HR Attrition Analysis

CAN WE PREDICT IT?

Background

Many employers today are taking great strides to identify what factors are most important to employees. This analysis examined a number of those factors and identified which ones are most relevant to prevent attrition.

Importance

What is employee attrition?

The loss of employees due to life events such as retirement, resignation initiated by employee, elimination of a position or other similar event.

Attrition for reasons initiated by employee is the category to be analyzed.

How can employers prevent attrition?

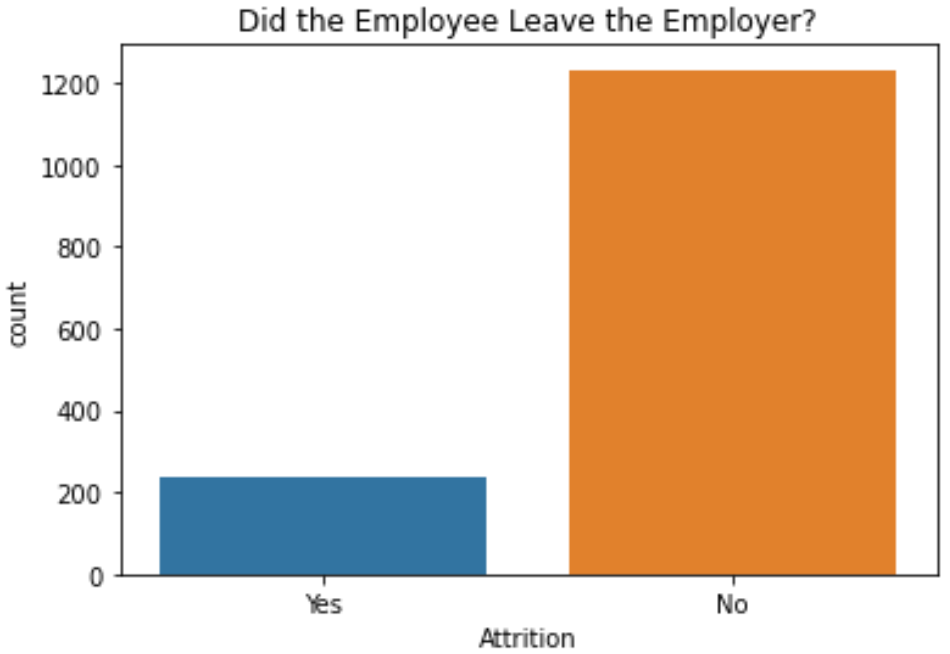
1. Encourage Flexibility
2. Give Incentives
3. Meet individual needs

The Process

This presentation aims to give insights into an HR dataset and develop a machine learning algorithm to predict attrition.

- This dataset was retrieved from Kaggle and is from the IBM HR department and contains details about employee attrition and performance. This is a fictional dataset created by IBM data scientists.
- It contains 1470 observations and contains 35 variables relating to each observation
  - Examples of the variables include age, attrition, pay data, education, job roles, performance, and years of employment
- Some questions that were developed and refined include:
  - *What are the most important factors for attrition?*
  - *Does increased salary reduce attrition?*
  - *If an employee lives further from the office, does that impact if they stayed on the job?*
  - *Does information such as marital status, gender, or educational level contribute to attrition?*

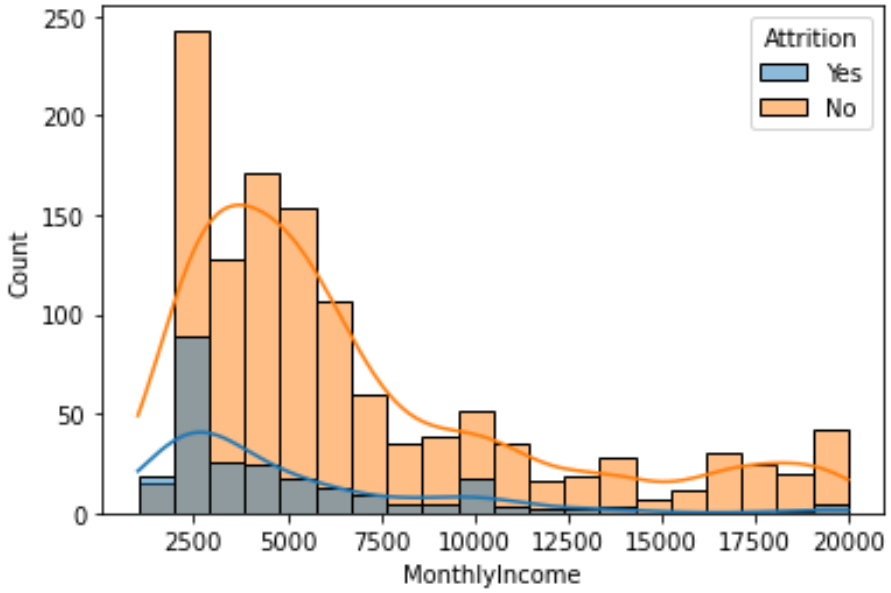
EDA Process and Results

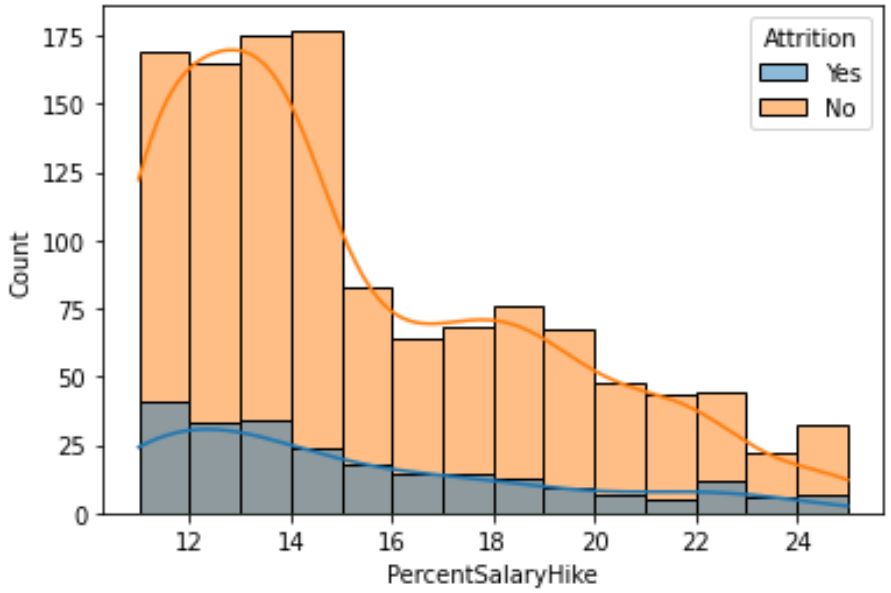


EDA - Income Drives Attrition!

Monthly Income proved to be one of the biggest factors that affect attrition.

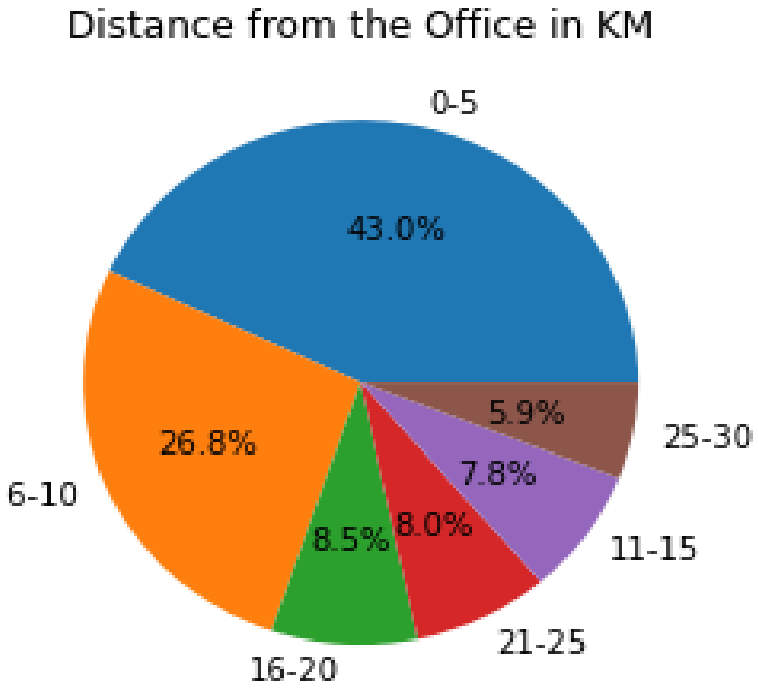
- *When employees are paid more, they are less likely to leave.*

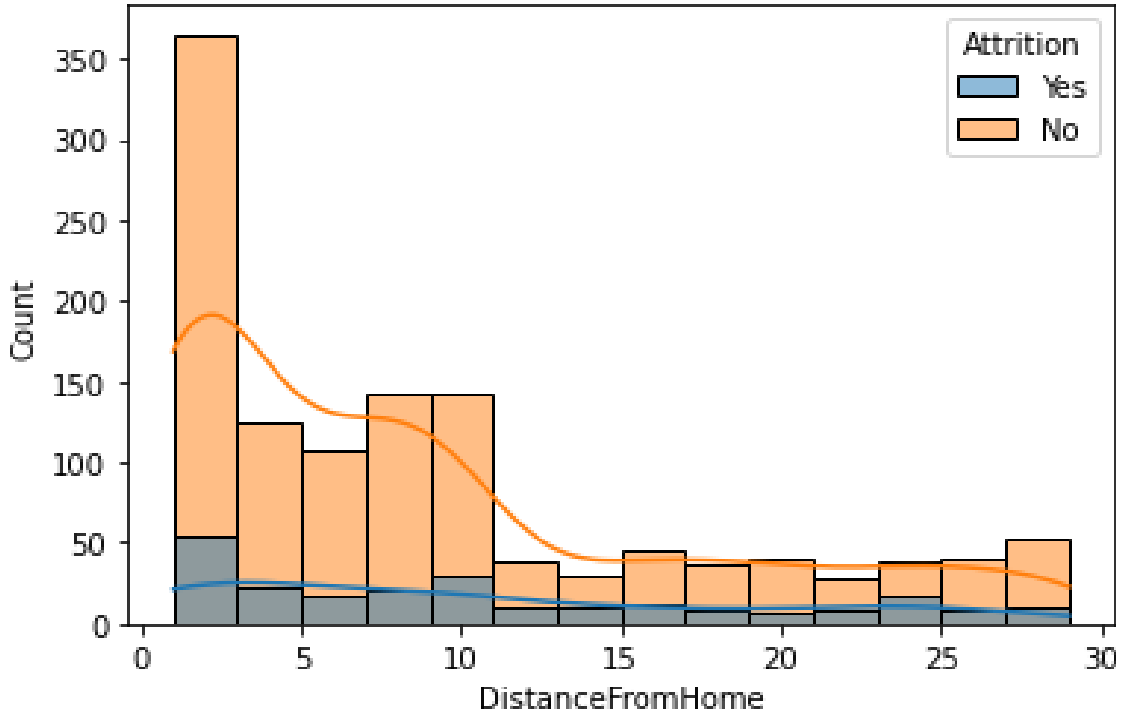




EDA – Distance does play a factor!

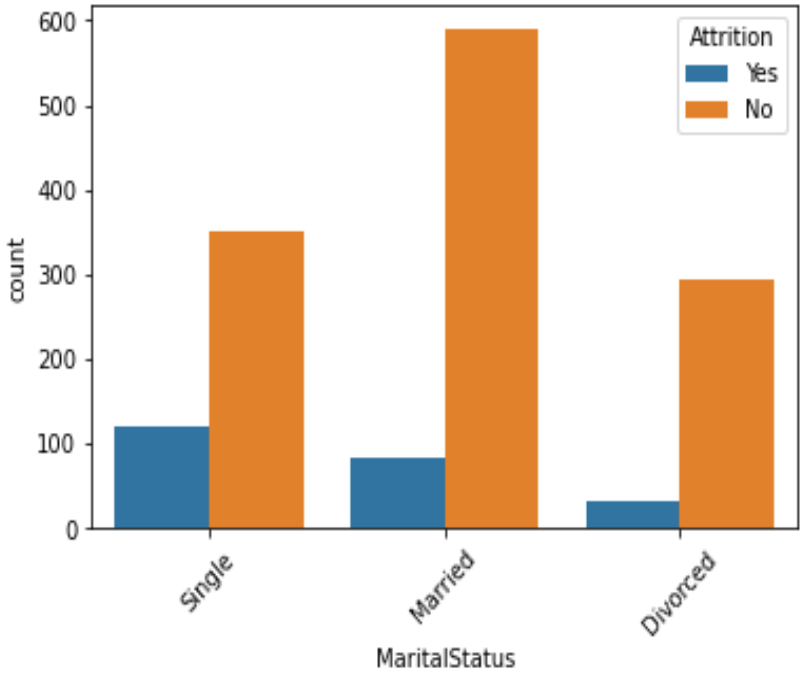
- *~70% of employees live within 10km from the office. The proportions of people that leave the job increase significantly if they live more than 10km from the office!*

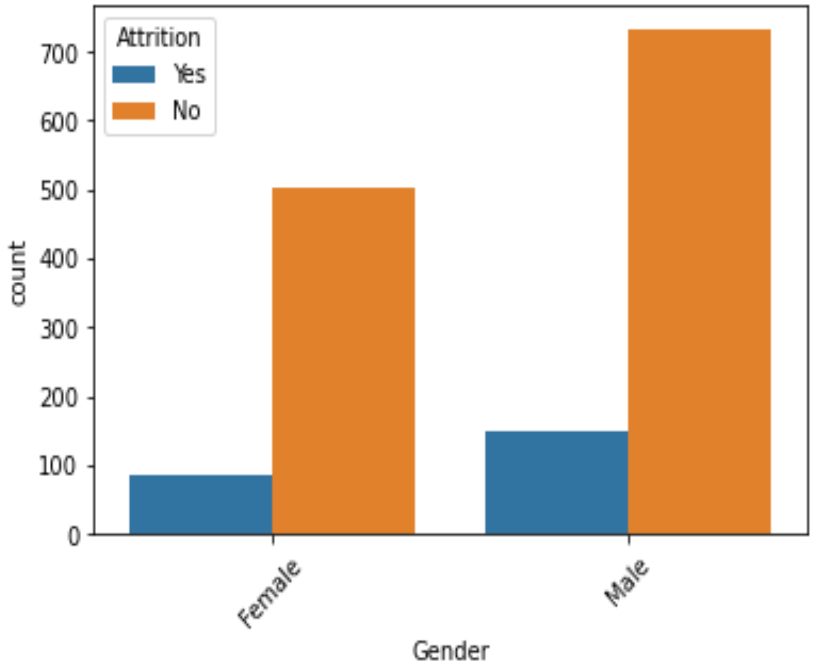


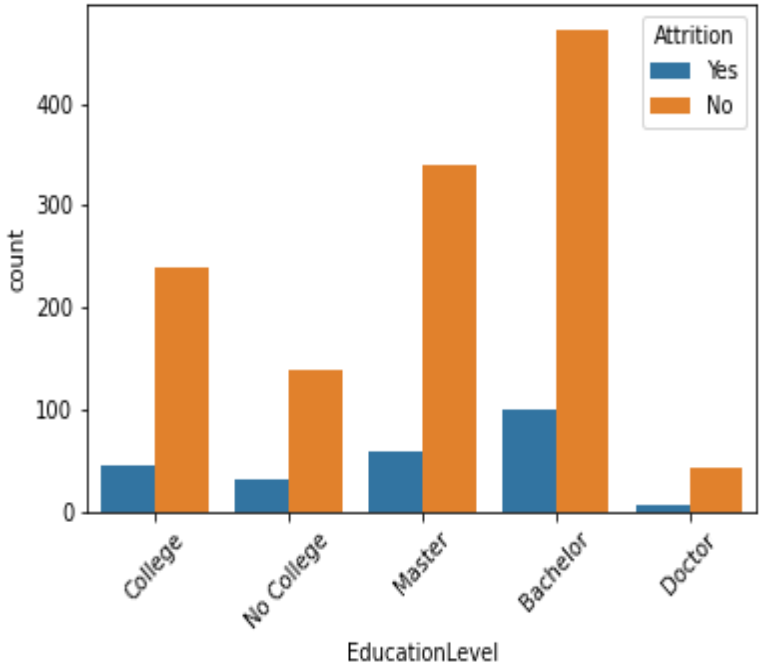


EDA – Marital Status MAY be a predictor

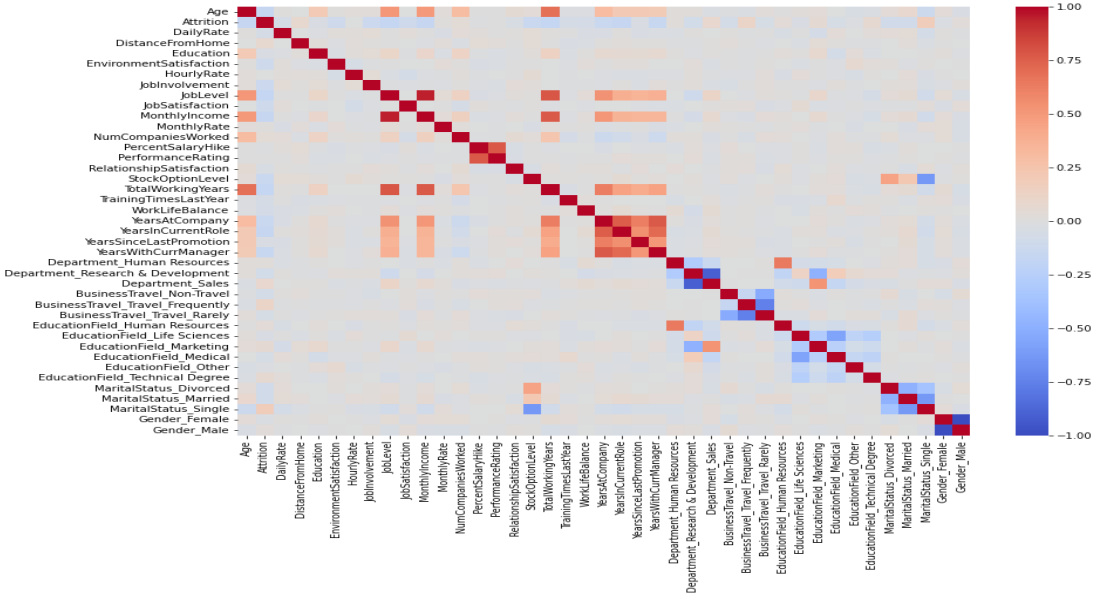
- *It appears that individuals that are single may have a higher rate of attrition compared to the other two marital status categories. Proportionally, gender and education do not appear to have correlations with attrition.*







Data Pre-Processing



Some of our features are highly correlated!

What is multicollinearity?

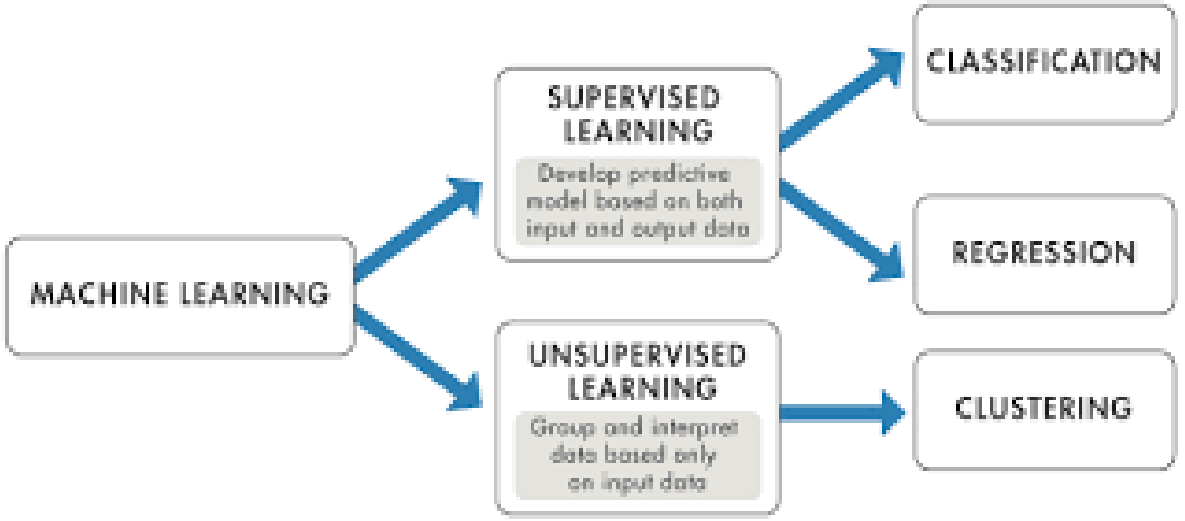
Why should it be dealt with prior to training the model?

Remove highly correlated features → Encoding → Scaling

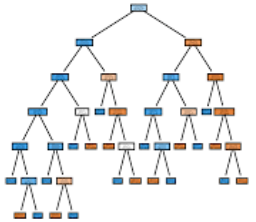
Data Pre-Processing Steps:

- Remove highly correlated variables
- Some of our variables are categorical
  - For this model, the pd.get_dummies was utilized for the categorical features
- The rest of the integer variables were scaled after the train/test split (85% / 25% respectively in this analysis)

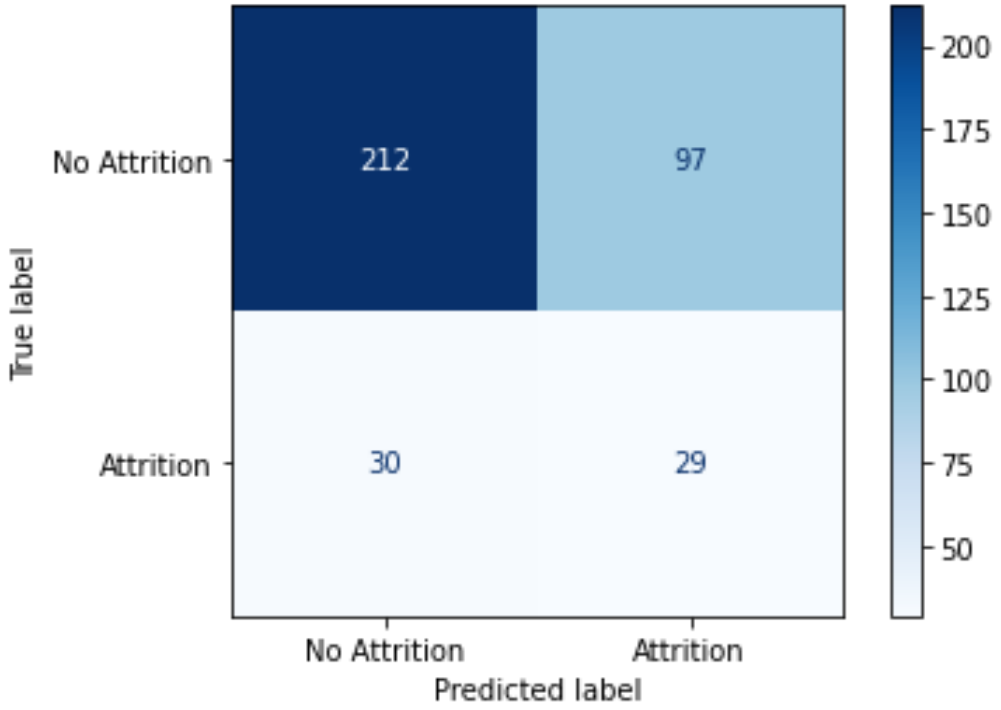
Models and Evaluation

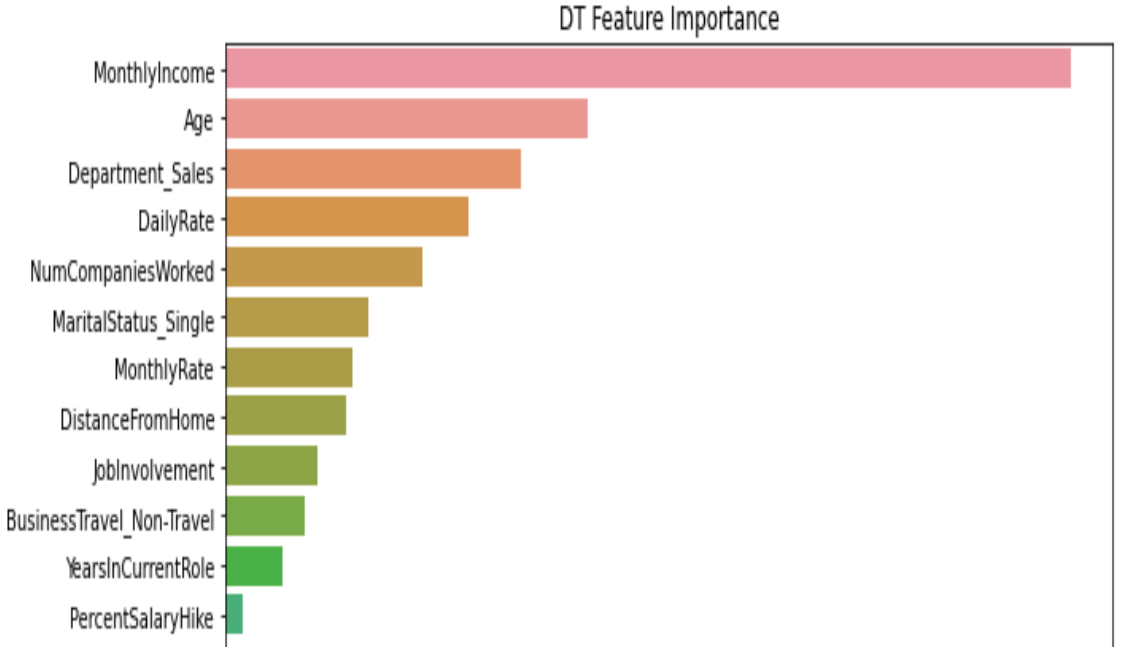


Decision Tree Model Evaluation

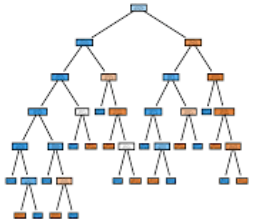


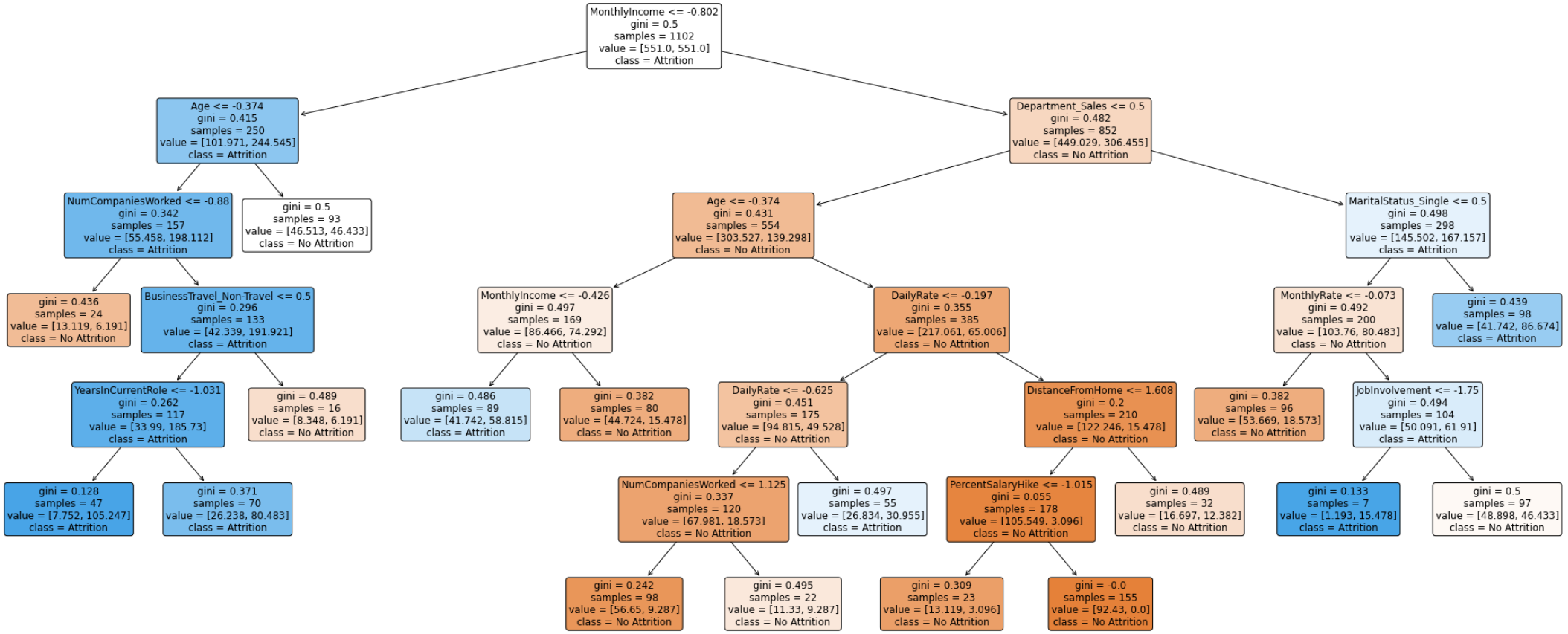
- *The training accuracy of this model was 72% and the test accuracy was 65%*



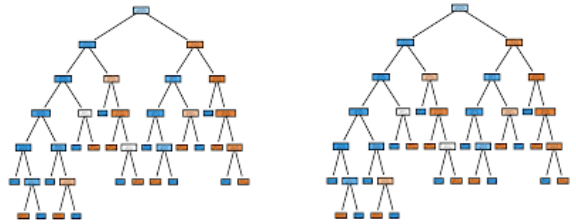


Decision Tree Model Evaluation

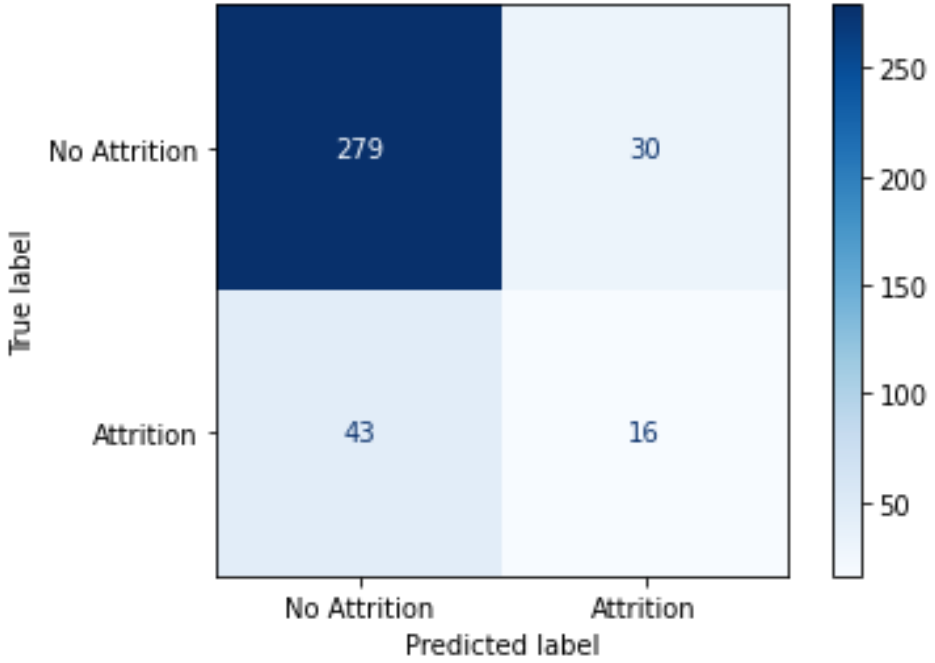


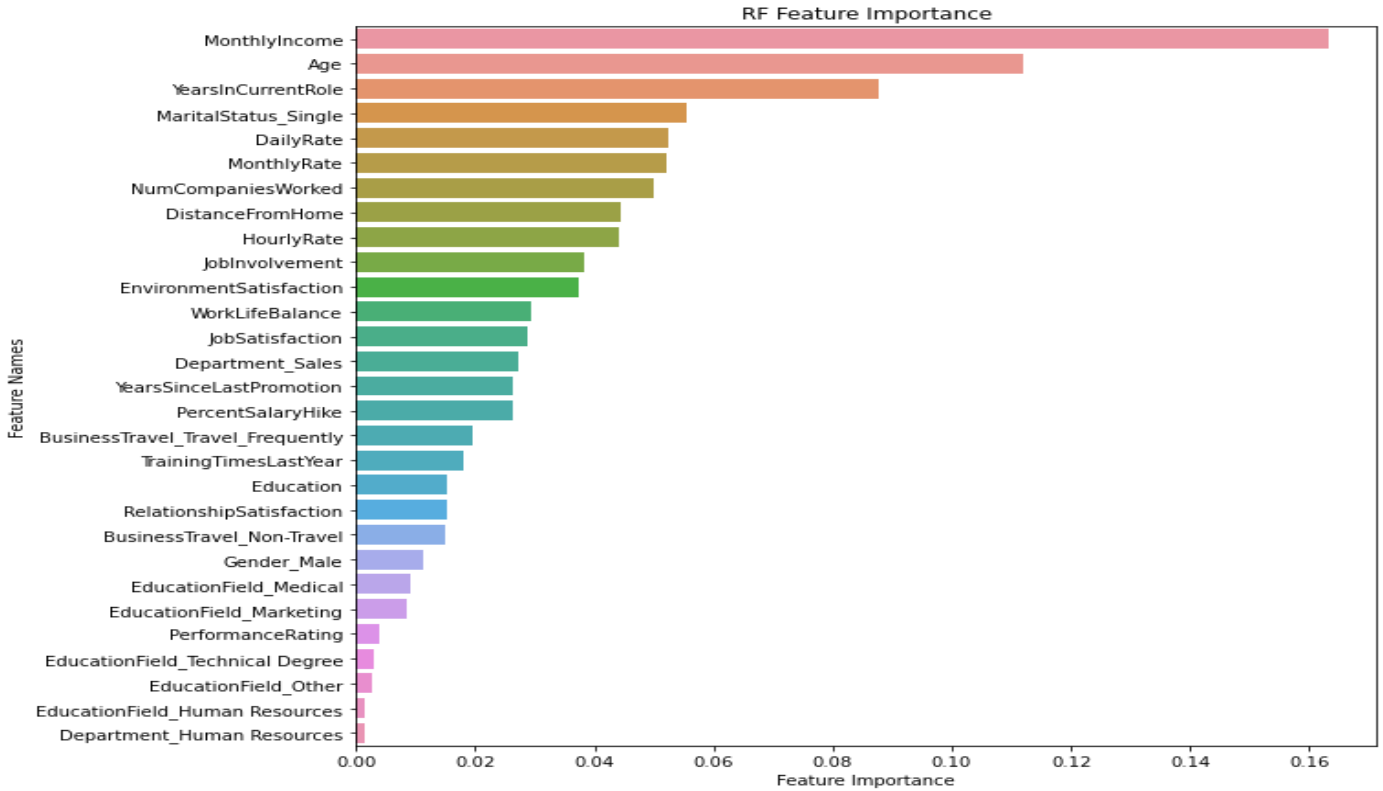


Random Forest Evaluation

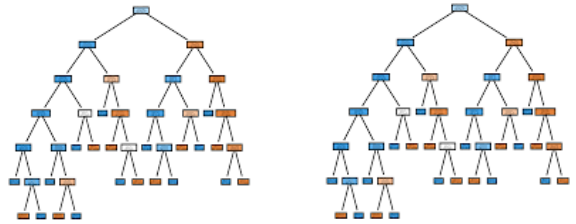


Random Forests are another supervised machine learning algorithm. They are made up of multiple decision trees and basically combines the output of the trees to reach a single result.



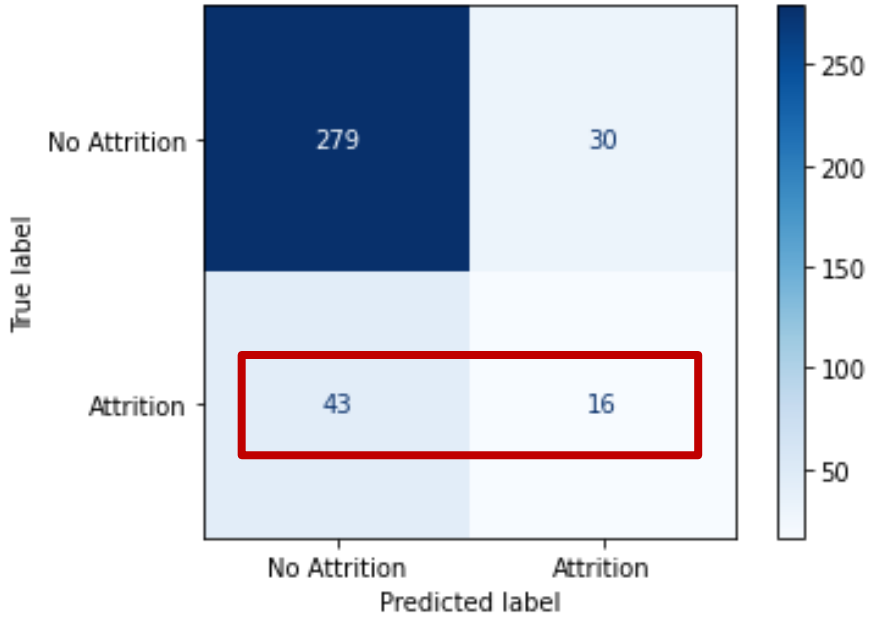


Random Forest Evaluation

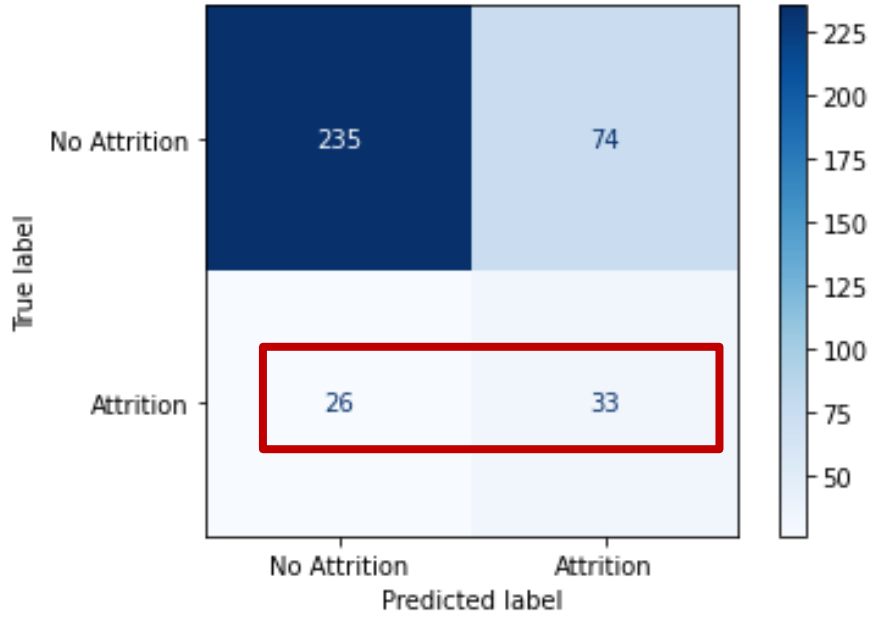


The second random forest model with adjusted parameters performed better at predicting attrition, but more poorly at predicting people that stayed

Original RF Model

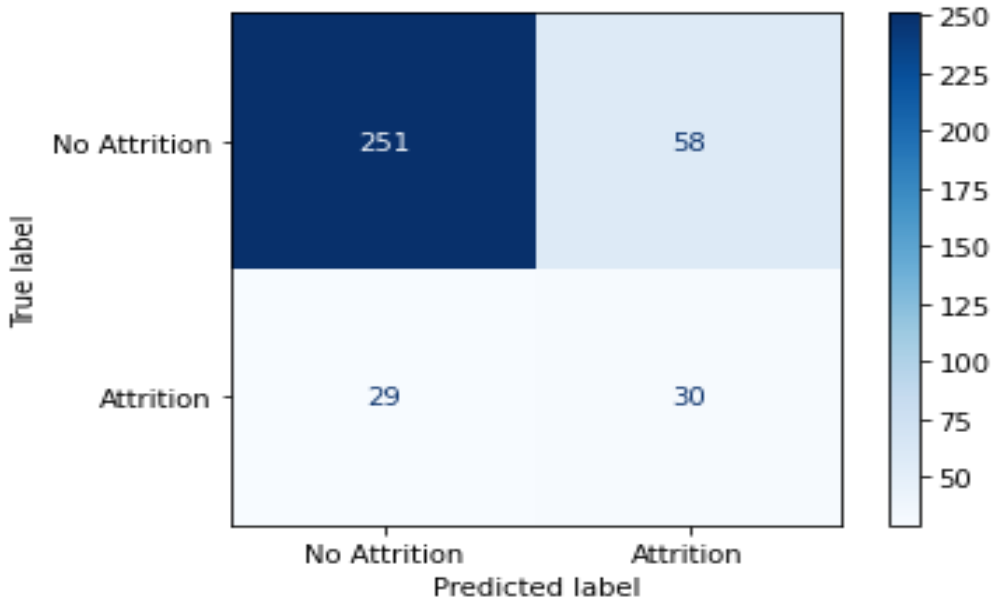


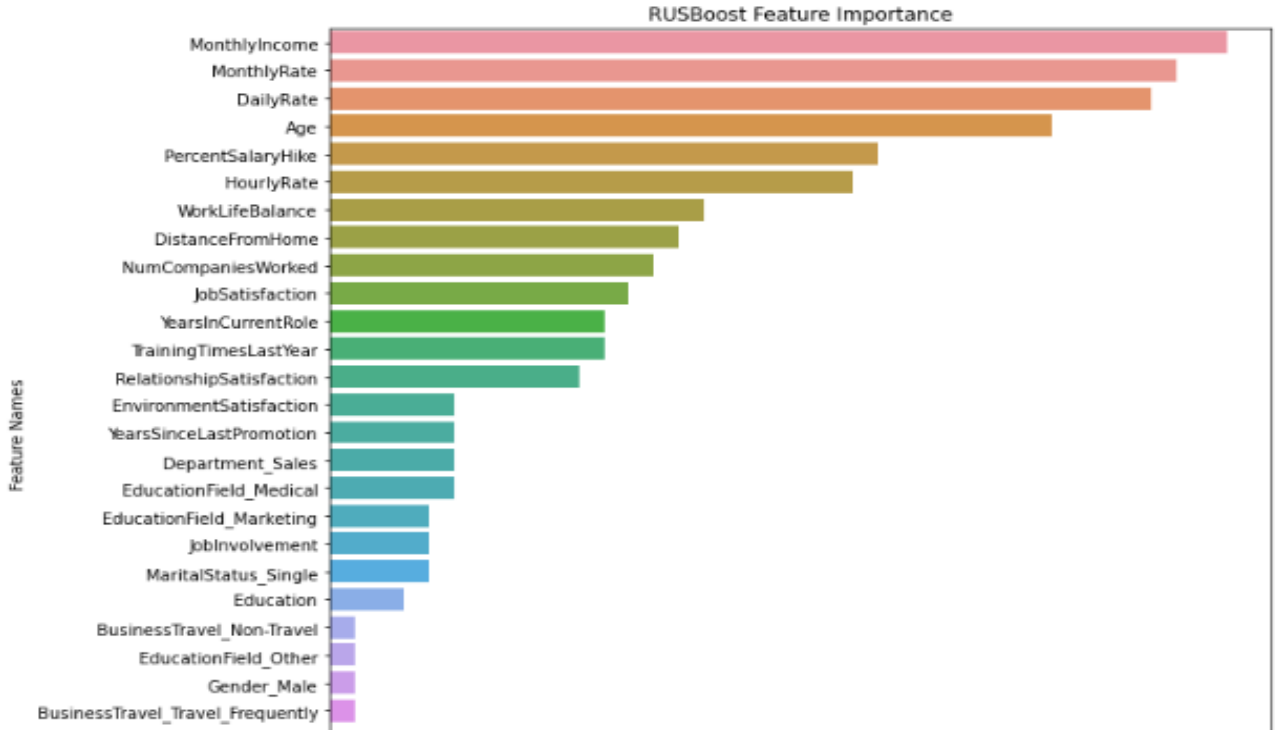
Adjusted RF Model



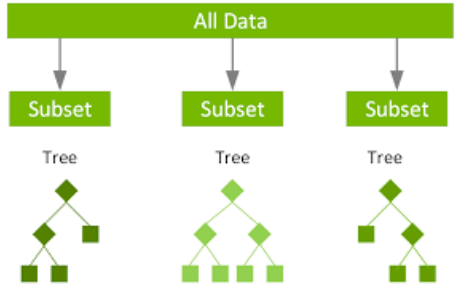
RUSBoost Classifier Model Evaluation

The RUSBoost algorithm uses a combination of random under-sampling and a standard boosting procedure to better predict minority classes.

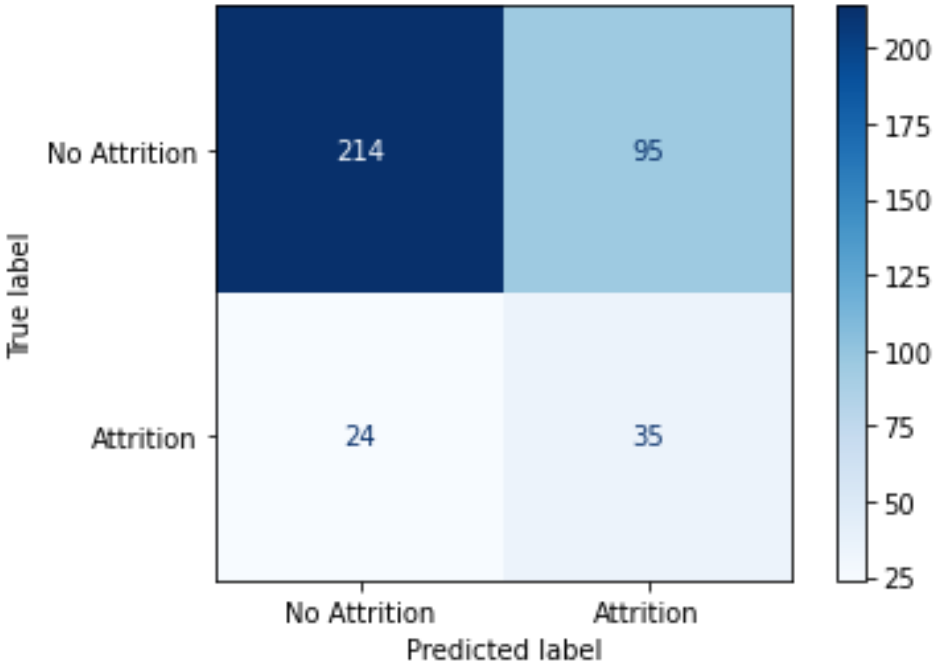


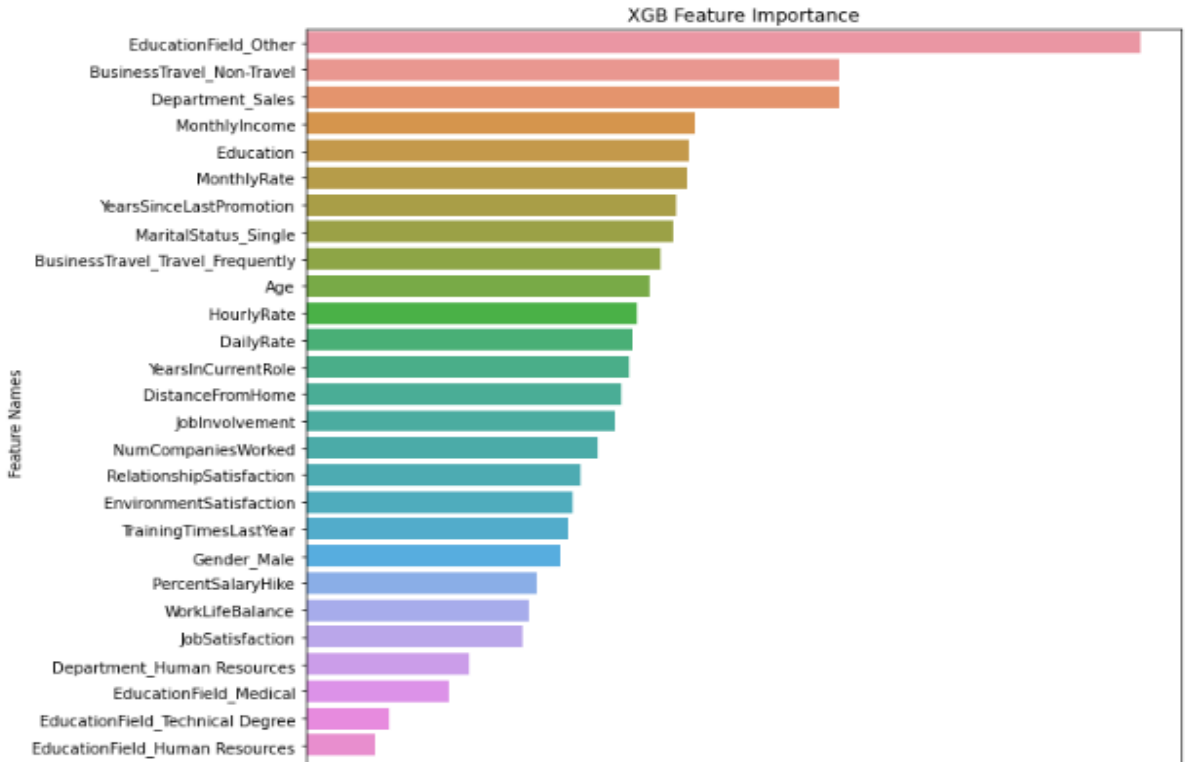


XGBoost Model Evaluation

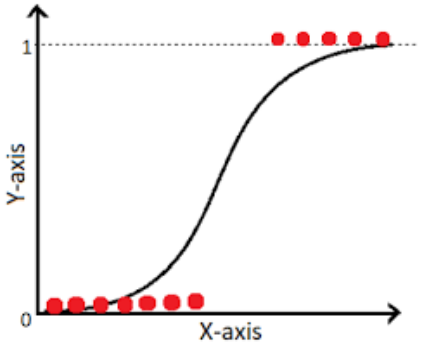


The XGBoost algorithm performed a little better than the RUSBoost model on the minority class, however it did worse on predicting people that stayed.

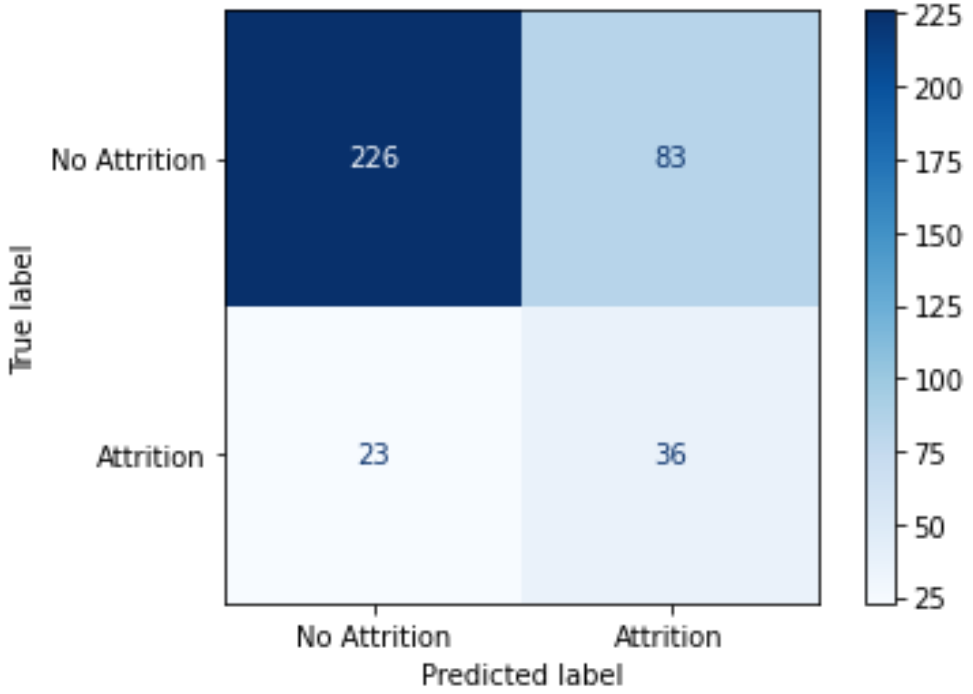


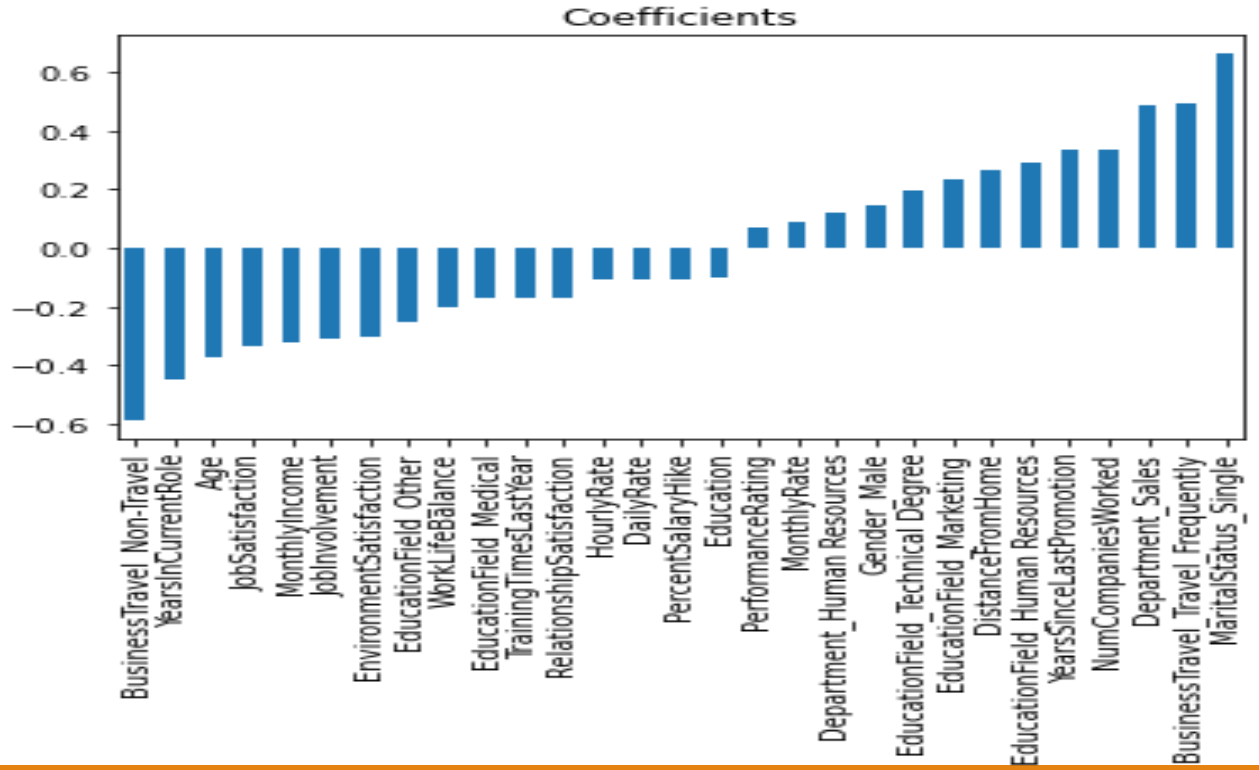


Logistic Regression Evaluation



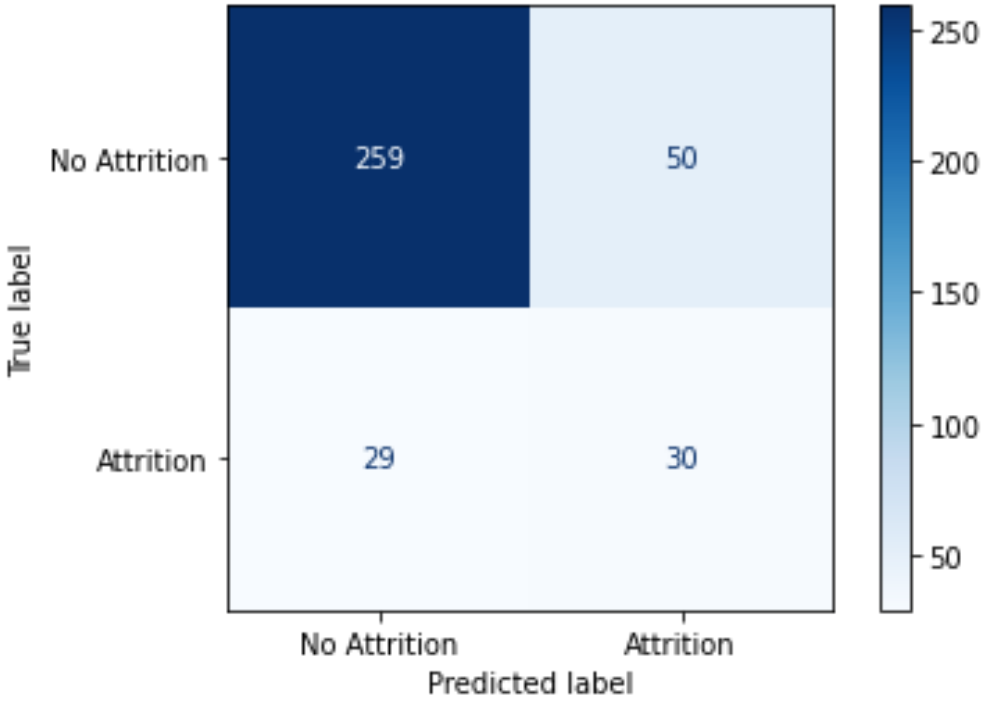
If the primary goal is to predict individuals that left the job, logistic regression performed the best on the test dataset as it correctly identified 36 people that left the job (more than any other model).

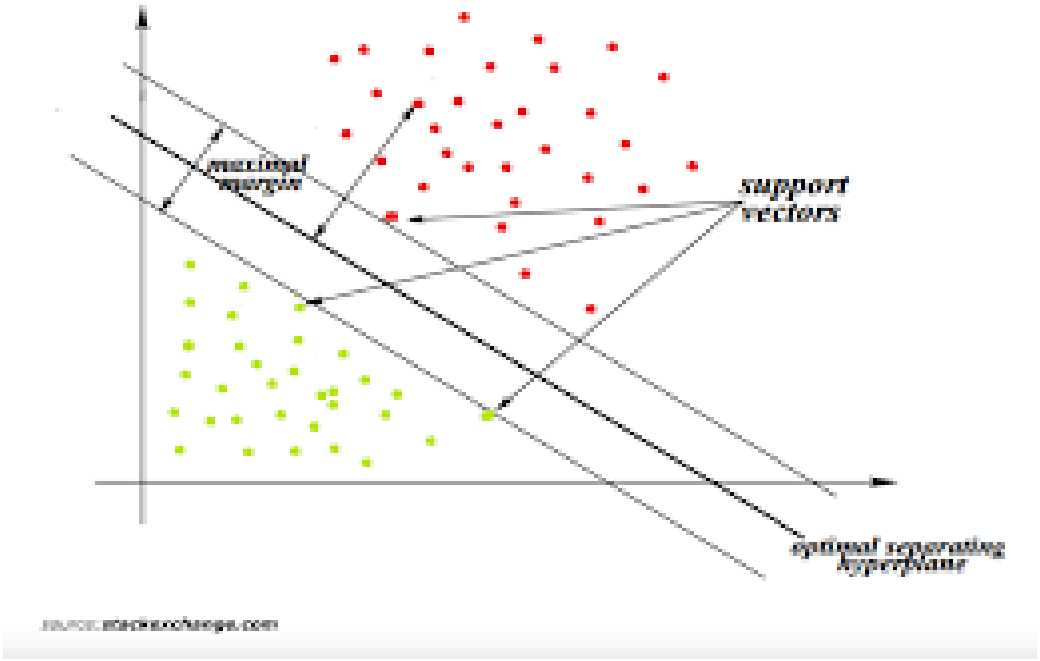




Support Vector Machine Model

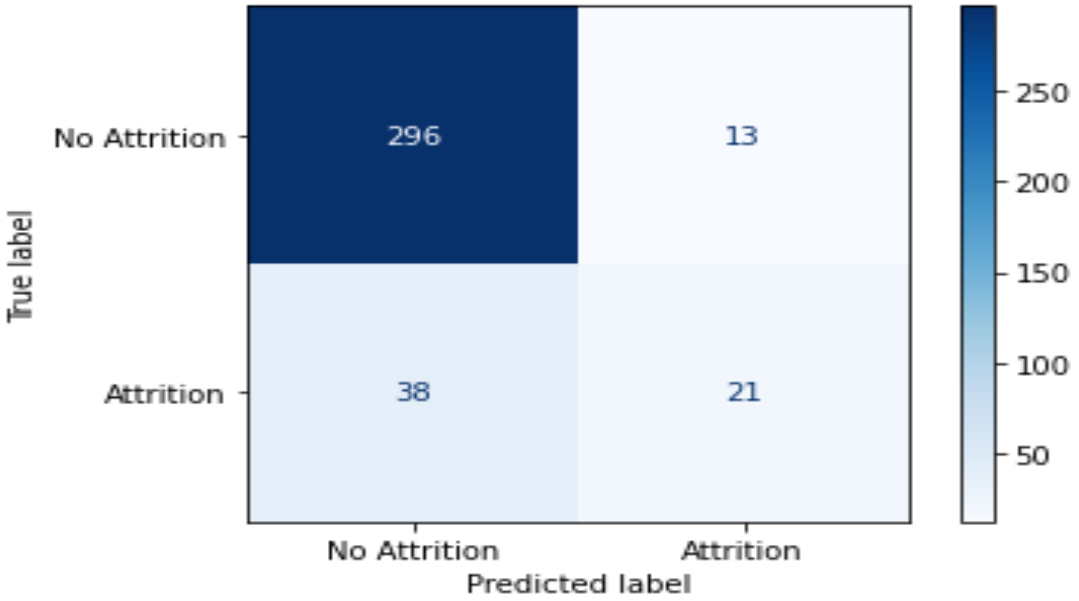
- *This model's train accuracy was 90% whereas the test accuracy was 79%*

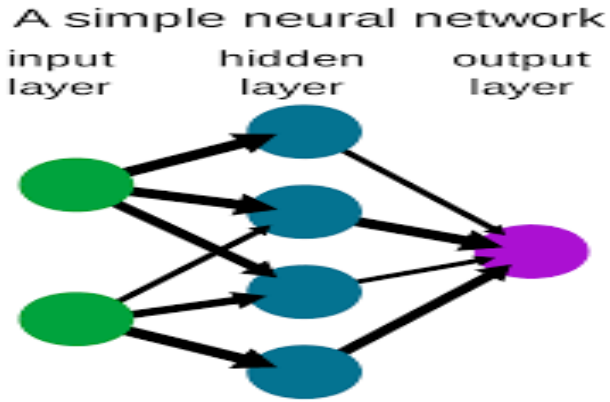




Neural Network Model Evaluation

- *This model performed better than all the other models so far at predicting people that stayed*



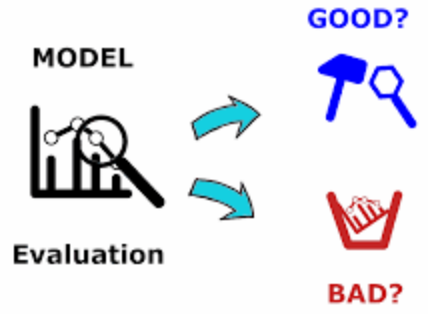


	precision	recall	f1-score	support
0	0.89	0.96	0.92	309
1	0.62	0.36	0.45	59
accuracy			0.86	368
macro avg	0.75	0.66	0.69	368
weighted avg	0.84	0.86	0.85	368

This model has the highest train/test accuracy

The accuracy of the train set is: 0.8747731397459165
The accuracy of the test set is: 0.8614130434782609

Final Model Evaluations



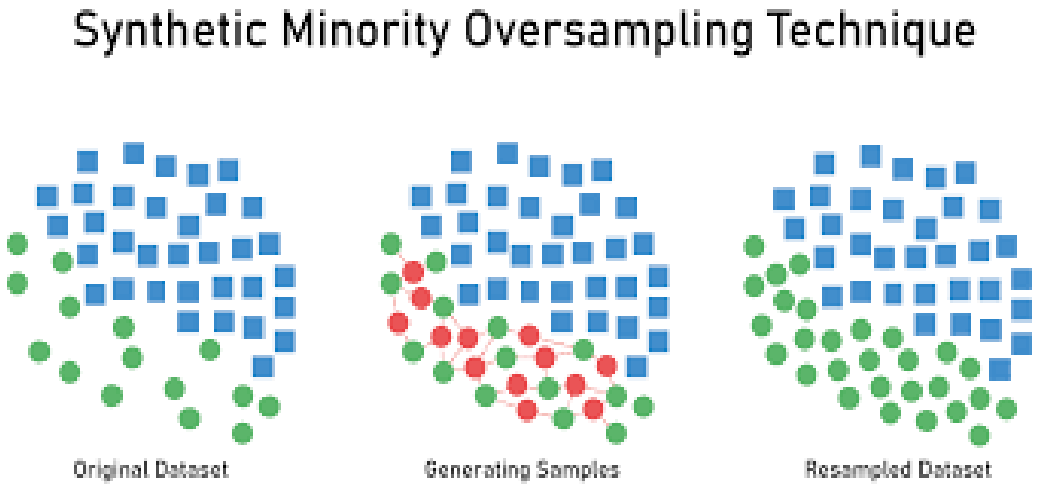
		Accuracy	Precision	Recall	F1-Score	AUC
DT	train	0.727768	0.333333	0.685393	0.448529	0.710662
	test	0.654891	0.230159	0.491525	0.313514	0.588805
RF	train	0.910163	0.723164	0.719101	0.721127	0.833035
	test	0.801630	0.347826	0.271186	0.304762	0.587050
RF1	train	0.774955	0.380952	0.629213	0.474576	0.716122
	test	0.728261	0.308411	0.559322	0.397590	0.659920
LogReg	train	0.726860	0.344304	0.764045	0.474695	0.741871
	test	0.711957	0.302521	0.610169	0.404494	0.670781
XGBoost	train	0.847550	0.514451	1.000000	0.679389	0.909091
	test	0.676630	0.269231	0.593220	0.370370	0.642888
RUSBoost	train	0.838475	0.500000	0.870787	0.635246	0.851519
	test	0.763587	0.340909	0.508475	0.408163	0.660386
SVM	train	0.905626	0.638060	0.960674	0.766816	0.927848
	test	0.785326	0.375000	0.508475	0.431655	0.673331
NN	train	0.874773	0.717391	0.370787	0.488889	0.671324
	test	0.861413	0.617647	0.355932	0.451613	0.656931

- *Some of these models are likely overfitting*
  - *Train versus test accuracy in DT, RF, XGBoost, RUSBoost, & SVM models*
- *Accuracy scores can be misleading*
  - *Neural Network is the most accurate model…however, Logistic Regression has a slightly higher AUC score on the test set.*
- *Goal = Predict Attrition*
  - *Logistic Regression is recommended model for deployment due to simplicity and highest test set AUC score*

SMOTE Oversampling

Imbalanced classification can cause issues with many machine learning models. One method to deal with this imbalance is creating new synthesized samples from the minority class. This is called Synthetic Minority Oversampling Technique or SMOTE for short.

		Accuracy	Precision	Recall	F1-Score	AUC
SMOTE Log Model	train	0.771104	0.765079	0.782468	0.773676	0.771104
	test	0.706522	0.271028	0.491525	0.349398	0.619549
Original Log Model	train	0.711039	0.714758	0.702381	0.708515	0.711039
	test	0.711957	0.302521	0.610169	0.404494	0.670781
Smote RUSBoost	train	0.933442	0.955631	0.909091	0.931780	0.933442
	test	0.834239	0.475000	0.322034	0.383838	0.627036
Original RUSBoost	train	0.719156	0.783217	0.606061	0.683344	0.719156
	test	0.763587	0.340909	0.508475	0.408163	0.660386



- *This did improve the SMOTE RUSBoost model accuracy, although there may be some overfitting of the train set*

Conclusion

- In this case, the final recommendation would be to utilize the logistic regression model as it was one of the better performing models at predicting attrition. It is also a less complex model which is preferred in many cases.
- There are more advanced over/under sampling techniques that could be used; however, they add complexity to the models.
- Additionally, the utilization of a hyperparameter optimizer such as Hyperopt or GridSearchCV could help optimize the models for better performance.
- Consistently, the most important predictor variable turned out to be monthly income. As employees' salaries increase, they are less likely to result in attrition.