Predicting Attrition of IBM employees

Yasemin Ceyhan

Capstone Project 1
Nisan 13, 2018

CONTENTS

Introduction

The employee attrition and turnovers are always issued for either small or large businesses. Now, more than ever, business leaders need to extract more inside from trends changing with employee attrition which impacts revenue and profits of todays' organizations. The biggest challenge in this case is retention of valuable employees in order to provide the continuity of steady state organizations. The department of the human resources is responsible for not only to set reasonable immediate pay for employees but also to seek motivational factors to raise employees performance, which has a great contribution on business. By identifying factors that have impact on employees' attrition and performance, the company will set HR policies during the decisions making process. Otherwise, the company would spend more time and money to hire a new employee.

The department of human resources of IBM (International Business Machine Corp.), Technology Company, represents the client of this project. Their concern is how they can detect the top key drivers of attrition and their importance by size to predict the attrition in order to make or tailoring some reform policies to prevent attrition, if an employee is more likely to leave from the company.

Data Acquisition and Cleaning

We acquire data from ibm.com. IBM data scientist created the data as a fictional data related to the IBM Human Resources (HR). Before exploring the data set, some initial examinations were applied during the data wrangling process. Data consists of 35 columns and 1470 observation. There is no missing data detected. Data contains features all related to IBM employees from regardless of belonging of any particular time frame,

so there is no attribution related to time. The data set is separated in two groups under “Attrition” column as one portion belonging employee who turned over as “Yes” and other portion consisting of employee who stays in job displaying as “No”. Each row in the attrition data set corresponds a unique employee with details such as age, gender, marital status, Business Travel, Distance from home, job level, department and many other features related to years of experience in the company and performance.

Data Exploration

During the data cleaning process, numerical, categorical and binary variables were identified. Some columns such as “Over18”, “EmployeeCount” and “StandardHours” contain one type of value, so they were deleted. Also, since “EmployeeNumber” didn’t generate anything, it was deleted too. There were no missing values or non-unique values examined. There were 1233 employees who stayed in job and 237 employees who turned over. In order to see related features of employees who were in attrition, new data set called 'df_yes' by filtering the data set 'data' where Attrition='Yes'.

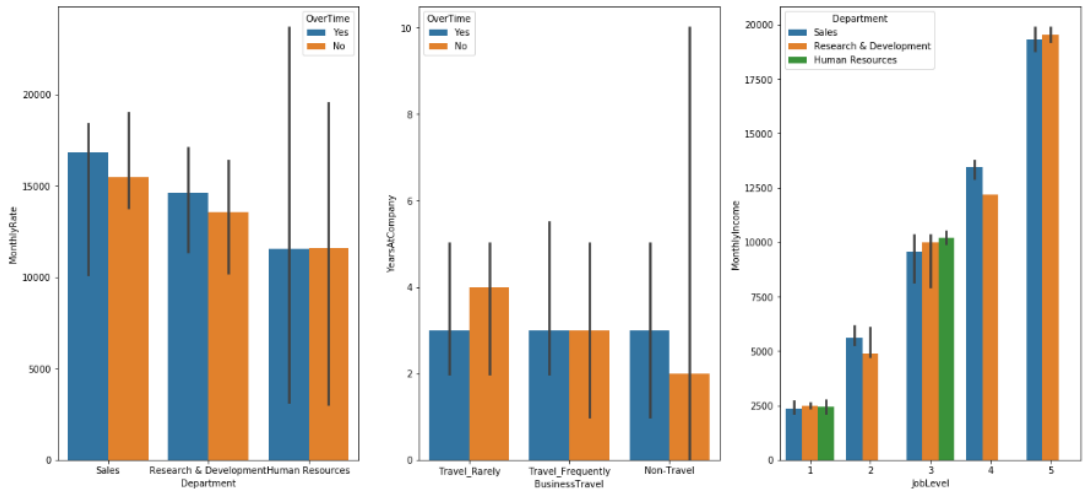

Figure1 – Various Factors for Employees in Attrition

According to the figure 1 displayed below, from left to right, OverTime is characteristics factor to examine MonthlyRate in different Department units such as “Sales” and “ Research & Development”, but not in the department of Human Resources. On the other hand, employees who made no OverTime and about 3 years at company are from either travel_frequently or Non_Travel. Employees who had JobLevel 5 from Research & Development Department having around 20000 monthly income are more likely to leave employees from Sales department having similar categories.

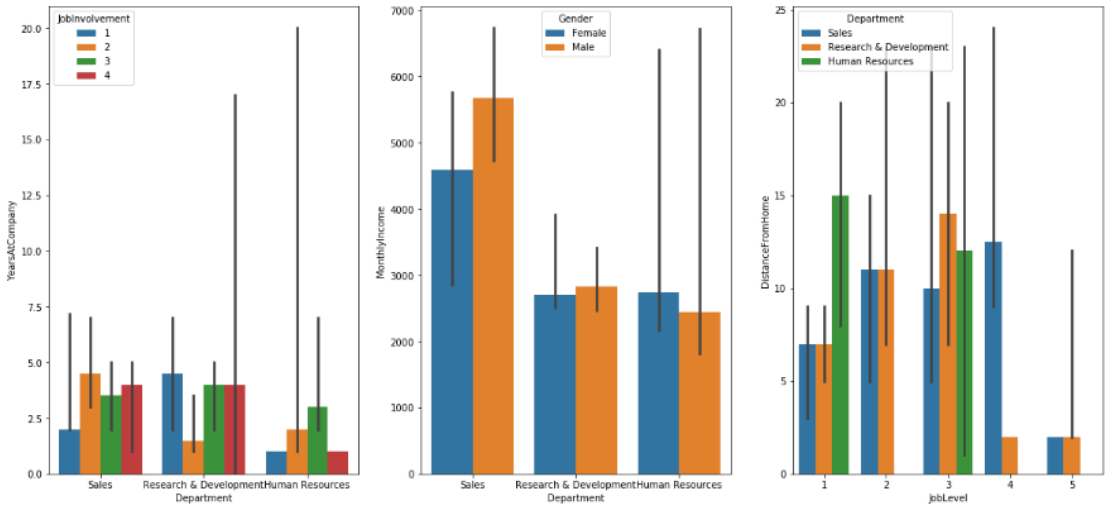

Figure 2. Various Factors for Employees in Attrition

According to the figure 2:

- Low job involvements from Sales and Research & Development department are more likely to leave.
- Males having around 5500 monthly incomes from Sales department are more likely to leave from females having 4500 monthly incomes in the same department.
- Job Level 1 in Human Resources having 15 miles distances from home are more likely to leave. Distance From Home in high job levels can be an issue for people from Sales Department.
- Average Monthly Rate for females is %4 higher than Males. We know that attrition in males is higher than females', so monthly rate can be an influential factor for that reason.

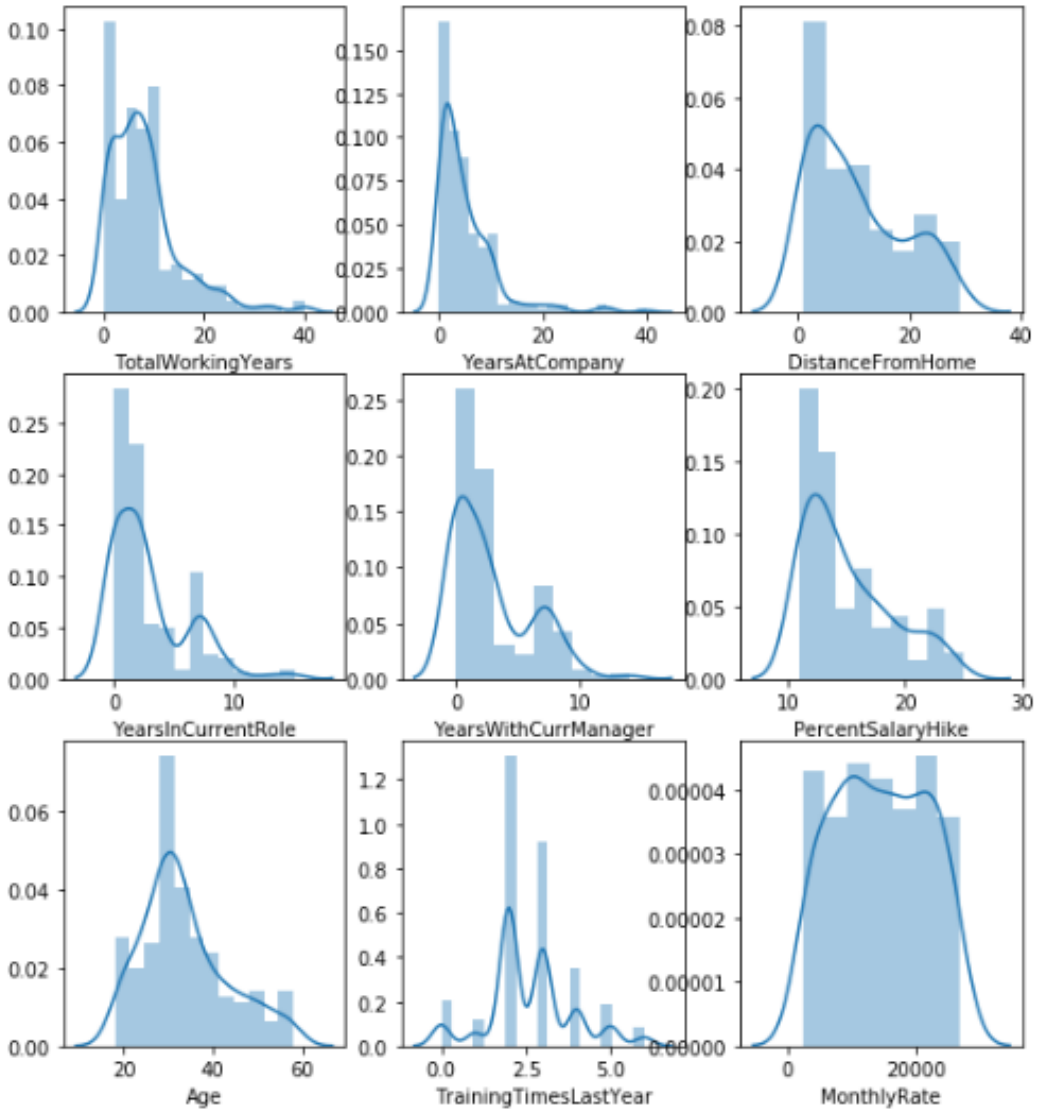

Figure 3. Trends of Employees in Attrition with Individual Factors

According to the figure 3:

- Number of employees in attrition decreased after 20 years of TotalWorkingYears and YearsAtCompany.
- Most of employees who left were had 20 miles or less distance from home to work.

- PercentSalaryHike mostly changes between 10 and 20 percent for employees who left.
- Number of employees who turned over decrease after age between 30 and 40.
- Most of employees turned over did 2 and 3 times of training last year.
- Employees in attrition are likely to be also from high monthly rate positions displayed on the last plot.

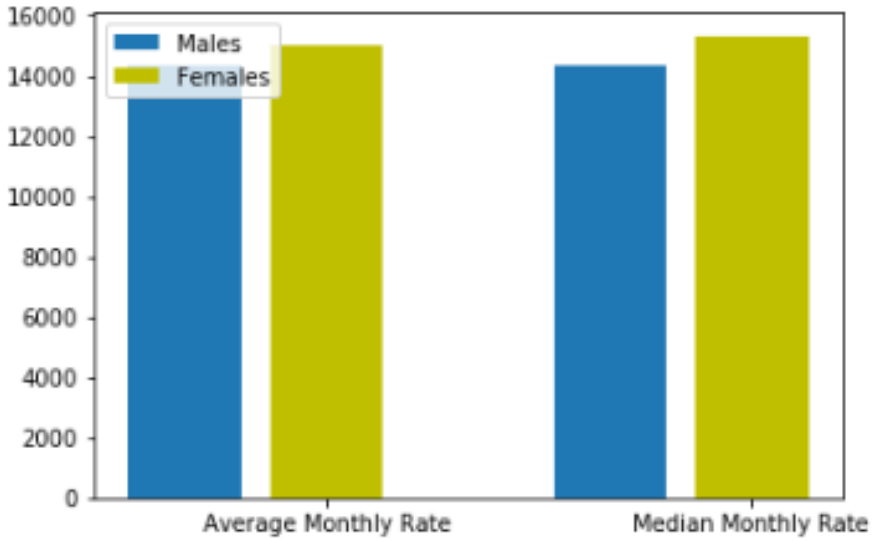

Figure 4. Change in 'MonthlyRate' by Gender

According to the figure4 displayed above and some calculations, average Monthly Rate for female 4% higher than Males. We know that attrition in males is higher than females, so monthly rate can be an influential factor for that reason.

Data Pre-Processing

Before running machine learning algorithms, there are some pre-processing steps to apply attrition data set. The purpose for those steps is to get more accurate and unbiased results for our predictions.

1. Label Encoding: There are some categorical and binary variables in this data set. We eliminated all label words and encoded with numbers. 'Attrition', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'OverTime' and 'BusinessTravel' are variables encoded with certain numbers. Also, in logistic regression models, encoding all of the independent variables as dummy variables allows easy interpretation and calculation of the odds ratios, and increases the stability and significance of the coefficients.

2. **Imbalanced Data in Binary Classes:** The main case required to solve during data wrangling is imbalanced binary variables. As our concern in this study is predicting employee attrition, we need to construct our statistical model around 'Attrition' variable as dependent variable. In this case, some data manipulation requires for this feature. 'Attrition' is a dichotomous variable with 'Yes' or 'No' responses, so there are particular statistical models are required to solve this classification problem. First, we encoded 'Yes' responses as '1' and 'No' responses as '0'.

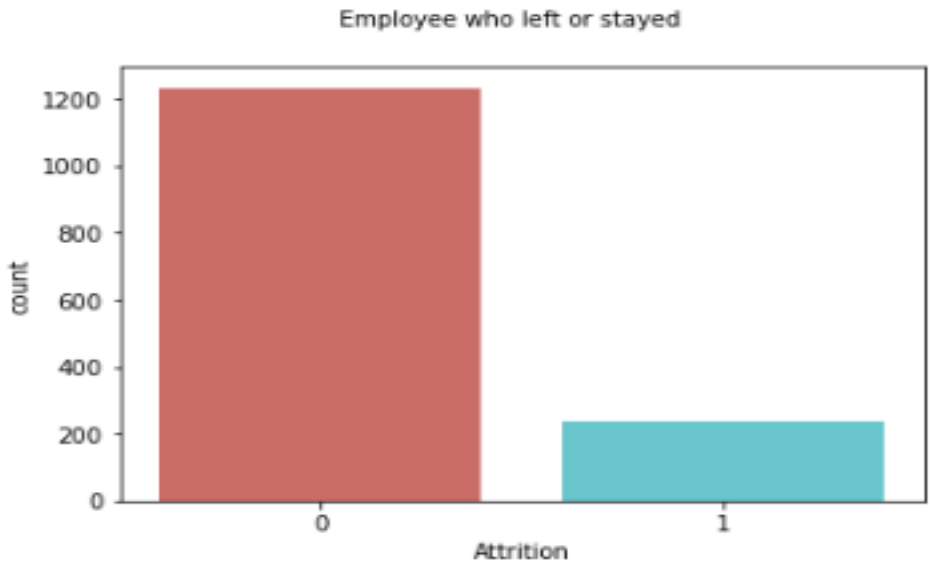

Figure 5. Number of Observations for 'Attrition'

As seen on figure 5, there are 1233 employees who stayed labeled as '0' and 233 employees who left from job labeled as '1'. Thus, the gap between two kinds of observation is really large and machine learning algorithms tend to predict majority of class. In this case, an up sample for minority class, '1' in our data, will be applied in order to reinforce its signal. Then, a new data frame consisting of up-sampled minority class will be created. As a result, number of observation for both classes is 1233.

Modeling

Logistic regression, random forest classification and support vector machines were applied for this classification problem. Before applying those machine learning algorithms, our data set was splitted as training and test set. The %33 data validated on test set. By using confusion matrix, we describe the performance of our model on test data.

Logistic Regression

After splitting our data sets as training and test sets, we get about 74% accuracy score on test set for the logistic model. K-fold resampling method was applied for testing the 10th fold and got 76% accuracy score.

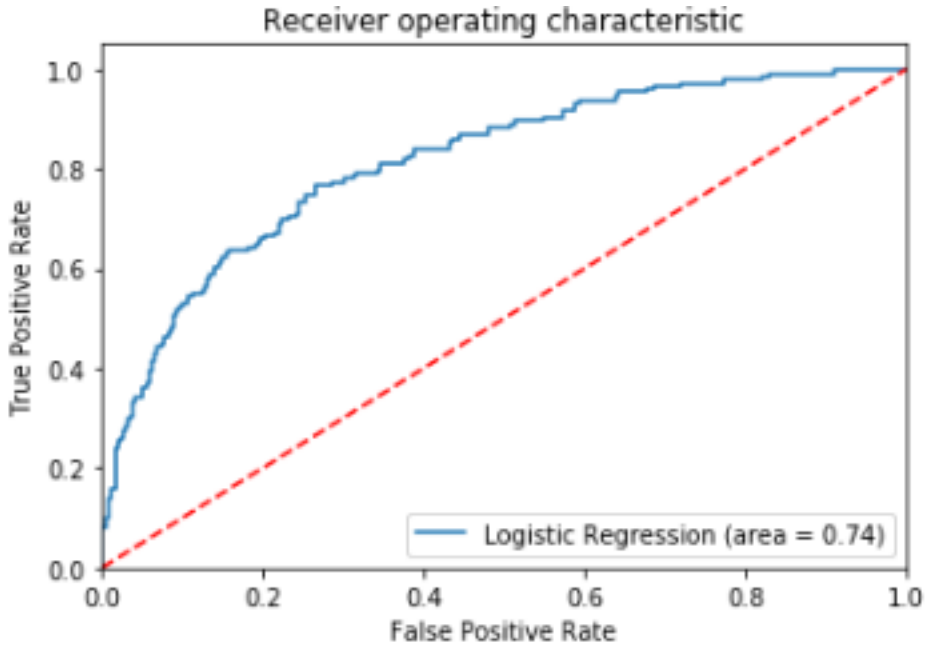

Figure 6. ROC Curve for Logistic Regression Model

The accuracy of the test depends on how well the test separates the group being tested into those with and without the attrition in question. The ROC curve compares the model true positive and false positive rates to the ones from a random assignation. If the model roc is above the baseline, then the model is better than random assignation, so our model is represented good and larger than 0.5 of AUC (area under curve) score as above the baseline.

Random Forest Classification

For the Random Forest Model, we used the same portion of data splitting for both training and test sets. The advantage of using random forest algorithms is providing feature importance of independent variables for the prediction of outcome variables displayed on figure 7. All variables were numerated based on their position on data set, represented on x-axis.

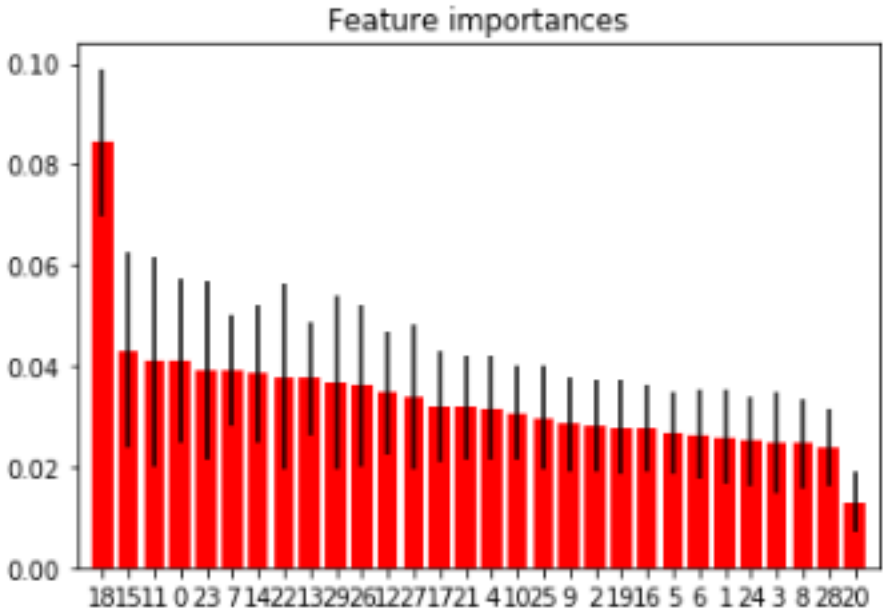

Figure 7. Feature Importance of variables

Support Vector Classifier (SVC)

SVC is the version of SVM for categorical response variables and after using the same ratio for splitting of training and test sets, we fit the model. Although the accuracy of the model is really high, the precision of the confusion matrix is really low compare to other applied models.

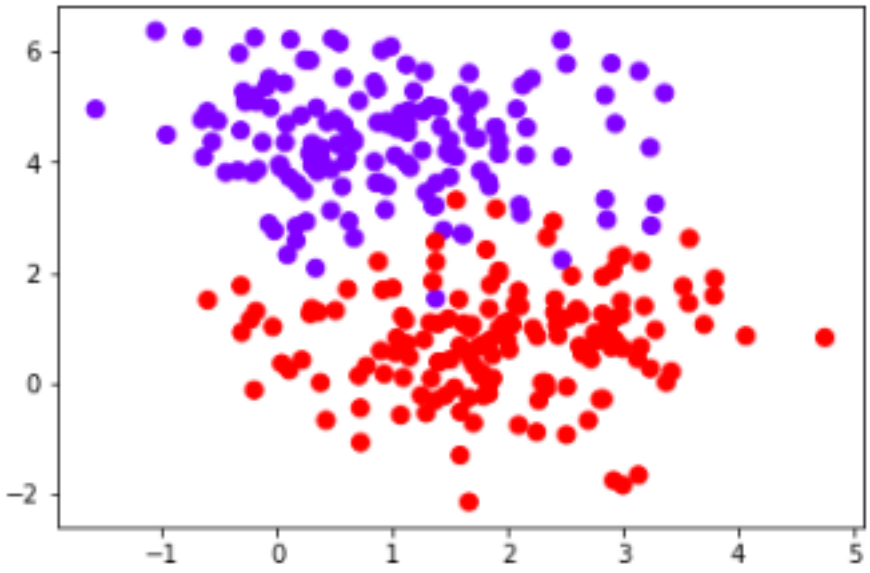

Figure 8. Scatter plot of outcome classification

Based on the scatter plot from Figure.8, there is no well separation between existence of attrition or no attrition. As a result, his plot supports the results of confusion matrix in determination of frequency of having attrition.

Using Model Recommendations

When we compare three classification models, Logistic Regression, Random Forest and Support Vector Classifier, the accuracy scores for all three models are between 97% and %99. However, this doesn't really tell us anything about where we're doing well. A useful technique for visualizing performance is the confusion matrix. This is simply a matrix whose diagonal values are true positive counts; while off-diagonal values are false positive and false negative counts for each class against the other. In this case, Support Vector Classifier makes the best prediction and random forest model is second best machine learning algorithm for this classification model.

Conclusion

Employee attrition is one of the biggest concerns of companies. Especially, losing talented employees causes more challenges for department of human resources. In this project, we purposed to find what kind of factors could initiate the existence of attrition for a potential employee with the determination of leaving.

After examining the problem, we explored the data and applied appropriate data manipulation steps to prepare our data for modeling in order to receive optimal predictions. Three machine-learning algorithms were applied for this classification problem. Based on model evaluation assumptions, Supported Vector Classifier performed the best results for predicting. According to the feature importance provided by Random Forest algorithms, 'NumCompaniesWorked', 'JobInvolvement' and 'MaritalStatus' are variables having the highest rankings for prediction. On the other hand, based on the output of logistic regression, those variables are significant variables to predict the target variable. Also, application of two-sample t-test confirms models results in some cases during the model pre- processing. More details can be seen on project coding output.