Estimating Lending Club Probability of Default

Introduction

Lending club is a platform that allows individuals to borrow money from individual/institutional investors. The platform uses technology to assess applications and determines if a loan is to be granted and the interest rate that will apply to the loan. Individual investors are then able to select to invest in specific loans

Innovation transforms lending

Lending Club is the world's largest marketplace connecting borrowers and investors, where consumers and small business owners lower the cost of their credit and enjoy a better experience than traditional bank lending, and investors earn attractive risk-adjusted returns.[4]

Here's how it works:

- **Customers interested in a loan complete a simple application at LendingClub.com**
- **We leverage online data and technology to quickly assess risk, determine a credit rating and assign appropriate interest rates. Qualified applicants receive offers in just minutes and can evaluate loan options with no impact to their credit score**
- **Investors ranging from individuals to institutions select loans in which to invest and can earn monthly returns**

The entire process is online, using technology to lower the cost of credit and pass the savings back in the form of lower rates for borrowers and solid returns for investors.

I used a dataset containing 37,869 loans extended between June 2007 and June 2011 in the US, made publically available by Lending Club.

The goal of the project is to estimate the probability of default

Model Creation Process

Removed Variables with Collinearity or Lacking Predictive Power

Interest Rate was not used because it was based on Lending Club's predictive data. Tested for collinearity Loan Amount and Instalment have a 0.955 collinearity as instalment is a direct function of loan amount and terms.

Location Data was removed due to lacking data points per category, and due to limitations of XLMiner. Other information such as employment title and description could be useful, but would require further NLP processing to convert into relevant data.

DTI was removed as it was based on Loan Amount. DTI had some useful data, as the description implied it considered I tried multiplying DTI by Reported Income to simply measure total debt: DTI * Annual_inc. However, there were still issues with significance, so it was removed.

Tested for Interaction Terms:

Interaction terms help model the way in which the combination of two variables affect the predicted variable, beyond their independent predictive power. While it appeared that some could interact (For example, it appeared that Grade interacted with term amount in relation to default relates), the p-values were not significant.

Check P-Values

Annual Income has a 0.43 P-value. However, it makes intuitive sense to use annual income to determine whether a person can pay their loan.

Delinquency, Employment Length, and Home Ownership also had high p-values. It makes intuitive sense that they would affect the model. However, the relationship is less clear than annual income (where more money is obtained, and the loan is composed of money owed). Thus, they were removed to have a more statistically significant graph.

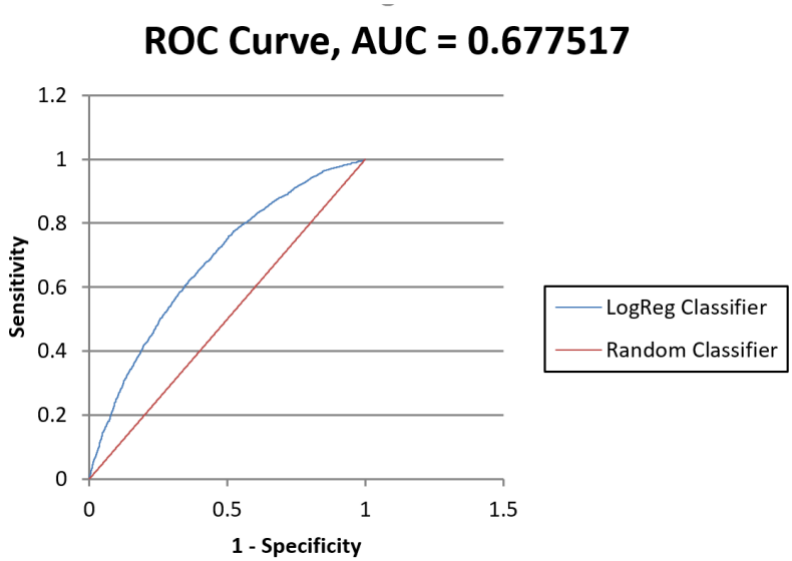
The final variables are: Loan Amount, Annual Income, Term length, and Grade.

Grade was converted into dummy variables, as it is categorical data.

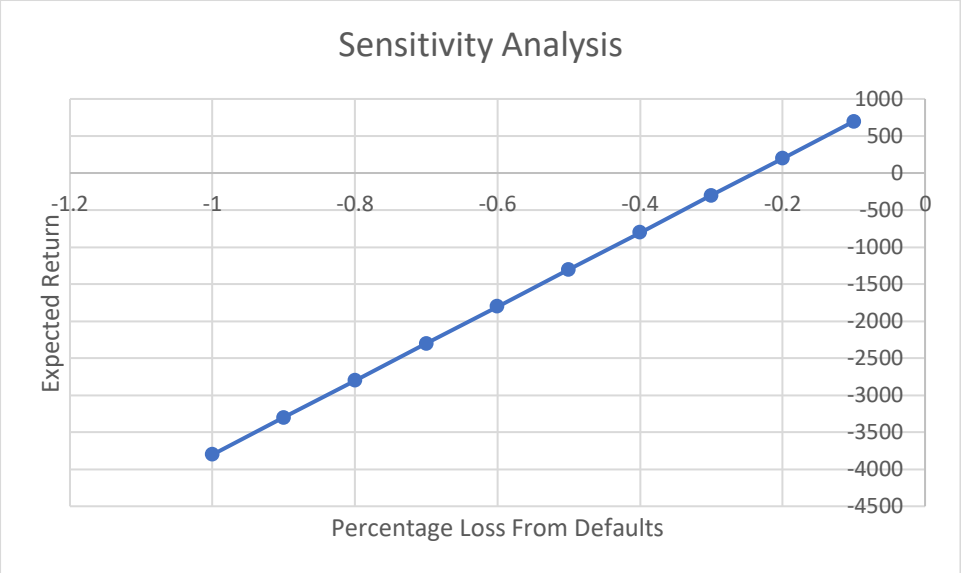
Run the Model:

Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds	CI Lower	CI Upper
Intercept	-0.19986	0.081427	6.024690481	0.014107	0.818842	0.698054	0.96053
loan_amnt	3.25E-06	3.3E-06	0.969756318	0.324741	1.000003	0.999997	1.00001
annual_inc	-4.8E-06	5.83E-07	68.61921533	1.19E-16	0.999995	0.999994	0.999996
term (mont	-0.51661	0.051159	101.9735741	5.63E-24	0.59654	0.539627	0.659456
grade_B	0.64638	0.06493	99.10170307	2.4E-23	1.908618	1.680546	2.167643
grade_C	1.027372	0.068402	225.5897036	5.46E-51	2.793713	2.443195	3.19452
grade_D	1.242494	0.075149	273.3620823	2.1E-61	3.464242	2.989792	4.013982
grade_E	1.337975	0.092257	210.3298064	1.16E-47	3.811318	3.180872	4.566718
grade_F	1.555456	0.129092	145.1832052	1.96E-33	4.737247	3.678262	6.101118
grade_G	1.725628	0.221575	60.65297003	6.81E-15	5.616045	3.637688	8.670331

Evaluate Results:



Observing the graph below, the expected value varies drastically depending on the chosen return on loan defaults. As seen in the data table, expected value becomes positive between -0.3 and -0.2.



The main hypothesis in selecting ideal loans is that by calculating lending Club's Implied probability, as well as the probability of this model, one can invest in the loans with the highest difference. This is because these loans will have a disproportional interest rate, while being likely to be fulfilled.

To select a subset of loans, the implied probability of Lending Club loans was predicted. This was done by determining p = 1 / (R / recovery) + 1, where R =(Instalments x term) / loan amount. Recovery was set at 0.5.

The probability of the model was then selected. This was done by multiplying the coefficients of the logistic regression with the values of each loan, and adding the y-intercept. The logistic regression formula was then used: Probability = 1 / (1 + exp(regression output)).

Finally, the probability of the model was subtracted by Lending Club's implied probability. This project stipulated investing in 200 loans, so the top 200 differences were selected.

Limitations:

The Loan Amount variable is only significant with 68% confidence. The model relies on the assumption that it is significant, as it makes sense that this factor would affect default rates.

The recovery amount was arbitrarily chosen. This would cause limitations if the model was used to select an arbitrary amount of investments. However, as 200 investments were required, using a sorting method for the best investments meant that the one did not have to worry about what amount of investments were worth investing on, just which were the best 200.

Future Improvements:

Location data could have added additional insight. The post codes could have been correlated with average income in the area. This this could have high collinearity with yearly income. However, as yearly income is self-reported, discrepancies in yearly income with postcode could provide insight to potential risky investments.

Another source for information could be tokenizing employment titles. Perhaps workers at FTSE500 companies have a lower probability of default, for example.