Will a Loan Get Funding Quickly on Kiva?

Yuan Yin

07/05/202

INTRODUCTION

More than 1.7 billion people around the world are unbanked and can't access the financial services they need. Kiva is an international nonprofit organization with a mission to expand financial access to help underserved communities thrive. They do this by crowdfunding loans and unlocking capital for the underserved, improving the quality of financial services, and addressing the underlying barriers to financial access worldwide.[1]

After a borrower applies for a loan and before a loan is posted for lenders to support, the loan must go through the underwriting and approval process, where Kiva comes in. There are different ways to help Kiva set investment priorities, for example, building more localized models to estimate the poverty levels of residents in the regions where Kiva has active loans.

- Problem Statement*

This project aims to build a machine learning model that can accurately predict whether a loan posted on Kiva will get funding quickly or not. Kiva can prioritize the borrowers whose loans are more attractive to the lenders, which would be helpful, especially during tight funds. Also, this project can help to inform Kiva what an attractive loan looks like, such as if the borrower's poverty level would impact the funding speed or not, whether a loan needs at least one tag, and so on.

In this project, we set the threshold of funding speed at $50 per day, which is the median of the funding speed of our data. That is to say, on average, a loan that raised at least $50 per day was considered a loan that can be funded quickly and would be set as positive. If a newly issued loan is predicted as positive, Kiva will prioritize the borrower to post the loan on the Kiva website. What we have to be aware of is that the threshold is flexible, which Kiva can adjust based on the funds status.

- Target Feature Definition

According to the Kiva website, loans on Kiva usually have 30 days to fundraise successfully, and Kiva has expanded funding options for those loans missing their funding goals[2]. The fundraising days ranged from less than 1 day to 420 days, and about 93% of the loans posted on the Kiva website got fully funded. Because we don't

*: All the data used for modeling exclude the loans from the United States, Virgin Islands, Guam, and Puerto Rico.

know how long can every single loan expand its fundraising period, here, for those loans failing to get fully funded (the information of funded time was missing), we default their fundraising cycle as 48 days**. So for the fully-funded loans, the funding speed was equal to the funded amount divided by the fully-funded days; for the less-funded loans, the funding speed was calculated by the funded amount divided by 48 days.

- Metrics Selection

An error metric will help us figure out when our model is performing well, and when it's performing poorly.

Except for accuracy, we're primarily concerned with false positive rate and true positive rate[4]. With a false negative, we predict that a loan won't get funding quickly, but it actually would get at least $50 funding per day. This loses the borrower's potential financial aid, since Kiva didn't approve their loan request that actually would have gotten funding quickly. With a true positive, we predict that a loan will get funding quickly, and it really attracts lenders. This ensures that as many borrowers as possible are approved to post loans on the Kiva website.

It’s obvious that we should optimize for a high true positive rate, or recall in statistics, and low false positive rate, or fall-out in statistics. But generally, if we reduce false positive rate, true positive rate will also go down, i.e., lowering false positive rate coming at the expense of lowering true positive rate. So which metric will be prioritized depends on Kiva’s status. Trying to lower the false positive rate makes sense to Kiva if they want to ensure that limited resources can be fully utilized. If Kiva prefers ensuring as many satisfied borrowers as possible are approved to post loans on the Kiva website, a higher true positive rate will be their primary concern. It's worth keeping in mind the tradeoffs.

Beyond the single true positive rate or false positive rate, we also import the AUC score[5]. The higher the AUC value for a classifier, the better its ability to distinguish between positive and negative classes.

**: According to statistics, any observation outside the range of [Q_1-1.5*IQR, Q+1.5*IQR] can be considered outliers. Here, 48 days is the upper outliers of the getting-fully-funded days.[3]

DATA WRANGLING

Data Wrangling

- **Datasets**
  - The loans dataset was retrieved from Kiva's dataset on Kaggle, including all the loans posted on the Kiva website from 01/01/2014 to 07/25/2017.
  - The nation-level Multidimensional Poverty Index(MPI) dataset was downloaded from Oxford Poverty & Human Development Initiative (OPHI). Three attributes, i.e., MPI value, country name, and world region, were imported.
- **Main Steps**
  - Removed the columns that were redundant or not closely related to the project.
  - Imputed the missing values of categorical attributes. For example, filled the missing values of the tags factor with 'No_tag'.
  - Added fully-funded days attribute derived from two original attributes, i.e., posted time and funded time.
  - Merged the MPI dataset to the loans dataset.

DATA EXPLORATION ANALYSIS

Data Exploration Analysis

- **Step 1**: Compared the loans from the U.S. with other countries. (Note: Here, the U.S. refers to the United States, Virgin Islands, Guam, and Puerto Rico; the other countries refer to all the countries except the U.S. group.)

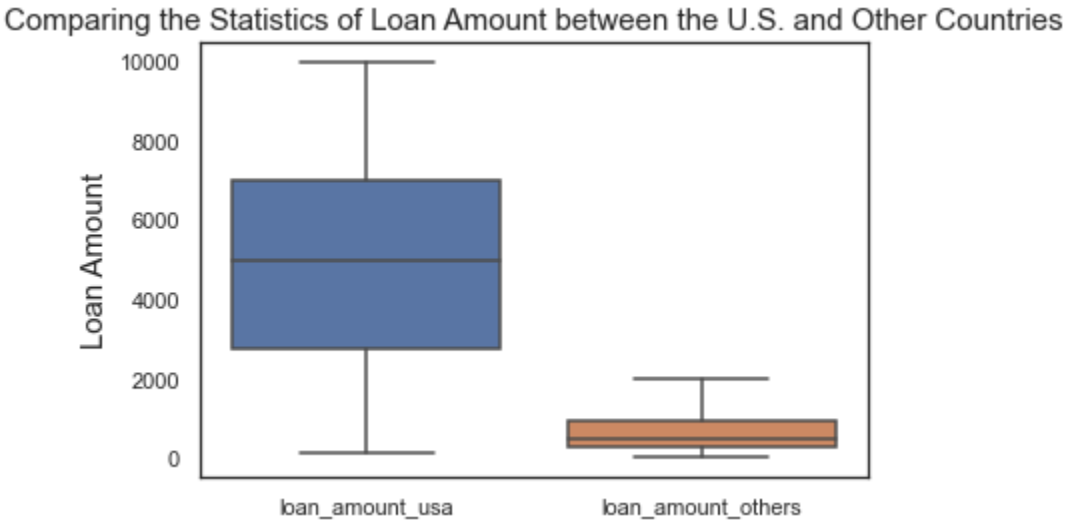

Figure 1

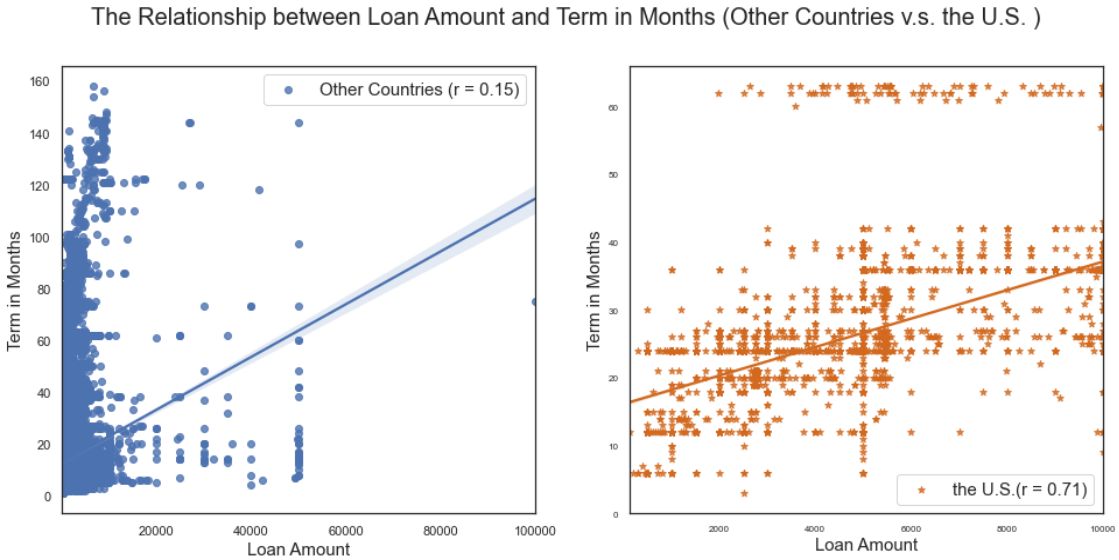

Figure 2

We can see big differences between the loans from the U.S. and the other countries [Figure 1]. The mean and median of loan amounts of the U.S. were much higher than the other countries. There was a significant correlation between the loan amount and term in months of the loans from the U.S., while such a relationship did not exist in the other countries [Figure 2]. About 98.7% of the loans from the other countries had field partners. But this ratio of the loans from the U.S. was only 17.7%, which means most of the loans from the U.S. were direct loans that can reach borrowers without field partners

involved. Besides, the fully-funded rate of the loans from the other countries was as high as 93.8%, while this rate was only 64% for the U.S.[Table 1]

Country With Partner		Fully-funded Rate
the U.S.	17.7%	64.0%
Other Countries	98.7%	93.8%

Table 1

Based on the significant difference between the loans from the U.S. and the other countries. All the entries related to the U.S., including the loans from the United States, Virgin Islands, Guam, and Puerto Rico, were removed. We only focus on the loans whose borrowers come from developing countries.

- **Step 2**: Analyzed the attributes of the datasets, excluding the entries from the U.S.
    - Distribution of Numeric Attributes

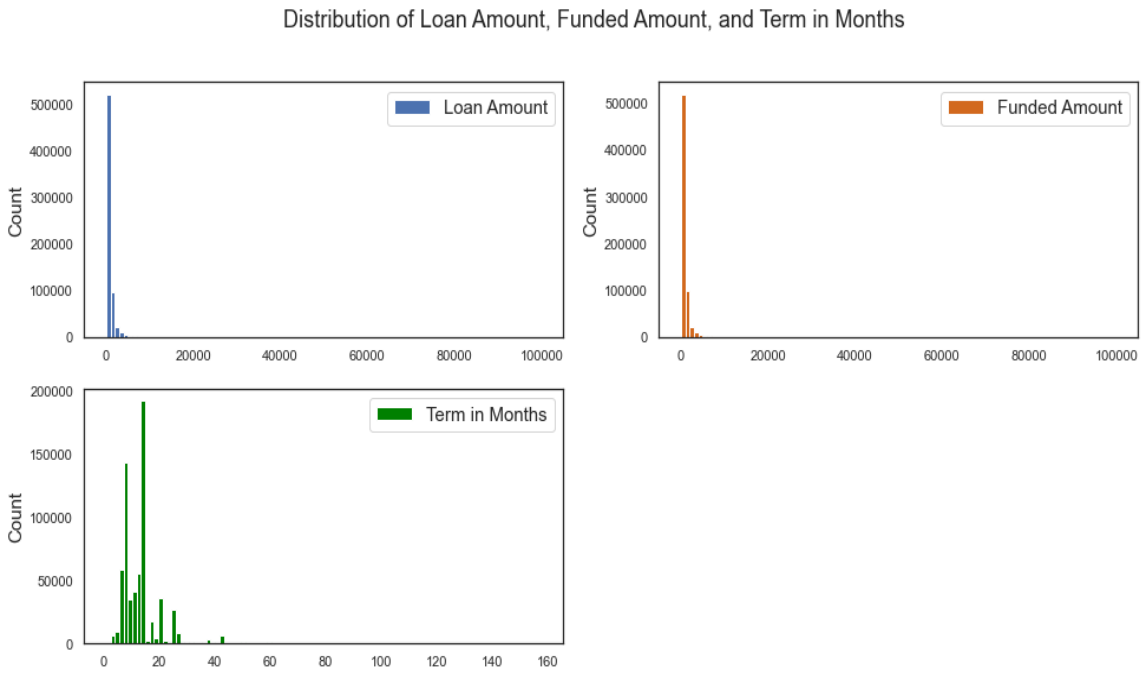

Figure 3

The distributions of numeric attributes [Figure 3], including loan amount, funded amount, and term in months, were all severely skewed to the right. The overwhelming majority of

the loans posted on Kiva were microloans (Microloans are normally defined as any loan for $50,000 or less); about 75% of the loans were less than $1000.

- Total Loan Amount v.s. Average Loan Amount

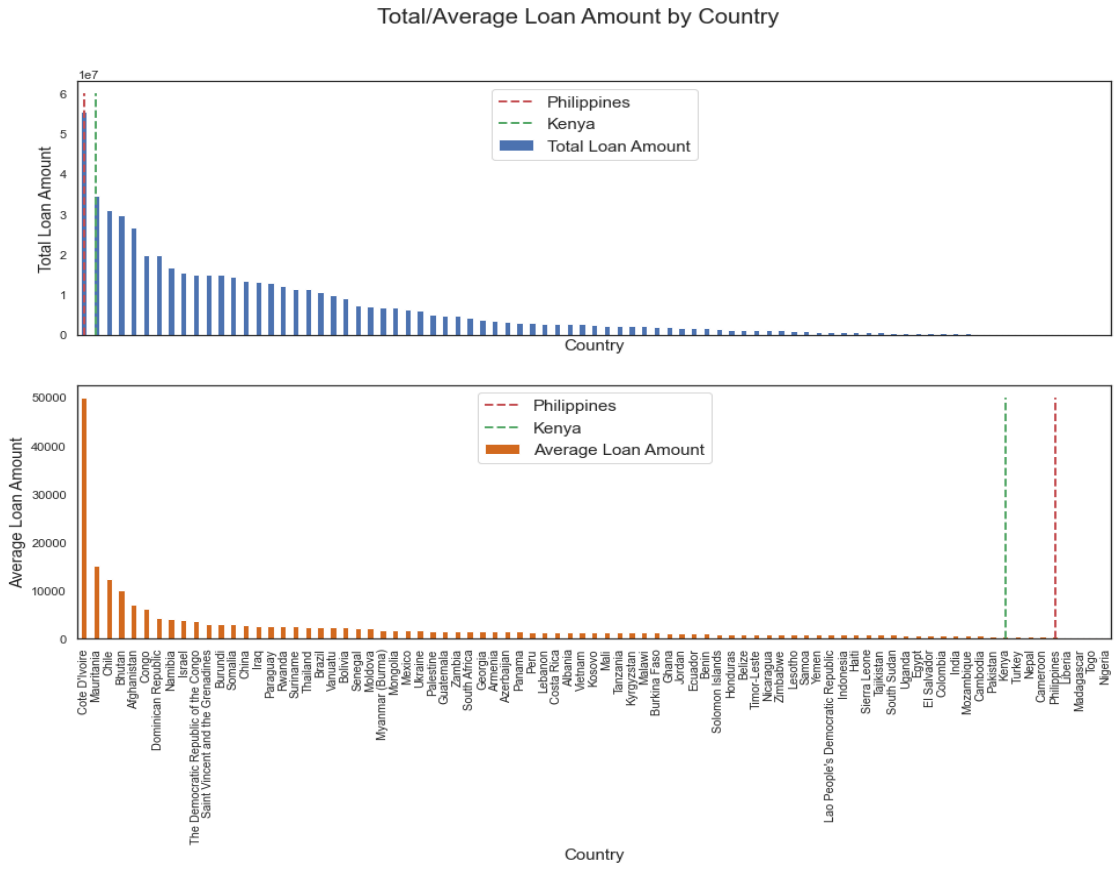

Figure 4

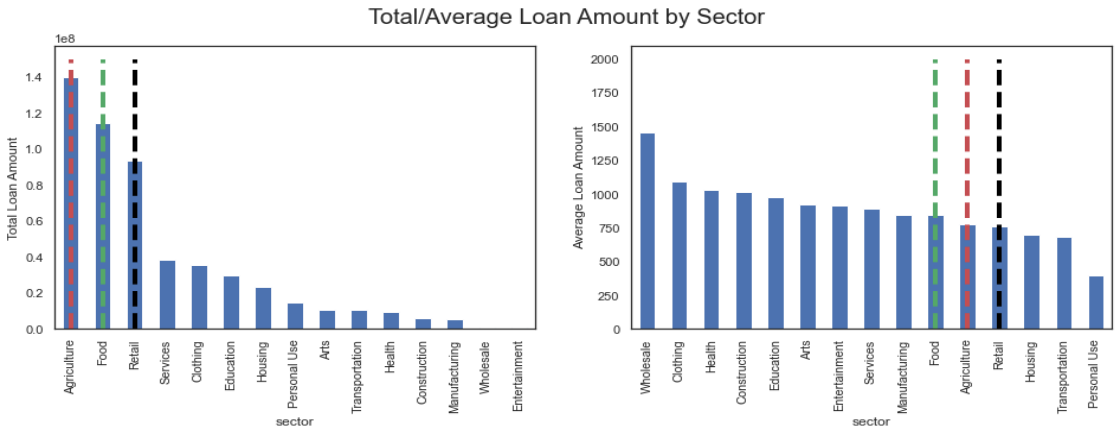

Figure 5

The two countries (Philippines and Kenya) with the highest total loan amount had very low average loan amounts [Figure 4]. The top three sectors by total loan amount were Agriculture, Food, and Retail, which account for 50%, while as far as the average loan value is concerned, these three sectors ranked low. About 73% of the loan borrowers were female individuals or female groups [Figure 6]. The average amount of the loans borrowed by the female was the lowest among the four gender groups, although the total amount of the female's loans was much higher than the others [Figure 7].

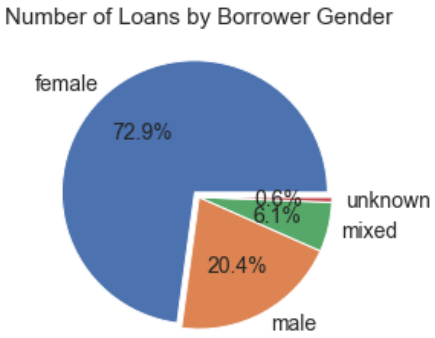

Figure 6

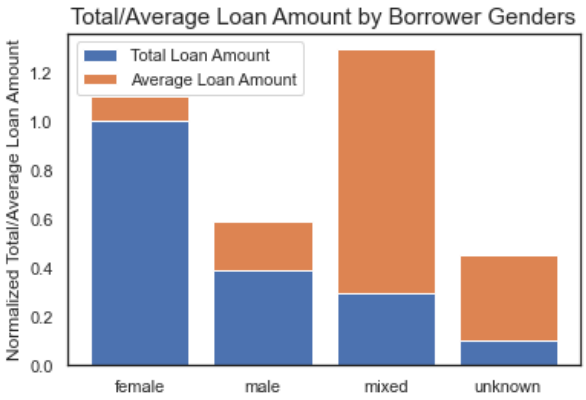

Figure 7

- **Step 3**: Defined the target feature derived from the processed attributes

The goal of this project is to build a machine learning model that can accurately predict if a loan will get at least $X fund amount per day on average when it posts on the Kiva website?

We defined the target feature as funding speed, i.e., the funded amount divided by fundraising days.

FEATURE SELECT AND MODELING

Data Pre-processing

Data Modeling

- Feature Selection

We removed the strongly correlated features to reduce the collinearity based on the

heatmap of the features [Figure 8.1, Figure 8.2]. Then we did feature ranking with recursive feature elimination and cross-validated selection(RFECV). Dropping correlated features deduced the number of features from 33 to 19, and finally, 5 important features were selected.

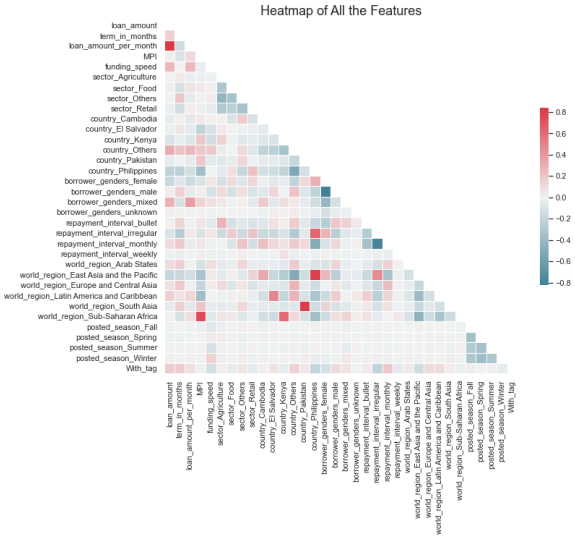

Figure 8.1

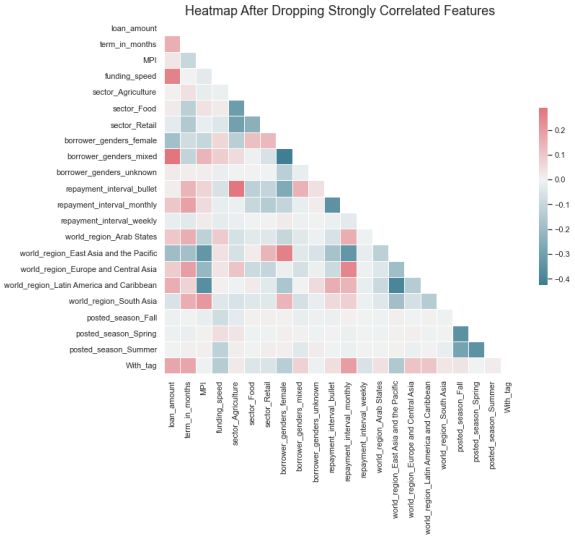

Figure 8.2

We compared the cross-validation accuracy scores between the dataset with 19 unrelated features and the dataset with 5 important features. We found the accuracy score of the 5-important-features dataset was a little bit lower than the 19-unrelated-features dataset and all-features dataset [Figure 9]. But if we take the time consuming into account, the dataset with only 5 important features would be a better choice.

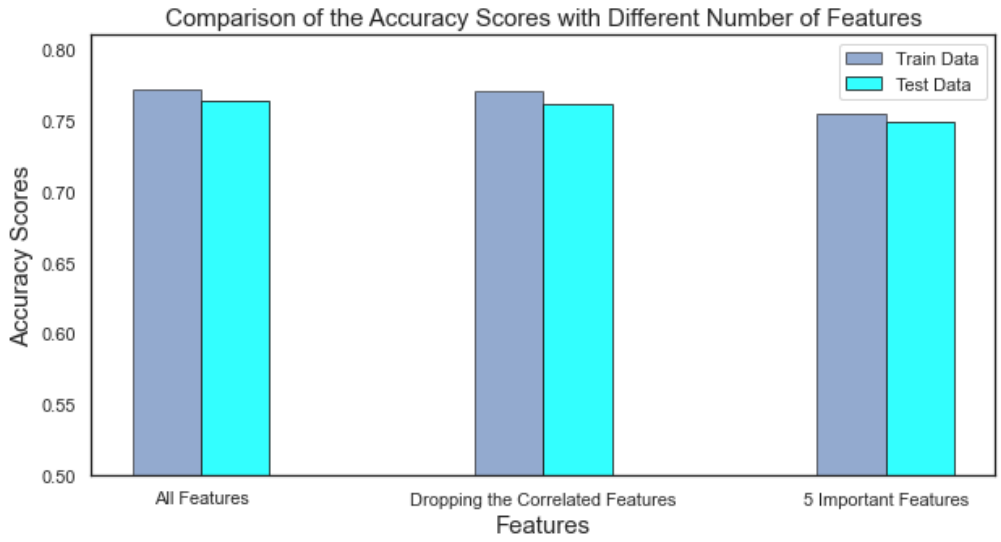

Figure 9

- Modeling

This project is a classification problem in supervised learning. We trained four classification models, optimized the hyper-parameters, calculated and compared the accuracy scores, AUC scores, true positive rate, and false positive rate.

1. Logistic Regression
2. K-Nearest Neighbors
3. Random Forest
4. Gradient Boosting

    - Comparing Accuracy Score[Table 2][Figure 10]

Model	Accuracy_train	Accuracy_test
Logistic Regression	0.721429	0.636667
KNN	0.658571	0.630000
Random Forest	0.735714	0.670000
Gradient Boosting	0.728571	0.620000

Table 2

Figure 10

- Comparing AUROC Score, True Positive Rate, False Positive Rate [Table 3][Figure 11]

Model	AUROC	True Positive Rate	False Positive Rate
Logistic Regression	0.736	0.630	0.309
KNN	0.709	0.500	0.204
Random Forest	0.761	0.630	0.228
Gradient Boosting	0.759	0.652	0.265

Table 3

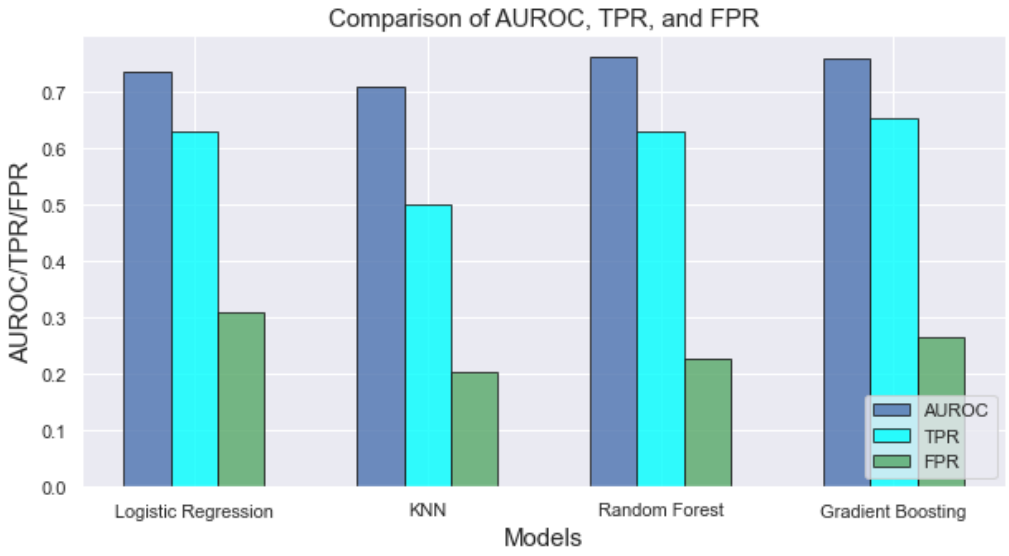

Figure 11

CONCLUSION

FUTURE WORK

1. The borrower's poverty level is critical for Kiva to set investment priority so it's better to estimate the welfare level of borrowers in specific regions instead of basing on the nation-level MPI. We would combine the analysis of loans information and localized methods for assessing borrower's poverty levels to provide a more valuable way to prioritize loan requests.
2. The threshold of the target feature is flexible. In this project, we set the point at 50%, i.e., the funding speed higher than the median being considered positive and the funding speed lower than the median being considered negative. If we set it at other points based on different situations, such as 90%, we need to consider the imbalance issue.

3. Overall, even using the model with the highest score, the performance of our prediction was not good enough. Except for future engineering or hyper-parameter adjustment, a better practice is to design what data should be selected and build models based on the well-designed data instead of existing data.

REFERENCES

1. Kiva Website: About us
2. Kiva Website: FAQs
3. Outlier
4. Confusion matrix
5. AUC-ROC Curve in Machine Learning