|  |
| --- |
| Practical Data Science |
| Project report |
|  |
|  |
| **Team # 3**  **Suresh Rangarajan**  **Charlotte George**  **David Cohodes** |
|  |

**12/19/2012**

|  |
| --- |
| [Type the abstract of the document here. The abstract is typically a short summary of the contents of the document. Type the abstract of the document here. The abstract is typically a short summary of the contents of the document.] |

Summary

**Data Source**

Kiva

URL: <http://build.kiva.org/>

Kiva provides a RESTful Web-service API for accessing data on lenders, loans and other related Kiva objects. Kiva API returns the response in XML and JSON formats depending on the URL accessed. We downloaded the bulk data provided by Kiva from the URL <http://s3.kiva.org/snapshots/kiva_ds_json.zip> to do an initial analysis of the data and came up with a few models and questions that we can answer with the data. We created a training and test database with tables for the major kiva objects. We loaded the bulk data for lenders and loans using python. We also created python scripts to fetch some live data from Kiva API and loaded that data as well into the training and test databases to ensure that the models apply to old as well as new data. We also loaded other tables such as country, loan->lenders and lender -> loans to analyze the relationships between the objects.

We reused some of the data model from <http://www.kivadata.org/> website.

**1.**

**We used python visualization to map the locations of lenders and borrowers to see the trend in kiva loans.**

**We could not do a mapping between lenders and borrower’s demographics since data about lender demographics is not available except for their location and occupation.**

**4.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sector** | **Funded** | **Not funded** | **% Funded** |
| Agriculture | 67459 | 3419 | 95.17622 |
| Arts | 7389 | 167 | 97.78984 |
| Clothing | 22617 | 1276 | 94.65952 |
| Construction | 6478 | 211 | 96.84557 |
| Education | 2838 | 70 | 97.59285 |
| Entertainment | 599 | 15 | 97.557 |
| Food | 85674 | 3610 | 95.95672 |
| Health | 2734 | 79 | 97.19161 |
| Housing | 9047 | 1175 | 88.50518 |
| Manufacturing | 4601 | 100 | 97.87279 |
| Personal Use | 3074 | 248 | 92.53462 |
| Retail | 75379 | 4026 | 94.92979 |
| Services | 25878 | 1193 | 95.59307 |
| Transportation | 11064 | 599 | 94.8641 |
| Wholesale | 805 | 32 | 96.17682 |

**Word cloud of lender loaning reason and loan uses:**

**Activity/Sectory histogram/bar chart**

1. Kiva runs promotion campaigns to incentivize lenders to make loans. We would like to identify patterns in the existing data to make these campaigns more effective.
   1. For example, we would try to map the locations of the lenders and the borrowers of loans and identify patterns to promote specific borrowers’ loans to lenders. Here, visualization may help a campaign sponsor to make subjective decisions.
   2. Map the demographics of the lenders and borrowers to target the promotion campaigns effectively.
2. Build a model that calculates certain probabilities:
   1. Probability of a loan (Kiva loan score?) being fully funded given the industry, location and other relevant parameters.
   2. Probability of a borrower receiving a loan given an industry, location and other relevant parameters from a specific lender
3. Where are the most loans made
   1. Aggregate data by country and potentially city
   2. Graphically plot where the most loans are made on a map
4. Analyze the demographics of borrowers to find out the following:
   1. Look at whether specific industries attract more loans in general.
   2. Are there specific triggers in an individual’s profile that make them more likely to get a loan (specific key words, location, industry, age)?
5. Analyze the demographics of lenders to find out the following:
   1. For a lender who has registered with Kiva but has never made a loan, what is the best approach to help him make a loan?
   2. For a lender who has not made a loan recently, what would be the best approach to help him make a loan again?
6. Analyze the data on teams and loans to determine the following:
   1. Which teams make the most loans?
      1. Analyze data set to sort loan volume by team and aggregate this information
      2. Graph teams and corresponding loans
   2. Is there anything in common between the teams that make the most loans?
      1. Use real expressions to see if there are any key words that tip off whether a specific group will make loans
   3. Analyze whether there are indicators as to whether a specific team will make multiple loans to the same individual
   4. Analyze whether the number of members of team has any impact on the volume of loans made.