Joshua Burden

Bellevue University

DSC550 Data Mining

Dr. Brett Werner

11/19/2022

Introduction

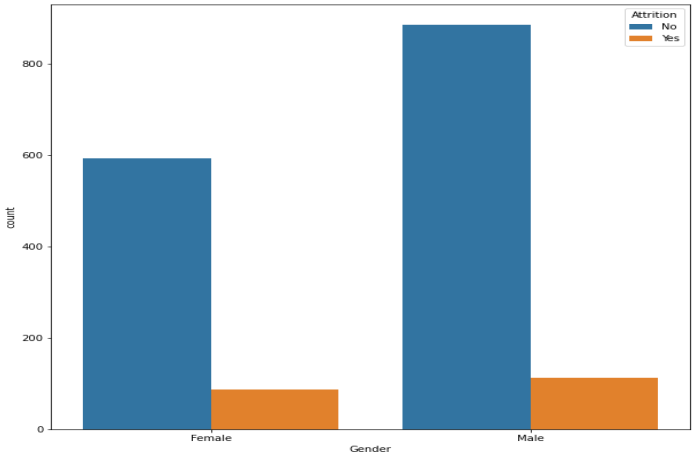
During the pandemic beginning in 2020, there became an evident need for an increase of health care professionals was required. The dataset collected is a modified synthetic dataset from IBM's Watson to show a useful insight into the attrition rate for healthcare workers. The idea behind this data set is to discover whether certain roles within the healthcare industry, hours worked, age of an employee, or any other qualifying data points stand out as to why the healthcare industry had any determining factor on whether a person was to leave their field, while also predicting whether the employee was eventually replaced.

The data set includes information about the attrition rate for employees within the healthcare field. The meaning of employee attrition is the departure of employees from the organization for any reason whether that be voluntary or involuntary, including resignation, termination, death, or retirement. Companies, to avoid attrition rates being too high, need to replace those who are either leaving voluntarily or involuntary. The data set should provide insights into whether a company in the healthcare field was replacing their employees that were leaving the field, or if they continued to have a gradual but deliberate reduction in staff for any reason.

This dataset contains employee and company data useful for supervised ML, unsupervised ML, and analytics. Attrition, whether an employee left or not, is included and can be used as the target variable. The data is synthetic and based on the IBM Watson dataset for attrition. Employee roles and departments were changed to reflect the healthcare domain. Also, known outcomes for some employees were changed to help increase the performance of ML models.

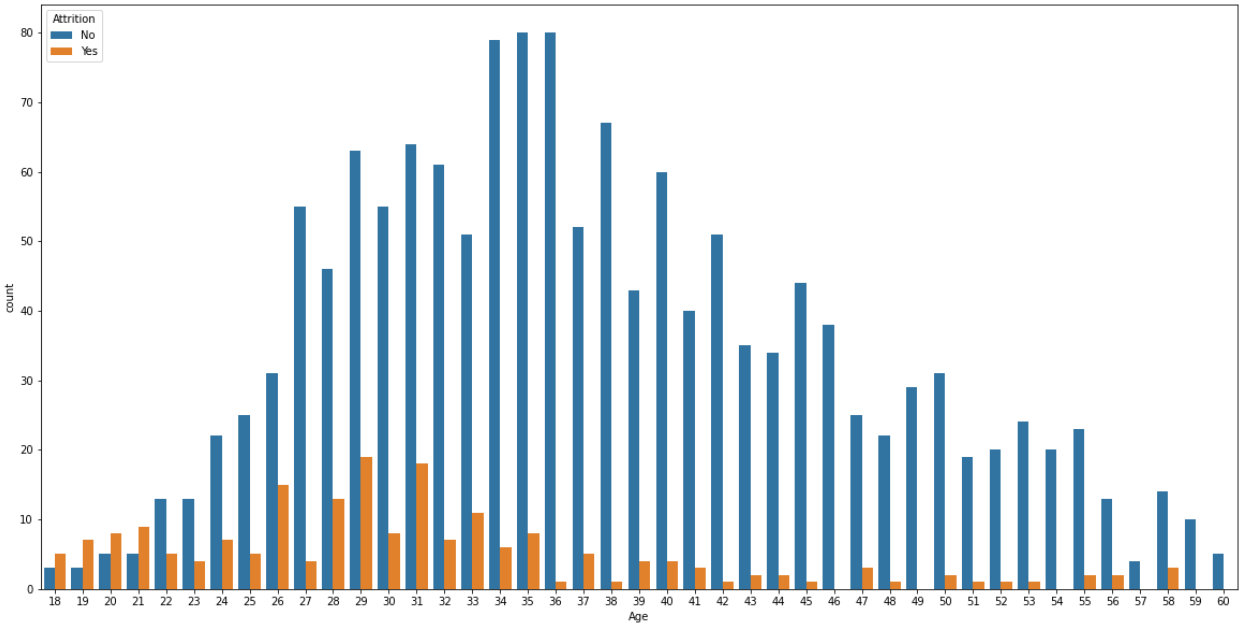
Milestone 1

An analysis of the datasets found that the health care department with the highest rate of attrition was the Maternity departments, followed by the Cardiology and Neurology department. The health care department with the lowest rate of attrition was the Human Resources department. Within the field of education for health care, Life Sciences was the education field with the highest amount of attrition.

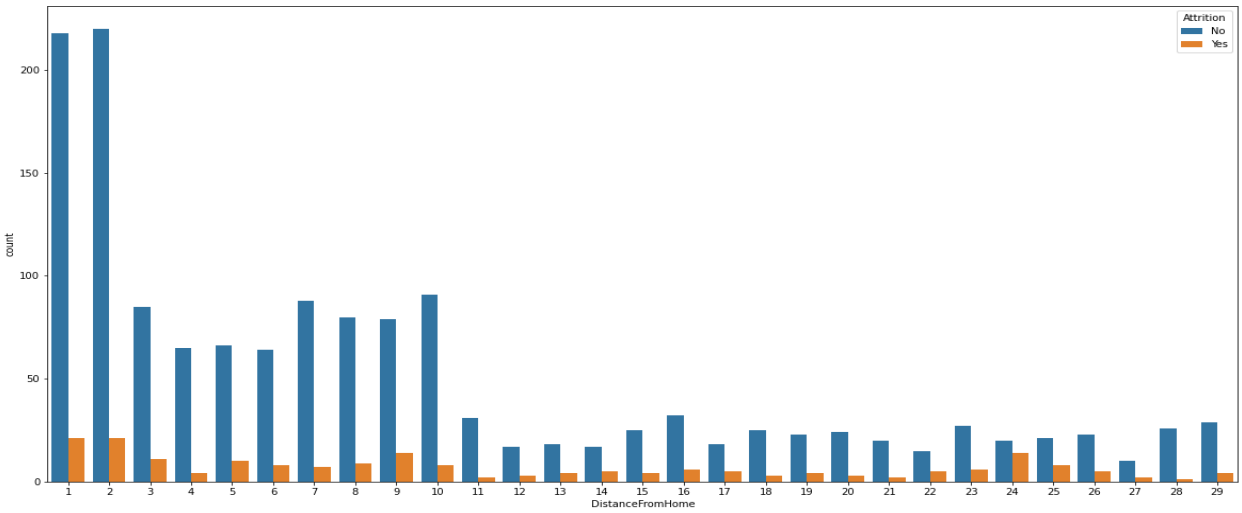


In examining the variable of age, the analysis found that attrition rates peaked at their highest for 29-year-olds. Those between the ages of 26 to 35 possessed the highest range of attrition. While

those 42 years of age and older saw the least attrition rates.



In examining the variable of gender, the analysis found that men were more likely to leave than women, but men were also more accounted for than women in the health care field. Finally, the analysis found that the people who lived closer to their jobs were more likely to leave.



Milestone 2

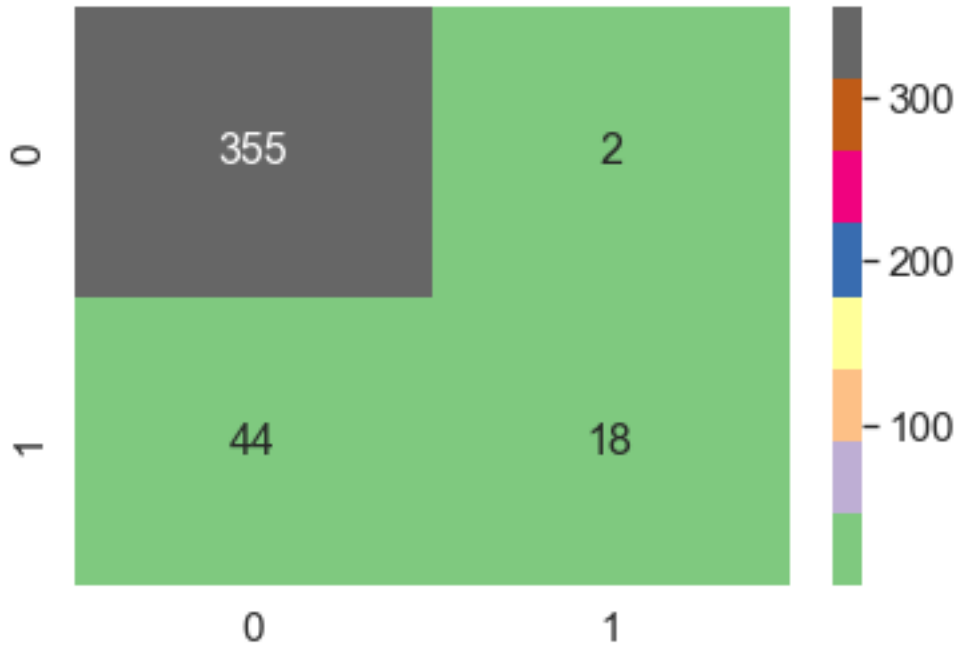
Within Milestone 2, the analysis first sought to drop data columns that did not provide much value or context to the data. The columns removed included; Employee ID, Over 18, Employee Count, and Standard Hours.

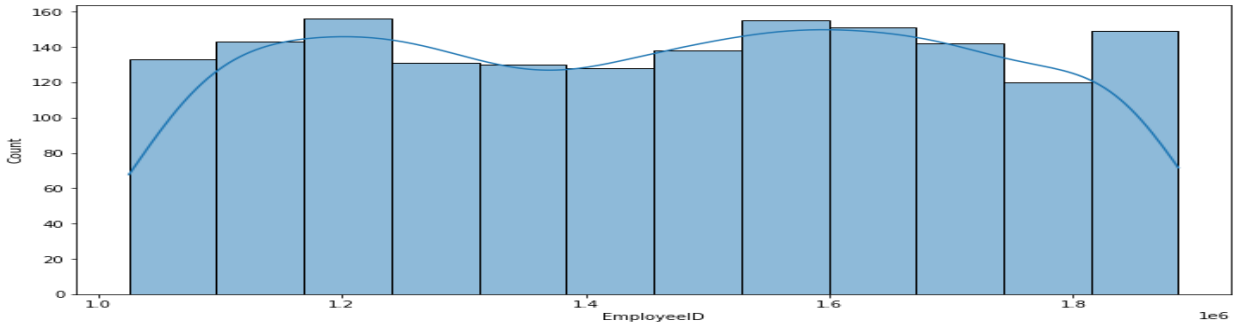
The analysis than converted attritions yes/no values and replaced them 0/1 an set the type to int.

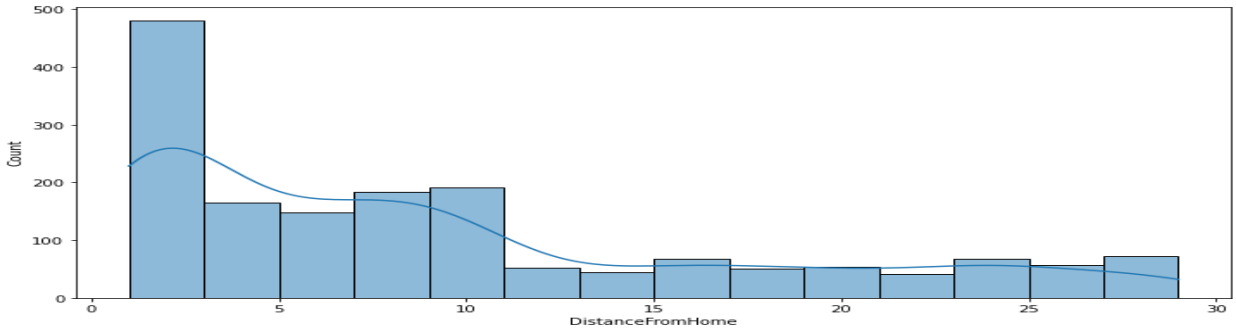
The researcher than checked the correlation of data columns.

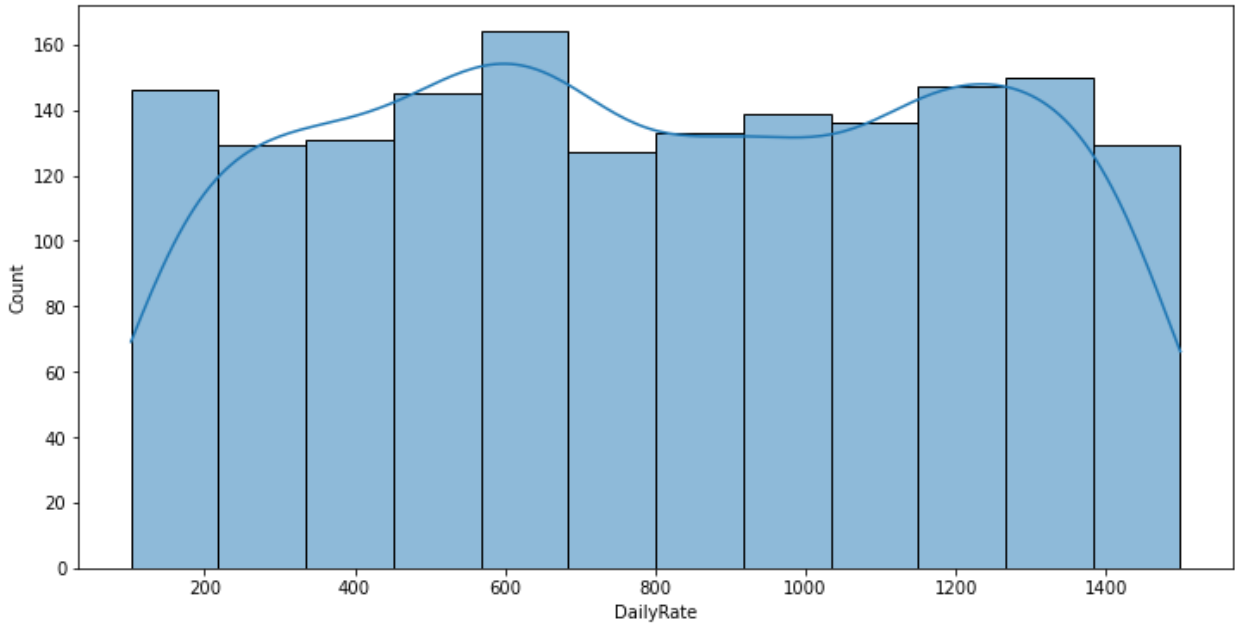
The researcher than dropped all rows that exceeded a threshold of 0.2.

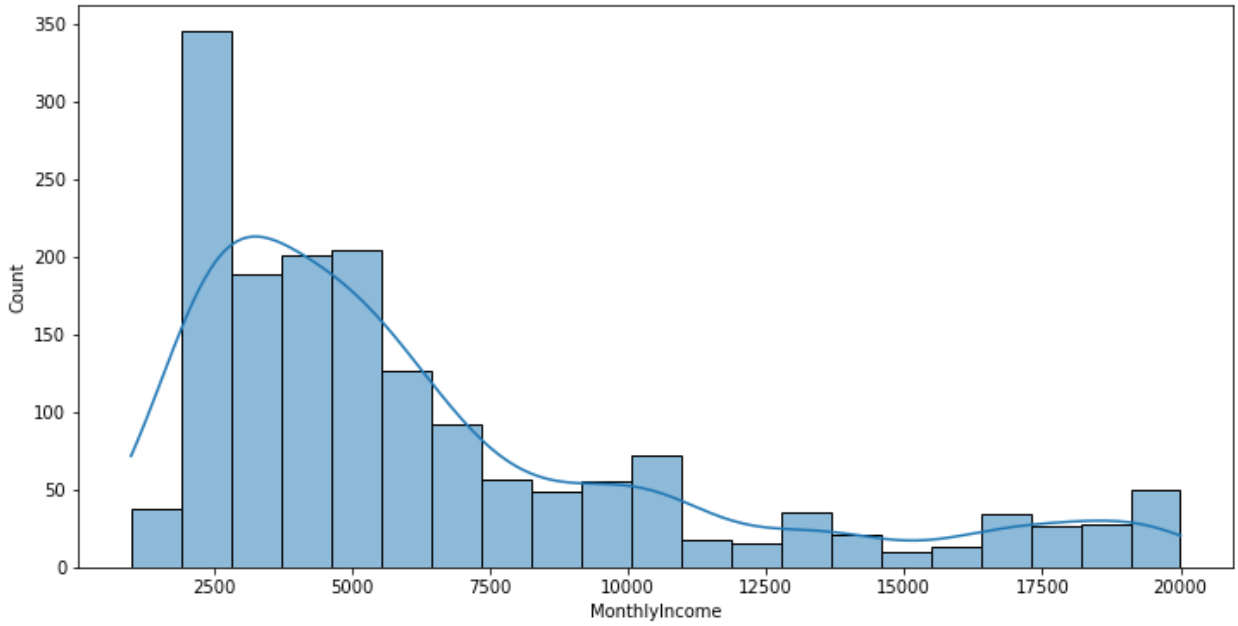
A training and test model set was than created.

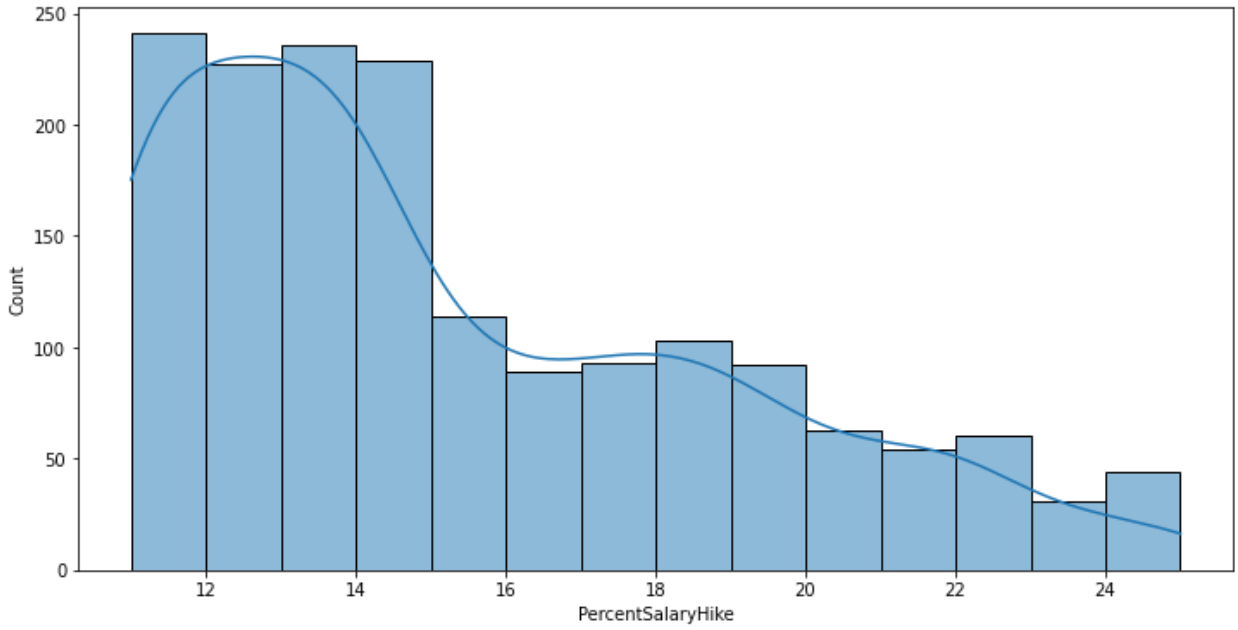

355
2
44
18
0
1
0
1
300
200
100

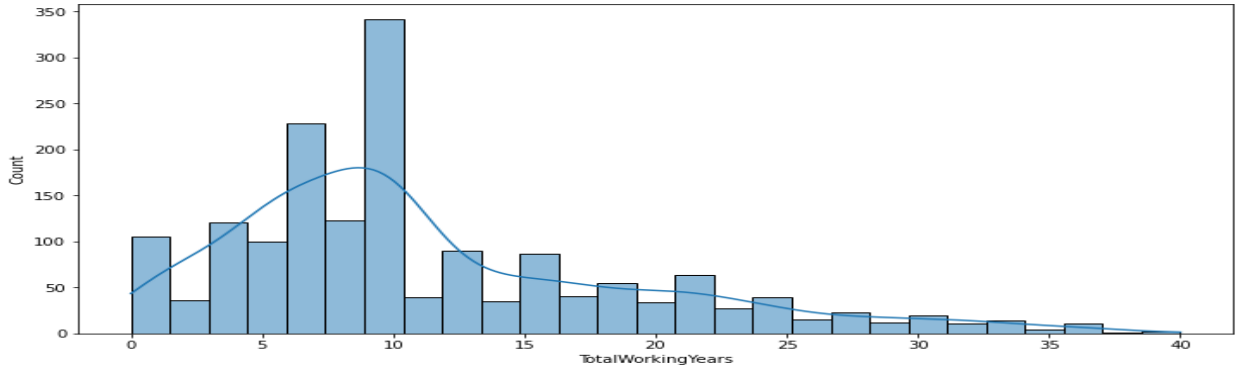

Count
EmployeeID
1e6

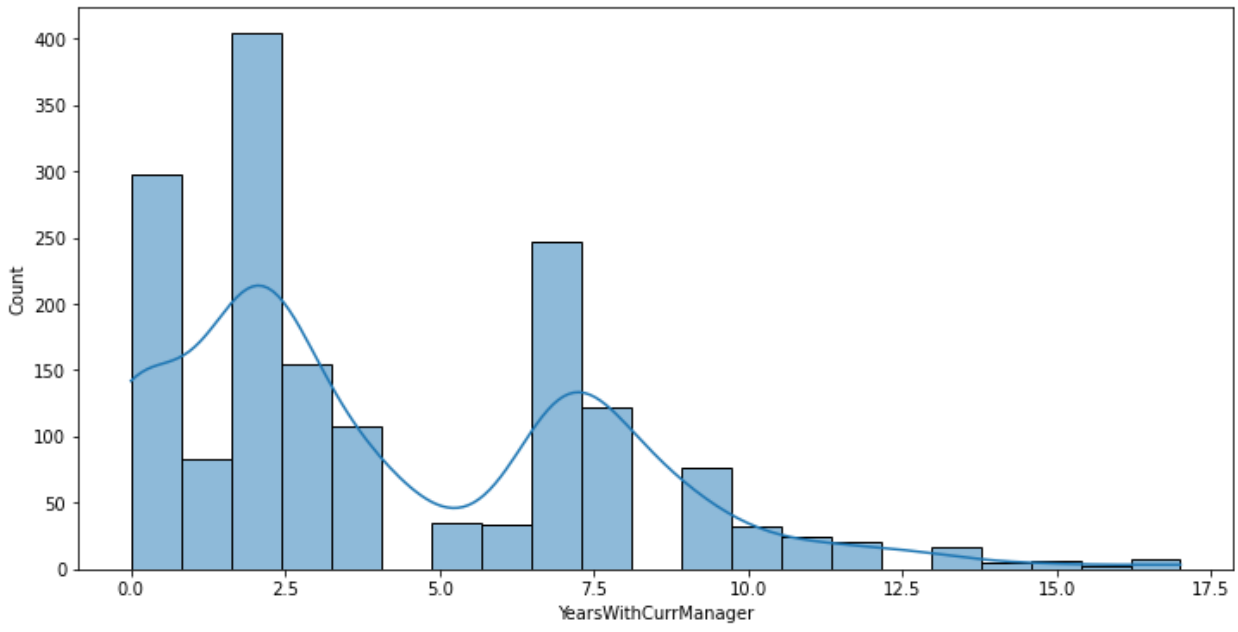

Count
DistanceFromHome

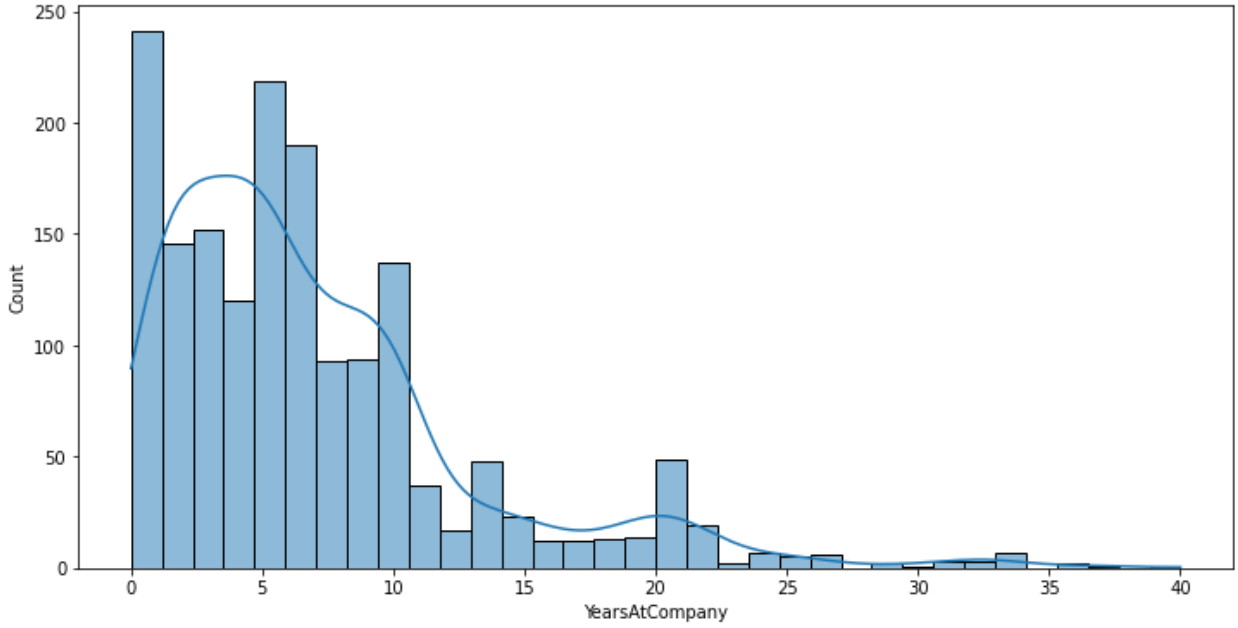

Count
DailyRate
0
20
40
60
80
100
120
140
160
200
400
600
800
1000
1200
1400

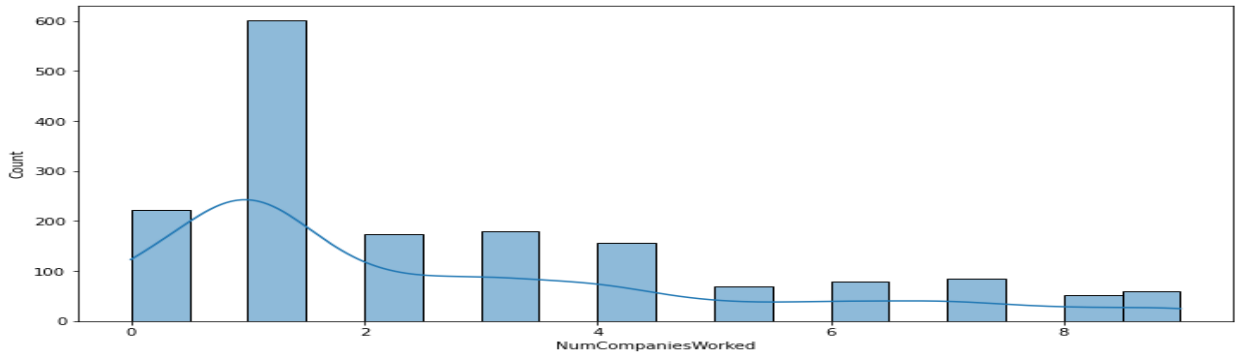

Count
MonthlyIncome
0
50
100
150
200
250
300
350
2500
5000
7500
10000
12500
15000
17500
20000

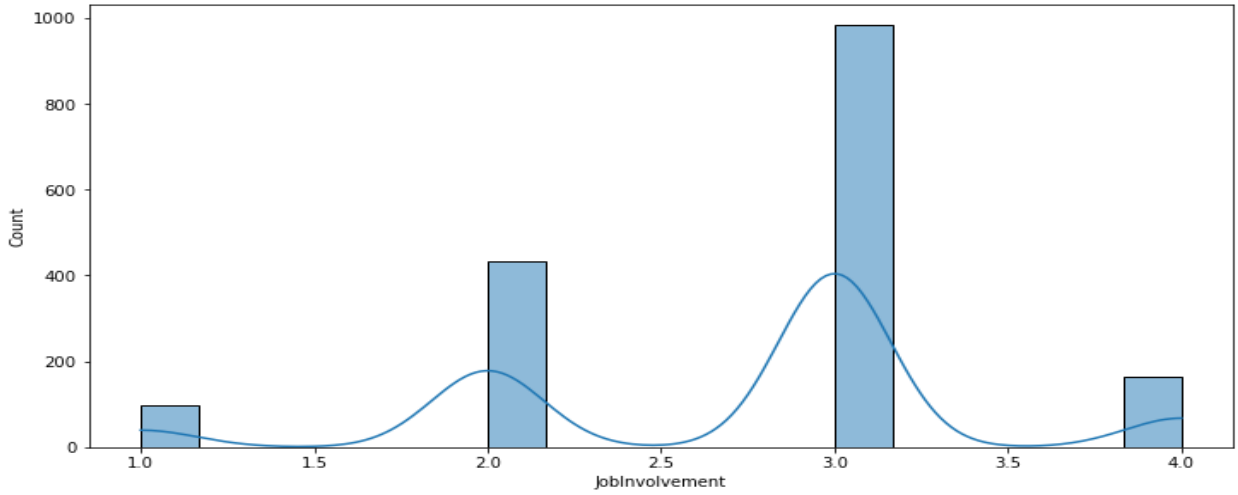

Count
PercentSalaryHike

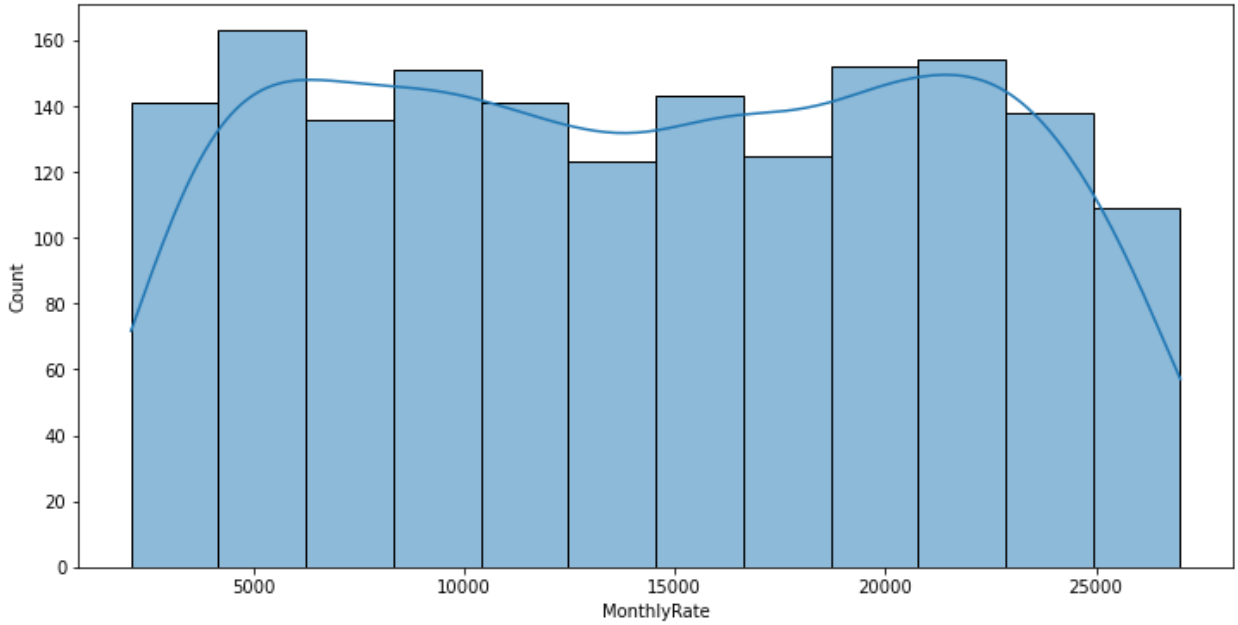

Count
TotalWorkingYears

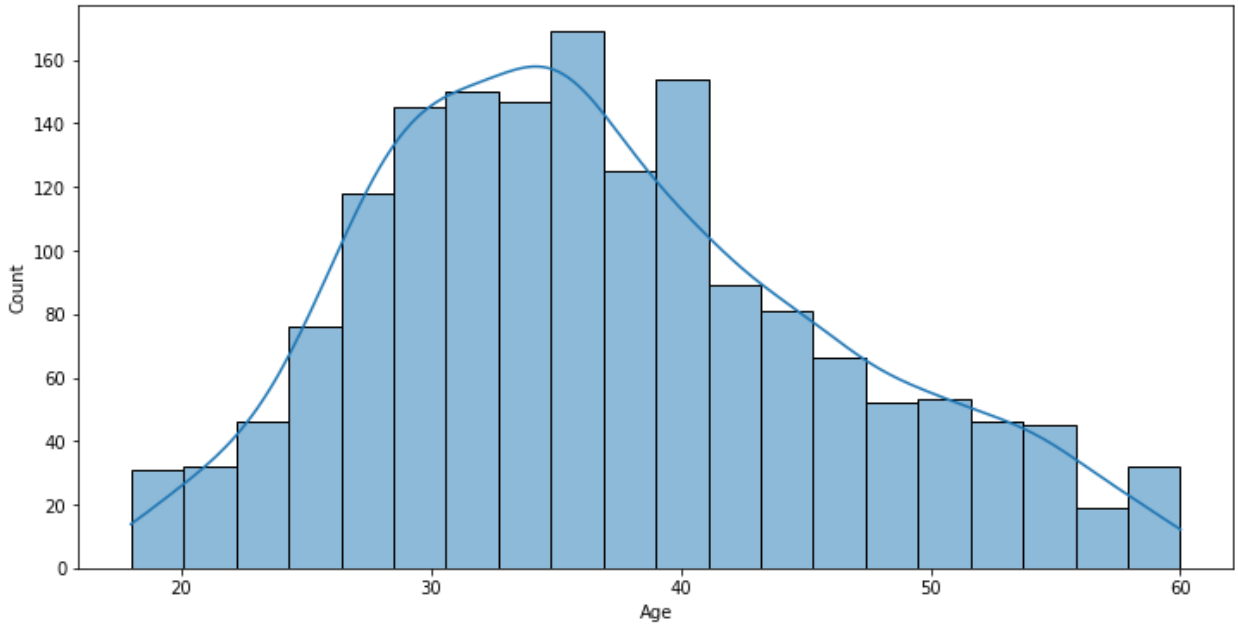

Count
YearsWithCurrManager


Count
YearsAtCompany
0
50
100
150
200
250
0
5
10
15
20
25
30
35
40


Count
NumCompaniesWorked
0
100
200
300
400
500
600
0
2
4
6
8

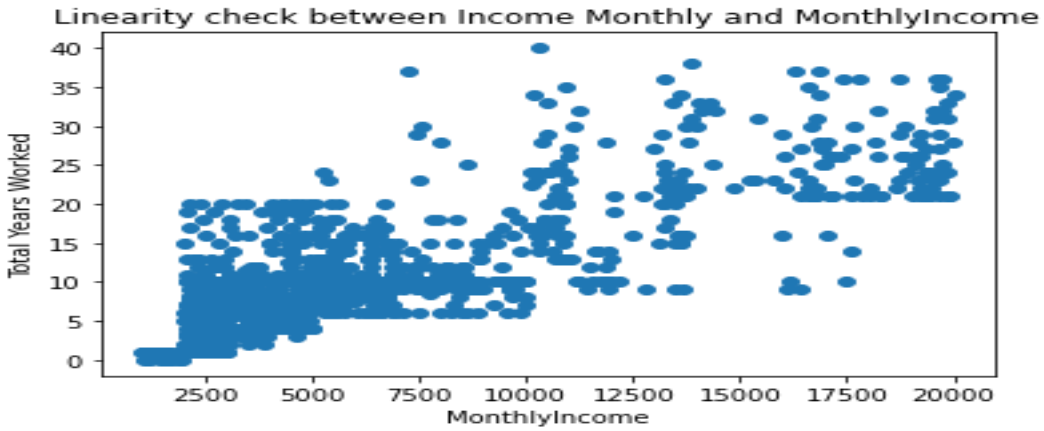

1000
800
600
400
200
0
Count
1.0
1.5
2.0
2.5
3.0
3.5
4.0
JobInvolvement

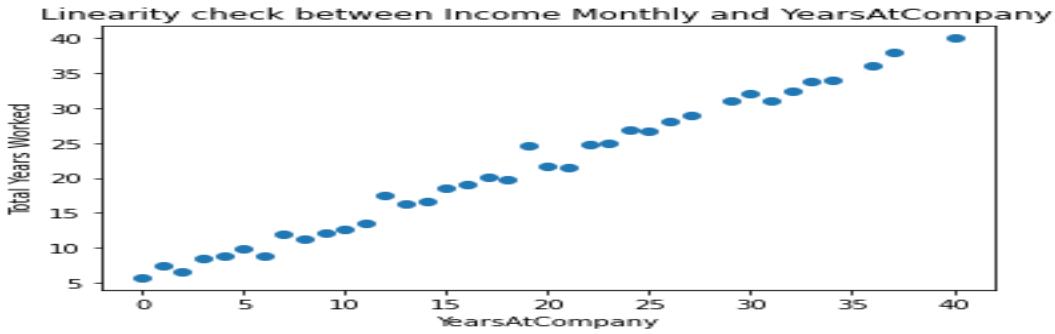

160
140
120
100
80
60
40
20
0
Count
5000
10000
15000
20000
25000
MonthlyRate

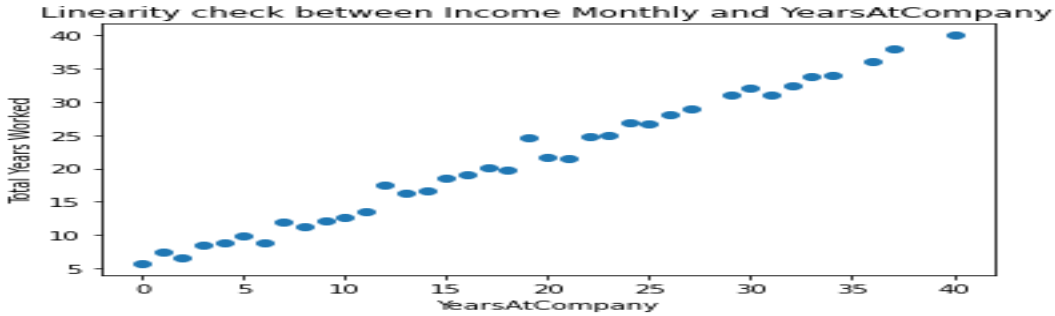


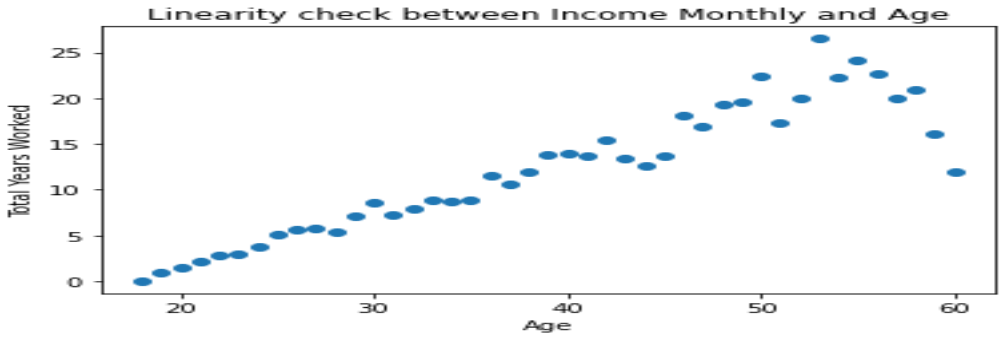
Milestone 3

The analysis found that the Total Years worked had a high positive correlation with the following: Job Level, Monthly Income, and Years at Company. This means that for the number of years worked within a company, there is a high likelihood these individuals are paid well, and have spent a long while in a prestigious career within their company and are less likely to leave. There were a number of interesting interactions found within this analysis.

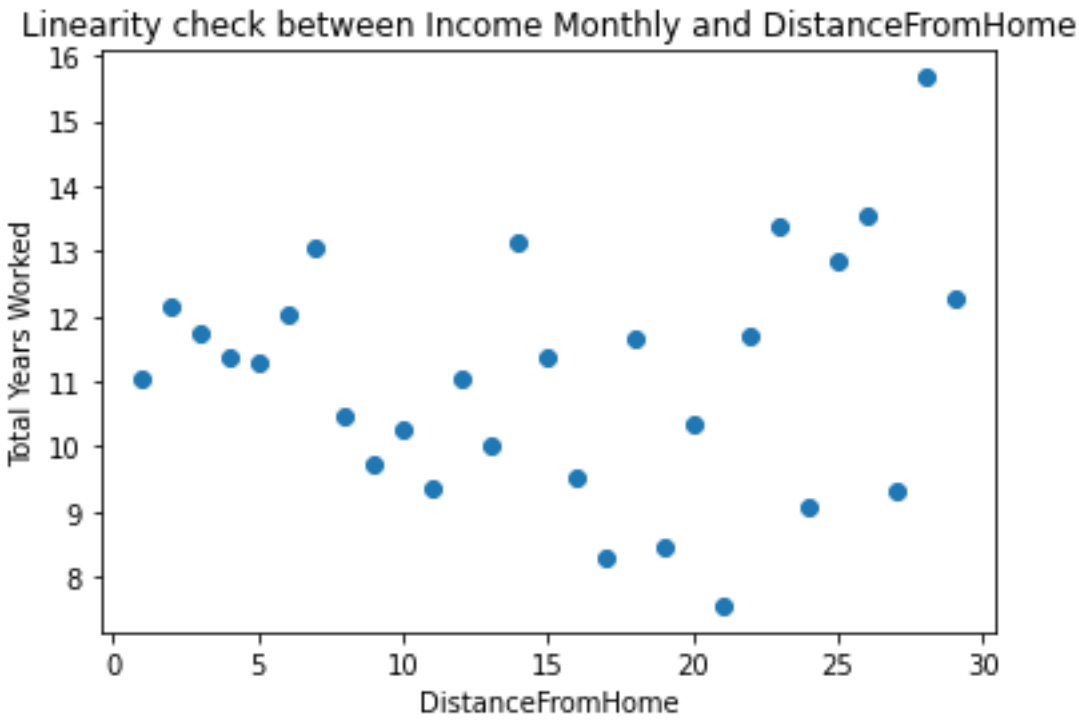








In looking at the MonthlyIncome versus TotalWorkingYears, the total years worked versus the monthly income was linear in value as even the lowest of incomes still increased more than the previous year worked. In examining Age versus TotalWorkingYears, an interesting observation was noted, at the age of 50-55 where the total years worked versus age significantly drops, indicating that the median age for people in this dataset leaving the workforce is around 55 while outliers do exist that make it past the working age of 60. Examining YearsAtCompany versus TotalWorkingYears found that the longer a person had total working years, it appeared that there is a linear climb of individuals working with the same company.



Another interesting interaction was DistanceFromHome versus TotalWorkingYears. This appears somewhat linear, but in two different ways, individuals who live closer work at the same place longer, and people that live farther away tend to leave sooner after 10 miles of distance, while at the same time an equal amount of people after 10 miles will continue to work for the same institution. Finally, an analysis of the YearsInCurrentRole versus TotalWorkingYears, found that those who work the same role tend to stay in the same institution, so a linear relationship is created.

Conclusion

Of all the variables, once a person was 42 and older, saw the smallest attrition rates while 26–35-year-old workers often left their company. The model was able to identify with an accuracy of 89%, so I think it is ready to be deployed. To keep people in the work force, my recommendations would be to give incentives for those who live closer to work, along with

people within certain age brackets, more reasons to stay, weather that be faster tracking towards career advancement, or even more competitive wages. Shuffling employees around departments also had a large impact on attrition rates, so keeping the same individuals in the same role seems to also keep people from leaving for different opportunities.