**Attempt to Derive a Logistic Regression Equation and a Decision Forest Model to Predict Employee Attrition**

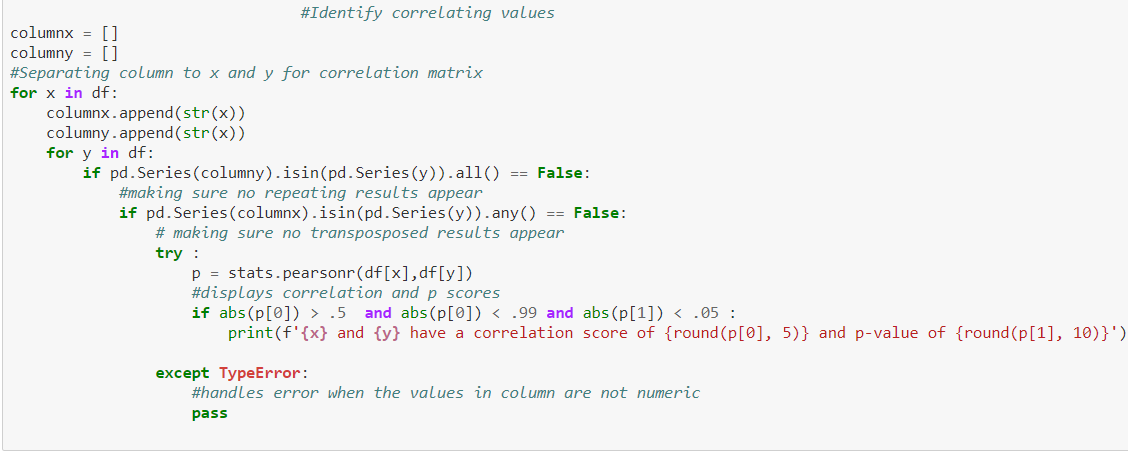
**Introduction**

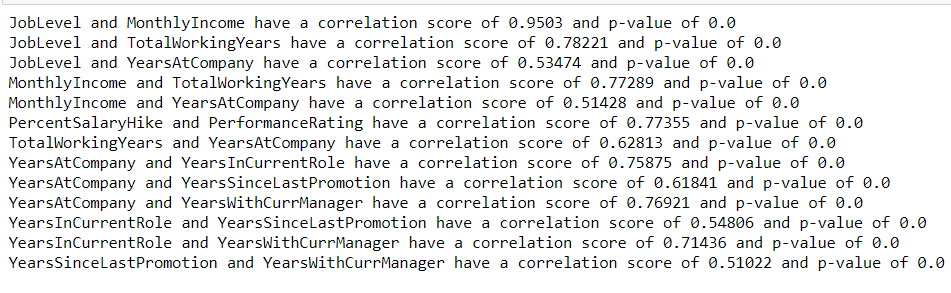
For the analysis project, I obtained data from a company’s Human Resources Unit, which lists all current and previously employed employees throughout the company’s history. The data is claimed to be fictional by IBM,(the source), but I strongly believe it is disclosed as fictional for legal reasons, since IBM does have a large consulting unit, and having access to such data for IBM is highly possible.

The data set consists of a total of 1470 unique employees. 237 of 1470 have left the company, and the rest are still employed. The data for each employee consists of 35 different variables which range from basic demographic data (age, race, gender, etc.), to internal data (department, years with company, evaluation scores, etc.) to personal historical (number of previous employers, total working years)

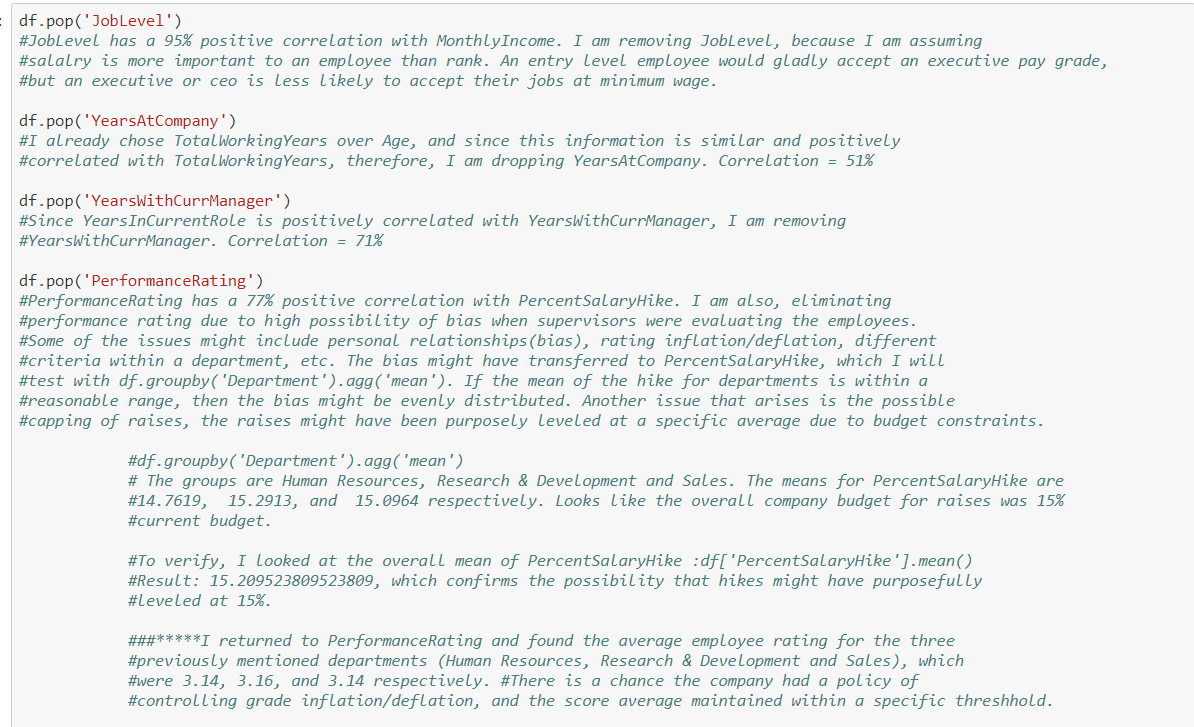
By utilizing SPSS and Python, I will attempt to derive a logistic regression equation (Python and SPSS) and decision forest model (Python) in order to predict the likelihood of existing employees leaving the company.

**Exploratory Analysis (Using Python)**

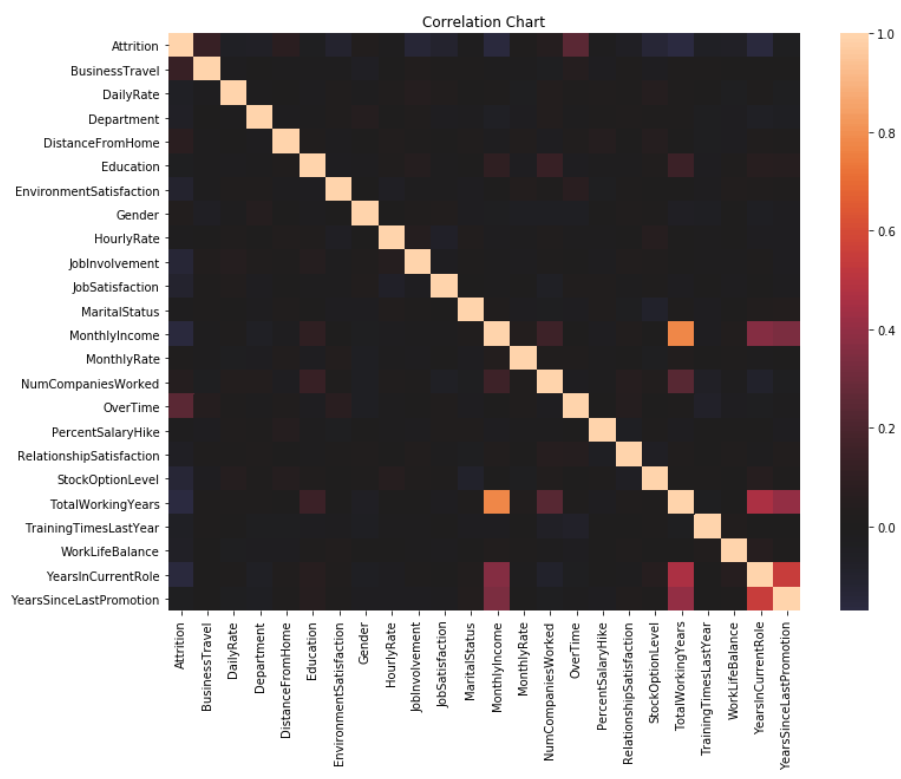
I started by searching for any correlated categories and removing them. The removal was chosen with no specific approach, 

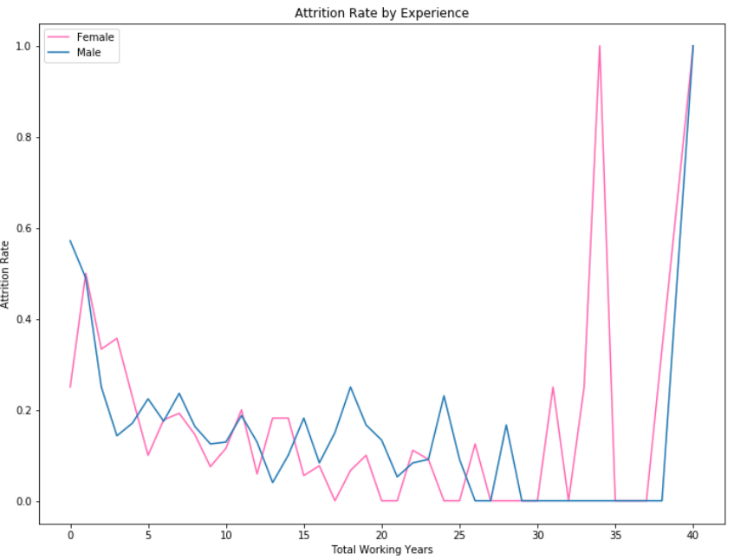
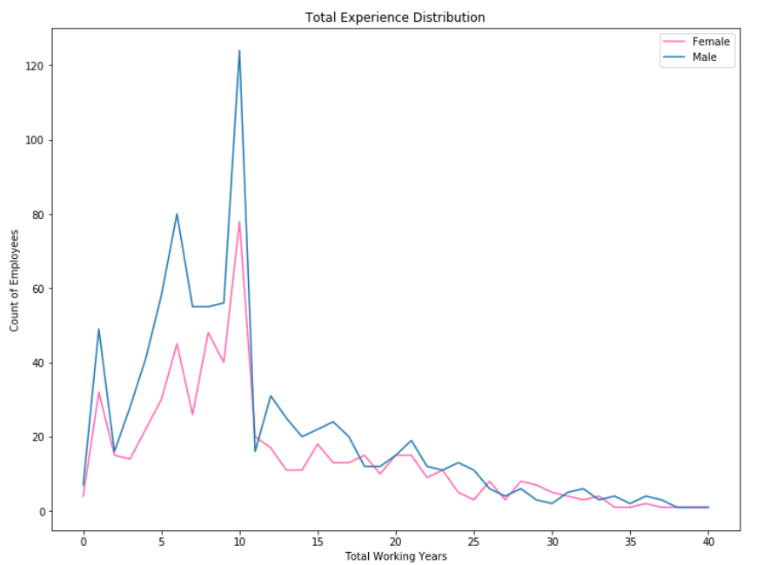
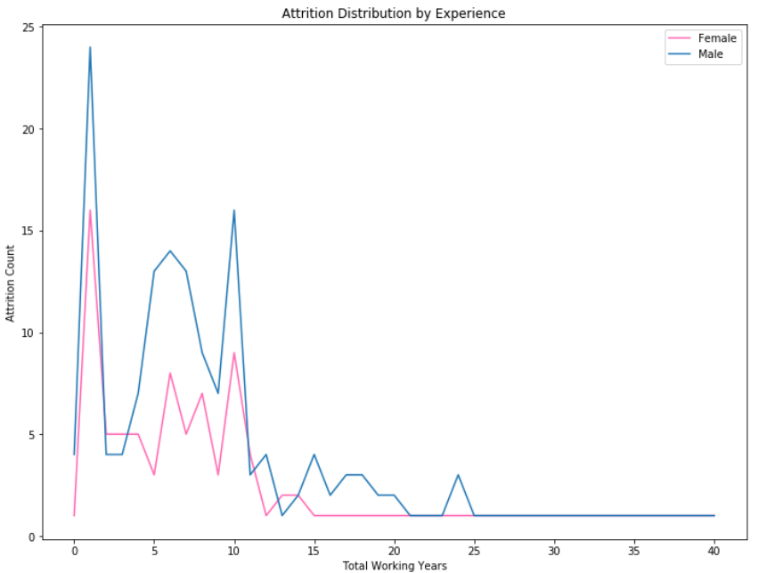
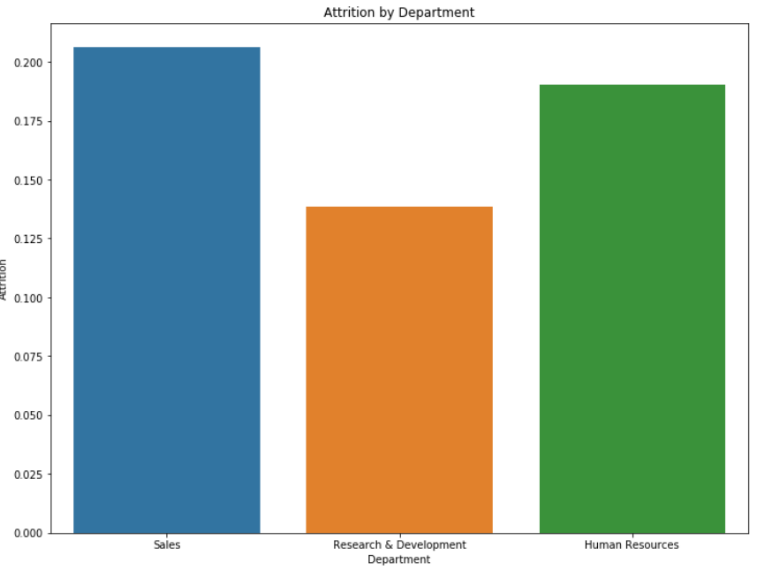
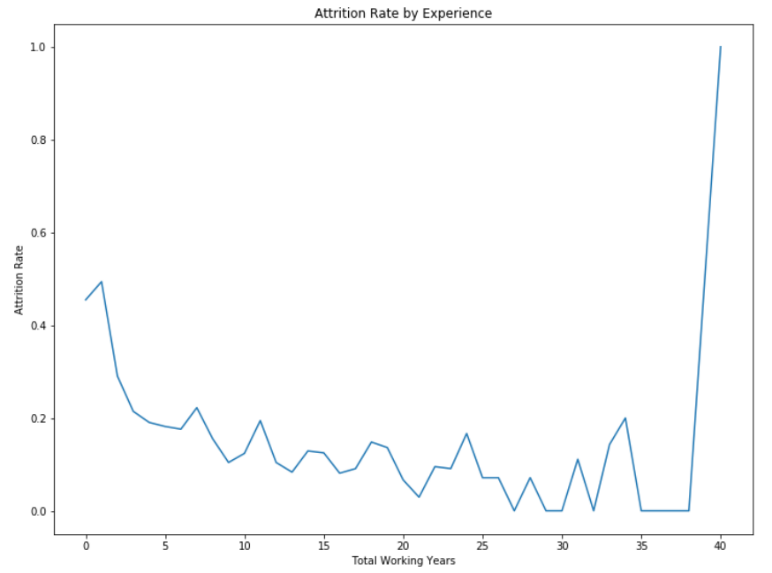
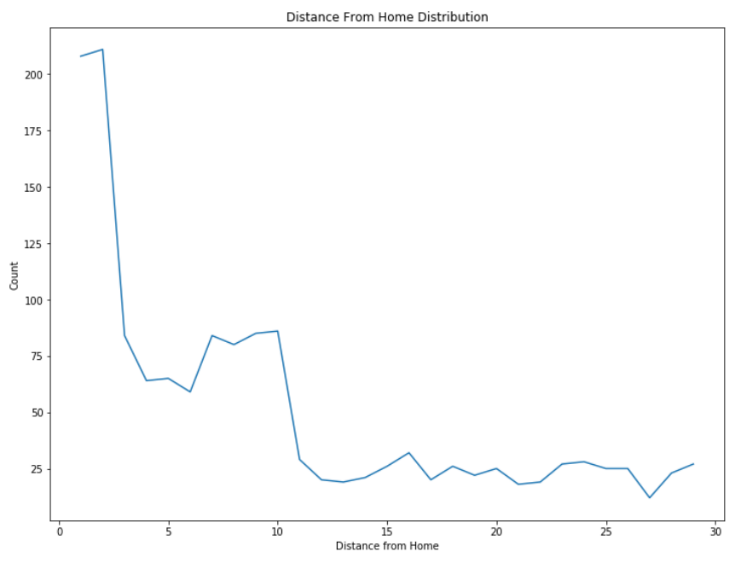
Which yielded the following results:

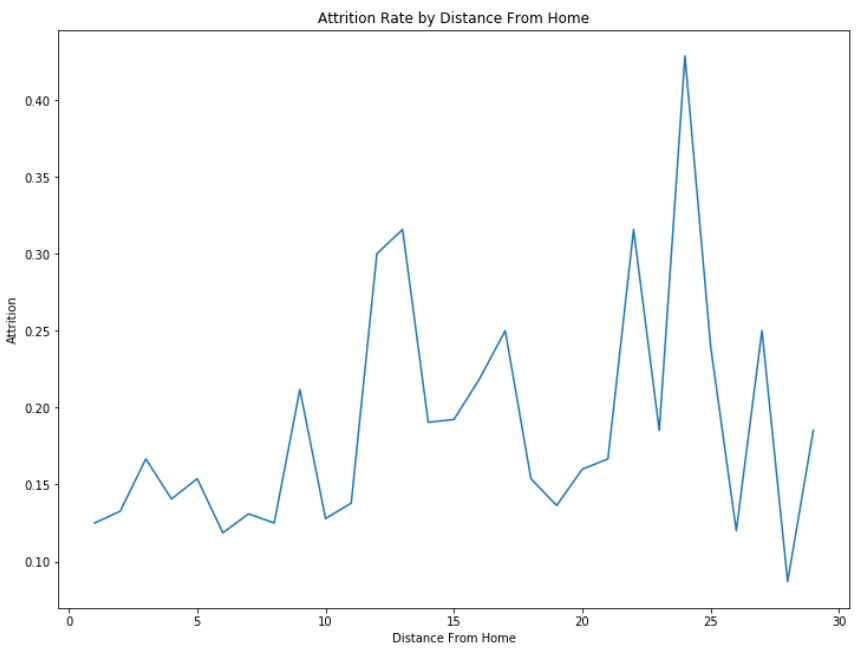
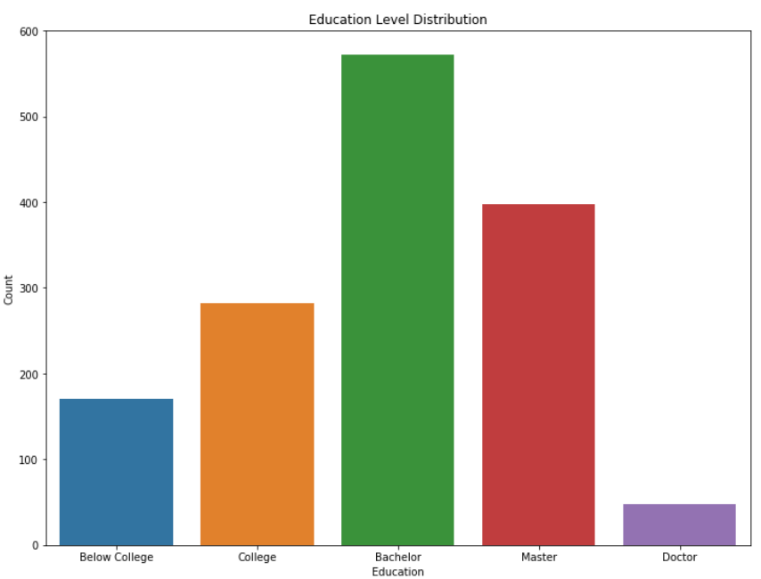
Removal of correlating variables:

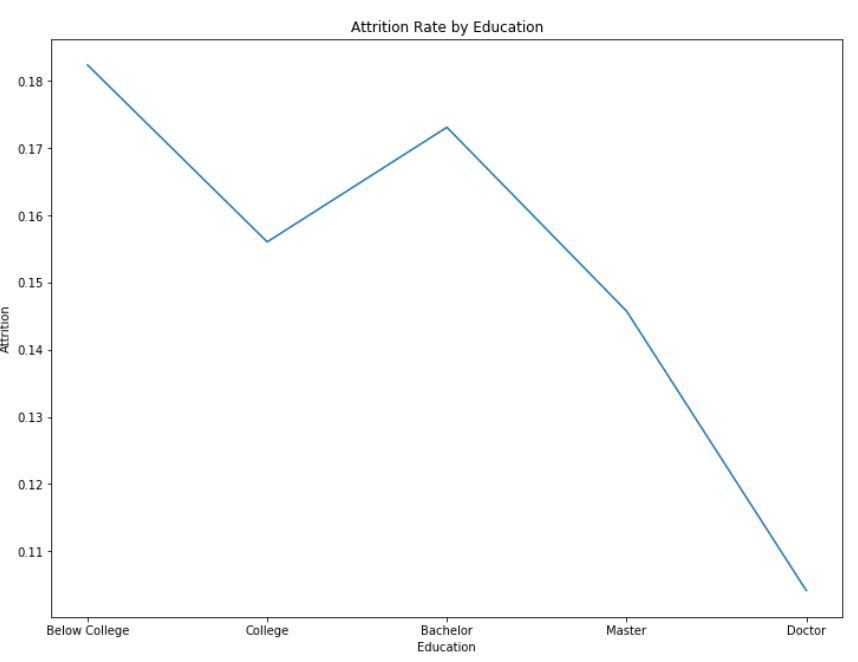
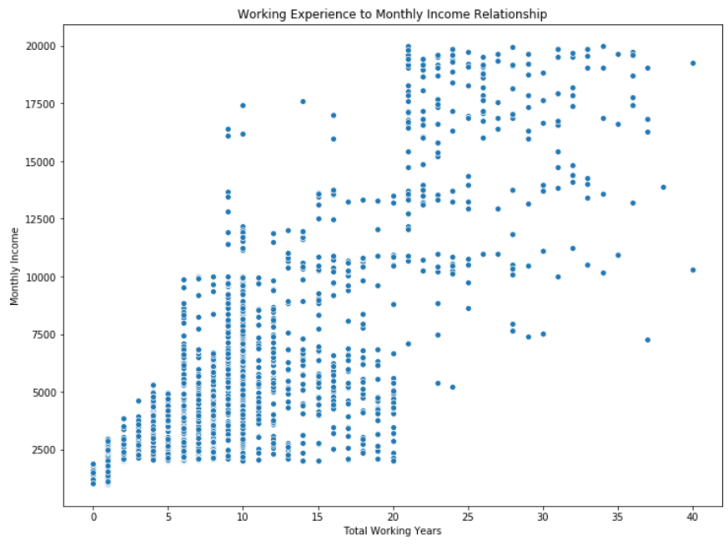


Now for some charts:

Correlation Heat Map, just to make sure nothing else stands out:



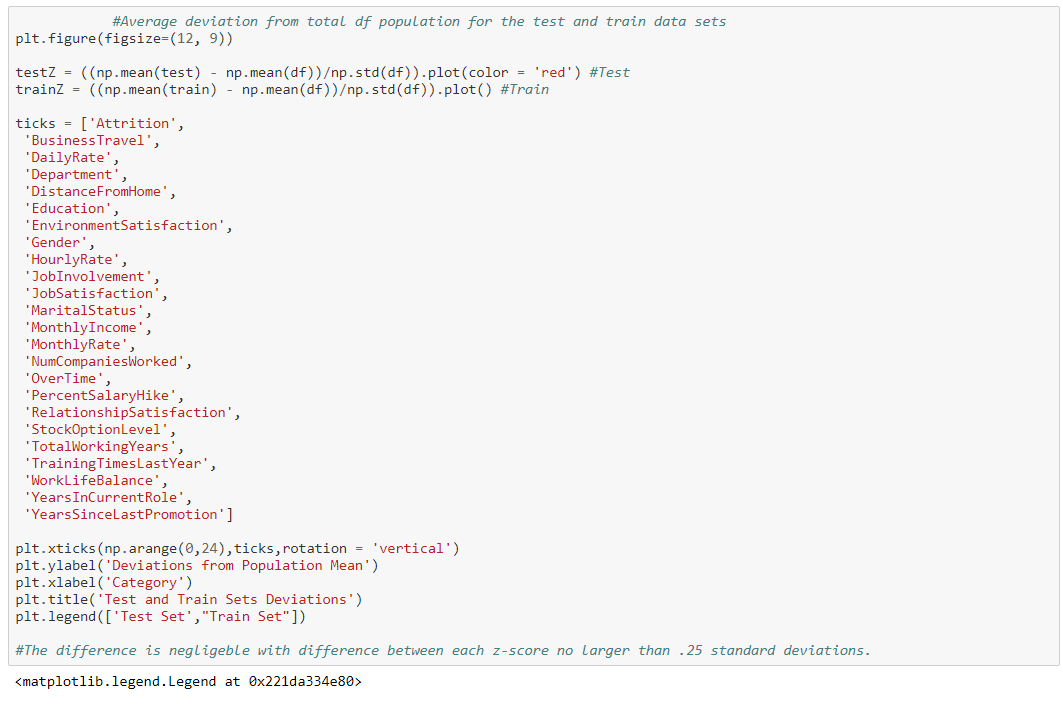
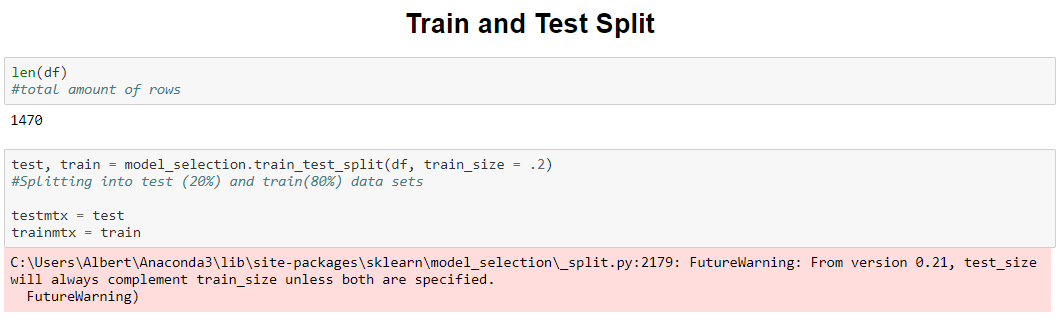




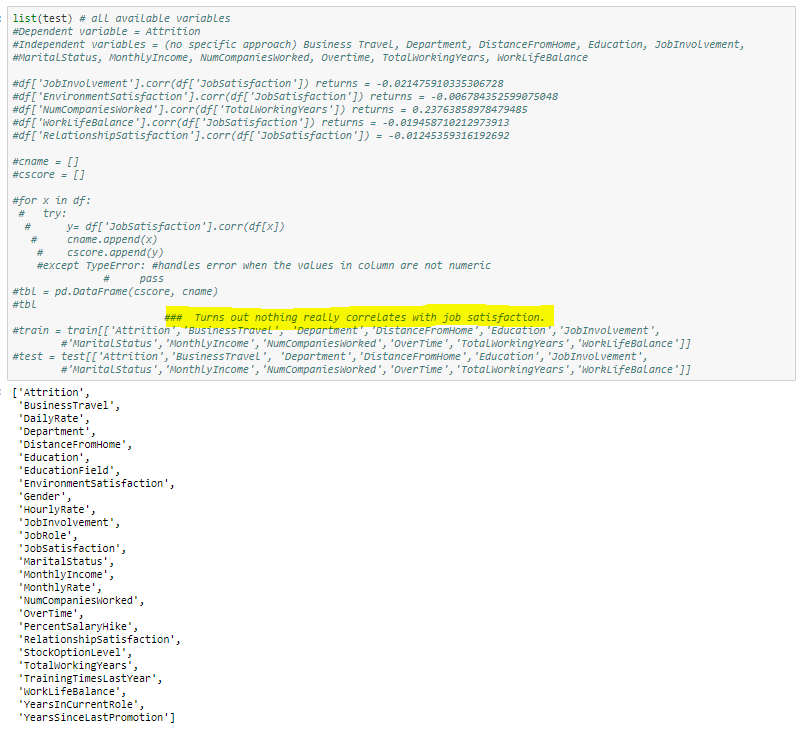
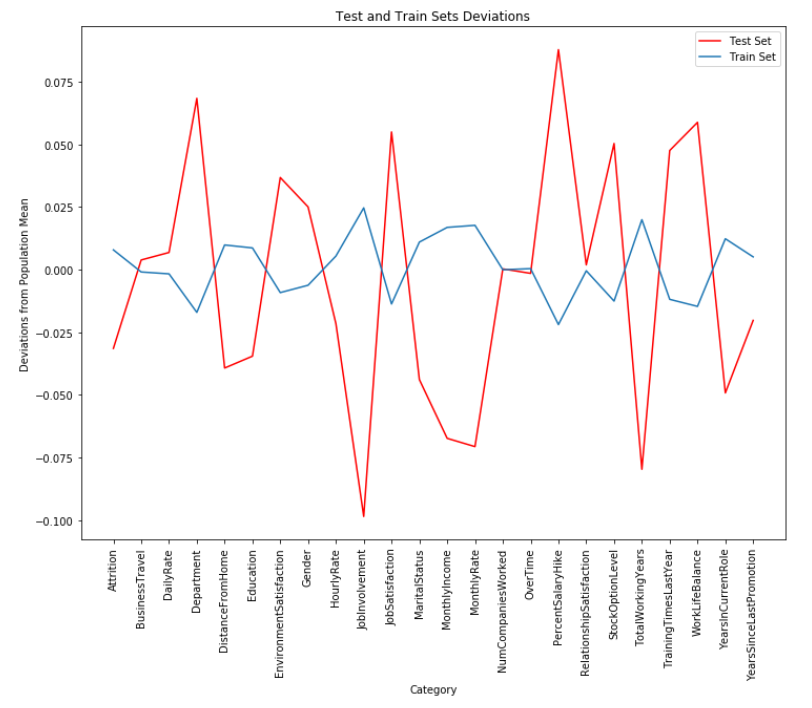
Some observations(in order of charts):

* We can see that attrition decreases between the working years amount of 0 through 35.
* After 35 years of work, we see a spike in attrition, probably due to retirement.
* We see Female attrition rates spike about 5 years earlier than Male. Note that average male age in the database is 36.65 and average Female age is 37.3.
* We see that the proportion of men to women in the organization is relatively close to 50%.(Men are 60% of population and women are 40%)
* We see that the attrition is proportional to the gender ratio of the company. From the individuals that have left the organization, around 63% were men and around 37% were women.
* Total Working Years distribution of employees, looks like its skewed to the right (positive skewness) Mode<Median<Mean
* Research and development has a noticeably lower rate of attrition than Sales and Human Resources
* We can see higher attrition rates as the distance of a commute from home increases.

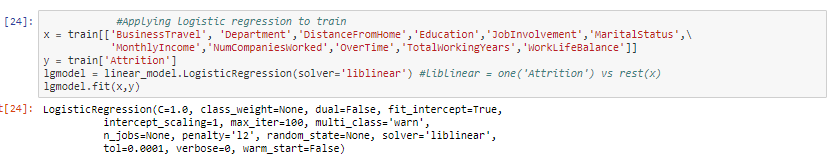
**Logistic Regression (Python)**

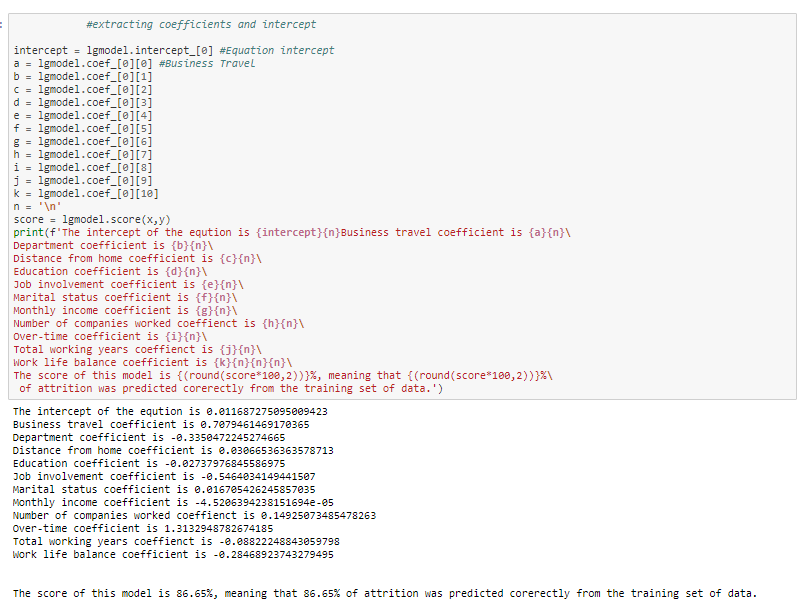


(Actual chart on the next page)

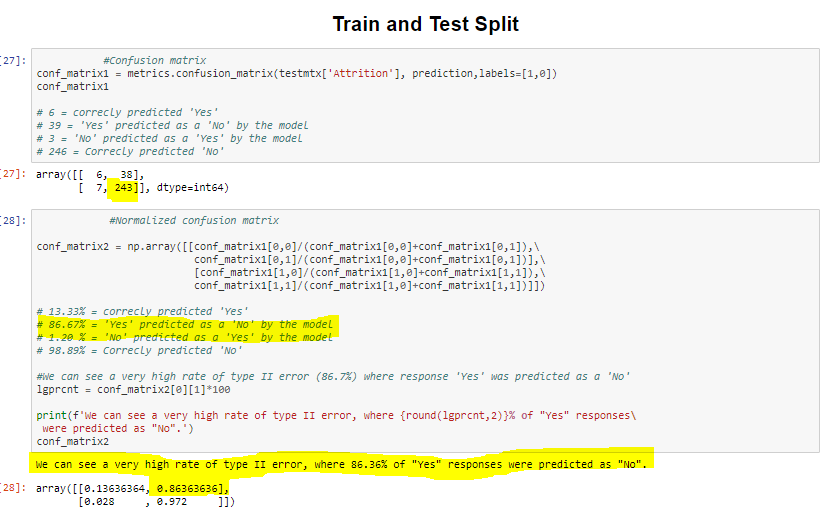
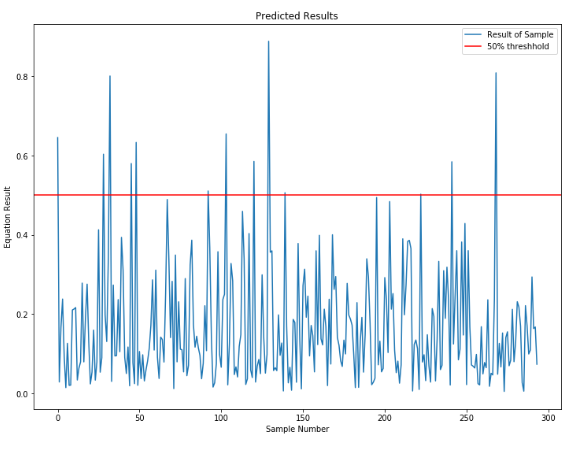


I ran a separate test to just see what correlates with job satisfaction, turns out there’s no panacea for job satisfaction here.

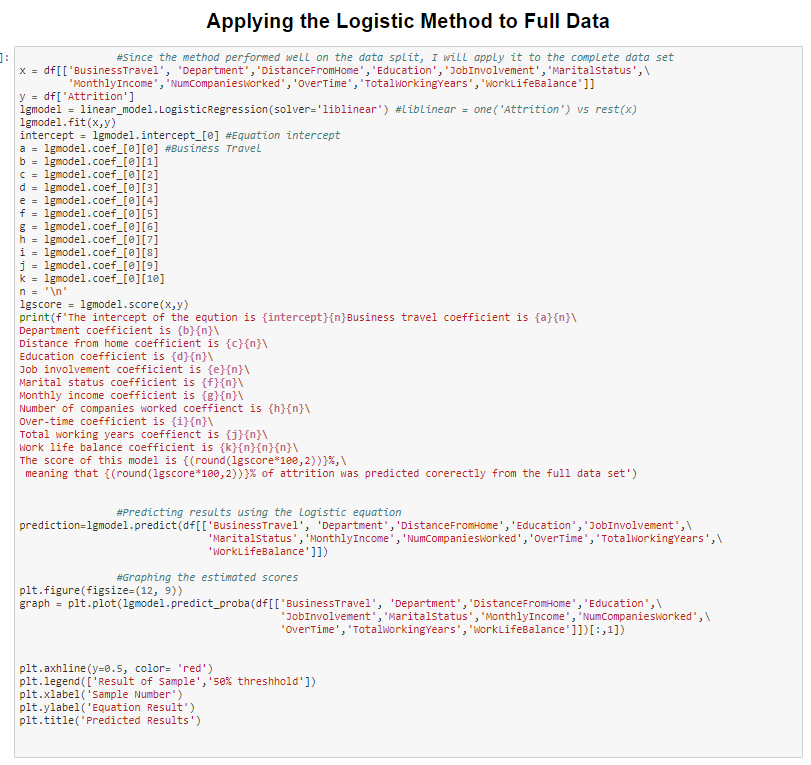


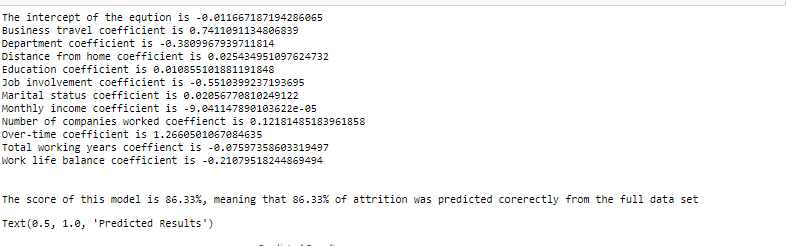


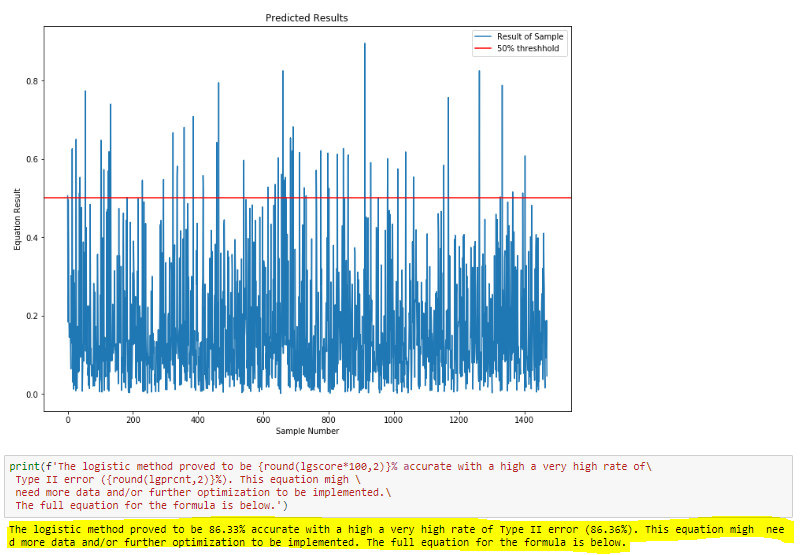
(Chart on the next page, cool but not very useful)

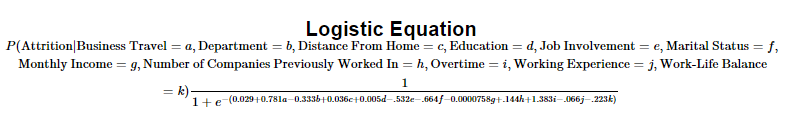


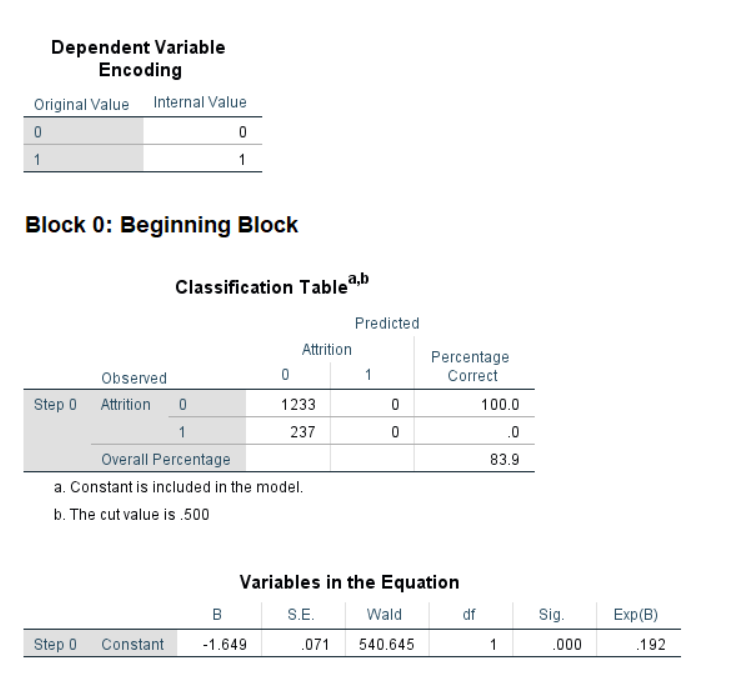
The train-test model was successful 86.36% of the time, but had a very high rate of Type II error for attrition (Predicting an employee who left as an employee who is still working)

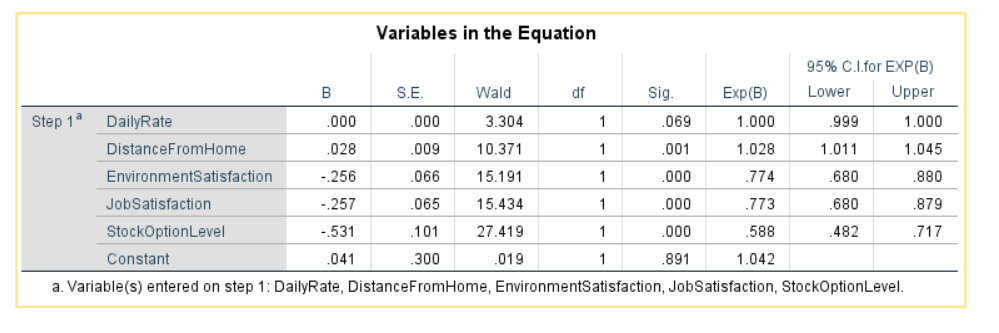
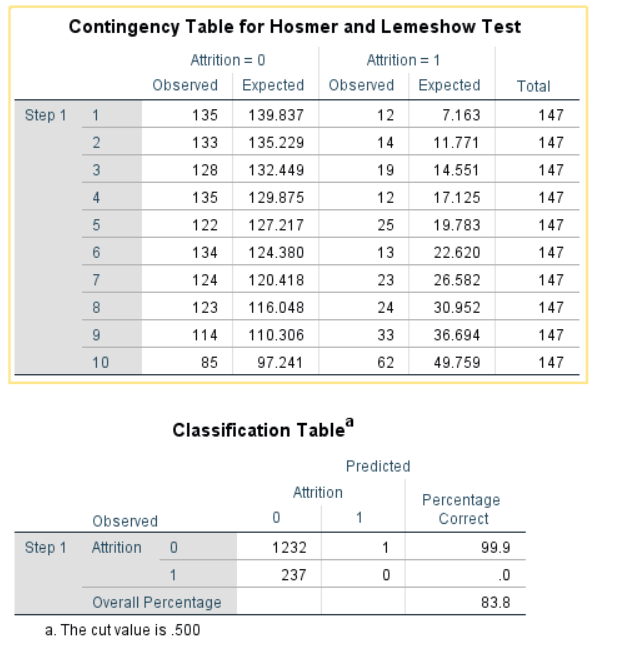
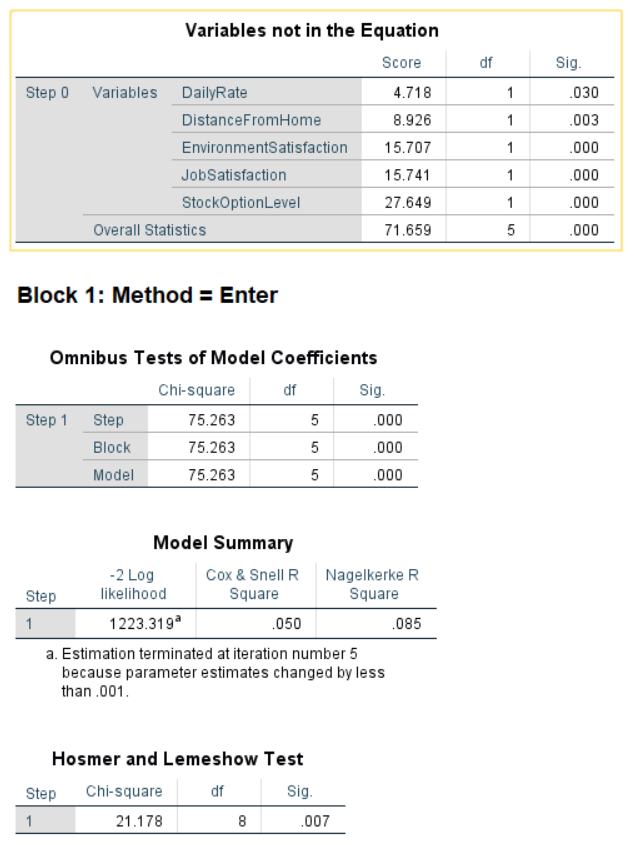








**Logistic Equation in SPSS**



**Interpretation of SPSS Logistic Regression Output**

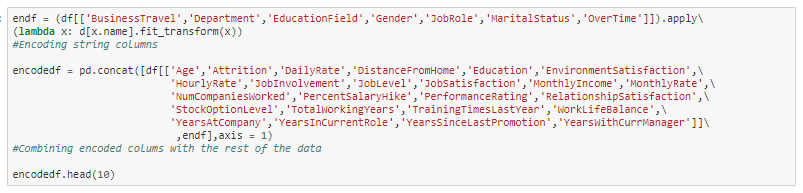
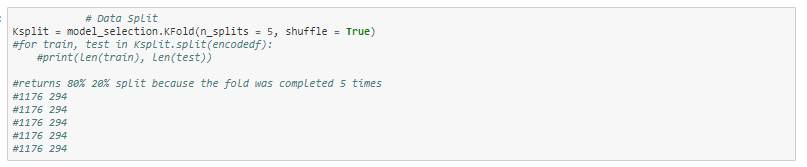
**Block 0:** SPSS presents a baseline, stating that if it predicted all employees to not have left the company, it would be correct 83.9% of the time. (Compared to the python model, python only has a 2.34% advantage, disappointing).

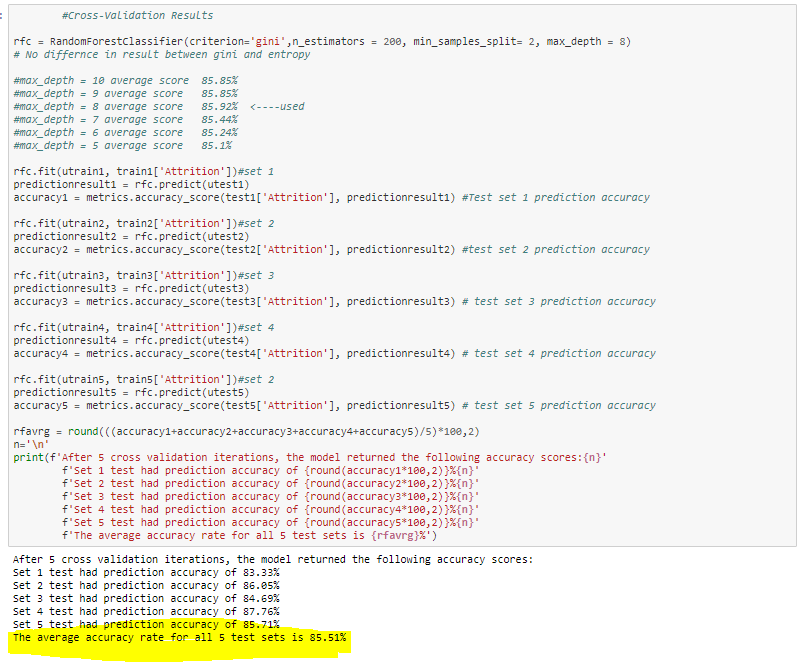
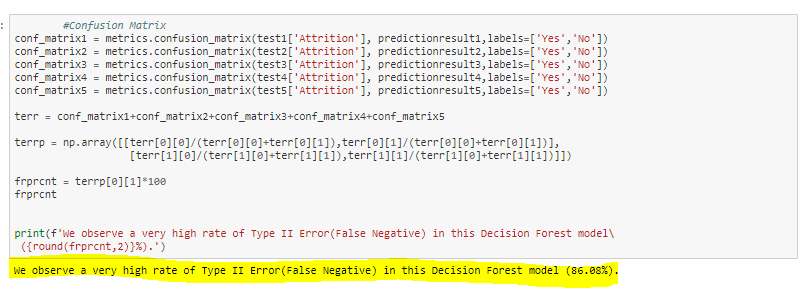
**Block 1:**

* **Omnibus Tests of Model Coefficients:** Present a significant improvement in the model, p<.0005. Which mean that the model is better method of prediction than the baseline at Block 0. The chi squared value is 75.262 with 5 degrees of freedom.
* **Hosmer-Lemeshow Test:** this test indicates poor fit of the model if the p value is less than .05. Unfortunately, my model scores a p value of .007, and therefore, falls under a poor fit category.
* **Model Summary:** The failure of my model doesn’t just stop at the Hosmer-Lemeshow Test. The Cox & Snell and Negelkerke values are .05 and .085 respectively. Which means that the model explains less than 10% of the variability by my chosen set of variables(DailyRate, DistanceFromHome, EnviromentSatisfaction, JobSatisfaction, StockOptionLevel)
* **Variables in the equation:** All chosen variables were significant, except for the “DailyRate”**.**

Both models, from Python and SPSS, did not show any impressive results. This is most likely due to lack of data, and lack of time that I had for optimization. At this point I will go into the Decision Forests Model, which hopefully will yield more significant results.

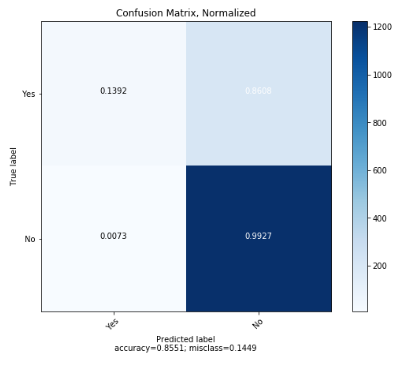
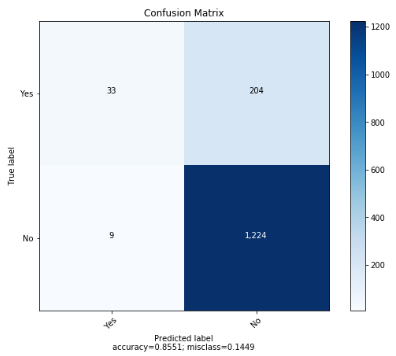
**Random Forests (Python)**

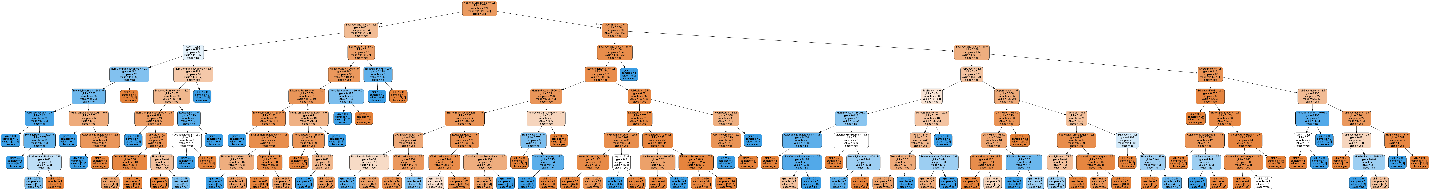
  

**Confusion Matrix**

Code obtained from (<https://www.kaggle.com/grfiv4/plot-a-confusion-matrix>)



Again, we observe a very high rate of Type II error, Identifying a positive value (positive attrition) as non-attrition.



Random forests have identified “YearsInCurrentRole” as the most import factor in identifying attrition. The full chart in .PNG format will be attached separately for this project. Unfortunately, this model is not valid as the Python and SPSS logistic regressions because of a high Type II error rate.

**Conclusion**

All three of the attempted methods in this project have failed to deliver significant results. The failure is attributed to the high rates of Type II errors, which are most likely caused by lack of data. I suspect, that even if I did spend additional hours on identifying the significant variables, optimizing features, parameters, codes and approaches for the models, there will still be significant Type II error rates.

In addition, if any of the models turned out to be extremely accurate in predicting attrition, because this model is based on demographics, releasing such a model would be unethical, and probably illegal to apply in the real world (within US), since this would probably lead to indirect discrimination.