Predicting Creditworthiness with LendingClub Data

Combiz Khozoie, Ph.D.

January 19, 2018

This document contains preliminary analyses and modelling of creditworthiness in a LendingClub dataset. A binomial multiple logistic regression model is constructed and its performance evaluated using standard metrics.

The LendingClub dataset provided comprises 24,724 records with 20 predictor variables and a single binary dependent variable specifying creditworthiness[1]. The data is first processed to (i) improve amenability of data inputs to model building, and (ii) engineer new features with predictive value. A cross-validation approach is applied, with partitioning of data into a training data set to train a model, and a test data set to evaluate its performance. Statistical sampling techniques are applied to improve class balance while ensuring test data remains unseen during training. Candidate logistic regression models are screened exhaustively, or using a Genetic Algorithm, to identify the best model. Model performance is evaluated using standard classification model validation metrics, including receiver-operating characteristic (ROC) curves, a confusion matrix, Gini coefficient, and Kolmogorov-Smirnov (KS) score. The final preliminary model achieves a AUC of 0.712 and Positive Predictive Value (PPV) of >90%.

Data Preprocessing

Wrangling

The dichotomous dependent variable, Class, is binary encoded (*Creditworthy*:1, *Uncreditworthy*:0). Several variables include numeric data encoded in string format (e.g. "One", "Two"), and where appropriate these values are recoded as discrete numeric variables [2]. The *Earliest.Credit.Line.Opened* variable is a POSIXct encoded date representing the month the borrower's earliest credit report was opened. This date was converted to a standard (YYYY-MM-DD) format and used to calculate the time (months) elapsed since this date, generating a new predictor variable (*credit_age*).

The *Loan.Application.Description* numeric variable specifies the word count of the loan description provided by the borrower (range 0-1000). The histogram revealed edge peaks at 0-1 and 1000, suggesting truncation and grouping at 1000+ and a large number of borrowers choosing not to enter any information or only a single

[1] Loan.Amount, Loan.Term, Employment.Length, Home.Ownership, Annual.Income, Loan.Purpose, Address.State, Debt.To.Income.Ratio, No..Delinquencies.In.Last.2.Years, Earliest.Credit.Line.Opened, FICO.Credit.Score, No..Inquiries.In.Last.6.Months, Months.Since.Last.Delinquency, No..Of.Credit.Lines, No..Adverse.Public.Records, Total.Credit.Balance, Use.Of.Credit.Line, Total.Number.Of.Credit.Lines, Loan.Application.Description, No..Of.Public.Record.Bankruptcies

[2] No..Delinquencies.In.Last.2.Years, No..Adverse.Public.Records, No..Of.Public.Record.Bankruptcies

word /letter for the loan description (Fig. 1.1). Exploratory analyses revealed marked differences in creditworthiness for these borrowers; therefore, this variable was recoded as a categorical variable (*descr_cat*). Categorization cutoff values were selected with reference to English language features (e.g. typical sentence length) and population distribution (e.g. disproportionate blank and one word descriptions)[3].

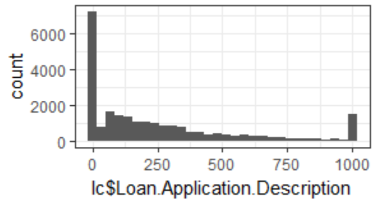


Figure 1: Histogram of the *Loan.Application.Description* numeric variable.

Several numeric variables with limited range and few records outside two majority values were simplified with binary encoding[4].; similarly, the *Home.Ownership* variable with five degrees of freedom was binarized (*Home.Owner* =1 ∨ 0). Next, a conditional density plot of *Annual.Income* against creditworthiness was examined, revealing inconsistent conditional distribution trends across all ranges of *Annual.Income*(Fig. 1.1). To facilitate categorization of *Annual.Income*, a decision tree was constructed, revealing suitable salary cutoffs (derived from entropy / information gain metrics) to stratify individuals into three salary categories[5].

A contingency table of counts for each US State specified in *Address.State* was produced and used to identify States with limited sample sizes. The median and mean counts for each State were 268 and 495, respectively. To avoid generalizing from limited sample sizes, any State with 75 or fewer records was relabeled as *Other*[6].

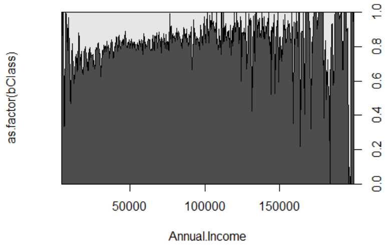


Figure 2: Conditional density plot of the *Annual.Income* numeric variable.

Train-test split

Initial partitioning of data into train (70%, n = 17307) and test (30%, n = 7417) splits was achieved using proportional stratified random sampling according to proportions of creditworthy (82%) and uncreditworthy (18%) records in the original data. As class imbalance can impair model performance, the minority class within the train data set was oversampled to achieve 50/50 class balance (n = 28138, Creditworthy: n=14180, Uncreditworthy: n=14138). By oversampling the train split (rather than the full data set before partitioning), contamination of test samples into the train split is avoided and the test split data remains unseen.

Logistic Regression Modeling

To determine which combination of 20 candidate predictor variables produces the best logistic regression model, automated model selection was performed. As an exhaustive search of all possible non-redundant models (excluding interaction terms) considers 2^n models (1048576 with 20 predictors), and is prohibitively computationally intensive, a more rapid Genetic Algorithm approach was utilized. After

[3] Word count cutoffs were as follows: Blank (=0), Word(=1), Short (≥2 & ≤15), Medium(≥16 & ≤350), Long(≥351 & ≤999), and Life-Story (≥1000).

[4] No..Delinquencies.In.Last.2.Years, No..Adverse.Public.Records, No..Of.Public.Record.Bankruptcies

[5] Low: ≤35000, Medium: >35000 & ≤76000, and High: >76,000

[6] States assigned to 'Other' due to limited sample sizes were as follows:- AK:57, DE:71, IA:4, ID:6, IN:11, ME:3, MS:19, MT:53, NE:5, SD:37, TN:18, VT:36, WY:55

1200 generations, the performance of the top 100 models was plotted according to Akaike information criteria (AIC). The best model (lowest AIC) was selected for subsequent predictions (Fig. 2). Potentially improving the model by incorporating interaction terms is a task for future work.

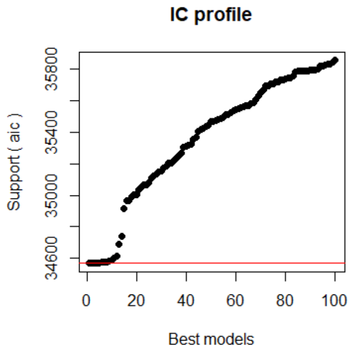
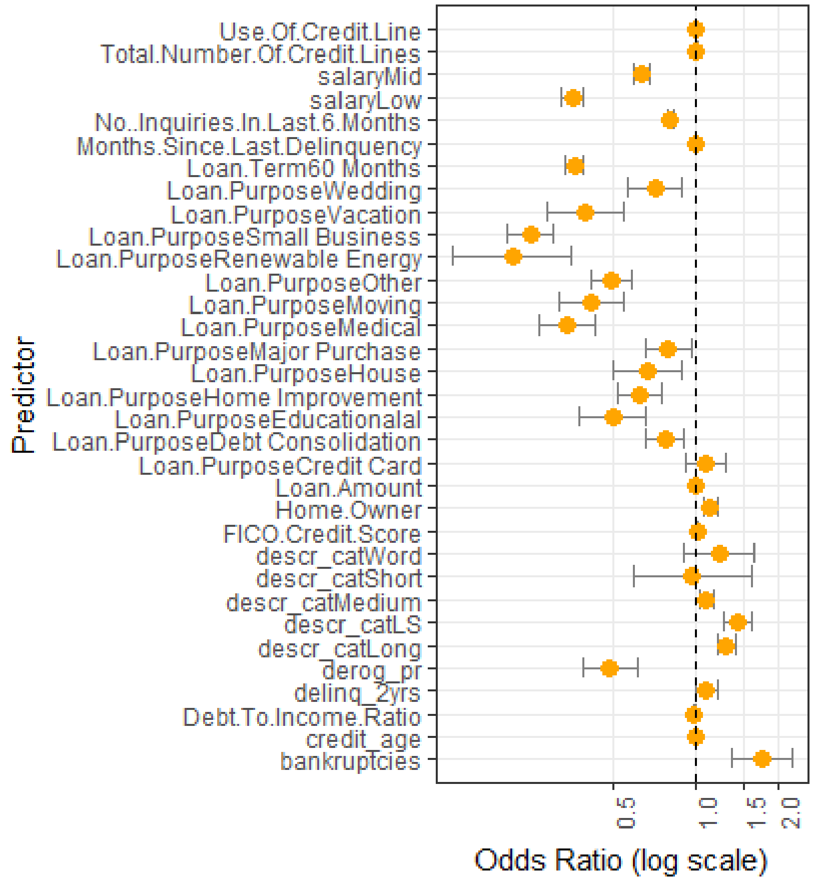


Figure 3: Akaike information criteria (AIC) for top 100 models after 1200 generations of Genetic Algorithm iteration.

Visualizing predictor influence

To visualize the influence of each predictor variable on classification, odds-ratios (OR) together with 95% confidence intervals (CI) were calculated. To aid visualization, the large number of ORs for US States were omitted from the initial plot (Fig. 1). Together, these reveal predictive factors associated with increasing risk (e.g. loans taken out over 60 months, loans for medical reasons, etc.) and decreasing risk (e.g. higher salary).



The final logistic regression model included the following 17 predictor variables:-
Loan.Amount,
Loan.Term,
salary,
Loan.Purpose,
Address.State,
Debt.To.Income.Ratio,
FICO.Credit.Score,
No..Inquiries.In.Last.6.Months,
Months.Since.Last.Delinquency,
Use.Of.Credit.Line,
Total.Number.Of.Credit.Lines,
delinq_2yrs,
derog_pr,
bankruptcies,
credit_age,
descr_cat,
Home.Owner

Figure 4: Creditworthiness risk predictors. *All model predictors excluding geographic predictors.*

Visualizing geographical predictor influence

ORs for each US State were calculated from the model coefficients and used to produce a choropleth map with regions shaded according to OR, representing relative increase or decrease in default risk (Fig. 2). An interactive version of the plot is available: `https://goo.gl/WMi79b`.

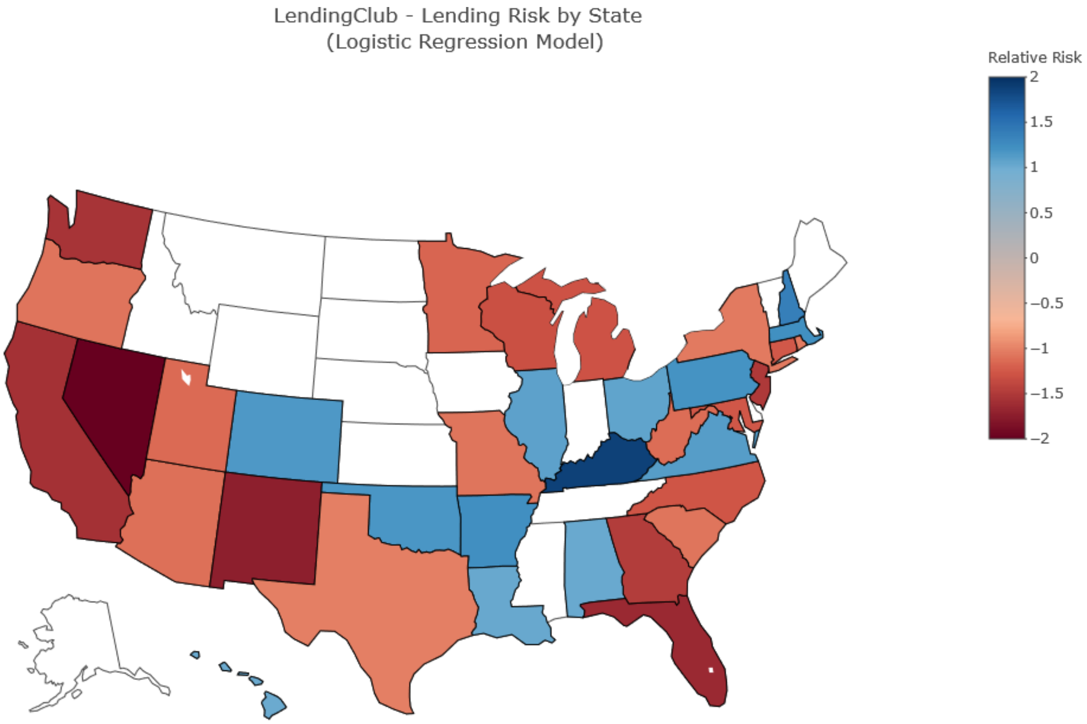

LendingClub - Lending Risk by State
(Logistic Regression Model)
Relative Risk
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2

Figure 5: Choropleth map of LendingClub loan relative default risk by US State. *Note that States without a fill colour were omitted due to low sample sizes.*

Model Evaluation

The model was used to predict creditworthiness on the test data set. Binomial model prediction output is within range 0-1 (Fig. 3). To binarize the prediction output, a threshold value was selected to maximize Youden's Index (t = 0.5264). Next, a classification matrix was produced, revealing the model correctly predicted uncreditworthy individuals in 90.22% of cases (Fig. 4). Evaluation metric results include a Gini-index of 0.4244, Recall of 0.63, KS score of 0.314, and ROC plot with AUC of 0.712 (Fig. 4).

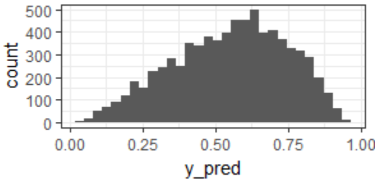
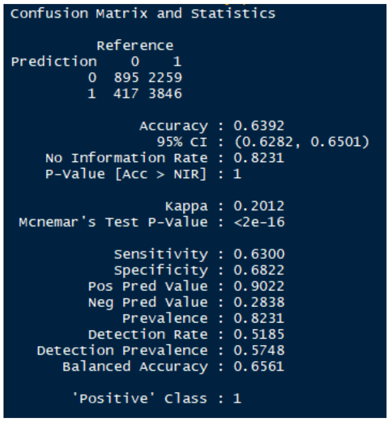


Figure 6: Histogram of binomial creditworthiness prediction output on the test data partition.

Concluding remarks

This preliminary analysis revealed useful insights into LendingClub data and facilitated prediction of creditworthiness from predictor variables. Opportunities for follow-up work are extensive and include modelling interaction terms in the logistic regression model, assessing the performance of machine learning based approaches (e.g. disjunctive and conjunctive terms within a decision tree based model), and incorporating natural-language processing (NLP) on the applicant's loan description text. The optimisation of model performance to meet investment goals (e.g. rate of return, trading frequency, etc.) is another area for future work.



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  895 2259
         1  417 3846

               Accuracy : 0.6392
                 95% CI : (0.6282, 0.6501)
    No Information Rate : 0.8231
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2012
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6300
            Specificity : 0.6822
         Pos Pred Value : 0.9022
         Neg Pred Value : 0.2838
             Prevalence : 0.8231
         Detection Rate : 0.5185
   Detection Prevalence : 0.5748
      Balanced Accuracy : 0.6561

       'Positive' Class : 1
```

Figure 7: Classification matrix. *Note that Creditworthy is encoded as 1.*

NOTE

THIS IS A WORK IN PROGRESS AND IS NOT FOR PUBLICATION OR DISTRIBUTION WITHOUT THE AUTHOR'S EXPLICIT PERMISSION.

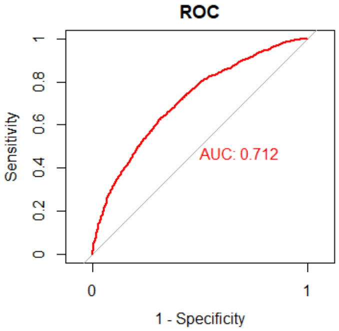


Figure 8: Receiver-operating characteristic curve with area under the curve (AUC).