Keiuntae Smith
DSC680
Applied Data Science
19 Oct 2022

Project 2: Milestone 3 White Paper

Topic:
The topic of choice for my second project is to predict employee attrition through machine learning tactics using HR data created by IBM data scientists.

Business Problem:
I have always been intrigued by what makes employees stay with jobs versus deciding to leave a job: personal motivation, professional motivation, challenges with the workforce, or even poor employee-to-job fitment. Employee attrition is the gradual reduction in employee numbers that diminishes a company's force over time. Five different types of employee attrition exist:

1. Attrition due to retirement
2. Voluntary attrition
3. Involuntary attrition
4. Internal attrition
5. Demographic-specific attrition (BasuMallick, 2021)

Predicting employee attrition can assist a company in making the necessary preventative measures to decrease it and aid in making better hiring decisions. HR departments can consequently use employee data to predict attrition and the possible reasons behind it and take appropriate measures to prevent it.

Datasets:
I obtained my data through the Kaggle website (PAVANSUBHASH, n.d.). This data is a fictional data set created by IBM data scientists. It consists of 35 features selected by the data scientists that are various potential factors that could lead to employee attrition throughout a company. These employee features are listed below:
Age
Attrition
Business travel
Daily rate
Department
DistanceFromHome
Education
EducationField
EmployeeCount
Employee number
EnvironmentSatisfaction
Gender
Hourly rate
JobInvolvement
JobLevel

Keiuntae Smith
DSC680
Applied Data Science
19 Oct 2022

Job role
Job satisfaction
Marital status
Monthly income
Monthly rate
NumCompaniesWorked
Over18
OverTime
PercentSalaryHike
Performance rating
relationship satisfaction
StandardHours
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
Work-life balance
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion
YearsWithCurrManager

Methods and Analysis:

The plan for my analysis begins with doing an initial EDA on the dataset to explore it, then analyzing it and making predictions on whether an employee would likely leave the company using machine learning. Succeeding the EDA, I used the Random Forest Classifier algorithm to learn from the data and see how accurate it was. Random forest is a commonly-used machine learning algorithm that combines the output of multiple decision trees to reach a single result. The three benefits of using random forest are that it has a reduced risk of overfitting, provides flexibility, and is easy to determine feature importance (IBM Cloud Education, 2020).

The analysis begins with loading and taking a closer look at the data to see how many columns and rows exist using 'df.shape' and examining the data types of each column using 'df.dtypes'. It shows that the data consists of 1,470 rows and 35 columns. Afterward, I wanted to count the number of empty values, which concluded in no missing data and look at some basic statistical details.

Keiuntae Smith
DSC680
Applied Data Science
19 Oct 2022

	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	HourlyRate	JobInvolvement
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.0	1470.000000	1470.000000	1470.000000	1470.000000
mean	36.923810	802.485714	9.192517	2.912925	1.0	1024.865306	2.721769	65.891156	2.729932
std	9.135373	403.509100	8.106864	1.024165	0.0	602.024335	1.093082	20.329428	0.711561
min	18.000000	102.000000	1.000000	1.000000	1.0	1.000000	1.000000	30.000000	1.000000
25%	30.000000	465.000000	2.000000	2.000000	1.0	491.250000	2.000000	48.000000	2.000000
50%	36.000000	802.000000	7.000000	3.000000	1.0	1020.500000	3.000000	66.000000	3.000000
75%	43.000000	1157.000000	14.000000	4.000000	1.0	1555.750000	4.000000	83.750000	3.000000
max	60.000000	1499.000000	29.000000	5.000000	1.0	2068.000000	4.000000	100.000000	4.000000

Moving forward, I wanted to obtain a count of the number of employee attrition by taking a closer look at the employees that stayed, which is identified by (no) vs. the number of employees that left the company (yes). I was able to visualize this count using the seaborn Python library:

See Appendix, Figure 1: Bar Chart of the Count of Employees that Stayed

I also wanted to look at how age and monthly income affect attrition regarding the data. To do this, I wanted to show the number of employees who left and stayed by age and examine the monthly income using 'sns.pairplot'. The data shows that the highest count of employee attrition ranges between the ages of 29 and 31, and the highest retention age is between 34 and 35.

See Appendix, Figure 2a: Bar Chart of Age of People That stayed vs. Left

See Appendix, Figure 2b: Pair Plot of Age and Monthly Income

In addition, I also wanted to show the distribution of educational backgrounds among the employees. To visualize this, I created a pie chart of the Education fields.

See Appendix, Figure 3: Pie Chart of Education Background Distribution

I then printed out the object data types to take a look at the unique values they possess. In doing so, I realized I did not need some columns for the machine learning process portion. I deleted the columns' StandardHours,' 'Over18', and 'EmployeeCount' since these columns contain only one value in every row and did not add any additional information to the model. Once the columns were removed from the data set, I acquired the correlation of the remaining columns. This correlation was visualized through a heatmap using the seaborn Python library.

See Appendix, Figure 4: Correlation Heatmap of Features

'TotalWorkingYears' is strongly correlated with 'JobLevel' and 'MonthlyIncome'. In general, we can see that many variables are poorly correlated. It is desirable to train a predictive model with features that are not highly correlated with each other.

To prepare the data for the model, I transformed the non-numeric columns into numerical ones using the LabelEncoder from sklearn.preprocessing. I then moved on to conduct a minor data

transformation to create a new column to store the age value. This transformation is only to put the age values at the end of the data set. Then remove the column 'Age' from the front of the data set so that the target column of 'Attrition' is first. To commence the modeling process, I split the data into independent 'X' and dependent 'Y' data. The data set was divided into 75% training and 25% testing. After splitting the data, I then applied the Random Forest Classifier to learn from the training data and see how accurate it was on that data.

Conclusions:

I obtained an accuracy score of 0.9791288566243194 using the function forest.score(X_train, Y_train). The model is about 97.91% accurate on the training data. When the confusion matrix and accuracy for the model on the test data were computed, I achieved a score of 0.8641304347826086. This test demonstrates that the model correctly identified 86.41% of the employees that eventually departed the company. The following graph visualizes this:

See Appendix, Figure 5: Confusion Matrix

Assumptions:

To see what the model perceives as the essential attrition features in the data set, I produced code that will return the feature importances. The results yielded that 'Monthly Income' had the highest importance score, followed by 'Age_Years' and that 'Business Travel' and 'Gender' were identified as the lowest importance. I generated a bar chart to visualize the feature importance in descending order.

See Appendix, Figure 6: Bar Chart of Feature Importance

Challenges/Issues:

I am concerned that my data may be too robust for the chosen algorithm. One of the main challenges faced with using the Random Forest algorithm is that many trees can slow the algorithm and render it ineffective for real-time predictions.

Future Uses/Additional Applications

I believe that this method is an excellent opportunity to predict employee attrition. Predicting turnover helps companies understand which employees are at risk of departing and hint at what interventions could be implemented to reduce the attrition change. With machine learning, HR managers can see not just what happened, but understand why it happened, what will happen next, and how to adapt their workforce strategy to align with company objectives. This model can help foresee vacant positions, team budget needs, what employee benefits they can improve to keep employees happy, what departments are the most and least likely to stay at a job for a long time, and more.

Ethical Considerations:

Since I am working with fictitious data, I don't think there are any ethical concerns regarding performing this particular analysis. If this were data about actual employees, privacy would be a significant concern. Safeguarding their information would be a high priority and should be

treated with care. HR professionals must address what is desirable and ethical when collecting and using employee information in a data-rich world. Any employee database is going to have susceptible information.

References:

BasuMallick, C. (2021, March 11). *What Is Employee Attrition? Definition, Attrition Rate, Factors, and Reduction Best Practices*. Retrieved from Spiceworks: https://www.spiceworks.com/hr/engagement-retention/articles/what-is-attrition-complete-guide/

IBM Cloud Education. (2020, December 7). *Random Forest*. Retrieved from IBM.com: https://www.ibm.com/cloud/learn/random-forest

PAVANSUBHASH. (n.d.). *IBM HR Analytics Employee Attrition & Performance*. e Retrieved from Kaggle: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Keiuntae Smith
DSC680
Applied Data Science
19 Oct 2022

Appendix

Figure 1: Bar Chart of the Count of Employees that Stayed

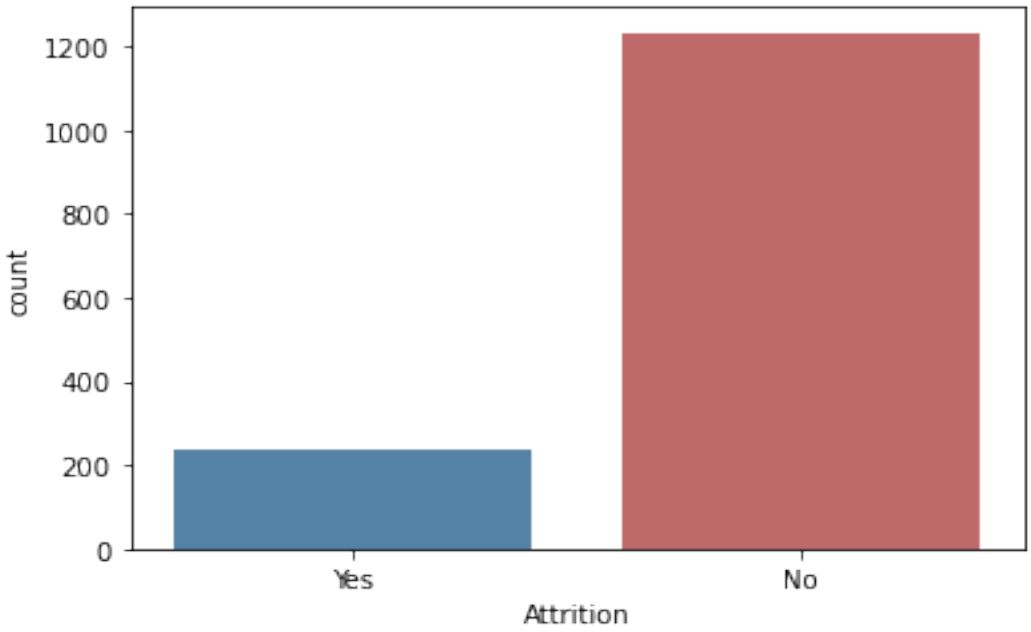
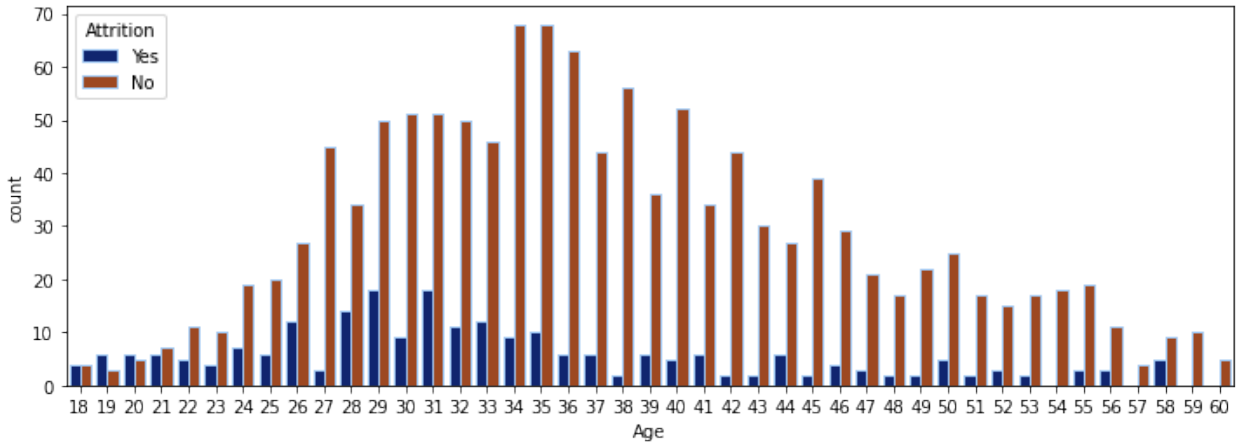


Figure 2a: Bar Chart of Age of People That stayed vs. Left



Keiuntae Smith
DSC680
Applied Data Science
19 Oct 2022

Figure 2b: Pair Plot of Age and Monthly Income

Keiuntae Smith
DSC680
Applied Data Science
19 Oct 2022

Figure 3: Pie Chart of Education Background Distribution

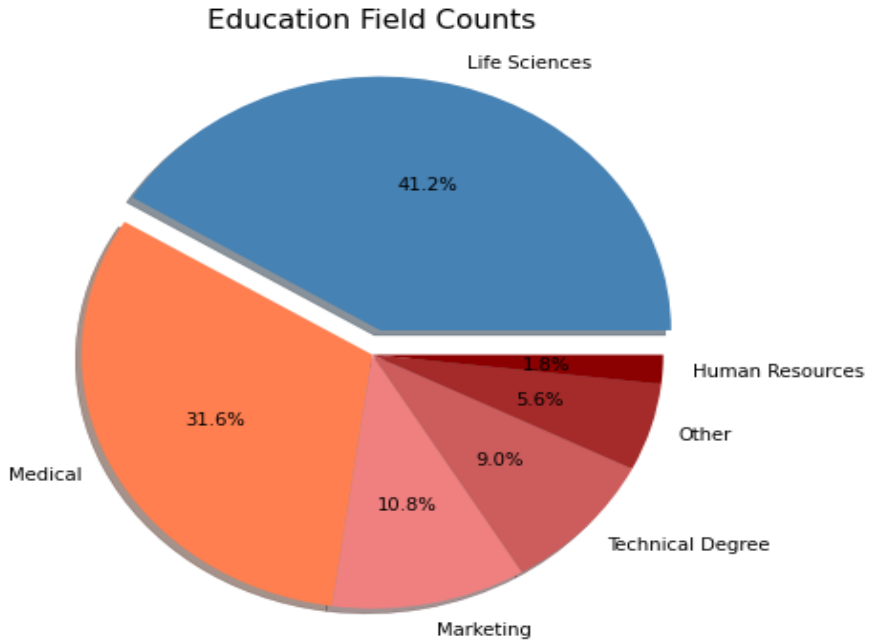
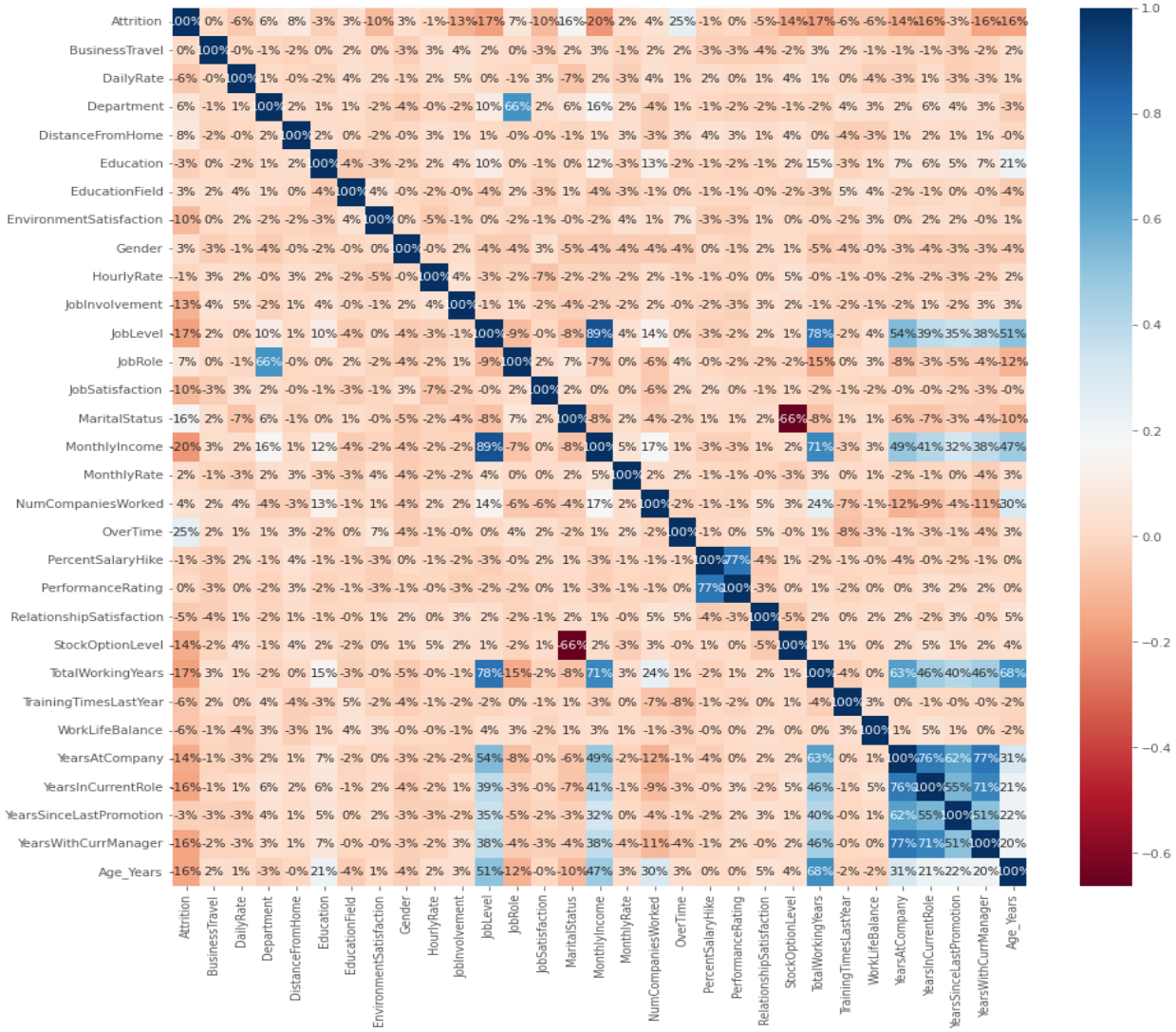


Figure 4: Correlation Heatmap of Features

Keiuntae Smith
DSC680
Applied Data Science
19 Oct 2022



Keiuntae Smith
DSC680
Applied Data Science
19 Oct 2022

Figure 5: Confusion Matrix

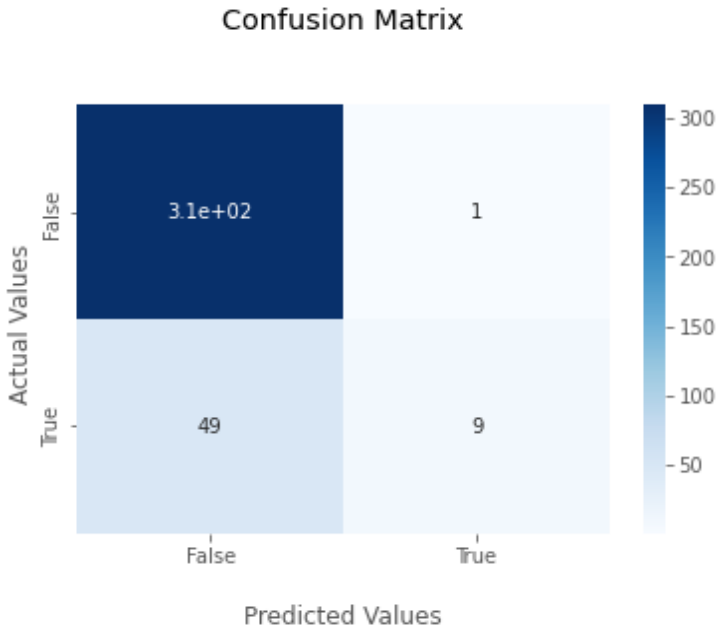


Figure 6: Bar Chart of Feature Importance