Predicting User Preferences, Creating Online P2P Lending Teams in Kiva

Julio Sotelo

Introduction

Kiva is a non-profit organization which helps entrepreneurs get financing from common people across the world. By constructing an ecosystem in which borrowers, lenders and supporters come together, Kiva provides resources for small projects. In order make the ecosystem available to as many people as possible, small amount investments are available for anyone to fund and help others start their projects. The minimum investment needed to participate is 25 American dollars. Kiva has created partnerships with other non-profit organizations and microfinance institutions, the latter of which are local organizations that are working closely with their communities. The system gains deeper knowledge of the projects through the field partners, whom are responsible for underwriting this process.

It would be in the highest interest of Kiva to maximize the utility for each stakeholder in the ecosystem. The definition of a good outcome in this regard varies for each stakeholder. A borrower wants to be certain that his funding needs are met. The lender may have different goals, making the benefits for this party different. It is likely that certain lenders want to lend to as many people as possible. In these cases, good repayment rates would help, so that these persons can continue lending, since resources are limited. There may be other goals for a lender, like wanting to support specific countries or activities, while others may wish to fund one project at a time. Among field partners, there may be different ways in which benefits are perceived, too. Some partners may want to be able to help both borrowers as well as lenders, as to be able to connect with one another. For those who fund the borrower in advance, the main concern would be concentrated on getting the funding from Kiva. Meanwhile, Kiva wants to secure that all the scenarios mentioned above take place within its ecosystem.

Related work

This section discusses related works on previous analysis with regards to Kiva data and recommendations for microfinance.

Team membership can improve the amount invested by lenders in a significant way, but does not affect the frequency in which a lender is actively funding projects [1]. Chen, Roy et al. [2] provided a deeper analysis pointing in this direction. Working with Kiva, they implemented a random test, consisting of 22,333 experiments. The authors concluded that goal-setting and coordination are effective mechanisms to increase both lender activity, as well as the invested amount. The study also

shows that once a team is created, the activity it produces is concentrated in the first few months. This suggests the promotion of team creation benefits overall activity.

While the forming of teams may promote overall activity, this is not an everlasting reaction. Chen, Roy et al. show that the activity of a team is high during the initial stage, and then has a rapid decline. The data used in their analysis, a data dump from April 2013, shows that only 25% of all teams were active at that moment. According to the same data, 50% of all teams had not funded a loan in the last year, and almost 90% of them stopped posting at the Kiva team's forum. Forums are used as the team's communication channel; lenders may utilize them to promote loans and coordinate team activity.

According to Choo, J. et al. [3] Kiva lending teams are more detail-oriented when selecting loans that need to be funded. Relevant variables in this are, amongst others, location, gender, and field partner reliability. In addition, the authors found that team behavior can vary based on a lender's background and interests (such as occupation, region, ethnic, and occupational aspects). The paper presents a model for team recommendation to lenders who have no affiliation with a team. The model computes the similarity of a lender and the 200 most popular teams, using data coming out of the lender's loans funded outside the team. To measure the performance, a rank is created for each lender, with a maximum ranking value of 1. On average, the model ranks as a 0.0851

Proposal

In this paper, I propose a recommender algorithm for Kiva, whose goal is to improve activity, by recommending teams to lenders. In addition, and in support the of the recommended algorithm, I have compared how the lender's reasons to fund are matched to team's goals and could be used to promote the creation of teams.

There are two main approaches to recommender systems: content-based filtering and collaborative filtering. Content-based approaches use data created within a system, as to be able to provide recommendations for its users. If the system handles products, then it would take information about the product, such as category, price, color, brand and more, to match the user profile, and then select some products to be suggested to that particular user. The collaborative filtering method uses a similar mechanism between users, herewith suggesting products to each user. In a system where two users are similar (because they, for example, both liked similar movies) the system would suggest something to one user, based on the information of what the second user has seen and liked.

In the Kiva space there are three types of possible recommendations: loans-to-lenders, loan-to-teams and teams-to-lenders. All three recommendation approaches are possible. Since I want to improve activity, I would want to recommend the teams-to-lenders or loans-to-teams. Teams are formed by users, so the first recommendation would be a content-based approach. The loan-to-

teams recommendation can be defined as a collaborative approach, since we need to see what other teams or lenders are doing in order to make recommendations to teams or lenders.

Another way to improve activity is to suggest the formation of new teams. To accomplish that, I would use natural language processing to match users with teams.

Kiva Data

The Kiva data set used for the analysis includes the following data:

- lenders 547,248 in total,
- loans 165,452 in total,
- teams 11,885 in total,
- partners 141 in total.

The attributes within the relations include geo-spatial, categorical, continuous, and unstructured text data. Regarding stakeholders, the attributes contained are as followed: for the lenders, the data has information regarding location, occupation, sign up date, and loan count, as well as information on the number of loans funded by the user, its invitee count, and the number of invitations sent to other users to fund a loan, because the latter is one of the reasons to be a part of Kiva. The team data has category selected from a list of options provided by the system, described as free text loan, because this is a brief description of the overall team goal, loan count, loan amount, member count, membership type (open or closed), date of creation and location. There are no restrictions to join a team with regards to location, but it helps to find affinities: when a new user would like to join a team, the region he or she is in could become one of the first reasons to join. Loan data makes up the largest relation, as it includes the status of the loan with detailed information about delinquency rate, repayment status, sector and more. Activity is a sub sector type of attribute, loan use as a free text to state the purpose of the loan, location, currency and amount.

In addition, the data set has the relations between lenders and teams, lenders to loan, which are many-to-many. A lender is not required to have a team affiliation, nor is he restricted to join only one team. There may be lenders that have joined several teams. The two main relationships I am interested in are:

- lender to teams, has a total of 341,973 edges
- lender to loans, has a total of 2,495,435 edges

The data set used in this paper is similar to those reviewed in the related work section. General statistics of the datasets are compared in Figure 1. In general, 65% of registered lenders have funded at least one loan via Kiva. Following the best case scenario from those lenders, 17% have joined a team. Teams are very important in the Kiva ecosystem. Previous researchers have found that 50% of all activity is produced by teams[3] .What these numbers show, is that about 8% of registered lenders support half of all transactions happening within the Kiva ecosystem.

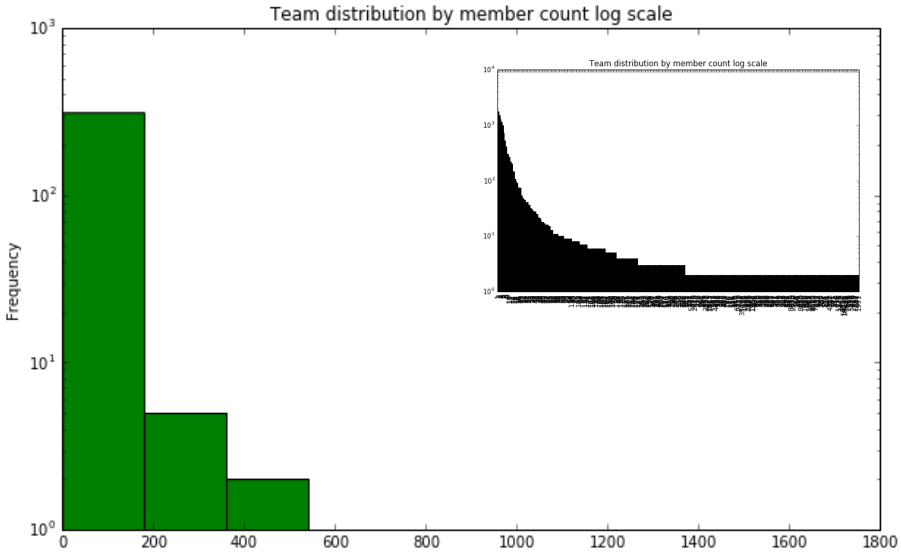
Figure 1 Dataset comparison

	Our dataset	Reference
Lenders that have given at least 1 loan	65%	64-66%[1]
Have no team	81%	85%
Have joined only one team	18%	12-17%
Overall activity from teams	33%	50%[2]
Lenders with more than 2 teams	4%	8%
Lenders in teams that have given more than 1 loan	69%	

I know that promoting team creation improves activity, leading to more funding with more frequency. Kiva would benefit from recommending teams to users that have never joined a team, or by matching lenders to promote team formation. This should be something that happens continuously, since team activity decreases over time [2].

The following section describes Kiva's data. At the initial phase of experimentation and review of related work, I was focused on analyzing the relationship between the reasons to loan (as stated by the lenders) and the objective of each separate team. To investigate this, I would only concentrate

Figure2 Team distribution by member count

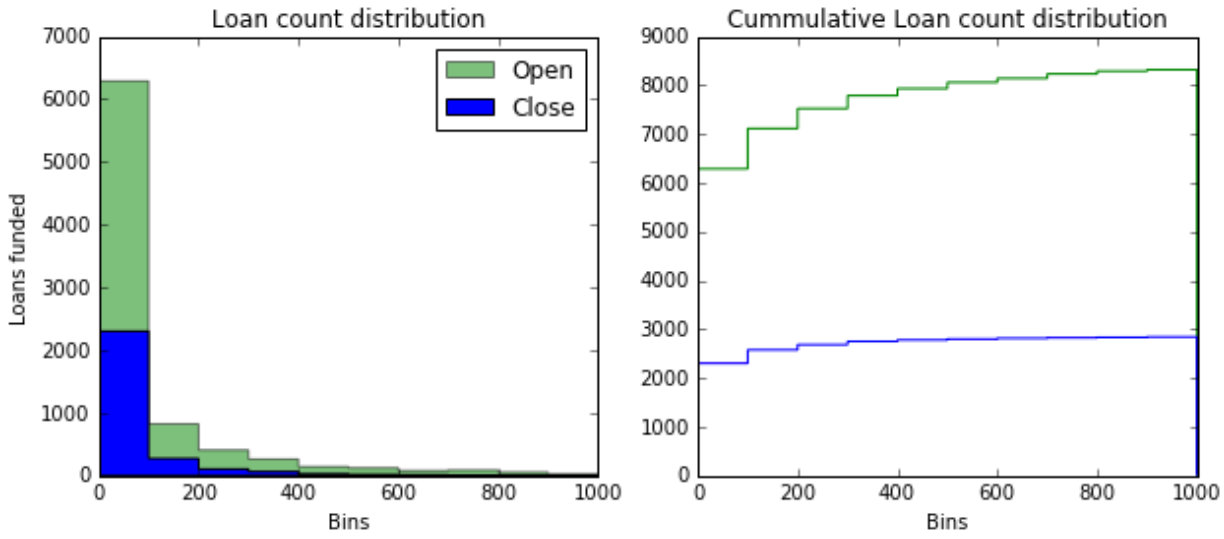


my research on lenders and teams that have stated their reason to loan. Taking that constrain into consideration, the team data gathered from the Kiva API corresponds with the data of teams created within the same time space as the lenders in the dataset. There are 11,835 teams within that space, making up a total of 99.5% of all team data. On average, a team has 32 members, with a standard deviation of 627 and a median of 4 members. The distribution is highly skewed to the right; 73% of teams have less than 10 members. Figure 2 shows team logarithmic distribution by member count.

There are two types of teams: those that are open for anyone to join, and those that require prospective members to be approved by the administrators. Each type is thus identified as either open or closed. Of all teams in the dataset, 74% are open, and 26% of the closed type. Which type of team membership contributes more to Kiva? Closed teams in average fund 223 loans, while open teams invest in 906, with a deviation of 4,090 and 13,509 loans respectively. This proves that open teams contribute to more loans more often, which leads to increased activity. Liu, Y. et al. [4] showed the same results for data from December 2010. The next question would be revolving around the terms of the lent amount. Which type of membership gives more per loan? Since the data does not reveal how much each lender gives with each loan, most Kiva-related papers make the assumption that each lender provided an equal amount to each loan. In the dataset, the open membership type averages a lent amount of $32.12 dollars per user. That is the same amount lent as for teams of the closed membership type. In terms of the amount lent, there is no visible difference in the amount derived by membership type. This leads us to the same findings as other related work.

There is a significant difference in the activity each type contributes to the ecosystem. Figure 3 shows the distribution of loans funded by each membership type. The distribution is similar, but the right tail of the open membership is longer, meaning more loans get funded by this type of teams. Kiva should promote the creation of open teams.

Figure 2 Loan count distribution by team membership



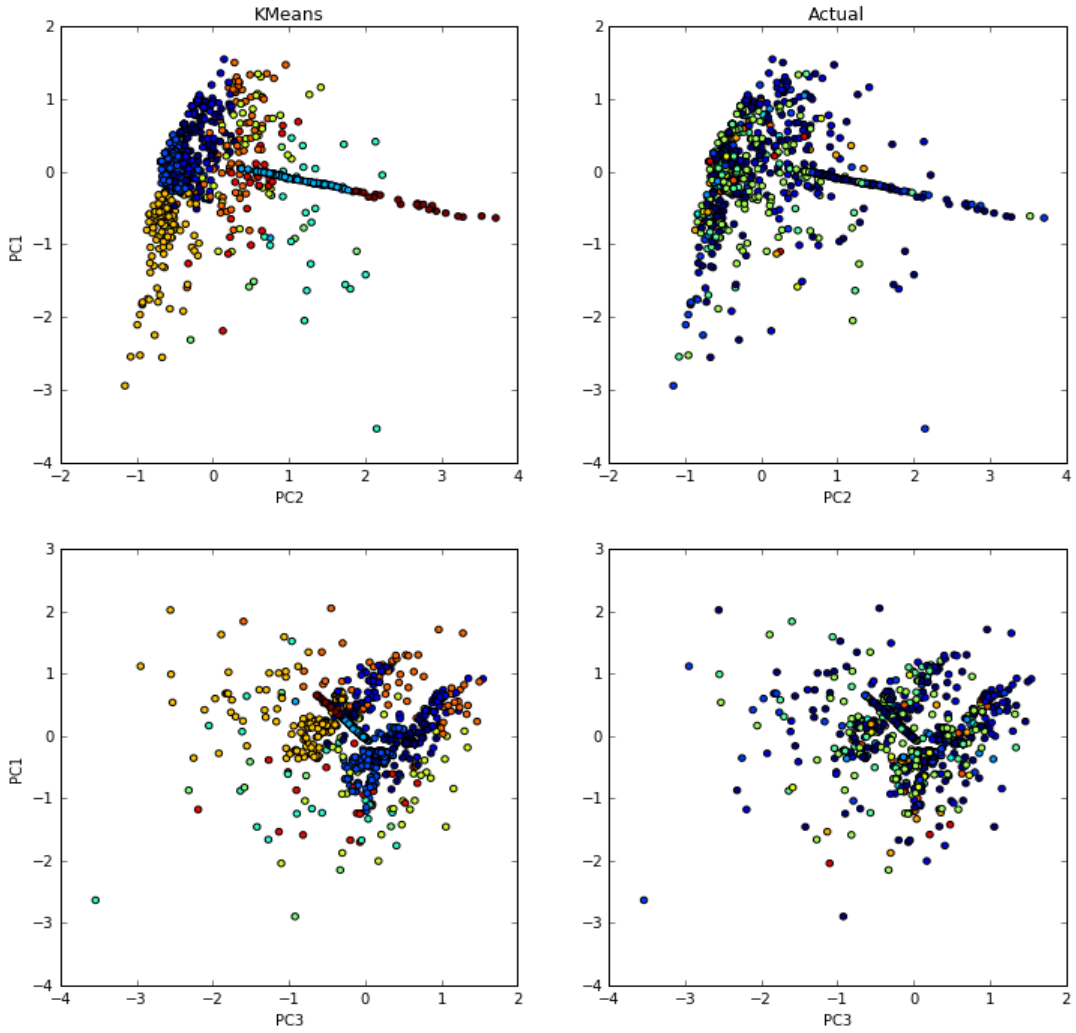
Experiments

Some additional experiments performed on the Kiva data are shown in the following section. It includes clustering, dimensionality reduction and filtering.

To support the main goal of producing more activity in the Kiva ecosystem by forming teams, we need to find a source of reasons from which teams my be created, hereby identifying similar lenders. One direct way of doing that is to investigate the loans and try to determine clusters of loans from which we may create a reason to lend. Imagine that we identify clusters in which a certain sector is relevant, or perhaps a combination of attributes such as country and sector. We could theoretically create clusters, and identify which reason to loan they would satisfy. From here, we would have to identify which teams or lenders match with these clusters in order to make a recommendation.

To cluster the loans, I have implemented a Kmeans algorithm enhanced with principal component analysis. Clusters are created using the loan purpose that was given by the borrower. This is an open text attribute where the borrower defines the use of the money. As I was preprocessing, I have

Figure 4 Loan Clusters

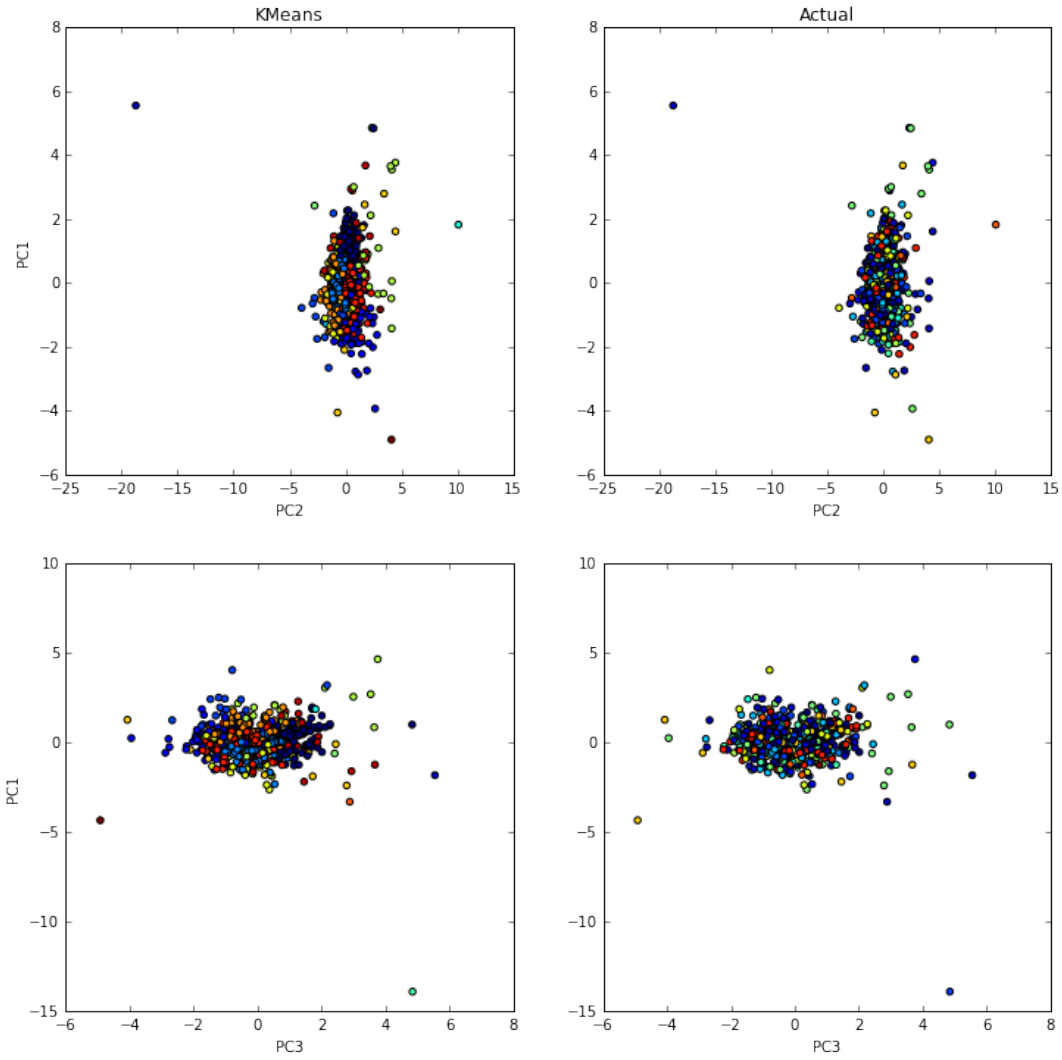


created vectors of words representing each loan. Since we know that each loan is categorized into sectors, I set the number of cluster to 12, hoping to see a relation. However, the results show good cluster, but no relation to the sector. As we can see in figure 4 below, the clusters created using the algorithm are not at all the same as the actual clusters created by the actual sector.

The figure shows clusters of loans created by the algorithm in the first column. The second column shows the loans colored by sector. In the top left graph we can see that the two main principal components cluster the loans in a good way. This effect is not aligned to the sector of each loan.

With the same idea of identifying clusters, I reproduced the experiment over teams. The attributed reason to loan is stated by the teams, which is also an open text field. Unfortunately the clusters created for teams are not as good. Teams may be seen as filters, if there is one for fishing, having two or more is unlikely. Arranging teams together in clusters is a harder task. In addition, many teams have very specific goals, such as “We want to get bicycles to people who have no method of transportation other than walking.” Others may have objectives that are focused on helping fishermen, and so on. Figure 5 shows the teams clusters.

Figure 5 Team Clusters



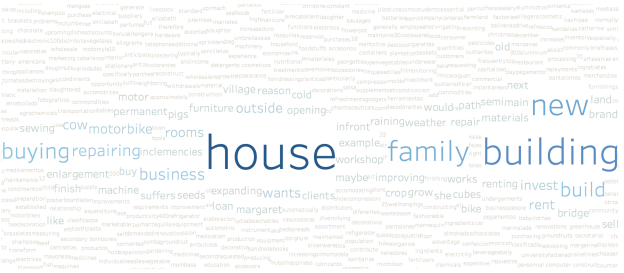
The following figures show some of the centroids of the clusters that showed a clear objective. As mentioned before, the languages in the dataset vary, with English being the most common language, followed by Spanish. Out of the twelve clusters, 7 of them are English and 5 are in Spanish. The complete set of images can be seen in the appendix.

Figure 6 shows Cluster number 9, with the most relevant words for the cluster being clothing related items. It has the type of cloth for women, men and children the type of article, such as pants, shirts, shoes, blouses, and so on. In addition, the cluster has some food related words like lard, rice, milk and other ingredients. This cluster would suggest creating teams and finding lenders that are interested in providing clothing and increase food inventory for business owners.

Figure 6 Cluster 9

Another interesting cluster is number 12, which refers to housing. The most relevant word is “house”, and we may also see building, build, buying, repairing and rooms. It also includes some business related terms for real estate. From this cluster we can identify users that may want to fund infrastructure, housing and real estate projects.

Figure 7 Cluster 12



One of the Spanish related clusters is shown in Figure 8. This cluster is all about food, and since Kiva finances projects, we can think that these clusters aim to building inventories, creating a convenience store or even setting up a restaurant.

Figure 8 Cluster 5

This approach could yield good results if we are able to identify a reason to fund within teams or lenders that corresponds to these clusters. For future work, it would be helpful to optimize this approach in order to create more possible purposes and identify teams and lenders that have funded similar loans in the past.

Reason to loan similarities

Since both lenders and teams provide open statements about their own motivation to lend, based on related work previously mentioned, this suggests that teams increase activity in the ecosystem. I decided to review similarities between the two, as to be able to enhance or create recommendations. If it is easy for a lender to explore teams, and is reasonable to think that they would join the one that is most similar to them, we should expect to see a natural match between lenders in a team.

All participants are encouraged to post everything in English, so that exposure to lenders is greater and becomes more effective. However, not every team and every lender states their purpose in a single language. The data set has most of the information in English, but it also has Spanish, French, and Portuguese, amongst other languages. There may be a problem translating each text to English in case expressions are being used that can not be translated directly. For that reason I decided to run the analysis using original languages. To measure the similarities I implemented *tf.idf* approach.

The approach to follow is the same as used search engines use. In first place a dictionary is built, based on the collection of documents to be indexed. When a new document is added, the dictionary is updated. When a query is executed, the dictionary is used to determine which document is most similar to it. I have a fixed number of teams that provide input to build the dictionary from, which is based on the team's goals. I do not need to update the dictionary, since there are no new teams and I am matching lenders to teams based on their reason to loan. Finally, the query would put forward the reason for each lender to lend. To measure the similarities, I have used cosine similarity weighted by *tf.idf.*

Since the overall goal of this paper is to provide a team recommendation, a subset is created selecting lenders that meet the following constraints: a) having provided a reason to loan and b) are members of more than one team. When applying these criteria, a total of 7,844 lenders are selected.

The similarity of each lender was measured against each of the 11,835 teams, recording a vector with the 100 most similar teams for each user. If we predict only one team for a given lender, only 1.20% of the recommendations based on the reason to loan similarities are correct, with an average cosine similarity of 0.56 (the average maximum similarity is 0.74, that is if the most similar team was the one joined by the user).

Another method to evaluate this approach is to produce as many recommendations as possible from the list of the 100 most similar teams, by setting a similarity threshold. When doing this, overall accuracy can be improved significantly to 6.6%, but this still forms a low portion of the total number of lenders. Figure 9 shows the accuracy at different thresholds.

This means that lenders do not select their teams based primarily on their reason to loan. There is no easy way for a lender to find his most similar team based on his or her reason to loan via the Kiva website[1]. The only recommendations that are given are based on what other people search. In

[1] https://www.kiva.org/teams/my-teams

Figure 9 Accuracy by cosine similarity

Cosine Similarity Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Accuracy	6.59%	6.56%	6.43%	5.33%	3.68%	2.49%	2.56%	0.86%	0.46%	0.05%
Avg. Cosine Similarity	0.7174	0.7177	0.7261	0.7591	0.8154	0.8635	0.9209	0.9606	0.9887	1

addition, the first teams shown are the ones that have lend more in all Kiva's history. Liu, Y. et al. [4] found that intra-team similarity, measured as the cosine similarity among team members, is high: 1,000 out of 1,185 teams they evaluated. However, the team data analyzed in this paper is 6 years older and ten times the size of that what was reviewed by the authors. Either selecting a team is coherent among similar users but has no significant relation to team goal, or lender behavior shifts in time. Choo, J. et al. [3] showed that teams evolve over time. Their evidence is the constant change on the leaderboard, where Kiva shows the top 10 teams based on the amount they lend and their new members.

Content-based team recommendations

The proposed model is explained in the following section. It includes a description of data preprocessing, model construction and evaluation.

The similarity analysis in the previous section demonstrated the need to gather additional information, as to be able to build a better model. In order to attain this, I followed a content-based approach. For each lender, a vector with data describing its lending activity was created and used to predict which team a given lender would join.

The data extracted for the final model includes the following data: number of loans funded, number of invitations send to fund a loan, geo-region of the user, occupation, registration date to Kiva, greatest, smallest and average loan amount, most frequent activity, sector, geo-region and field partner funded, and finally two vectors with the codes and frequencies of activity, sector, continent and partner of funded loans. These final vectors have an equal size for each lender, since all codes of each attribute are included in the recording of the frequency in which the lender has contributed to such attribute.

Registered lenders at Kiva have no interpersonal relations, and teams may be formed regardless of any preset conditions. For that reason I would expect that the selection of a team is independent among lenders. If this assumption is right, I may use Naive Bayes as baseline model with Gaussian distribution.

The attributes in the baseline are those that are related directly to each lender. The model grew because of potential improvements identified at each iteration and revision.

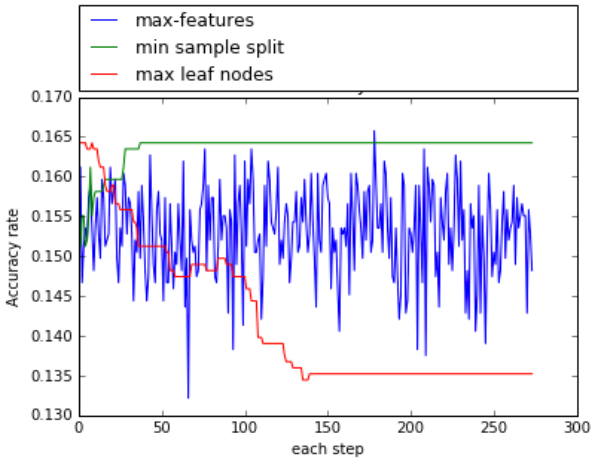
Figure 10 Recommender model results

Model Figues in %	Model Prediction Accuracy		Adjusted Predition Accuracy		Model Prediction Precision		Adjusted Predition Precision		Model Prediction Recall		Adjusted Predition Recall	
	Training	Testing	Training	Testing	Trainig	Testing	Trainig	Testing	Trainig	Testing	Trainig	Testing
Baseline	12.17	0.69	12.99	1.22	16	3	28	7	12	1	13	1
Decision trees	18.61	16.73	24.27	22.15	6	5	24	18	19	17	25	22
Naïve Bayes	14.21	5.12	18.21	8.56	28	5	36	12	14	5	18	9
SVM	17.96	16.50	24.89	22.92	6	3	29	22	18	17	25	23
LDA	18.23	14.59	24.97	20.78	6	4	32	27	18	15	25	21
LDA all variables	33.97	8.63	37.52	11.31	55	4	59	14	34	9	37	9
Random Forest	99.20	11.99	99.10	16.5	99	5	99	13	99	12	99	17

To test, a single prediction was made for each lender. That recommendation was compared to the first team joined by the user. In addition, an adjusted prediction was recorded. If the prediction was present in the list of teams that a lender had joined, the prediction was correct. Figure 10 shows results for all five models for the first prediction and the adjusted prediction. The same data was used to train and test all models. The baseline model is overfitted. One initial problem was the distribution used for the model. So I implemented Naïve Bayes with a Bernoulli distribution, training results improved, but nevertheless the model was still overfitted. This situation happened for random forest and LDA. Decision trees provided the best results; Figure 11 shows the process tuning result.

The tuning process for the final decision tree model was done in two steps. In the first step I trained several trees while changing 1 parameter at a time. The parameters that were tested are maximum number of features to use at each split, minimum samples at each split (resulting instances for a leaf), and the maximum number of leaf nodes. The default value of each parameter was 4, 100 and 30 respectively. The accuracy of each tree was recorded. From the first step the parameter that produces the simples and more accurate tree was extracted, resulting in a maximum number of features of 4, maximum leaf nodes of 50 and the minimum samples at each split of 200 were selected. In the second step I followed a similar approach. Again training several trees, but this time the default value of the parameters was the one as identified in the first step. The best model is reached with 2 features at each split, with maximum leaf nodes of 20, and samples minimum of 100.

Figure 11 Tuning Decision Trees

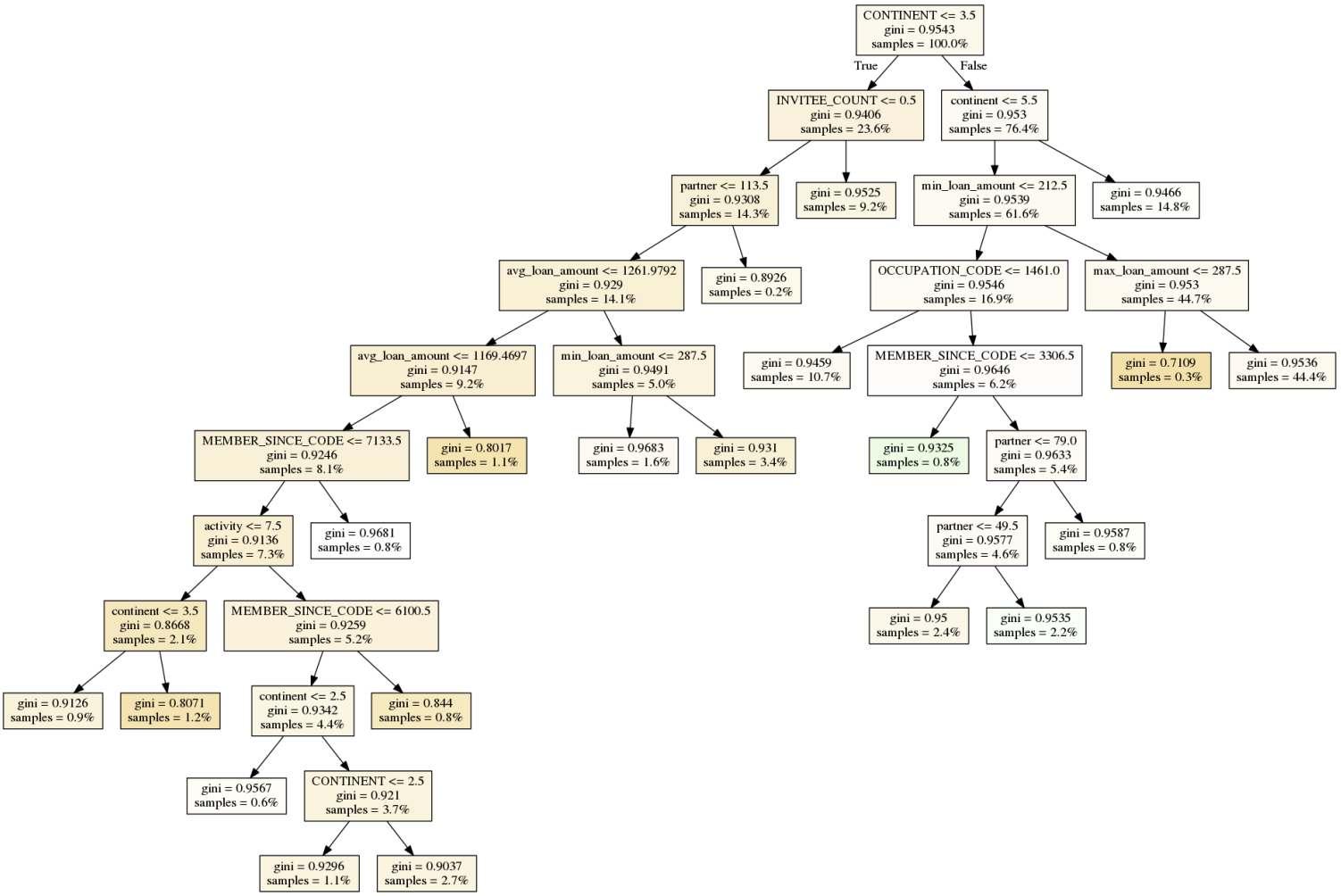


The final model tree is presented in Figure 12, producing 20 rules to produce a recommendation given the a lender or a team profile.

Conclusion and future work

I have shown that people who join Kiva as lenders and have since joined a team, do not select this team based on its purpose alone. A plethora of other factors are necessary for that to happen. It may be that people forget why they joined Kiva in the first place, or that their objective changes along the way and is not updated in their profiles. Using their profile as investor did improve the recommended model. For the final model, recall is good, so we may replicate what users do somehow. However, there is still much room for improvement. For future work, I would recommend to look for users that are actively joining different teams and try to find a pattern of why and how they decide to change teams.

Figure 12 Final model decision tree



References

[1] Shaaf, P., 2013. The effect of team membership on lending behavior on Kiva.org, Erasmus University Rotterdam, Master Thesis in Sociology.

[2] Chen, R., Chen, Y., Liu, Y., & Mei, Q. (2015). Does team competition increase pro-social lending? Evidence from online microfinance. Games and Economic Behavior.

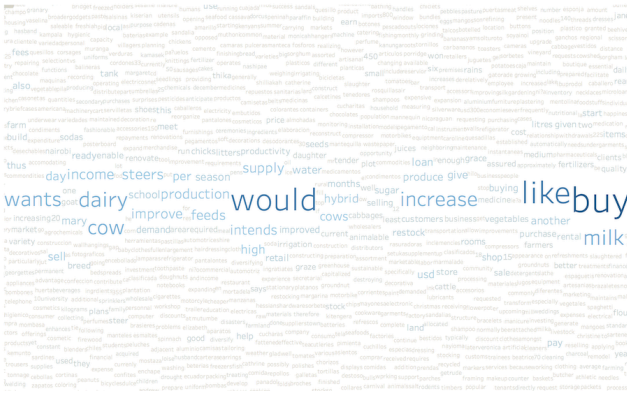
[3] Choo, J., Lee, D., Dilkina, B., Zha, H., & Park, H. (2014, April). To gather together for a better world: understanding and leveraging communities in micro-lending recommendation. In Proceedings of the 23rd international conference on World wide web (pp. 249-260). ACM.

[4] Liu, Y., Chen, R., Chen, Y., Mei, Q., & Salib, S. (2012, February). I loan because...: Understanding motivations for pro-social lending. In Proceedings of the fifth ACM international conference on Web search and data mining (pp. 503-512). ACM.

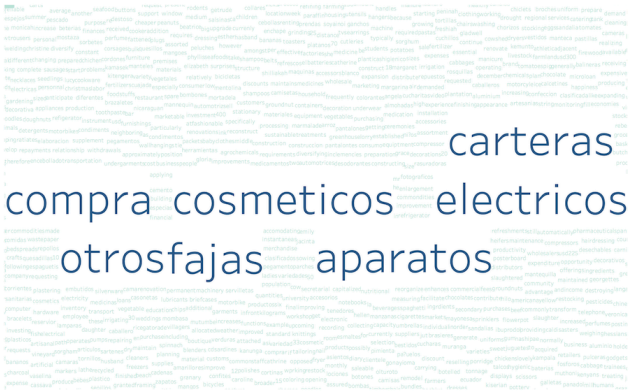
Appendix.

Experiments, reason to loan Clusters

Cluster 1



Cluster 3



Cluster 2

Cluster 4

Cluster 5

Cluster 8

Cluster 6

Cluster 9

Cluster 7

Cluster 10

Cluster 11

Cluster 12