Lending Club Loan Risk Assessment Milestone Report

Access to credit is a vital component of the modern economy. To facilitate a health market for credit, credit must be priced appropriately, like any other good. The price of credit can be partially determined by assessing the risk that a given borrower will default on a loan, based upon his or her past behavior. Here, I am to develop a model of loan risk.

My client is a hypothetical online lending platform that wants to make loans to customers in the United States. Their goal is to use big data to assess loan risk and price/grant loans in an automated fashion. They are an early stage start-up and do not enough internal data to create an accurate risk model. A good risk model is vital to their pricing system, so they hired a number of consultants to help them craft a good risk model.

Once they have a good model or ensemble of risk models, they can begin predicting the riskiness of potential loans. That risk assessment will allow them to set interest rates and deny loans. Along the way, they may add other features of their service based upon risk. These features include targeted and timed micro-interventions with risky borrowers, novel loan re-payment structures, and services designed to help borrowers improve their credit and acquire lower interest rates. All of this, however, is contingent upon initial success with a solid risk model.

In order to build a loan risk model, I have been using the publicly available portion of the Lending Club’s data set (<https://www.lendingclub.com/info/download-data.action>). This data set consists of millions of loans and a number of features relevant to risk calculations. Loans in the data set are either current, paid, late, in default, or charged off. My modeling focuses on “completed” transactions and, therefore, utilizes the paid and charged off loans. As “default” and “charged off” are essentially the same from a loss perspective, I consider them to be the same in this analysis. In this data set, the rate of charged off loans is ~18%.

Data Features

Several features in this data set are based off the Lending Club’s own risk models. These include the interest rate on the loan, as well as the grade and sub-grade of the loan. These features are highly predictive of a charged off loan. However, as these represent results that the client would want to see rather than features that the client has, they are not relevant to the current analysis. Additionally, features such as the loan amount and the funded amount are indirectly based on the Lending Club’s models and are not considered here.

One of my preliminary hypotheses was that financially secure borrower pose a smaller risk. Features such as income, debt, utilized credit, etc. may reflect financial security may reflect the financial security of the borrower. Specifically, low incomes, high debt-to-income ratios, and high credit utilization may indicate an increased default risk. Additionally, a history of previous defaults, derogatory public records, number of delinquencies, and credit inquiries could be used to predict default. The employment history of the borrower may also represent a source of information on default risk.

This data set also includes geographic information that could be used to predict loan risk. Borrowers from more economically insecure areas may be riskier than those from secure areas. The Lending Club provides the borrower’s state, as well as a zip code that has been anonymized to the regional level. While these fields may be useful, sparseness in the data is an issue. Particularly remote or rural regions tend to have few loans represented in the data. This presents a problem for using the state or zip code alone as a feature.

Beyond financial and geographic data, the loan purpose may be a strong predictor of default risk. Specifically, loans for physical property may be less risky, as the borrower is likely to have more “skin in the game.” Conversely, loans for small businesses may be riskier, as larger economic trends may have a greater impact on the ability of the borrower to pay.

Limitations

This data set has several limitations that result from features that have inconsistent meanings, have high dimensionality and sparsity, and have been deprecated. Some features were introduced later in the data set, but this limitation can be overcome via imputation. Critically, the inconsistency and dimensionality/sparsity make some features difficulty to use as features.

With respect to inconsistency, the employer title is a potentially useful field that has little predictive power. Specifically, this field changed from the name of the employer to the job title in the middle of the data set. Additionally, this feature is a text field with tremendous variability. For example, “Manager” is the most common entry. Even when punctuation, leading/trailing text, and capitalization have been stripped, this entry represents less than 1% of entries. Given how inconsistent the feature is and how rare the elements are, it seems like this feature is not promising with respect to the goal of this analysis.

With regards to dimensionality/sparsity, the geographic features are both high in number and low in representative cases. While the dimensionality is lower than that of the titles, it is still high enough that it renders prediction difficult. Additionally, the number of loans from any one geographic location is such that it has limited predictive potential. That said, the geographic features could be turned into other features based on the qualities of those location (i.e., average regional income, unemployment rate).

Finally, deprecation has imposed some major limitations on the analysis. Critically, the description field was deprecated. This is a potentially rich data source where factors such as reading level, key words, and length could be used to predict loan risk. Unfortunately, this feature was deprecated over privacy concerns. While there is enough data to have some predictive power, it does not make sense to use as a feature going forward, given the privacy concerns. That said, as a P2P lending service, it is understandable that the Lending Club deprecated the field for privacy reasons. If the client intends to use only internal risk models, then I would recommend that the client collect descriptions for later analysis and modeling.

Cleaning and Wrangling

Several features have intentionally missing values that complicate the analysis of these data. Specifically, the months since the last delinquency, last public record, and the last major derogatory record are all null when there hasn’t been a delinquency, public record, or derogatory record. While this makes sense logically, it will not work for most machine learning methods. Additionally, as the data are missing in a structured fashion, simply filling in the field with the minimum, the mean, or the maximum does not make sense.

To solve this problem I considered two options. First, I considered discretizing the range of times and adding a “never,” treating the variable as categorical. I elected not to do that, as it would render my models brittle to my choice of bins the values. Next, I considered imputation based on linear regression. This method is limited, as it assumes a linear relationship between the features and the variable, but it is not as limited as binning in my opinion. That said, the relations among the variables are almost certainly non-linear, so other methods may be appropriate. In general, a linear model accounts for a decent amount of the variance, so I have decided to use one for the time being.

A number of other features also have missing values. However, these values seem to be missing at random and represent a small portion of the data. As such, I elected to impute those values with the mean.

Beyond imputation, I had to convert the categorical variables into dummy variables to make sure that all data are numeric. Additionally, a number of fields consisted of strings and dates that I have processes. Again, I have made these fields into numeric data that can be easily modeled.

After cleaning and imputation, I normalized the data with a min-max scaling function. This method normalizes the data such that all values lie between zero and one. While this is a simple method, I am considering switching to using a z-score as my normalization method.

External Data Sets

As previously mentioned, the zip code and state data are both high dimension and sparse. This makes them more of a challenge to work with and renders them less useful for modeling. As I hypothesize that indicators of economic health in the region will be most predictive of default risk, I have added an IRS data set with tax return data to this project.

The IRS data set has tax return data for all zip codes and includes features such as the average income and average unemployment compensation. I anonymized the zip codes in this data set using the same method as the Lending Club data and then added the relevant values to the Lending Club data. This allowed me to reduce the dimensionality of the geographic data and also capture the features that probably matter most for the risk assessment. That said, there are certainly other data sources that I could add. For example, the BLS has data on unemployment and employment in different sectors that could be useful. The US Census also has other rich data sources that cover regional differences, but the IRS data seem sufficient for the present case.

Preliminary Findings

Many of the features that have a strong relationship with the charged off rate are related to financial security. These include the borrower’s income and debt-to-income ratio. These effects are modest (14 and 16%, respectively) and are statistically reliable based on a bootstrapping analysis. Similarly, the number of delinquencies in the past two years, number of credit inquiries in the past six months, and number of delinquent accounts have reliable effects (13, 22, and 37%, respectively).

With respect to the estimates linked to geography, the regional income, proportion of itemized tax returns, proportion of returns with unemployment compensation, and proportion of returns with an educational credit all have small, but reliable effects (3, 4, 2, and ~2%). These geographically derived features do not have a strong relationship with the charged off rate, but they provide more reliable information than the individual zip codes or states are able to. In this sense, adding the IRS data has made the location features more useful. That said, it remains to be seen how important these features are in any model of default risk.

Finally, with respect to the purpose of the loan, there are some fields for which there are major differences in the default rate. Default rates are reliably higher than the base rate for small business loans (61%). Conversely, default rates for major purchases and cars are reliably lower than the base rate (21 and 45%). There are several other loan purposes that have a reliable relationship with the default rate. Furthermore, there are a number of aspects tied to home ownership that can predict the default rate.

Conclusions

Preliminary data analyses have revealed a number of useful features for modeling. Additionally, data cleaning, the addition of external sources, and imputation has all made more of the features useful to that end. Initially, I had planned much more intensive natural language processing, but that has not proved useful for this application. I am ready to begin developing predictive models of loan risk for this data set. I aim to examine the performance of regularized logistic regression, Naïve Bayes, and Random Forest classification methods in order to predict loan performance. My performance metric will be the area under an ROC curve, as the signal (default) is relatively rare.