National College of Ireland

A Fairness-based Recommender System for Charitable Lending Platform Kiva Using Classification and ϵ-Greedy Policy

MSc Research Project
Data Analytics

Krisztina Hapek
Student ID: x17126631

School of Computing
National College of Ireland

Supervisor: Dr. Majid Latifi

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Krisztina Hapek
Student ID:	x17126631
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Majid Latifi
Submission Due Date:	23/09/2021
Project Title:	A Fairness-based Recommender System for Charitable Lending Platform Kiva Using Classification and ϵ-Greedy Policy
Word Count:	7548
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	21st September 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Fairness-based Recommender System for Charitable Lending Platform Kiva Using Classification and ϵ-Greedy Policy

Krisztina Hapek
x17126631

Abstract

A large portion of the world's population does not have access to traditional financing through banks and financial institutions, thus they have less opportunities to improve their lives, businesses and to grow. The recent appearance and proliferation of peer-to-peer financing platforms has been offering a solution to this problem in the last over a decade. Among them, Kiva, a charitable online platform attracts well-meaning donors who want to help alleviate poverty. Previous studies into peer-to-peer lending, however, discovered that even charitable online lenders are driven by their preferences and biases, which leaves some groups of people still lacking financing opportunities. This paper proposes a fairness-aware loan recommendation system to diversify charitable loans among a larger group of people. The presented solution is based on the initial classification of loan applications using naïve Bayes and logistic regression classifiers followed by an implementation of the ϵ-greedy policy, which is a simple yet efficient strategy for the multi-armed bandit problem. The results show that in spite of the difficulty in correctly classifying loan applications, the ϵ-greedy strategy could be a viable option for the diversification of loan recommendations.

1 Introduction

In 2017, less than 70% of the world's adult population had access to a money account either at a financial institution or at a mobile money provider (*The Global Findex Database 2017*; 2017). The distribution of account ownership is not even among all countries and regions: developed countries have approximately 94% account ownership on average, while less developed regions hardly reach 63%. The lack of access to a money account means less opportunities for income, savings and borrowing.

In order to narrow this gap between different regions, several micro-financing institutions and peer-to-peer (P2P) lending platforms appeared in the past 15 years and loan portfolios covering low-income clients have been increasing uninterruptedly. The hundred largest institutions held 76% of the global micro-loan portfolio with 77.1 billion US dollars in 2017 (*Microfinance Barometer 2019*; 2019).

Especially P2P platforms have gained popularity in recent years. These platforms allow individual lenders to contribute funds to private individuals or small businesses that do not have access to traditional financing. These platforms operate either on a for-profit or charitable basis. The purpose of the first one is to earn interest on the

lent amount, while the goal of charitable platforms is to support small businesses or individuals in need.

The focus of this research is Kiva, a non-profit organisation that enables charitable P2P financing on its online platform. The goal of the organisation is to allow small businesses and private persons in poverty to receive small loans to improve their earning potentials or living conditions. Potential donors can browse and choose from thousands of application profiles. The profiles provide details of the borrowers including a picture, short description of their situation, what they request the loan for and repayment terms. Donors can lend a minimum of $25, thus a loan is usually financed by several donors.

While it is assumed that lenders are motivated by poverty alleviation, it has been noted that certain types of borrower profiles are more popular and receive more funding than others. This raises the question of how fairly these funds are distributed. Fairness is a complex and subjective term, therefore, it can be interpreted in multiple ways, one of which is – in relation to charitable lending - a more even distribution of funds among applicants. In order to narrow the financing gap between different causes, this paper proposes a fairness-based loan recommender system to potential donors. There have been prior proposals for recommender systems in this domain, but most of them focused on personal recommendations with added fairness considerations in some cases. This study emphasises the fairness concern based on the assumption that donors cannot objectively decide which applicant needs the funds more or where the support will have a bigger impact, therefore a more even distribution is a more desirable approach.

1.1 Research Question

RQ: "Using the available loan attributes, can a fairness-aware loan recommendation system be implemented using machine learning methods and randomised selections (Naive Bayes, Logistic Regression, ϵ-greedy algorithms)"

Research Objectives

- Objective 1: Understand the biases and previously developed recommender systems in the P2P charitable lending domain.
- Objective 2: Develop a recommendation framework through the combination of classification and randomisation techniques.
- Objective 3: Evaluation of results and feasibility of the proposed framework.

This paper is structured as follows: Chapter 2 provides an overview of **Related Work** in the P2P lending domain and previously proposed recommender systems; Chapter 3 describes the **Research Methodology**; Chapter 4 presents the details of the **Design Specification** followed by **Implementation** in Chapter 5; solution **Evaluation** is outlined in Chapter 6; Chapter 7 draws **Conclusions** and proposes further research opportunities in this area.

2 Related Work

The recent growth of online P2P lending platforms attracted significant interest from researchers. There is limited number of studies available relating to the Kiva platform, but there is a rich literature about for-profit lending platforms. In many emerging countries,

borrowers may not be able to receive loans from traditional financial institutions and are forced to avail of P2P lending opportunities. China is one example where the number of online lending platforms exploded and China's P2P lending industry is larger than that of the rest of the world combined, with outstanding loans of US$217.96 billion[1]. Therefore, most of the available literature concerns Chinese for-profit platforms. While their approach to lending decisions can be partially applied to demonstrate the existing biases in the P2P lending market, fairness concerns play little role unlike in charitable lending. The reviewed papers have been published in the past 12 years and demonstrate the evolution of the research from identifying biases to offering solutions.

2.1 P2P lending research and identifying biases

Initial studies in pro-social lending focus on the impact of such loans on poverty alleviation and on predicting if a loan application will be successful based on borrower characteristics. One of the first studies by Desai and Kharas (2009) compare the determinants of aid supply on two charitable online platforms, Kiva and GlobalGiving and conduct a survival analysis examining the funding speed of projects.

They investigate the aid distribution's effect on poverty alleviation considering macroeconomic factors along loan data, furthermore, they examine donor fragmentation based on the risk of loan default using a seemingly-unrelated regression model. The study concludes that lenders are less influenced by macroeconomic factors than by individual project profiles, and that loans perceived as higher risk take longer to fund than applications seen as low-risk. The research is limited to statistical methods, but explores interesting correlations using data sets related to external factors.

Ly and Mason (2012) partially contradicts these findings with statistical analysis and logistic regression. They examine the impact of borrower profile details on funding speed on the Kiva platform. They conclude that loan default risk does not play a role in lending decisions and find that loans to women get funded significantly faster than loans to men. The disadvantage of this study is that it does not consider that larger loan amounts may require longer funding time in spite of the application's relative popularity.

Cai et al. (2016) predict loan application success on a for-profit P2P platform based on the borrowers' transaction history using a logistic regression model. The authors argue that there is an information asymmetry since lenders get unverified and filtered information about borrowers. The research is based on the signalling theory that posits that observable entity attributes can serve as a signal of quality. Signal cost is an important element of this theory, suggesting that a signal would be costly if the associated monetary spending is high. They find that borrowers' transaction history is a good predictor of application success. Results show that loan duration does not significantly affect repayment while the loan amount is positively correlated with application success for first time borrowers. Both of these findings seem counter-intuitive and would be worth investigating further. It has to be noted that the analysis was conducted on a quite small sample size of less than 1% of the original dataset.

Paruthi et al. (2015) investigate how local field partner's risk rating impacts application success as P2P loans are distributed by local field partners. They calculate Pearson's correlation coefficient for ratings and lending activity and find a strong correlation. They observe that lenders linked to highly rated field partners contribute to more loans while show lower lending complexity than lenders linked to lower rated partners. This research

[1] `https://www.finextra.com/blogposting/17107/the-rise-and-fall-of-p2p-lending-in-china`

combines external country data with loan data and finds that Kiva appears to be successfully driving lending activity towards countries where job creation is harder to achieve. The limitation of this study is that it focuses on field partner ratings as the main indicator of lending activity not considering that their ratings may be linked to their geography and their time in business.

The geographical and cultural aspect of charitable lending is addressed by Burtch et al. (2014) who investigate lending actions between country-pairs. Their model considers physical and cultural distance and wealth disparity. They find that both cultural differences and physical distance have a significantly negative association with lending, consistent with expectations. Singh et al. (2018) research the same based on the "flat-world" hypothesis, a theory that globalisation will eventually equalise global economy. They use a regression model to predict bias in country–pair transactions and find that per capita GDP difference and migration between country-pairs, and the historical presence of a colonial relationship are all positively and significantly correlated with lending volumes. While the study presents a new approach towards the analysis of P2P lending attitudes, the findings are not new.

The impact of cultural distance on lending activity is confirmed by Park et al. (2019) who focus on the correlation between facial expressions in borrower profiles and the cultural preference of the lenders' countries. They conclude that lenders from nations that value excitement and other high-arousal positive states such as the United States, would lend more to borrowers who show excitement in their profile photos as lenders perceive them to be more trustworthy. They find that people use their culture's affective values to decide with whom to share resources, and lend less to borrowers whose emotional expressions do not match those values regardless of their race or sex. The limitations of this study are the small sample size of lenders; the fact that facial expressions were manually rated by independent raters, furthermore, most of the lenders on Kiva are from the United States, therefore it is difficult to separate decisions based on facial expressions since borrowers from low excitement cultures also get most likely funded by Americans.

Wang et al. (2019) use soft information from borrower profiles to predict application success and repayment default on a Chinese for-profit P2P platform. Soft information includes age, gender, demographic characteristics, posted picture, social network and textual description. The authors use a Tobit regression model and find that soft factors are important predictors of both application success and probability of default. This research uses a novel approach, but has somewhat limited use for charitable platforms.

Jenq et al. (2015) demonstrate that physical appearance and assumed borrower traits do influence charitable lenders. They use a small sample of manually coded borrower profile pictures and apply a regression model and find that lenders strongly favour more attractive, less overweight and lighter skinned borrowers. They also conclude that the assumed "creditworthiness" does not have a significant effect on charitable lending decisions unlike assumed "trustworthiness". In spite of the small sample size and the subjectivity of the manual coding, the findings raise an interesting point and confirm yet another possible source of lender bias.

These findings are partially contradicted by Moss et al. (2015) who apply the signalling theory to explain how information asymmetries influence decision making. Signals reflect unobservable characteristics and behavioural intentions that are not explicitly expressed in the borrower's statement and help lenders fill in information gaps. They apply text analysis on borrower profiles followed by a survival analysis to examine if a loan got funded, how quickly it was funded and if a loan was repaid. The study finds that virtuous

characteristics, such as trust are less appreciated and rewarded by Kiva lenders, furthermore, ventures signalling conscientiousness, courage, warmth, or zeal are less likely to repay their loans. In contrast, entrepreneurial characteristics were positively associated with loan funding, but did not have significant correlation with loan repayment. The limitations of this study are that it uses a specific dictionary for the text analysis; it focuses on word count and does not consider the context. Finally, the experience on the Kiva platform confirms that lenders pay little attention to repayment risk as majority of the repaid funds are never withdrawn[2].

2.2 Recommender systems in loan recommendations

The question of recommendations proves rather complex as there are various factors to consider. One approach is to create a recommender system that focuses on personalised loan recommendations for potential lenders, while another is to focus on the fairness of loan recommendations, however, the definition of fairness itself is complicated. "There are three classes of systems, distinguished by the fairness issues that arise relative to these groups: consumers (C-fairness), providers (P-fairness), and both (CP-fairness)" (Burke et al.; 2018).

Yan et al. (2018) use text mining methods to improve personalised loan project recommendations on Kiva by classifying and matching lender and borrower motivations. The study uses averaged word vectors and use them on pre-trained Twitter word vectors. Lending reasons are manually classified into lending categories. They use various classification methods to match the two sides and achieve mixed results. The possible reasons are explained as partially hand-coded data; furthermore, the data is significantly imbalanced in terms of observations for the different lending motivation categories. The authors conclude that the more active a lender is, the more difficult it is to predict his or her behaviour, which is counter-intuitive. The other weakness of the study is that the manual classification requires substantial resources.

Zhang et al. (2018) examine how borrowers can find lenders within a short time-frame. The authors propose a hybrid Random Walk (RW) approach with a sliding window solution for a recommender system. They apply various models: collaborative filtering (CF), which uses user-item relationships, content-based methods (CB), which build a preference profile for users based on previously selected items, and hybrid systems. RWH outperforms the other models on precision performs best at predicting funding percentage, however, the model run time is significant, therefore small datasets are used, which affects precision, therefore implementation may not be feasible.

Lee et al. (2014) propose a fairness-aware recommender system for the Kiva platform. Their model recommends certain loans to certain lenders in order to maximise lending. The study presents a novel method, Bayesian Personalised Ranking (BPR) optimisation together with matrix factorisation (MF) called Fairness-Aware BPRMF. The model is trained in Stochastic Gradient Descent (SGD) using tuples that include the lender, a positive example of a loan (chosen) and a negative one, which in this case is an otherwise popular loan, so that the recommendation can be trained toward less popular loans while trying to maintain recommendation accuracy from the lender's perspective. The performance is evaluated with AUC and the number of a loan's recommendation for

[2]Matt Flannery, Kiva founder: "They are just keeping the money in their [Kiva] account. Maybe they didn't know it was a loan." (Kiva: Improving People's Lives, One Small Loan at a Time. Knowledge@Wharton Podcast, May 28th 2008.)

fairness. They conclude that that the model achieved good fairness with minimal loss of accuracy. While this is a strong proposal, it cannot be applied on first-time or relatively new lenders if personalising is a goal.

Burke et al. (2020) present a theoretical work on algorithmic fairness that considers lender's personal preferences with added ranking based on the sofcial choice theory, meaning that it aims collective fairness. They assume that fairness concerns consist of groups over which recommendation results should be made fair, which may not conflict the lender's personal preferences. This study remains on a theoretical level and does not offer a practical solution. The social choice theory is implemented in another proposal by Sonboli et al. (2020) in their model called the SCRUF [3]. Instead of making the re-ranking decisions based solely on protected subgroups, they consider both the user and the protected group as actors with preferences. They observe fairness over time after multitudes of recommendations. They propose a system with multiple re-ranking functions where the choice of the re-ranking function is based on user history and either a fixed lottery (equal odds), a dynamic lottery (probability proportionate to the re-ranker's weight) or allocation lottery (randomised allocation with each choice eventually having the same probability). They tested the model on a movie and a loan dataset and managed to improve the fairness score slightly using the allocation lottery re-ranker while losing accuracy. While the approach is novel, it has two main limitations for lending recommendations: the large number of first-time lenders in real life and the fact that the model only used a subset of the loan attributes and completely omitted loans that did not received funding, that is the loans that were "treated most unfairly".

The evaluation of a recommender system is quite challenging due to its dynamic nature, but Beutel et al. (2019) propose a set of evaluation metrics to measure algorithmic fairness. They define fairness as equality of odds where the fairness of a classifier is quantified by comparing either its false positive rate and/or false negative rate. They propose a regularisation approach and randomised pairwise comparisons where the regulariser calculates the correlation between the residual between the chosen and not chosen items and the group membership of the chosen item. As a result, the model is penalised if its ability to predict which item was chosen is better for one group than the other. While this approach may work under certain circumstances, it may not suit the evaluation of recommendations in the pro-social lending domain for various reasons: each loan has several structured and unstructured attributes, which makes it difficult to assign them to one single group, furthermore, lenders may not be active enough to observe a significant pattern in their choices.

2.3 Summary of Related Work

In summary, the reviewed literature highlighted several biases in charitable lending, such as based on gender, purpose of loan, geography, facial expression, attractiveness, cultural differences, skin colour, weight among others, which defies the noble purpose of charitable giving. Several recommender systems have been proposed in this domain, but most of them focus on personalised recommendations. Fairness concerns have only recently emerged and mostly as an added feature to personal recommendations. A number of the proposed fairness-aware recommender systems may not provide an adequate solution due to substantial resource requirements or relying upon the lender transaction history. Table 1 summarises the proposed recommender systems. The present study attempts to

[3]Social Choice for Re-ranking Under Fairness

provide a solution to this problem by proposing randomised recommendations leaning toward disadvantaged applications.

Table 1: Summary of reviewed recommender systems in P2P lending

Authors	**Domain**	**Methods**	**Contributions**	**Limitations**
Yan et al. (2018)	Motivation-based recommender system	Text analysis, Gradient Boosting, Random Forest	Proposes a motivation-based recommender system, achieves improved recommendations for inactive lender groups.	Recommender system did not perform well on active, recurring lenders.
Zhang et al. (2018)	Recommender system to shorten funding speed	Hybrid Random Walk with temporal loans and lenders; KNN	Tests various models based on user history and item popularity.	- Needs significant resources - implementation may not be feasible
Lee et al. (2014)	Fairness-based, personalised recommender system	Bayesian Personalised Ranking (BPR) optimisation with matrix factorisation, called Fairness-Aware BPRMF	Achieves good fairness with small loss of accuracy.	- Changes in popularity over time not considered - Cold-start problem for first-time lenders.
Sonboli et al. (2020)	Fairness-aware recommender system	Social Choice for Re-ranking Under Fairness (SCRUF)	Personalised recommendations with multiple, fairness-based re-rankers.	- Model trained on a subset of loan features and completely omits unsuccessful applications - No solutions to the cold-start problem.
Beutel et al. (2019)	Evaluation metrics to measure algorithmic fairness	Multi-layer neural network	Regularisation approach and randomised pairwise comparisons	- Item groups based on one attribute only, may not work with Kiva loans - Assumes active lenders with identifiable lending patterns.

3 Methodology

The aim of this project was to propose a recommender system that introduces more fairness into charitable P2P lending by distributing funds more evenly on the Kiva platform. Prior recommender systems placed the emphasis on personalising recommendations with an added fairness element, however, Kiva has significantly less returning lenders than new ones, therefore personalisation may be difficult or less efficient as shown by related studies. Furthermore, each loan application has a finite funding time and finite target amount meaning that the loans cannot be recommended to a large number of users. This poses a unique situation in terms of recommender systems as recommendations are usually based on user transaction history or item popularity or similarity. This paper proposes a novel approach to loan recommendations based on the initial classification of applications by their popularity or funding status using **Naïve Bayes** and **Logistic Regression** classification models, followed by the adapted application of the **ϵ-greedy algorithm**, which proposes a solution to the Multi-Armed Bandit Problem. This study loosely followed the Cross Industry Standard Process for Data Mining framework without the deployment step. This chapter describes the research methodology whose steps are shown in figure 1.

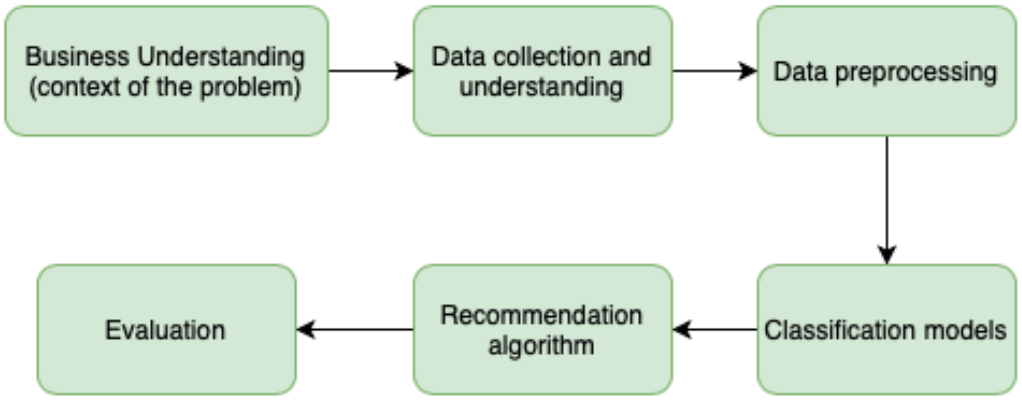


Figure 1: Research Methodology Steps

3.1 Business Understanding

P2P lending has been playing an increasing role in the financing of populations without access to traditional financial institutions. While credit risk is an important factor in for-profit P2P lending, charitable lenders are driven by the intention to help. In spite of this, there are borrowers that are disadvantaged even on charitable lending platforms. This paper proposes a model-based recommender system that puts forward potential borrowers whose applications have a low probability of getting funded based on historical transactions.

3.2 Data Understanding

The datasets used in this project were acquired from the Kiva platform through their API. The platform regularly prepares data snapshots, which are available for download

in csv format [4]. The snapshots include three separate datasets, one about loans, another one about lenders and the third one about loan-lender interactions. The present project was primarily built on the loans dataset.

After importing all three datasets, an exploratory data analysis was conducted. The initial inspection and descriptive statistics of the lenders dataset revealed that the number of lenders is 80% of the number of loans, furthermore, the mean number of loans per lender was 20, median was 3 and the mode was 1. These findings highlighted that most lenders only lend 1-3 times, therefore recommendations could not be based on user transaction history, instead it had to consider the cold-start problem. As discussed in chapter 2, Burke et al. (2020) and Lee et al. (2014) noted that their recommender systems did not work well with new users, which is the majority of the users on the Kiva platform.

The loan dataset included the details of nearly two million loans, each of which had 34 features describing various attributes of the loans and the applicants.

3.3 Data Cleaning and Preprocessing

The data exploration process included inspecting the features, data types and missing values. 15 columns were removed from the dataframe as they were deemed not to contribute to the potential model as they contained too specific information, such as names, video or picture IDs, or redundant information, such as duplicates.

Missing values in the various attributes were handled in different ways: where the proportion of missing values was low and the attribute was deemed significant, NA observations were removed or variables were imputed based on mean, mode or unique value if those values needed to be distinguishable. Attributes with large proportion of missing values were removed.

The original pandas dataframe stored features describing dates as object data type. These were cast to 'datetime' type for further computations. A new attribute was computed from the date variables to quantify funding period, and use it in regression models to predict funding speed. The same variable was used to create binned popularity classes based on the funding speed for a classification approach.

The descriptive statistics of the newly computed funding period column revealed that there were observations with negative or long funding periods, which were not in line with the platform's terms and conditions. Such outliers were removed. The fundraising period was further inspected by computing another feature for planned fundraising period. Grouped summary statistics were generated, which showed that applications distributed by local field partners had a shorter fundraising period than direct applications where the borrowers posted their applications themselves.

Previous research, such as Ly and Mason (2012) indicated that loans to women get funded significantly faster than loans to men, thus the feature containing gender information needed to be carefully transformed. The original dataset listed the gender of each applicant separately in an 'object' feature, which resulted in 23895 unique combinations of female and male, single and group applications. This was reduced to a smaller number of unique values using pandas' string methods: counting the number of occurrences of 'fe' and 'ma' strings in each record where 'fe' indicated the number of women and 'ma' indicates the total number of applicants; the number of males was calculated as a difference of these. The gender variable was eventually transformed into 'female', 'male' and 'mixed' categories for better classifications.

[4] https://www.kiva.org/build/data-snapshots

The categorical variables - sector, country, distribution method, gender and repayment terms - were one-hot encoded using pandas get_dummies method. Numerical features, loan amount and lender term, were normalised to a range between 0 and 1 using Scikit-learn's MinMaxScaler function to avoid that these features weigh more heavily in the classification model.

3.4 Models

This research project utilised two types of models sequentially to achieve fairness-based recommendations: the first one aimed to predict the success of loan applications through classification, while the second one made recommendations for potential lenders.

3.4.1 Classification

Two approaches were considered to infer application success: regression to predict funding period and classification either to predict funded status (funded or expired) or 'popularity', a five-class classification where the popularity classes were derived from the binned funding days variable. The regression model would have allowed a more accurate calibration of applications. Pearson's correlation was calculated and visualised using seaborn's heat map to verify correlations between numeric attributes. While female attribute showed strong negative correlation (r = -0.86) with funding speed indicating that female applicants receive the funds quickly, other features, including loan amount and lender term showed weak correlations with the length of funding (r = 0.15 and r = 0.23 respectively), therefore it was concluded that numeric features were not strong predictors of funding speed, thus regression was discarded and the classification approach was implemented.

Logistic regression and Naïve Bayes classifications were utilised for the prediction of loan success due to the accuracy, speed, simplicity and suitability of these algorithms for high-dimensional datasets. It has to be noted that the nature of this project - fairer and more even distribution of funds - allowed less accurate predictions as the second stage introduced a level of randomness into the recommendations.

Naïve Bayes algorithms are based on Bayes' theorem that describes the probability of an event based on prior knowledge of conditions that might be related to the event. "Given a hypothesis (H) and evidence (E), Bayes' theorem states that the relationship between the probability of the hypothesis before getting the evidence P(H) and the probability of the hypothesis after getting the evidence P(H|E) is" [5]

$$P(H|E) = \frac{P(E|H)}{P(E)}P(H) \tag{1}$$

The model is called naive because it assumes the independence among attributes; the presence of a particular feature in a class is unrelated to the presence of any other feature and all of the features independently and equally contribute to the probability of the class.

Two naïve Bayes approaches were tested: one based on Bernoulli distribution, which only accepted discrete values in binary form; and the other one based on multinomial distribution. "The multinomial distribution describes the probability of observing counts among a number of categories, and thus multinomial Naïve Bayes is most appropriate for

[5] https://brilliant.org/wiki/bayes-theorem/

features that represent counts or count rates" (VanderPlas; 2016). Scikit-learn's Bernoulli Naïve Bayes implementation takes the argument "binarize=True" that maps floats to booleans such that the two numeric features were also included in the classifier.

The naïve Bayes models were contrasted with logistic regression whose assumptions were also met by the dataset: the dependent variable was dichotomous (funded vs. expired); there were no outliers (continuous predictors were standardised) and there was low level of multicollinearity.

The models were trained and tested with various feature selections and different splits between the training and test sets to find the one with the highest accuracy of prediction. A high training split ratio may lead to overfitting, however, the Naïve Bayes algorithm is less prone to overfitting due to its simpler hypothesis function when compared to other algorithms. Overfitting would only occur if the class had too many values and a relatively small data set was trained. The dataset used in this research project did not suffer from these restrictions, thus Naïve Bayes was a suitable classifier. The models were tested with two different target variables: the previously mentioned two-class (funded status) and five-class (popularity) options.

The classification models were evaluated as described in subsection 3.5.

3.4.2 Recommendation System

In the past years, the rapid evolution and growth of large-scale online services with more and more items to choose from made recommender systems an important tool for information filtering. They help both the users by allowing quicker decisions and online platforms to sell more and better suited items. While all recommender systems have the same goal - to make relevant recommendations -, there are several approaches to achieve the best results. Isinkaye et al. (2015) summarise the main recommendation techniques in figure 2 .

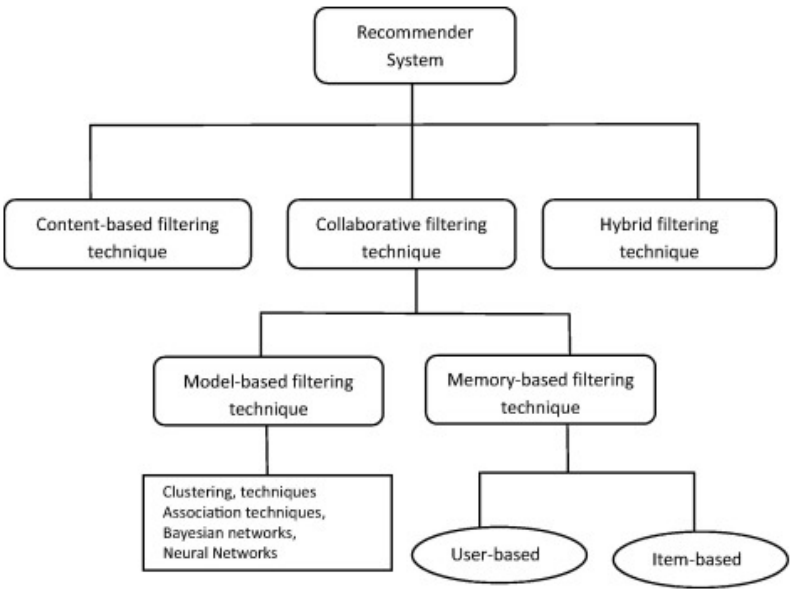


Figure 2: Recommendation techniques

Various recommendation techniques had been considered to answer the research question. In content-based filtering, the recommendation is made based on the user profiles

using features extracted from the content of the items the user has evaluated in the past. This technique was not suitable for this project as most lenders are either new to the Kiva platform or have only lent a few times before, furthermore, this technique would reinforce the user's previous behaviour.

In collaborative filtering, the model builds a user-item matrix of preferences for items by users. It then matches users with relevant interests and preferences by calculating similarities between their profiles to make similar recommendations within such user neighbourhoods. This technique was dismissed for similar reasons as the content-based filtering. In addition, loan applications have a limited lifespan on the platform: once the requested amount is reached or the funding period expires, loans are removed.

The above techniques do not offer an adequate solution for the cold-start problem, a scenario where a recommender system does not have sufficient information about a user or an item in order to make relevant predictions. This paper attempted to solve this challenge with the application of the **ϵ-greedy algorithm**, which is a strategy for the solution of the multi-armed bandit problem. Decisions are often made to maximise some expected numerical reward, but they can also help discover new knowledge that could be used to improve future decisions. The multi-armed bandit problem is an instance of this dilemma. A multi-armed bandit is similar to a traditional slot machine (one-armed bandit), but has more than one lever. "When pulled, each lever provides a reward drawn from a distribution associated to that specific lever. Initially, the gambler has no knowledge about the levers, but through repeated trials, he can focus on the most rewarding levers" (Vermorel and Mohri; 2005). The dilemma is how to optimise the overall reward by balancing known rewards and discovering new options that may bring in higher rewards. This process is referred to as exploitation versus exploration trade-off in reinforcement learning.

There are multiple strategies dealing with this problem. The ϵ-greedy policy follows a greedy arm selection policy and chooses the best-performing arm (highest reward) with 1-ϵ probability, but ϵ percent of the time it chooses a random arm (Auer et al.; 2002).

As Vermorel and Mohri (2005) indicates, ϵ-greedy in its simplest form cannot reach optimum results with a constant ϵ, that is a constant exploration percentage since the algorithm learns which are the most rewarding options, therefore their selection probability should increase. This project attempted to turn this handicap of the ϵ-greedy algorithm into an advantage in the achievement of the research goal. More fairness was attempted by using a constant ϵ exploration probability in order to avoid moving toward more reward. Moreover, reward, which, in other domains could be represented by positive feedback, was replaced by a dummy score attribute, which was derived from the predicted status of the application. Loans predicted as not-funded were generally assigned higher but random scores than loans predicted as successful. In order to avoid bias, the proportion of the exploration phase was similar to that of the exploitation phase. Figure 3 shows the components of the ϵ-greedy algorithm.

3.5 Evaluation

The performance of classification models was evaluated using accuracy, precision, recall and the F-measure. Accuracy measures the percentage of instances correctly predicted by the model. Precision expresses the fraction of relevant instances among the retrieved instances, while recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. These metrics are expressed in the following

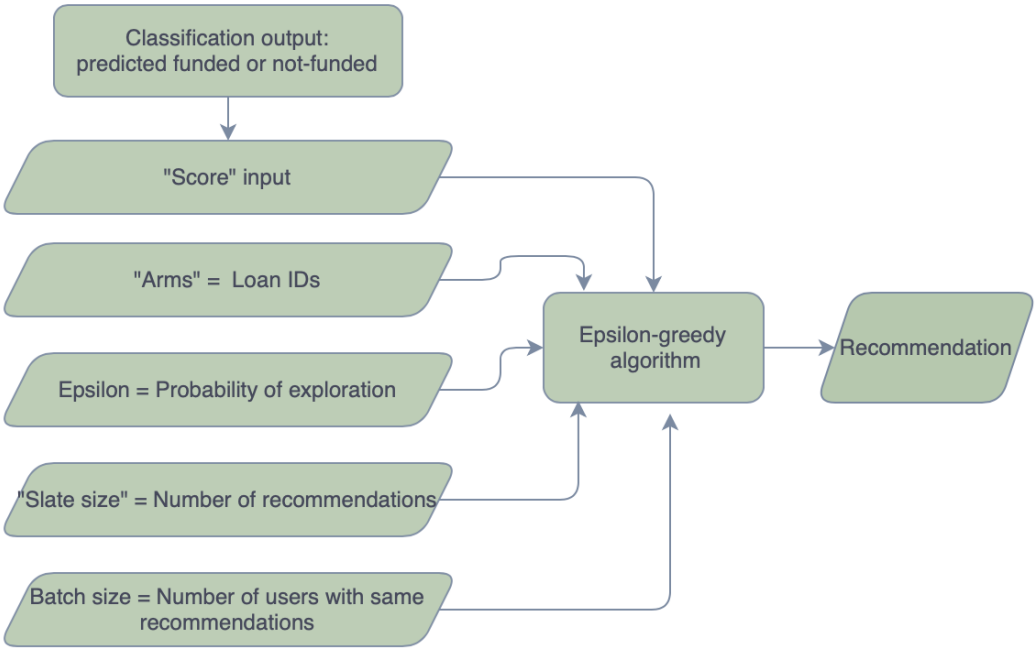


Figure 3: ϵ-greedy recommendation process

formulas:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{4}$$

F-score is the harmonic average of precision and recall. The closer the F-measure is to 1, the more accurate the prediction is.

$$F1score = 2 * \frac{Precision * Recall}{(Precision + Recall)} \tag{5}$$

The evaluation of recommender systems is a challenging task as their usefulness can only be observed once they are implemented. Typical evaluation metrics are accuracy, the fraction of correct recommendations out of all possible options, and coverage that measures the fraction of items that the recommendation system is able to provide recommendations for (Isinkaye et al.; 2015). Measuring accuracy requires knowledge of previous user-item interactions, which was not available for the present study. Coverage seemed a more appropriate approach to evaluation, however, the randomness included in the algorithm made this metric hard to quantify. A standard metric for multi-armed bandit algorithms is regret, which is the difference between the reward of the chosen arm and the reward of the best possible arm (LeDoux; 2020). Since the reward of the arm that was not chosen is unknown in this case, the most appropriate evaluation metric was cumulative reward. This was based on a dummy score, which was assigned to each loan based on their predicted funded status where the score of predicted not funded loans was distributed in a higher range than that of predicted funded loans. This means that

the model can be measured on the cumulative or mean reward achieved over several iterations.

4 Design Specification

This research project was implemented using Python as programming language in the Jupyter Notebook environment through Anaconda package manager. The data was downloaded from the developer page of the Kiva platform where it was presented in csv format. Both data preprocessing and the models - classification and recommendations - were performed in Python. Figure 4 shows the design flow chart.

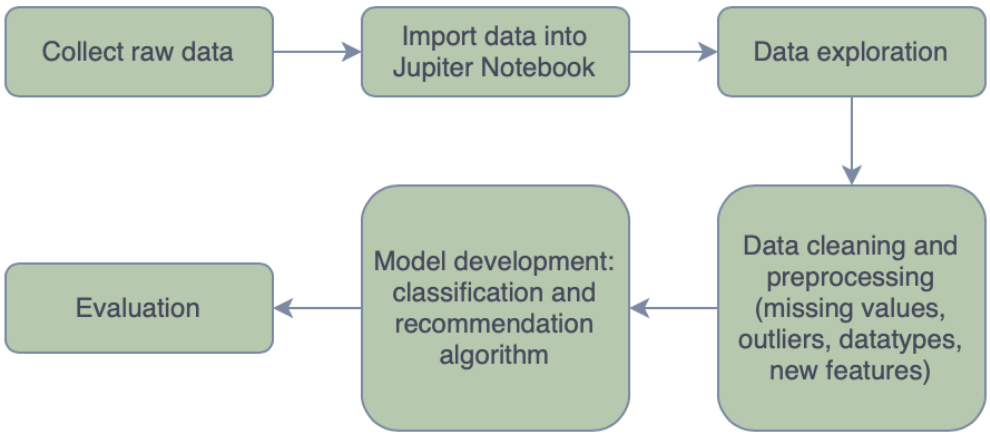


Figure 4: Design Flow Chart

The study made use of several Python libraries for the various tasks: Pandas and Numpy were used throughout the whole project for data preprocessing, data manipulation and analysis; Matplotlib and Seaborn were applied for visualisations; Scikit-learn's preprocessing, classification and metrics packages helped the model development and evaluation phases; finally, the Random package was utilised in the implementation of the recommender system.

The classification phase utilised standard Python packages, while the recommender system was based on an individual solution, which had not been mentioned in the related literature. The recommender system design included the following steps:

- Creating a new feature as a substitute for the "reward" feature in the ϵ-greedy algorithm. This new feature, "dummy_score", was derived from the predicted funding status of applications. A random integer was generated in way of "dummy_score" where the score of "expired" applications moved within a higher range than the score of "funded" applications.

- In the greedy strategy, arms of the multi-armed bandit have to be selected. The arms were represented by the "loan id" feature in this application.

- The algorithm took various other arguments. These were ϵ, representing the probability of the exploration phase where loans with lower dummy_score were selected,

while 1-ϵ indicates the probability of the exploitation phase when loans with a high dummy_score were selected. The slate_size argument referred to the number of recommendations to be made in each round and batch_size indicated the number of users the same recommendation could be made for.

- The algorithm first decided if it was exploration or exploitation phase by randomly selecting one of them from a binomial distribution.
- In the exploration phase, the model randomly chose the predefined number of loan applications from all applications using the Numpy random.choice function.
- In the exploitation phase, however, it sorted applications based on their dummy_score and chose the ones with the highest scores.
- In the final step, list recommendations was extracted.

5 Implementation

The main raw dataset used for this project had a shape of 1951124 x 34 and it required extensive preprocessing for the implementation of the classification models. 14 features were selected for the algorithms. The variables that were dropped included date attributes, original language, detailed description of the loan application, tags and the number of lenders to each loan, which could not be used as a predictor as it was not a prior attribute of the loan. The selected variables contained both numerical and categorical data types and included the following:

- Loan amount - float;
- Lender term - float, the length of loan repayment period;
- Sector - categorical, what field the loan will be used in, e.g. agriculture, retail, personal use and more with 15 unique categories;
- Country - categorical with 99 unique values;
- Repayment interval - categorical with three values;
- Distribution method - categorical with two categories: field partner and direct;
- Partner id - categorical variable including the individual field partners';
- Female - integer, indicates if the borrower or borrowers were female;
- Male - integer, indicates if the borrower or borrowers were male;
- Mixed - integer, indicates if the applicants were a mixed group;
- Group - integer, indicates if it is an individual or group application
- Funding days - float, the number of days between posting date and full funding date
- Status - object, shows if an application reached full funding or expired.

- Popularity - object, alternative target variable, which indicates how quickly an application got funded.

The classification algorithms were run using Scikit-learn and various settings were tested to find the best performing model.

Both naïve Bayes algorithms and the logistic regression model were implemented with different train - test splits from 65 - 35% to 80 - 20%. The applied algorithms handle high-dimensional datasets well, therefore all features that had low proportion of missing values and could potentially contribute to the classification were included initially.

The Bernoulli naïve Bayes model was implemented using the BernoulliNB method from Scikit-learn. It provides the argument "binarize=True", which transformed the loan amount and lender term features into binary form, while the other features were already one-hot encoded.

Multinomial naïve Bayes was implemented with Scikit-learn's MultinomialNB method.

Scikit-learn's LogisticRegression function was applied in the third classification model. For the binary classification, it was implemented with parameter "solver = liblinear" as it applies automatic parameter selection. For the multi-class classification, the "solver = saga" parameter was utilised, which is a variant of Stochastic Average Gradient Descent and it is recommended for multi-class predictions for large dataset.

All of the classifications were evaluated with accuracy, precision, recall, F1 score, specificity, AUC and confusion matrix with the help of sklearn.metrics package.

The recommendation system was implemented by first assigning dummy reward scores to each record based on their funded status as predicted by the multinomial naïve Bayes model combined with random numbers. Numpy's "randint" function was used to generate the scores in a way that applications predicted as not funded were assigned a random number in a higher range than records predicted as funded. This approach allowed predicted unsuccessful applications to have an assumed higher reward; and thus to be prioritised by the ϵ-greedy recommender system during the exploitation phase. The arms of the multi-armed bandit were represented by the "loan_id" attribute. The algorithm first decided if it was exploration or exploitation phase using Numpy's "random.binomial" function in a way that the probability of choosing exploration was ϵ. During exploration, the recommender system selected records from the pool of all applications using the "random.choice" function. For the exploitation phase, applications with the highest reward scores were picked for recommendations. This model was tuned by adjusting the probability of exploration (ϵ) versus exploitation, the number of recommendations to be made in each implementation and the number of times the same recommendation should be made.

6 Evaluation

This chapter presents the evaluation of the output of the main implemented models using the previously described metrics.

6.1 Multi-class classification / Predicting popularity levels

Multinomial naïve Bayes and logistic regression were implemented to predict popularity classes derived from the number of funding days. A 70 - 30% training and test split

provided the best results. Table 2 summarises the main average weighted metrics of these models.

Table 2: Multi-class classification results

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Multinomial NB	0.38	0.37	0.38	0.36
Logistic Regression	0.39	0.39	0.39	0.36

Both models achieved weak results with accuracy, precision, recall and F1-score all below 40%, therefore the use of the output of the multi-class classifications as input for the recommendation system was dismissed.

6.2 Binary classification / Predicting funding status

Three algorithms were run for the binary classification problem: Bernoulli and multinomial naïve Bayes and logistic regression. Table 3 summarises their evaluation results.

Table 3: Binary classification results

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Bernoulli NB	0.95	0.91	0.95	0.93
Multinomial NB	0.93	0.93	0.93	0.93
Logistic Regression	0.95	0.93	0.95	0.93

All three models performed similarly based on accuracy, precision, recall and F1-score with the Bernoulli naïve Bayes slightly falling behind on precision. Results were further evaluated using ROC curve and AUC. Figure 5 shows the comparison between the three implemented models and a random prediction. The Bernoulli naïve Bayes model's prediction does not exceed that of a random prediction, thus it performs poorly. Logistic regression achieved an AUC of 0.838 while the multinomial naïve Bayes algorithm produced an AUC of 0.813 indicating that logistic regression was the best performing classifier.

Considering the research goal, which was more representation of "unpopular" loan applications in recommendations to potential lenders, it was important to capture a large proportion of the true negative predictions, which can be expressed in the specificity metric, therefore confusion matrices were created and compared. Figure 6 shows the confusion matrices for the best performing classifiers based on the previous evaluation metrics. The Bernoulli naïve Bayes model did not identify any negative instances. It has to be noted that there was a class imbalance in the dataset as over 90% of the instances were funded.

The multinomial naïve Bayes model identified 4725 true negatives out of 26288 negative instances (specificity = 0.18), while the logistic regression prediction resulted in 175 true negatives out of 26308 negatives (specificity = 0.007). For the purpose of the fairness-aware recommender system, the multinomial naïve bayes classifier was chosen as it served the purpose of prioritising disadvantaged applications more, however, none of the classifiers performed well overall.

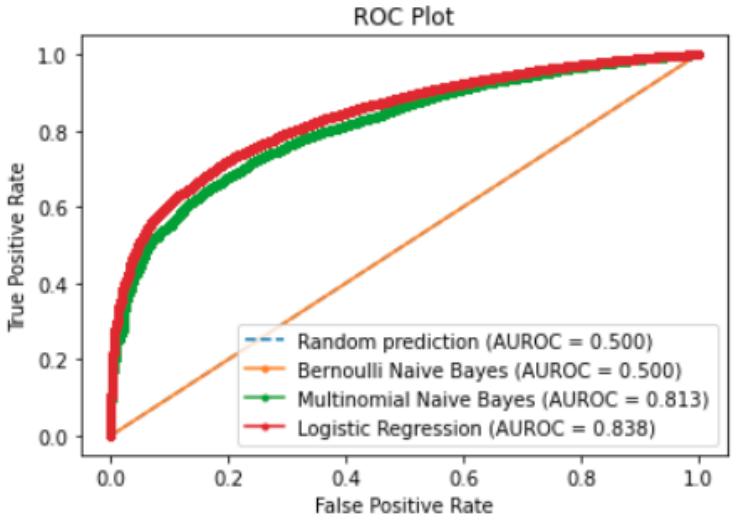
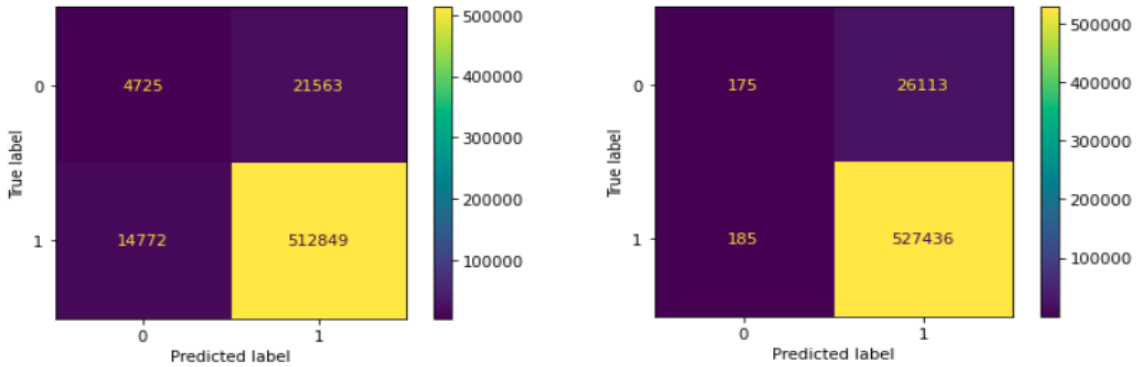


Figure 5: ROC and AUC



(a) Multinomial naïve bayes confusion matrix (b) Logistic regression confusion matrix

Figure 6: Confusion matrices for the binary classification

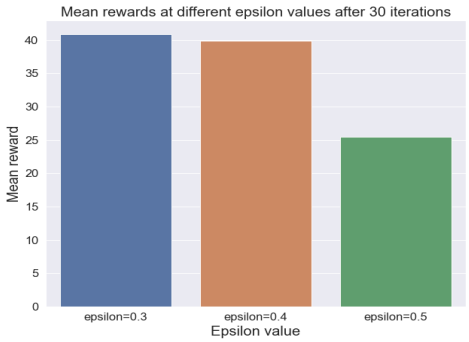
6.3 Recommender system

The ϵ-greedy policy- based recommender system was evaluated on the assumed reward (dummy score) of each choice. These dummy scores were arbitrary as they were substitutes for the "clicks" or "likes" that items would get from users in a live system. The recommender system was tested with three different ϵ settings: the tested models included $\epsilon = 0.3$, 0.4 and 0.5 exploration probabilities. The same number of iterations, batch and slate sizes were applied for each iteration. Batch size refers to the number of lenders the same recommendation should be served for and slate size indicates the number of loans in each recommendation. Figure 7 shows the achieved rewards with different ϵ settings.

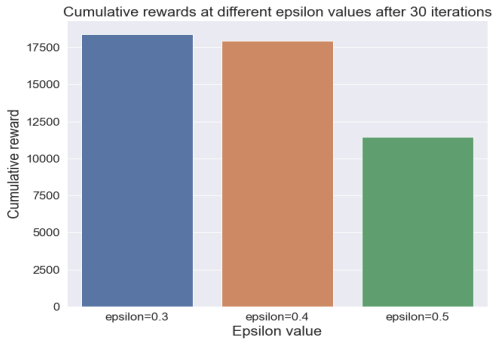
Rewards were highest when the probability of exploration was 30% (mean reward = 41) and lowest when the probability of exploration was 50% (mean reward = 26).

The model was also evaluated on the recommendations of predicted and actual status of loan applications as shown in figure 8.

The mean reward for loans predicted as not funded was higher than for loans predicted as funded, which was in line with the expectations, however, the comparison with the rewards of the actual funding status shows that the rewards of the two groups are nearly equal, successful applications have a slight advantage. This may be the result of the sub-

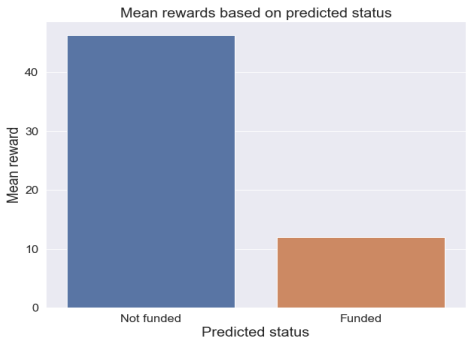


(a) Mean rewards at different ϵ settings

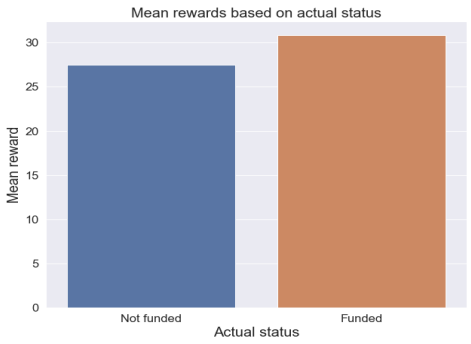


(b) Cumulative rewards at different ϵ settings

Figure 7: Rewards after 30 iterations



(a) Mean rewards of recommended loans by predicted status



(b) Mean rewards of recommended loans by actual status

Figure 8: Rewards based on predicted and actual funding status

optimal performance of the previous classification models. The last evaluation metric was the percentage of not funded loans included in the recommendations. Table 4 summarises the proportion of predicted and actual unsuccessful applications.

Table 4: Recommendations by funding status

ϵ **value**	Predicted not funded %	Actual not funded %
0.3	71%	6%
0.4	51%	5.3%
0.5	52%	5.1%

The numbers show that there was a significant gap between the proportions of predicted and actually not funded applications, which, again, may be due to the performance of the classification models.

6.4 Discussion

The above results represent the final and best performing models tested during this project. The evaluations indicate that the ϵ-greedy algorithm achieved a good result

in recommending "predicted not funded" applications based. A significant proportion of recommendations were selected from this group. After further clarification of the concept of fairness in a charitable lending context, the parameters of the model can be tuned further to achieve the desired results as it proved to be a feasible option for cold-start recommendations. In contrast, the classification models achieved poor results in identifying true negative instances, thus the input for the recommendation algorithm was of poor quality, which lead to a significant discrepancy between the recommendations of predicted and actual not funded instances. In the future, the classification models may be improved by including further loan attributes, such as text analysis of loan descriptions or the existing tags.

7 Conclusion and Future Work

The main objective of this paper was to propose a fairness-based recommender system for P2P platform, Kiva. A novel approach was proposed in this domain; instead of utilising historical user-item interactions, which is the basis of personalised recommendations and reinforces users' existing lending patterns, a generic recommender system was developed based on predicting application success and subsequently attempting to include more items in recommendation that were predicted to be unsuccessful. The evaluation of the applied method, the ϵ-greedy policy showed that the model was capable of offering diverse recommendations, which can be further tuned according to specific requirements. However, the overall performance of the recommender system was negatively impacted by the subpar quality of the classification models.

The limitations encountered during the development of the solution included difficulty in correctly predicting the success of applications due to the lack of strong predicting attributes beside gender, and the built-in randomness, which, on one hand, allows more applications to be covered by the recommender system, on the other hand, cannot guarantee a steady proportion of disadvantaged applications.

As shown in the literature review in chapter 2, charitable lending by definition is essentially driven by emotions and perceived usefulness as opposed to rational, financial decisions, therefore future work would benefit from collaboration with psychologists or marketing professionals in developing fair recommendations. A possible approach would be the implementation of certain tags and specific wording for less popular items, which would make them look more attractive to aspiring lenders. Similarly, further text analysis on applications could be beneficial to understand potential associations between the loan appeals and application success.

References

Auer, P., Cesa-Bianchi, N. and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem, *Machine learning* **47**(2): 235–256.

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H. et al. (2019). Fairness in recommendation ranking through pairwise comparisons, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2212–2220.

Burke, R., Sonboli, N. and Ordonez-Gauger, A. (2018). Balanced neighborhoods for multi-sided fairness in recommendation, *Conference on Fairness, Accountability and Transparency*, PMLR, pp. 202–214.

Burke, R., Voida, A., Mattei, N., Sonboli, N. and Eskandanian, F. (2020). Algorithmic fairness, institutional logics, and social choice, *Harvard CRCS Workshop: AI for Social Good*.

Burtch, G., Ghose, A. and Wattal, S. (2014). Cultural differences and geography as determinants of online prosocial lending, *Mis Quarterly* **38**(3): 773–794.

Cai, S., Lin, X., Xu, D. and Fu, X. (2016). Judging online peer-to-peer lending behavior: A comparison of first-time and repeated borrowing requests, *Information & Management* **53**(7): 857–867.

Desai, R. M. and Kharas, H. (2009). Do philanthropic citizens behave like governments? internet-based platforms and the diffusion of international private aid, *Wolfensohn Center for Development Working Paper* (12).

Isinkaye, F. O., Folajimi, Y. and Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation, *Egyptian informatics journal* **16**(3): 261–273.

Jenq, C., Pan, J. and Theseira, W. (2015). Beauty, weight, and skin color in charitable giving, *Journal of Economic Behavior & Organization* **119**: 234–253.

LeDoux, J. (2020). Offline evaluation of multi-armed bandit algorithms in python using replay.
URL: *https://jamesrledoux.com/algorithms/offline-bandit-evaluation/*

Lee, E. L., Lou, J.-K., Chen, W.-M., Chen, Y.-C., Lin, S.-D., Chiang, Y.-S. and Chen, K.-T. (2014). Fairness-aware loan recommendation for microfinance services, *Proceedings of the 2014 international conference on social computing*, pp. 1–4.

Ly, P. and Mason, G. (2012). Individual preferences over development projects: Evidence from microlending on kiva, *Voluntas: International Journal of Voluntary and Nonprofit Organizations* **23**(4): 1036–1055.

Microfinance Barometer 2019 (2019).
URL: *https://www.convergences.org/en/104906-2/*

Moss, T. W., Neubaum, D. O. and Meyskens, M. (2015). The effect of virtuous and entrepreneurial orientations on microfinance lending and repayment: A signaling theory perspective, *Entrepreneurship theory and practice* **39**(1): 27–52.

Park, B., Genevsky, A., Knutson, B. and Tsai, J. (2019). Culturally valued facial expressions enhance loan request success., *Emotion* .

Paruthi, G., Frias-Martinez, E. and Frias-Martinez, V. (2015). Understanding lending behaviors on online microlending platforms: The case for kiva.

Singh, P., Uparna, J., Karampourniotis, P., Horvat, E.-A., Szymanski, B., Korniss, G., Bakdash, J. Z. and Uzzi, B. (2018). Peer-to-peer lending and bias in crowd decision-making, *PloS one* **13**(3): e0193007.

Sonboli, N., Burke, R., Mattei, N., Eskandanian, F. and Gao, T. (2020). " and the winner is...": Dynamic lotteries for multi-group fairness-aware recommendation, *arXiv preprint arXiv:2009.02590* .

The Global Findex Database 2017 (2017).
URL: *https://globalfindex.worldbank.org/basic-page-overview*

VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*, " O'Reilly Media, Inc.".

Vermorel, J. and Mohri, M. (2005). Multi-armed bandit algorithms and empirical evaluation, *European conference on machine learning*, Springer, pp. 437–448.

Wang, C., Zhang, W., Zhao, X. and Wang, J. (2019). Soft information in online peer-to-peer lending: Evidence from a leading platform in china, *Electronic Commerce Research and Applications* **36**: 100873.

Yan, J., Wang, K., Liu, Y., Xu, K., Kang, L., Chen, X. and Zhu, H. (2018). Mining social lending motivations for loan project recommendations, *Expert Systems with Applications* **111**: 100–106.

Zhang, H., Zhao, H., Liu, Q., Xu, T., Chen, E. and Huang, X. (2018). Finding potential lenders in p2p lending: A hybrid random walk approach, *Information Sciences* **432**: 376–391.