2016

An Analysis of Accuracy using Logistic Regression and Time Series

Edwin Baidoo  
*Kennesaw State University*

Jennifer L. Priestley  
*Kennesaw State University*, jpriestl@kennesaw.edu

Follow this and additional works at: [http://digitalcommons.kennesaw.edu/dataphdgreylit](http://digitalcommons.kennesaw.edu/dataphdgreylit)

Part of the [Statistics and Probability Commons](http://digitalcommons.kennesaw.edu/dataphdgreylit)

**Recommended Citation**  
[http://digitalcommons.kennesaw.edu/dataphdgreylit/2](http://digitalcommons.kennesaw.edu/dataphdgreylit/2)

This Article is brought to you for free and open access by the Ph.D. in Analytics and Data Science Research Collections at DigitalCommons@Kennesaw State University. It has been accepted for inclusion in Grey Literature from PhD Candidates by an authorized administrator of DigitalCommons@Kennesaw State University. For more information, please contact digitalcommons@kennesaw.edu.
An analysis of Accuracy using Logistic Regression and Time Series

Edwin Baidoo, Ph.D Student in Analytics and Data Science
College of Science and Mathematics
Kennesaw State University

Jennifer Lewis Priestley, Ph.D
College of Science and Mathematics
Kennesaw State University

Abstract—This paper analyzes the accuracy rates for logistic regression and time series models. It also examines a relatively new performance index that takes into consideration the business assumptions of credit markets. Although prior research has focused on evaluation metrics, such as AUC and Gini index, this new measure has a more intuitive interpretation for various managers and decision makers and can be applied to both Logistic and Time Series models.

Index Terms—Binary, Classification, Credit Scoring, Logistic Regression, Time Series

I. INTRODUCTION

A primary concern for any lending institution is that of default. Accurate assessment of defaults and risk are particularly complex, when one factors in the different forms of financial credit instruments that have gained popularity over the last two decades, including the various forms of credit securitization instruments (Asset Backed Securities, Credit Default Swaps, etc). The figure below shows the general trend of consumer credit between 2006 and the first quarter of 2016.

Fig. 1: Consumer Credit

The primary question therefore, for many lenders is this: how can default risk be supervised or at least attenuated? Without resorting to insurance techniques (like risk sharing and risk avoidance), there exists two basic tactics or screening devices lenders employ: interest rate and the terms of the contract.

Stiglitz and Weiss (1981) concluded that these mitigation methods are not sufficient because the interest rate can lead to the adverse selection effect, while the terms of the loan may induce the borrower to take actions not in the interest of the lender. The authors found that given these two screening devices, there may exist an equilibrium state of credit rationing.

The other approach of tackling credit default risk is through statistical modeling or more specifically through Credit Scoring. The goal of Credit Scoring is to estimate or predict the probability of credit default, assuming that credit has been extended to an applicant. In the course of application of the model, all applications are scored and a decision to reject or accept is made based on an established cutoff value (Verbraken, et al., 2014). This is usually followed by performance measurement analysis to determine the robustness of the designed model.

The aim of this paper is twofold: to explore the use of logistic regression and time series analysis within the framework of the analysis of prediction accuracy of credit default. Secondarily this paper will establish the importance of a profit-based classification measure as a viable alternative to many performance measurement KPIs. The literature review, empirical analysis and conclusions will follow.

II. LITERATURE REVIEW

The literature for default prediction - or more specifically, for generic binary classification - has grown substantially over the last 40 years. Altman (1968) and Beaver (1967) analyzed the concept of business failures and customer defaults through the use of financial ratios. Using Multiple Discriminant Analysis (MDA) on a set of 66 firms they were able to predict that 33 of them failed while 33 did not.

For many years MDAs were the model of choice, until it became increasingly clear that some of the model assumptions
were being violated. Also, the standard coefficients did not lend themselves easily into the interpretation of slopes in the classic regression equation. In response to this issue, Ohlson (1980) introduced the conditional logistic model. Since then, the most common statistical model used in credit analysis has been the logistic regression (Nargundkar et al. 2004). Given a set of independent variables $x_i \forall k \in N$, the dependent variable $y_i$ can only assume a limited possible outcomes $i \in \{0, 1\}$. Typically, $y_i$ would indicate the occurrence of some random event - in this case the event of a default - where $y_0$ would suggest the opposite. From here, a model is derived such that for $n$ observations the likelihood function below is maximized.

$$\prod_{i=1}^{n}(p_i^{y_i}(1-p_i)^{(1-y_i)})$$  \hspace{1cm} (1)

where $p_i$ corresponds to the probability that $y_i = 1$ (Finlay, 2009)

Recently with the advent of faster computing power, there has been a surge of newer techniques that has been developed and refined. These include Decision Trees, k-Nearest Neighbor, Neural Networks, Clustering Analysis and even Genetic Algorithms. According to Nargundkar et. al. (2004), the varying assumptions underlying each technique may also lead to different results even when the same datasets are being scrutinized.

Frey et. al (2001) addressed credit default in the settings of financial portfolio optimization through the concept of copulas. They argued that a default will occur if certain latent variables are dependent. For example, if the borrowers asset falls below the threshold of the loan amount. In this case, the correlation matrix of the latent variables can be constructed from factor models. They found that the dependence structure of latent variables can determine joint default probabilities for a group of borrowers.\(^1\)

### III. ANALYSIS OF PERFORMANCE MEASURES

#### A. Background

Assuming that there are two class labels $i \in \{0, 1\}$, and $p_i$, the classification process works as follows:

1. A classification rule first assigns a score $s(x)$ to all applications based on a transformation of the probabilities.
   - $f_i(s)$ represents the probability density function (PDF)
   - $F_i(s)$ represents the cumulative distribution function (CDF) of the scores, where of-course $\sum_{i=0}^{1} p_i = 1$ and $0 \leq t \leq 1$ represents classification threshold

2. A decision is made based on $t$
   - If $s > t$ the observation is classified as belonging to class 1

\(^1\)For a thorough treatment of the modeling techniques please see (Lessman et. al. 2013)

- If $s \leq t$ the observation is classified as belonging to class 0

#### B. Overview

While many classification techniques exists, its just as important, if not more, to have KPIs that correctly measures the performance of the proposed model. To this end, the fundamental assumptions underlying a model are often questioned by the researcher to further ascertain the right tool to use.

One of the most widely used tool to measure classifier performance is the Area Under the Receiver Operating Characteristics Curve (AUC) and the related Gini coefficient. It was originally used for signal detection to demonstrate the trade-off between false and accurate predictions (Fawcett, 2005).

To fully maximize the benefit, the AUC is often analyzed in context of the confusion matrix or the contingency table. A contingency table can be constructed in the following manner.

<table>
<thead>
<tr>
<th>Predicted Good</th>
<th>True Good</th>
<th>True Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Good</td>
<td>$p_0 F_0(t)$</td>
<td>$p_0 (1 - F_0(t))$</td>
</tr>
<tr>
<td>Predicted Bad</td>
<td>$p_1 F_1(t)$</td>
<td>$p_1 (1 - F_1(t))$</td>
</tr>
</tbody>
</table>

A useful benefit of the matrix is that other metrics can be directly derived.\(^2\) For example:

- sensitivity = $F_0(t)$\(^3\)
- specificity = $(1 - F_1(t))$\(^4\)

The AUC therefore can be constructed where $F_0(t)$ is the familiar y-axis and $F_1(t)$ corresponds to the x-axis. One of its attribute is that its insensitive to the changes in the distribution of the classes (Fawcett, 2005). The AUC represents the average sensitivity - assuming that all values for the specificities are equally likely.

The AUC can be defined as follows (Hand, 2009)

$$\text{AUC} = \int_{-\infty}^{\infty} F_0(s) F_1(s) ds$$ \hspace{1cm} (2)

According to Fawcett (2009) the AUC has been found to be related to the Gini coefficient by the formula

$$\text{Gini} + 1 = 2(\text{AUC})$$

The concept of misclassification is equally as important. In the confusion matrix above, each cell has both a cost $\pi_i$ and a probability of misclassification $\beta_i$ associated with correct

\(^2\)For a comprehensive list of metrics from the confusion table, the reader is advised to see Fawcett(2005)

\(^3\)The sensitivity is defined as the probability corresponding to the event of True Good being predicted as good

\(^4\)The specificity is defined as the probability corresponding to the event of True Bad being predicted as bad
or incorrect classification. A loss function \( L \) can therefore be constructed as follows (Nargundkar et al., 2004)

\[
L = p_0 \pi_0 \beta_0 + p_1 \pi_1 \beta_1
\]  

(3)

The equation above works well when the misclassification parameter is equal across all cells, which raises an obvious issue when a researcher tries to model a behavior which has various forms of classification costs. In other words, where each outcome cell of the confusion matrix has a different economic impact.

C. The Expected Maximum Profit: A profit based performance measure

Hand (2005) demonstrates that indeed the traditional KPIs such as the AUC measure, the Gini Coefficient and the KS statistic may not be appropriate in many situations for the very reason that the business aspect is often overlooked.

Similar sentiments echoed through recent studies suggests (Finlay, 2009; Bravo et al., 2012; Verbraken et al., 2014) that sufficient attention should be directed to the business aspect of the topic rather than treating it as a purely statistical or machine learning exercise. When both is combined, the empirical evidence now becomes thus: the performance measure should be able to capture the profit of the firm, along with model specificities - this leads to the proposed Expected Maximum Profit (EMP) measure for credit analysis.

The EMP measure starts with the question of lender profit per each loan borrowed. For lenders to fully understand and maximize such profit, the model should fully incorporate costs and benefits associated with each classification. Let the costs and benefits be represented as \( \pi_i \) and \( b_i \) respectively. Then, the average classification profit per borrower is defined as:

\[
P(t; b_0, \pi_1, \pi^*) = (b_0 - \pi^*) p_0 F_0(t) - (\pi_1 + \pi^*) p_1 F_1(t)
\]  

(4)

In this way, maximizing the above equation on the cut-off value of \( t \) gives the maximum profit measure, MP defined as:

\[
MP = \max\{P(t; b_0, \pi_1, \pi^*)\}
\]  

(5)

The value \( t \) that maximises MP will be denoted by \( T \) and referred to as the optimal cut-off value. It satisfies the following equation.

\[
\frac{f_0(T)}{f_1(T)} = \frac{p_1 (\pi_1 + \pi^*)}{\pi_1 (b_0 - \pi_0)}
\]  

(6)

According to Verbraken et al (2014), the MP on its own can be used as a performance measure with the visible advantage that it can help select the model with the highest profit. Additionally, the straight-forward nature of the cut-off value makes it appealing. The Expected Maximum Profit Measure (EMP) is defined as follows:

\[
EMP = \int_{b_0}^{T} \int_{\pi_1}^{\pi^*} P(T(\theta); b_0, \pi_1, \pi^*) \cdot h(b_0, \pi_1) d\pi_1 db_0
\]  

(7)

where

- \( h(b_0, \pi_1) \) is the joint probability distribution of the cost and benefit
- \( \theta = \frac{(\pi_1 + \pi^*)}{(b_0 - \pi_0)} \)

To empirically put the EMP to use, the following parameters \( (b_0, \pi_1, \pi^*) \) and \( h(b_0, \pi_1) \) will need to be defined. The following methodology expanded by Verbraken et.al (2014) will be used to calculate each of them. First, the parameter \( b_0 \), representing the benefit of correctly identifying a "bad" borrower can also be interpreted as the percentage of the loan amount that cannot be recovered, given default.

\[
b_0 = \frac{LGD \cdot EAD}{P} = \lambda_i
\]  

(8)

where \( 0 \leq \lambda_i \leq 1 \) and

- \( LGD \) is the loss given default
- \( EAD \) is the exposure at default
- \( \lambda \) is the principal of the loan

The parameter \( \pi_1 \) is the cost of incorrectly classifying a non-defaulting borrower. This can be thought of as the return on investment (ROI) on the loan that was foregone (an opportunity cost). For a principal \( P \) and interest rate \( i \) along with maturity \( M \) the following definition exists for the ROI:

\[
ROI = \frac{i M}{1 - (1 + i)^{-M}} - 1
\]  

(9)

Where both \( M \) and \( i \) are constant for a given period. In some applications in finance, \( i \) is thought to be generated by some stochastic process that can also be modeled, with its own distributional properties. However, in this context \( i \) has been shown to have little to no variation.

The parameter \( \pi^* \) measures the cost of the action. In this case, when a borrower is rejected, there is no additional costs generated, other than the fixed costs of building the model (at least fixed in the interim). For this reason, it can be safely omitted (Verbraken et al., 2014). That is, \( \pi^* = 0 \)

Finally, regarding the probability distribution \( h(b_0, \pi_1) \). The cost \( \pi_1 \) has been established as being constant over time. However, \( \lambda_i \) may assume different distributions because for a given default, recovery rates ranges from 0% to 100% of the principal on the loan. Therefore at best, an empirical distribution \( H(\lambda_i) \) will be constructed for a given \( \lambda_i \) or a family of \( \lambda_i \).

As an example, \( \lambda_i \) usually falls in three categories:

- \( \lambda = 0 \) with probability \( \gamma_0 \), where the borrower redeems the defaulted loan, with a recovery rate is 100%
- \( \lambda = 1 \) with probability \( \gamma_1 \), where the lender recovers 0% of the defaulted loan
• \( \lambda \) follows a uniform distribution with \( H(\lambda) = 1 - \gamma_0 - \gamma_1 \)

Clearly the assumption of uniformity is made to simplify calculations. Factors affecting borrower repayment can be quite complex, with many variables spanning personal and macro-economic behaviors. Given the points above, the EMP measure can be simplified as follows:

\[
EMP = \int_0^1 P(T(\theta); \lambda, ROI) \cdot h(\lambda) d\lambda \quad (10)
\]

where

\[
P(t; \lambda, ROI) = \lambda \cdot p_0 F_0(t) - ROI \cdot p_1 F_1(t) \quad (11)
\]

\[
\theta = \frac{ROI}{\lambda} \quad (12)
\]

Theoretically \( \theta \) can range from \( ROI \) to \(+\infty\) for \( 0 \leq \lambda \leq 1 \).

The problem it presents is that the integration of the EMP will not consider, as its domain, the entire plane of the ROC curve because the curve stretches from the origin \((0,0)\) to \((1,1)\). This means that an EMP measure can give a different result for the AUC of the ROC.

IV. METHODOLOGY

For this experiment, a logistic regression and a time series model were constructed. The reason for this selection is that logistic regression has been established to be the foundational model for credit scoring analysis (Nargundkar et. al., 2004). However, the time series model was chosen largely because of the time component of the data. Also because credit analysis has components that may have seasonal, trend or cyclical elements that may be better explained using the tools available in time series.

Traditionally, a time series model is constructed to depict or describe some form of a univariate stochastic behavior. However, it can be safely stated that credit analysis is not univariate in the sense that multiple variables contributes to default. Therefore a multivariate time series model can also be constructed. In this way, time series in general is flexible enough to handle such changes.

For the empirical aspect of this paper, a total of 36 separate datasets were used, granted by a large credit scoring agency. Each dataset represented identical quarterly consumer information from 2006 to 2014. On average, each dataset had over 11 million observations and 305 explanatory variables in the following categories:

- Consumer Non-Financial Accounts
- Consumer Telecommunication Accounts
- Consumer Utility Accounts
- Consumer Service Accounts
- Consumer Industry Accounts
- Consumer Liabilities
- Consumer Liens

A. Data Exploration

Many techniques aide in the selection of relevant features within a dataset. Often, using a combination may optimize variable selection particularly when the data exhibits high dimensionality and/or cardinality. For this analysis a combination of the traditional variable clustering and the Augmented Backwards Elimination (ABE) method was used.

The third quarter dataset of 2006 was chosen for exploration and cleaning purposes before extending similar scrutiny to the remaining dataset. The next step was to select a response variable. Since the approach of this paper was to incorporate element of time series analysis, the motivation was to select a variable that can easily extend such properties. Using SAS to provide basic statistical plots (bar graphs, etc), the dependent variable chosen was the number of new non-financial account per every three months - see figure below.

![Distribution of New Non-Financial Account opened within the last three months](image)

During the process, it became easily apparent that there exists significant amount of missing and coded observations across all dataset (The target variable shown above shared the same pattern). For example more than 50% of the observations in the target variable were coded and therefore deleted.
The intention was to use the complete observations within the target variables as the basis to eliminate other variables with significant missing information. To this end, factors missing more than 75% of information were deleted as well. It should be noted that such deletion simply addressed the issue of excessive missing information rather than the coded values. The figures below show variables with more than 75% missing information.

The figures below show variables with more than 75% missing information.

Fig. 3: Highest Non-Financial Account Limit Reported in Last 3 Months

Fig. 4: Total Number of Utility Accounts 2+ Cycles Past Due or Charge-Off Currently

The distribution of these variables represents a cross sectional representation of other variables within the dataset. It could not be inferred that one variables distribution represented the behavior of the entire class. For example, it could not be deduced that the distribution in figure 3 (Highest Non-Financial Account Limit Reported in Last 3 Months) can be extended for all Non-Financial Account variables. The same could also be said for Utility Accounts in figure 4. However, it is apparent, through several exploratory plots that the variables, in any random category exhibit such distribution for missing and coded observations.

This revealed the extent to which missing and coded observations plagued the dataset. It also provided a solution on which method of imputation could potentially be prescribed. For example, by looking at the variables above, a mean imputation could significantly reduce the amount of variation expected in the dataset. The insight gained from this influenced the method of imputation, which is discussed further in the next section.

Next, a single master file was created by merging all 36 datasets, resulting in 85 million observations and 175 variables. It should be noted that the file still had incomplete observations and coded values. The figure below shows the missing data pattern inherent in the file for a subset of 800,000 observations and five variables. The main contribution of this is to provide further insight into the pattern of missing observations. For example, it can be seen that the percentage of telecommunication accounts charged off within the last 24 months, makes up the majority (27.33%) of the missing observation. This raises an obvious question of the contribution of this variable within the wider scheme of default prediction, especially when the variable contains information relating to a charged-off account.

Fig. 5: Missing Data Pattern

B. Data Cleaning And Imputation

To avoid a static univariate imputation over the entire master file, the approach used was to reduce the dimension by using variable clustering. Doing so yielded 60 variables with approximately 94% of the total variation retained within the dataset.

Another dimension reduction technique, the Augmented Backwards Elimination (ABE) method was used. It builds upon the idea of “purposeful variable selection” proposed by Hosmer (2013). ABE combines the power of the backwards elimination method, along with the change-in-estimate criterion associated with a given statistical significance (alpha). In this way, non-significant variables will be kept if

For a detailed analysis the reader is referred to http://www.meduniwien.ac.at/user/georg.heinze/abe/techrep.pdf
their removal causes drastic shifts in the parameter estimates. This method works best if some type of variable pre-screening has already been done (in this case, clustering analysis). In the end, the method selected 12 variables, from which any analysis can be done.

For imputation, the Markov Chain Monte Carlo Method (MCMC) was used in SAS to impute missing cases. The method assumes that variables within the data have random and arbitrary missing patterns. It further assumes that the variables have a joint multivariate normal distribution. It therefore fills in the missing data by selecting from the conditional normal multivariate distribution via Markov Chains [13].

V. MODEL DEVELOPMENT AND COMPARISON
The primary model used for analysis was the logistic regression. From the 12 variables, the following estimates and their p-values are displayed below.

![Fig. 6: Estimates for Logistic Model](image)

Also, the analysis of the odds ratio gives the following

![Fig. 7: Odds Ratio for Logistic Regression](image)

Staying true to the objective of the paper, the analysis of the accuracy of the model is also given below.

![Fig. 8: Logistic accuracy model](image)

As it can be seen, the percentage of those classified correctly is 92.31%. From the table, other measures of classification and mis-classification can be deduced. Using the same analysis, it can be seen that the logistic analysis far outperforms the time series.

![Fig. 9: Time series accuracy model](image)

In building this model, five variables were selected based on their correlation with the dependent variable, as shown below. As stated above, the time series model can be flexible enough to incorporate another variable.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Correlation with Dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td>totNFA2CPDC3mon</td>
<td>0.561</td>
</tr>
<tr>
<td>totIA3CPDC12mon</td>
<td>0.558</td>
</tr>
<tr>
<td>NoSAbalance3mon</td>
<td>0.501</td>
</tr>
<tr>
<td>NoNFA3mon</td>
<td>0.499</td>
</tr>
<tr>
<td>NoIAc12mon</td>
<td>0.496</td>
</tr>
</tbody>
</table>

The performance analysis of the time series model is also shown below:

![Table of time series accuracy model](image)
VI. CONCLUSION

The EMP measure shows promising signs of being a key performance index. The straightforward interpretation lends itself to greater use and scrutiny. It incorporates factors such as the recovery rates and attempts to construct a distribution of repayment factors.

Because it directly incorporates profit analysis per classification, it makes it easier to use as a decision making tool not for analysts but also for executive management as well. A possible next step will be to test it empirically and compare the results against common benchmarks.

The time series model can be carried out further where each classification bucket of 0s and 1s can be analyzed. In this context, segmenting the good vs bad can even give clearer insight as to where to push a certain product for a specific season.

VII. REFERENCES


3) Thomas Verbraken Development and Application of consumer credit scoring models using classification measures


5) Hand, David. "Measuring Classifier Performance: A coherent alternative to the area under the ROC curve."


10) Edward, Altman and Gabriele, Sabato. "Modelling Credit Risk for SMEs: Evidence from the US Market."


