How to Use the AUC Calculator.

In the final project you will develop your own binary classification model. Your model will use historical data to forecast how probable it is that a new applicant for a credit card will default.

The best way to measure the effectiveness of any binary classification model (whenever the relative cost of False Positive and False Negative classification errors is not known) is by using the metric known as the "Area Under the Curve," or "AUC."

You will use the **AUC Calculator** spreadsheet to complete the final course project. You will use it to test and improve your binary classification model. Improving your model will increase the Area Under the Curve. You will also confirm the validity of your model using the **AUC Calculator** Spreadsheet by observing whether, when forecasting based on new input data not used to develop the model, the value of the AUC remains relatively stable or is much lower with new data. If the AUC is much lower on new data, it indicates that your model is *overfit* to the old data.

AUC was first demonstrated in the **Bombers and Seagulls** Spreadsheet [Cell I28].
A step-by-step review of how AUC is calculated is provided in the **Review of AUC Curve** Spreadsheet, which is attached to the **AUC Calculator** Spreadsheet in the same workbook.

You can of course create your own spreadsheet to calculate the false positive rate and true positive rate at each threshold score of your binary classification model, and then measure the area under the resulting curve – the AUC.

The added convenience of the **AUC Calculator** Spreadsheet is that it allows you to input a list if scores and outcomes and calculate the AUC *without needing to reorder the scores from highest to lowest.*

This feature makes it efficient to try many variations quickly while developing your model. Simply insert your model scores in [Column A] and your outcomes in [Column E], and read the correct Area Under the Curve in [Cell G8].

However, in order to take advantage of this spreadsheet's convenience you may need to make one or two adjustments to your score data.

First Adjustment: you may need to convert the inputs scores created by your model to values that fit in the limited number of columns of the spreadsheet – namely values between [-3.5, 3.5].

Convert the inputs to numbers that all fall within a -3.5 to 3.5 range by *standardizing* them. [Column B, cells 23 to 222].

You can confirm that the values fit within the range because the Maximum after standardization is given in [Cell B17] and the minimum after standardization in [Cell B18]. Further adjustment can be made if the values are still too spread out by dividing by a number greater than 1 in [Cell C20].

Second Adjustment: In your binary classification, outcome 1 means "default," and outcome 0 means "no default. You always want to associate your model's *highest* scores with the greatest probability of a "1" outcome.

However, some input variables have the property that the *lowest* score is more associated with the positive outcome of default. Age is example, because *younger* people are more likely to default. As an example, one simple binary classification "model" is to use the ranking provided by the age of each borrower. The age of 200 borrowers is given as an example [Column A, rows 23 to 223].

If the *lowest* value is most associated with "1" outcomes, the area calculated as the sum of rectangles in the AUC calculator will be less than 0.5. In the example shown, it will be .36 [Cell G2]. However, this is not the Area Under the Curve, which must always be greater than or equal to .50. The actual Area Under the Curve is shown in cell [G8].

You may need to correct your data so that the *highest* scores are associated with default. Simply insert -1 in [Cell C20] – to divide all model scores by -1. This will make the area under the rectangles [Cell G2] the same as the correct area under the AUC curve [Cell G8].

Note that with the **AUC Calculator** Spreadsheet you can also enter your chosen costs per False Positive classification in [Cell I2] and costs per False Negative Classification in [Cell I4] and immediately identify the threshold for "Positive" classification that has the *minimum cost per event*. The minimum cost per event is given in [Cell J2] and the corresponding threshold in the column in Row 17 that has the same value as [Cell J2]. In the example given, if you don't first divide standardized ages by -1 [Cell C20], the minimum cost per event will not be accurate.

Multiplying all scores by -1 gives an accurate minimum cost per event of $975, [Cell BH17] which matches to a threshold of .9 - shown in the same column [BH] of Row 10.