Capstone Proposal

Analysis of Kiva Microloan Data

29 June 2016

Peter B. Pearman

Introduction. Kiva (https://www.kiva.org/) is a 501(c)3 nonprofit organization in the United States that provides micro-loans to people who would otherwise likely not have access to opportunities for getting business loans. The loans vary in size, but more than half of them have been less than about 500 USD. Constituting a type of crowd funding, one can participate as a lender in the Kiva loan process directly at the Kiva website or with one of several apps for mobile devices. Lenders are motivated to participate because of feelings of social responsibility at a grand scale, because lenders do not receive interest on the funds they provide. I chose to work with the Kiva data because I am interested in development issues, curious about the Kiva loan program, and would be happy to find eventually a data science position in finance or a development organization.

Client. The clients for this project are the potential lenders, who want to know where kiva microloans are occurring, what sectors they are impacting, and how they have changed through time. The project will help contribute transparency to the Kiva program.

Data. Snap-shots of Kiva loan and lender data are produced irregularly and are publically available for download. There are three types of files: (1) loan files that provide information on the recipient, the purpose for which the loan will be used, and loan details of amounts and payments; (2) lender files that provide variables that describe a little about the people who provide the funds for the loan, their approximate location, and a statement of motivation; and (3) loans-lender files that match the loans and the identity of the lenders who contribute to the loans. The loan files are nearly 2100 in number and each contains descriptions of 500 loans. There is a similar number of lender files. Initial examination of the data indicates that the two countries with the highest Kiva activity are the Philippines and Kenya. I will focus initially on Kenya because the size of the corresponding data frame (about 0.7GB) is more manageable on a notebook computer than that of the Philippines (over 2GB). In both countries, loans have been labeled as being made to a number societal sectors (activities), and to individuals and groups of varying sizes. Previous work indicates that recipients are primarily female, although a breakdown by sectors and countries is not readily available.

The goal of this project is to contribute to understanding loan activity that is conducted through Kiva. I present several research questions to stimulate work toward this goal. As a response to each question, I outline achievable objectives and briefly describe the work that will accomplish them. I also provide a working hypothesis (an expectation), the testing of which will guide inference regarding the trends revealed during the analysis.

Question 1. **How do different sectors and activities receive Kiva funding and has the distribution of financial support to these sectors changed over time?**
Objective: Develop analyses that present the temporal course of the distribution of both (a) the number of loans to the various sectors and (b) descriptive statistics of those loans as partitioned to the sectors.
Any understanding of Kiva loan activity depends on quantification of the distribution of loans in the differing economic sectors of society. The analyses to address this issue are both graphical and statistical. First, the sectors represented in the data will need to be aggregated because they are currently vague, indistinct and potentially overlapping. I will develop plots of counts to sector and descriptive statistics. These plots will incorporate a temporal axis, either directly or as a temporal succession of sub-plots. I expect that loans to different segments or activities have been constant over time, since I have no basis for suspecting otherwise.
Modeling and statistics: Based on physical examination, I will fit linear or non-linear models as necessary to the data, then test for either non-zero slopes or improved fit provided by the non-linear components. Where sufficient data exist, I should be able to determine differences among data partitions in terms of the higher moments of the distribution in loan size by applying a Kolomogorov-Smirnov test, or similar.
Outcome: These analyses will reveal any changes in the distribution of microloan activity to sectors and and the temporal evolution of these differences.

Question 2. How is gender of recipient and the composition and size of recipient groups associated with loan characteristics and has this changed over time?
Objective: Develop analyses that characterize loans by recipient characteristics and represent these while incorporating a temporal component.
Loan recipients in the Kiva program are over 80% female. These analyses are similar to the above in that they are also graphical and statistical. I will partition counts of loans by sector, a temporal axis (e.g. year) and gender of the recipients (in the case of single recipients). By partitioning single-recipient loans by gender I will be able to determine whether changing loan activity is gender-specific. I will also examine the relationship between number of participants, gender composition of recipient groups, and loan amounts. If possible, I will examine these characteristics with respect to societal sector by further partitioning the data by sector. I expect that there will be some variation among sectors/activities in the proportion of recipients that are women, since there is no reason to assume an invariant proportion across the activities. Because most approved recipients are women, I naively expect that gender proportion of unfunded loans is the same as for funded loans, since I have no reason to assume otherwise.
Modeling and statistics: I will examine graphical presentations of the data for temporal trends in the proportion of recipients that are women and fit linear or non-linear models as appropriate. Similar to the previous case, I will test for linear temporal trends, or significantly improved fit provided by non-linear parameters.
Outcome: These analyses will potentially illuminate changes in Kiva loan activity as associated with gender. As women in developing lands likely continue to have reduced participation in cash economies in comparison with men, these analyses could potentially illuminate trends in the ability of Kiva loans to impact activity in economically disadvantaged segments of society.

Question 3. How is the financial opportunity provided by Kiva distributed spatially within a focal country?

Objective: Conduct a spatial analysis of the distribution of loan activity.

The rapid growth of urban areas suggests that economic activity, including the lender-recipient activity of Kiva, will be concentrated in urban centers. The issue is whether this activity is disproportionately focused on urban society. One might expect loan size to correlate positively with community population. Kiva loan data records are labeled with indicators of the rural/urban location of the activity. Some records have geographically-informative coordinate data (some coming with GPS-level precision). In other cases, the community of the recipient is stated, while many records may lack geographically-informative information. I will collect additional data on community population size from on-line sources. I then will present loan characteristic descriptors on a country map, using such 2-D representations as variably sized circles. I will also develop plots to show the relationship between loan activity and community size. I expect that sparsely populated areas will show disproportionately low levels of funding, because of the isolation of potential loan applicants.

Modeling and statistics: I will develop linear and non-linear models of relationships between loan characteristics and community population size. Geographic distribution of loan activity will be shown superimposed on country maps.

Outcome: These analyses will illustrate how urban/rural differences are associated with loan characteristics. Rural communities generally have less access to many services, suggesting a potential for discrepancies in loan access. Map areas with low Kiva activity could point to areas in which improved loan access could have relatively strong impacts on local economies.

Question 4. Is repayment success related to any characteristics of the loans, as represented in the data?

Objective: Identify potential correlates of poor loan performance or default.

Default rates for Kiva loans are seemingly low; in the data for Kenya less than 5% of loans are listed as being in default. However, each of those loans represents a loss for willing investors, in spite of the recipient having passed a qualifying process. These analyses will develop predictive models of loan default. I will develop a set of predictor variables and apply these in a predictive machine learning exercise. In addition, I will make a series of graphs that will plot modeled probabilities of default against predictor variables. I expect that there will be only weak signals in the data, with regard to default potential, because loan recipients have already been scrutinized for credit worthiness by Kiva's local partners and have obtained (non-monetary) sponsorship from local community entities as part of the application process.

Modeling and statistics: I will use predictive modeling algorithms available in Python. Model training and optimization will be based on 10x cross-validation of performance statistics, including ones derived from confusion matrices (specificity, sensitivity, AUC-ROC and TSS). The data set will be randomly divided into 10 partitions. In each training trial, nine partitions will be pooled and used for training and the remaining partition used for validation. Model performance will then be reported in terms of means of the 10 trials. Model pruning will be conducted to produce relatively parsimonious models that should perform well outside of the current data set.

Outcome: A quantitative basis for identifying default risk in approved applicants could both improve the evaluations conducted for loan qualification, and potentially direct additional resources and assistance (such as planning or accounting help) to recipients

that experience higher risk of loan default. If no functionally predictive relationships are found within the data, this in itself will be informative. In this case, other predictors would need to be identified.

Deliverables. The project will produce python code, in the form of one or more ipython notebooks, and a pdf of a report, illustrated with graphics.