Predicting Employee Attrition

DSC680 Applied Data Science

Business Problem

Employee attrition is one of the critical factors which affects the organization. Affecting the organization can be of different reasons which include employee knowledge is lost and can be taken to rival organizations. Some of the reasons which affect an organization less are job mismatch, retirement, etc. Attrition is not only the cost of losing the resources, but we also invest in training the newly hired resource. For organizations to be successful, it is essential that the employer and the employee have a good relationship. If the employee decides to leave the organization, there will be challenges the organization faces. It will impact productivity, revenue, experience, and time invested in training. So, we can use some of the machine learning techniques to predict the same. Thus, it is essential for an organization to understand why the employee is leaving them.

Background/History

According to the FinancesOnline Review for Business, There are around 18.9 million Americans either exit the labor force or change occupation every year. Around 3.5 million workers quit their jobs at the beginning of 2020. The average turnover rate in the U.S. is about 20% annually. In February, the number of total separations initiated changed to 6.1 million and 4.1 percent. Attrition will cause a big issue for organizations both in-process and in cost. If an organization performs successfully, the organization and the employee should have a good relationship and understanding. When an employee decides to quit, there will be a lot of challenges for the organization. It will impact their productivity, revenue, experience, and time invested in training the employee. There is around $630Billion in overall costs of employee turnover in the US. The cost of replacing an employee who resigned is around 33% of the employee's annual salary. So, it is important for employers to understand why employees are leaving the company. Using the dataset from Kaggle to uncover the factors that lead to employee attrition and explore important questions such as ‘The time is taken to travel from home to work is a factor for attrition, Among critical factors, is Monthly income a factor to keep employees happy, Will the highly educated employees more tend to leave the organization, Work-life balance relates to employee retention.

Data Explanation (Data Prep/Data Dictionary/etc)

The dataset is a fictional dataset created by IBM data scientists. Looks like, there isn’t much scope with regards to data cleaning and preparation. However, as a step in preparing to model, the categorical features contain numeric values as they were also important features that increased the accuracy of the classification models. The dataset is explored to look at different variables. The dataset has1470 records and 35variables. Attrition – is the predictor variable that contains values Yes / No. ‘Yes’ indicates employee attrition and ‘No’ indicates employee staying with the company.

Methods

With a higher number of columns, I need to identify which columns would be beneficial for the analysis. I will need to perform data exploration and cleaning activities for missing values and perform Exploratory Data Analysis on the dataset data. Identify any outliers using some data visualizations. The features that impact the attrition will be analyzed. Perform Feature reduction and selection and then build some classification models to make predictions on attritions of employees. With the support of the data, the plan is to analyze and apply necessary data preparation techniques which could be useful for creating an effective model. We followed the process as follows:

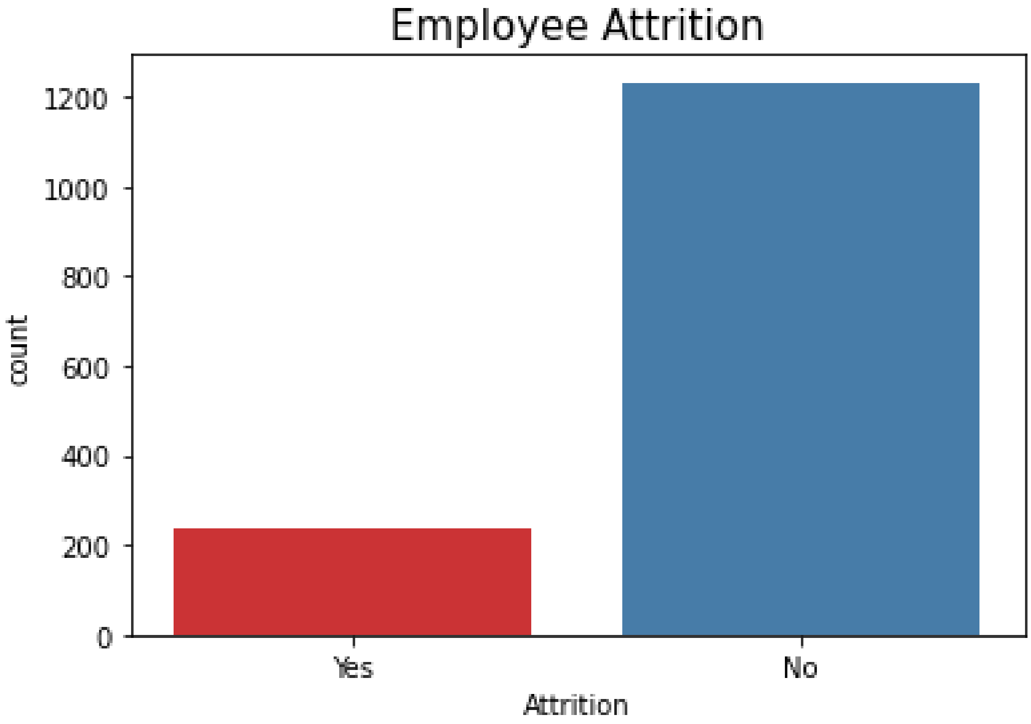
1. Load the dataset.

2. Converted column names for readability.
3. Dropped unnecessary columns.
4. Handled missing and null values.
5. Converted non-numeric to numeric value.
6. Exploratory Data analysis – We created multiple graphs and analyzed the dataset.
7. Train and test the machine learning model – We used sklearn's model selection for splitting dataset into train and test set.

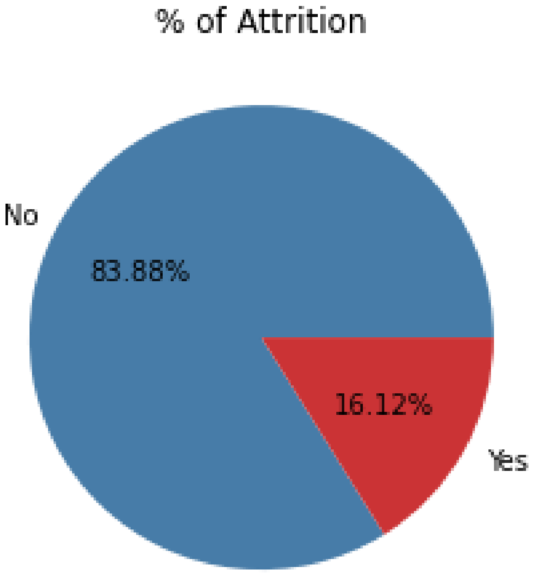
We are training a supervised learning model where Attrition is our target variable. Before going through the modeling process, we dropped columns that were irrelevant for our models such as numerical columns and categorical columns. These columns do not have any correlation and are deemed unnecessary.

EDA Analysis: Here are examples of few graphs created for EDA.

Employee Attrition count:

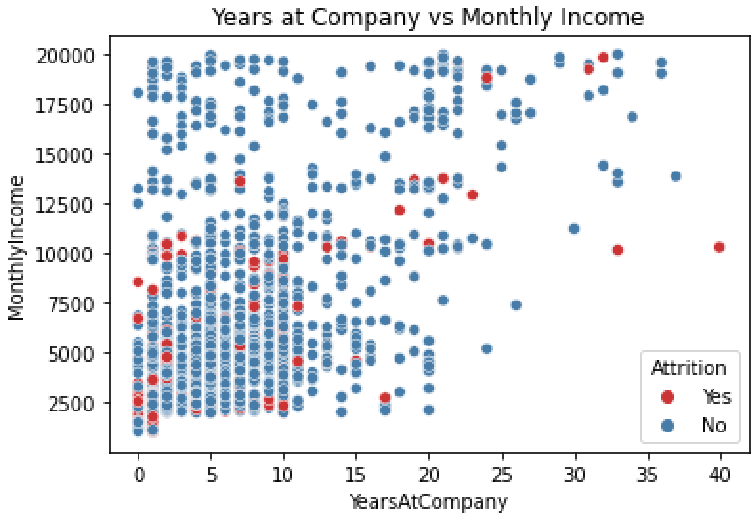


Employee Attrition Percentage:16.12% of the employee leave the organization



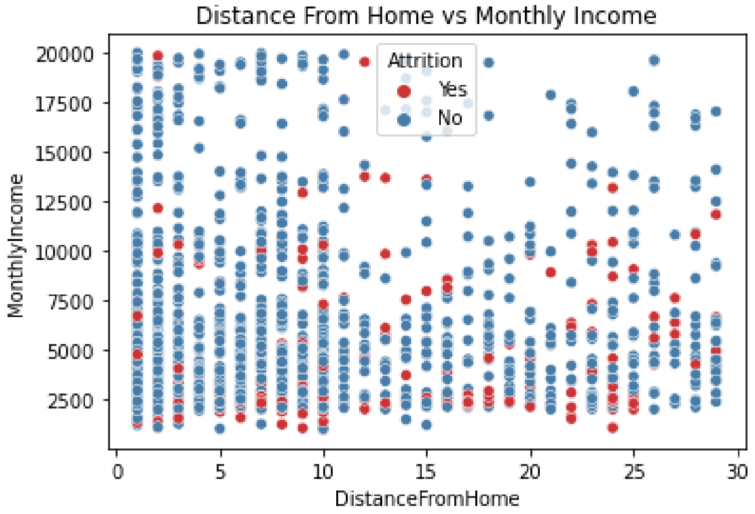
Years at Company vs Monthly Income:

More people are likely to leave at early stage of the company or during there 10 years approx..



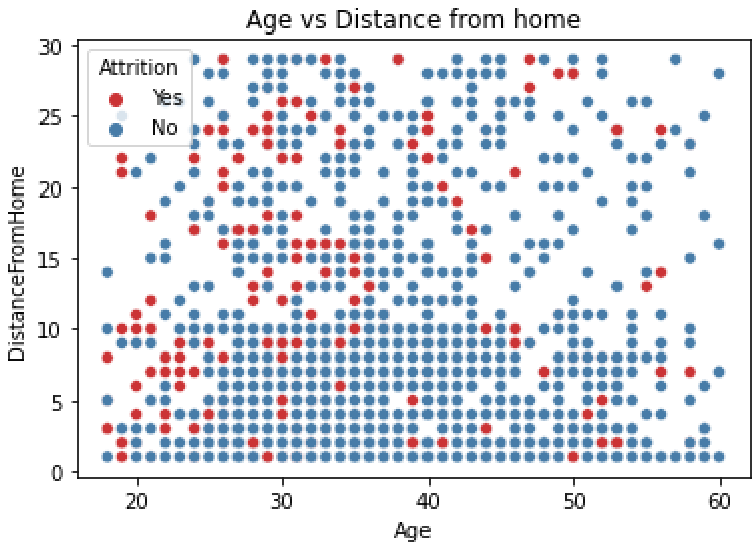
Distance From Home vs Monthly income:

Irrespective of distance from home, If monthly income is low. The employee will leave the organization.



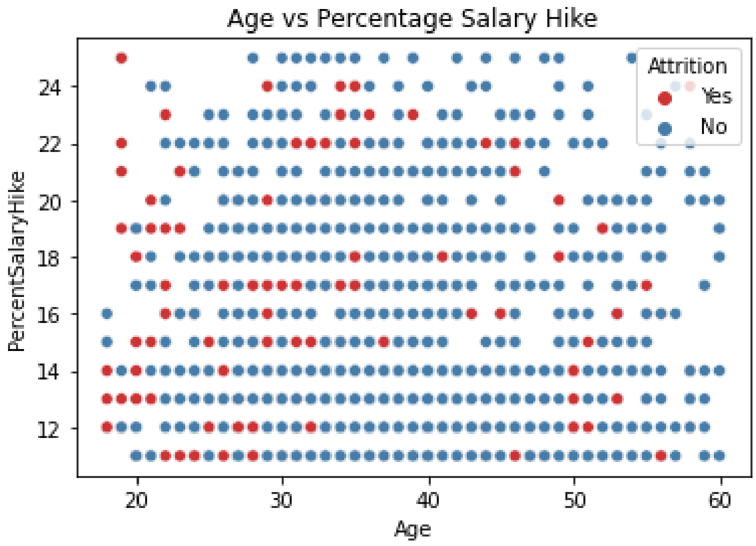
Age vs Distance from Home

If the distance is higher and age is lesser then, there is more probability of leaving.

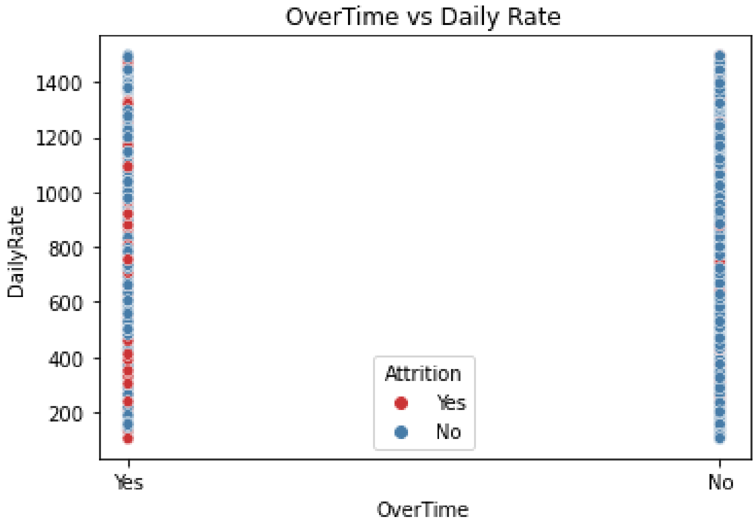


Age vs Percent Salary Hike:

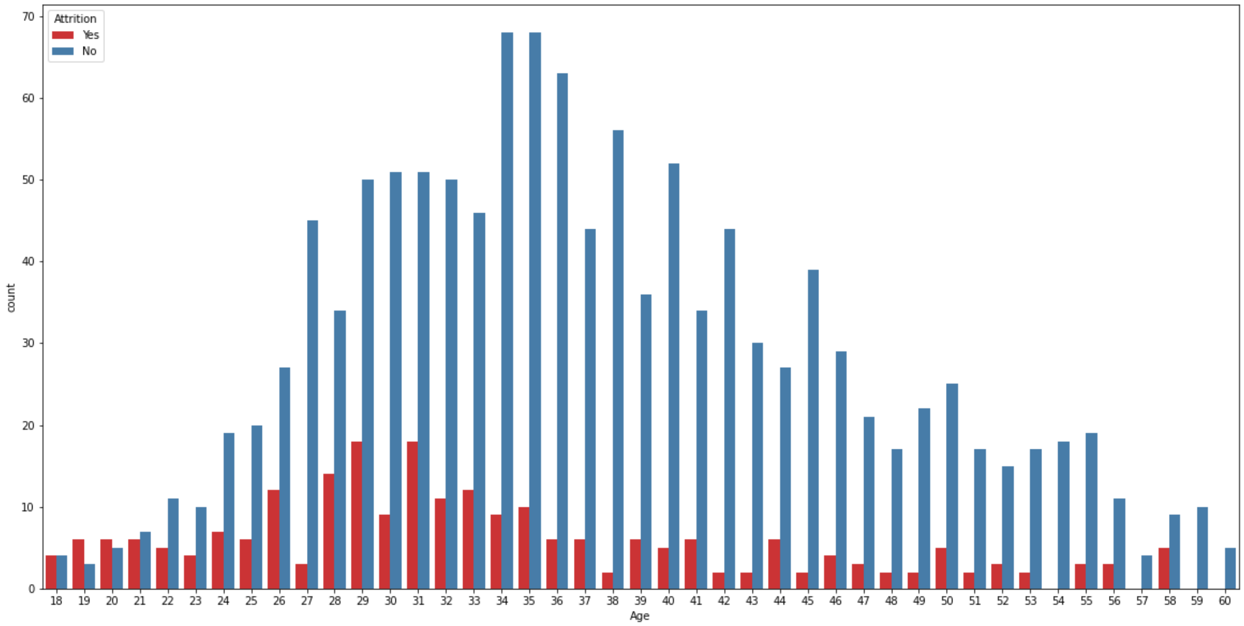
At the early age more people are leaving the organization.



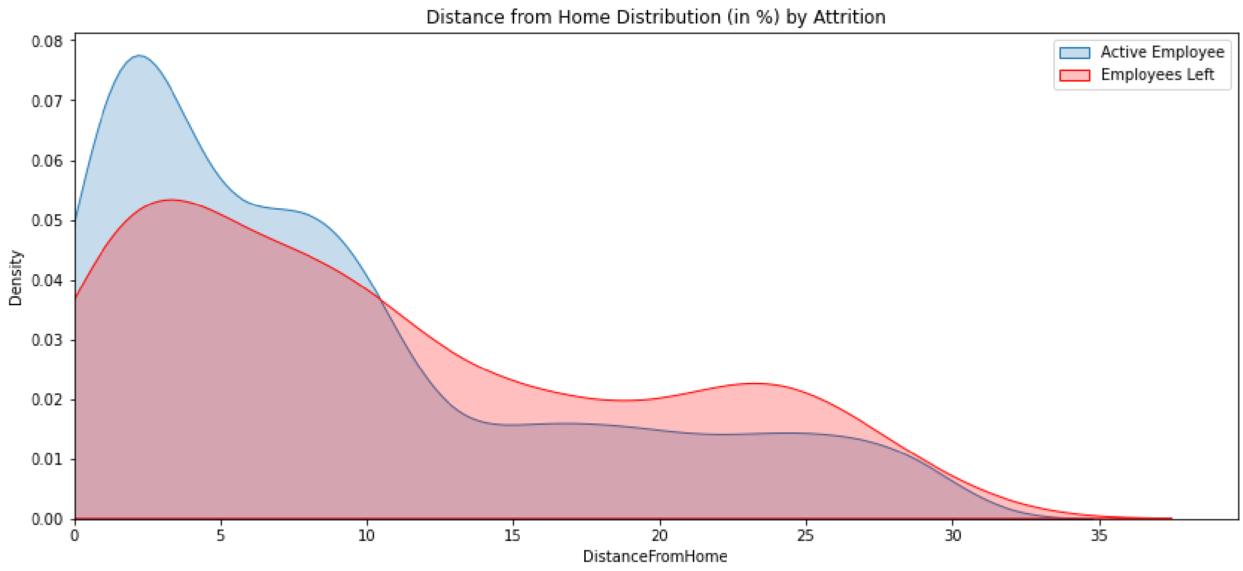
Over Time vs Daily Rate: Irrespective of daily rate, the attrition happens when there is overtime.



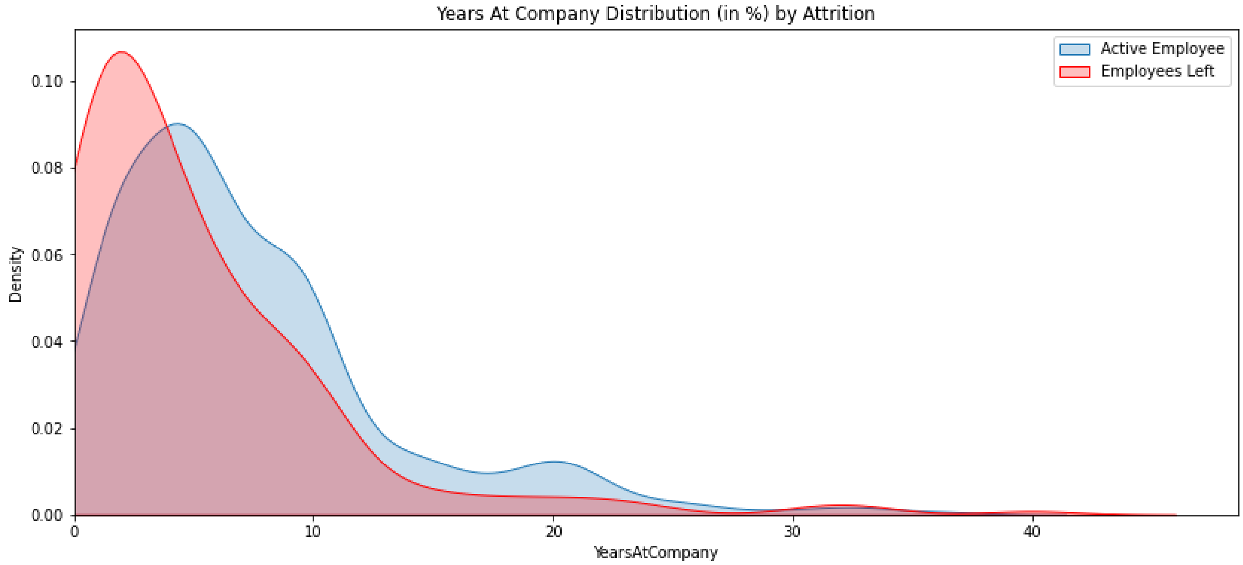
Age factor vs attrition count: People with age of 26 to 35 tend to leave the organization.



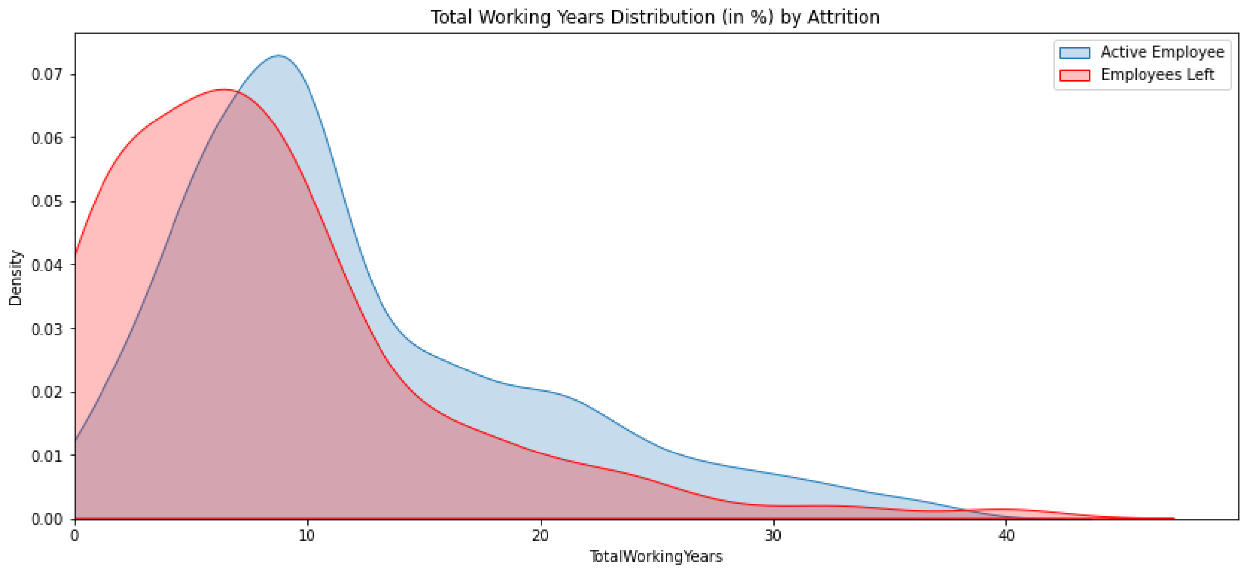
Distance from home distribution by Attrition: The travel distance is more than 10miles from home, the employee is more likely to leave the company.



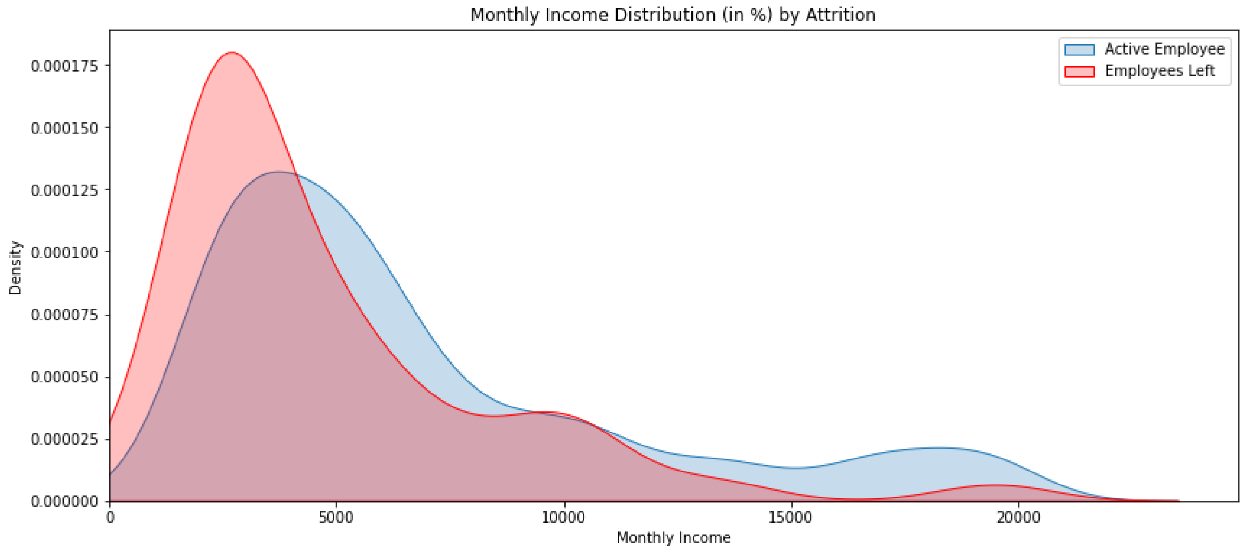
Years at company distribution by Attrition: Employee under 4 years at company is more likely to leave the company



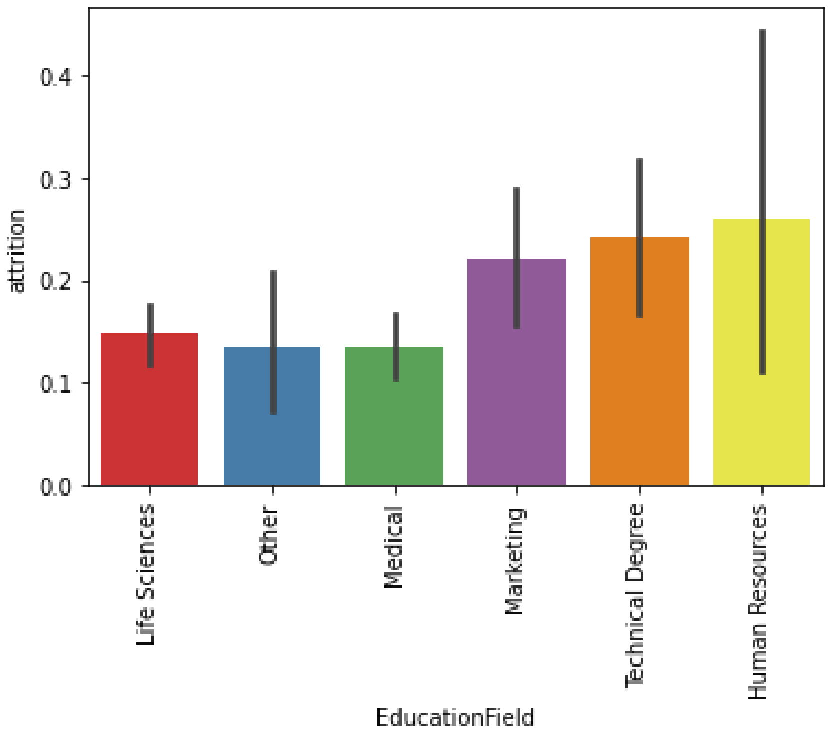
Total Working experience by Attrition: Employee under 8 years of total work experience are more likely to leave the company



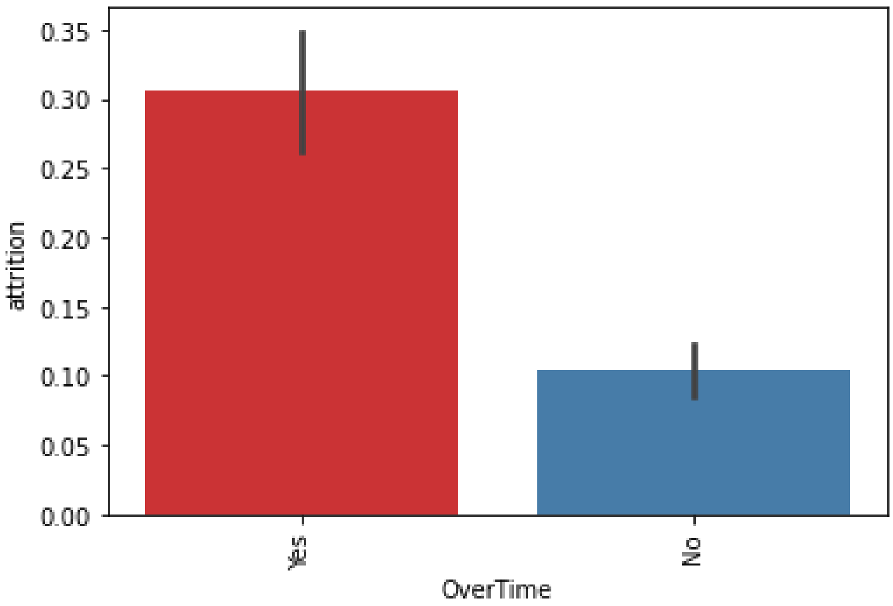
Monthly income distribution (in %) by Attrition: Employees with a monthly income of less than 5,000 are more likely to leave the company.



EducactionField with highest attrition: The top 3 fields where the attrition is higher are HR, Technical Degree and Marketing.

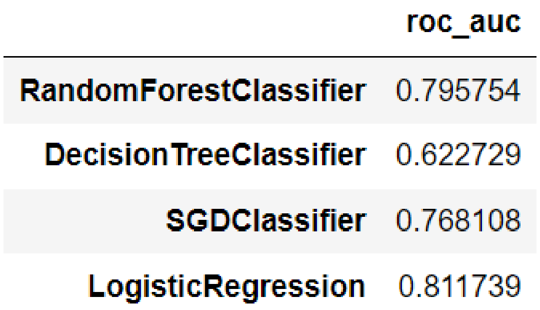


Overtime vs attrition: Employees who do overtime are more likely to leave the company



Models

To prepare and apply a model to this dataset, we split the dataset into two subsets. The first is the training set on which we developed the model. The second is the test dataset which we used to test the accuracy of our model. We allocated 80% of the items to Training and 20% items to the Test set. For modeling we used Logistic regression, Decision tree, Random Forest and SGDClassifier models. 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. Below is table which shows the accuracy based on used machine learning algorithm used.



	roc_auc
RandomForestClassifier	0.795754
DecisionTreeClassifier	0.622729
SGDClassifier	0.768108
LogisticRegression	0.811739

Analysis

Logistic Regression is selected for further performance tuning. Logistic Regression model is tuned with hyper parameters using Grid Search CV. The Grid SearchCV on Logistic Regression classifier increased the ROC-AUC score to 81%.

Conclusion

The employees are the backbone of the organization. If the employee leaves, he takes away most of the stuff. Which includes domain expertise and knowledge. For an organization, it is important to understand the causes of employee attrition. It would be better to have the attrition rate below the 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 attempt to retain them, and also determine the factors that lead to employee attrition. I believe ingesting more data into a machine learning model will help us get better results from what we have achieved here in our research.

Limitations

The employee benefit, better payroll, and the situation of covid have brought in more work-from-home options. This will invalidate the model.

Challenges

I might find many columns with null data or outliers or with only one value. With so many columns, I need to pick the right and required columns for the analysis. Need to check if the data is highly imbalanced.

Future Uses/Additional Applications

A similar model can be used for other applications where the customers are involved.

Recommendations

The efficiency of the models can be improved if the dataset is larger, and balanced so, that the sampling method is not needed. If the original values of the dataset are known, then we can know how the data is correlated and which features are important and train accordingly. In the future different methods can be used to improve the results and more parameter tuning can be done.

Ethical Assessment

While looking at the dataset, I see the age, distance from Home, Role and Martial Status will help in identifying the person. This will an ethical challenge to the privacy. Also, this is a fictional dataset created by IBM data scientists.

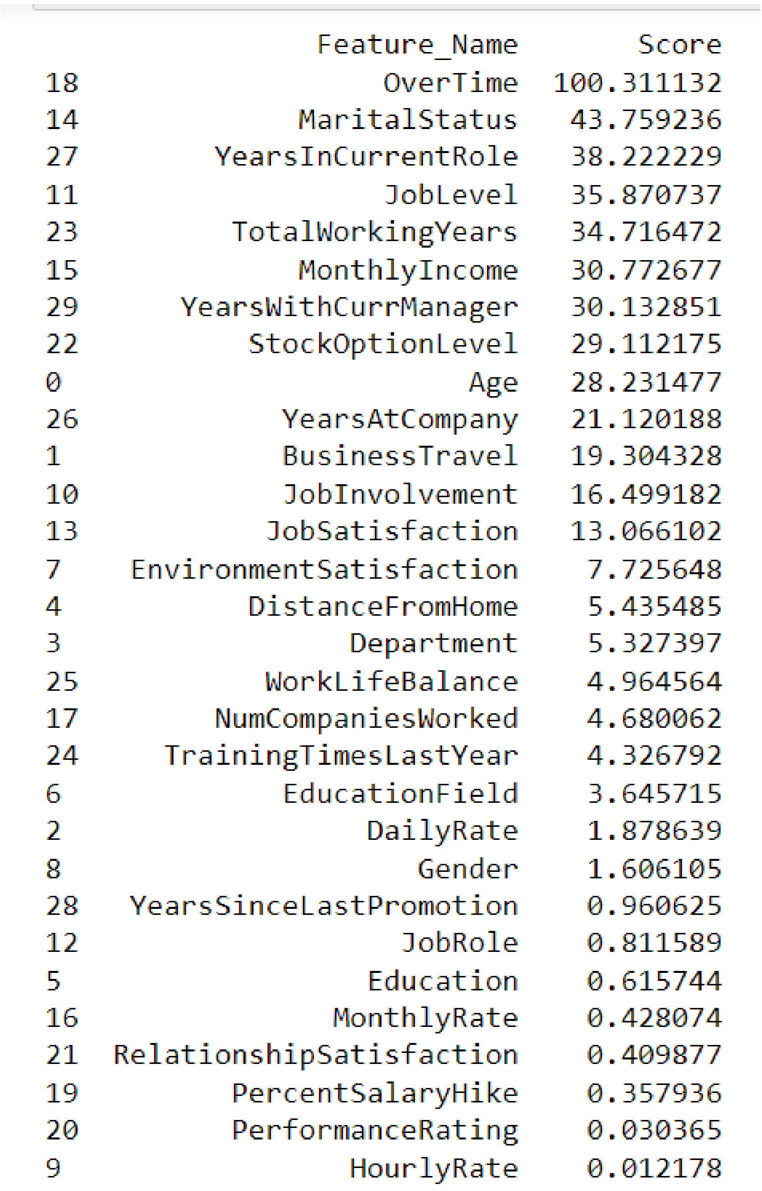
Questions

1. What is the percentage of attrition in the provided dataset?
2. Does overtime impact the attrition?
3. Does age impacts attrition?
4. Does Business travel affect attrition?
5. Which Martial status are more likely to leave the organization?
6. Does distance impact in attrition?
7. If employee stays in company for less than 2 years. What is the probability of he/she leaving?
8. Does Promotion show any impact on attrition?
9. Does total working years impact attrition?
10. Does monthly income impact attrition?
11. Which education field has more attrition?

12. Which is the important feature for attrition?

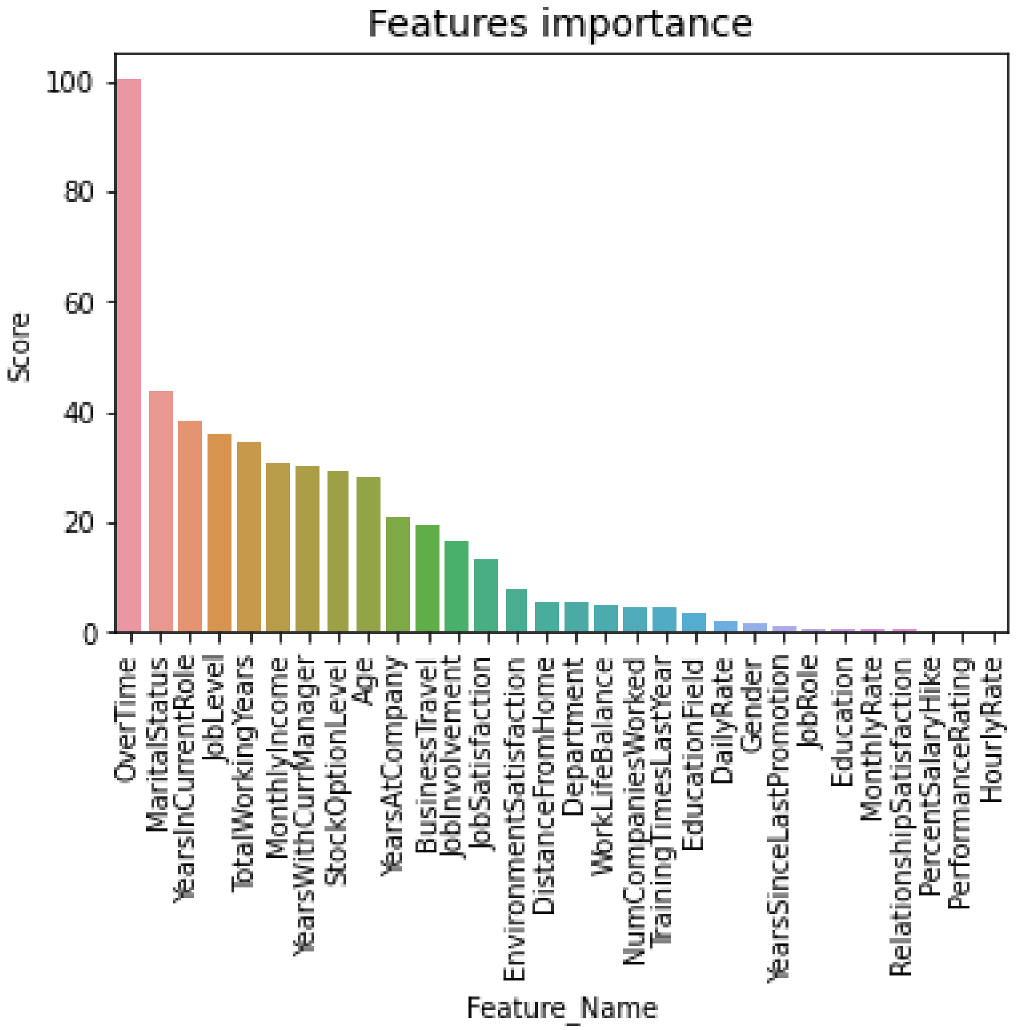
13. Which is the best algorithm for the analysis?

Appendix A:



```
                Feature_Name       Score
18                  OverTime  100.311132
14             MaritalStatus   43.759236
27        YearsInCurrentRole   38.222229
11                  JobLevel   35.870737
23         TotalWorkingYears   34.716472
15             MonthlyIncome   30.772677
29      YearsWithCurrManager   30.132851
22          StockOptionLevel   29.112175
0                        Age   28.231477
26            YearsAtCompany   21.120188
1             BusinessTravel   19.304328
10            JobInvolvement   16.499182
13           JobSatisfaction   13.066102
7    EnvironmentSatisfaction    7.725648
4           DistanceFromHome    5.435485
3                 Department    5.327397
25           WorkLifeBalance    4.964564
17        NumCompaniesWorked    4.680062
24     TrainingTimesLastYear    4.326792
6             EducationField    3.645715
2                  DailyRate    1.878639
8                     Gender    1.606105
28   YearsSinceLastPromotion    0.960625
12                   JobRole    0.811589
5                  Education    0.615744
16               MonthlyRate    0.428074
21  RelationshipSatisfaction    0.409877
19         PercentSalaryHike    0.357936
20         PerformanceRating    0.030365
9                 HourlyRate    0.012178
```

Appendix B:


Features importance
Score
100
80
60
40
20
0
OverTime
MaritalStatus
YearsInCurrentRole
JobLevel
TotalWorkingYears
MonthlyIncome
YearsWithCurrManager
StockOptionLevel
Age
YearsAtCompany
BusinessTravel
JobInvolvement
JobSatisfaction
EnvironmentSatisfaction
DistanceFromHome
Department
WorkLifeBalance
NumCompaniesWorked
TrainingTimesLastYear
EducationField
DailyRate
Gender
YearsSinceLastPromotion
JobRole
Education
MonthlyRate
RelationshipSatisfaction
PercentSalaryHike
PerformanceRating
HourlyRate
Feature_Name