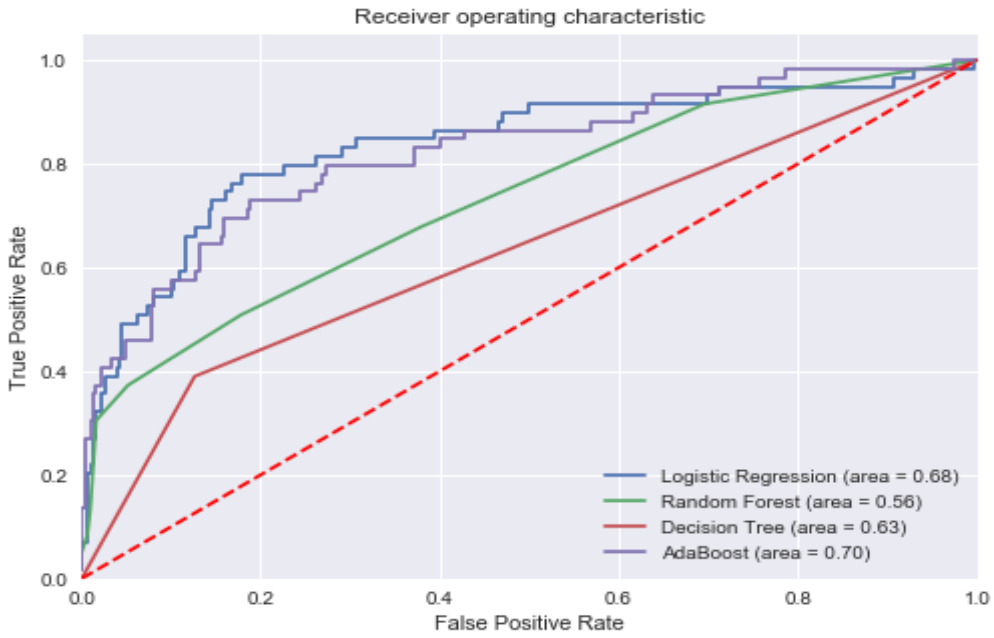

Figure 9

From the above ROC Graph (Fig. 9), 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.

Top 5 positive coefficient

```
OverTime_Yes                  1.105801
JobInvolvement_Low            0.965035
MaritalStatus_Single          0.874477
EnvironmentSatisfaction_Low   0.869302
RelationshipSatisfaction_Low  0.850258
Name: positive, dtype: float64
```

Top 5 negative coefficient

```
OverTime_No                   0.965188
JobSatisfaction_Very High     0.701842
BusinessTravel_Non-Travel     0.630154
JobInvolvement_Very High      0.606149
WorkLifeBalance_Better        0.584350
Name: negative_abs, dtype: float64
```

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 associated with 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 associated with 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).

Conclusions

Conclusion before Resampling Dataset

```
All Results across Models on Imbalanced Dataset
--------------------------------------------------------------------------------
Name    Class     Precision              Recall                 Fscore    Support
LR   : Class 0 0.8922155688622755 0.9644012944983819 0.9269051321928461 309
LR   : Class 1 0.6764705882352942 0.3898305084745763 0.49462365591397844 59


ADA  : Class 0 0.9012345679012346 0.9449838187702265 0.9225908372827805 309
ADA  : Class 1 0.6136363636363636 0.4576271186440678 0.5242718446601942 59


DTREE: Class 0 0.8679245283018868 0.8932038834951457 0.8803827751196173 309
DTREE: Class 1 0.34 0.288135593220339 0.3119266055045872 59


RF   : Class 0 0.8547486033519553 0.9902912621359223 0.9175412293853074 309
RF   : Class 1 0.7 0.11864406779661017 0.20289855072463767 59
```

Figure 11

Above mentioned classification report (Fig. 11) shows performance score of models on test set of imbalanced dataset. AdaBoost classifier with recall score of 0.46 showed highest among all models and hence was recorded as the best model for this dataset. Even though

accuracy score of AdaBoost classifier (0.87) was good than resulting model 'M', the recall score of class 0 (majority) and class 1(minority) had still a huge difference which showed that majority class had over-influenced the model.

I also built the top five features that may impact employee attrition and top five features that influences employee retention using feature importance which was one of the requirement of the client.

Conclusion after Resampling Dataset

```
All Results across Models on Balanced Dataset using SMOTE
--------------------------------------------------------------------------------
Name    Class     Precision             Recall                Fscore    Support
LR   : Class 0 0.9389312977099237 0.7961165048543689 0.861646234676007 309
LR   : Class 1 0.4056603773584906 0.7288135593220338 0.5212121212121211 59


ADA  : Class 0 0.9102564102564102 0.919093851132686 0.9146537842190016 309
ADA  : Class 1 0.5535714285714286 0.5254237288135594 0.5391304347826087 59


DTREE: Class 0 0.8778135048231511 0.883495145631068 0.8806451612903227 309
DTREE: Class 1 0.3684210526315789 0.3559322033898305 0.3620689655172414 59


RF   : Class 0 0.8724637681159421 0.9741100323624595 0.9204892966360857 309
RF   : Class 1 0.6521739130434783 0.2542372881355932 0.36585365853658536 59
```

Figure 12

Above mentioned classification report (Fig. 12) shows performance score of models on test set of balanced dataset. We see from the above comparison table that the resulting model 'M1' (Logistic Regression) show high recall score of 0.73(Fig.13) with a good increment while comparing with the previous model (AdaBoost) with low recall score of 0.52 indicating when employees left the organization model predicted correctly most of the time. This shows unpredicted loss of employees in an organization.

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

          0       0.84      0.81      0.83       924
          1       0.82      0.85      0.83       924

avg / total       0.83      0.83      0.83      1848

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

          0       0.94      0.80      0.86       309
          1       0.41      0.73      0.52        59

avg / total       0.85      0.79      0.81       368
```

So the model developed now will generalize well and 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.

Recommendations for the Client

Above findings can help the client in different aspects and can become instrumental in deciding the employees retention. Feature Importance is another finding which did provide the top features that contributed to employee attrition and also top features that helped employees to stay in the organization for longer duration of time. This will help HR team to work on the employee-retention program by giving more importance on those features having high weightage value by saving attrition rate and keeping employees focused and happy.

Model created can also be used to generalize and predict employee attrition for new employee data. Model can predict if an employee is going to leave the organization or not. This will really help company to prepare in advance and improve on the area which is impacting negatively to the employees.

So, client should know that employees are the biggest assets of an organization. If an organization take good care of their employees they will replicate this feeling to their clients and customers making an organization stand out in the crowd.

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.

Resources Used

- https://towardsdatascience.com/a-data-science-workflow-26c3f05a010e
- http://contrib.scikit-learn.org/imbalanced-learn/stable/
- https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/
- http://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html
- http://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html
- http://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html
- https://machinelearningmastery.com/feature-selection-machine-learning-python/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3648438/
- https://web.stanford.edu/class/psych252/cheatsheets/chisquare.html#way-classification-contingency-test
- https://stackoverflow.com/questions/34052115/how-to-find-the-importance-of-the-features-for-a-logistic-regression-model
- http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/