Overview of the Analysis – Brian Marowsky

In This Assessment of our Credit risk classification data. I used 2 different types of machine learning models to compare data. **Receiver Operating Characteristic**. It's a graph that shows how well a model can distinguish between two classes (like 0 and 1). **Decision Tree** is like a flowchart that makes decisions step by step. It splits data into smaller groups based on rules until it reaches a final decision or prediction. I decided to try ROC first being that the data

In Looking at the Dataset, It appeared that the loan status column was the column I could tell with the most difference between the lowest number and the highest number in the set. In doing the value count, I saw that over 96% of the data was in the first group (0) vs the second group (1). This told me there was a big imbalance in the 2 groups which would mean the model would have a hard time predicting the group 1 vs group 0 where a majority of the data was kept.

In order to make the model work the best, cleaning of the data as well as using a standard scaler was a must being there a big difference between the lowest and highest values.

Once the scaler was complete, it was time to do classifications or –- sort into groups. Once complete, Running the Logistic Regression first created the test/ train metric for ROC.

The ROC model returned

Once I had the results, I decided to run a Decision Tree model to compare if there would be any differences to the sets of data run in the model. After running the model, the conclusion is DT model does a good job but not as strong as for Class 0:

- Precision (0.89): 89% of predicted "1s" are correct.
- Recall (0.83): Captures 83% of actual "1s." more room for improvement for sure!
- F1-score (0.85): Balances precision and recall

I would say predicting the 1’s would be the most important as the 0 are a test environment where the model is learning vs the 1’s which is the actual run of the data which you want to be the most accurate.