# Jason “Scott” Person

4.1 Learning Response: Advanced Model Evaluation

20250429

# Response Questions

### From Data Science for Business

1. What is a confusion matrix?
   1. Provide a basic definition.
   2. Find and paste a good image of a confusion matrix for a simple binary classification problem. (This matrix has only two rows and two columns.)
   3. Define and clarify these terms:
      1. True positives
      2. False positives
      3. False negatives
      4. True negatives
   4. Include this in your notes:
      1. Greater **precision** means a reduced number of false positives.
      2. Greater **recall** means a reduced number of false negatives.
   5. Why is the confusion matrix useful for helping us think more clearly about how to determine our best models in light of our business goals?

A confusion matrix is an n x n matrix with the actual values labeling the columns and the predicted values labeling the rows. The intersecting cell contains the number of instances for that combination of actual and predicted values.

Figure 1 below shows a good example of a confusion matrix.

A diagram of a person's negative values

Description automatically generated

Figure : Binary Confusion Matrix

For a 2 x 2 confusion matrix there are four possibilities:

1. True Positive – found in the upper left cell. The actual value is positive, and the predicted value is positive. The example in Figure 1 is when a woman is actually pregnant, and she was predicted to be pregnant by the model.
2. False Positive – found in the upper right cell. The actual value is negative (the man is not actually pregnant) but the model predicted positive (model predicted that man was pregnant). This is also known as a Type-1 error.
3. False Negative – found in the lower right cell. The actual value is positive (the woman is actually pregnant) but the model predicted negative (model predicted that the woman was not pregnant).
4. True Negative – found in the lower right cell. The actual value is negative (the man is not actually pregnant) and the model predicted that the man was not pregnant.

Precision and recall are common measures of model accuracy. Precision measures the accuracy of a model when it predicts positive. In other words when the model predicts positive, what are the chance is it correct. The formula is TP/(TP+FP). Recall is a measure of how accurate the model is at finding true positives. The formula is TP/(TP+FN). By inspecting the formulas, we can see that Precision will be increased with fewer false positives (thus a smaller denominator) and Recall will increase with few false negatives.

The completed confusion matrix provides metrics for benefit (TP and TN) and costs (FP and FN) associated with applying a model.

1. As the text discusses, many binary classification problems have very unbalanced classes (0 or 1) for their target variables. What are three specific examples the authors give of binary classification problems in which positive cases are quite rare?

The authors provide several examples of binary classification where the positive cases are rare:

1. Fraud – the number of fraudulent transactions/interactions is generally quite low.
2. Defective parts – a well running fabrication line will produce a low number of defects.
3. Customer offer response – the number of customers who will respond to an offer.
4. Consider the three above examples, plus the additional example of predicting cancer (discussed under “Problems with Unequal Costs and Benefits”). Now list those cases, and for each describe which is more important and why:
   1. Higher precision, with fewer false positives?
   2. Higher recall, with fewer false negatives?
   3. And for each, why ...

* Cancer diagnosis – in this case we would definitely want **higher recall**. Higher false negatives (lower recall) would result in missed cancers, which has a very high cost. In the event of false positives, follow up tests would eliminate the errors.
* Fraud detection – I think this one is quite situational. I am sure that lots of fraud goes undetected so in practice there is low recall, but I would generally think **high recall would be desired**. We want to catch the fraud because it is expensive. Interestingly, with my own accounts, I’ve gotten a much higher number (several) of calls about potential fraudulent transactions than have actually been detected (zero).
* Defective parts – as with cancer diagnosis, we want to minimize the missed positives and thus want **higher recall**. Defective parts that are missed will go to customers and/or end up installed in downstream assemblies with significant cost.
* Customer offer response – we would want **higher precision**. Missing a few potential customers is generally fine provided the customers you do reach are of high quality.

### From McCormick, Predictive Analytics Essential Training: Estimating and Ensuring ROI

**NOTE:** In these assigned videos, McCormick is on the same page with the authors of your text, who repeatedly remind us that we must work to remember the business imperatives as we evaluate the results of our machine learning models. McCormick brings some special clarity to this relationship in his discussion of the confusion matrix and the related business imperatives — many of which can be estimated in terms of dollar values.

1. In approximately 100-200 words, reflect on the ways in which the confusion matrix and ROI estimations — in relation to specific business need s — help us determine the comparative value of a specific metric (such as precision or recall) in evaluating our machine learning models.

The confusion matrix provides actual metrics for the various combinations of correct and incorrect predictions. Using these values, using what we know or can gather about the business, we can multiply by the approximate cost of each type of success and failure of the model. This will allow us to determine a cost/benefit for successes and failures as well as for the model as a whole. If the cost of false positives significantly exceeds that of false negatives, then we want to emphasize precision over recall. Alternatively, if false negatives are more costly then recall is more important.

1. For each of the examples McCormick discusses in these videos, which is the more important metric:
   1. Higher precision, with fewer false positives?
   2. Higher recall, with fewer false negatives?
   3. And for each, why ...

* Insurance fraud – missed fraud is very expensive both because of the direct cost of paying an invalid claim, but also because of potential litigation expense in the recovery. Because of this, false negatives are expensive, and we want to minimize them. Recall is more important.
* Predictive maintenance – damaged machinery has a very high cost both in terms of repair cost as well as potentially unavailability outages during a repair. There is usually a comparatively much lower cost for further testing well as a shorter outage in the event of a false positive. In this case, we want to eliminate the false negatives, which would occur when a machine is damaged due to unperformed maintenance. Recall is more important.
* Proactively offering refinancing to troubled borrowers – refinancing is expensive in time and thus dollars. It is important that we don’t attempt to refinance poor candidates. Here we would want a balance considering the cost of refinancing and the benefit of keeping the loans performing. Overall precision would be more important.

### From Vanderblock, Mistakes to Avoid in Machine Learning – LinkedIn Learning

**NOTE:** I have chosen to assign these videos by Vanderblock for these reasons: (1) to introduce you to this title, which — like McCormick’s — will be valuable if you go on to participate in a machine learning project; (2) to introduce you to some advanced feature engineering concepts, including:

* **Standardizing (aka Scaling) data** — While this step is sometimes important, DO NOTE that only certain models benefit from this. Vanderblock mentions KNN. SVM is another. (By contrast, Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting mechanisms do not *need* feature scaling, though they are not harmed by it either.) [See this article](https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35) if you want to learn more.
* **Imbalanced sampling —** especially SMOTE oversampling, which is typically the best approach — and is often needed to adjust for data in which the positive case is rare. Indeed, SMOTE oversampling will even benefit our Churn modeling project, since the positive case is proportionally much smaller than the negative.
* **The limitations of accuracy** as a sole metric — Vanderblock’s video helps drive this point home with additional illustrations and examples. In addition, he discusses the helpfulness of the ROC curve and AUC score and provides examples of how to implement an alternative approach in Python.

1. Define these key terms:
   1. Standardization (aka scaling) data – scaling data puts the values of all the features onto the same scale. This prevents the algorithm from overweighting features that have large values. There are two methods standardization, which centers the data so that the mean is 0 and standard deviation is one and normalization, which resizes the data to a fixed range usually 0 to 1 inclusive.
   2. Imbalanced sampling – this occurs when the number of records is disproportionally distributed across the range. An example of this is when the desired value of the target variable doesn’t appear often in the data.
2. When is *oversampling* the appropriate approach, as opposed to undersampling, and vice versa? We can use oversampling and undersampling with unbalanced sets to control the size of the training set. If the disproportionately small classes are very small, oversampling is a good choice. Conversely when our majority class is very large, we can use undersampling to constrain the size of the training set.
3. Why does the F1 score and/or the AUC score provide a more robust model evaluation metric than accuracy alone? Accuracy simply measures the percentage of correct predictions, which can be quite misleading when the data is imbalanced. If a positive result doesn’t show up often in the data, then a model evaluated simply on accuracy would appear to be quite good simply by always predicting a negative result. F1 and AUC provide a more complete measure of model behavior particularly for imbalanced data.

F1 – balances precision and recall to account for both false positives and negatives.

AUC – measures how well the model can distinguish between positive and negative results.

### From Radecic, How to Effortlessly Handle Class Imbalance with Python and SMOTE

1. What does SMOTE stand for? And when is it useful?

SMOTE: Synthetic Minority Over-sampling Technique

SMOTE is best used when the positive result set is much smaller than the negative result set and when we have concerns about overfitting.

1. How and why is SMOTE a more useful method by comparison to other oversampling methods? SMOTE introduces new data into the dataset. This will help prevent the model from simply memorizing the dataset.
2. In relation to the fraud data example Radecic provides, what is the weakness in the performance of the Random Forest model *without* SMOTE?
   1. Is this weakness in Accuracy, Precision, or Recall?
   2. What would be the related impact on the corresponding business imperative? (In other words, why is this a problem?)

Without SMOTE, the Random Forest missed 91% of the fraudulent transactions. In other words, there were many false negatives, which resulted in a very low recall. While the model is 98% accurate, it doesn’t add a lot of value. The whole point of this model is to identify fraud, yet the model predicted almost no fraud.

1. Using Smote, what improvement did Radecic achieve in model performance? Why will this benefit the corresponding business imperative?

With SMOTE, the model has accuracy and recall of 0.99. In other words, almost all fraudulent transactions were identified and there were very few false positives as well. Since the purpose of the model is to identify fraud without causing an inordinate number of eventually fruitless investigations, the model meets business needs.