Optimizing Lending Club’s Financial Risk
Rachel Ahn, Michael Morrissey, Edward Xu

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.

\[
\text{Profit: } P(t,b,c) = \pi_1 F_1(t)b - \pi_2 F_2(t)c \\
\text{EMP: } EMP = \int_0^1 P(t^*, \lambda, \text{ROC}) f(\lambda)d\lambda
\]

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: \( \frac{1}{n} \sum_i (-y_i \log h_\beta(x_i) - (1 - y_i) \log(1 - h_\beta(x_i))) \)

Regularized Logistic Regression
- L2 norm with optimal lambda \( \lambda = 0.01 \)

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 i-th output of layer i is: \( o_{ji} = g(W_{ij}^T x + b_{ji}) \)
- We duplicate positive data points and use weighted cross entropy loss to counter imbalance of the dataset: \( o = -(w y \log \hat{y} + (1 - y) \log(1 - \hat{y})) \)

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'.

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Train</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.5637</td>
<td>0.5696</td>
<td>0.5643</td>
</tr>
<tr>
<td>L2 LR</td>
<td>0.6002</td>
<td>0.6498</td>
<td>0.6685</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.7611</td>
<td>0.707</td>
<td>0.7292</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0.5625</td>
<td>0.5903</td>
<td>0.8765</td>
</tr>
</tbody>
</table>

Profit:

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.7</td>
<td>0.8</td>
<td>0.71</td>
</tr>
<tr>
<td>L2 LR</td>
<td>0.74</td>
<td>0.65</td>
<td>0.68</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.77</td>
<td>0.71</td>
<td>0.73</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0.72</td>
<td>0.76</td>
<td>0.73</td>
</tr>
</tbody>
</table>

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.

Abstract

Data Overview

Features

Methodology

Models

Discussion

Results

Future Work