Loan Approval and Quality Prediction in the Lending Club Marketplace

Yondon Fu, Shuo Zheng and Matt Marcus

Introduction

Lending Club is a peer-to-peer lending marketplace where individual investors can provide arms-length loans to individual or small institutional borrowers. Lending Club performs the loan evaluation and underwriting, and investors such as you or I would fund the loans (in a way similar to KickStarter).

As a creditor, Lending Club performs loan underwriting in a much different method from traditional consumer loan creditors such as a consumer bank. Lending Club receives applications from individuals looking to borrow money, and evaluates the loan decision exclusively based on the information provided by the applicant; in-person evaluations are not involved. The company then assigns a rating of the loan, similar to how a rating agency such as Moody's assigns a rating to a publicly traded security, which significantly determines the interest rate on the loan. Lending Club then makes the loan available on the marketplace, where individual investors are able to evaluate the loan before making a decision to invest.

We are interested in encapsulating Lending Club's loan approval and rating assignment process using machine learning algorithms. Lending Club makes publicly available data about the loan applications they receive and the loans that are subsequently financed. We plan to apply machine learning techniques to this data to predict which loans they will approve, what grades they will assign them, and whether the loan will be a good investment.

Methods

Our goal is to predict which loans will be approved or denied, the grades they will receive and whether they will be good or bad investments. Given that our data is labeled and each of our predictions consists of placing loans in various categories, we will clearly be tackling a

classification problem. Furthermore, our data set is fairly large. Consequently, suitable methods for this problem need to be able to perform well with large training sets. Some of the methods that we plan on testing are listed below:

Ordered logistic regression

- ➢ Version of logistic regression for multiclass problems
- ➢ May be useful for the grading prediction since grades can be A, B, C, D...etc.
- ➢ Provides probabilities for each classification
- ➢ Has been used to predict bond ratings before, so this method might be suitable for predicting loan grades as well
- ➢ Data must meet proportional odds assumption: logarithms of the proportion of loans that receive each grade form an arithmetic sequence (i.e. number added to each logarithm to get the next is the same)

Relevance vector machines

- ➢ Similar to support vector machines, but provides probabilistic classification
- ➢ High accuracy
- ➢ Performs well in high dimensional spaces
- ➢ Odds associated with each classification
- ➢ http://www.tristanfletcher.co.uk/RVM%20Explained.pdf

AdaBoost

- ➢ Can be used with another learning algorithm to boost performance
- ➢ Output is a weighted sum generated from the combination of outputs from other learning algorithms (“weak classifiers/learners”)
- ➢ The final model will converge to a strong learner, so as long as the individual learners perform better than random guessing they can be weak
- ➢ Decision trees may be a good choice as weak learners since they are the easiest to implement
- ➢ Issue: susceptible to noise and outliers

Random forests

- ➢ Ensemble learning method that outputs the most frequently occurring class by creating a large amount of decision trees while training
- ➢ Solves the issue of decision trees tending to overfit to training set data
- ➢ Easy to interpret
- ➢ Non-parametric (no assumptions about how the data is distributed) so we don’t have to worry about outliers in the training data

Data

We'll be using the data Lending Club has made available here: https://www.lendingclub.com/info/download-data.action.

The data is already structured in CSV files. There is data for about 400,000 loans that Lending Club approved and 3 million that they denied. Of the approved loans, about 280,000 are currently in progress, 85,000 have been fully paid, 20,000 have defaulted, and 15,000 were paid late.

The data set is very comprehensive and contains a number of features that we'll consider. These include: zip code, annual income, number of collections over last 12 months, loan description, debt-to-income ratio, employment length, employee title, fico ranges, loan amount, loan grade, home ownership status, interest rate, last payment amount and date, current loan status, number of times creditors asked for payment, months since last delinquency, remaining principal, payments received to date, interest received to date, late fees received to date, and principal received to date

Milestone

By the milestone due date, we plan to have completed our prediction system for which loans Lending Club will approve as well as for the grades that Lending Club will assign individual loans. This will include appropriate feature selection for both approval and grading prediction as well as the implementation of the relevant learning algorithms. Our goal is to have successfully trained our algorithms with the training set and tested them on a test set by the milestone.

References

*Table from data dictionary xlsx file that can be found at https://www.lendingclub.com/info/download-data.action

zip_code	**The first 3 numbers of the zip code provided by the borrower in the loan application.**
addr_state	**The state provided by the borrower in the loan application**
annual_inc	**The annual income provided by the borrower during registration.**
collection_recovery_fee	**post charge off collection fee**
collections_12_mths_ex_med	**Number of collections in 12 months excluding medical collections**
delinq_2yrs	**The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years**

desc	Loan description provided by the borrower
dti	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
emp_title	The job title supplied by the Borrower when applying for the loan.*
fico_range_high	The upper boundary of range the borrower’s FICO belongs to.
fico_range_low	The lower boundary of range the borrower’s FICO belongs to.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
id	A unique LC assigned ID for the loan listing.
initial_list_status	The initial listing status of the loan. Possible values are – W, F
inq_last_6mths	The number of inquiries by creditors during the past 6 months.
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
is_inc_v	Indicates if income was verified by LC, not verified, or if the income source was verified
issue_d	The month which the loan was funded
last_credit_pull_d	The most recent month LC pulled credit for this loan
last_fico_range_high	The last upper boundary of range the borrower’s FICO belongs to pulled.
last_fico_range_low	The last lower boundary of range the borrower’s FICO belongs to pulled.
last_pymnt_amnt	Last total payment amount received
last_pymnt_d	Last month payment was received
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan
member_id	A unique LC assigned Id for the borrower member.
mths_since_last_delinq	The number of months since the borrower's last delinquency.
mths_since_last_major_derog	Months since most recent 90-day or worse rating
mths_since_last_record	The number of months since the last public record.
next_pymnt_d	Next scheduled payment date
open_acc	The number of open credit lines in the borrower's credit file.
out_prncp	Remaining outstanding principal for total amount funded
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors

policy_code	publicly available policy_code=1 new products not publicly available policy_code=2
pub_rec	Number of derogatory public records
purpose	A category provided by the borrower for the loan request.
pymnt_plan	Indicates if a payment plan has been put in place for the loan
recoveries	post charge off gross recovery
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
sub_grade	LC assigned loan subgrade
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
title	The loan title provided by the borrower
total_acc	The total number of credit lines currently in the borrower's credit file
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
total_rec_prncp	Principal received to date
url	URL for the LC page with listing data.

[1] https://www.lendingclub.com/info/download-data.action
[2] http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/
[3] http://en.wikipedia.org/wiki/Ordered_logit
[4] http://www.cs.princeton.edu/~schapire/talks/picasso-minicourse.pdf
[5] http://www.tristanfletcher.co.uk/RVM%20Explained.pdf