**Project Proposal: Machine Learning Analysis of Lending Club**

David Stevens

Lending Club is fairly new company that offers peer-to-peer (P2P) lending. P2P lending is an attractive option for people in need of a loan, as it cuts out the traditional financial intermediaries (banks, credit unions, etc.) leading to lower interest rates. It is also an interesting alternative for investors, providing a different set of risks and considerations than the stock and bond markets.

Though Lending Club provides tools to prospective investors to help analyze available loans and select a portfolio that matches their specific risk tolerance, it is my understanding that credit store and interest rate are the only factors these tools take into consideration. However, they publish historical records of all accepted, rejected and pending loans with all related published data for each in a nice .csv format. This includes major factors like loan amount, interest rate and credit score, as well as many other factors, like employment status/history, salary, and even education. I believe there is great potential in applying machine learning techniques to this data to try to get a deeper assessment of the risk/reward potential of a given loan and could even be used to improve investment performance.

There is also a potential to take the counter perspective and apply machine learning techniques to try to benefit the potential lendee. Though I’m sure Lending Club’s application is rigorous enough that most of the features would be hard to fake (they already reject about 90% of applications and likely do a thorough background screen in addition to a credit assessment), the description field is particularly susceptible to manipulation as the applicant has more freedom to manipulate it without being deemed fraudulent. If this attribute is found to have a significant impact on loan acceptance and funding (interest rate is almost exclusively controlled by credit score and loan amount/length), it would be interesting to look at what features would help to maximize the lendee’s chances.

**Other considerations**

Though the published datasets provided help to identify a feature set, there are quite a few more factors to consider for this project. One of the most important is how to measure *success*. A simple model for the investor would be that a loan was a successful if it did not default and unsuccessful otherwise. More sophisticated models could try to predict a default rate and even take interest rate into account to develop a better understanding of risk than that provided by Lending Club. From the lendee’s perspective success can be measured by acceptance and percent funded.

Another important consideration is what methods are most applicable/efficient. A quick Internet search has found a few articles about applying logistic regression and support vector machines to this problem. If these approaches show room for improvement, I can attempt to make the approach more sophisticated, or I can try other methods (e.g. random forest). Either way, a thorough analysis will require me to try a few different approaches. If I get particularly good results, I can put my model to the ultimate test and try and use it to make some money!