I appreciate the comments you made about bias. Honestly, the bias in the prediction can have more than one interpretation depending on which side you look at. Also, recently there have been some updates in Alteryx that affected the Model Comparison Tool and the values it presents.

NOTE:
Here want to make some clarifications about the Model comparison tool and the bias in each models prediction.

First please note that in the recent versions of Alteryx the Model Comparison tool computes the creditworthy and non-creditworthy accuracies differently than the older versions. Previously, in the output of the Model comparison tool creditworthy accuracies ranged from about 75% to about 80% while non-creditworthy accuracies ranged from about 45% to 87% depending on the models. Now In the most recent versions, creditworthy accuracies are in the range from about 86% to 97%, while non-creditworthy accuracies are in the range from 35% to 45%. All of that happens because the Model Comparison tool used the values from the confusion matrix to calculate the different type of accuracies. In the previous versions, the Model Comparison too computed Positive Predictive value (also called Precision) and Negative Predictive value as creditworthy and non-creditworthy accuracies, respectively. Now, the new versions in the Alteryx the accuracies that are computed are the True Positive Rate and True Negative Rate. You can read more about the different metrics here - https://en.wikipedia.org/wiki/Confusion_matrix. All of those metrics can be derived from the confusion matrix. The values from the confusion matrix have not changed or are changed insignificantly.

The thing is that in our case we have an imbalanced dataset. There are a lot more creditworthy applicants than non-creditworthy. So in our case to be able to select a model for prediction we are interested in the overall accuracy and the Positive Predictive value (also called Precision) and Negative Predictive value. We are not that interested in True Positive Rate, and True Negative Rate also called Recall and Specificity respectively that the current versions give us. The good thing is that we can calculate Positive Predictive value (also called Precision) and Negative Predictive value ourselves from the confusion matrix that the Model Comparison tool gives us in no time. But first, let’s see what Recall, Specificity, Positive Predictive value (also called Precision) and Negative Predictive value mean.

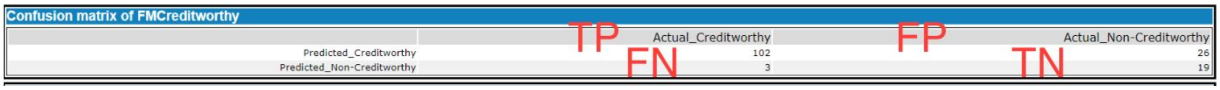
- **Recall** tells us if the person is creditworthy how often will the model predict that he is creditworthy.

- **Specificity** tells us if the person is not-creditworthy how often will the model predict that he is not-creditworthy.

- **The Positive Predictive Value (PPV)** tells you how likely it is for someone who was predicted to be creditworthy actually is creditworthy. It answers the question, “The customer was predicted to be creditworthy. Does this mean they definitely are creditworthy?”

- Equally the **Negative Predictive Value (NPV)** tells you how likely it is for someone who was predicted to be non-creditworthy actually is non- creditworthy. I.e., it answers the question “The customer was predicted to be non-creditworthy. Does this mean they definitely are non-creditworthy?”

Let’s see how we can calculate the PPV and NPV value for the Forest model from the confusion matrix. We have the following values in the confusion matrix:



Confusion matrix of FMCreditworthy

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

True Positives - 102
False Positives - 26
False Negatives - 3
True Negatives - 19

Recall = True Positives / (True Positives + False Negatives) = 102/ (102 + 3) = 0.97
Specificity = True Negatives / (True Negatives + False Positives) = 19 / (19 + 26) = 0.42

In general, models that give high Recall will have low Specificity. In other words, they are good for catching actual cases of creditworthy people but they also come with a fairly high rate of false positives.

For the first two metrics, we took the values vertically. And for the next two metrics, we will take them horizontally from the confusion matrix.

PPV = True Positives / (True Positives + False Positives) = 102/ (102 + 26) = 0.8
NPV = True Negatives / (True Negatives + False Negatives) = 19 / (19 + 3) = 0.86

The higher the value for the PPV/NPV, the more accurate the model prediction is. So having, those values in mind now this model is the strongest in terms of overall accuracy, and it also had the best predictions of non-creditworthy individuals (86%). This model predicted whether an individual was creditworthy or not at almost an equal percentage, indicating little to no bias in this model. You can compute the PPV and NPV for the other 3 models to verify that following the same type of calculation. The other models have lower PPV and NPV. You will see that the Boosted model has similar performance as the Forest but not better. If we use the Decision tree and Logistic regression models we would deny a loan to many creditworthy individuals as it classifies many creditworthy applicants as non-creditworthy. The two models have high PPV but low NPV. So we might say that those models are biased towards predicting individuals who are creditworthy, as they do not predict individuals who are not creditworthy nearly at the same level as those who are.

Another useful metric here is the F1 score which is often used too when comparing binary model performance. An F1 score looks for the bias in "Creditworthy" and "Non-Creditworthy" and calculates a single score: the higher the score, the better. You can read more about the F1 score here: https://en.wikipedia.org/wiki/Precision_and_recall You'll need to read more about precision and recall in order to understand the F1 score. As you can see the Forest model has the highest F1 score here.

To sum up we are interested in the overall accuracy, PPV, NPV and the F1 score to determine the best model. We want max values on all of them because they describe the predictive power of the model. The higher the values the more accurate the model.