Dheeraj Navani
Gaurang Patel
Varun Sharma

Assignment 2

Task 1 - Exploration

Introduction/Abstract

Kiva is a service that provides microloans to people in third-world countries without access to traditional banking systems. Through a classification analysis we were able to determine certain attributes which make a potential loan candidate superior to others.

Dataset Description

The Kiva dataset provided by their API was used as the basis of this work. The dataset being used is a combination of Kiva loan data files in JSON format, which include every single Kiva loan that was ever given out. This totals to 886,541 loans. Loan entities, which have a rich set of information, are described by a loan description, a loan sector (e.g., agriculture, food, retail, etc.), a list of borrowers requesting the loan, a field partner, a geo-location, a loan amount, a description, and posted/funded/paid timestamps, amongst others.

There are a few objects in the Kiva world that are important: a loan, borrower, lender, and partner. A loan is requested by a borrower and supplied by either an individual lender or a partner, a microfinance institution that partners with Kiva.

The full dataset has 886541 loan listings. Our goal was to represent the full dataset in a smaller subset by representing the weights of different features, like gender, loan status, sectors, and country of origin, with similar weights. Of course, each of the feature percentages might be difficult to preserve, so we did some exploration to determine how our subset should compared to the full dataset. A couple of statistics stood out in the full dataset. One, females were the majority of borrowers, as depicted in Figure 1. This led to us looking directly at the dataset. A manual scan showed that the word ‘children’ was a frequently mentioned word at an average rate of 1.03 occurrences per description, which gives preliminary indications that the loans were needed for families.

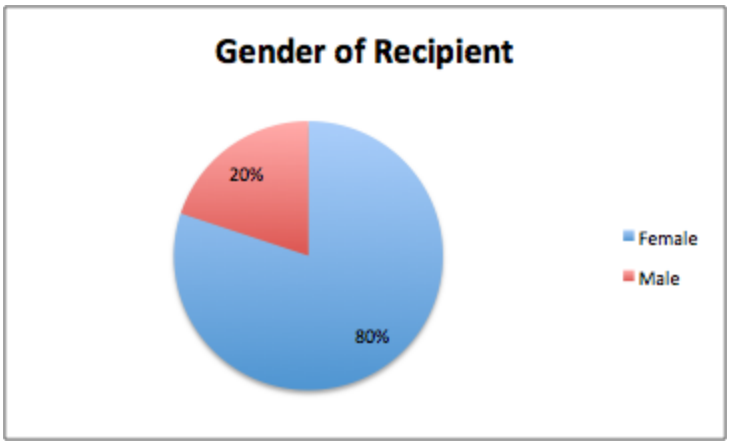

Figure 1. Borrowers gender graph.

Secondly, the majority of loan requests came from Peru, Cambodia, the Philippines, Eastern Europe, Central America, and parts of Africa. Seeing that the majority of loans were provided to *all* and *any* regions considered 'developing', this statistic was initially thought to not be too important to maintain.

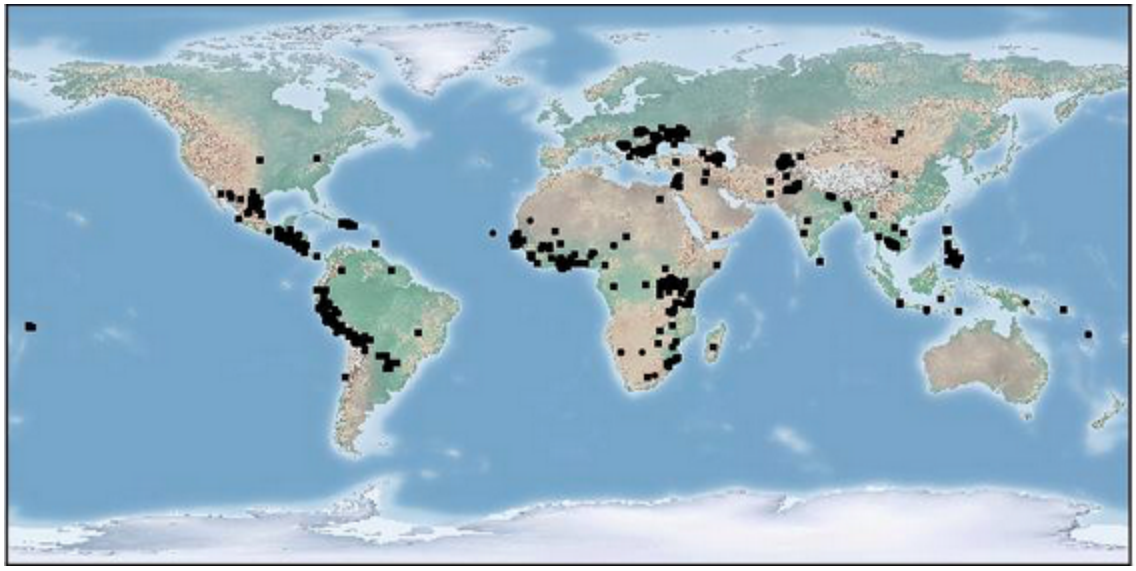

Figure 2. World map plotting of all loan recipients, based on their geo-coordinates.

Thirdly, loans were provided across many sectors. Borrower's self-reported indications reveal that the loans were primarily needed for retail, agriculture, food, or clothing. While those four needs reflect a total 76% of the loans, it should be noted that educational loans were only needed for 2%. It seems that borrowers tend to request money for business providing basic needs like food and economic sufficiency. This fact along with the aforementioned findings indicate that these individuals could be high risk recipients who may not be able to pay back the loans.So this brings up the question: how likely are these loans to be paid back?

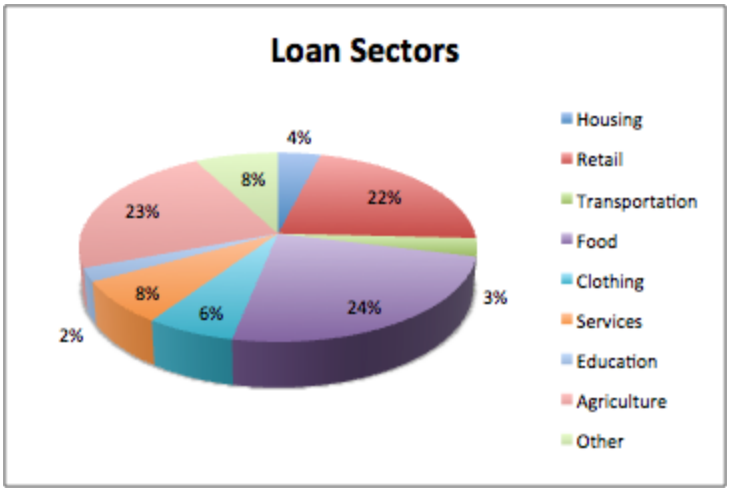

Figure 3. Loan distributions across sectors.

This led to an analysis of the status of these loans. Only 2% of the total loans provided were in default, and the rest were either in repayment or paid.

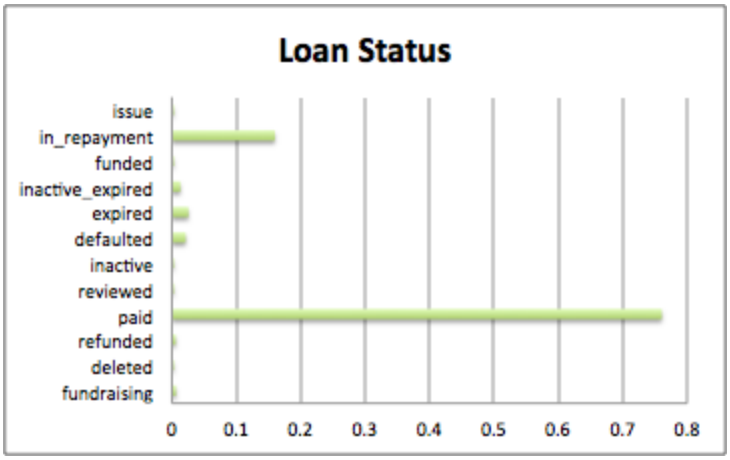

Figure 4. The status of loans represented as a decimal percentage.

Our analysis attempts to predict what features of a loan listing are likely to determine whether a loan will be defaulted or not. The number of defaulted loans in the entire dataset was very low, so the smaller subsets were adjusted to balance the weight of default loans and paid loans (see predictive task for more details).

Task 2 - Predictive Task

The predictive task we identified is to predict if a loan will default based on some core features. We were presented with a large amount of features but we quickly realized many of the features lacked independence and that Kiva strips out certain features for defaulted borrowers like the loan description. This led us to an examination of much of what we thought were core features, such as gender, loan amount, comma use, and length use, amongst others. The following analysis of these features determined which of them would be best to include in our model.

First, a distribution of funded loan amounts revealed that the majority of defaulted loan borrowers did not request large amounts, often less than $1500, whereas the loans that were paid back often asked for more money. See figure 5. Next, we analyzed the difference across genders for defaulted vs paid loans. After normalizing for the fact that 80% of the all borrowers are female, we still found that women tended to both default more often and pay back loans more often than men did. This consistency does not lead to any conclusion about which gender is more likely to default on a loan. Therefore, we removed gender as a predictive feature, increasing our prediction accuracies across all classifiers.

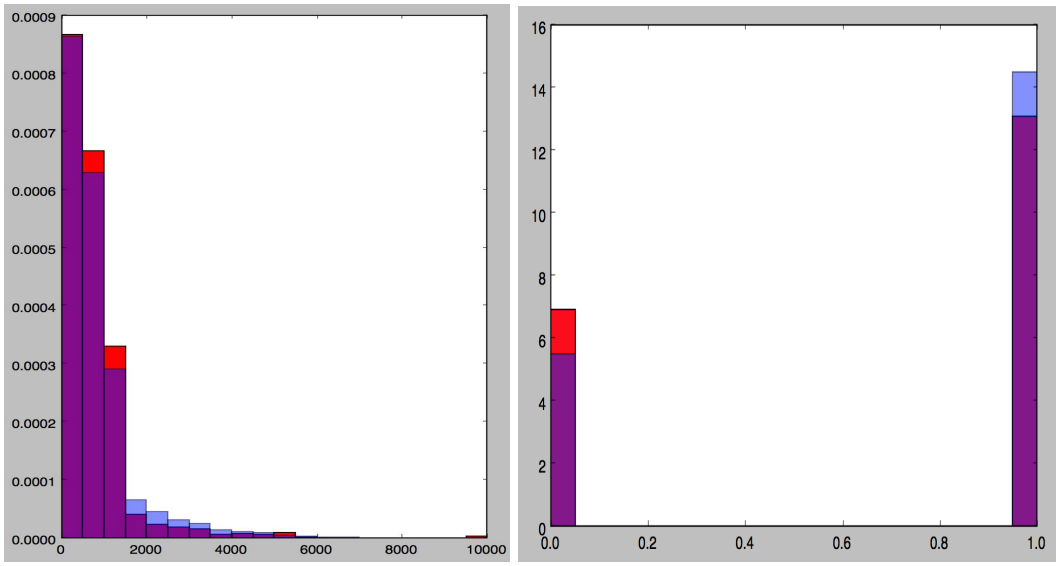

Figure 5. In both graphs, defaulted are represented as red bars and paid are blue. On the left: Histogram distribution of funded loan amounts, where the y-axis is percentage and the x-axis is loan amount. On the right: Gender Histogram, where 0 is males, and 1 is females on the x-axis. The y-axis represents a normalized fraction of all borrowers.

Looking at the length of the use descriptions, we found that defaulted loans distributed normally around a median of 30-40 characters, while paid loans were constantly distributed from 0 to 100 characters. This difference in use descriptions across the two statuses of loans led us to believe that this feature would be important for classification. See figure 6.

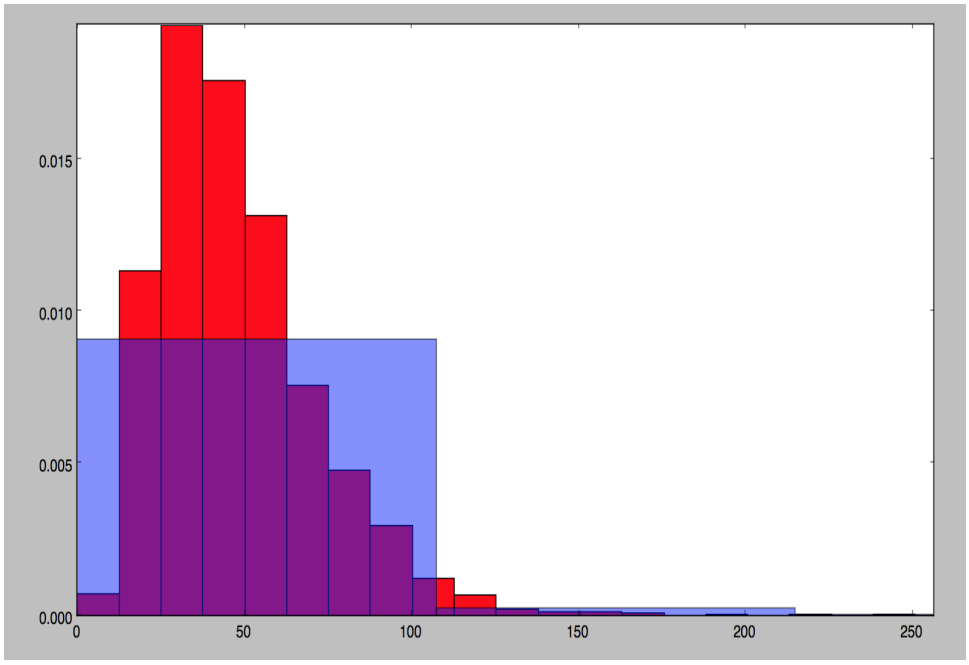

Figure 6. Distribution of the length of the use description of a loan in character count. Defaulted = Red, Paid = Blue.

Before continuing to describe which features were implemented in our classifiers, we will first describe our subset data, baselines, and how we evaluated our models.

We needed 3 data subsets, one that we could train on which would be equally distributed between defaulted and paid loans call that the train dataset. This train dataset would take up 50% of the positively labeled and negatively labeled data. Next we needed a test dataset which would reflect the true distribution of the defaulted and paid loans. This would reflect the total unseen dataset the other 50% of data. Lastly we would want a 50% defaulted and 50% paid dataset which we could evaluate our classifier on. This would have to reflect a much smaller percentage of the total data.

The first major problem we faced was generating the test dataset which was equally distributed between positively and negatively labeled points. Of the full dataset we had 18,899 defaulted points and only 674,294 paid loans so only 2.7% of our data was negatively labeled. This test dataset would take up 50% of our total data set so so we randomized the order of both the defaulted and paid loans and split the data in half.

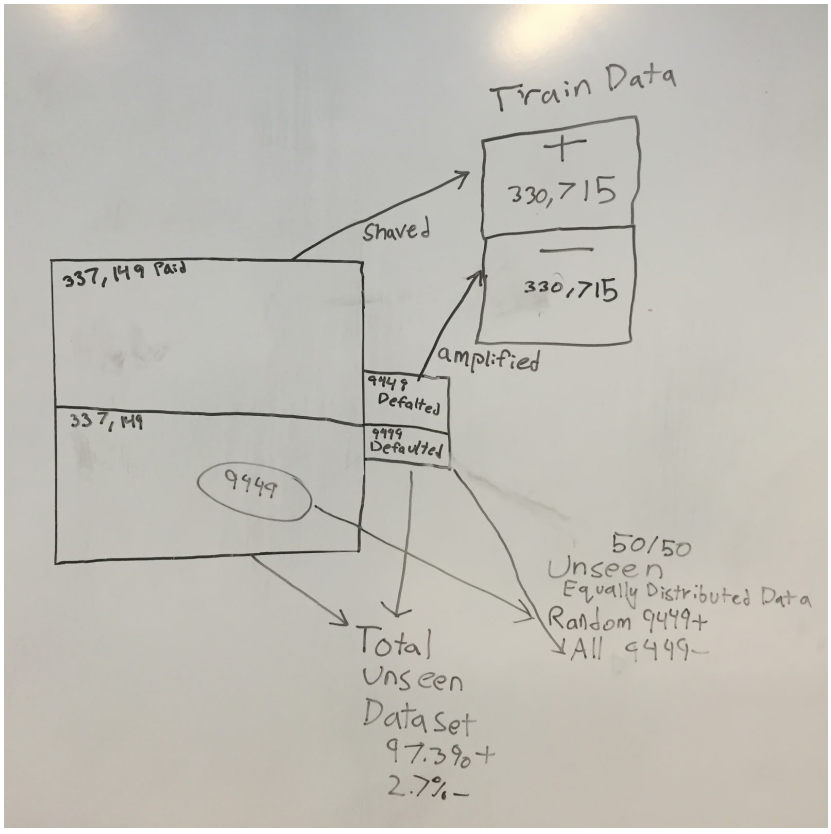

Figure 7 Construction of Datasets

After dividing the total dataset in half we retained the imbalance of data in our training subset with 337,149 paid loans and 9449 defaulted loans. We started to balance the two sets by amplifying our negatively labeled points, repeating the negative points 35 times resulting in a

new negatively labeled subset of 330,715 points. Now there was 330,715 negatively labeled points for testing and 337,149 positively labeled points. Obviously this data is still imbalanced so we randomly shaved 6,434 from the positive training data to arrive at 330,715 positively labeled points to pair with our 330,715 negative points so we finally arrived at a 50% positive/50% negative training set.

To construct the total test dataset we used the 50% of non trained data was and it retained its distribution of 97.3% being paid and 2.7% defaulted. Lastly we needed to construct an equally distributed testing dataset so we took the 9449 points we never trained on from the negative dataset and 9449 random points positive total unseen dataset and the to construct the equally distributed unseen dataset. Refer to figure 6 for a diagram of how the datasets were divided.

With these 2 test datasets we can establish some basic baselines with which we should compare our model against. The total dataset where the distribution of loans is unbalanced the naive baseline would be to predict paid every time which would result in a 97% accuracy, for the equally distributed dataset the baseline is a random classifier performing on average 50% correct.

Since this was a classification task we set up a few basic metrics to evaluate our classifiers by beyond the basic accuracy. The primary ones used were the True Positive, False Positive, False Negative, and True Negative rates, and the balanced error rate. By using these 5 metrics we could determine how our predictor was obtaining its results. For example, if we naively predicted one for the unequally distributed test dataset we would obviously get 97.3% accuracy. Since that represents the percentage of all paid loans our true negative rate would be 0. These metrics helped us gauge exactly how our model was performing on each subset of data.

To form our predictions, we first used a Logistic Regression classifier. We chose this classifier because it was the simplest to implement: we had previously learnt about the operation of the classifier in class and simply had to plug in our feature and label vectors into the scikit library. The Logistic Regression classifier attempted to describe the positive and negative labels by fitting a linear hyperplane onto the feature data that would separate positive instances from negative. It would then produce a series of ratios or weights that described the importance of each feature. The scikit library automatically calculated all the ratios necessary to fit the data to the model and handled the predictions based on the model as well. Initially, the classifier performed exceptionally well. It predicted whether or not the loan would be paid off with 99.89% accuracy. This figure seemed to be too perfect, and any other classifier we used had a similarly high accuracy. So, we examined the features that were given the highest weights. This showed us that the length of the description was extremely influential. Intuitively, this made sense because it indicated that those who invested the most effort in writing a detailed description would be much more likely to pay off their loan than others.

However, upon further examination, we realized that the descriptions for defaulted loans were nullified. So, the incompleteness of the data was used to characterize our predictions. When we removed this feature, the linear classifier performed poorly on the training data: 50% accuracy, which meant that it was equivalent to a random predictor. We then experimented to modify our features to decrease the training error. After a marginally successful attempt, we examined the data again.

Graphing the distributions of the features, as shown above, showed us that there was not a simple linear cut through the data that would clearly separate the negative and positive values. Instead, we chose to change our classifier. We wanted something more flexible and adaptive to the training data so we selected the Decision Tree classifier. Instead of forming one cut across all the features, the Decision Tree algorithm formed multiple small cuts along each feature, which customized the predictor further. This classifier significantly outperformed the logistic regressor on our training data. Based on this validation, we deduced that our assumption was correct. The feature data was not linearly separable so the Decision Tree algorithm was more suited for the data. To us what that means is there might not be one dimension or a set of dimensions which many defaulted loans share rather there certain subsets and ranges of attributes which share defaults in common.

Since adding flexibility to the predictor promoted its accuracy, we decided to use an even more flexible model: K-Nearest Neighbor. This model stores all of the training data. To classify a test point, it computes the Euclidean distance between itself and all other training points and selects the majority label corresponding to the K minimum distances. This model provides us with more flexibility than decision trees because the boundaries of decision trees are always vertical or horizontal cuts in regards to the graph of two features. On the other hand, the decision boundaries of K-Nearest Neighbor can be much more complex because they can involve diagonal and curved cuts as well.

To select the features we used in our classifiers, we first used a list of features that we thought would be most indicative: the loan amount, the sector, the length of the description, the lender count, the repayment term, the borrower count, and the gender of the lender. When we found out about the description issue (referred to above), we removed the length of the description from our feature vector. Also, we used logistic regression to find the weights of the least relevant features and remove those from our feature vector while adding in new features, as well. Our final feature vector included funded amount, repayment term, gender, activity, country, loss liability, length of the “use” field, number of commas in the “use” field, and number of periods in the “use” field.

Task 3 - Literature

It is also important to discuss the origin of the data set. This data represents actual data collected by Kiva’s website. All requests and contributions are made via their online interface, and Kiva simply compiles their database into JSON files on a monthly basis (1). However, a critical analysis of the dataset reveals (unfortunately after we made our predictions) that

intermediate microfinance partners, like MAXIMA, cover for lenders who default to keep the repayment rates of Kiva higher than usual (2). This insight explains why so few of the reported loans are in default even when the actual percentage of defaulted loans is much higher. Though we do not blame Kiva for this discrepancy, had we had access to all defaulted loans, we could have predicted more realistically, but we made do with what we were given.

We found other researchers who did too, but to the best of our knowledge, our predictive task has not been resolved. The most closely-related work is the classification of lenders' motivations for lending, using the same Kiva dataset. This group classified the lenders' self-stated motivations into ten categories with human coders and machine learning based classifiers. They employed text classifiers using lexical features, along with social features based on lender activity information on Kiva, to predict the categories of lender motivation statements (3). This study, along with (4) and (5), prove that the dataset is large and rich enough for data analysis. Moreover, they implemented predictive binary classifiers, support vector machine classifiers, regression, naive bayes, bag-of-words models, and multi-class classifiers. All of which were considered in our study, since our task is primarily one of binary classification and some text analysis.

However, other studies have attempted to understand what constitutes a reliable borrower in domains other than microfinance charities, such as in banking. Since these loans come with interest and Kiva loans do not, there may be some discrepancies that are worth noting. Nonetheless, these studies try to understand personality traits associated with defaulting on a loan, i.e one study suggests that conscientiousness, extraversion, and neuroticism have significant positive associations with defaults, while cognitive ability has a negative one (6). Another study indicates the exact opposite (7), so no real conclusion has been made on what personality traits characterize a good borrower. Other studies explore reasons for defaulting, i.e. low income is seen as a usual condition that leads to non-payment (8). Nonetheless, the question of what qualifies a good borrower from Kiva, who offers loans with no interest, is yet to be answered, and therefore, is the scope of this paper.

Task 4 - Results and Conclusions

Final Results

	Logistic Regressor	Decision Tree	5-Nearest Neighbor
Training Accuracy	0.768900715117	0.987455966618	0.961906777739
Test Accuracy	0.766906550958	0.767435707482	0.791935654567
Total Unseen Accuracy	0.751128867252	0.950082901189	0.898090332898

After running all the models on the test set, we found that the 5-nearest neighbor model performs the most accurately by a slight margin. The decision tree model performed much better on the total unseen data set, while the logistic regressor performed at a consistent ~75%

accuracy. At the least, each of these classifiers outperform the baseline that predicts with 50% accuracy.

However, the second baseline which predicts 1 each time would outperform each of these classifiers. This comparison makes sense of the fact that the majority of the Kiva dataset loans are in a paid status, whereas only a few are defaulted. In the case that we were predicting on a dataset similar to the original one, this baseline or the decision tree classifier would probably be a stronger choice. However, given a dataset similar to our training subset, which has a 50-50 split, we would benefit from using the 5-nearest neighbor predictor since it outperforms the others on the test set. The logistic regressor is unfortunately not adaptive enough to the variation in features.

	LR Train	DT Train	5-NN Train	LR Test	DT Test	5-NN Test	LR Total	DT Total	5-NN Total
True Positives	248069	323013	306674	7060	9216	8757	248074	317897	299289
False Negative	82646	7702	24041	2389	233	692	82641	12818	31426
True Negative	260505	330120	329560	7433	5287	6209	7433	5287	6209
False Positives	70210	595	1155	2016	4162	3240	2016	4162	3240
True Positive Rate	0.75010	0.97671	0.92731	0.74717	.97534	.92676	0.75011	.96124	.90498
False Positive Rate	0.21230	0.00179	0.00349	.21336	.44047	.34289	0.21336	.44047	0.34289
True Negative Rate	.78770	0.99820	0.99650	.78664	.55953	.65711	0.78664	.55953	.65711
False Negative Rate	.24990	0.02328	0.07269	.25283	.02466	.07324	0.24988	.03876	.09502
Balanced Error Rate	.23110	0.01254	0.03809	.23309	.23256	.20806	.23162	.23961	.21896

Based on the table above its pretty clear that all the models we try fail almost all in same ways just in different degrees. The false positive, false negative, true positive and true negative rates are pretty consistent across the three classifiers showing that it wasn't the model which made the biggest difference rather it was feature construction.

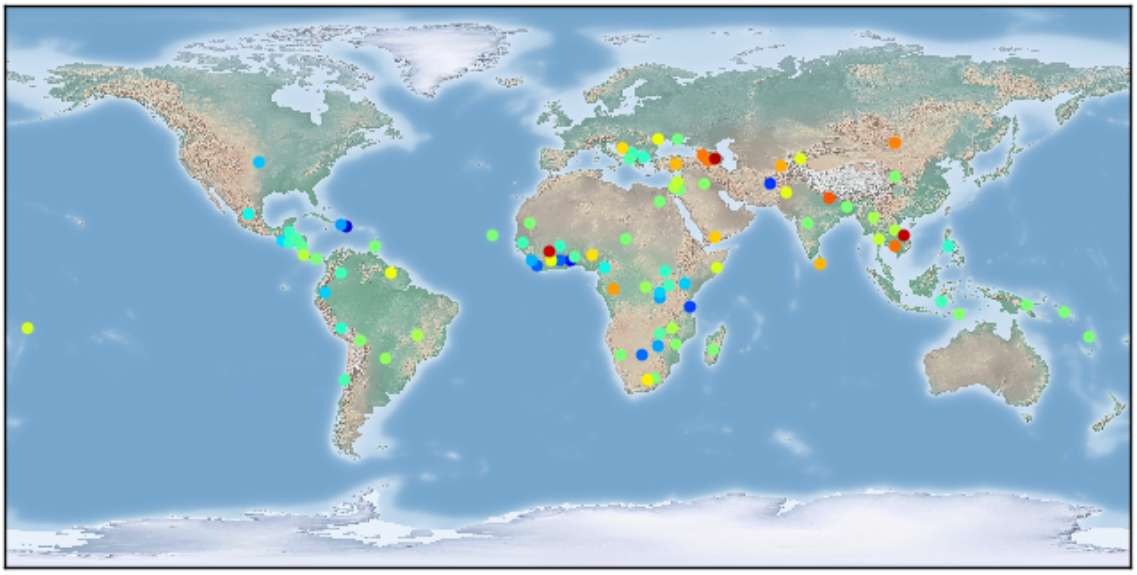

Figure 8 Likelihood of Loans To Be Paid Back by Country
on a scale from blue (defaulted) to red (paid)

It surprised us that the country of origin of the borrower played a large impact on whether a loan would be defaulted or not, while gender was insignificant to our predictions. Above is a plot of the countries with the color of their pointers referring to whether the loan would likely be defaulted (blue) or paid (red). Countries highlighted with the green color did not have significant bearing on the results. The following countries were considered most likely to pay back and default:

Most likely to pay back	Most likely to default
Azerbaijan	Dominican Republic
Vietnam	Togo
Mali	Afghanistan
Nepal	Tanzania
Cambodia	Liberia

Next, our classifications indicate that the loans that are most likely to be paid back have these features: high funded amount; shorter repayment term; female; activities: Land Rental, Religious Articles, musical instruments, Personal Medical Expenses, Consumer Goods; country: country in top country list; partner; and longer "use" description. Religious articles being a strong predictive feature might indicate the moral duties are highly followed by religious borrowers. It's likely that females are predictive since they represent a majority of the dataset. Land rentals probably generate a stable source of revenue so that income could explain why they were able to pay back the loans often.

What is most interesting is that a loan financed by a partner institution was more likely to be paid back than a loan that wasn't. While outside the scope of this class, this supports our initial suspicions that Kiva does not accurately report the loans that go into default. Partner institutions may be covering for Kiva by paying off the defaulted loans so that Kiva maintains a high repayment statistic across their dataset and that the partners retain that stream of money coming which they can use for their lending purposes.

References

(1) Kiva data set -> http://build.kiva.org/docs/
(2) Roodman, David. "Kiva Is Not Quite What It Seems." Center For Global Development. Center for Global Development, 2 Oct. 2009. Web. 02 June 2015.
(3) Chen, Roy W. Social Identity and Cooperation. Doctoral Dissertation. The University of Michigan, 2012.
(4) Liu, Yang, et al. "I loan because...: understanding motivations for pro-social lending." Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012.
(5) Hartley, Scott. "Kiva. org: Crowd-Sourced Microfinance & Cooperation in Group Lending." Harvard University Working Paper, 2010.
(6) Rustichini A, De Young C, Anderson J & Burks S. (2012). Toward the integration of personality theory and decision theory in the explanation of economic and health behavior. IZA Discussion Paper 6750.
(7) Klinger B, Khwaja AI & Carpio CD (2013). Enterprising Psychometrics and Poverty Reduction, Springer.
(8) Bhardwaj & Bhattacharjee (2010). Modeling money attitudes to predict loan default. IUP Journal of Bank Management, 9(1/2).