IBM HR Analytics Challenge

Anuprava Chatterjee

Executive Summary

Employee retention is a hottest issue in 2017 for business leaders. Overall economy has gradually recovered and unemployment rate continuously reduces. Employees are easier to find new jobs in the competitive marketplace these days, leading companies suffer by costly employee replacement. A company loss of employee attrition is way more than we can imagine. However, this problem will be massive lessened if those companies will perceive in advance who in the company are likely to resign and early implement a proper retention plan individually. This idea motivates us to create employee retention software tool in order to tackle this problem.

Main purpose of our product is to predict employee attrition in a company and further seek out specific reasons affecting employee attrition problem in that company. For our model prototype, we use IBM HR analytics employee attrition and performance dataset. We implement many techniques to tackle the problem such as logistic regression, random forest, LDA, QDA, KNN and Neural Network. Based on this dataset, the results show that logistic regression is outperformed with 89% overall accuracy.

Part 1: Business Part

Motivation

The importance of employee retention is overlooked by most companies but not for successful companies. A successful company realises that retaining valuable employees is essential for maintaining institutional knowledge, high morale, satisfied customers, and even sales growth. In general, a company would like to offer best in both tangible and intangible incentive to make their talents engaged to the company. However, any company cannot invest this offer to everyone. This offer usually happens when HR is notified that an employee has already decided to leave which is, most of the time, not in time to change the employee's mind; and eventually the company lose their valuable employees again and again.

Moreover, intrinsic cost of employee attrition is way more than a company realization. A study[1] by the Society for Human Resource Management stated that replacing one employee, a company needs to spend about six to nine months of that employee's salary. What that high cost comes from? One obvious cost is hiring process. This includes advertising, interviewing, screening, and hiring. A company also needs to spend for onboarding a new person, including training and management time which at least cost about 10% to 20% of an employee's salary and can be even more for more technical training. Lost productivity is another intrinsic cost since a new employee may take one to two years to reach the productivity of an existing person. New employees may also take longer time to be skillful which makes company loss of customer service and errors as well. Furthermore, the attrition of these employees can lead a butterfly effect on other employees and even further worsen company's productivity.

To make it more clear, assume that an employee who earns $80,000 a year is quitting, cost of the company will be $40,000 - $60,000 for replacement process. This is

[1]https://www.huffingtonpost.com/julie-kantor/high-turnover-costs-way-more-than-you-think_b_9197238.html

the cost of only one employee resigns. Please imagine if your company is losing more employees, like 10?, 50? or 100?. How much cost is for that?

This problem can be significantly reduced if those companies can exactly identify who else in their company are likely to quit in advanced. So that the company can handle everything in time and never lose their talents also their money ever again. This motivated us to develop a software tool that can tackle attrition problem. We aim to predict employees' attrition with specific identify who they are and also discover specific reasons that affect a company's attrition problem.

Product Idea

Our product is a software tool that can convey a company attrition situation. The product can predict how many potential employees in that company are likely to resign. We can further identify who are in those potential quitting list. Moreover, since the attrition issue in each company might be different, we can make a specific recommendation for each company that what exactly the characteristics most affect their employees' attrition and how. Therefore, HR team will be able to strongly develop retention plan and also be able to accurately focus a potential quitting employee.

Product Price

Price structure will be separated into 2 part. One is based price which is fixed at $10,000. Second part varies depending on the size of the company.

Based price = $10,000
Varied price = $1,600 * #employees

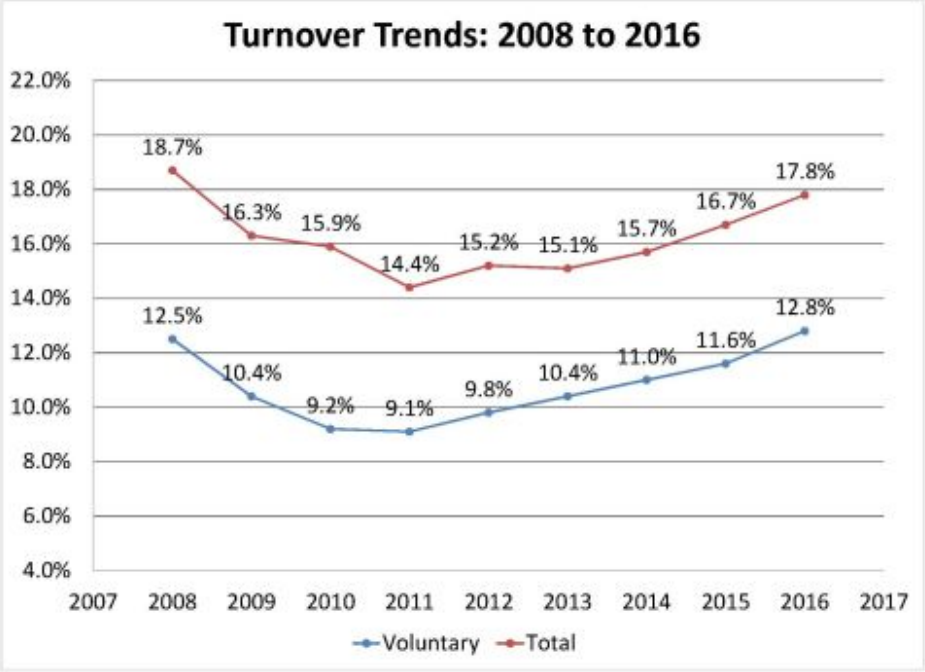
For example, price for a company with 100 employees = (10,000+160,000) = $170,000.

Market Opportunity and Competitiveness

In 2017, employee retention is reported as the biggest priority and concern for business leaders. A study[2] by Future Workplace and Kronos stated that 87% of employers said that improving retention is a critical priority for their organization. In the U.S., according to Salary Budget Survey[3] in 2016-2017 by WorldatWork, an association of total rewards professionals, it is reported that employee salaries are projected to grow by 3% in 2017. Also, the U.S. unemployment rate has gradually reduced to 4.1% (October, 2017[4]). All these facts reflects that overall economy continuously improves and employees have more opportunities for their jobs. This is the incentive for employee turnover to highly increase. A survey by CompData Surveys[5] shows that total turnover in all industries in 2016 is high at 17.8% and turnover trends has continuously gone up since 2011.

[2] http://fortune.com/2016/12/28/employers-2017-employee-retention-unemployment/
[3] https://www.worldatwork.org/adim/pub/2016-2017-top-level-results.pdf
[4] Bureau of Labor Statistics, Eurostat
[5] http://www.compensationforce.com/2017/04/2016-turnover-rates-by-industry.html

2016 Total Turnover	
All Industries	17.8%
Banking & Finance	18.1%
Healthcare	19.9%
Hospitality	28.6%
Insurance	12.2%
Manufacturing & Distribution	16.0%
Not-For-Profit	15.7%
Services	16.8%
Utilities	8.8%



Source: http://www.compensationforce.com/2017/04/2016-turnover-rates-by-industry.html

This situation shows that employee retention is blowing the market up leading to highly opportunity for our product in the market. In addition, since we are the first mover introducing this product, we are highly competitive in the market.

Target Customers

Our target customers are literally all companies. However, in first stage, we aim companies suffering from employees' attrition. According to total turnover report in market opportunity part above, we target companies in hospitality, healthcare and banking and finance as our first priority.

Minimum Viable Product

A minimum viable product (MVP) is a development technique in which a new product or website is developed with sufficient features to satisfy early adopters. The final, complete set of features is only designed and developed after considering feedback from the product's initial users. So our MVP would have a similar body like:

1. Early adopters - IBM HR
2. Provide IBM with two models as interactive software/mobile application.
    a. Classify attrition amongst employee with inferential factor analysis.
3. Receive Feedback from the HR and incorporate changes quickly.

Next steps after success of MVP we can take our model and test them on the data of other MNCs like Microsoft, Apple, Google and Amazon so that we can increase the robustness of our product.

Initial Investment

To set up our company, we need $465,500. The details are shown below.

Legal	1,000
Stationary	2,500
Renting	2,000
Equipment	10,000
Cash Balance on Starting date	450,000
Intial Investment	**465,500**

Part 2: Mathematical Part

Problem formulation in Math

According to product idea in business part, we can state out problem as following.

- Target variable: employee's attrition status (Yes, No)
- Predictors: many variables can be included. However, those variables should contain in these four categories as following.
    1. Personal characteristic such as age, education level, gender, marital status, distance from home, total working experience in years and the number of companies has been working for.
    2. Job characteristic such as salary,job level, department, working year experience, the number of years with current manager, the number of years in current role, the number of years at company, frequency of business travel and training time last year.
    3. Job benefit such as buying company stock option and %salary hike.
    4. Working condition such as environment satisfaction, job satisfaction, relationship satisfaction, work life balance and performance rating.

Our aim is to find a function using those predictors to predict target variable (attrition). Since attrition is categorical variable, we will apply classification methods as following.

Method 1: Logistic Regression
Method 2: Random Forest
Method 3: LDA
Method 4: QDA
Method 5: KNN
Method 6: Neural Network

Math Solution

Logistic Regression:

Logistic regression is a linear model measuring the relationship between a categorical dependent variable and one or more independent variables. It applies the logit function $g(t) = log(\frac{p}{1-p})$ to transform the probability as the following:

$$log\ \frac{p(x}{1-p(x)} = \beta_0 + \beta^T X \text{ where } p(x) = \frac{exp(\beta_0+\beta^T X)}{1+exp(\beta_0+\beta^T X)} = \frac{1}{1+exp(-(\beta_0+\beta^T X))}$$

Traditionally, the result from the logistic regression model can be interpreted with the odds ratios of the independent variables. The odds ratio is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. Given a 2-by-2 contingency table below, the odds ratios of a c can be computed as the following: the odds of a man drinking is p_{11}/p_{10}, the odds of a woman drinking is p_{01}/p_{00}, and therefore, the odds ratio is $\frac{p_{11}p_{00}}{p_{10}p_{01}}$ meaning that the odds of men drinking are $\frac{p_{11}p_{00}}{p_{10}p_{01}}$ times of the odds of women drinking.

X: Gender (1:Male) Y: Drinking (1:Yes)	Y = 1	Y = 0
X = 1	p_{11}	p_{10}
X = 0	p_{01}	p_{00}

Random Forest:

A random forest constructs multiple decision trees then takes the majority vote to construct a classification tree. To choose a variable that best splits the data, the decision tree use different metrics to measure the impurity: how well the data are separated or the homogeneity of the target variable within each subgroup of the split data. Two widely used measures are the following:

- Gini measure: Gini(D) = $-\sum_{i=1}^{k} p_i log p_i$
- Entropy measure: H(D) = $1 - \sum_{i=1}^{k} p_i^2$

where D is the data entries, D_i is the data entries classified as i, and p_i is the ratio of instances classified as i, for i = 1, 2, …, k.

LDA & QDA :

Both LDA and QDA can be derived from simple probabilistic models which model the class conditional distribution of the data $P(X|y=k)$ for each class k. Predictions can then be obtained by using Bayes' rule:

$$P(y=k|X) = \frac{P(X|y=k)P(y=k)}{P(X)} = \frac{P(X|y=k)P(y=k)}{\sum_l P(X|y=l) \cdot P(y=l)}$$

and we select the class k which maximizes this conditional probability.

More specifically, for linear and quadratic discriminant analysis, $P(X|y)$ is modelled as a multivariate Gaussian distribution with density:

$$p(X|y=k) = \frac{1}{(2\pi)^n |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(X-\mu_k)^t \Sigma_k^{-1}(X-\mu_k)\right)$$

To use this model as a classifier, we just need to estimate from the training data the class priors $P(y=k)$ (by the proportion of instances of class k), the class means μ_k (by the empirical sample class means) and the covariance matrices (either by the empirical sample class covariance matrices, or by a regularized estimator: see the section on shrinkage below).

In the case of LDA, the Gaussians for each class are assumed to share the same covariance matrix: $\Sigma_k = \Sigma$ for all k. This leads to linear decision surfaces between, as can be seen by comparing the log-probability ratios $\log[P(y=k|X)/P(y=l|X)]$:

$$\log\left(\frac{P(y=k|X)}{P(y=l|X)}\right) = 0 \Leftrightarrow (\mu_k - \mu_l)\Sigma^{-1}X = \frac{1}{2}(\mu_k^t \Sigma^{-1}\mu_k - \mu_l^t \Sigma^{-1}\mu_l)$$

In the case of QDA, there are no assumptions on the covariance matrices Σ_k of the Gaussians, leading to quadratic decision surfaces.

KNN:

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

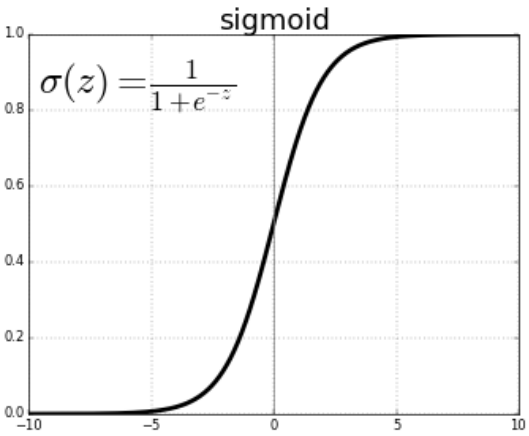
Neural Network:

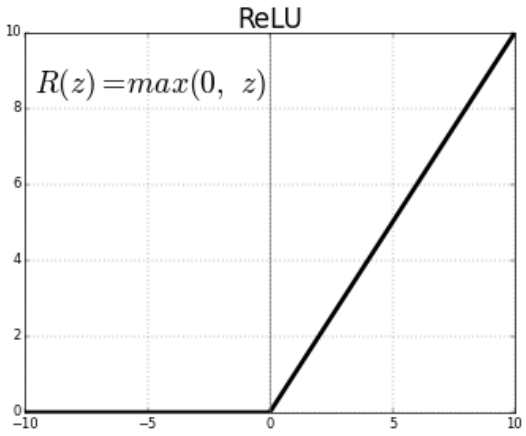
A **sigmoid function** is a mathematical function having an "S" shaped curve (**sigmoid curve**). Often, *sigmoid function* refers to the special case of the logistic function shown in the first figure and defined by the formula.

$$S(t) = \frac{1}{1 + e^{-t}}.$$

The **ReLU** is the most used activation function in the world right now.Since, it is used in almost all the convolutional neural networks or deep learning. In the context of artificial neural networks, the **rectifier** is an activation function defined as the positive part of its argument:

f(x) = max (0, x), where *x* is the input to a neuron.





Implementation

These mathematical concepts will be implemented in python applying on a prototype data of IBM HR analytics employee attrition and performance.

Part 3: Hacking Part

Data Description

The dataset of IBM HR analytics employee attrition and performance is obtained from Kaggle. It contains 1,470 records. Each record comes with 35 features. One of these 35 is the target attribute indicating an employee's attrition status, yes or no. The rest 34 attributes are 26 numerical and 8 categorical variables.

Data Preprocessing

We drop the following features:

- EmployeeCount, StandardHours, and Over18 because of invariant
- EducationField, JobRole because of redundancy

Data Analysis and Results

To tackle the problem, various algorithms are implemented based on model description here.

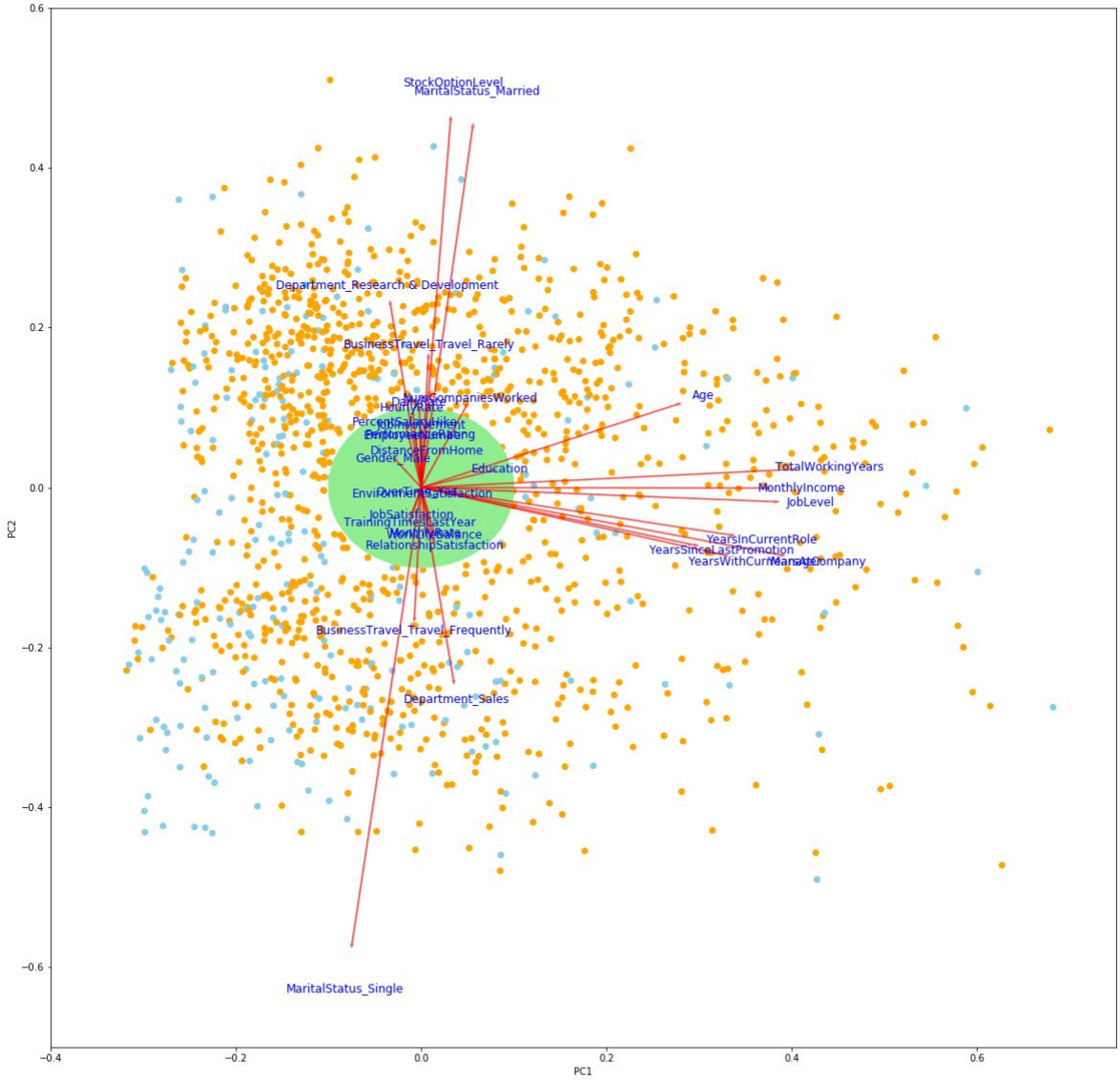
Model Objective: To predict employees' attrition

Target Variable: Attrition (Yes, No)

Predictors:

- Age, Gender, Marital Status, Education
- Daily earning rate (DailyRate)
- Distance from home (DistanceFromHome)
- Employee Number
- Environment Satisfaction
- Hourly earning rate (HourlyRate)
- Job Involvement
- Job Level
- Job Satisfaction
- Monthly Income
- Monthly earning rate (MonthlyRate)
- The number of companies has worked for (NumCompaniesWorked)
- Percent Salary Hike
- Performance Rating
- Relationship Satisfaction
- Stock Option Level
- Total Working Years
- Training Times Last Year
- Work Life Balance
- Years At Company
- Years In Current Role
- Years Since Last Promotion
- Years With Current Manager
- Business Travel
- Department
- OverTime

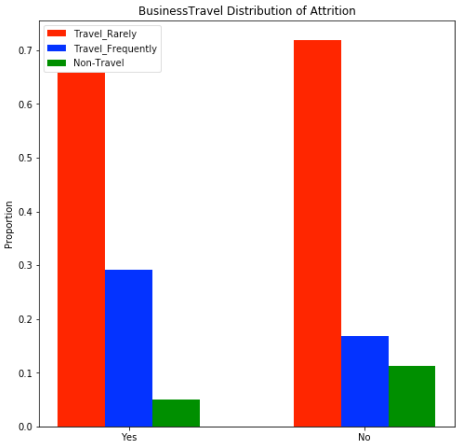
First, we try to understand the relationship among the features. We apply PCA and create biplots. The first biplot is constructed using the first two principal components of the Attrition model. Even though result does not separate the attrition in a 2-dimensional plot, we can observe some features from the biplot shows as the following:

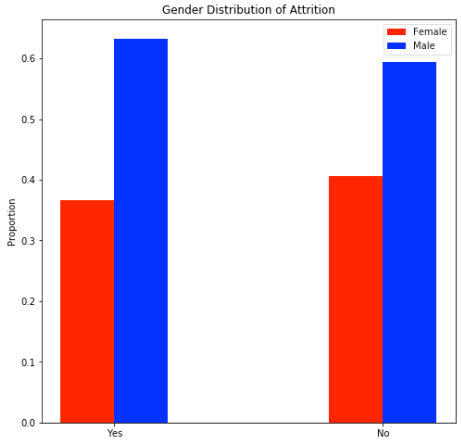


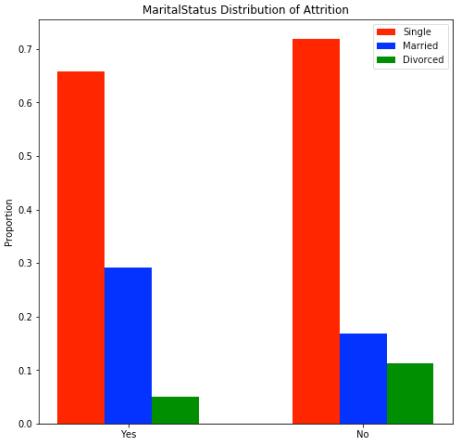
- Age, TotalWorkingYears, MonthlyIncome, JobLevel, YearInCurrentRole, YearSinceLastPromotion, YearWithCurrentManager, and YearAtCompany are associated.
- Employees who are single and in Sales Department travel frequently.
- Finally, employees who are married and in Research & Development Department rarely travel and have high StockOptionLevel.

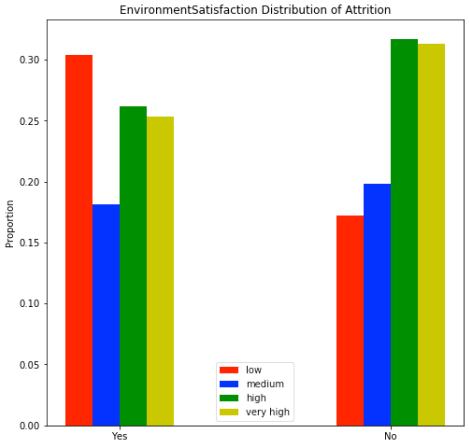
Logistic Regression:

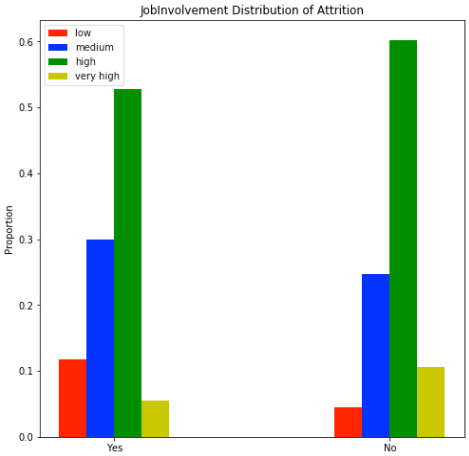
Given that the target is Attrition, the result shows that the following features are significant: DistanceFromHome, EnvironmentSatisfaction, JobInvolvement, OverTime, JobSatisfaction, NumCompaniesWorked, PerformanceRating, TotalWorkingYears, YearsWithCurrentManager, RelationshipSatisfaction, TrainingTimesLastYear, Gender, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, BusinessTravel and MaritalStatus. We further investigate these features to find out how they affect the Attrition.

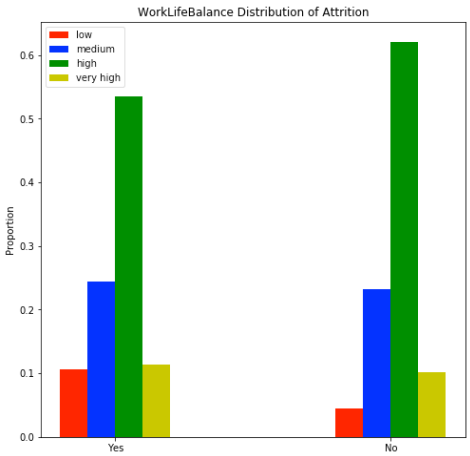




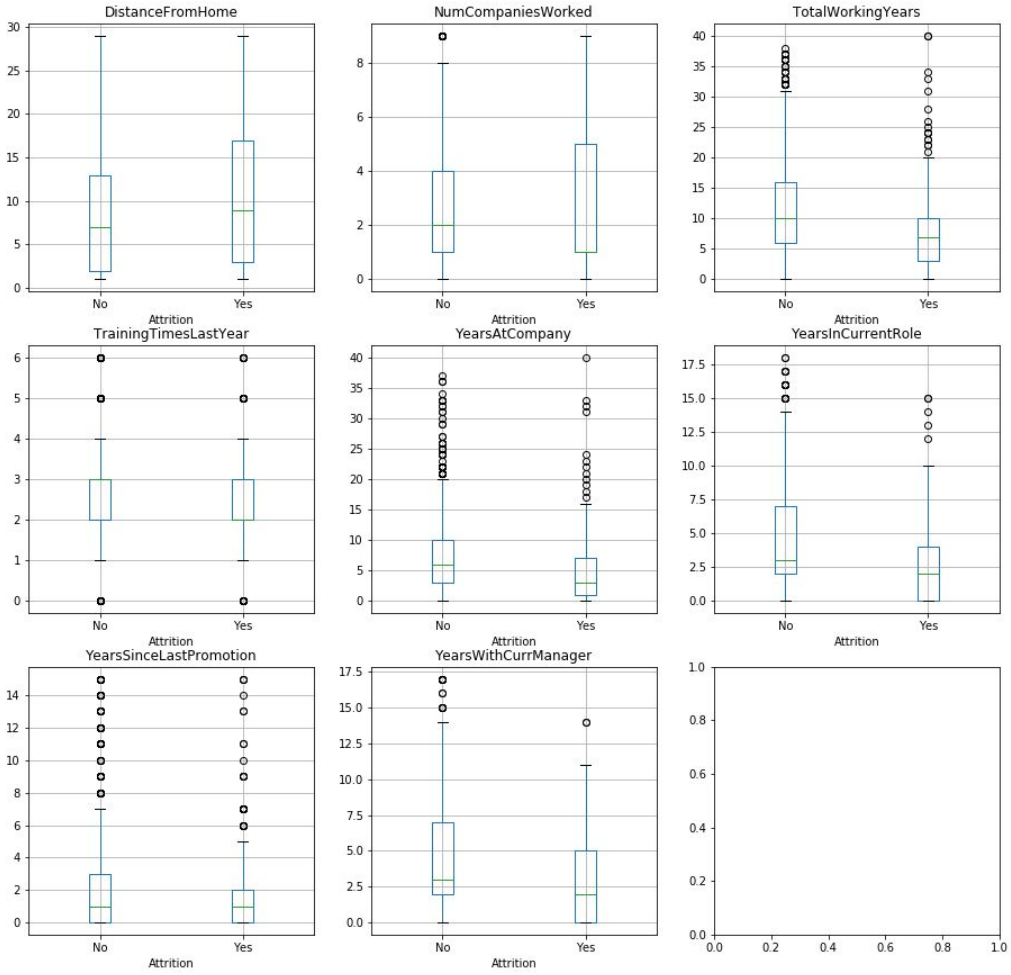








Out of employees who left their job, over 60% rarely travel for business, over 60% most of them are male, over 60% of them are single, over 60% work overtime, about half of them do not satisfy with the job environment. over half of them highly involve in the job, and interestingly more than half of them have high WorkLifeBalance. In addition, employees who left their job have higher DistanceFromHome, but lower NumCompaniesWorked, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, and YearsWithCurrManager.



The result of logistic regression model is shown below.

Variable	coef	std err	z	P>|z|
Age	-0.0230	0.013	-1.816	0.069
DailyRate	-0.0002	0.000	-1.072	0.284
DistanceFromHome	0.0460	0.010	4.434	0.000
Education	0.0233	0.084	0.278	0.781

EmployeeNumber	-9.78e-05	0.000	-0.681	0.496
EnvironmentSatisfaction	-0.3741	0.079	-4.746	0.000
HourlyRate	0.0028	0.004	0.672	0.501
JobInvolvement	-0.4878	0.118	-4.125	0.000
JobLevel	-0.3112	0.281	-1.106	0.269
JobSatisfaction	-0.3851	0.077	-4.973	0.000
MonthlyIncome	-4.268e-05	6.71e-05	-0.636	0.525
MonthlyRate	7.657e-06	1.2e-05	0.640	0.522
NumCompaniesWorked	0.1873	0.037	5.047	0.000
PercentSalaryHike	-0.0577	0.037	-1.569	0.117
PerformanceRating	0.7844	0.311	2.518	0.012
RelationshipSatisfaction	-0.2263	0.079	-2.852	0.004
StockOptionLevel	-0.1339	0.147	-0.913	0.361
TotalWorkingYears	-0.0605	0.028	-2.163	0.031
TrainingTimesLastYear	-0.1450	0.070	-2.081	0.037
WorkLifeBalance	-0.2421	0.115	-2.096	0.036
YearsAtCompany	0.0944	0.038	2.498	0.012
YearsInCurrentRole	-0.1474	0.044	-3.354	0.001
YearsSinceLastPromotion	0.1748	0.041	4.235	0.000
YearsWithCurrManager	-0.1348	0.046	-2.931	0.003
BusinessTravel_Travel_Frequently	2.1857	0.423	5.162	0.000
BusinessTravel_Travel_Rarely	1.3485	0.392	3.438	0.001
Department_Research & Development	-0.4145	0.401	-1.033	0.301
Department_Sales	0.4123	0.417	0.988	0.323

Gender_Male	0.4925	0.178	2.761	0.006
MaritalStatus_Married	0.3886	0.261	1.490	0.136
MaritalStatus_Single	1.2659	0.331	3.830	0.000
OverTime_Yes	1.8175	0.183	9.918	0.000

Odd Ratios interpretation:

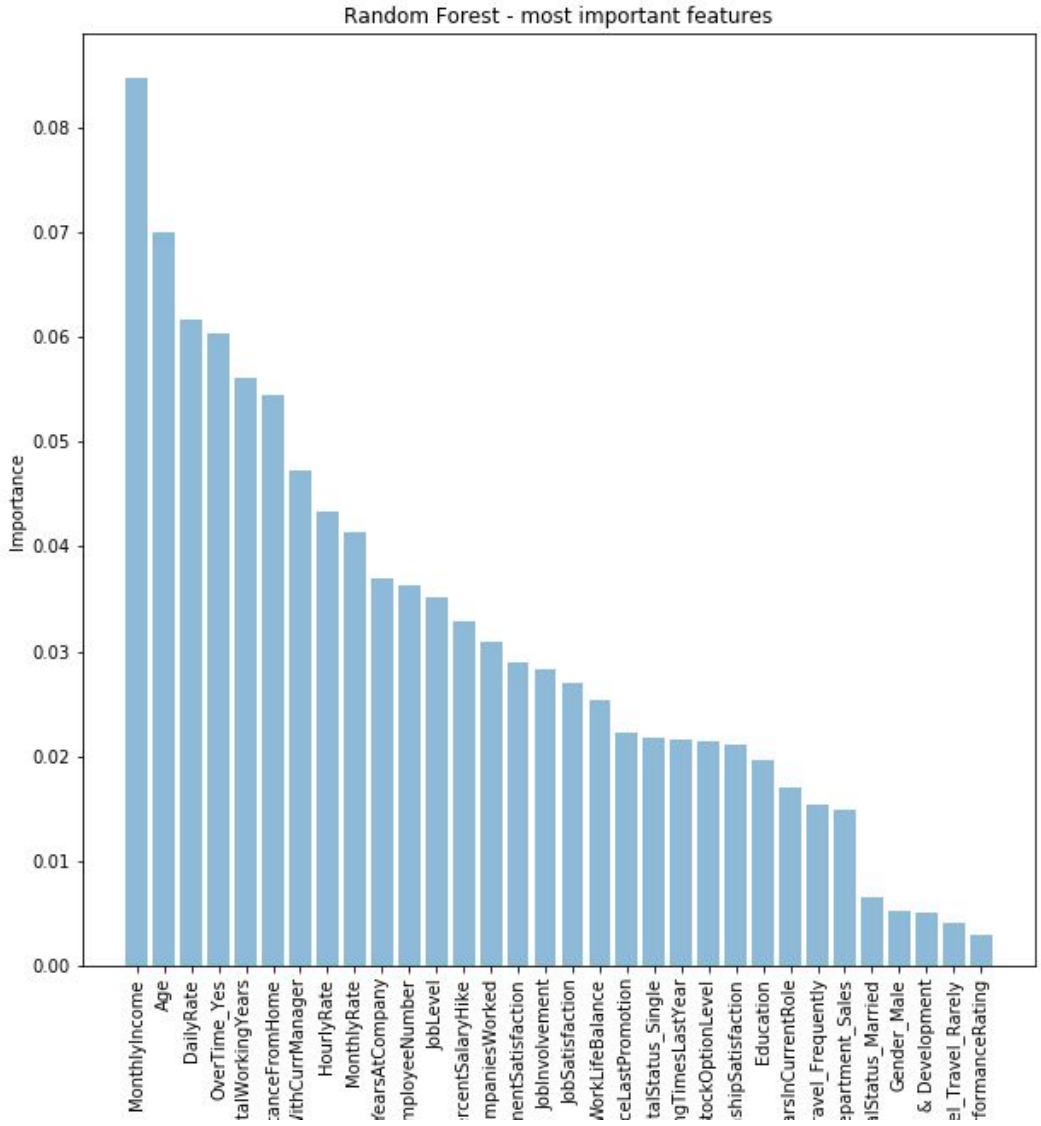
The result shows that

- For one unit increase in **DistanceFromHome**, the odds of an employee leaving the job is $e^{0.046} = 1.047$ compared to the odds of an employee retaining the job. Or for one unit increase in **DistanceFromHome**, the odds of an employee quitting increase about 5%.
- For one unit increase in **EnvironmentSatisfaction**, or for one unit increase in **JobSatisfaction**, the odds of an employee staying increase by about 43%.
- For one unit increase in **JobInvolvement**, the odds of an employee staying increase by 67%
- For one unit increase in **NumCompaniesWorked**, the odds of an employee quitting increase by 20%.
- For one unit increase in **PerformanceRating**, the odds of an employee quitting is more than twice of the odds of an employee retaining.
- For one unit increase in **RelationshipSatisfaction**, the odds of an employee staying increase by about 25%.
- For one unit increase in **TotalWorkingYears**, the odds of an employee staying increase by about 6%.
- For one unit increase in **TrainingTimesLastYear**, the odds of an employee staying increase by 18%.
- For one unit increase in **WorkLifeBalance**, the odds of an employee staying increase by 28%.
- For one unit increase in **YearsAtCompany**, the odds of an employee quitting increase by 10%.
- For one unit increase in **YearsInCurrentRole**, the odds of an employee staying increase by 16%.
- For one unit increase in **YearsSinceLastPromotion**, the odds of an employee quitting increase by 20%.
- For one unit increase in **YearsWithCurrManager**, the odds of an employee staying increase by 15%.
- For one unit increase in **BusinessTravel_Travel_Frequently**, the odds of quitting for an employee who frequently travel for the business is almost 9 times larger than the odds of an employee who never travel for the business. In addition, for one unit increase in **BusinessTravel_Travel_Rarely**, the odds of quitting for an employee who rarely travel for the business is almost 4 times

larger than the odds of an employee who never travel for the business. Therefore, an employee who travel for the business is likely to quit the job.

Random Forest:

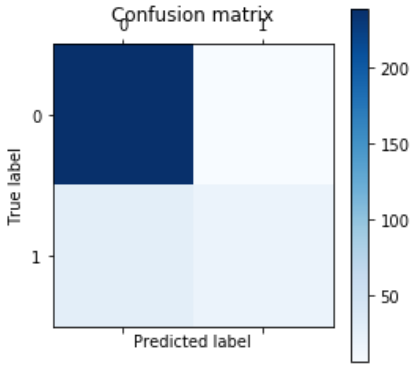
Given that the cutoff is at 0.04, random forest shows that MonthlyIncome, Age, DailyRate, MonthlyRate, TotalWorkingYears, DistanceFromHome, HourlyRate, YearsWithCurrentManager, and OverTime importantly affect the Attrition.



LDA:

Before performing LDA and QDA we check if the variables are collinear by applying a VIF test. If VIF > 5 then the variable is collinear, and we could eliminate the variables whose VIF is greater than 5.

	VIF Factor	features
0	2.118022	Age
1	1.024380	DailyRate
2	1.022946	DistanceFromHome
3	1.065968	Education
4	0.000000	EmployeeCount
5	1.030529	EmployeeNumber
6	1.023357	EnvironmentSatisfaction
7	1.022335	HourlyRate
8	1.028209	JobInvolvement
9	11.560367	JobLevel
10	1.026852	JobSatisfaction
11	10.812384	MonthlyIncome
12	1.020697	MonthlyRate
13	1.272930	NumCompaniesWorked
14	2.533932	PercentSalaryHike
15	2.537125	PerformanceRating
16	1.031773	RelationshipSatisfaction

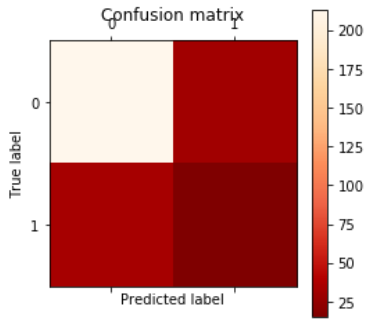


	precision	recall	f1-score	support
0	0.89	0.98	0.93	245
1	0.77	0.41	0.53	49
avg / total	0.87	0.88	0.87	294

Accuracy of the LDA Model is 0.8809

QDA:

We can see that LDA is performing better than QDA, since we have a few training observations, we can observe that LDA is performing better than QDA, if we had a large training data set, then QDA would have performed much better.

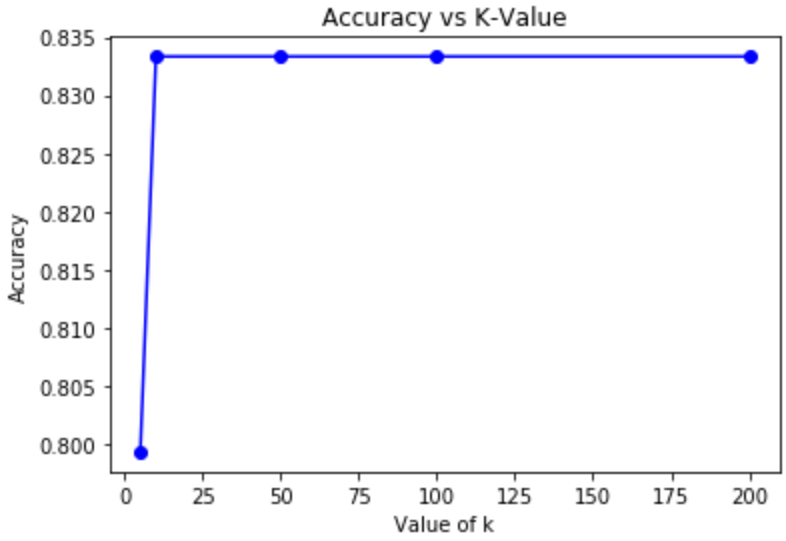


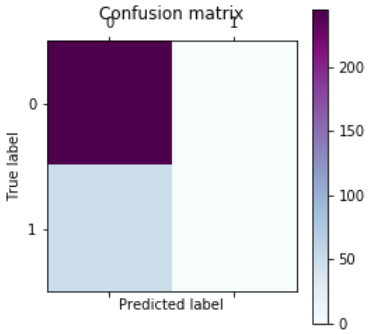
	precision	recall	f1-score	support
0	0.86	0.87	0.87	245
1	0.32	0.31	0.31	49
avg / total	0.77	0.78	0.77	294

Accuracy of the QDA Model 0.806

KNN:

We then performed KNN by plugging in different values of k, and from the below graph you can observe that from k=10, the accuracy remains constant, so we use k=10 and apply KNN. Below we can also see the Confusion matrix and precision recall values for both the classes. Accuracy of KNN model is 0.833.

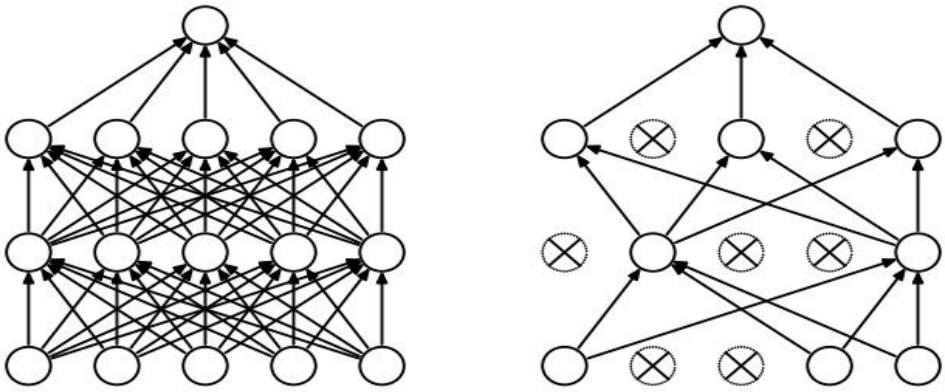




	precision	recall	f1-score	support
0	0.83	1.00	0.91	245
1	0.00	0.00	0.00	49
avg / total	0.69	0.83	0.76	294

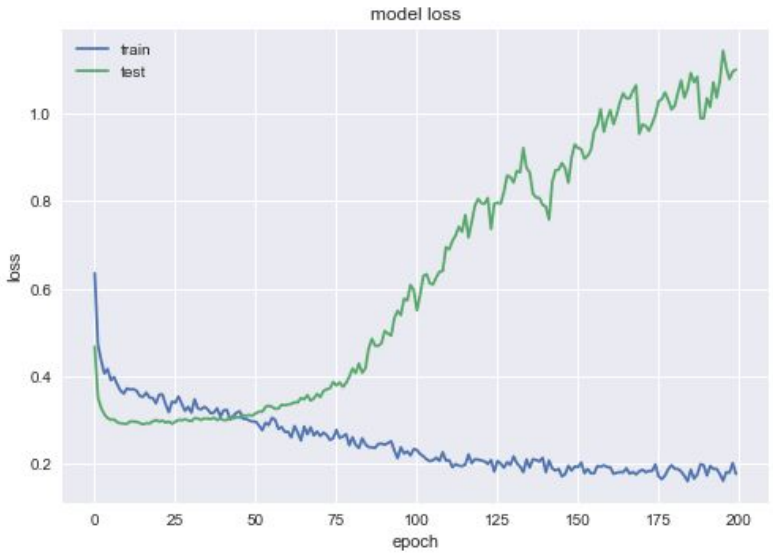
Neural Network:

We build our Neural network with 3 dense layers. In Our first setting we overfitted the data and classification was biased towards class 0. Then we added dropouts in our neural network which are explained below. Using dropouts we were able to reduce the overfitting of the data and now both the classes were well classified.



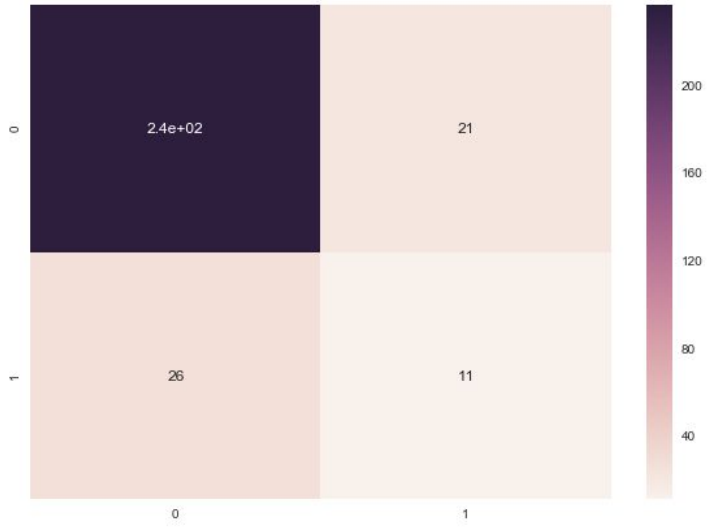
(a) Standard Neural Net (b) After applying dropout.

Below is the train(blue) and loss and the test(green) loss over 200 epochs. We can see that after 50 epochs the test loss increased but the recall for class 1 actually improving that was our target class.



Below is the train(blue) and accuracy and the test(green) accuracy over 200 epochs. We can see that after 50 epochs the test accuracy decreased but the recall for class 1 actually improving that was our target class.

Below is the confusion matrix we can see that for class 1 (“yes for attrition ”) we are able to correctly predict 11 people who will be leaving company.



Model Summary:

Attrition Prediction:

Class 0 : No Attrition

Class 1 : Yes Attrition

	Attrition Prediction				
Technique	Accuracy	Precision		Recall	
		Average	Class1	Average	Class1
Logistic Regression	0.89	0.89	0.84	0.89	0.43
Random Forest	0.84	0.81	0.58	0.84	0.22
LDA	0.88	0.87	0.77	0.88	0.41
QDA	0.77	0.87	0.77	0.88	0.41
KNN (k = 10)	0.79	0.71	0.00	0.80	0.00
Neural Network	0.85	0.85	0.40	0.85	0.43

Summary

Employee attrition has been a major concern across industries. Our product is offer to predict potential quitting employees for any company to help a company retention plan or early replacement process. To create a prototype model, we used the dataset of IBM. We applied many statistical methods and attempted to uncover the nonlinearity. One of the problems were unbalanced data set as target labels were biased towards “No” attrition by 84%. Logistic regression and QDA provide highest recall for class “yes”.

Logistic regression model suggests that IBM should increase the quality of working environment and also increase the job involvement of employees to retain them. IBM should recruit people who personally like to work in the company and like the job they are working. This is intuitive. Employees who are happy with the job and the company will work more efficiently and productively; thus, the company will gain more profit. Since we observe that employees with higher performance rating tend to quit more, we suggest that IBM should value their work and treat them more special to retain them in the company. In addition, if IBM wants long-term employees, they should put the distance from home into consideration. Lastly, the result shows that employees who change their job more often and employees who have less experience working in the company are more likely to quit their job. Therefore, IBM should appreciate employees who have worked in the company for a long time and loyal to the company.