Employee Attrition

Lara Mechling, Corrina Hanson, Isaac Liem

Bellevue University

DSC 450 Applied Data Science

Professor Alsaleem

November 13, 2022

Employee Attrition

Table of Contents

Introduction

Employee attrition – turnover – is often very costly to a business. It is even estimated that the cost to replace an employee could range 50% - 200% of that employee's salary. (Ismail, 2022) While this fact alone is sufficient to encourage companies to reduce turnover, cost is not the only factor in losing an employee. High employee attrition leads to losses in productivity, skills, essential knowledge, and synergy. Moreover, these things only exacerbate the overall cost to the company. For the purposes of this project the terms attrition and turnover will be used interchangeably and refer to an employee who left voluntarily or was terminated.

"Fifty-two percent of voluntarily exiting employees say their manager or organization could have done something to prevent them from leaving their job." (Mcfeely, 2019) With this in mind, is it possible to understand what causes employees to leave and make positive changes to alleviate the business losses in order to help a company improve and thrive – ultimately making a company more profitable? This analysis seeks to uncover if an employee will leave the company and what features most heavily influence this prediction.

This analysis is being performed on fictional employee data. (IBM HR Analytics Employee Attrition & Performance | Kaggle) The goal is to create a sound model that may be used on any employee dataset to predict attrition. Understanding employee attrition is beneficial to companies and it imperative to long term growth and business success.

Business Problem

Businesses can sustain costly losses through employee attrition. In order to mitigate this loss, it is important to understand what causes employees to leave. For businesses to thrive, they must continually look at their bottom line. One way to improve upon this is to cut or mitigate employee attrition losses.

Method

The business problem to determine employee attrition was a classification problem with 1 being where the employee stayed with the company and 0 being where the employee left the company. The library utilized was sklearn which is an open-source library built for Python which houses functions built on SciPy, NumPy, and Matplotlib to provide simple and effective tools for predictive data analysis (scikit-learn, n.d.).

In order to determine the model type with the best potential for effectively predicting attrition some of the most widely used classification model techniques were tested on the dataset. These include:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Gaussian Naïve Bayes

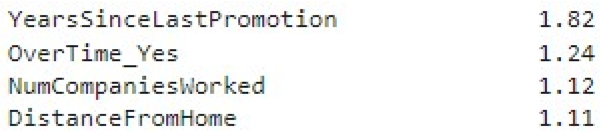
Each of these models has different statistical models behind it and manipulates the data in a different manner thereby producing different results. To test the initial validity of each model the accuracy score was utilized. The logistic Regression model had the highest initial accuracy of 89.8%. This function computes the accuracy, or the number of times the predicted Y results exactly match the actual Y results for the dataset. Though accuracy of a model is important it is not the only method of evaluating a model's effectiveness so further evaluation is needed. To look at other aspects of the model's performance the confusion matrix, overall precision, overall recall, and the breakdown classification report were benchmarked.

Once the initial model benchmarks were created, we proceeded to hypertune the model to further drill down to achieve the best possible results. Grid Search was used to iterate over possible attributes of the Logistic Regression model in different combinations. The results of the Grid Search were printed, and it was determined that the parameters with the most efficient results were a Logistic Regression model with a C of 1.0, a penalty of l2, and a solver style of liblinear. With this information the model was re-trained on the data with the updates to the Linear Regression function.

Once the model was retrained the model’s performance was re-evaluated with the same metrics to include accuracy, overall precision, overall recall, confusion matrix, and the breakdown classification report. The updated metrics were compared with that of the initial Linear Regression model, and it was determined that the hyper tuned model had better results.

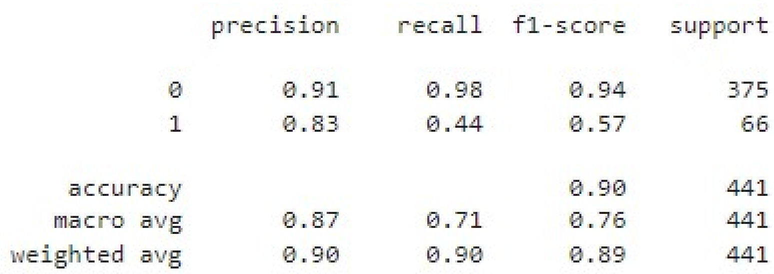
Results

The feature that had the most importance though was if the individual had received a promotion. This was followed by if overtime was present and then distance from home.



```
YearsSinceLastPromotion          1.82
OverTime_Yes                     1.24
NumCompaniesWorked               1.12
DistanceFromHome                 1.11
```

The accuracy came in at a whopping 90.25%, however, when reviewing the classification report it was overly weighted with those who did not leave the company.

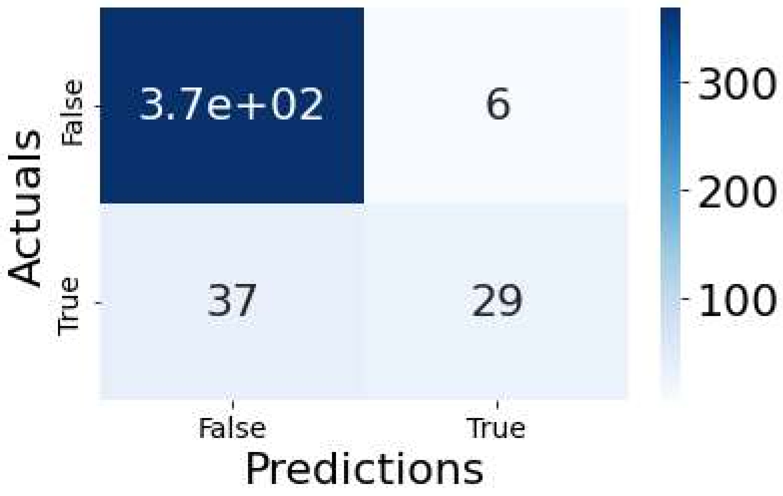


```
              precision    recall  f1-score   support

           0       0.91      0.98      0.94       375
           1       0.83      0.44      0.57        66

    accuracy                           0.90       441
   macro avg       0.87      0.71      0.76       441
weighted avg       0.90      0.90      0.89       441
```

The confusion matrix eludes the previous f1-score where the actuals for attrition are not being captured accurately as seen in the lower-left quadrant.



All-in-all there was not one feature that contributed to attrition, but a multitude of factors.

Recommendations and Ethical Considerations

More data points to be considered could go a long way. Implementing surveys to understand work satisfaction as well as stress levels at work could provide even more clarity into attrition. This has some ethical implications as well since the more data about a given person is provided, the less privacy they can retain.

Conclusion

With employee attrition being a very costly business venture, determining how best to prevent turnover is a high priority among various enterprises. When a business turnover rate is high it causes losses in productivity, skills, essential knowledge, and synergy. Using a machine learning model, the goal is to help aid businesses in determining what the best recourse it to keep their attrition rate low. Using the Kaggle employee dataset several employee characteristics were reviewed and utilized in a Logistic Regression model to determine which factors played a role in attrition. Promotions, overtime, and distance from home played a significant role in determining if an employee would remain with the company. The model was very accurate in predicting the employees who would stay, but the predictions of the employees who would leave the company need to be improved upon. Going forward, more research needs to be done in the area of what truly causes an employee to leave. More data points and employee satisfaction surveys would greatly increase the model’s ability to better predict attrition.

References

Ismail, K. (2022). *Forecasting Attrition With HR Data*. Can HR Analytics Predict Attrition? (reworked.co)

Mcfeely, S., & Wigert, B. (2019). *This Fixable Problem Costs U.S. Businesses $1 Trillion.* This Fixable Problem Costs U.S. Businesses $1 Trillion (gallup.com)

Scikit-Learn Machine Learning in Python. scikit-learn. (n.d.). Retrieved November 13, 2022, from https://scikit-learn.org/stable/

Appendix

Kaggle Dataset Variables:

Age, Attrition, BusinessTravel, DailyRate, Department, DistanceFromHome, Education, EducationField, EmployeeCount, EmployeeNumber, EnvironmentSatisfaction, Gender, HourlyRate, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, MonthlyRate, NumCompaniesWorked, Over18, OverTime, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StandardHours, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager