Employee Attrition Prediction

Milestone 4: Project-1 Report

Vinay Nagaraj

DSC680, Summer 2021

Bellevue University, NE

https://vinaynagaraj88.github.io/DataScience_Portfolio/

Abstract

Attrition in business describes a gradual but deliberate reduction in staff numbers that occurs as employees retire or resign and are not replaced. A reduction in staff due to attrition is often called a hiring freeze and is seen as a less disruptive way to trim the workforce and reduce payroll than layoffs. The term is also sometimes used to describe the loss of customers or clients as they mature beyond a product or company's target market without being replaced by a younger generation. It is generally perceived as a negative because of the costs and challenges involved in hiring new employees to take over jobs. However, not all attrition is bad in the long run.

A data-driven approach to analyzing employee turnover can empower your organization to identify why people leave and boost retention. Through this project we will identify the factors that lead to employee attrition and build a classifier model that would help an organization in predicting the employees that can leave the organization so that they can work with employees and take corrective measures to reduce the attrition problem related to employees leaving the organization.

Introduction

According to the U.S. Bureau of Statistics [1], the average turnover rate in the U.S. is about 12% to 15% annually. According to LinkedIn, an average annual worldwide employee turnover rate is 10.9%. However, some industries, such as retail and hospitality, have above the average turnover rates.

Attrition can make a big dent in your organization's bottom line as well as its culture. For an organization to perform successfully, it is important that the employer and the employee have a good relationship and understanding. It is known that if a strong relationship is in place employees will be more productive, more efficient, create less conflict and will be more loyal.

When an employee decides to quit, there will be a lot of challenges for the employer. It will impact their productivity, revenue, experience and also time invested in training the employee. So, it is important for employers to understand why employees are leaving the company.

As part of this project, I will be using the dataset from Kaggle to help uncover the factors that lead to employee attrition and explore important questions such as 'show me a breakdown of distance from home by job role and attrition' or 'compare average

monthly income by education and attrition'. This is a fictional data set created by IBM data scientists.

Methods

The project will be carried out by utilizing the CRISP-DM model. It stands for Cross Industry Standard Process for Data Mining. The process contains 6 steps that will be followed throughout the project.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

As part of this paper, first 4 steps (mentioned above) will be covered in this methods section. Evaluation and deployment steps are detailed in the results and conclusion section.

- Business Understanding - A high employee attrition rate is an expensive problem. When employees leave, a company has to replace them with new hires. Replacing employees costs a lot of money. The reason why replacing employees' costs so much money becomes much clearer when you consider all the expenses associated with employee replacement. First, you need to find and hire new employees. Then you have to onboard new hires and train them. You

should also count in the ramp up time. Inexperienced employees tend to be less productive. Not to mention the time wasted to find, hire and train new employees until they are fully productive. As part as of this project, we are trying to identify the factors that lead to employee attrition and build a classifier model that would help an organization in predicting the employees that can leave the organization.

- Data Understanding - This is a fictional data set created by IBM data scientists. In this case study, we are trying to solve employee attrition issue and feature attrition looks like the feature is most interesting, and this will be the target variable. This tells about whether attrition happened or not. So, the target variable will be Attrition, and the remaining features will be input attributes which will help us find the target attribute.

  As part of this step, we performed graph analysis to see the distribution of the variables.

  Fig 1: Below plot shows the count of Employee Attrition in our dataset.

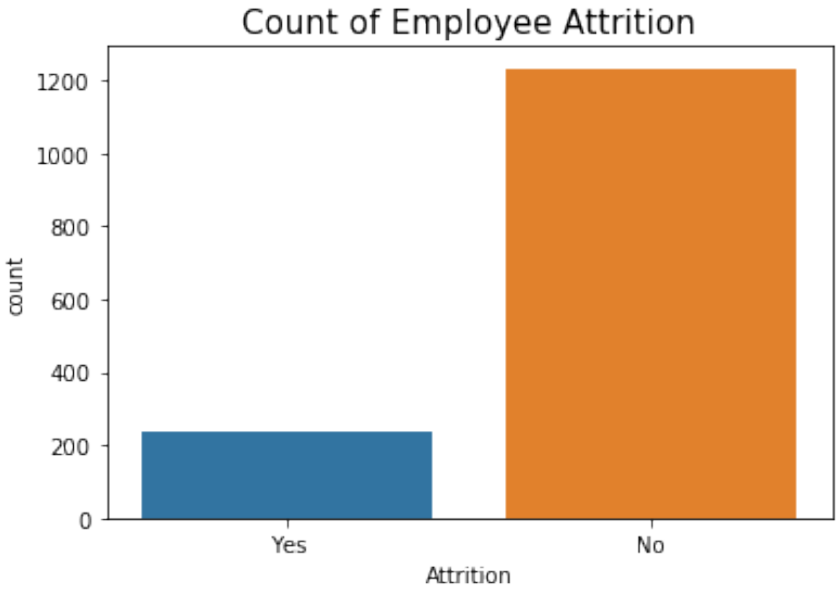


Fig 2: Below plots shows the distance from home distribution (in %) by Attrition. Employees travelling more than 10miles from home are more likely to leave the company.

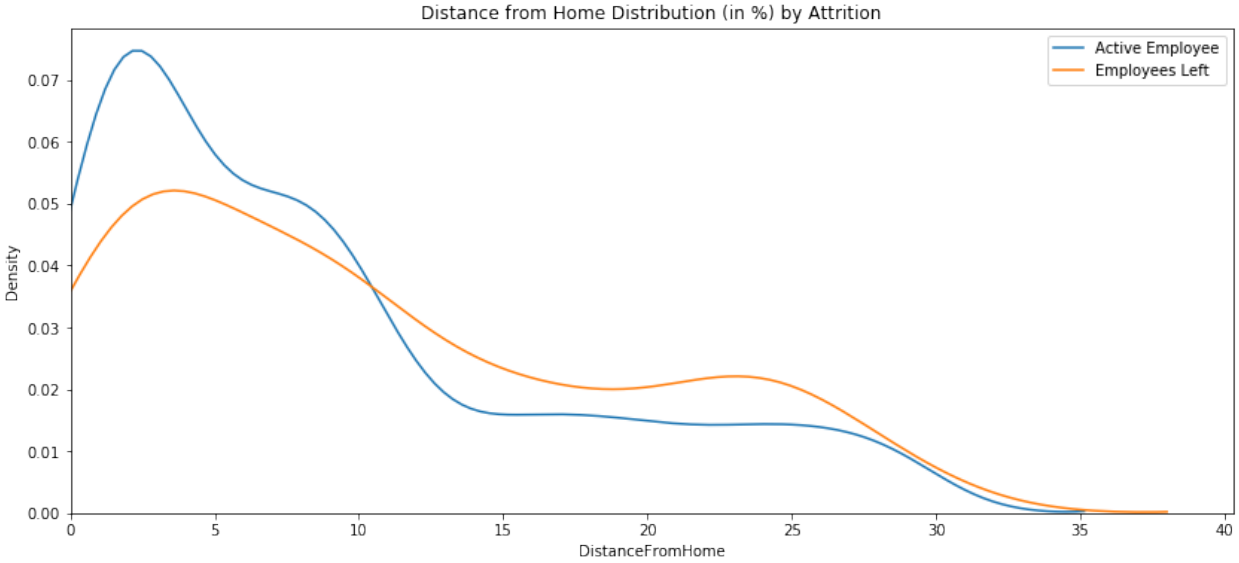


Fig 3: Below plots shows the monthly income distribution (in %) by Attrition. Employees with a monthly income of less than 5,000 are more likely to leave the company.

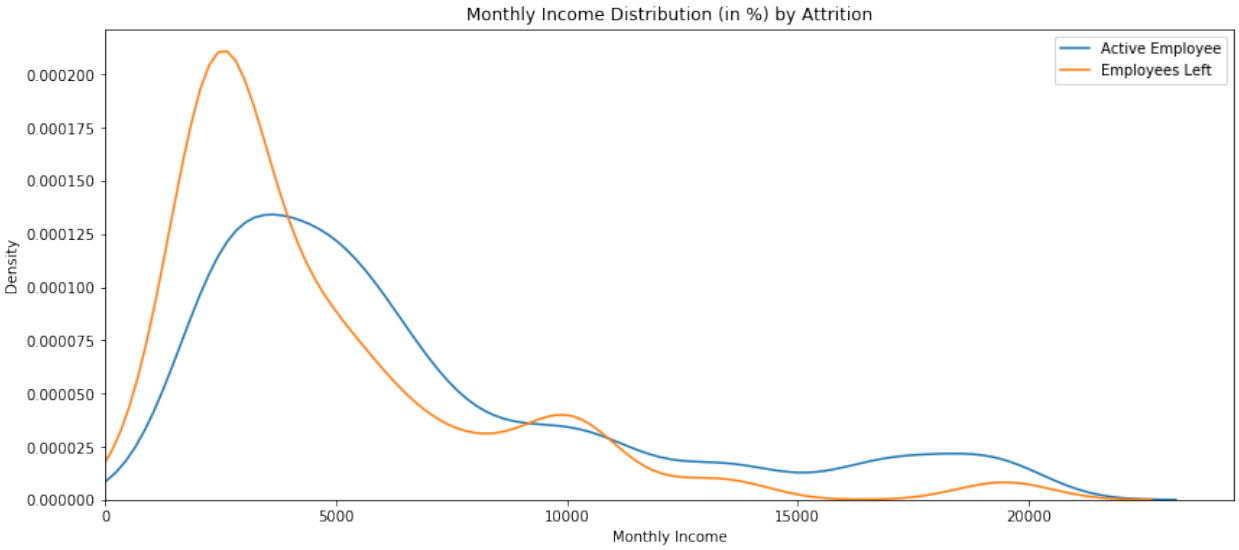
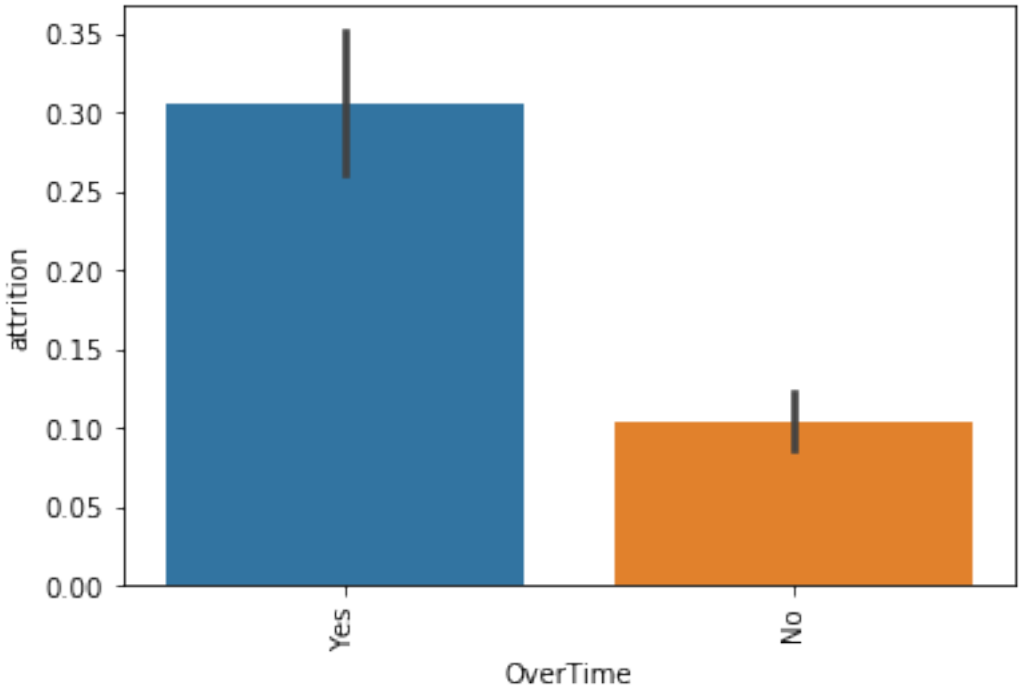


Fig 4: Below plots shows how working overtime affects the Attrition rate. Employees who do overtime are more likely to leave the company.



- Data Preparation – As this dataset is a fictional data set created by IBM data scientists there wasn't much scope with regards to data cleaning and preparation. However, as a step preparing to modeling, I modified all the categorical features to contain numeric values as they were also important features which increased the accuracy of the classification models.
- Modeling – As part of modeling, I perform feature selection which helps us select the features in our dataset which contributes most to the target variable. I executed code for both Variance threshold approach and SelectKBest approach. Variance threshold approach determined that 'OverTime' was not one of the important features and by experience I didn't feel that to be correct. So, I used

the SelectKBest technique to determine 15 best features for the model. This step helps to improve our model accuracy and reduces training time. The SelectKBest scores of all the features of our dataset can be found in Appendix A and the plot that shows the features in the order of their importance can be found in Appendix B.

Using the top 15 features determined by SelectKBest technique, we normalized the data and performed model comparison.

	roc_auc
RandomForestClassifier	0.805351
DecisionTreeClassifier	0.637815
SGDClassifier	0.770046
LogisticRegression	0.816674

Based on the roc_auc values for all the 4 models we compared, RandomForestClassifier and LogisticRgression had the better scores and I decided to pursue these models further.

Models

Based on the above roc_auc scores, I decided to run Random Forest model and Logistic Regression model.

1. **Random forest model** creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to

decide the final class of the test object. It provides us a high true prediction values for our dataset. We tested this model on the test dataset.

This model has an AUROC of 0.783 which means that the model has very good discriminatory ability. 371 cases are true negative, meaning they are non-attrition employees and model predicted them as non-attrition. 11 transactions are true positive, meaning they were predicted as attrition employees and are truly attrition employees. 50 instances are false negative and 9 are false positives.

```
Confusion Matrix
[[371   9]
 [ 50  11]]

Classification report
              precision    recall  f1-score   support

           0      0.881     0.976     0.926       380
           1      0.550     0.180     0.272        61

    accuracy                         0.866       441
   macro avg      0.716     0.578     0.599       441
weighted avg      0.835     0.866     0.836       441

Scalar Metrics
        AUROC = 0.783
```

2. **Logistic regression model** takes a linear equation as input and use logistic function and log odds to perform a binary classification task. We tested this model with the test dataset. This model has an AUROC of 0.75 which means that the model has very good discriminatory ability. 370 cases are true negative, meaning they are non-attrition employees and model predicted them as non-attrition. 13 transactions are true positive, meaning they were predicted as

attrition employees and are truly attrition employees. 48 instances are false negative and 13 are false positives.

```
Confusion Matrix
[[370  10]
 [ 48  13]]

Classification report
              precision    recall  f1-score   support

           0      0.885     0.974     0.927       380
           1      0.565     0.213     0.310        61

    accuracy                          0.868       441
   macro avg      0.725     0.593     0.618       441
weighted avg      0.841     0.868     0.842       441

Scalar Metrics
        AUROC = 0.750
```

Results

Organization's performance is heavily based on the quality of the employees. So as an organization it is important to understand the causes of employee attrition and see how the rate can be kept below a certain acceptable threshold. With the help of machine learning algorithms, employers will be able to predict employees who are at risk of leaving the company and also determine the factors that lead to employee attrition.

- As proven by Graph Analysis and SelectKBest, ‘OverTime’ plays a major factor in employee attrition.
- Using Random Forest Model our model will correctly predict if the employee would leave the company or not 78.3% of the time.

- Logistic Regression Model our model will correctly predict if the employee would leave the company or not 75% of the time.
- Random forest model has fewer false positives than logistic regression making it a better model.

I believe ingesting more data to machine learning model will help us get better results from what we have achieved here in our research.

Acknowledgements

I'm indebted to the communities behind the multiple open-source software packages on which we depend. I would like to thank my family for their understanding of our time in this endeavor. Last by not least, thanks to Prof Catie Williams and my classmates for their guidance and feedback.

References

[1] Pavansubhash. (2017, March 31). *IBM HR Analytics Employee Attrition & Performance*. Kaggle. https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

[2] Zojceska, A. (2020, April 3). *HR Metrics: How and Why to Calculate Employee Turnover Rate?* Blog. https://www.talentlyft.com/en/blog/article/242/hr-metrics-how-and-why-to-calculate-employee-turnover-rate

[3] randerson112358. (2020, March 31). *Predict Employee Attrition*. Medium. https://medium.com/analytics-vidhya/predict-employee-attrition-a34e2c5a972d

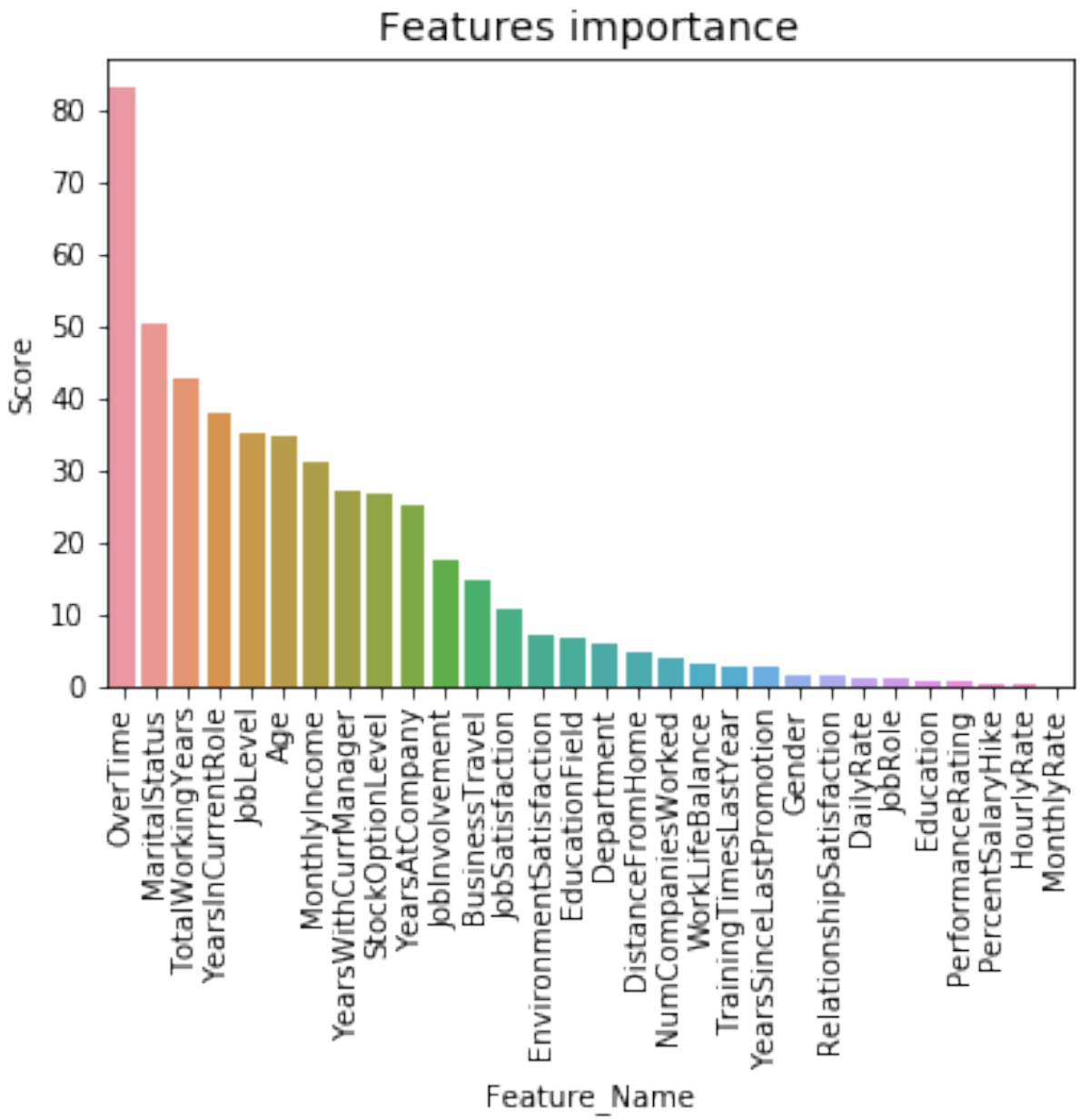
[4] Singh, A. (2020, July 8). *Exploratory Data Analysis-Employee Attrition Rate*. Medium. https://medium.com/swlh/exploratory-data-analysis-employee-attrition-rate-591ce8e7518f

Appendix

Appendix A:

	Feature_Name	Score
18	OverTime	82.898389
14	MaritalStatus	50.198765
23	TotalWorkingYears	42.490621
27	YearsInCurrentRole	37.803235
11	JobLevel	35.027917
0	Age	34.776314
15	MonthlyIncome	31.140654
29	YearsWithCurrManager	27.173446
22	StockOptionLevel	26.475752
26	YearsAtCompany	25.088237
10	JobInvolvement	17.444769
1	BusinessTravel	14.574032
13	JobSatisfaction	10.818824
7	EnvironmentSatisfaction	7.117858
6	EducationField	6.737346
3	Department	5.626314
4	DistanceFromHome	4.436932
17	NumCompaniesWorked	3.941519
25	WorkLifeBalance	3.052695
24	TrainingTimesLastYear	2.750407
28	YearsSinceLastPromotion	2.646695
8	Gender	1.591793
21	RelationshipSatisfaction	1.406934
2	DailyRate	1.201109
12	JobRole	0.936297
5	Education	0.501570
20	PerformanceRating	0.475371
19	PercentSalaryHike	0.127847
9	HourlyRate	0.058277
16	MonthlyRate	0.008998

Appendix B:


Features importance
Score
0
10
20
30
40
50
60
70
80
OverTime
MaritalStatus
TotalWorkingYears
YearsInCurrentRole
JobLevel
Age
MonthlyIncome
YearsWithCurrManager
StockOptionLevel
YearsAtCompany
JobInvolvement
BusinessTravel
JobSatisfaction
EnvironmentSatisfaction
EducationField
Department
DistanceFromHome
NumCompaniesWorked
WorkLifeBalance
TrainingTimesLastYear
YearsSinceLastPromotion
Gender
RelationshipSatisfaction
DailyRate
JobRole
Education
PerformanceRating
PercentSalaryHike
HourlyRate
MonthlyRate
Feature_Name