Final Project: Employee Promotion Prediction

By Margarita Espinoza

GitHub Repo: https://github.com/maes2069/HR-Analytics-Bayesian-Classification.git

What problem do you solve?

This project is a **classification problem,** set to predict whether an employee in a multinational company will be promoted or not.

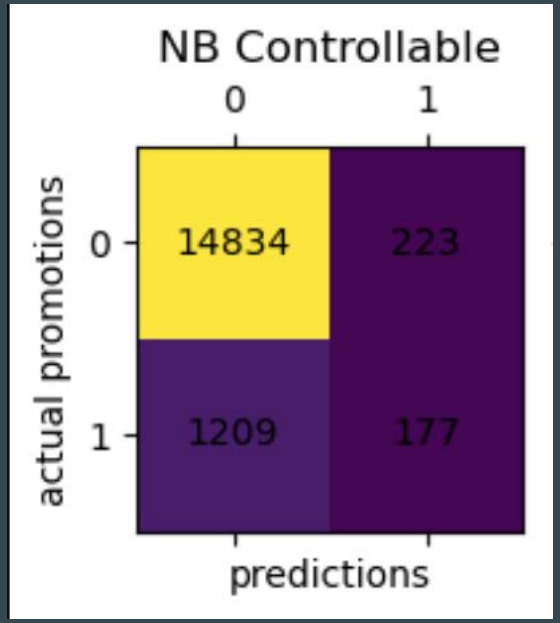
My goal was to learn details about this company's promotion criteria and how features that are **controllable** by the employee (such as education and number of trainings completed) and **uncontrollable** (like age and gender) impact their probability of being promoted.

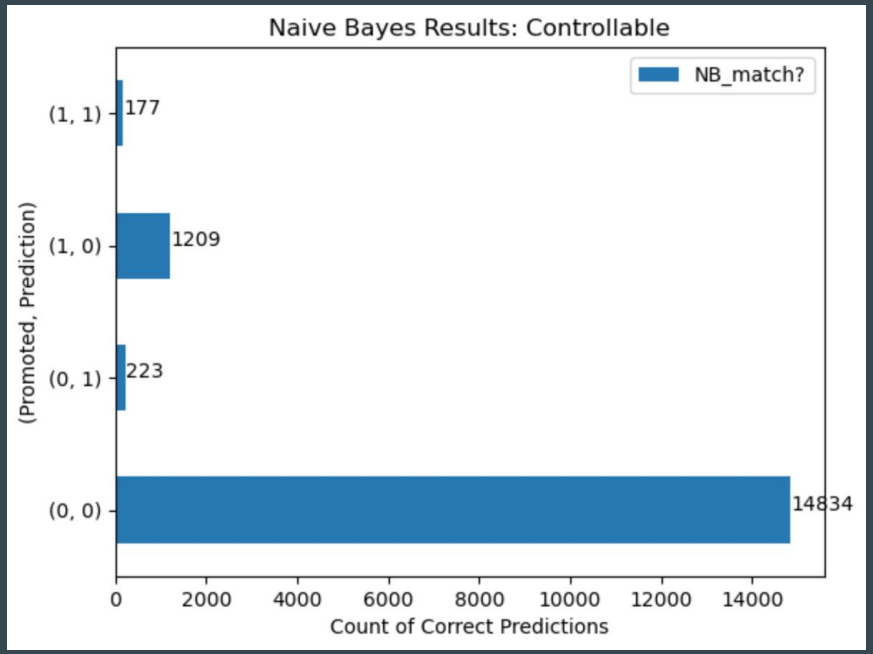
Machine Learning Approach

I split the dataset into 2 dataframes, categorized by features that are controllable and uncontrollable.

1. **Naive Bayes Classification** - uses Bayes Theorem to calculate the probability of predictions.
    a. **Controllable Features**: Education, number of trainings complete, previous year rating, length of service, if awards are won, average training score
    b. **Uncontrollable Features:** Age, gender
2. **Logistic Regression** - a classification algorithm used to predict a binary outcome based on a set of independent variables.
    a. **Uncontrollable Features:** Age, gender
        i. Tested different thresholds between 0.1 and 0.99

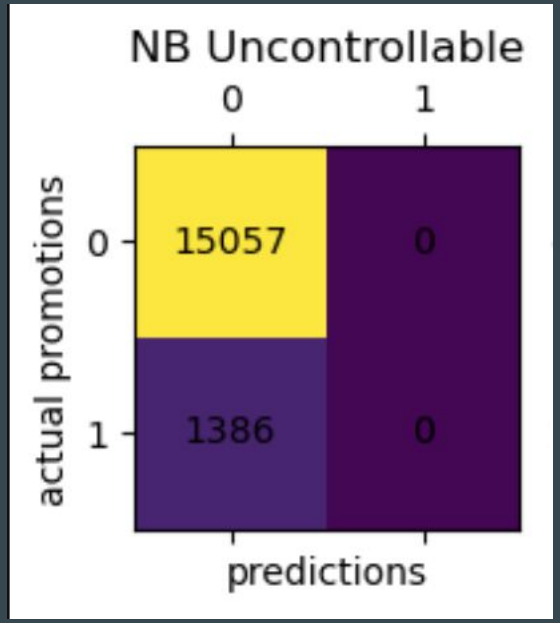
Results: Naive Bayes on Controllable Features

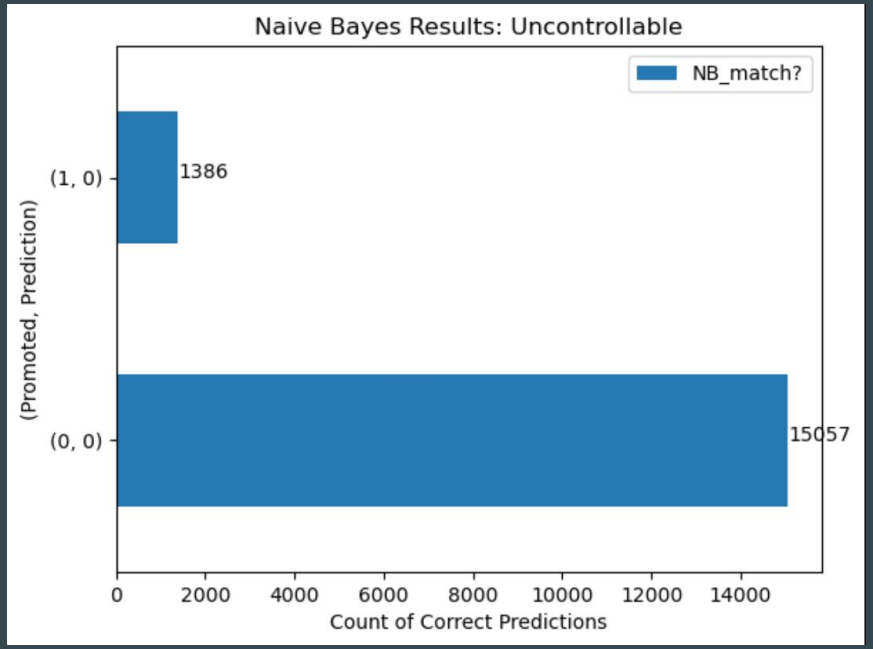




Naive Bayes score: 91.3 %

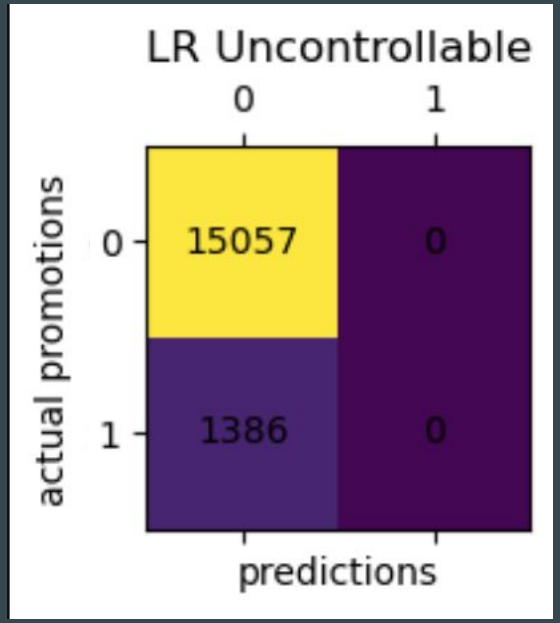
Results: Naive Bayes on Uncontrollable Features

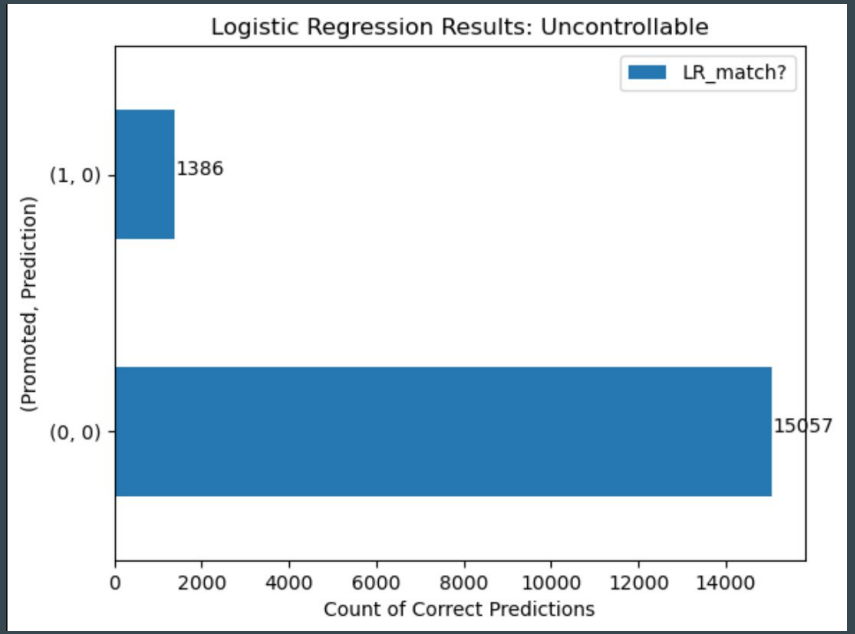




Naive Bayes score: 91.5 %

Results: Logistic Regression on Uncontrollable Features





Logistic Regression score: 91.5 %

Results: Logistic Regression with Adjusted Thresholds

```
******** For i = 0.1 ******
Our testing accuracy is 0.9157088122605364
[[15057     0]
 [ 1386     0]]

******** For i = 0.2 ******
Our testing accuracy is 0.9157088122605364
[[15057     0]
 [ 1386     0]]

******** For i = 0.3 ******
Our testing accuracy is 0.9157088122605364
[[15057     0]
 [ 1386     0]]

******** For i = 0.4 ******
Our testing accuracy is 0.9157088122605364
[[15057     0]
 [ 1386     0]]

******** For i = 0.5 ******
Our testing accuracy is 0.9157088122605364
[[15057     0]
 [ 1386     0]]
```

```
******** For i = 0.6 ******
Our testing accuracy is 0.9157088122605364
[[15057     0]
 [ 1386     0]]

******** For i = 0.7 ******
Our testing accuracy is 0.9157088122605364
[[15057     0]
 [ 1386     0]]

******** For i = 0.8 ******
Our testing accuracy is 0.9157088122605364
[[15057     0]
 [ 1386     0]]

******** For i = 0.9 ******
Our testing accuracy is 0.9157088122605364
[[15057     0]
 [ 1386     0]]

******** For i = 0.99 ******
Our testing accuracy is 0.9157088122605364
[[15057     0]
 [ 1386     0]]
```

We don't see any change in results as we adjust the threshold from 0.1 to 0.99.

Conclusion

Did my models work? Yes and No.

My first Naive Bayes Classification model on controllable features was wildly successful at predicting promotions for each employee, however that same model failed at predicting any promotions for uncontrollable features. After Naive Bayes failed to produce accurate predictions, I attempted a Logistic Regression model while adjusting the threshold to test if the issue was due to a poor model or poor dataset - which also failed. I can confidently say that my models failed because the uncontrollable features were simply not enough correlated data to generate positive predictions.