Employee Attrition Analysis

The cost of training an employee for them to be situated in the company has a financial impact because the time it takes for the new hire takes at least 12 months to be fully productive. When an existing employee decides to leave the company. There are many factors that affect an employee’s attrition and performance: family, kids, work-life balance, job satisfaction, complexity of the work, relationship status, etc.

Developing a classifier that more strongly weighs relevant factors using Logistic Regression, Decision Trees, or Random Forest could help solve this problem. This is an invaluable method because, if successful, it will be extremely important to improve employee satisfaction with the company and improve attrition and performance. The scope of this project is to find ways to improve employee satisfaction best and retain attrition. Other companies can benefit from similarly specified factors especially when it comes to improving the company's environment to retain employees.

Data

Having the IBM data as a reference point will allow us to use a representation of the workforce at a large-scale company. Given IBM's growth in products and individuals, IBM would serve as a good representation as a baseline for other companies whether to that size and magnitude or hopes to become of that size. Since the data has many variables as below:

- Work-life balance
- Relationship satisfaction
- Performance rating
- Job satisfaction
- Environment satisfaction
- Education level
- Manager
- Recent promotions

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

Exploratory Data Analysis

We graphed the different variables to determine the attrition of an employee:

- Overtime (figure 1)
- Number of companies worked (figure 2)
- Business Travel (figure 3)
- Stock option (figure 4)

Figure 1

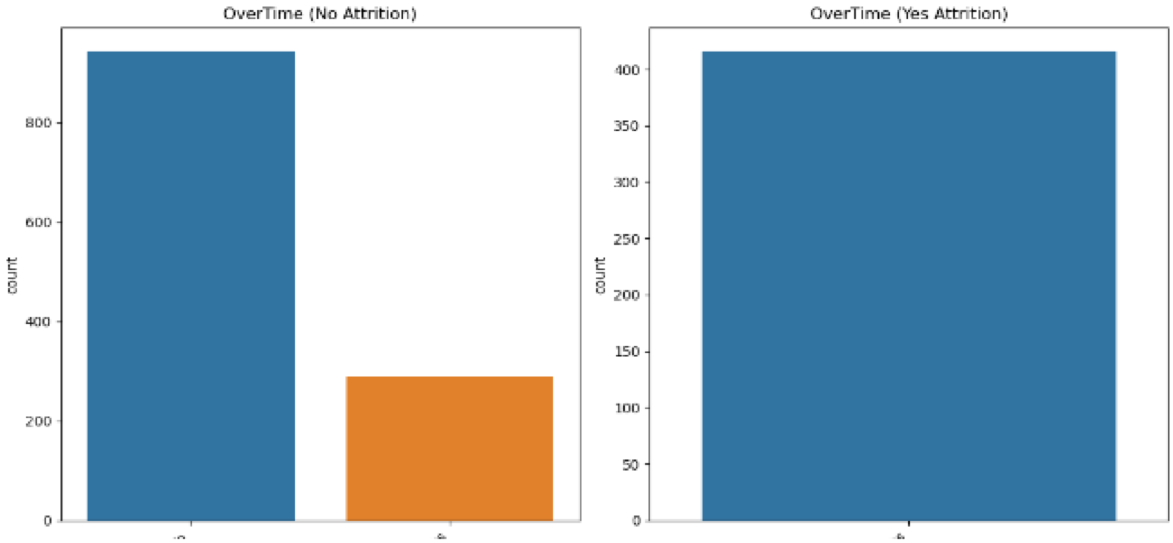


Figure 2

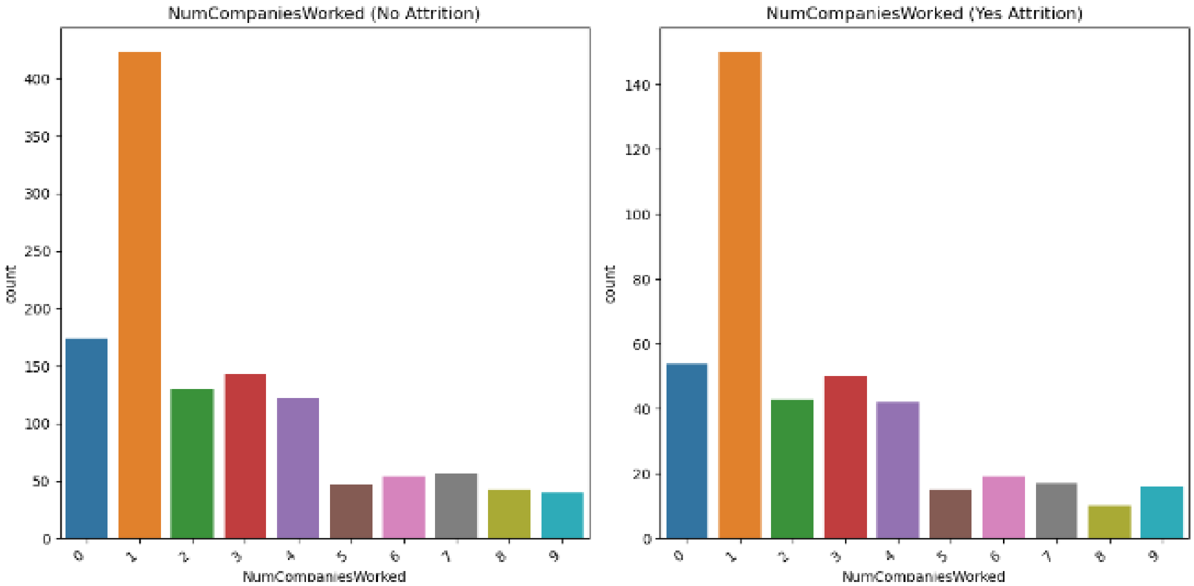


Figure 3

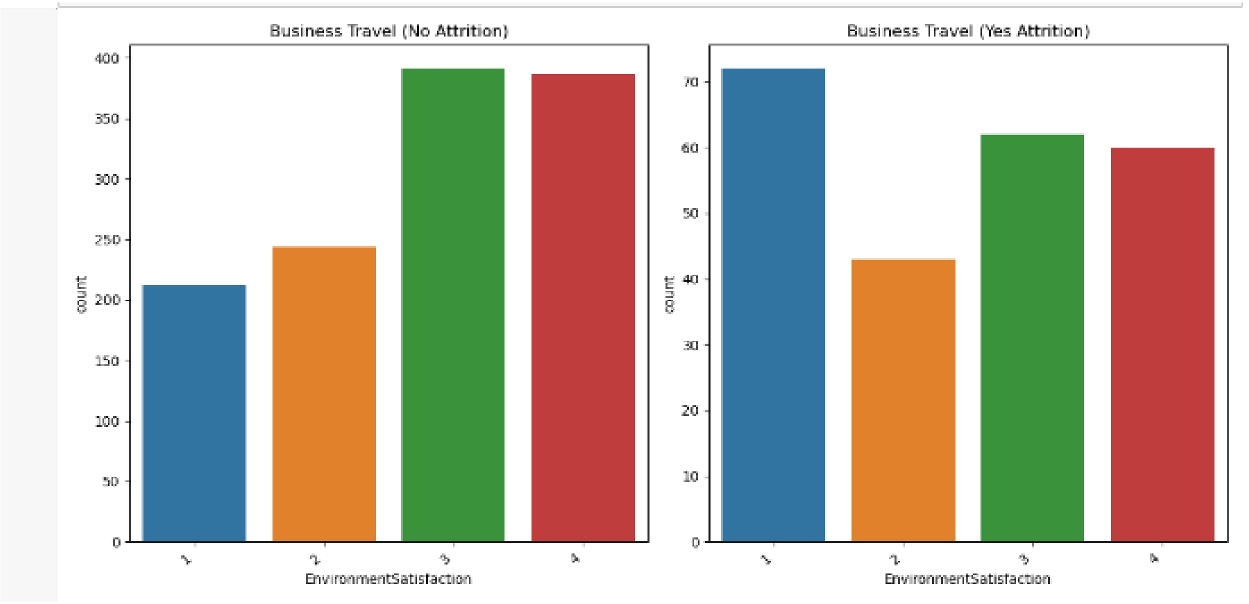
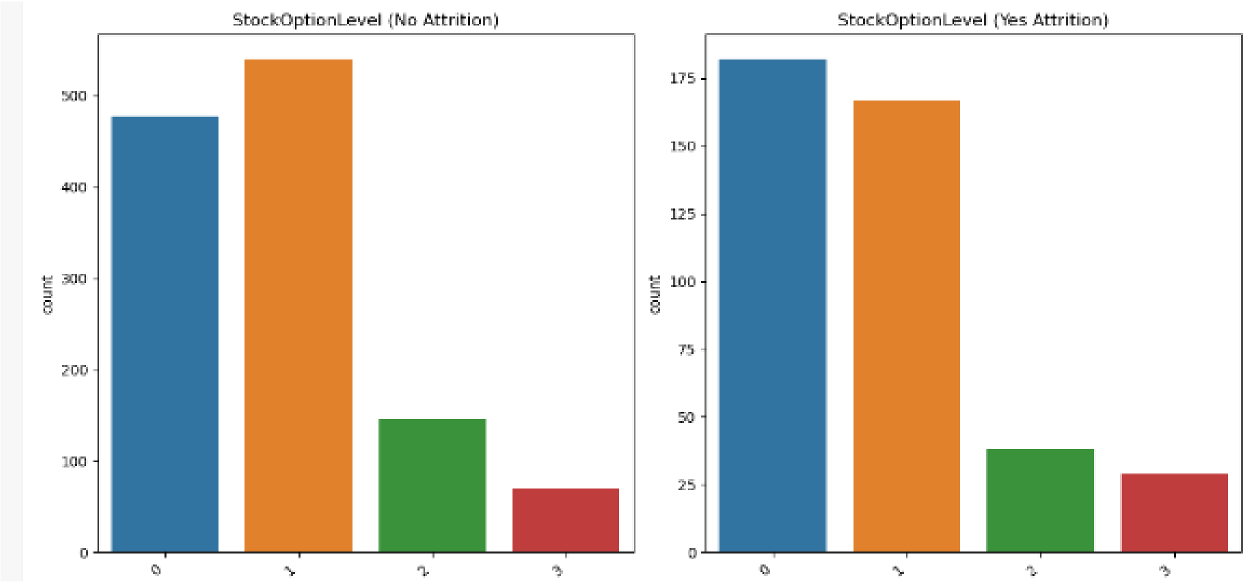
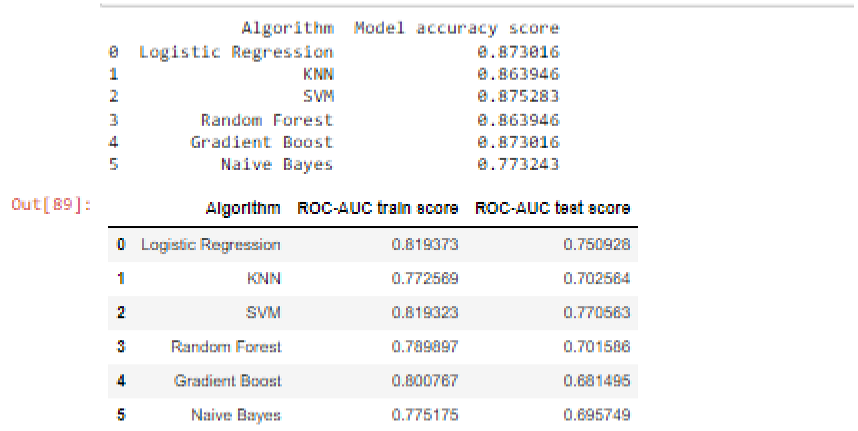


Figure 4



Method/Model

We determined the algorithm and model that would best determine whether an employee would stay at the current company. When finding that, we found that the Logistic Regression was the most accurate model to determine that.



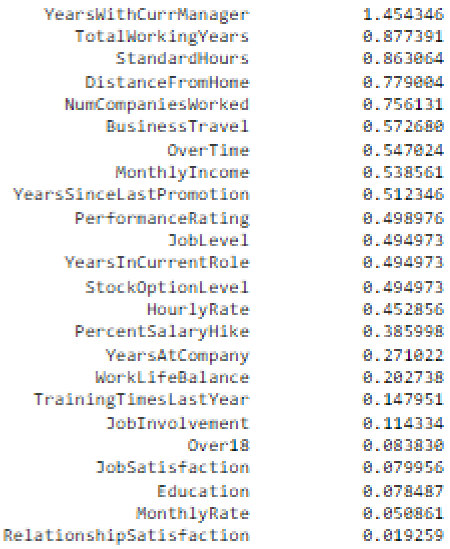
```
              Algorithm  Model accuracy score
0  Logistic Regression              0.873016
1                   KNN              0.863946
2                   SVM              0.875283
3         Random Forest              0.863946
4        Gradient Boost              0.873016
5           Naive Bayes              0.773243
```

Out[89]:

	Algorithm	ROC-AUC train score	ROC-AUC test score
0	Logistic Regression	0.819373	0.750928
1	KNN	0.772569	0.702564
2	SVM	0.819323	0.770563
3	Random Forest	0.789897	0.701586
4	Gradient Boost	0.800767	0.681495
5	Naive Bayes	0.775175	0.695749

Feature Importance

In determining the best attributes/features that would determine whether the employee would stay at the company. We found that the most important features were years with the current manager, total working years, standard hours, distance from home, etc..



```
      YearsWithCurrManager           1.454346
         TotalWorkingYears           0.877391
             StandardHours           0.863064
          DistanceFromHome           0.779004
        NumCompaniesWorked           0.756131
            BusinessTravel           0.572680
                  OverTime           0.547024
             MonthlyIncome           0.538561
   YearsSinceLastPromotion           0.512346
         PerformanceRating           0.498976
                  JobLevel           0.494973
        YearsInCurrentRole           0.494973
          StockOptionLevel           0.494973
                HourlyRate           0.452856
         PercentSalaryHike           0.385998
            YearsAtCompany           0.271022
           WorkLifeBalance           0.202738
     TrainingTimesLastYear           0.147951
            JobInvolvement           0.114334
                    Over18           0.083830
           JobSatisfaction           0.079956
                 Education           0.078487
               MonthlyRate           0.050861
  RelationshipSatisfaction           0.019259
```

Future Improvements

While IBM is a good source of data, having one organization's data prevents the users of the data from predicting for small or medium-sized companies. As companies continue to grow, some other variables and factors could impact an employee’s decision to stay at the company such as mission statement, growth, stock price, etc. Incorporating these variables will allow for a holistic view and a better prediction of whether an employee stays at a company.