Lending Club Loan Risk Assessment

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 help them to assess loan risk, I will build a risk model. Given their limited data, I will be 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 loan statuses and debt grades for over three quarters of a million borrowers. The data set is messy, with many self-reported text fields, and contains many missing items. Critically, the data also includes items pulled from recent credit reports, such as the debt-to-income ratio and history of previous delinquencies. It is interesting to note that no credit scores are included in these data. Such scores will be useful for my client’s own internal models, but without access to those values now, the current model will have to do without that information.

To approach this problem, I will begin by attempting to extract new features using natural language processing. This data set includes text fields where the borrower includes a brief statement of the purpose of the loan. This is a potentially rich area with many indicators of risk. I will also need to clean the employment title field, as its use seems to have changed over time. Nevertheless, it is a valuable information source. Ultimately, the lack of credit scores means that I will have to put a lot of work into engineering new features given the available data.

Once I have created the new features, I will start with a regularized logistic regression approach to this problem. Logistic regression one of the simplest risk assessment tools available and regularization is helpful in avoiding the problem of over-fitting. In my modeling, I will also consider Naïve Bayes and Random Forest methods. Eventually, I may opt to create an ensemble of many models with a voting structure in order to most accurately assess borrower risk. All of these models will be trained and tested using k-fold cross-validation in order to ensure good model performance while avoiding over-fitting.

When I have developed a solid model or ensemble of models, I will deliver a suite of Python functions that will predict the loan risk of a borrower, given some information. The suite of functions will be tolerant of missing data and should also provide an assessment of confidence in the risk estimate. Additionally, I will provide a short written description of my findings, as well as a slide deck. These documents will help explain my methods and findings and suggest next steps for my client.