**October 17, 2016**

Yan Wu

[Yanwu26@gmail.com](mailto:Yanwu26@gmail.com)

415-754-3290

**Preface:**

I was asked to deliver an evaluation of machine learning techniques for people without a deep machine-learning background.

1. **Evaluate this model's current performance:**

Null accuracy: is this model better than random? (Simply guessing the most likely probability)

In-sample accuracy: is this model predictive on in sample data?

Prediction accuracy: is this model predictive out of sample? (Overfitting)

1. **Evaluate the structure/integrity of the model/methods used**

Introduce Confusion Matrix

Measure stability of classifier across time series

Use ROC curve to find appropriate level of sensitivity

Measuring business metrics

1. **Improve on the model's performance (using any available levers, including but not limited to data sources, methods used, process to create the model, staff, etc.)**

Ways to improve model:

* Understand AUC
* Improve model parameters
* Different classification models
* Boosting/Bagging/Stacking

Other ideas:

* New data acquisition
* Other ideas

\* All code notation within this document is in python

1. **Evaluate this model's current performance:**

Question to Answer: will the lessee be delinquent on lease payments time-t in the future?

Dependent variable: (0,1) where:

0 will not be delinquent

1 will be delinquent

Data:

X – independent variable matrix

y – vector of dependent variables (0,1)

For each period:

Ho = 0; Lessee not delinquent in time t

Ha = 1; Lessee delinquent in time t

**Null Accuracy: Probability of being correct simply by guessing the most likely outcome?**

max(y.mean(), 1- y.mean()))

The null accuracy establishes a baseline for simply guessing the most likely outcome. If 5% of the lessees become delinquent within time t, then guessing 0 every time will yield 95% accuracy and 5% error. While a prediction model might be 90% accurate, it’s useless when compared to the null accuracy.

**In-sample accuracy: is this model predictive on in sample data?**

Train on in sample data, predict on in sample data. Comparing the output of the in sample accuracy to the null accuracy will show if the model is predictive in sample. If the in sample accuracy is 70% and the null accuracy is 95%, then that means we are better off simply guessing the most likely outcome. If this occurs, we should stop and examine the model, data and set up.

Below, we assume using Logistical Regression as the classifier:

from sklearn.linear\_model import LogisticRegression

from sklearn import metrics

Logreg=LogisticRegression()

logreg.fit(X,y)

y\_pred\_is=logreg.predict(X)

print('In Sample Acc:', metrics.accuracy\_score(y, y\_pred\_is))

**Prediction accuracy: is this model predictive in out of sample?**

Most basic way is to train/test-split, where a portion of the matrix X and vector y is segregated into training sets (X\_train, y\_train) and test sets (X\_test,y\_test). After the model is fit to the training set, the predicted y (y\_pred) is predicted by feeding the X\_test. y\_pred is then compared to y\_test for accuracy.

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y)

logreg=LogisticRegression()

logreg.fit(X\_train,y\_train)

y\_pred=logreg.predict(X\_test)

print('Train/Test Split Accu:',metrics.accuracy\_score(y\_test, y\_pred))

print('True:', y\_test.values[0:25])

print('Pred:', y\_pred[0:25])

Our goal is to maximize the generalize out of sample prediction ability using the least complex model. If the out of sample prediction is 25% and the in sample prediction is 95%, that means we’ve simply overfit the model, which creates false sense of accuracy.

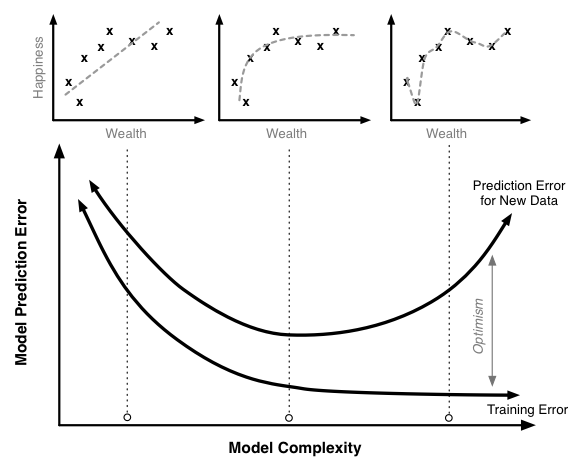


Figure 1: Source: http://scott.fortmann-roe.com/docs/MeasuringError.html

**Summary:**

The **null accuracy** measures the natural probabilistic binary outcome.

The model’s **in sample accuracy** should be better than the null accuracy.

The model’s **out of sample accuracy** should be similar to the in-sample.

1. **Evaluate the structure/integrity of the model/methods used:**

We introduce a confusion matrix, which describes the attribution of the classification accuracy.

For each period, hypothesis:

Ho = 0; Lessee not delinquent in time t

Ha = 1; Lessee delinquent in time t

**Confusion Matrix:**

|  |  |  |
| --- | --- | --- |
|  | **Ho True** | **Ha True** |
| **Accept Ho** | True Positive (TP) | Type 2 Error (FN) |
| **Accept Ha** | Type 1 Error (FP) | True Negative (TN) |

TP – predicted no default, lessee did not default

FN – predicted no default, lessee default

FN – predicted default, lessee did not default

TN – predicted default, lessee default

Precision = TN / (TN + FN)

Recall = TN/(TN+FP)

True Positive Rate (TPR) = TP/ (TP + FN) (sensitivity)

False Positive Rate (FPR) = FP/ (FP + TN) (1-specifity)

True Negative Rate = TN/(FP + TN) (specificity)

from sklearn import metrics

confusion = metrics.confusion\_matrix(y\_test, y\_pred)

TP = confusion[1, 1]

TN = confusion[0, 0]

FP = confusion[0, 1]

FN = confusion[1, 0]

**Rolling Confusion Matrix:**

We calculate the out of sample metrics, using a rolling training period and t forecast period. For example, sample data are presented below for a model that uses a rolling training period of 1 year and then forecasting a delinquent lessee 1, 2, and 4 weeks out:



Due to the asymmetric downside cost of delinquent lessees versus the profitability of the lessee, the most important metric in delinquency prediction is **sensitivity**, since false positives (non-delinquent lessees predicted as delinquent) are more acceptable than false negatives (delinquent predicted as good standing). In order to achieve the necessary sensitivity level, we will need to adjust the threshold.

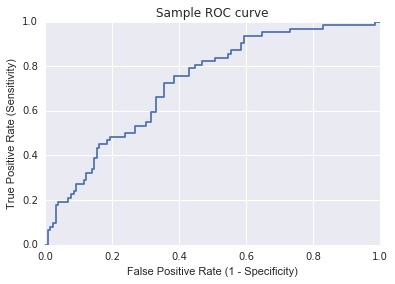
**Threshold and ROC curve:**

In order to improve the **sensitivity**, we will need to accept an increase of false positives due to an increase of type 1 errors. In the case of lessee delinquent detection, we will accept this tradeoff.

In order to control the **sensitivity,** we adjust the threshold in the classification model. A classification model uses an underlying probability to predict the classifications; if the probability for 1 is above a threshold of 50%, the classifier chooses 1, otherwise 0 is predicted. In this scenario, we can lower the threshold to 20%, and the classifier will choose 1 for any probability above 20%. Intuitively, while we will improve the sensitivity, we will likely end up with more type 1 errors.



We can compute both sensitivity and false positive rates at various threshold levels, and then plotting the relationship, called a ROC curve. Using the ROC curve, we can select the desired level of sensitivity.



In the above example, if our business requires a sensitivity of 80%, we must accept a 41% probability of a false positive.

**Measuring Business Impact:**

In measuring business impact, we look at a simplified framework for estimating

* Measured weekly due to weekly lease payments.
* If our credit model predicts the lessee will become delinquent, we will immediately repo the vehicle. We estimate that it will take 2 weeks to repo the vehicle, and then redeploy to a new lessee. (TN)
* If our credit model does not predict the lessee will become delinquent and the lessee misses a payment, we estimate it will take a total of 4 weeks to recognize delinquency, repo the vehicle, and redeploy to a new lessee. (FP)

Symbology:

R = Revenue, weekly

C = Cost, weekly

TP – True Positive

TN – True Negative

FP – False Positive

FN – False Negative

Measurements:

Weekly profit without forecast:

Profit = (TP + FN) \* (R-C) – 4\*(FP + TN)\*C

Weekly profit with forecast:

Profit = TP\*(R-C) – 2\*(TN+FN)\*C – 4\*FP\*C

Opportunity Cost:

Opportunity Cost = 2 \* FN \* (R-C)

Total Savings From Correct Identification:

TN\*C\*(4-2)

Total Savings:

Total Savings from Correct ID – Opportunity Cost

= TN\*C(4-2) – 2\*FN\*(R-C)

If the classifier adds value, then the total savings should be positive.

**Summary:**

* The confusion matrix measures attributes of the predictability
* Using the confusion matrix, time series analysis measuring model stability across multiple t intervals.
* Lowering the threshold improves sensitivity while increasing type 1 errors. Find allowable level using ROC chart.
* Translate statistical measures into business metrics.

1. **Improve on the model's performance (using any available levers, including but not limited to data sources, methods used, process to create the model, staff, etc.)**

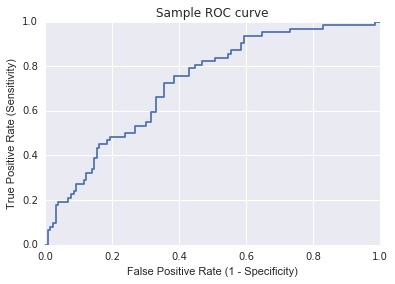
In general, model performance is evaluated comparatively using out-of-sample testing area under the curve (AUC) of the ROC plot. The higher the AUC, the better the model.

AUC

Worthless

Line

Model Improvements



**Improving Existing Model Parameters:**

* K-Fold cross validation – divide sample data into K sections, choose 1 section as the test set, and train on others. Repeat to see distribution.
* Grid Search Cross Validation – Create exhaustive set of model input combinations for test scoring accuracy. IE Combination of 20 neighbors, and 2 different weighting schemes.
* Randomize Search Cross Validation – Random sampling of parameters using GSCV due to resource / time constraints.
* Holdout Dataset: Regularly collect golden copies of data not used for model calibration. Usage for model calibration.

**Different Classifiers:**

Run similar tests as in section 2 for different classification models:

* Logistical regression
* K-Nearest Neighbors
* Support Vector Machines
* Decision Trees
* Random Forest
* Ensemble Classifier

**Other data sources:**

* Anonymized data from banking partner to understand factors that cause revolving creditors to default.
* Partner with social media sources, collect sentiment data

**Other ideas:**

* Upload data to open source and create competitions

**Sources:**

Fawcett, Tom (2004). ROC Graphs: Notes and Practical Considerations for Researchers, HP Labs.

Ieong, Samuel (2006). Probability Theory Review for Machine Learning, Stanford.

Florentin Butaru et al (2015). Journal of Banking and Finance: Risk and risk management in the credit card industry. 72, 218-239