Will a Loan Get CrowdFunding Quickly

Springboard Data Science Career Track
Capstone Project

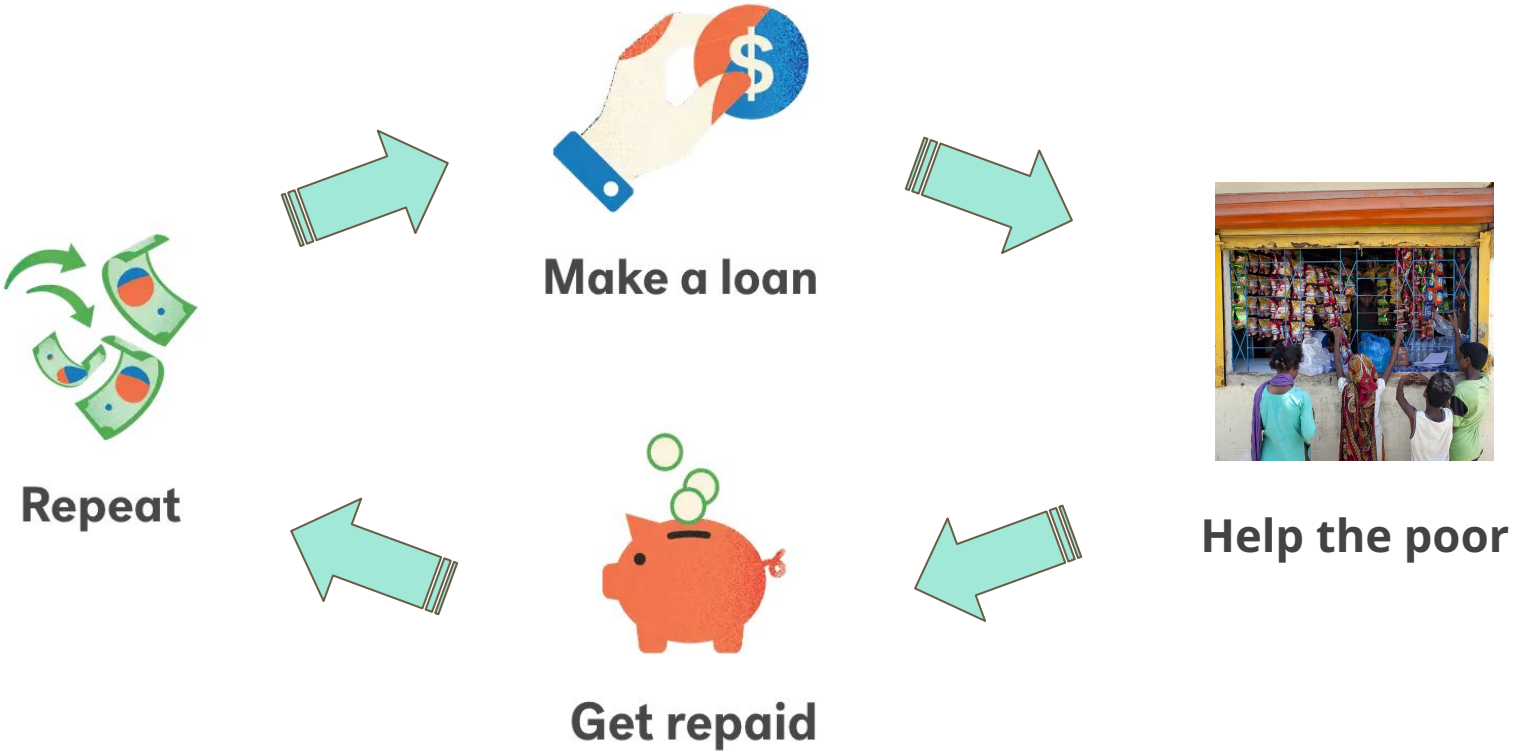
Yuan Yin, Akshay Jhawar(Mentor)

Overview

Hayat Kheir Imriri took out a microloan of $800 to buy stock for her store in a refugee camp for Palestinians in Beirut, Lebanon. *Sam Tarling/Corbis via Getty Images*

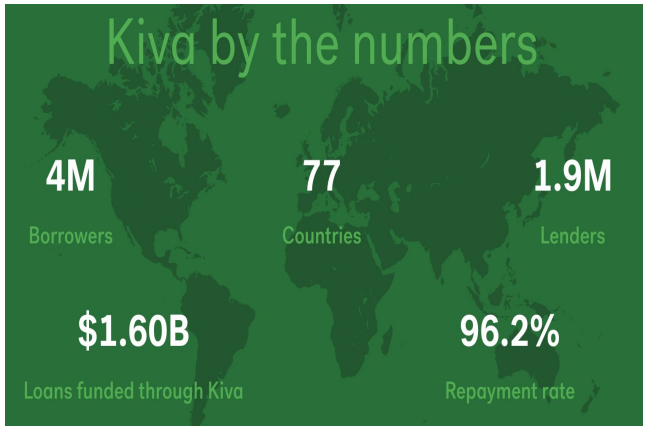
- **Microloan** is a small loan to someone or group in a poor country, which would lift them out of poverty.
- The number of microloan borrowers worldwide skyrocketed to **211 million** by 2013.[1]
- More than **3,000** organizations all over the world offer microloan.[2]

How Microloan works



Problem Statement

- Kiva.org is a crowdfunding platform to extend financial services to poor around the world.



- This project aims to build a machine learning model that can accurately **predict whether a loan posted on Kiva will get at least $50 funding per day.**

Datasets

- All the loans posted on the Kiva website from 01/01/2014 to 07/25/2017*
  - **665041 entries**
  - **20 attributes**
- The nation-level Multidimensional Poverty Index (MPI) dataset
  - **103 entries**
  - **3 attributes**

* : We excluded the loans from the United States, Virgin Islands, Guam, and Puerto Rico from the final dataset

Target Feature Definition and Metrics Selection

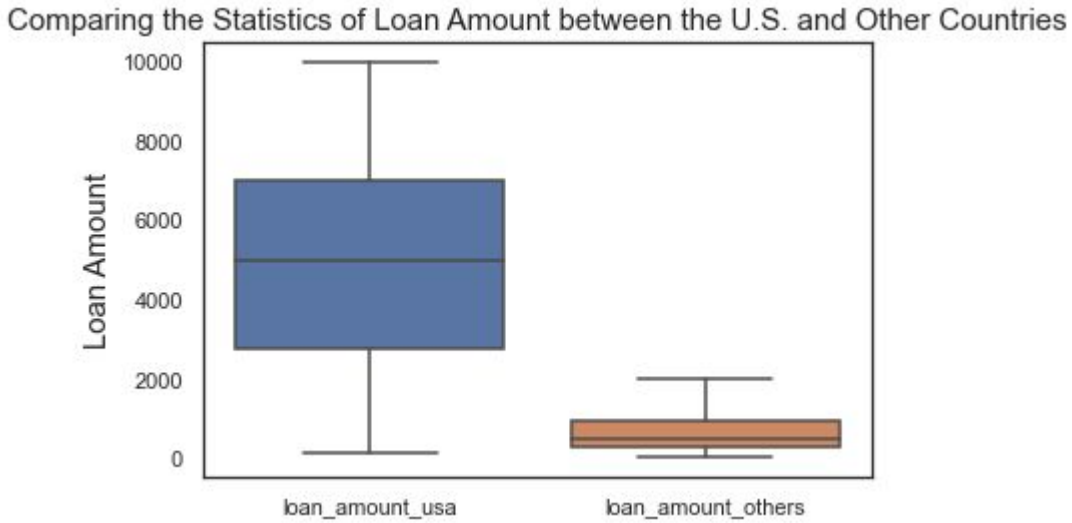
- Derived Target Feature

$$Funding\ Speed = \frac{Funded\ Amount}{Fundraising\ Days}$$

- Metrics
  - ROC_AUC Score
  - True Positive Rate / False Positive Rate / Accuracy

Data Exploration Analysis

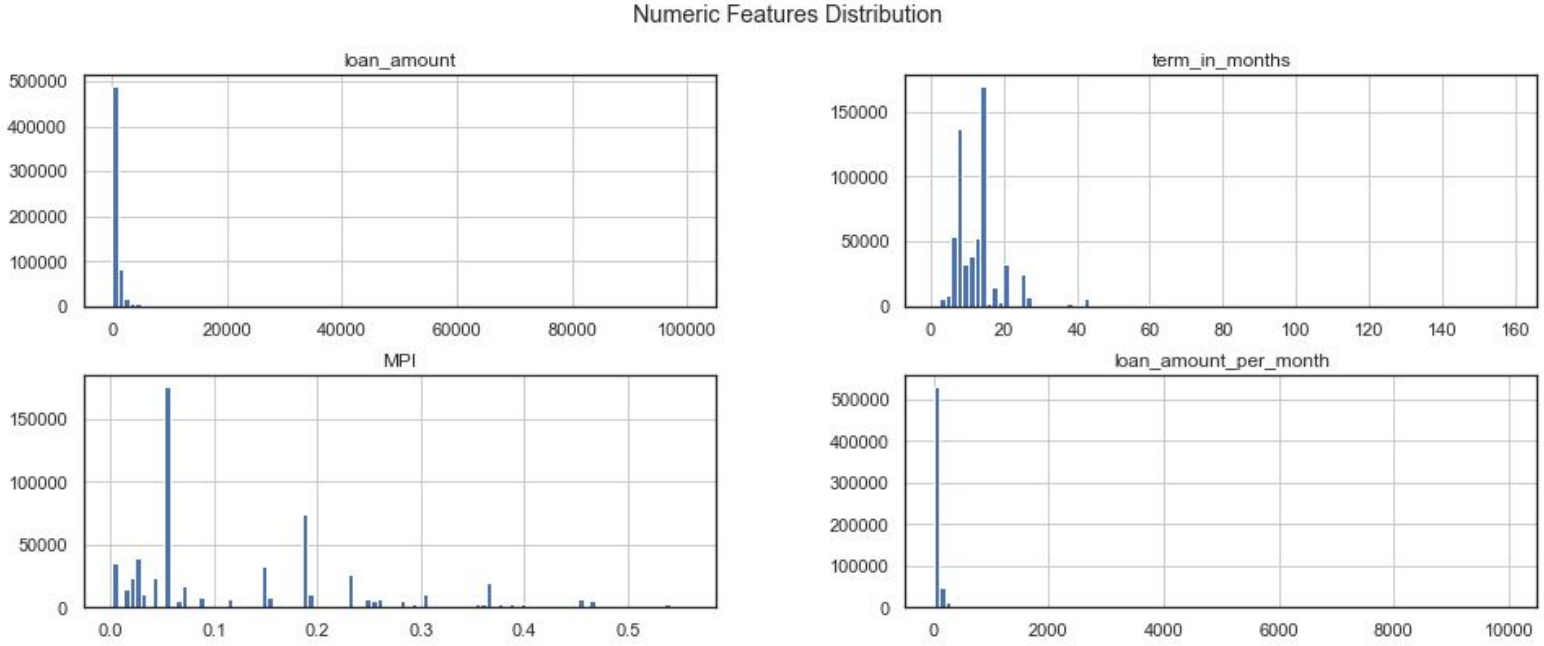
- The U.S. related countries v.s. the other countries
  - There were big differences between the two data groups
  - For this project, we will focus on **the other countries***



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

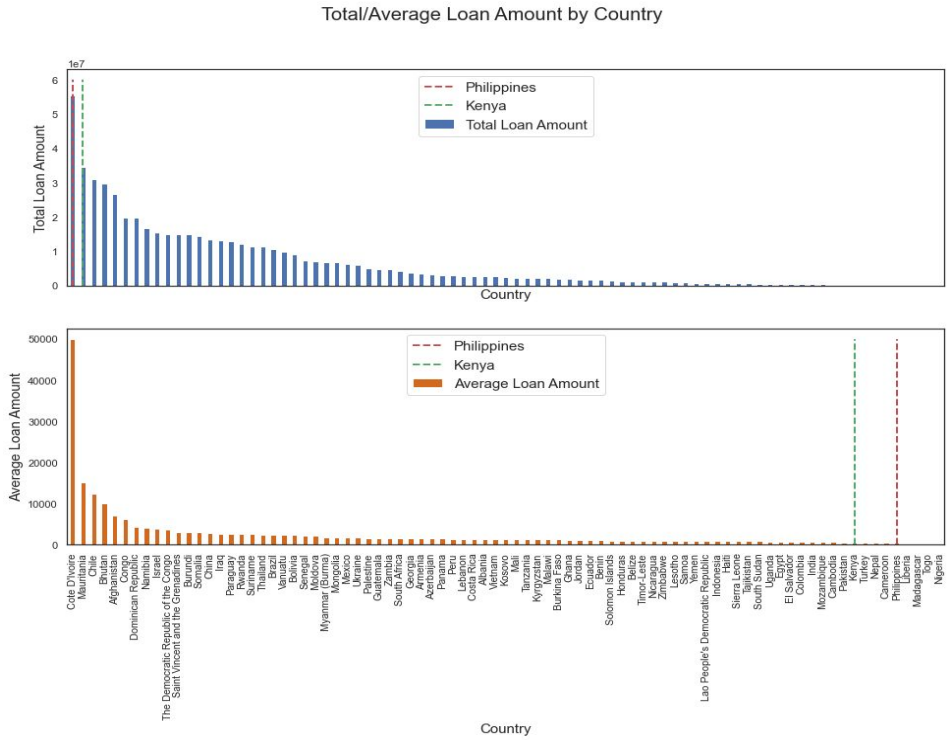
*: The other countries refer to all the countries where borrowers posted their loans on Kiva website, excluding the United States, Virgin Islands, Guam, and Puerto Rico.

- Numeric Attributes Distribution



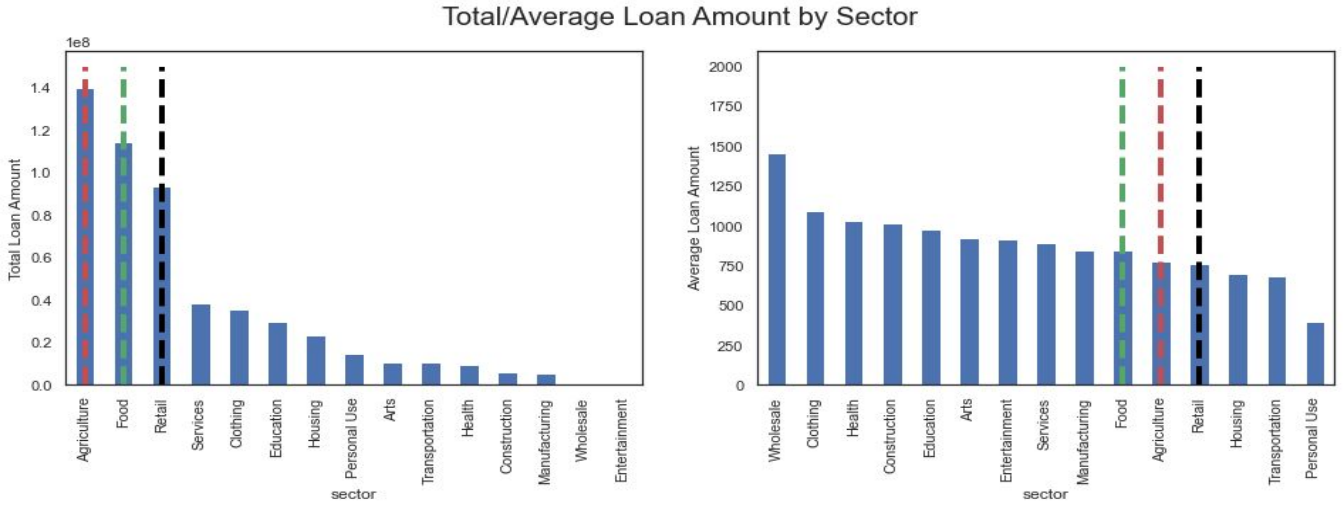
All the numeric features were **skewed to the right**.

- Total Loan Amount v.s. Average Loan Amount
  - **By Country**



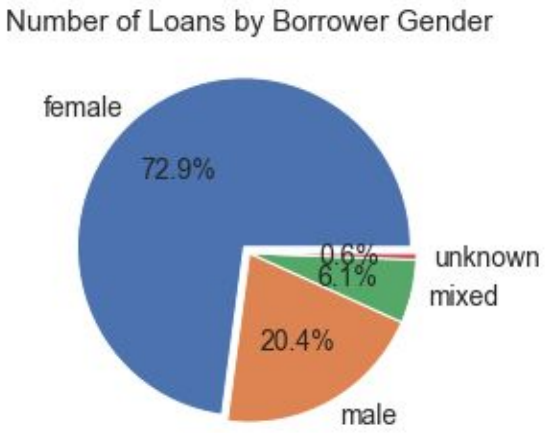
The top two countries with the highest total loan amount **ranked low** in the average loan amount.

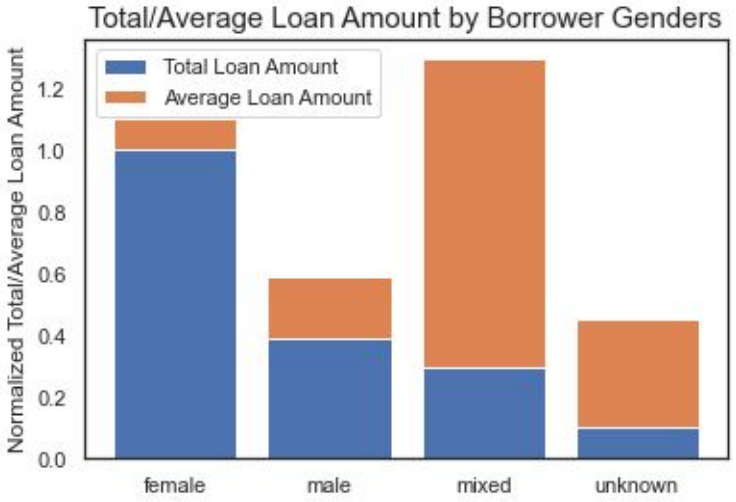
- Total Loan Amount v.s. Average Loan Amount
  - **By Sector**



The top three sectors with the highest total loan amount **ranked low** in the average loan amount.

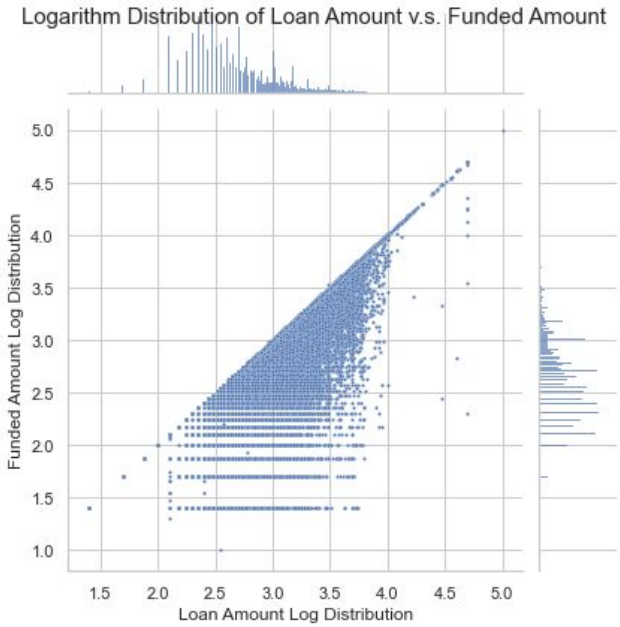
- Total Loan Amount v.s. Average Loan Amount
  - **By Borrower Genders**





The female(s) borrowers, accounting for the highest total loan amount, **ranked low** in the average loan amount.

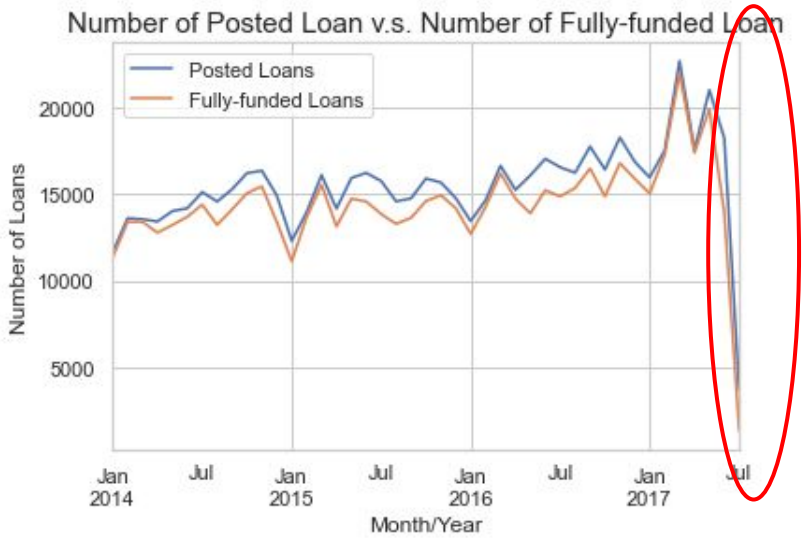
Loans Amount v.s. Funded Amount



Note: More than **93%** of the loans posted on Kiva got fully funded.

Fully-funded Loans

- **Timeline**

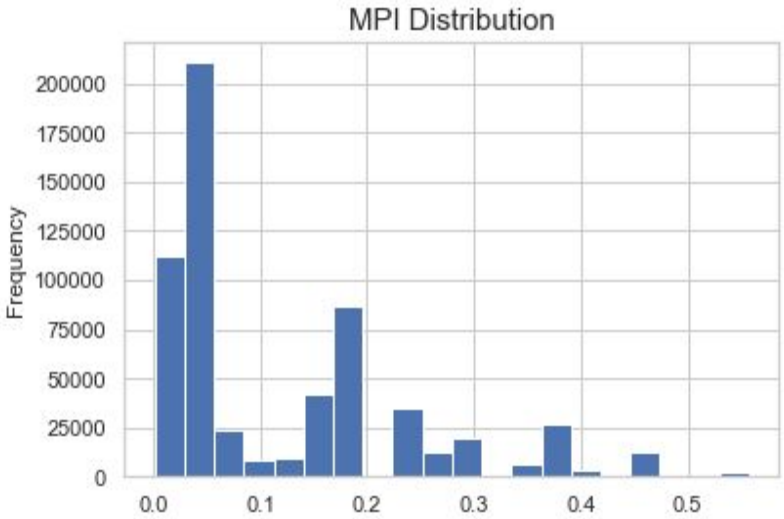


- **Statistics of Fully-funded Days**

count	618933.000000
mean	14.482962
std	14.108709
min	0.001389
25%	5.152674
50%	9.522847
75%	22.320324
max	420.573264

All the loans posted after 2017-06-07 will be dropped.

Multidimensional Poverty Index (MPI) Distribution



count	612393.000
mean	0.124
std	0.116
min	0.001
25%	0.052
50%	0.054
75%	0.187
max	0.557

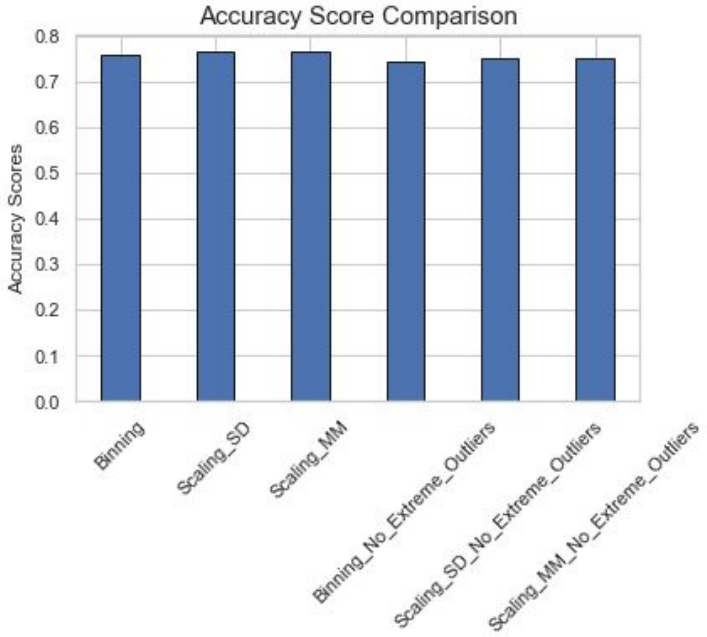
Most of the countries have very low MPL value.

Data Preprocessing

Problems	Solutions
Skewed Numeric Features with Extreme Outliers	Standard Scaler
Categorical Feature	One-hot
Feature Elimination	RFECV

Data Preprocessing: Numeric Features Scaling

- **Comparing different ways to process numeric features**
  - Binning
  - Removing extreme outliers
  - Scaling with StandardScaler and MinMaxScaler

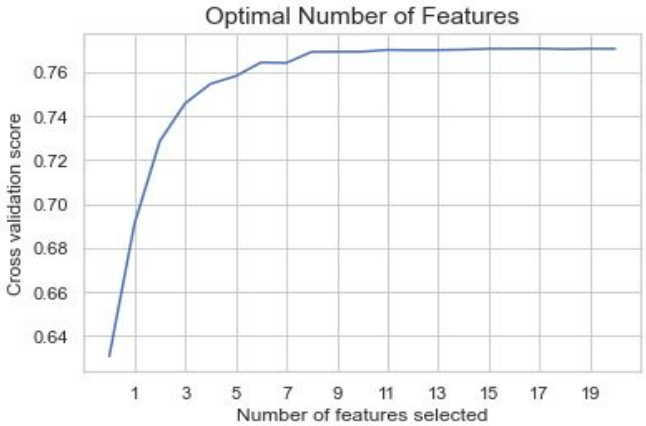


No significant difference

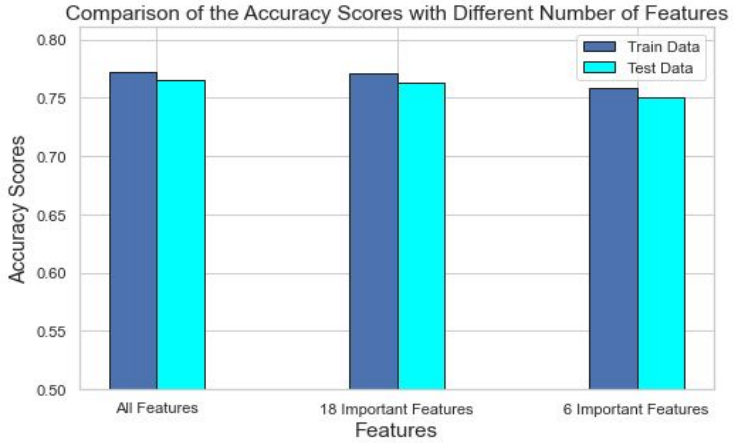
Data Preprocessing: Feature Selection

- **RFECV**

  18 Features are selected



- **Comparing Accuracy with different number of features**



Six features dataset has a little bit lower accuracy score but high computational efficiency.

Data Preprocessing: Feature Selection

- **Six Important Factors**
  - Loan amount
  - Term in Month
  - MPI
  - With tag or not
  - Posted in Summer or not
  - Borrower is female(s) or not
- **No feature is strongly correlated with the target feature, i.e., funding speed**

Target Feature: Customized Threshold

- **Statistics of Funding Speed**

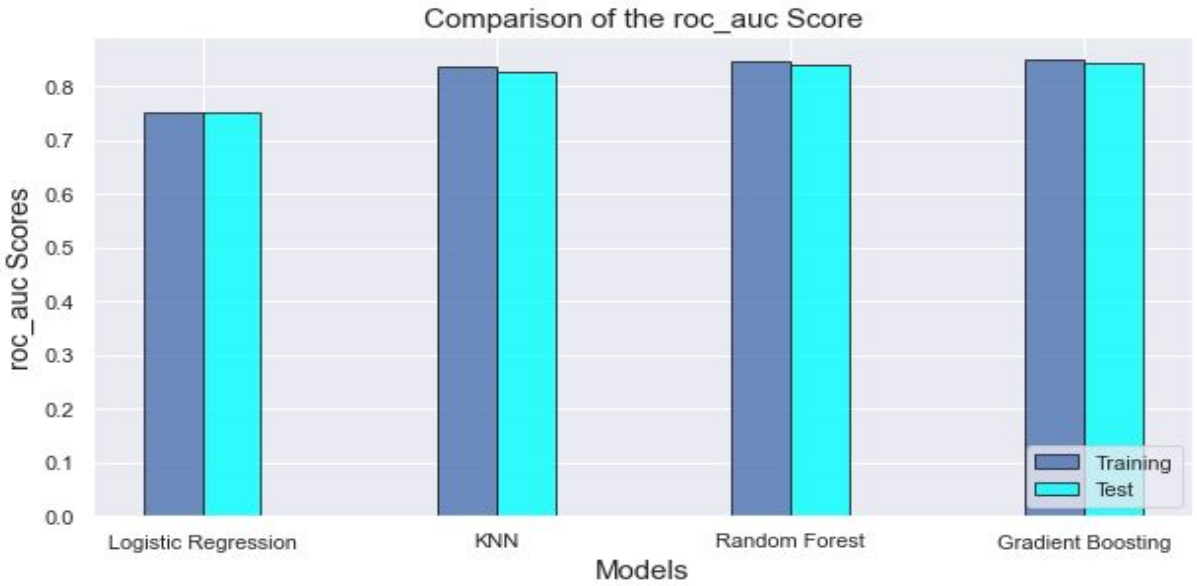
count	612393.000
mean	146.612
std	1078.561
min	0.000
25%	20.000
50%	40.000
75%	92.000
max	176327.000

- **Funding Speed Threshold**
  - For this project, we set the threshold at the **median**: the funding speed over 40 will be set as **1**; the funding speed that is less or equal to 40 will be **0**
  - We can **adjust the threshold** based Kiva's financial status
  - We can **predict the funding speed** with regression models.

Modeling

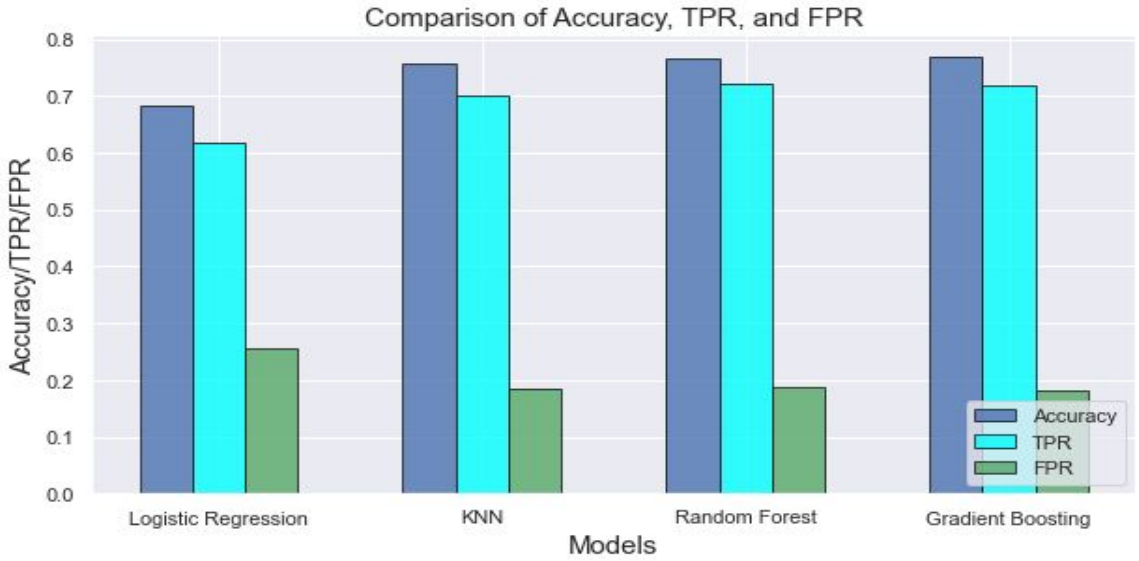
- **Algorithms**
  - linear: Logistics Regression
  - Non-linear: KNN
  - Enthemble: Random Forest
  - Enthemble: Gradient Boosting
- **Metrics**
  - ROC_AUC Score
  - True Positive Rate(TPR)
  - False Positive Rate(FPR)
  - Accuracy

Train/Test Data ROC_AUC Score Comparison



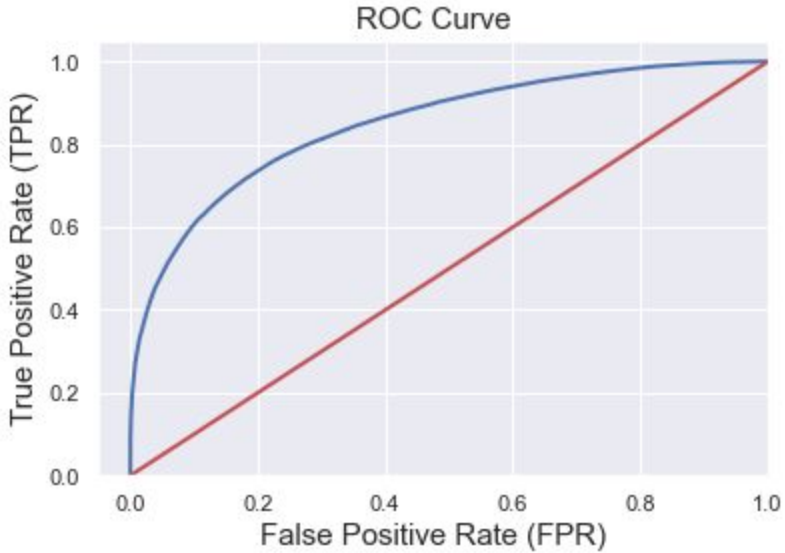
Model	roc_auc_train	roc_auc_test
Logistic Regression	0.750	0.751
KNN	0.835	0.828
Random Forest	0.846	0.841
Gradient Boosting	0.848	0.844

Accuracy/TPR/FPR Comparison



Model	Accuracy	True Positive Rate	False Positive Rate
Logistic Regression	0.682	0.619	0.256
KNN	0.757	0.700	0.185
Random Forest	0.767	0.722	0.187
Gradient Boosting	0.768	0.719	0.183

Best Solution: Gradient Boosting



- **Other Metrics**

	precision	recall	f1-score	support
0.0	0.75	0.82	0.78	92327
1.0	0.80	0.72	0.76	91391
accuracy			0.77	183718
macro avg	0.77	0.77	0.77	183718
weighted avg	0.77	0.77	0.77	183718

Summary and Recommendation

- In this project, we focus on the loans from the countries **outside the U.S.**
- We used a **adjustable threshold** of funding speed here; We can select suitable threshold based Kiva's financial status or predict the value of funding speed with regression models.
- Six features play the more important role in Predicting funding speed: **loan amount, term in months, MPI, tag, post in summer, borrower gender.**
- There's no significant difference between KNN, Random Forest, and Gradient Boost; **Gradient Boost is slightly better**, of which the roc_auc score is 0.844, accuracy score is 0.768, true positive rate is 0.719, false positive rate is 0.183.