Learning Points from Final Project: Part 1

This project requires you to develop your own binary classification model for credit card default, using a "Training Set" of 200 applicants, and then to test your model for over-fitting on "Test Set" of an additional 200 applicants.

The Area Under the Curve (AUC) performance metric is threshold-independent, which makes it the best measure of the effectiveness of a binary classification model whenever the cost per False Negative classification and cost per False Positive classification are not known. You design your model to maximize "sustainable" AUC: AUC that does not fall significantly from the Training Set to the Test Set.

You confirm that your model does not suffer from excessive over-fitting by measuring the AUC of the model on the Test Set.

In general, models with "too good to be true" AUC on a Training Set will perform much less well on new data. If your model had a dramatically lower AUC on new data, you should go back and redesign it. It is better to have a model with lower AUC that is consistent across both data sets – it is more likely to have robust performance with new applicants.

Below is an example of an effective and robust scoring model. This is not the only model, or even the best model, possible. You should develop your own model to learn how to experiment with the data inputs.

(-6*(raw cc debt + raw auto debt)/(raw income))
-(.3*(standardized age + standardized years at employer + standardized years at address)) – 2.9

This model has an **AUC** of **.77** on the **Training Set** and **.79** on the **Test Set**.

The model fits the Test Set data as well or better than it fits the Training Set. This does not *guarantee* there is no over-fitting, because better performance on Test Set could be due to chance, or random noise fit by the model could be unusually similar in the Training Set and Test Set, so the model could still have lower performance in future – but the probability of that is lower with consistent performance on data on which the model had not been trained.

Once estimates of cost per False Negative ($5,000) and cost per False Positive ($2,500) [as given in the **AUC Calculator** Spreadsheet] become available to you, the ideal performance metric for your model *changes* from AUC to *average cost-per-event* at the model's threshold score. Because you do not know future outcomes ahead of time, you must keep the threshold that is optimal on the Training Set unchanged when calculating cost-per-event on the Test Set.

Comparing cost-per-event on the Test Set and Training set is a more challenging test of model robustness than comparing the two Areas Under the Curve.

If the cost-per-event increases dramatically, that is a sure sign of model over-fitting and the model should be redesigned. It is better to have a model with higher cost-per-event that is consistent across both data sets.

At the above costs, the example model has a minimum cost-per-event of **$800** on the **Training Set** at a **threshold of .25** [Column BU on the AUC Calculator Spreadsheet].

On the **Test Set**, at the same threshold of .25 in Column BU, the new, sustainable cost-per-event is **$875** [Cell BU17]. The increase between $800 and $875 is small and suggests minimal over-fitting.

It is of course common for the AUC to be somewhat lower, and the cost-per-event somewhat be higher, on the Test Set. It is a matter of judgment and experience whether the differences are so large that you should go back and redesign your original model.

It is important to understand what is meant by the *base rate*, and the cost-per-event at the base rate.

The condition incidence in the Training Set and Test Set is that 25% of all outcomes are defaults, and 75% are non-defaults.

A model with no predictive power will have an AUC of .5. Why? Because the true positive rate will equal the false positive rate at every possible threshold. The Area under the Curve will be the area under the line x = y from (0,0) to (1,1).

At a cost per default of $5,000 (False Negative) doing no forecasting results in a cost-per-event of (.25*$5000) = $1,250. We can conclude that *base rate cost-per-event* is $1,250.

For any classification model to have information gain (reduce uncertainty) it must outperform a base-rate forecast. This means that on the Test Set:

The AUC must be greater than .5,
More than 25% of events classified as “positive” must be defaults,
More than 75% of events classified as “negative” must be non-defaults, and
The cost-per-event must be less than $1,250.

The incremental value of the model is measured in savings per event. Savings equal the difference between the $1,250 cost-per-event at the base rate and the average cost-per-event with your predictive model.

The model above will save ($1,250 - **$875**) = **$375 per event**.

At 1,000 applicants per day, the payback period for the original $750,000 cost of the experiment used to gather the 600 data points is 750,000/1000*(savings per event).

For the example model above, the payback period is $750,000/$375,000 = **2 days**.