FEBRUARY 28, 2025

PROJECT 3: MILESTONE 3

WHITE PAPER

SUBMITTED TO: PROFESSOR FRANK NEUGEBAUER
SUBMITTED BY: MARIA AI PAULA COUTO
DSC 680 APPLIED DATA SCIENCE - Bellevue University

Topic: HR analytics- Predicting Employee Turnover. Using Machine Learning techniques to understand why employees leave their organization.

Business Problem

This Project focuses on the use of Machine Learning techniques to determine employee turnover. In every business, Human Resources (HR) manages the people within the workforce. Data science and analytics allows us to use behaviors and patterns from people are keys to maximizing HR performance. We see the application of Machine Learning algorithms in HR functions. Perhaps, the most critical function of talent management is keeping and developing talent to support organizational goals. A study by O'Connell and Kung quantifies the cost of an employee leaving the company through productivity loss, finding a replacement, and training the new employee. (2007) There are also indirect expenses involved in attrition that cannot be measured monetarily but can still have a significant impact on an organization. Therefore, it's important to be able to anticipate if an employee would be highly likely to leave. If we can predict metrics that drive turnover, then we can develop strategic measures to retain employees.

Datasets

I will be using the IBM HR analytics Employee Attrition & Performance dataset from Kaggle. With 35 columns, 1500 rows, and the attrition column (True/False) as the target variable. This dataset contains different features that could potentially influence employee turnover such as travel, pay rate, satisfaction level, etc. I find this dataset interesting because it includes details about the employee's current job such as travel, distance from home, and job role. It also has details about the employees' performance and job satisfaction- which could be important factors in an employee's decision to stay or leave their current job.

Data Explanation and Prep

With 35 features in the dataset, I would like to narrow down the variables that are most likely to affect attrition. I used exploratory data analysis to help understand how the information relates to each other and which ones are the most important. To get this done, I built a summary statistic to show the details of the columns in the data set that are int types. Here, I found some insights that could be useful to determine attrition. For example, the mean total working years of employees was 11 but the mean Years with Current Manager and Years in Current role is 4. Which tells us that most employees have some type of movement within the organization. It's also an indicator that some variables may be highly correlated to each other. One way to optimize a machine learning model is to reduce the number of features in a dataset. However, we need to do this in a way that we don't lose meaningful variables. So, finding variables that are highly correlated to each other would help us eliminate features that have high-collinearity and narrow down the measures we use to determine attrition. For the integer type variables, I used a correlation matrix, along with a heatmap to better identify features that have a high correlation, In this exercise, we see that JobLevel, Monthly Income, and Monthly Rate with a score of 0.95 correlation. I set the threshold of correlation to 0.80 so I dropped the JobLevel, and Monthly Income column. Other columns that have a high correlation are Years at Company and Years with Current Manager, Total Working Years and JobLevel, PercentSalaryHike and Relationship Satisfaction. However, these columns did not meet the threshold, so they are included in the dataset. For the categorical/object datatypes, I used chi-square test to determine the features' relationship with each other. Here I found that the values that have a strong relationship to attrition are Job Role and Overtime. Using EDA, I was able to eliminate the following features:

1. EmployeeCount
2. StandardHours
3. EmployeeNumber
4. JobLevel
5. Gender
6. Over18

Methods

This project uses Logistic Regression and Random Forest Classification to predict features that affect employee turnover. The target variable, Attrition, was treated with Label encoder to fit both models. For the categorical data, the pandas get_dummies function was used to transform the variables into numerical values that can be applied in our algorithm. The dataset was split into an 80/20 train-test split and was treated with a scaler to normalize the values so that metrics that have a high value do not affect the performance of the models. Finally, even though only 2 models were used, the Random Forest classification was repurposed with hyperparameter tuning to see if finding the right parameters would improve the performance of the model.

Analysis

Between the two models, even with the addition of hyperparameter tuning from the Random Forest Classification algorithm, the Logistic Regression model performed better using the accuracy score and recall to evaluate the results.

Table 1. Logistic Regression Classification Report Training Data

ACCURACY SCORE:
0.8997
CLASSIFICATION REPORT:

	precision	recall	f1-score	support
0	0.91	0.98	0.94	986
1	0.82	0.48	0.61	190
accuracy			0.90	1176
macro avg	0.86	0.73	0.78	1176
weighted avg	0.89	0.90	0.89	1176

Table 2. Logistic Regression Classification Report Test Data

Predicted

ACCURACY SCORE:
0.8639
CLASSIFICATION REPORT:

	precision	recall	f1-score	support
0	0.91	0.98	0.94	986
1	0.82	0.48	0.61	190
accuracy			0.90	1176
macro avg	0.86	0.73	0.78	1176
weighted avg	0.89	0.90	0.89	1176

Using the Logistic Regression model, we see that both the training and test data had a high accuracy score (90% and 86%, respectively) which means that it is able to distinguish patterns within the dataset which is able to identify what belongs to the attrition and no attrition class. However, the recall for 1 (attrition_yes) is at 48%. We see this in the correlation matrix where the models predicted a high negative, attrition_no, which mirrors the characteristics of our dataset. However, we see a lot of false negatives where the model predicted attrition_no and the actual value is an attrition_yes. This means that while our model performs well at predicting attr_yes, it struggles to identify attr_no, which is the key value we want to identify in this project.

Conclusion/Assumptions

Based on the performance of the models, I will be extracting the feature importance – the variables in the dataset that has the most effect in the attrition variable using the coefficients from the logistic regression models.

Table 3. Feature Importance

	Feature	Coefficient	Abs_Coefficient
47	OverTime_Yes	0.535080	0.535080
20	YearsSinceLastPromotion	0.516164	0.516164
36	JobRole_Laboratory Technician	0.489830	0.489830

This tells us that the employees who work overtime are at most risk for attrition. Combined with factors such as YearsSinceLastPromotion and if their JobRole was a Laboratory Technician makes them a candidate for the possibility of leaving their current position within this company.

Challenges/Issues/Limitations

The biggest concern that I am anticipating with this dataset is that the attrition rate is at 16%. While a low attrition rate is good for the organization, it will definitely affect the performance of my algorithm. It may make sense to use a more balanced dataset, however, the overall turnover rate for the U.S. in 2024 is less than 5%. So, in the real world, we are more likely to use an imbalanced dataset which makes our project applicable.

Future Uses/Additional Applications/Recommendations

While this project has provided insights into identifying employees that are at high risk for attrition, what an organization does with the data has significantly more impact on turnover than a machine learning algorithm. This model can be used to communicate to managers the factors that may indicate an employee leaving so that it can be prevented. Initiatives to help employees avoid working overtime could help as it is the most significant factor in attrition. Professional development programs may also be helpful so that employees would feel empowered moving up in their career so they can stay within the organization.

Ethical Considerations

In working with HR data, one of the most critical ethical concerns is that people analytics contain sensitive information protected under privacy laws. Data anonymization can be a viable solution in handling any personal information that could identify employees. HR institutions should perform de-identification so the data cannot be linked to a specific person and access to employee data must be limited so only qualified people can gain information about employees. Another ethical consideration that needs to be addressed is to protect employees from biases that may have been inherited from historical data. A regular audit on the data and models should help maintain the integrity of using machine learning models to predict outcomes. Finally, this project aims to predict factors that may be correlated with attrition, not the cause of it. Information gleaned from the dataset should be used to help employees stay within an organization instead of unjust removal or termination.

References

Datasource: https://www.kaggle.com/datasets/raneemoqaily/ibm-hr-analytics-employee-attrition-performance/data

Algorithm:

Logistic Regression- I will be using logistic regression for this model. The IBM website provides an explanation for how the algorithm works and different ways to interpret the outcome.

https://www.ibm.com/think/topics/logistic-regression

Random Forest Classification – This article from Geeks for Geeks was a great resource in helping me understand and implement Hyperparameter tuning

https://www.geeksforgeeks.org/random-forest-hyperparameter-tuning-in-python/?ref=header_outind

Python Libraries

Plotly Express in Python = https://plotly.com/python/plotly-express/

Seaborn = https://seaborn.pydata.org/

The Python Graph Gallery = https://python-graph-gallery.com/

Supporting Data: One of the benefits of being a Bellevue student is access to Gartner, which provides information from different industries. I will be cross-referencing HR articles from Gartner to help with the interpretation of the outcomes.

https://www.gartner.com/en/human-resources/glossary/hr-analytics

Cost of Employee reference from Business Problem: O'Connell, Matthew & Kung, Mavis (Mei-Chuan). (2007). The Cost of Employee Turnover.. Industrial Management. 49. 14-19.