Improving Underfunded Loans through Prediction and Recommendation

Group 9 He Dongying (A0142279B) Kwok Shi Ann Sheranne (A0218954U) Lee Kok Mun Delwynn (A0218929N) Mittal Sidharth (A0218815B) Poh Huizhen, Isabella (A0218865R) | **GitHub Link**

1. Executive Summary

Kiva is a non-profit crowdfunding loan platform to help communities in need. This project adopts a machine learning approach to improve loan funding success rate. Potential underfunded loans are first predicted prior to loans expiry. Various machine learning models are attempted and XGBoost is selected as the final model, providing a recall score of 0.82. Next, a content-based filtering recommendation system is developed to promote these underfunded loans to lenders. The recommended loans along with its corresponding brief description, generated using text summarisation techniques, are proposed to be shown to lenders on Kiva's landing page and in a personalised e-newsletter to improve the funding success of such loans.

2. Introduction

2.1 Introduction to Kiva

Globally, more than 1.7 billion people in the world are unbanked[1]. Without a bank account, these people are often unable to get access to financial services, such as financial loans to tide over difficult periods or start a new business. Founded in 2005, Kiva is a non-profit organization that aims to help meet the financial needs of these communities through their loan crowdfunding platform. Kiva currently supports borrowers in 77 different counties and works closely with a global network of field partners (such as local non-profit organizations). These field partners often provide the loan to the borrowers first (pre-disbursal) and the loan is then posted on Kiva for lenders to contribute. Generally, if a loan does not get fully funded, the field partner needs to come up with other sources of funding to cover the rest of the loan amount. Till date, over $1.54 billion loans have been funded through Kiva, helping over 3.8 million unbanked borrowers[2].

2.2 Loan matching process at Kiva

As seen in Figure 1 below, borrowers can apply for a loan and after going through an underwriting process, the loan is posted on Kiva. A lender coming on the website can then discover loans they are keen to support through searching for a particular region or cause (Refugees, Women, Conflict zone etc). The crowdfunding success rate on Kiva is relatively high, with 95% of loans being funded. However, the remaining 5% of loans is equivalent to more than 10,000 loans in 2019 and would increase even more as Kiva scales in future. Thus, the impact of improving funding success rate would increase in business value moving forward.

Figure 1: Kiva loan matching process

3. Project Objective

Given that all loans undergo a strict underwriting and approval process, we believe that some loans remain underfunded not because they are of inferior credit quality, but rather due to insufficient exposure to the most interested lenders. With more targeted publicity of the loan descriptions, we hope to further improve funding success rate. This project thus aims to increase exposure of underfunded loans on Kiva's landing page and create a personalised e-newsletter that reaches out to lenders with loan descriptions that could potentially be of their interests. To achieve this, the project can be broken down into 3 sections as follows:

3.1 Prediction of Underfunded Loans

By predicting the probability of the loan being underfunded before the loan expires, we can take on a proactive approach and find opportunities to push these loans to lenders such as displaying the loans on the site's landing page and curating a newsletter to be sent out to lenders via email.

[1] The World Bank's The Global Findex Database 2017. https://globalfindex.worldbank.org/sites/globalfindex/files/chapters/2017%20Findex%20full%20report_chapter2.pdf

[2] Kiva's about page. https://www.kiva.org/about

3.2 Recommendation System

To increase the conversion rate of email newsletter, we aim to do more than mass-targeting. Given that most mass targeted newsletters are ignored, it is of value to curate a newsletter that speaks to lenders to increase the effectiveness of a newsletter. Using content-filtering methods in recommender systems, the lenders are more likely to receive loan descriptions that align with their personal interests, thereby increasing the likelihood of lenders contributing to the loan.

3.3 Text Summarization

Taking it a step further, we summarise the loan descriptions on Kiva into short and concise snippets to insert into the newsletter using Natural Language Processing (NLP) methods. Brief descriptions are more likely to capture the attention of lenders.

4. Literature Review

There is a plethora of literature around online micro-finance platform. Hence, we refer to various research papers for our prediction, recommendation, and text summarisation models.

4.1 Prediction of Underfunded Loans

For the prediction of underfunded loans, we want to identify the different factors that influence lenders' choice to fund a particular loan on crowd-funding platforms, such as Kiva. It has been found that entrepreneur's narrative, i.e., description of a loan, makes a difference. (Allison et al., 2015). Hence, we generated the sentiment scores for the loan descriptions as an additional feature for our prediction model.

The function of hashtags has evolved over time, from information sorting and aggregating tool based on topics, to become an important part of message that serves communicative functions, which includes expressing attitudes, socializing, topic-marking and initiating movements (Laucuka, 2018). Inspired by this, we created new feature by counting the number of hashtags used in the loan profile and use it in prediction model.

4.2 Recommendation System

Recommendation systems can broadly be classified along dimensions such as input information and filtering techniques. Firstly, input information can exist in the form of explicit feedback such as user ratings for products or movies, or implicit feedback such as purchase history. Secondly, recommendation systems can also be classified into common filtering techniques such as content-based filtering which depends on the attributes of items to determine similar items and collaborative filtering which associates users with similar purchases or rating preferences (Isinkaye et al., 2015).

However, recommendation algorithms that merely produce a suggested list of similar items fail to recognise that the recommendation list would become too monotonous and thus reduce users' interest in these suggestions (Zhang and Hurley, 2008). In order to make the recommendations more diverse, we intend to use re-ranking methods based on maximal marginal relevance (MMR) to reduce intra-list similarity within the final recommendations (Carbonell and Goldstein, 1998), which may appeal to users who seek slightly differing loans for their next donations. The MMR approach can be adopted to introduce a loan to a recommendation list one at a time, with a "penalty" term for a loan that is too similar to loans already in the list. The general formula is reproduced below.

$$MMR = Arg \max_{D_i \in R \setminus S} [\lambda(Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j))]$$

Where Q is a query or user profile; D is a document; R is a ranked list of documents retrieved through recommendation; S is the subset of R already selected and Sim_1 and Sim_2 are similarity metrics; and λ is a weight parameter to adjust the extent of dissimilarity within the recommendation list.

4.3 Text Summarisation

With the increasing volume of data and information online, text summarization is an important and useful technique that can help users quickly consume and understand only information that is relevant to them.

Text summarization techniques can be broadly classified into 2 categories – abstractive and extractive. Extractive models produce summaries of the original text by choosing a subset of sentences in the original text, while abstractive models attempt to improve coherence and reduce redundancies of the passage. While abstractive models use linguistic concepts to produce summaries and can be more concise, it is more complex and requires labelled summaries to provide a deeper understanding on the contextual meaning. (Chandra Khatri et al., 2018).

For our project, we summarized the loan description as provided by borrowers. Considering the limitation of absence of labelled data for our summaries, we used extractive summarization. In extractive models, sentence scoring is the most commonly used method in text summarization. Each sentence in the passage is first given a measure or weight based on their representation of the entire passage. The summary is built by selecting a number of sentences based on the sentence scores.

5. Dataset

5.1 Overview and Data Description

The key dataset comes from the "Data Snapshots" section[3] of Kiva's website, which contains lender profiles, loan profiles and a loan-lenders file for mapping. The files are updated on an ad-hoc basis, which includes all the loans and lenders information.

KIVA_Loans	Data type	KIVA_Lenders	Data type
BORROWER_PICTURED	boolean	OCCUPATION	categorical
COUNTRY_CODE	categorical	COUNTRY_CODE	categorical
ACTIVITY_NAME	categorical	MEMBER_SINCE	date
SECTOR_NAME	categorical	PERSONAL_URL	link
FUNDED_AMOUNT	continuous	LOAN_BECAUSE	text
LOAN_AMOUNT	continuous	Name	text
DESCRIPTION	text		
VIDEO_ID	text		
POSTED/PLANNED_EXPI			
RATION/RAISED_TIME	date		

Table 1: Data and its types in data snapshots

5.2 Web Scrapping

As mentioned in Section 1.1, Kiva engages field partners for loan underwriting and approval. Each loan story page contains information about the field partners such as how long they have been with Kiva and their past delinquency rate, which can influence a lender's contribution decision.

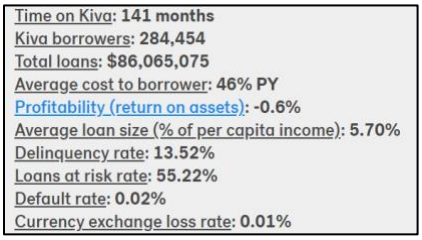

Figure 2: Field partner's information

There are also various interest groups on Kiva and lenders can choose to join a group which values resonates with them. The interest groups a lender join will thus contain useful information on the areas they are keen on.

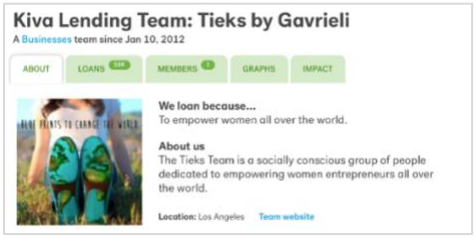

Figure 3: Sample interest group

Given that the existing available dataset does not contain the information, we supplemented the dataset by scraping the various sections on the Kiva website. Python package BeautifulSoup4 was used to scrape the data under the relevant HTML tags in every loan page.

5.3 Data Connections

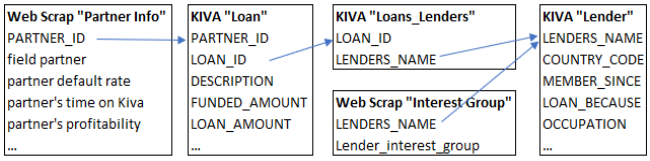

Figure 4: Data overview

Given the main dataset and further datasets collected, we will set up the dataset connection as shown above.

5.4 Exploratory Data Analysis

Data exploration can help us understand the features that might be important for classifying a loan as funded or underfunded.

5.4.1 Loan Amount

Loans that were underfunded are typically higher in loan amount (median value of $1,000). This is intuitive as loans with higher amount will need more lenders and will be more difficult to get fully funded.

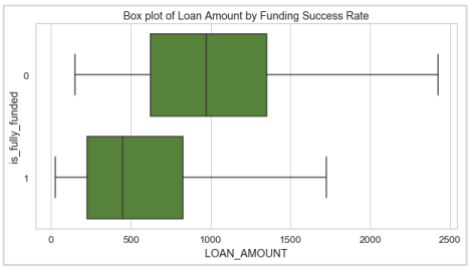

Figure 5: Loan amount by success rate

5.4.2 Sector-specific funding

Some sectors can have higher funding rates as people have different preferences and there can be a preferred sector for majority of the lenders. We find that Transportation and Wholesale sectors have the highest underfunded loans, indicating low interest from lenders in such sectors. On the other hand, Education loans are almost guaranteed to be fully funded with 99.6% funding success rate.

[3] Kiva Data Snapshot: https://www.kiva.org/build/data-snapshots

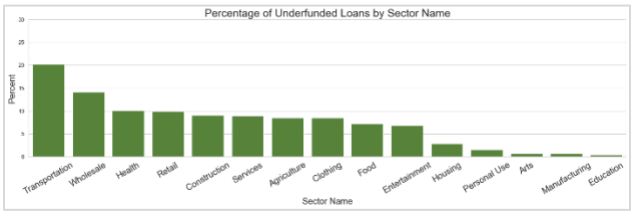

Figure 6: Percentage of underfunded loans by sector

5.4.3 GEOGRAPHIC-SPECIFIC FUNDING

We study both country and continent by grouping countries. For country, the percentage of underfunded is highest for Armenia (25%), and lowest for Egypt (0.2%). For continent, the percentage of underfunded is highest for Oceania (14.5%), and lowest for Asia (4%).

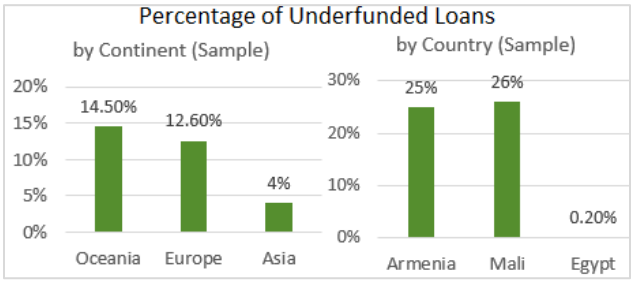

Figure 7: Percentage of underfunded loans by continent

5.4.4 TERM-SPECIFIC FUNDING

Loans with shorter-lender terms are more likely to be fully funded as lenders will prefer loans where they are likely to receive their money faster.

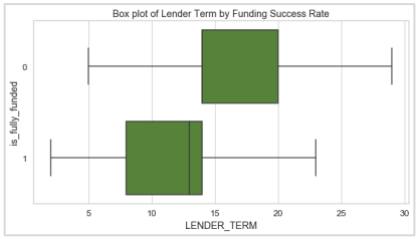

Figure 8: Lender term by funding success rate

5.4.5 GENDER-SPECIFIC FUNDING

Lenders generally prefer lending to loans where the borrowers' gender are known as the probability of a loan being underfunded more than doubles when the gender information is not available.

5.4.6 LOAN DESCRIPTION FUNDING

In addition, we found out that fully funded loans have at least 1 less sentence and 40 less characters as compared to underfunded loans at median.

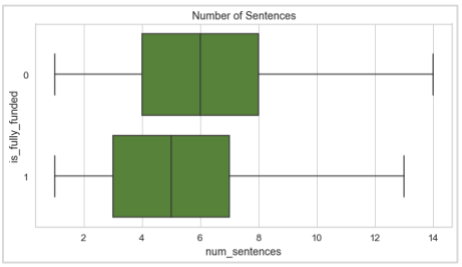

Figure 9: Number of sentences by funding success rate

Such results validate that a more concise loan description could help in improving the funding success rate as lenders generally prefer loan descriptions with lesser number of sentences and characters.

6. Hypothesis

Given our project objectives and available data, we formulate 3 hypotheses in this project.

6.1 Loans are underfunded due to certain features

The data exploration section revealed differences in funding success across countries and sectors. This could be due to a bias from lenders against or towards certain regions or activities. As such, performing prediction of loan funding status based on loan features can allow more proactive actions to be taken to boost the success rate.

6.2 Lenders exhibit preferences towards loan types

We hypothesise that lenders have personal preferences towards certain loan categories. For example, a lender who is a retired farmer may support more agricultural related loans. As a result, we will be able to use loan histories and loan attributes to match borrowers to lenders, which forms the basis for our recommendation system.

6.3 Description length influence attention of lenders

As the loan description is an informative part of the loan profile, we hypothesise that the length of loan description could have impact on the lenders' perception towards the loan. A succinct loan description could potentially capture the attention of lenders. As such, performing a text summarisation will bring out the most impactful points within a loan description.

7. Feature Engineering

Further engineering is carried out on the existing dataset to extract useful features which will improve the model ability to predict underfunded loans.

7.1 Sentiment of loan description

Sentiment score of loan description is calculated using NLTK sentiment intensity analyser after text pre-processing steps, which include lowering case, lemmatisation, removing borrower name, number, punctuation, meaningless text '
', stop words and white space. The score computed is added as an additional feature and it is found that loan descriptions with more negative sentiments (e.g. borrower appears more desperate) tend to have higher funding rate.

7.2 Length of description and number of tags

As seen in data exploration, descriptions that are more concise tend to have better funding rate. Furthermore, loans with more hashtags also tend to have higher funding rate. The length of description and number of hashtags used are thus calculated and added as additional features.

7.3 Grouping of countries

In the original loan dataset, borrowers come from 67 different countries. After encoding this categorical variable, we have a very sparse matrix which is not ideal for models such as neural network. To overcome this, the countries were mapped to the 6 continents instead.

7.4 Posting duration

The difference between the loan posting date the loan and expiry date is calculated as a new feature and the former 2 date columns are then dropped.

7.5 Funding ratio

We construct a ratio of funded amount to requested amount to describe whether a fund is fully funded, which is the dependent variable of our prediction. This is an important ratio given that we are focused on underfunded loans.

8. Data Pre-Processing

Several data pre-processing steps are then carried out:

- Categorical variables with no ordinal ranking such as gender and loan sector are encoded
- As features have varying magnitudes, normalisation using standard scaler is carried out. This is especially important for models such as Logistic Regression and K-Nearest Neighbors (KNN) which are very sensitive to varying scales in features.
- A small minority of loans are not backed by field partners and field partner features, such as rating, default rate, and total number of loans supported were thus imputed as 0.
- Features that are very similar to other features, such as loan use (similar to loan description) are dropped. Features such as borrower's town, currency and repayment interval are also dropped as they are found to have little impact on predictions based on data exploration.
- Lastly, various sampling methods, such as random under sampling and random over sampling, are applied to address class imbalance.

9. Methodology and Results

9.1 Funding Status Prediction

9.1.1 Evaluation metrics

The selection of evaluation metric shall reflect project objective in Section 2.1, which is to predict funding status of loans, in order to take proactive measure to boost its funding ratio. In view of the class imbalance issue, accuracy may not objectively reflect the model's ability in identifying the minority class – underfunded loans. It is important to have a high true positive rate (i.e. high recall for class label 1 – underfunded loan), where underfunded loan are correctly identified. In addition, we also try to minimise the false positive rate for class label 1 (i.e. high precision), so as to minimise the opportunity cost incurred from pushing fully funded loan to potential lender.

9.1.2 Modelling

There are a total of 218,064 loans in 2019 used in modelling, which are split into train and test sets in a 4:1 ratio. The splitting is done in a stratified manner to ensure that the proportion of underfunded loan in test set follows the proportion of underfunded loans in full dataset.

Two set of features are explored – one set with the country feature encoded, thus resulting in 105 features in total; and another set grouping country by continent to reduce matrix sparsity, thus resulting in 44 features in total.

To ensure consistency, the random state of all models is set to be 7. The following classification models are attempted: logistic regression, naïve bayes, KNN, decision trees, random forest, XGBoost and deep neural network. As underfunded loans only comprise 5% of the entire dataset, we use sampling techniques, such as over-sampling the underfunded loans or under-sampling the fully funded loans to mitigate the issue of imbalanced classes.

9.1.3 Result analysis

The best model selected is XGBoost which has high recall, high F1 macro and decent precision scores.

Model	Precision	Recall	F1
Logistic Regression	0.39	0.92	0.91
Naïve Bayes	0.11	0.94	0.58
KNN	0.63	0.60	0.95
Decision Tree	0.73	0.70	0.96
Random Forest	0.91	0.23	0.93
XGBoost	0.77	0.82	0.97
Neural Network	0.80	0.81	0.97

Table 2: Prediction model results

To ensure the model predictions are fair across genders and sectors, error analysis is conducted. For gender, the model achieves above 94% accuracy for both male and female borrowers. Similarly for sectors, the accuracy across all sectors is above 93%. Therefore, the model does not seem to indicate any biasedness by gender and sector.

To understand feature importance that affect underfunding prediction, different interpretation methods are used. The important features of XGBoost (identified via SHAP) are also cross-checked against the important features of other models (such as decision tree features importance and the coefficient of logistic regression). This is to ensure the prediction based on the best model picked is reasonable and interpretable. Figure 10 below shows the global importance: the increase in loan amount, male borrowers, and sentiment score can reduce the funding chances of a loan, whereas increase in journal entries and days to expire, loans for education purpose can increase the funding success rate. Number of journal entries refers to the times the borrower updates the loan, hence, the more frequent borrower update, the more likely loan gets funded.

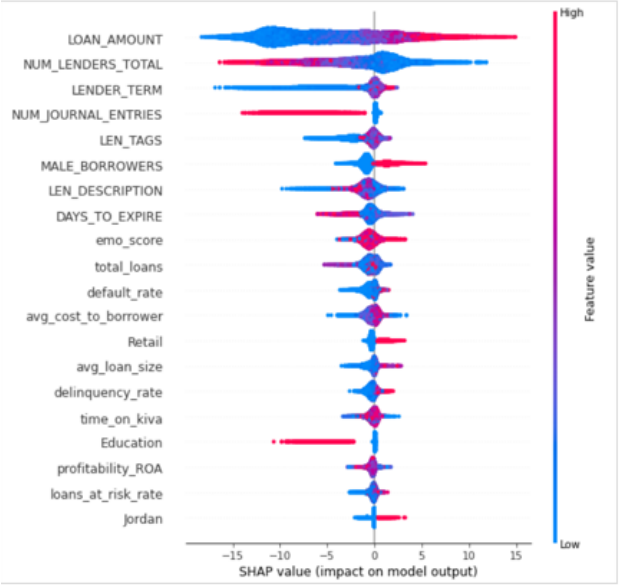

Figure 10: SHAP plot of important features

9.2 Recommendation System

The figure below summarises the approach to design, evaluation and addressing of issues for the recommendation system.

Figure 11: Recommendation system process

9.2.1 Design of Recommendation System

Following the prediction of underfunded loans, we deploy a recommendation system to suggest loans to potential lenders. A content filtering system is selected in lieu of collaborative filtering given that the utility matrix in the latter method will be mainly binary as majority of lenders (about 99%) donates only once per loan and is very sparse. On the other hand, the content filtering approach leverages on the rich loan attribute data.

We intend to incorporate available information about each loan to measure the cosine similarity between loans a user has previously contributed to (in year 2019) vis-à-vis underfunded loans (in period of Q1 2020). The information incorporated into the recommendation systems are listed below.

Gender	Country
Activity	Continent
Sector	Original language

Table 3: Attributes used in recommendation

An average of loan attributes is first computed based on the loan history by taking the mean of each attribute after encoding. For example, the user "barbara5610" has contributed to 4,990 loans in 2019. A vector of length 202 is then obtained with example of the first 5 attributes as shown below.

Attributes	"barbara5610"	Dataset average
Mixed gender loan	0.023046	0.019926
Male only loan	0.381964	0.191087
Female only loan	0.594990	0.788988
Agriculture	0.057515	0.029693
Animal Sales	0.007415	0.008321

Table 4: Example of 5 attributes for loans contributed by "barbara5610" against dataset average

Each attribute assumes a value between 0 and 1 as it is a mean of one-hot encoded dummy variables. Based on the above, "barbara5610" seems to have a stronger preference for male only loans (0.38 vs. avg. of 0.19) and agriculture (0.058 vs. avg. of 0.030) as compared to the dataset average. Comparing to the dataset average is important to consider distribution of the loan dataset and examine if a user's preference deviates from the average.

Thereafter, cosine similarities are computed between the average of the users' historical loans in 2019 and each loan in the underfunded category in Q1 2020. Principal component analysis is also deployed to reduce dimensionality from 202 to 32 to explain for 95% of the variance in order to improve the similarity measures and make the computations more efficient. The top 10 underfunded loans ranked by the cosine similarity score is proposed to each user. For "barbara5610", these were the recommended loans.

S/N	Gender	Activity/Sector	Locality
1	Female	Farming/Agriculture	Africa Kenya
2	Female	Farming/Agriculture	Africa Kenya
…	…	…	…
10	Female	Farming/Agriculture	Africa Kenya

Table 5: Recommended loans

9.2.2 INTRODUCING DIVERSITY

For the case of "barbara5610", an obvious problem was that every recommended loan belongs to the same combination of categories. Firstly, having such monotonous loans in a list of 10 is not very useful to the average user who may value some variety among the loan suggestions. Secondly, even though we speculate earlier that "barbara5610" has a preference for male loans, the fact that Kiva has a heavy proportion of borrowers that are female causes the recommendations to be largely centred around female loans. This will cause fairness issues if such a recommendation system is deployed.

Based on our earlier literature review, one potential method to introduce is the MMR method. This method requires a stepwise approach to recommending each additional loan as the similarity score has to computed taking into account loans already recommended. The following is the algorithm approach.

Algorithm MMR-based recommendation

Input: past loans for user x, underfunded loans y, number of recommendations m
Round #1
Calculate cosine similarities between x and every y_i
Sort similarities and add top ranking loan y_i to list S
Remove top ranking loan y_i from underfunded loans y
Round #2 to *m*
for $j = 2$ to m
Calculate cosine similarities sim_1 between x and every y_i in updated underfunded list y
Calculate cosine similarities sim_2 between every y_i and loans in already recommended list S
Sort similarities according to $(sim_1 - \lambda sim_2)$ and add top ranking loan to list S with weight λ
Remove top ranking loan y_i from underfunded loans y

Based on the algorithm above, "barbara5610" now has a more diverse set of recommendations.

S/N	Gender	Activity/Sector	Locality
1	Female	Farming/Agriculture	Africa Kenya
2	Female	Retail/Retail	Asia Jordan
3	Female	Beverages/Food	Africa Uganda
4	Female	General Store/Retail	Africa Ghana
5	Female	Shoe Sales/Retail	Africa Kenya
6	Mixed	Livestock/Agriculture	Africa Uganda
7	Female	Arts/Arts	America/US
8	Male	Food/Food	Asia Jordan
9	Female	Farming/Agriculture	Africa Kenya
10	Female	Farming/Agriculture	Africa Kenya

Table 6: Recommended loans with MMR

9.2.3 NLP AS FEATURES FOR CONTENT FILTERING

In addition to the 6 loan attributes described in 8.2.1**,** we also attempt to perform NLP on the loan description in order to construct more features for computing similarities. This premise on the assumption that similar loans will have similar words or terms used in the description.

We create a document term matrix based on term frequency-inverse document frequency (TF-IDF) with up to 2-grams and common English stop words removed. Thereafter, a similar recommendation system based on a cosine similarity is applied.

9.2.4 RESULTS AND ANALYSIS

To evaluate the accuracy of recommendation system, we adopt an offline approach as an online approach would require deploying the solution on the Kiva platform. Similar to evaluation metrics like precision@k, we attempt to determine if any underfunded loan recommended (out of 10 loans) to a user in Q1 2020 is indeed selected by the user, which we define as a "hit".

The evaluation requires users who have made some loans in the underfunded category in Q1 2020. Based on a cut-off of 100 loans in 2019 and more than 10 underfunded loans in Q1 2020, we select 14 users below for testing.

User	No. of underfunded loans in Q1 2020 contributed by user	No. of loans in 2019 contributed by user
kent3920	95	1,216
themissionbeltco	75	37,466
gooddogg1	72	41,091
trolltech4460	52	15,949
henry1547	31	5,379
barbara5610	31	5,005
rene3075	22	14,580
cliff5639	20	2,068
anonymous5138	19	21,331
anish7115	18	1,841
tristan4920	13	397
amirali5409	12	5,058
sharon047	12	338
am8748	12	5,055

Table 7: Summary statistics of test users

There are 319 underfunded loans in Q1 2020. From these 319, the number of hits per user out of a list of 10 recommendations are as follows.

User / No. of hits	Without MMR	Content-with MMR ($\lambda = 0.5$)	Without MMR and with NLP
kent3920	5	6	5
themissionbeltco	0	0	0
gooddogg1	0	0	0
trolltech4460	1	3	0
henry1547	0	1	0
barbara5610	5	4	6
rene3075	0	2	0
cliff5639	0	0	0
anonymous5138	3	2	1
anish7115	0	0	0
tristan4920	1	1	1
amirali5409	1	0	1
sharon047	1	1	0
am8748	2	3	0
Total	**19**	**23**	**14**

Table 8: Results of recommendation system

Content-filtering based on the MMR approach performs the best with the most hits among the 3 approaches. Using NLP as additional features did not perform well, likely because lenders do not choose loans based on specific words appearing in the loan description, but rather relying more on generic attributes such as activity, gender and country. Introducing diversity through MMR seem to improve results for some users, possibly because lenders do not invest in exactly the same type of loans and seek to have variety from time to time.

For the MMR approach, a weight of 0.5 for the penalty term for intra-list similarity is found to be the best performing. A search on a list of potential weights from 0 to 0.9 is performed with the corresponding total number of hits seen below.

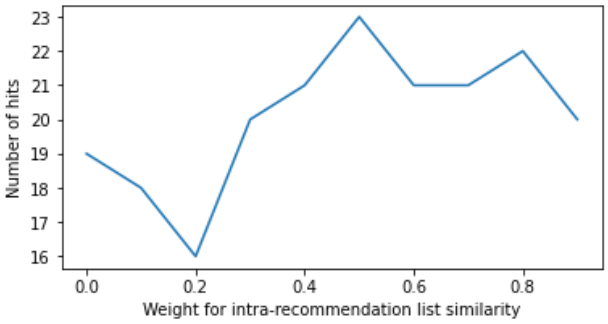

Figure 12: Performance of MMR over various weights

9.2.5 Results with predicted underfunded loans

As earlier mentioned, the prediction of underfunded loans will enable Kiva to pick up such loans earlier, so that these loans can be included in the pool to be recommended. The selected model, XGBoost is applied on the same Q1 2020 loan subset and 1,107 is predicted to be underfunded, as compared to 319 which eventually turn out to be underfunded. This is presented in below confusion matrix.

	Actual		
Predicted	Under-funded	Fully funded	Total
Underfunded	200	907	**1,107**
Fully funded	119	47,293	**47,412**
Total	**319**	**48,200**	

Table 9: Confusion matrix of XGBoost results

The content-filtering recommendation system is run again on these underfunded loans. However, the number of hits has decreased quite significantly from 23 to 18. This is likely because there is now a larger pool of underfunded loans to recommend from, and it is thus more difficult to obtain a hit within the same 10 loan recommendations as compared to earlier before. Furthermore, 119 loans are not predicted as underfunded despite being underfunded eventually. Nonetheless, as this is only an offline evaluation metric, we believed the overall recommendation will still be beneficial if assessed via A/B testing.

9.2.6 Recommendation by similar user

The recommendation system as elaborated thus far uses content-based filtering based on the lender's past loans. However, for new lenders, there may be a 'Cold Start' problem as they have very few lending histories. Hence, the loans recommended to them may not be well aligned with their interest. Hence for such users, we identify similar lenders with rich lending histories using content filtering on user profile. For this, we use the interest group and the lending reason information to find the similarity between frequent and infrequent lenders. The recommendation of the frequent lenders (identified in the previous section) will also be sent to the infrequent lenders as we hypothesise that lenders with similar user profile are likely to invest in similar loans.

Frequent lenders are defined as those who make more than 100 loans in 2019. The hypothesis is tested based on 1000 randomly selected infrequent lenders (less than 100 loans in 2019).

Firstly, identified similar frequent lender.

- Filter frequent lenders who have at least one interest group in common as the infrequent lender; and
- Tokenise the lending reason, remove stop words, compute the cosine similarity between the description of the infrequent lender and frequent lenders. Filter those with similarity higher than 0.2.

Secondly, count the number of loans which are invested by both the target infrequent lender and the similar frequent lender identified. Based on the 1000 random samples, 37%

of infrequent lender have at least one common loan with the frequent lender identified.

Infrequent Lender	Sample Similar Lender Identified	Common Loan ID
hassan76586349	tim4327	1659603, 1808367, 1864467
bettina4357	heg	1842288, 1819217
reg9953	shirley1905	1816366, 1830422, 1809381, 1769462
jeremy79228168	am8748	1779590, 1864910

Table 10: Sample output of 4 infrequent lenders

Following the recommendation from previous section, we get recommended loans for frequent lender with reasonable accuracy. Then, we identify similar frequent lender of the infrequent lender based on lender profile and share the recommended loan of the frequent lender to the infrequent lender, being the complement of previous section for lender with few lending records.

9.3 Text Summarisation by Sentence Scoring

After identifying the underfunded loans and potential lenders who may be interested in contributing to these loans, we deploy a text summariser, where the loan descriptions are summarised, and the resultant summary can be part of the newsletter sent to lenders who may be interested in the underfunded loans.

For our model, we adopt an extractive summarisation approach due to a lack of labelled dataset for developing an abstractive text summarizer. The extractive approach involves producing summaries of the original text by choosing a subset of sentences in the original text. We use a sentence scoring technique based on word frequency.

Before applying the text summarization, the loan descriptions are pre-processed in the following order:

1. Expand contractions such as “you’re”, “I’m”
2. Remove stop words such as “the”, “and”
3. Stem words using porter stemmer

The sanitised loan descriptions are then tokenised, and a table of words is generated by counting the frequency of words, also known as term-frequency table. Each word in the description is scored based on frequency and summed before being normalised by dividing by the number of words to give us a score for each sentence. We explored 2 methods of using this sentence score:

Method 1: Summarise by sentences with scores above average

This method generates the summary by selecting sentences with scores higher than the average of sentence scores in the passage.

Method 2: Summarise by sentences with top 3 highest scores

This method generates the summary by selecting the top 3 sentences with the highest scores.

The results of both methods are summarised in the figure below. We find that 17% of summarised loan descriptions returned extremely short description of <77 characters. This happens when loan descriptions are short to begin with and only 1 sentence is selected after applying the summarisation. Summarised loan descriptions that are too short may not be helpful as important information such as loan purpose or lender’s background could be omitted. As such, we decide to choose the summary generated by top 3 sentences for its ability to control for total length of description, which ensures sufficient information is retained in the summarised loan description.

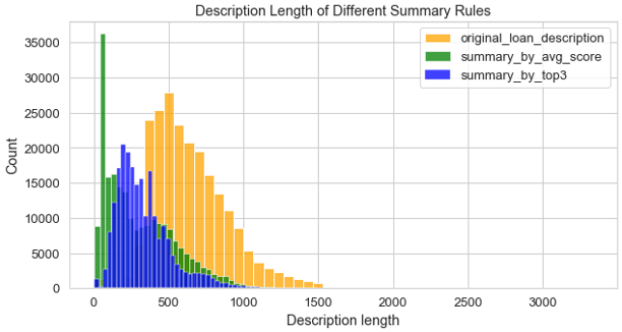

Figure 13: Description length of different summary rules

An illustration of original loan description:
Kennedy is a married man with five children. He describes himself to be honest and hardworking. (…) This is his first loan with SMEP Microfinance Bank. He will use the anticipated profits to buy more materials for making fish net. (150 words)
Loan description summarised by top 3 sentences:
He describes himself to be honest and hardworking. He operates a retail business where he sells fish net. He has been involved in this business for five years. (28 words)

10. Business Value

The recommendations of underfunded loans, coupled with summarised loan description can be used to improve loan funding success rate. This can be done via a two-pronged approached by targeting 2 groups of Kiva lenders:

Group 1: Lenders actively looking for loans to contribute

We propose for recommendations to be shown on the landing page of Kiva account upon lender’s log in. This solution facilitates real-time recommendations and allow lenders to quickly browse loans that they could be interested in.

Group 2: Lenders who have lent before but are not active

We propose to send out recommendations of summarized loan description in an e-newsletter with the aim of engaging lenders who do not log in to the platform recently. By actively reminding lenders of loans that can meet their lending interests, we can increase the frequency of loans while improving the funding rate.

A sample landing page (link) and newsletter (link) can be found in GitHub.

11. Limitations and Further Enhancements

Currently, the prediction model will not be able to predict the funding status of new loan categories (such as Covid-19). For such loans, we can skip the prediction part and directly include them in candidate loan list for the recommendation algorithm.

For the recommendation system, our current evaluation method is to test the results of recommendations on a holdout dataset (which contains users with a significant loan history). However, a better evaluation metric would be to carry out online experiments such as A/B testing, wherein our recommendation engine can be deployed for one group while a random recommendation list can be deployed for the second group. If statistically significant improvements can be observed for the former group, it could provide better assurance that the recommendation is working. Although, we have tried to address the cold start problem for new lenders by recommending them same loans which we had recommended to users which are similar to the new users. To calculate the user similarity, we used only few features due to limited user information. To further improve the user similarity process, we need to incentivise (such as complementing donations with small amount) users to provide more information about them so that we can better match them with the frequent lenders in the Kiva dataset.

Our current approach to text summarization is based on extractive algorithm; however, this is prone to information loss and absence of semantic order in the text summarization results. This can be further improved by using advanced methods, such as abstractive summarization (for which we would need labelled data) and use of Transformers and RNN-based models. Currently, our recommendation newsletters are in the English language; however, if we are able to get the information on lender's preferred language then we can further customize our newsletter to the lenders' preferred language, and hence allow Kiva to reach out to a wider group of lenders whose main language might not be English.

12. Conclusion

XGBoost (recall score 0.82) is used to predict loan funding status. Predicted underfunded loan are recommended to lenders using content-based filtering recommendation system with MMR (weight = 0.5) that generated diverse recommendations. Cold start problem is addressed by using content-based filtering to identify similar frequent lenders of infrequent lender. Same set of recommended loans are applied to both. Description of recommended underfunded loans are summarised using extractive summarisation technique by picking up the top 3 most important sentences. Summarised texts are sent to user to boost the exposure of underfunded loan and better capture lenders' attention. As such, the funding ratio of underfunded loans may likely be enhanced.

13. References

Carbonell J. and Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. SIGIR '98: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, 335-336, 1998.

Isinkaye F.O., Folajimi Y.O. and Ojokoh B.A. Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, 16, 261-273, 2015.

Zhang, M. and Hurley, N. Avoiding monotony: Improving the diversity of recommendation lists. *Proceedings of the 2008 ACM Conference on Recommender Systems – RecSys '08,* 2008.

Thomas H. Allison, Blakley C. Davis, Jeremy C. Short, Justin W. Webb. Crowdfunding in a Prosocial Microlending Environment: Examining the Role of Intrinsic versus Extrinsic Cues.

Laucuka A. Communicative Functions of Hashtags. *Economics and Culture, Sciendo, June 2018.*

Chandra Khatri, Gyanit Singh and Nish Parikh, Abstractive and Extractive Text Summarization using Document Context Vector and Recurrent Neural Networks

GitHub Link

https://github.com/sidharthmittal25/BT5153-Group-Project