Optimizing LendingClub's Financial Risk

Abstract

In this paper, we explore profit maximizing strategies for LendingClub based on a classification approach to predict defaults. Traditionally, loan-level risk is measured as probability of default which we use to measure the expected loss. Using various classification models, we find the maximum additional expected profit to be earned through classification, and give the optimal percentage of loan applications to reject at each risk level ('Grade') in order to reach this maximum.

1. Motivation

A key issue of loan origination in the financial services industry is balancing risk while increasing credit accessibility. In this project we aim to test a variety of models to understand how to accurately measure financial risk, as probability of default, for P2P loans using the LendingClub dataset. Traditionally, credit risk as a simple binary classification problem is sufficient; however we explore whether this applies to a fundamentally different bank-issuance loan structure. We propose that instead of credit score classification, we can use expected maximum profit, EMP, as a way to optimize financial risk for P2P loans.

1.1. Literature Review

Evaluating Credit Risk and Loan Performance in Online Peer-to-Peer (P2P) Lending looks at P2P loan characteristics to analyze the loan's credit risk and quantify its performance. Predicting the probability of default for P2P loans is standard practice for how loan risk is quantified and optimized, however in our paper we seek to analyze whether there is a better way to understand P2P risk.

Development and application of consumer credit scoring models using profit-based classification measures presents a new approach for consumer credit scoring, suggesting profit-based classification performance measure. This performance measure considers the profit or loss generated from accepting or denying loan applications. It is based on the Expected Maximum Profit (EMP) and is driven by the loss given default and the operational income given by the loan. The paper also explores the optimal cutoff given by EMP analysis.

A Novel Profit Maximizing Metric for Measuring Classification Performance of Customer Churn Prediction Models applies EMP to the problem of churn prediction and derives the cost benefit analysis structure that underlies the EMP metric.

2. Data Exploration

We analyzed a dataset of roughly 1,300,000 loans backed by the LendingClub between 2008 and 2018. We chose to only analyze loans that were paid off in full, charged off or defaulted in this case. The variables provided are borrower characteristics and loan characteristics at time of application and loan issuance.

The lowest 25th percentile of interest rates is 10%, the average rate is 13.5%, and the highest 75th percentile is 16%. For reference the average annual percentage rate for credit card payment is 15.96%. As a comparison, the average 30-year fixed rate for a mortgage is 4.62% and the average 15-year fixed rate is 4%.

The lowest 25th percentile of loan amounts is $\$8,000$, the average is $\$14,755$, and the highest 75th percentile is $\$35,000$. The maximum loan amount given by LendingClub is $\$40,000$. The lowest 25th percentile of income of loan applicants is $\$45,000$, the average is $\$65,755$, and the highest 75th percentile is $\$90,000$.

The preliminary data analysis plots are as shown below in Figures 1 and 2 based on metrics that we deemed empirically interesting. The one plot to note below is the Default Rate by Risk Classification. There seems to be considerable overlap until 2011, but by 2015 there is a clear distinction. To test whether the default rate distinction is statistically significant in 2015, we used a t-test. We stated the null hypothesis to be that there is no statistical difference between the six means for each of the six unique risk classifications, and we rejected the null hypothesis with a p-value of 0.002759 and t of 5.4799. Since this p-value is ≤ 0.05, we conclude that the difference between each risk classification is meaningful in 2015.

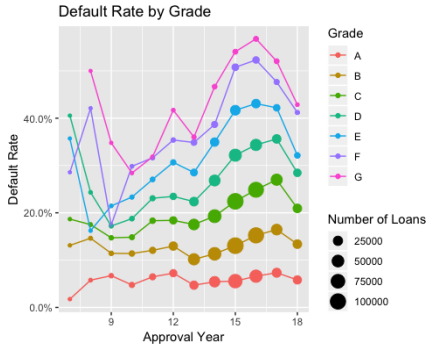

Figure 1. Default Rate by LendingClub Grade

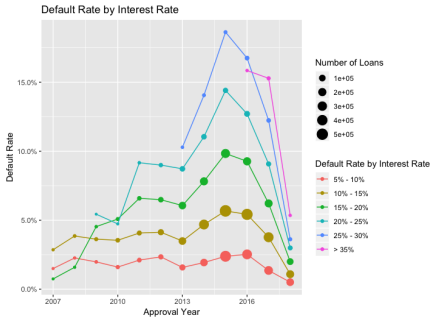

Figure 2. Default Rate by Interest Rate

2.1. Data Modifications

Initially, there were over 400 variables provided in the dataset, many of which were sparse. We dropped columns that were more than 20% sparse and ensured that the variables kept were only known during the time of application, keeping 347 columns in our final dataset.

For our data preprocessing, we did basic data exploration, looking at default trends of loan amounts by risk classification, interest rate by risk classification, loan amounts by income, etc. We dropped rows where income was greater than $\$500,000$, due to anomalies including an example where an applicant listed an income of $\$8,400,000$ with a job of "Mechanical Mobile Inspector." There are other examples with greater than $\$1,000,000$ income listed with jobs such as "Teacher" and "Dietary." We are treating these as outliers to be eliminated from the dataset. As standard practice, we normalized the continuous variables with mean 0 and variance 1 such as income, and one- hot encoded categorical variables such as borrower state.

We time split our data with approximate years for train: 2008-2015, validation: 2016, and test: 2017-2018 for a standard 80/10/10 split. The dataset has 347 columns with 1,097,529 rows for the train, 137,191 rows for the validation, and 137,191 rows for the test set. Initially, when we had a random shuffle train/val/test split we incorporated economic variables. For the final dataset we did a time split, so we did not incorporate economic variables in our analysis since the test dataset is only one year, and thus the predicting power of those economic variables has minimal effect. Additionally, since this time series did not include a full business cycle and the economy has been in an upswing, we did not think it was necessary.

3. Methodology

3.1. Prediction

As a first step, we experimented with several different models in order to make default predictions. These models included simple logistic regression, regularized logistic regression, random forest, and a neural network. Then, the relative value provided by the predictions from these models was evaluated using EMP estimation.

3.2. Expected Maximum Profit

EMP is a metric of comparison between classifiers. It can be interpreted as an upper-bound on the additional profit gained by using the classifier versus performing no classification.

This metric quantifies the value of a classifier by assigning a cost, c, to false positives and a benefit, b, to true positives. Negative predictions are ignored as no action would be taken on these points. Given a cutoff score t, predicted probability of default $F_1(t)$, predicted probability of no default $F_0(t)$, and empirical probability of default π_1, the profit is then

$$P(t, b, c) = \pi_1 F_1(t) b - \pi_0 F_0(t) c, \quad (\pi_0 = 1 - \pi_1)$$

In context of our dataset, true positives are correct predictions of default (deny application), false positives are incorrect predictions of default (deny application), and the baseline value is full acceptance of all applications, since this is a dataset of loans that were in fact granted by LendingClub.

The benefit, b, associated with denying a loan that will default is the avoided loss,

$$\lambda = \frac{\text{total amount paid back}}{\text{loan amount}}$$

λ is an uncertain quantity as the exposure at default and recoveries after the fact can both vary widely. We handle this value as a distribution $f(\lambda)$, where $f(\lambda)$ is the empirical distribution in our dataset.

The cost c of denying a loan that does not default is the lost ROI on that loan. This is calculated as the total $(A + I)/A - 1$, where A is the loan amount and I is the total interest paid. In our dataset, this value was essentially the same across loans within each risk classification. Thus for each loan grade we set c to be the average ROI for that grade.

Optimizing the average profit function over t gives the max-

imum potential profit generated by using the classifier at hand:

$$MP = \max_t P(t, b, c) = P(t*, b, c) = P(t^*, \lambda, ROI)$$

Since λ is defined over a distribution, we calculate the expected maximum profit,

$$EMP = \int_0^1 \max_t P(t, b, c) d\lambda = \int_0^1 P(t^*, \lambda, ROI) f(\lambda) d\lambda$$

$$= \int_0^1 (\pi_1 F_1(t^*(\lambda))\lambda - \pi_0 F_0(t^*(\lambda)) ROI) f(\lambda) d\lambda$$

where $t^*(\lambda) = \arg\max_t P(t, \lambda, ROI)$

Finally, the the expected optimal cutoff, corresponding to the optimal fraction of loans to deny, η_{emp} is computed as follows,

$$\eta_{emp} = \int_0^1 (\pi_1 F_1(t^*(\lambda)) + \pi_0 F_0(t^*(\lambda))) f(\lambda) d\lambda$$

3.3. Empirical Estimation

We split our data by grade ROC curves for our models predictions on each grade, g. We then compute the EMP and the η_{emp} for each grade based on the empirical loss distribution, $f_g(\lambda)$ and the average ROI for the given grade, r_g. Here we estimate EMP with a sum rather than an integral, using $\lambda \in \{0, 0.01, 0.02, ...0.99, 1\}$.

$$EMP = \sum_\lambda P(t^*, \lambda_i, ROI) p(\lambda \in (\lambda_i - 0.01, \lambda_i])$$

A more detailed version this process as well as the theoretical derivation of the metric can be found in (Verbraken, 2013).

4. Models

For all of these models, we balanced the dataset because of a large imbalance between the number of defaults vs survivals in the loan. Unbalanced, we found that the algorithm would always choose to make predictions that the loan would not default producing a high accuracy but low AUC. Additionally, we tuned the hyperparameters using the validation set by testing a range of $0.01 - 100$ λ for the regularized logistic regression, and range of $5 - 20$ max depth for the random forest. We chose the best performing model on the validation set and made predictions on the train/test sets.

4.1. Logistic Regression

In logistic regression, we choose a hypothesis of the form: $h_\beta(x_i) = \sigma(\beta^T(x_i)) = \frac{1}{1+e^{\beta^T x_i}}$

In order to pick the optimal parameter vector β, we minimize the empirical binary cross-entropy loss over our dataset:

$\mathcal{L}_{BCE}(x_{1:n}, y_{1:n}; \beta) =$

$\frac{1}{n}\sum_{i=1}^n (-y_i \log h_\beta(x_i) - (1 - y_i) \log(1 - h_\beta(x_i)))$

This minimization is performed using stochastic gradient descent.

4.2. Logistic Regression with L2 Penalty

In order to improve the prediction accuracy and prevent over fitting, we implemented regularization techniques. Since the features in our model are intuitively highly correlated and spare, theoretically we believed the coefficients would be poorly determined using ordinary logistic regression.

The Ridge Penalty, known as the L_2 norm, adds the squared magnitude of coefficients as a penalty term to the loss function. The optimal lambda was tuned to be 0.01.

$\mathcal{L}_{L_2}(x_{1:n}, y_{1:n}; \beta) = \mathcal{L}_{BCE}(x_{1:n}, y_{1:n}; \beta) + \lambda \left|\left|\beta\right|\right|_2^2$

4.3. Random Forest

Random forests are models consisting of a ensemble of decision trees trained through bagging, where B different trees are fitted using n samples with replacement from the training data. To perform classification with the ensemble, the majority result of the trees is used, which reduces the variance of fitting individual tree models. The optimal depth was tuned to be 15, outlined in our bias-variance trade-off analysis in Section 6.

4.4. Neural Network

We used a fully connected, 5-layer neural network. Our 5 hidden layers have shapes (89, 89, 45, 20, 2) and we use ReLU activation $g(z) = \max(0, z)$. The j-th output of layer i is calculated as:

$$a_j^{[i]} = g({W_j^{[i]}}^T x + b_j^{[i]})$$

for input x and activation function g. We use weighted cross-entropy loss

$$o = -(wy \log \hat{y} + (1 - y) \log(1 - \hat{y}))$$

where w is the weight of the loss for positive datapoints, and setting $w > 1$ decreases the false negative count. We used this weighted cross-entropy loss with $w = 5$ because the dataset had around four times the number of negative examples as positive examples. Furthermore, to counter this imbalance, we duplicate positive datapoints (label=1) in our training set.

Our prediction that the loan will default is based on the argmax of the output of our final hidden layer, which when softmaxed represents the predicted probabilities the loan will and will not default.

To find the optimal parameters, we train for 50 epochs with the Adam stochastic gradient-based optimizer, a learning rate of 0.0001, and a batch size of 128. To prevent overfitting, we add dropout on every layer with rate 0.5 and an ℓ_2 regularization penalty of 0.0001. We save the model at

the epoch with the best validation accuracy to use during testing.

5. Results

5.1. Binary Classification

We chose to measure the accuracy of each model with the area under the receiver operator characteristic curve (AUC). The AUC is a standard metric to measure the performance of binary classifiers measuring the false positive and false negative classification rates. Additionally, we included precision, recall, and F-1 scores. We are interested in looking for a high precision, which means that the algorithm returned substantially more relevant results than irrelevant ones, and a high recall, meaning that the algorithm returned most of the relevant results.

Estimation results are in tables 1 and 2. Some notable influential variables include interest rate, debt to income ratio, annual income, loan amount, loan term, and grade.

Table 1. AUC: True Positive vs False Positive Rate

Model	Train	Validation	Test
LR	0.5637	0.5696	0.5643
L2 LR	0.6602	0.6498	0.6685
Random Forest	0.7611	0.707	0.7292
Neural Network	0.8757	0.6074	0.5630

Table 2. Evaluation Metrics on the Test Set

Model	Precision	Recall	F1-Score
LR	0.7	0.8	0.71
L2 LR	0.78	0.65	0.68
Random Forest	0.77	0.71	0.73
Neural Network	0.72	0.76	0.73

5.2. Profit-Based Classification by Grade

Below are the results of the EMP analysis comparing the performance of the Random Forest model and Neural Network on each grade. EMP as a percentage gives the average additional return per applicant if the model is used. The EMP in dollars is the total profit gained from the additional percentage return, and the fraction to reject is the optimal percentage of applications to deny. The total estimated profit earned by the random forest model is $\$4.40$M corresponding to 0.116% in additional return on LendingClub's portfolio. The neural network is slightly less profitable generating about $\$4.36$M.

While the percent increase in profit increases with riskier loan grades, the majority of increased profit comes solely from 'F' grade loans. Aside from 'A' and 'B', 'F' is the largest category (642k loans) and also one of the riskiest. We also see that the optimal cutoff point is almost identical to 'G' (Fig. 4) indicating that LendingClub may be systematically underestimating the risk of these loans. This finding suggests that it would be beneficial for LendingClub to add more grades to the risky end of their classification scale with higher rates, or deny 40% of the loans in 'F' and 'G.'

Random Forest	Grade						
	A	B	C	D	E	F	G
EMP (%)	0.0013%	0.0011%	0.0022%	0.1033%	0.3308%	0.5174%	0.8498%
EMP ($)	$ 13,530	$ 11,180	$ 12,852	$346,263	$ 403,141	$ 3,322,769	$ 286,983
Fraction to Reject	3.57%	4.75%	16.54%	26.48%	32.88%	39.50%	40.06%

Neural Network	Grade						
	A	B	C	D	E	F	G
EMP (%)	0.0001%	0.0002%	0.0013%	0.1011%	0.3283%	0.5188%	0.8316%
EMP ($)	1,456	2,194	7,511	338,752	400,121	3,331,510	280,843
Fraction to Reject	0.66%	0.33%	6.42%	15.83%	23.99%	30.27%	30.52%

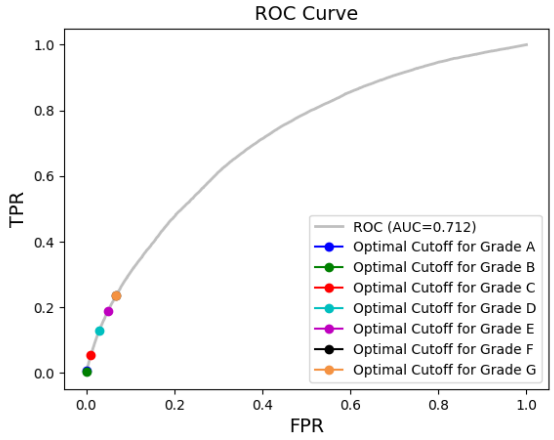

Figure 3. Optimal cutoff by Grade (Random Forest Model)

6. Bias Variance Analysis

By looking at the validation set predictions compared to the training set predictions, we are able to visualize the generalization gap for the Random Forest and Neural Network models. We produced graphs to compare this difference between in-sample and out-of-sample AUC for different hyperparameters, given in figures 4 and 5. We can see that the neural network has the highest variance, since there is

a large gap between the test and validation AUC, followed by the random forest. As expected, the random forest and neural network models have a larger generalization gap with weaker regularization that comes from a larger max depth or early stopping. The models do not indicate to be suffering from high bias. We suspect that the high variance of the neural network model largely stems from the imbalance of the dataset, since duplicating datapoints was able to increase the AUC of the training set but unable to generalize to the still imbalanced validation and testing set.

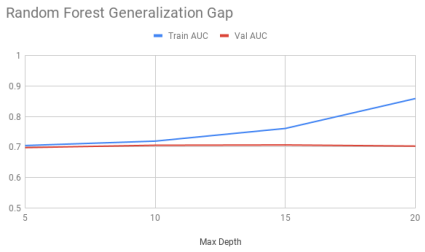

Figure 4. Measuring the generalization gap based on max depth: 0-20

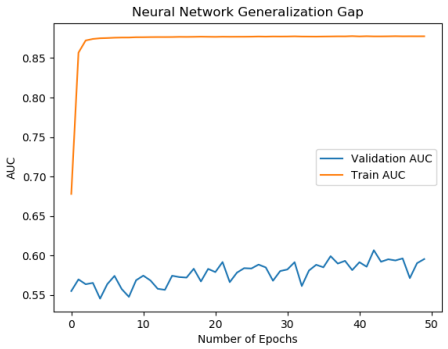

Figure 5. Measuring the generalization gap based on epochs

We also analyzed the misclassification rates for the most indicative variables such as interest rate, debt to income ratio, annual income, loan amount, loan term, and grade. Different loan terms had a 5% different misclassification rates, not enough to be interpreted as significant. The only significant difference in misclassification was for the grade, shown in figure 7. From top to bottom, this is the overall misclassification rate, false positive, and false negative.

7. Future Work

Since Random Forest has the best results, we are motivated to try additional tree structure algorithms. Similarly, the Neural Network performs well during training, but we can try to prevent overfitting by tuning more hyperparameters or finding methods to prevent dataset imbalance. Using this

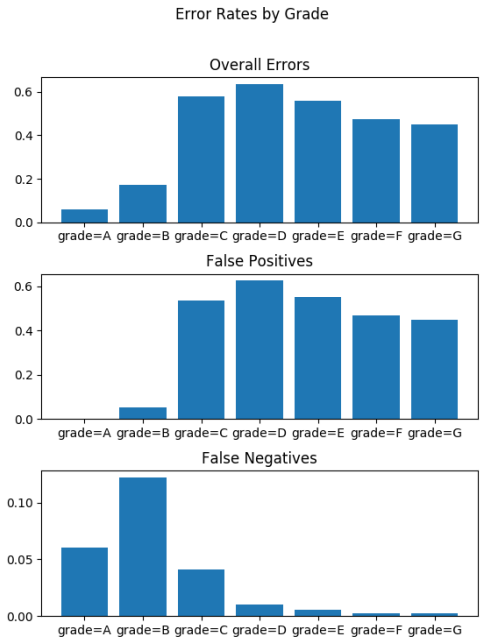

Figure 6. Granular Error Analysis by Grade

EMP evaluation metric, we can find subtleties for each grade and possibly classify loan default rate by grade. To further maximize firms' profit, we can try categorizing the loans in ways different from grade and applying the same EMP metric. Additionally, the EMP could be evaluated during training. Rather than calculating EMP after minimizing the loss function, gradient decent could be stopped at maximum EMP rather than minimum loss.

8. Conclusion

This paper applies a profit based performance measure, EMP, to the evaluation of classification models in the context of LendingClub's P2P lending business. We find significant opportunity for LendingClub to increase their profitability with the methods we lay out, and find our random forest model to be the most profitable choice. Furthermore, we identify that the majority of increased profit comes from denying selected loans in the 'F' risk class, suggesting that LendingClub should adjust their treatment of 'F' grade loans. More generally, there is a pattern of decreasing classification profitability per loan as loans get less risky. This is to be expected as the percentage of defaults is much lower in less risky classes, meaning that it is much harder to avoid false positives.

9. Citations and References

Riza Emekter, Yanbin Tu, Min Lu Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending

Hosmer, David W, and Stanley Lemeshow 2004, Applied logistic regression (John Wiley Sons)

Verbraken, Thomas, et al. "Development and Application of Consumer Credit Scoring Models Using Profit-Based Classification Measures. European Journal of Operational Research, vol. 238, no. 2, 2014, pp. 505513., doi:10.1016/j.ejor.2014.04.001.

Verbraken, T., Verbeke, W., and Baesens, B. "A Novel Profit Maximizing Metric for Measuring Classification Performance of Customer Churn Prediction Models," in IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 5, pp. 961-973, May 2013. doi: 10.1109/TKDE.2012.50

Carlos Serrano-Cinca, Begoa Gutirrez-Nieto, Luz Lpez-Palacios Determinants of Default in P2P Lending

Milad Malekipirbazari, Vural Aksakalli Risk assessment in social lending via random forests