Optimizing Lending Club's Financial Risk

Rachel Ahn, Michael Morrissey, Edward Xu

raahn@stanford.edu
mmorriss@stanford.edu
edxu24@stanford.edu

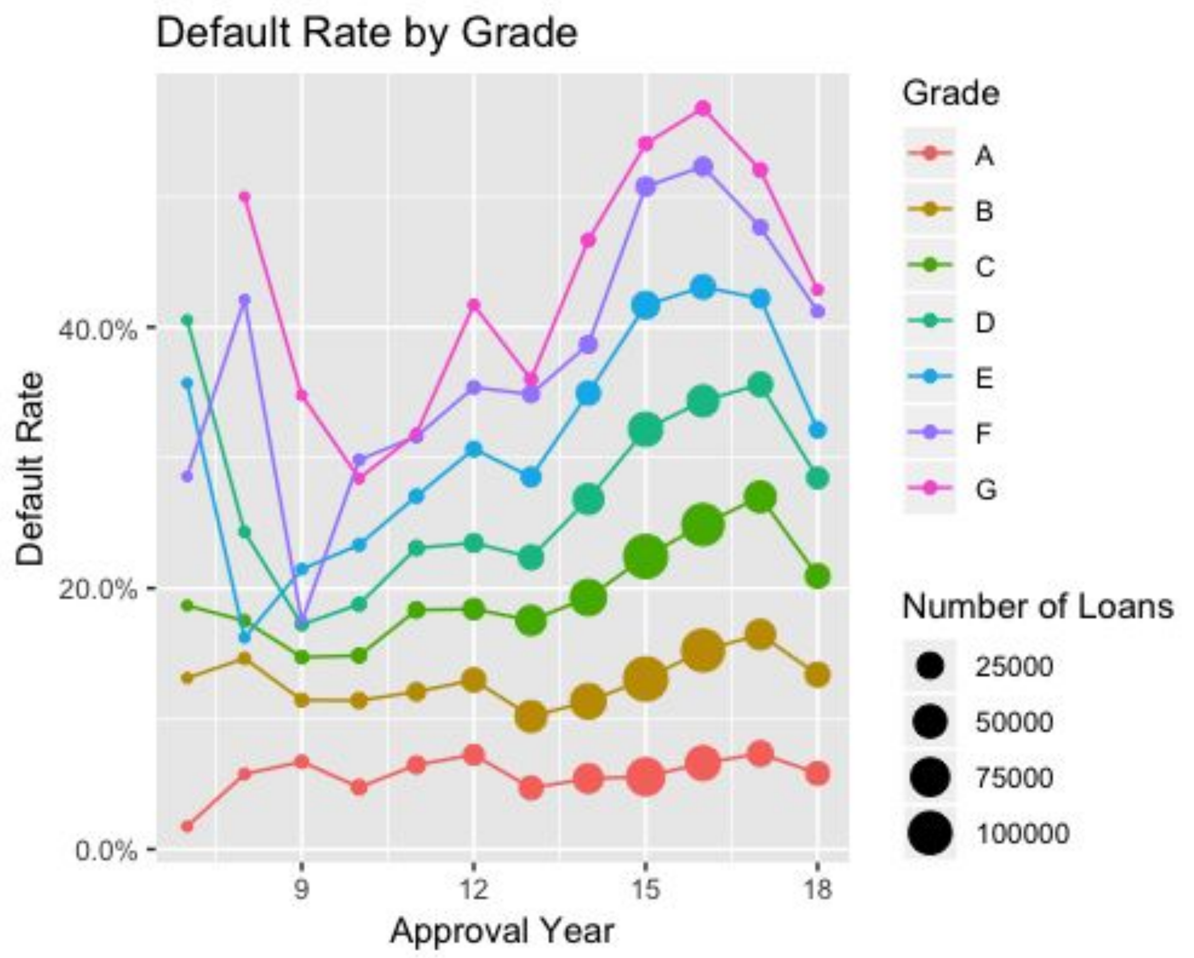
Abstract

Traditionally, loan-level risk is measured as credit risk—the probability of default to measure the expected loss. Using machine learning techniques, we modeled credit risk and expected payoff maximization on the ROC, to help LendingClub optimize their risk.

Data Overview

We analyzed LendingClub's dataset of roughly 2.2M loans between 2008–18. We chose to only analyze loans that were paid off in full, charged off or defaulted in this case. There are over 400 borrower characteristics at time of application and loan characteristics at time of issuance.

Features



Some notable influential variable include: interest rate, debt to income ratio, annual income, loan amount, loan term, and loan grade.

Methodology

1. Different models predict default probabilities.
2. The relative value provided by the predictions from these models was then evaluated using EMP estimation. 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.

Profit: $P(t, b, c) = \pi_1 F_1(t) b - \pi_0 F_0(t) c$

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

Models

__Logistic Regression__

- Our hypothesis has form: $h_\beta(x_i) = \sigma(\beta^T(x_i)) = \frac{1}{1+e^{\beta^T x_i}}$
- We use stochastic gradient descent and minimize cross entropy loss: $\frac{1}{n}\sum_{i=1}^{n}(-y_i \log h_\beta(x_i) - (1-y_i)\log(1-h_\beta(x_i)))$

__Regularized Logistic Regression__

- L2 norm with optimal lambda λ = 0.01.

__Random Forest__

- We use decision trees through bagging to do classification with optimal max depth 6

__Neural Network__

- We use a fully connected, 5-layer network with hidden layers of shape (89,89,45,20,2) and ReLU activation, where the j-th output of layer i is:

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

- We duplicate positive data points and use weighted cross entropy loss to counter imbalance of the dataset: $o = -(wy\log\hat{y} + (1-y)log(1-\hat{y}))$

Results

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	0.5625	0.5903	0.8765

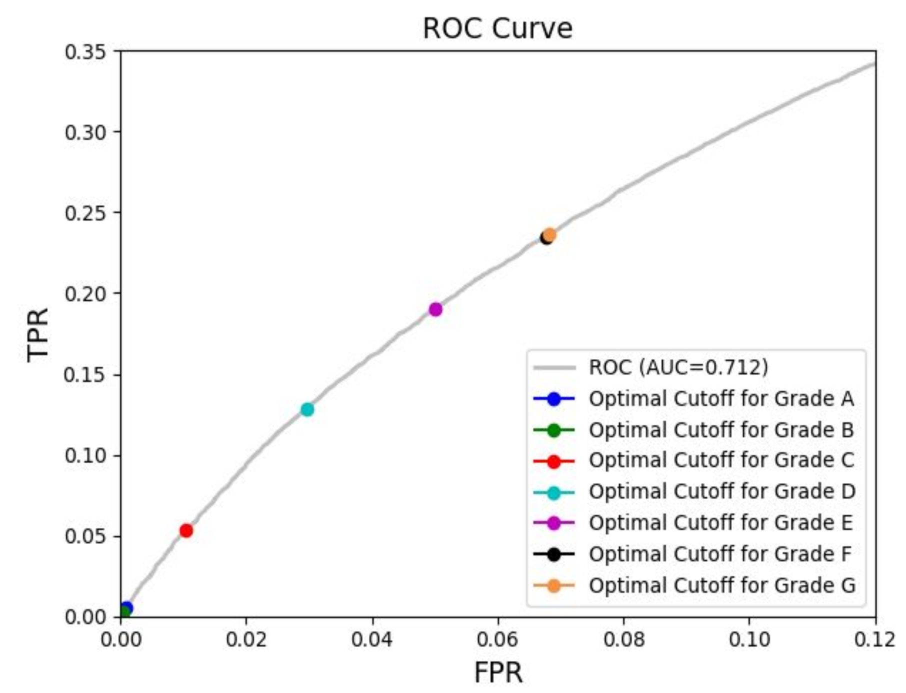
Table 2. Evaluation Metrics on the Test Set

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

EMP

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%



Discussion

The total estimated profit earned by the random trees model is $4.4M corresponding to 0.116% in additional return on LendingClub's portfolio. This finding suggests that it would be beneficial for Lending Club 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'.

Future Work

Based on the success of our random forest model, we can try other tree classification algorithms. Additionally, trying different types of Neural Network architecture could prove to be beneficial to correct for the class imbalance. We can apply the EMP metric for different combinations of results to optimize risk and do further misclassification analysis.