Project\_Part 3 - Executive Summary

Mark Kuether

April 28, 2020

## Section 1: Executive Summary

### Predicting Loan Success using Statistical Models

#### (Prepared for the Bank of Elbonia)

Based on data provided by the Bank of Elbonia, loss due to defaults represents an 89% reduction in potential profit from successfully repaid loans. This gap represents a large opportunity for improvement, and can be reduced by using a logistic statistical model for predicting loan success.

**Findings**

Using the test data provided by the Bank of Elbonia, we applied a logistic statistical model to predict successful and defaulted loans. With this application on the test data, we were able to see a profit increase of 163% over the current profit. This reduced the loss due to defaults from 89% to 72%.

**Recommendations**

We recommend the Bank of Elbonia use this model, optimized for profit, as an initial prediction model for their loan applications. While this will reduce the overall number of loans they accept, it should also reduce the number of loans that are likely to default. Default loans will still occur, but the profit from loans overall will increase.

The Bank of Elbonia will also need to continue developing and adjusting models based on the steps reviewed in this report. Future models may test the inclusion of additional data not used in our model, or change due to changing economic health of its loan customer base.

**Background**

This model predicts whether a specific loan application would result in a successfully repaid loan. This is independent of the grading system used by the bank, so the grades were not used in this analysis. This allows the bank to use this model without dependency on the other grading system.

The model was developed with data provided by the Bank of Elbonia. Active loans were removed from the data since the objective was predicting the end status of the loans. The remaining data was classified into two categories between loans fully paid, and loans that were in default or charged off. Logistic models work with binary results, such as “Yes” vs “No”, by generating a probability that a specific case belongs to a desired category. The bank of Elbonia can use this to identify loans that are likely to default and deny those applications. The model can also be tuned for maximum profit by adjusting what probability level corresponds to the split between the two categories.

## Section 2: Summary

Based on the data in this data set, about 22% of loans default. This percentage of defaults resulted in a loss of 89% of potential profit for the bank. Reducing the percentage of defaulted loans is critical to ensure a bank remains competitive in a modern market.

The first task involved with reducing this percentage of defaults is understanding the underlying factors affecting the value. To this end, the bank collects a large number of details for each loan applicant and the details of each loan. The bank has provided a subset of that larger data set. This subset contains the key factors from their larger set of details, as well as the resulting status of the loans. We are using this subset of data to further refine which factors are key to predicting whether a loan applicant will repay or default on the loan. The result of the loan is reflected in the “Status” field, which is our response variable. The initial data set contained 50,000 records with 32 variables. Below is a summary of the steps used to prepare the data for analysis.

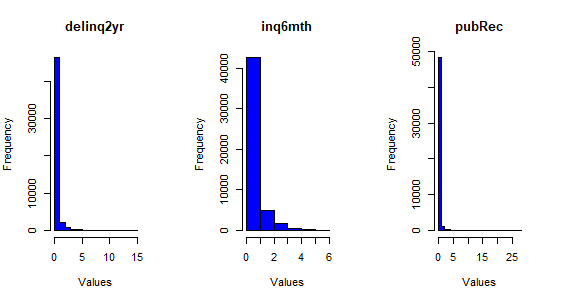
## Section 3: Preparing and cleaning the data

**Removing Fields from Raw Data**

The first step is to simplify the data where able. We start by removing any fields we feel certain would not contribute to the creation of our model. The following fields were removed from the initial data set:

* The loan ID field is not necessary to predict loan performance and provides no information for this analysis. This field was dropped from the data set.
* After reviewing the values of the summary data, the “delinq2yr”, “inq6mth”, and “pubRec” fields appeared to have mostly 0 or very low values. A histogram was used to confirm this (see below). These fields were dropped from the data set.
* The “employment” column contained 21401 levels. Breaking this into sub-categories effectively is not feasible for this analysis. This field was dropped from the data set.
* The “payment” field is a direct calculation of the amount, term and rate. Since that field is redundant, it was removed from the data set.
* The “grade” field is a loan health indicator based on the bank’s own criteria. This may present a confounding variable for our analysis, and was removed from the data set. Removing this also allows the bank to remove or replace their grading algorithm without affecting this model.

Removing these fields reduced the number of variables in the dataset from 32 to 25.



**Recoding Categories** In order to analyze variables for predicting successful loans, the “status” field values were recoded. Values of “Fully Paid” were recoded as “Good”. Values of “Default” or “Charged Off” were recoded as “Bad”. Since we are only interested in predicting future loans based on previously successful or defaulted loans, all rows not containing “Good” or “Bad” were dropped from the data. This reduced the number of records from 50,000 to 34,655.

After recoding and removing data, the data set was found to contain 27074 “Good” loans and 7581 “Bad” loans for a probability of success of 78%, and a probability of default to be 22%.

The “reason” field has 11 categories which makes it difficult to analyze. These were reduced to the following categories:

* Inv -> Expenditure which may improve value. These included “home\_improvement”, “house”, “small\_business” and “renewable\_energy”.
* Debt -> Expenditure to cover debt. These included “credit\_card” and “dept\_consolidation”.
* Cons -> Expenditure on consumer good with no value added. This includes “car”, “major purchase”, and “vacation”.
* Life Exp. -> Single or Life event which may represent a single cost. This includes “medical”, “moving”, and “wedding”
* Other -> This reflects the “other” category within the data set.

The states were grouped by the regions defined by the Bureau of Economic Analysis (<https://apps.bea.gov/regional/docs/regions.cfm>).

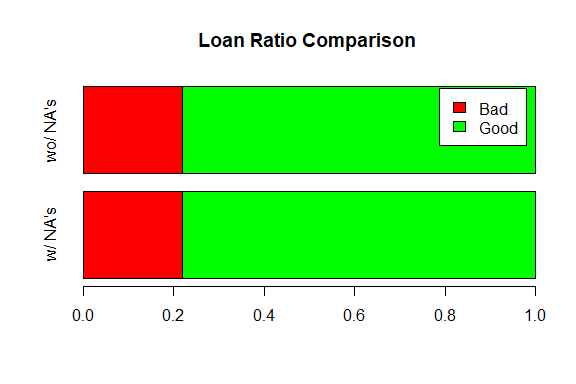
Inside the “verified” field, some records showed “Source Verified” which has the same meaning as “Verified”. Those records were combined with the value of “Verified”

The “length” field for the length of work was divided into segments of 0-4 years, 5-9 years, 10+ years. The fields containing NA were recoded as “Not Empl” and may either represent unemployed, retired, or temporary positions. This field was retained in the data set because it appeared to have a significant impact on the loan status.

**Handling Missing Values**

After recoding the categories, most missing data was found in the two fields of “bcRatio” and “bcOpen”. Comparing these rows to other bc fields did not show a relationship to allow the values to be imputed. There was not enough information in the data set to determine the cause or meaning of the missing data. This meant that we could either avoid using these fields, or we could remove the records with the NA values. Our next step was to determine if removing these records would cause a change in the ratio of “Good” loans.

The proportion of “Good” claims was calculated for the initial data set, which included records with missing values. The same proportion was calculated without the records containing those specific N/A values. The difference in the two proportions was calculated to be 0.00007024. Since removing these fields did not have an adverse effect on the overall proportion of “Good” claims, the records containing these missing values were removed. This reduced the data set from 34655 records to 34271.



In addition to these, there was a single record found where most values, including the loan amount, were missing. This single record was removed with the other bcRatio and bcOpen records.

## Section 4 - Exploring and Transforming the Data:

Reviewing the data after this work showed that the categorical fields still contained formerly removed levels. In addition, all categorical fields contained an empty category, which may have been a byproduct of the import process. A script was run on all categorical fields in the data set to remove un-used categories. This used the command “droplevels”.

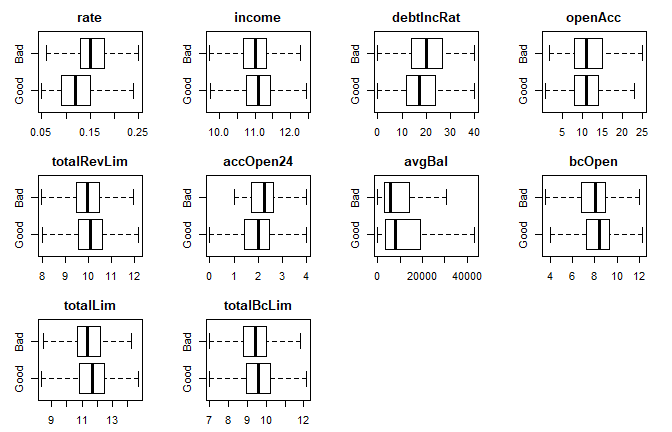
The numerical fields were analyzed to see if they could benefit from transformations for further analysis. A script was run to graph the histogram and boxplot for all remaining numerical fields to visually check on the distribution. The fields for “income”, “totBal”, “totRevLim”, “totalLim”, “totalRevBal”, “totalBcLim”, and “totalIlLim” were found to have a heavy right skew.

Another script calculated the log, square root, cube root, and inverse values of these fields. The script then graphed all transformed distributions with a histogram and boxplot. These graphs showed which transformations produced a distribution that was less skewed. The “income”, “totBal”, “totalLim”, “totalRevBal”, “totalBcLim”, and “totalIlLim” fields appeared to benefit most from a logrithmic transformation. The “totalRevLIm” benefited from a square root transformation.

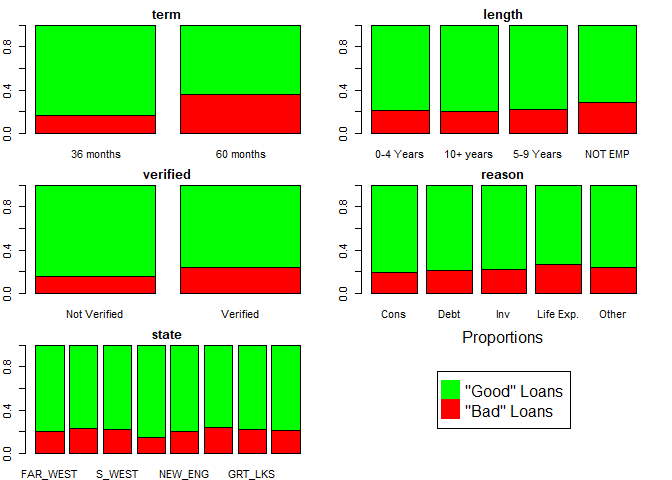
A final script then transformed those fields as determined from the earlier analysis.

**Data Analysis - Numerical Distributions:**

The initial analysis involved side by side box plots on all numerical fields comparing the distributions between “Good” and “Bad” loans. A script graphed the plots for initial visual review. Because of the high number of outliers, boxplots were made without the outliers. Boxplots with the lower quartile, median, and upper quartile measures offset between “Good” and “Bad” loans were chosen as variables to consider for the equation. The distribution of “Good” vs “Bad” loans appeared to be affected by the following variables: “rate”, “income”, “debtIncRat”, “openAcc”, “totalRevLim”, “accOpen24”, “avgBal”, “bcOpen”, “totalLim”, and “totalBcLim”



A similar procedure was performed with the categorical variables. Categories which showed an offset of the ratio between “Good” and “Bad” loans between specific categories were chosen as variables to consider for the equation. Reviews of these proportions showed that the following variables appeared to have an effect on the loan status: “term”, “length”, “verified”, “reason”, and “states”.



## Section 5: The Logistic Model

To create a logistic predictive model, we start by generating a model training data set, and a model testing data set. The data set was randomly divided with 80% forming the model training set, and 20% used as the model testing set. All results derived from the model reflect the model applied to the testing data set.

A logistic model was generated from the fields selected in the initial analysis, which resulted in an AIC score of 25,940. We also generated models using both a forward and backward step wise analysis of all of the fields. This generated two models with the same AIC value of 25,856. Since these models had a lower AIC value than ours, we decided to use these. With identical AIC values, we chose the one generated from the backward stepping analysis, which generated a formula of:

## status ~ amount + term + rate + length + home + verified + reason +   
## state + debtIncRat + openAcc + revolRatio + totalAcc + totalRevLim +   
## accOpen24 + bcOpen + bcRatio + totalLim + totalIlLim

Since we transformed our data before generating this model, the actual relationships would be represented by:

status ~ amount + term + rate + length + home + verified + reason + state + debtIncRat + openAcc + revolRatio + totalAcc + sqrt(totalRevLim) + accOpen24 + bcOpen + bcRatio + ln(totalLim) + ln(totalIlLim)

Transforming the data prior to this step makes the formula (without ln or sqrt) easier to work with.

This logistic model was applied to the model testing data using the predict function. Using the functions default success cutoff probability of 0.50, the model accurately predicted 79.1% of the test cases from the test data. The accuracy of the prediction was calculated by comparing the sum of good and bad loans successfully predicted against the total number of loans in the test set. The results are summarized by:

pred\_tbl

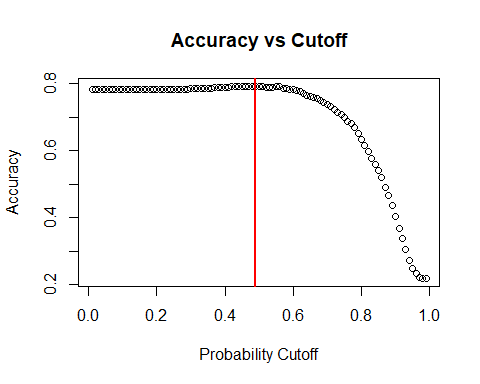
## pred\_result  
## Bad Good  
## Bad 224 1274  
## Good 161 5195

At this default cutoff probability level, the model prediction would result in a Type 1 error rate (incorrectly rejecting good loans) of 0.03, and a Type 2 error rate (incorrectly accepting bad loans) of 0.85. Since the Type 2 errors are primarily responsible for reducing the banks profit earned from good loans, this represents an improvement of about 15% over not using the logistic model.

## Section 6: Optimizing the Threshold for Accuracy

The model accuracy can be adjusted by changing the probability value used by the predict function to determine which loan applications are classified as “Good” and which are classified as “Bad”. To find the value that corresponds to the highest model accuracy, we loop through probability values from .01 to .99. At each increment of .01, we generate a contingency table and calculate the accuracy at that point. These accuracy values are stored in a vector for graphing and analysis. To find the probability corresponding to the maximum accuracy, we find the maximum value from the accuracy vector, and look up that value’s position in the vector. Since we started at a probability of .01, and incremented by .01, the position in the vector is equal to the probability value \* 100.

This approach found a maximum model accuracy to be 0.792 at a cutoff probability value of 0.49. This is shown in the graph below by the red line.



## Threshold Correct Incorrect Accuracy T1\_rate T2\_rate  
## 0.1 5356 1498 0.781 0.000 0.999  
## 0.2 5358 1496 0.782 0.000 0.998  
## 0.3 5378 1476 0.785 0.001 0.980  
## 0.4 5407 1447 0.789 0.008 0.939  
## 0.5 5419 1435 0.791 0.030 0.850  
## 0.6 5360 1494 0.782 0.074 0.735  
## 0.7 5060 1794 0.738 0.182 0.547  
## 0.8 4348 2506 0.634 0.381 0.310  
## 0.9 2757 4097 0.402 0.745 0.070

By reviewing the results of the different cutoff values, it’s clear that over estimating the number of good loans, with a smaller cutoff probability, will not greatly change the level of accuracy. Since the overall percentage of good loans is at about 78%, the model accuracy will not decline below that as your cutoff decreases.

Conversely, over estimating the number of bad loans, with a high cutoff probability, will affect the accuracy to a greater extent. In this case, the lower limit of the accuracy will match the number of bad loans at about 22%. This illustrates that setting the cutoff probability value closer to the value representing the greater population will yield a higher accuracy than the converse.

## Section 7 - Optimizing for Profit

We define profit as (Total Repaid - Loan Amount). “Good” loans generate a profit, while “Bad” loans generate a loss. Using the model’s predictions, we only calculate the profit for all loans predicted as “Good”, since those predicted as “Bad” would be denied. By changing the cutoff proportion used by the predict function, we change the profitability of the model.

To determine maximum profit, we use the same method as used to find the accuracy. Using this maximum profit level (shown below), we see what the model is capable of, and how much difference there is between our model and a “perfect” model with 100% accuracy.

Original Bank Profit (before model, accuracy = 78%): **$1,344,071**

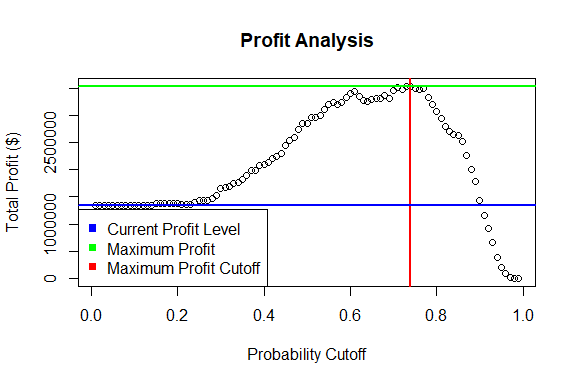
Profit allowed by model (cutoff = .5, accuracy = 79.1%): **$2,851,203**

Profit allowed by highest accuracy (cutoff = 0.49, accuracy = 79.2%): **$2,846,962**

Maximum profit provided by the model (cutoff = 0.74, accuracy = 70.6%): **$3,537,760**

Perfect model (accuracy = 100%): **$12,439,991**

From this, we can see that the largest profit comes from a proportion level which allows more bad loans to be classified as “Good”. However, this also allows in more actual good loans. Between the maximum accuracy at a cutoff proportion of 0.49 and this level, the increase in actual good loans offsets the loss by additional bad loans. Once the maximum profit level is reached, then the inclusion of more bad loans begins to reduce the profit generated by the inclusion of more good loans.



## Section 8 - Summary

Using our approach, we found a model which worked with the following variables to predict the binary status of each application:

## status ~ amount + term + rate + length + home + verified + reason +   
## state + debtIncRat + openAcc + revolRatio + totalAcc + totalRevLim +   
## accOpen24 + bcOpen + bcRatio + totalLim + totalIlLim

We used this model, with the predict function to generate the probability that a certain loan application would result in a good or bad loan. By adjusting a cutoff probability set in the predict function, we were able to affect the accuracy and profitability of the model. With the large loss in profit due to bad loans, it is clear from these results that predicting even a small percentage of bad loans can result in significant increases in profit.

By reviewing the different probability levels, and calculating the profit for each level, we found that setting the predict function to categorize loans at a probability of 0.74 maximized the profitability of the model. This level increased the profits of the model testing data from $1,344,071 to $3,537,760 for an increase of 163.2%. For the test data, this results in 4036 good loans correctly identified out of 5356, and 805 bad loans correctly identified out of 1498.

After using this model for a period of time, it’s important to review the results and re-test the model on new data. Changes over time may warrant adjustments or even a new model to help maintain a higher profitability. We recommend studying the steps we’ve laid out and using them to continue testing the current and new models for predicting loan outcomes.