

**Analytics Programming in Python**

**Project Title:**

*Prediction of Employee Attrition Probabilities in a Company*

**Team: Data Pirates**

*Abhishek Singla*

*Sriram Ranganathan*

*Subramanya Rao*

*Anshul Sarvate*

**School of Graduate Professional Studies**

MPS in Data Analytics

DAAN 862 – Analytics Programming in Python

(Fall, 2019)

**Abstract:**

Employee attrition is one of the major problems for a company’s growth. In this analysis we built two classification models, using two different approaches, to predict if an employee is likely to quit, which could greatly increase the HR’s ability to intervene on time and remedy the situation to prevent attrition. While these models can be routinely run to identify employees who are most likely to quit, the key driver of success would be the human element of reaching out to the employee, understanding the current situation of the employee and taking action to remedy controllable factors that can prevent attrition of the employee.

**Introduction**

Attrition is a gradual voluntary reduction of employees (through resignation or retirement) who are not then replaced. This means that attrition decreases the size of the workforce. High attrition significantly increases costs to the company. When attrition is high, employers lose a lot of money invested into recruitment, selection, onboarding, employee training etc. In addition, there is also a cost associated with having positions opened resulting in lower productivity related to attrition.

There are many causes of attrition, but here are a few most common examples of why attrition happens:

1. Unfair pay

2. Inability to grow and develop careers

3. Lack of work-life balance

4. Lack of employee recognition and awards

5. Poor management

6. Poor work conditions

7. Lack of benefits

8. Other reasons for attrition.

Employee attrition is one of the major problems for a company’s growth. We developed two different classification models, using two different approaches, to predict employee attrition using IBM HR Analytics dataset. This data set presents an employee survey from IBM, indicating if there is attrition or not. The data set contains approximately 1500 entries. Given the limited size of the data set, the model should only be expected to provide modest improvement in identifications of attrition vs a random allocation of probability of attrition.

**Dataset Characteristics**

We conducted our analysis on “IBM HR Analytics Employee Attrition & Performance” dataset which is available at “<https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>”. It is a fictional dataset created by IBM data scientists. IBM has gathered information on employee satisfaction, income, seniority and some demographics. It includes the data of 1470 employees.

Below table describes some of the columns definition:

|  |  |
| --- | --- |
| **Name** | **Description** |
| Attrition | Employee leaving the company (Yes, No) |
| BusinessTravel | (1=No Travel, 2=Travel Frequently, 3=Travel Rarely) |
| Department | (1=HR, 2=R&D, 3=Sales) |
| Education | 1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor' |
| EducationField | (1=HR, 2=Life Sciences, 3=Marketing, 4=Medical Sciences, 5=Others, 6= Technical) |
| EnvironmentSatisfaction | Satisfaction with the environment - 1 'Low' 2 'Medium' 3 'High' 4 'Very High' |
| Gender | (1=Female, 2=Male) |
| JobInvolvement | Job Involvement - 1 'Low' 2 'Medium' 3 'High' 4 'Very High' |
| JobLevel | Level of job – 1 to 5 |
| JobRole | (1=HC Rep, 2=HR, 3=Lab Technician, 4=Manager, 5= Managing Director, 6= Research Director, 7= Research Scientist, 8=Sales Executive, 9= Sales Representative) |
| JobSatisfaction | Satisfaction with the job - 1 'Low' 2 'Medium' 3 'High' 4 'Very High' |
| MaritalStatus | (1=Divorced, 2=Married, 3=Single) |
| PerformanceRating | Performance rating - 1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding' |
| RelationshipSatisfaction | Relations satisfaction - 1 'Low' 2 'Medium' 3 'High' 4 'Very High' |
| StockOptionLevel | Stock owned by employees - 0 'No' 1 'Low' 2 'Medium' 3 'High' |
| WorkLifeBalance | Time spent between work & outside - 1 'Bad' 2 'Good' 3 'Better' 4 'Best' |

**Data Collection**

The data collection process was easy, the dataset is available as a comma separated file (HR-Employee-Attrition.csv), so we downloaded the file and copied it into a Pandas data-frame using “read\_csv” function.

**Data Preparation**

After copying the dataset into a data-frame we explored the data and prepared it for modeling.

**Descriptive Statistics**

The descriptive analysis of the dataset showed that the shape of dataset is 1470 rows and 35 columns. It does not contain missing and duplicate values. We further checked data balancing based on Attrition column and found that IBM's attrition rate is (237/1470) = 16.12%, this means that, we have imbalanced dataset. Further we separated numerical and categorical variables into two different datasets and identified unique values in each column, based on those unique values we found that the following columns: EmployeeCount, StandardHours and Over18 had redundant data i.e. only 1 unique value, so we can drop these columns.

**Feature Engineering**

We performed feature engineering by sub-setting the dataset, labelling some categorical variables and scaling the numeric variables.

**Recoding**

We divided our dataset into two subsets based on the numeric and categorical data. As we apply different analytical techniques for different data types, this step helped us to thoroughly analyze our dataset. We changed the variable names by adding ‘\_N’ suffix for numeric and ‘\_C’ for categorical variables.

**Labeling**

We labelled the values of two categorical variables, Attrition (Yes = 0, No = 1) and OverTime (Yes = 1 and No = 0), for our analysis.

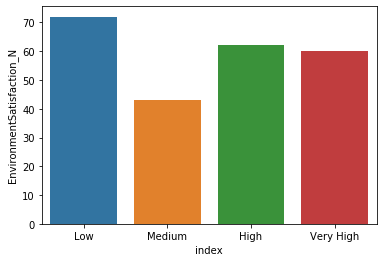
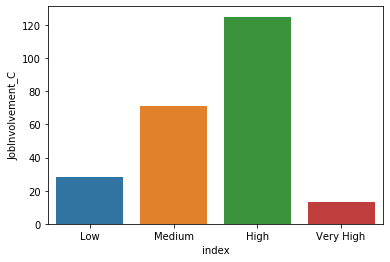
**Normalization**

We scaled the numeric variables using ﻿MinMaxScaler class of sklearn preprocessing library.

**Visualization**

After descriptive analysis we performed following exploratory data analysis to visually represent the dataset and check relationship between different variables.

Below graphs show the environment satisfaction rating (left) and job involvement (right) distribution for employees who left company. This indicates that most of the employees who left might be working in high stress environment.

**Data Distribution analysis**

We checked the skewness and kurtosis values and plotted the distribution plots for each numerical variable. We observed that our dataset does not have high skewness and kurtosis values. The distribution plots showed that the columns: DailyRate, HourlyRate and MonthlyRate, have uniform distribution. We further explored the effect of these columns on attrition and observed that these columns do not affect attrition. Below figure shows the comparative distribution analysis of HourlyRate for both the attrition labels.

A picture containing screenshot

Description automatically generated

**Correlation analysis**

We used **correlation matrix** to check the correlation between numeric variables and we found that 7 combinations have correlation value greater than 0.6. We further analyzed the correlated variables and checked their importance. For categorical variables we used **Chi Square test for independence** and based on the correlation results we can remove the highly correlated variables from our dataset.

**Interpretation**

Based on correlation analysis we made following interpretations:

* TotalWorkingYears is correlated with MonthlyIncome, Age and YearsAtCompany, so we can remove TotalWorkingYears.
* JobLevel is correlated to many other categorical variables so we can remove it.
* MaritalStatus and JobRole has significant dependency but both can play significant role in determining attrition of an employee. So, we should keep both features in our dataset.
* EducationField is somewhat similar to JobRoles so we can drop EducationField.
* JobRole is a sub-department of Department feature, so we can drop Department as we can get the similar information from JobRole.
* Gender, EmployeeNumber, RelationshipSatisfaction, PerformanceRating and Education do not affect Attrition in the dataset, so we can remove these columns.

**Modeling Techniques**

In our analysis we used two different ways to build the model:

First, was to gather domain knowledge using exploratory data analysis, use that information for dimensionality reduction, split data into 70:30 ratio of training and test datasets, build a logistic regression model using training dataset and test model performance using test dataset.

Second, was to scale the dataset, perform boxcox transformation to reduce skewness, perform Principal Component Analysis (PCA) for dimensionality reduction on transformed dataset, use PCA components to split the transformed dataset into 70:30 training and test dataset, perform SMOTE balancing model on training datasets, build Bernoulli Naïve Bayes model using balanced training datasets and test model performance using imbalanced test dataset.

**Type of model**

Logistic regression is a linear classification method that learns the probability of a sample belonging to a certain class. Logistic regression tries to find the optimal decision boundary that best separates the classes. So, we used it for the dataset in which dimensionality was reduced based on domain knowledge.

Naïve Bayes is a classification method based on Bayes’ theorem that derives the probability of the given feature vector being associated with a label. Naïve Bayes has a naive assumption of conditional independence for every feature, which means that the algorithm expects the features to be independent. So, we used it for the dataset in which dimensionality was reduced based on PCA.

**Discussion of results**

Below diagram shows the results obtained for Logistic Regression, the mean accuracy score for training and test datasets are almost similar, which shows that the model does not have any overfitting issue. But we can see that in this model we got very low recall score for ‘1’ i.e. Attrition = Yes and high for ‘0’, which means our model is not accurately predicting who will leave but it is accurately predicting who won’t leave the company. As we did not use the balanced dataset in this model, we can clearly see the impact of imbalanced data on our model.

A screenshot of a cell phone

Description automatically generated

Below diagram shows the results obtained for Bernoulli Naïve Bayes, the mean accuracy score for training and test datasets are almost similar, which means that the model does not have any overfitting issue. In this model we got 0.61 recall score for ‘1’ i.e. Attrition = Yes, and 0.77 for ‘0’, which means our model is 61% accurately predicting who will leave and 77% accurately predicting who won’t leave the company. We used the balanced dataset in this model, and we can clearly see that balancing improves model performance.

A screenshot of a cell phone

Description automatically generated

**Conclusion**

While some level of attrition in a company is inevitable, minimizing it and being prepared for the cases that cannot be helped will significantly help improve the operations of most businesses. As a future development, with a sufficiently large data set, it would be used to run a segmentation on employees, to develop certain “at risk” categories of employees. This could generate new insights for the business on what drives attrition, insights that cannot be generated by merely informational interviews with employees.