Capstone Project 2 Final Report

Problem statement:

The goal is to understand what are the reasons that are causing higher delays between the time someone has posted a loan to the time it gets funded by the lenders and the time when it gets disbursed to the borrower. It is important to know the reasons in a way to understand the important steps in the process and take necessary actions to resolve the issue. It would also help lenders to get return of the borrowed money and they can lend again many times. Kiva can think about the resolving steps depending upon the analysis.

Dataset:

- Data is obtained from

  https://www.kaggle.com/kiva/data-science-for-good-kiva-crowdfunding

  https://www.kaggle.com/codename007/a-very-extensive-kiva-exploratory-analysis/data

Data Cleaning:

After looking at the overview of the data, changed few column names for better understanding and changed the datatypes which has wrong dtype.

Variables Inspection and cleaning:

Main variables required for performing analysis are

1. Posted_time
2. Funded_time
3. Disbursed_time

Posted_time:

There were no missing values or incorrect data.

Funded_time:

There were few missing values and there might be chance those null values might correspond to loans that were funded, but there isn't a way to fill those values. But all the null values might not be a missing value, there are chances when a requested loan might not be funded.
So, removing the cases where **loan_amount – funded_time <= 0** and keeping the data that is >

Discrepancy check for remaining values:

Comparing the columns to check for any discrepancies in the data

Funded_time vs funded_amount:

- There were no discripencies when compared to funded_amount. However, there were data where there is difference between loan_amount and funded_amount. Removed the data where **loan_amount – funded_amount < 0** but the funded_time was filled. (funded_time is only filled when loan_amount is is fully funded by lenders)

funded_time vs status:

- Status has four types – funded, refunded, expired and fundraising
- Comparing funded_time with funded, refunded and fundraising status there were no issues.
- There was inconsistency in data when compared to expired status. There were a couple of cases where the status was noted as expired but funded_time has data filled. Checking the remaining column values it was known that loan was funded. Hence, changed the status from expired to funded.

Disbursed_time:

- There were some cases which were null. But there might be chances where loan was funded but not yet disbursed. Hence not removed the data for further analysis.
- The timing of the disbursal can vary. For most Filed partner loans, the money is pre-disbursed for the borrower to use. Hence, sometimes the disbursed_time is made before the posted_time. For direct loans, money is disbursed only after the amount is fully funded by lenders.

Posted_time vs funded_time:

As discussed above there is no order between posted_time and disbursed_time and also between funded_time and disbursed_time. But the bias can happen with data such as funded_time appears before posted_time. There were few similar cases where funded_time was way before the posted_time and all correspond to first posting date. Hence, removed all these cases from analysis.

Other Variables:

- funded_amount column was compared to num_lenders_total column and no issue found.
- If funded_amount >= loan_amount, then status != expired. No discrepancy found
- Check of misspelt words: No issues found for columns activity_name and country_name.
- lender_term has only 23 missing values and can be ignored

Search for Outliers:

Loan_amount and funded_amount:

In previous analysis we ignored the cases where funded_amount is greater than loan_amount as it can be the case. But, have to check for outliers as there might be a chance. Looking at general statistics there were no significant cases.

Posted_time and funded_time:

Earlier I excluded all the cases where funded_time was before the posted_time, hence it is not required to look for outliers. Considered all remaining data after making the comparison with expiration time column.

Posted_time and disbursed_time:

There can be chances where amount is disbursed before the posted as amount for the Filed partners. It is necessary to check outliers though. Considered the cases where disbursed_time is before posted_time.

- Used z-score function to detect outliers
- Few outliers were found with threshold 3. In all of those cases, the difference between disbursed_time and posted_time is minimal. It means that although disbursal was made before the posted_time, loan was funded immediately. This gives assumption that kiva handled those cases. Hence, considered those cases for analysis.

Funded_time and disbursed_time:

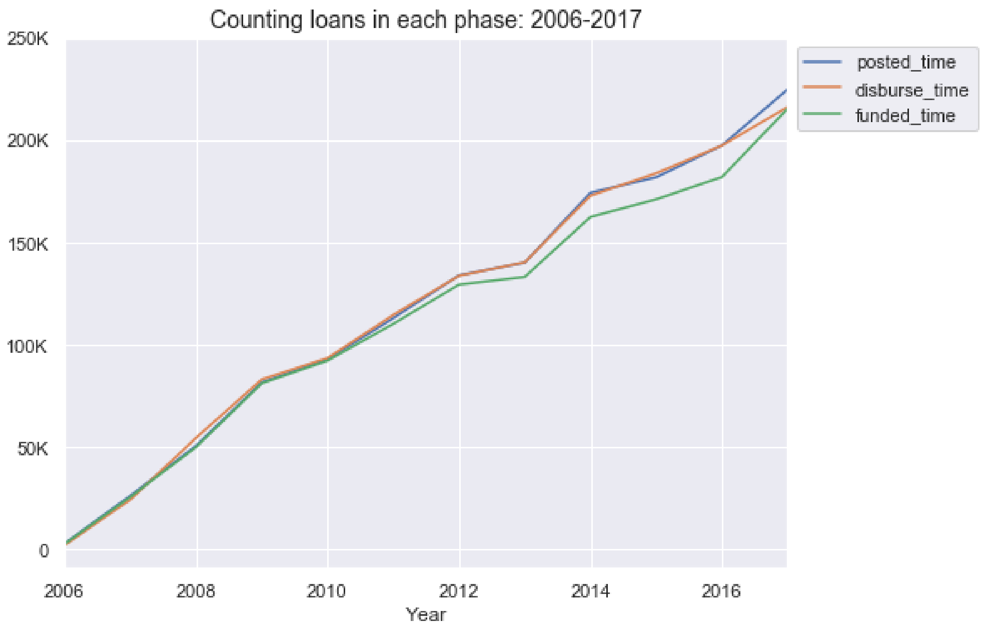
Considered only cases where disbursed_time appeared after funded_time. After dividing the cases where the field partner is involved or not, a few outliers were found by using the similar z-score function which was used earlier with threshold 3. Looking at the distribution of these cases they were among the small number of countries. So, decided to keep for the analysis.

There were few cases with missing disbursed_time but funded_time was filled. All of them are from Kenya and USA and the values were off the scale. Removed those values.

Exploratory Data Analysis

Now the data is clean, investigate data using few visualizations.

Characteristics of the main variables posted_time, funded_time and disbursed_time over the period. Considering the time period between 2006 and 2017 and looking at the evolution of number of loans there are three major distinctive time periods.



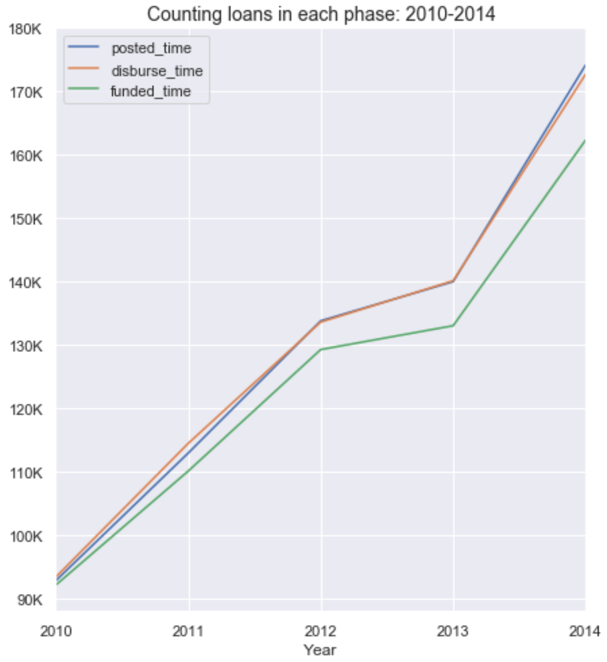
Findings:

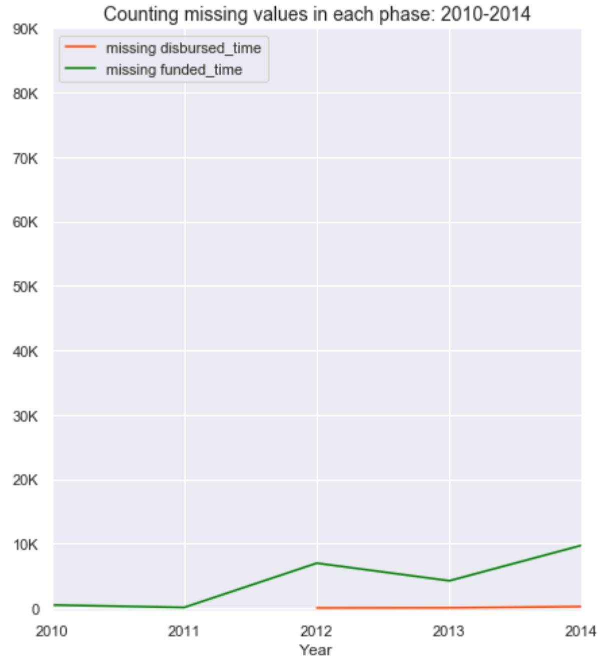
1. Until 2010, the number of posted loans are matching with number of loans funded and disbursed.
2. Between 2010 and 2014, there is a clear separation between each of the variables. It shows that loans that were being posted has close match with disbursement but they were losing efficiency with funding's.
3. After 2014, The differences continued between postings and funding and number of disbursals and postings became a bit volatile. But, in the last year funding gained proximity with number of disbursals.

There might be reasons as missing values or other factors. It needs a further exploration.

The period between 2010 and 2014 has discrepancies between postings and funding. Mainly between 2010 and 2011. This difference could be due to the cyclical nature of the process as it was later observed. The difference was mostly due to the funded missing values.

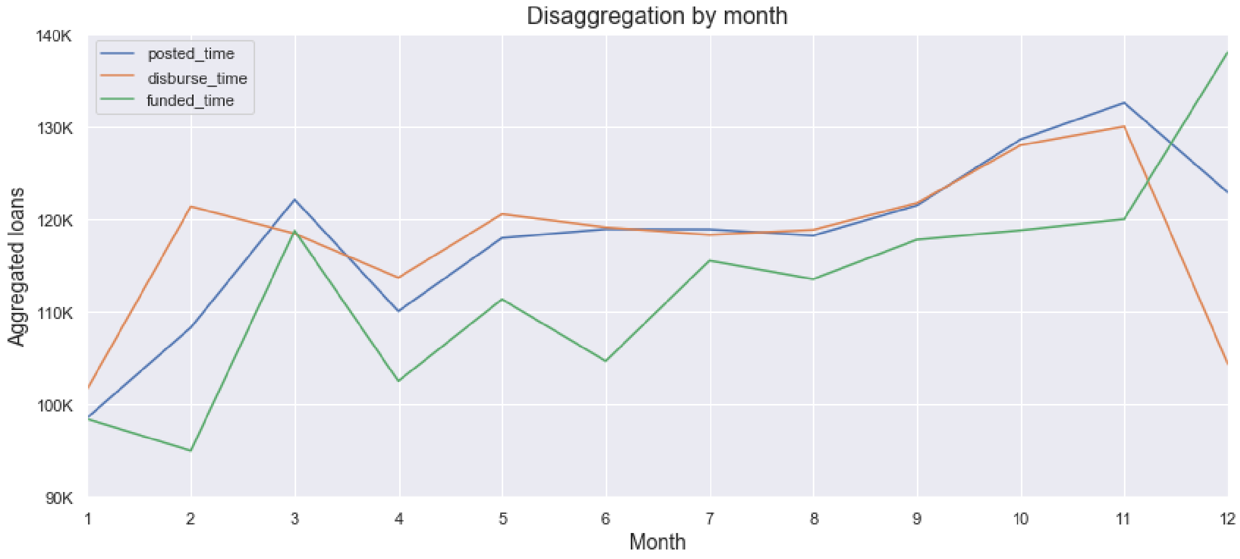
In the last phase, though the difference between the number of missing values of the disbursals and funding remained relatively constant, the number of funding's augmented more than disbursals equalizing them on final stretch.





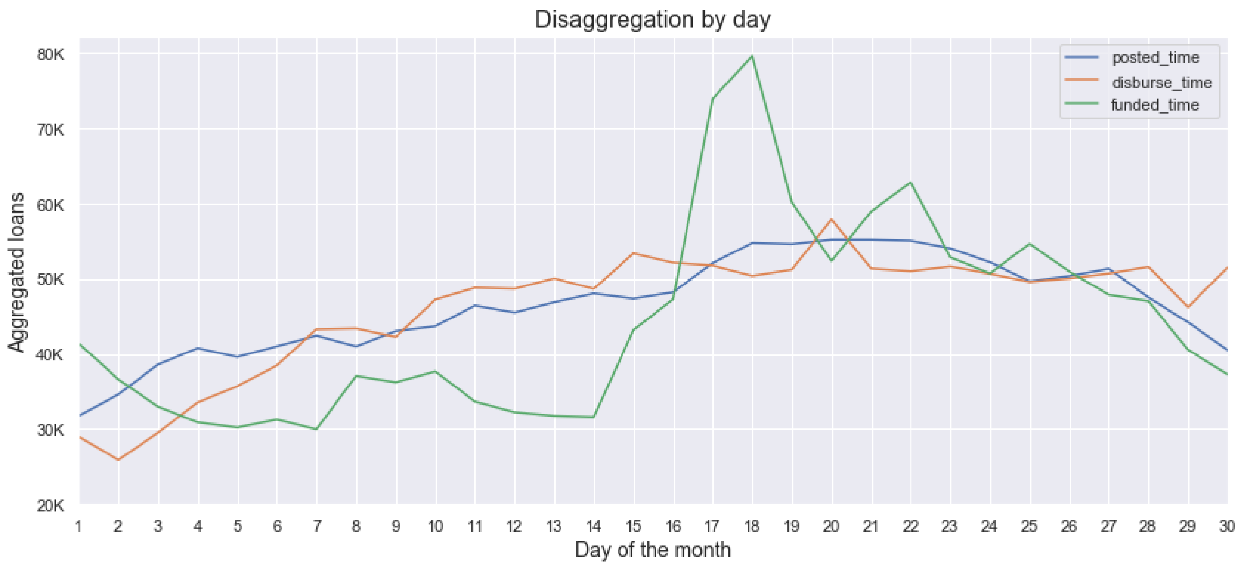
Between 2010 and 2014, the funding started to lose track of the posted loans mainly between 2010 and 2011. This difference, however, could be due to the cyclical nature of the process, as it will be shown below. Thereafter, the difference was mostly due to the funded missing values.

To further examine this cyclical nature, let us now display the aggregated values by month and finally, by day.



As seen, there is a slight tendency to post, disburse and fund as the year comes to an end.

While the posting and disbursing periods relatively aggregate the same amount of loans throughout the month, there is a strong pressure to fund loans during 3rd week of the month.

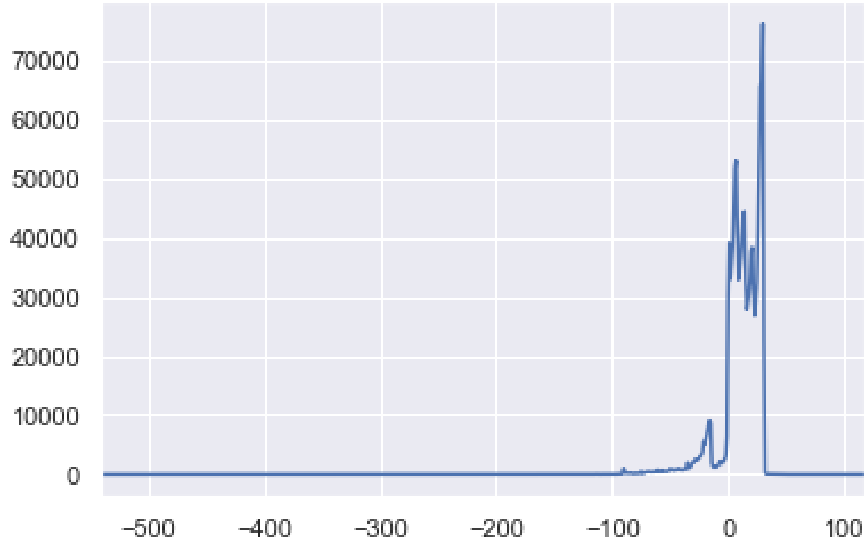


Now tried to get answers for below mentioned questions:

Time differences between main variables:

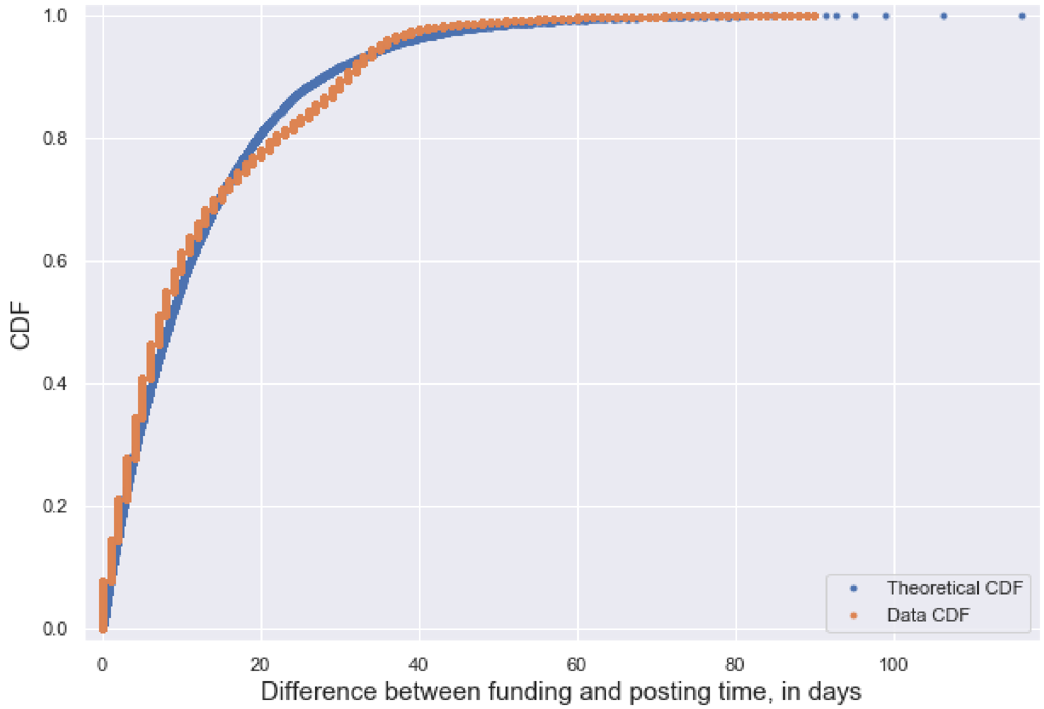
Posted_time and disbursed_time:

Looking at the glimpse of the distribution of posted_time - disbursed_time.



98% of the differences between the time of disbursal and posting of loans occur on a range of -35 days and +32 days. Moreover 90% of the differences are positive that is pre-disbursals. We could then immediately say that at least 90% of the disbursals are made by field partners, only ones who can pre-disburse.

When we look at that time a loan takes to get funded since the time it is posted on kiva website, we see that 99% of them are funded within 50days and almost 90% within a month. In fact, this behavior follows an exponential distribution as presented in the CDF:



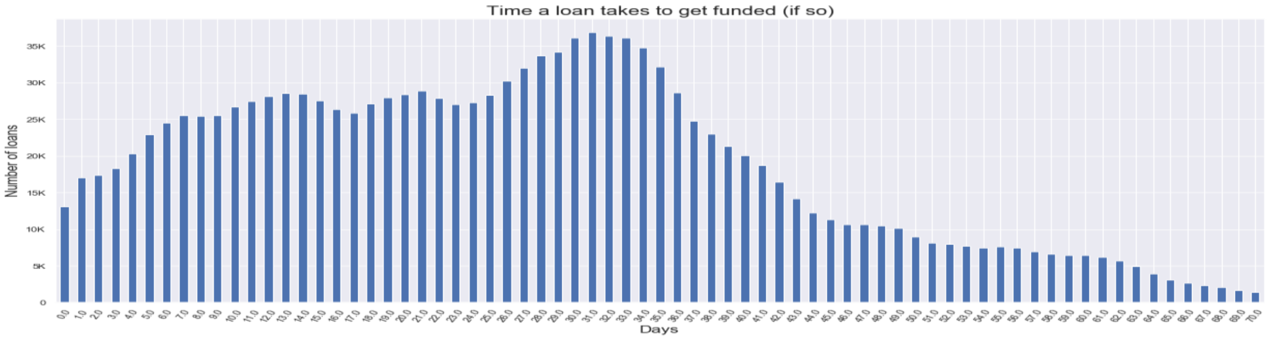
Close to 10% of the total loans are disbursed after the funding is complete and 5% occur 14 days after the time of funding. This clear spike suggests that 2 weeks could be a pre-staged arrangement. In most cases, a pre-disbursal occurs and only then funding period commences. I also confirmed that the 5% cases where disbursal occurs 14 days after funding are disturbing over time.

Then considering the real time a loan takes to get funded, that is, looking at the difference between the funding time and the disbursal or posting time, depending on which of these occurred first. The following facts were checked

- Close to 60% are funded within a month. Some irregularly can be seen during this period. After that they start to get the funding in an exponential fashion way.
- 97% are funded within 2 months.

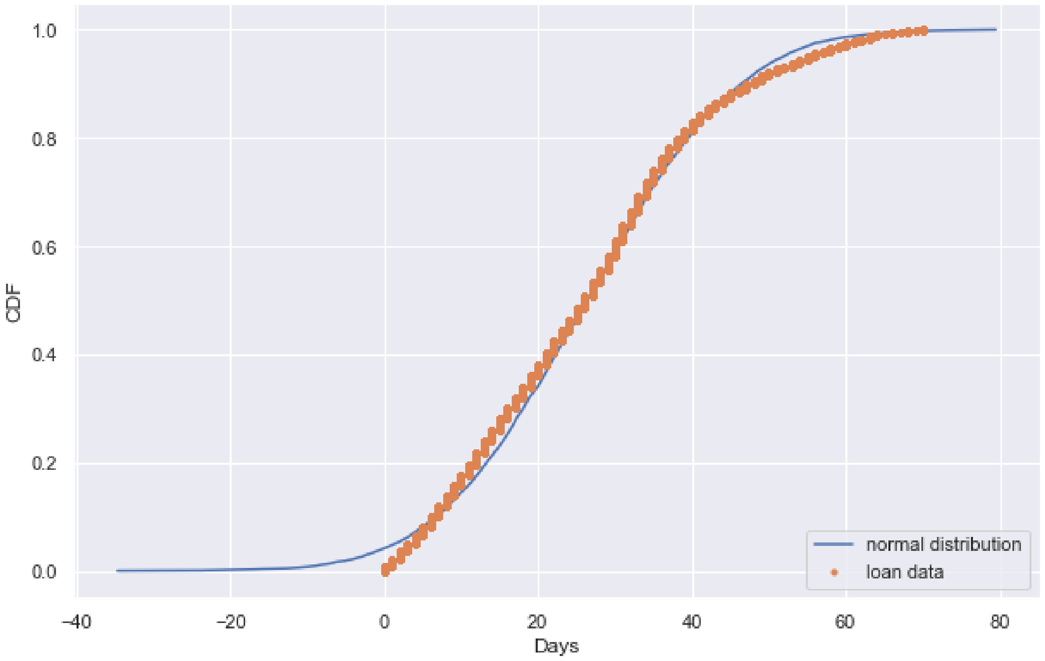
Also, to keep in mind that, 8% of the total loans are only disbursed after they are funded. Considering these cases means that 99% of the loans are funded with in 2 months and 2 weeks. Using D’Agostino’s K-squared test, we reject the hypothesis that the funding period follows a normal distribution – with a null p-value.

PDF:



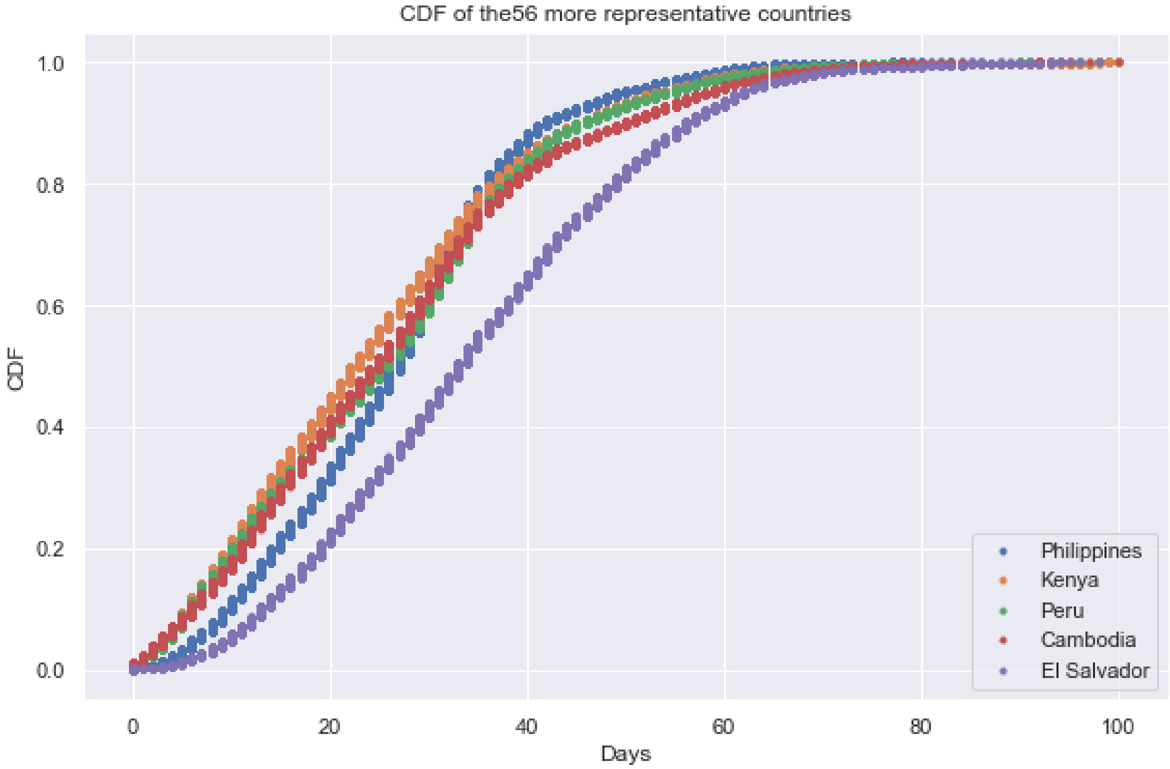
This test encompasses negative values, so when looking at the ECDF and comparing it with the theoretical CDF of a normal distribution, the resemblance is a bit more clear.

ECDF:

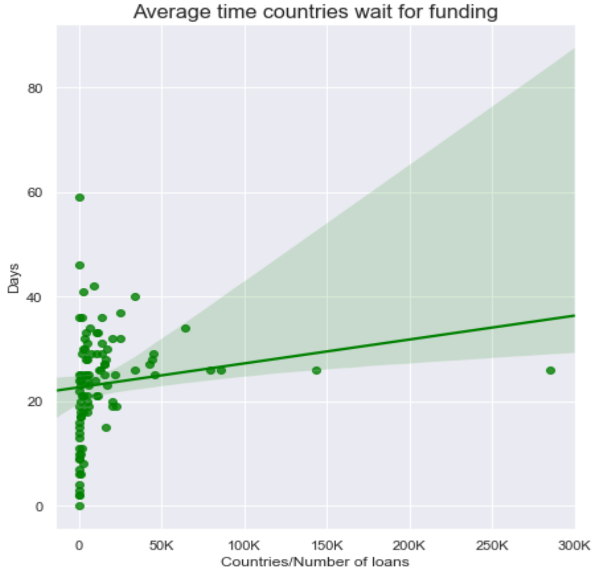


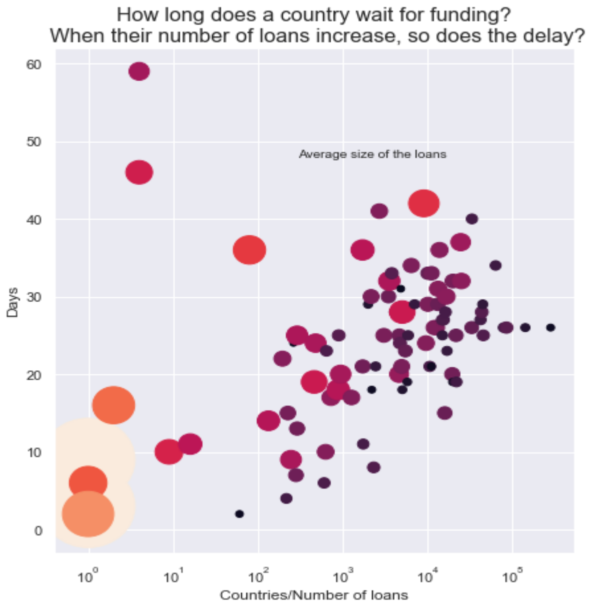
Do the countries have influence regarding delays?

Half of the loans go to 5 countries (Philippines, Kenya, Peru, Cambodia and El Salvador). When looking specifically at the 6 countries PDFs and CDFs, we clearly distinguish El Salvador takes 5% of the loan from other ones.



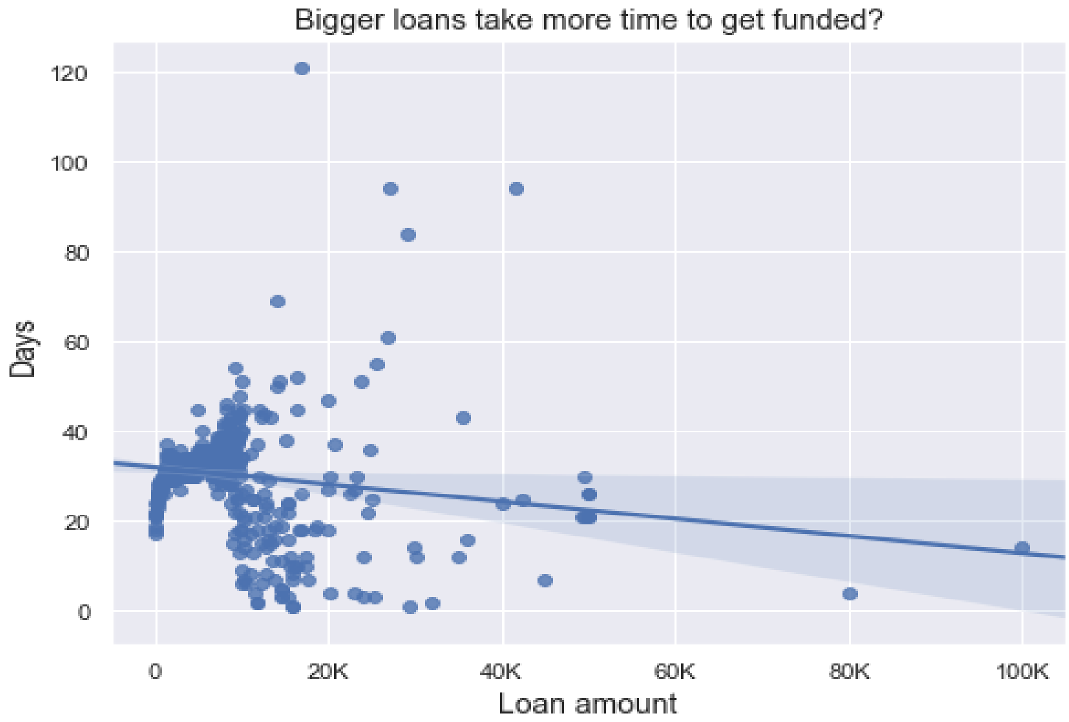
When a loan is funded 96% of the data, how long does it take to get that funding, on average, in each country? Does the number of loans have any impact?



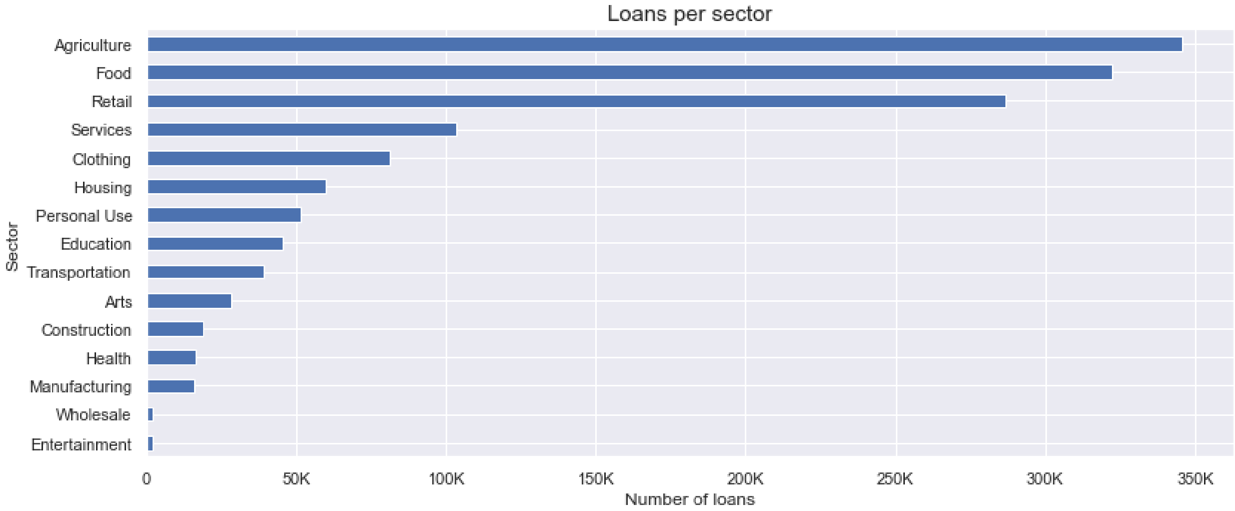


Do bigger amount of loans take more time to fund?

There is a significant negative correlation between the size of the loan and the time it takes to get funded. 99% of the data has loans inferior to 5000USD and in that case there is a significant positive correlation of 62%.



Do the sectors of the loan influences delays?

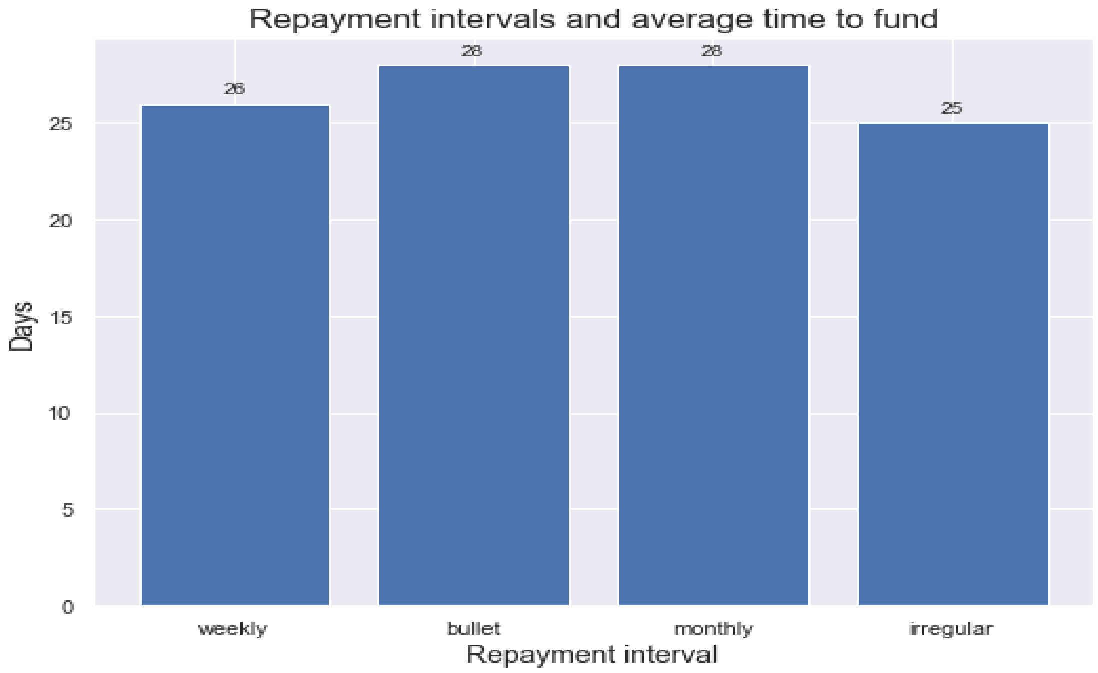


70% of the loans go to 3 sectors: Agriculture, Food and Retail. With further analysis regarding each sector main statistics, we can notice that the main 6 sectors are the ones who also influence the most a higher delay in funding's, where Agriculture has the biggest impact.

Does the repayment interval affect delays?

The main types of the repayment intervals are Monthly, Irregular and Bullet:

There are slightly different statistics for each type. Dividing the data into two groups one where the time to fund is below the median and the other above the median. Using a Chi-squared test, we conclude that there is a significant impact on the type of repayment interval. All in all, irregular repayments have a significantly better performance than monthly repayments, regarding delays.



Modelling

Initial Feature selection and extraction

As only few of the features has importance, selecting only certain columns of interest and excluded all other features where computation of real time of delay was not possible.

Features that were not considered are

‘Loan_id’, ‘funded_amount’, ‘status’, ‘activity_name’, ‘loan_use’, ‘country_code’, ‘town_name’, ‘posted_time’, ‘planned_time’, ‘planned_expiration_time’, ‘disbursed_time’, ‘funded_time’, ‘borrower_genders’

Countries: main 50 countries, which represent 97% of the data. I then turned the remaining countries into one variable named ‘country_name_other’

Partners: main 100 partners including (partner_id=1), which represents 85%of the data, and created ‘partner_id_other’ for any of the excluded partners.

Currency: main 25 currencies, which represent 90% of the data and create a new feature named ‘currency_other’for any of the excluded currencies.

Problem Statement:

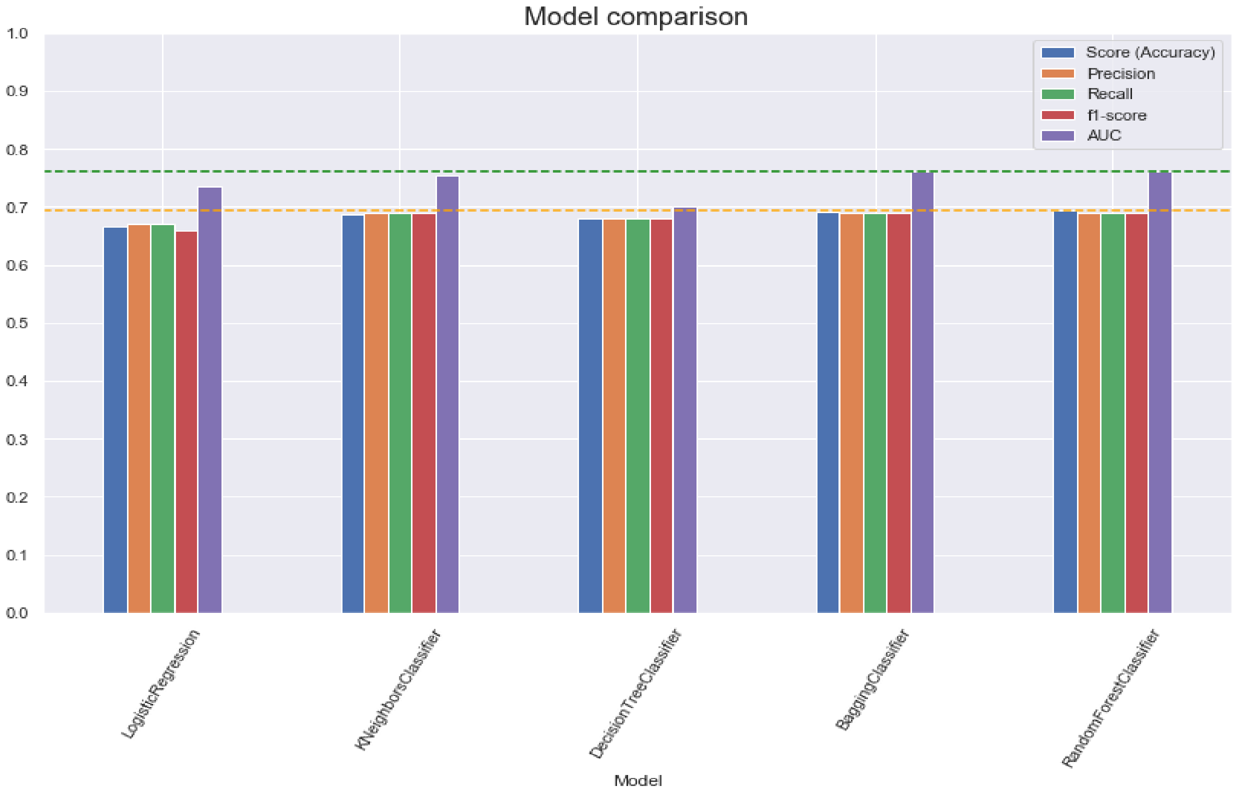
I considered it a classification problem. To classify a loan as either a “**high**” delay loan or “**low**” delay loan. “High” delay loan is a loan which takes more than median delay of all the loans in given data and “Low” delay loan is a loan which takes than median delay of the all the loans in given data. Median delay of the loans is 26 days.

- 1 – “high” loan delay
- 0 – “low” loan delay

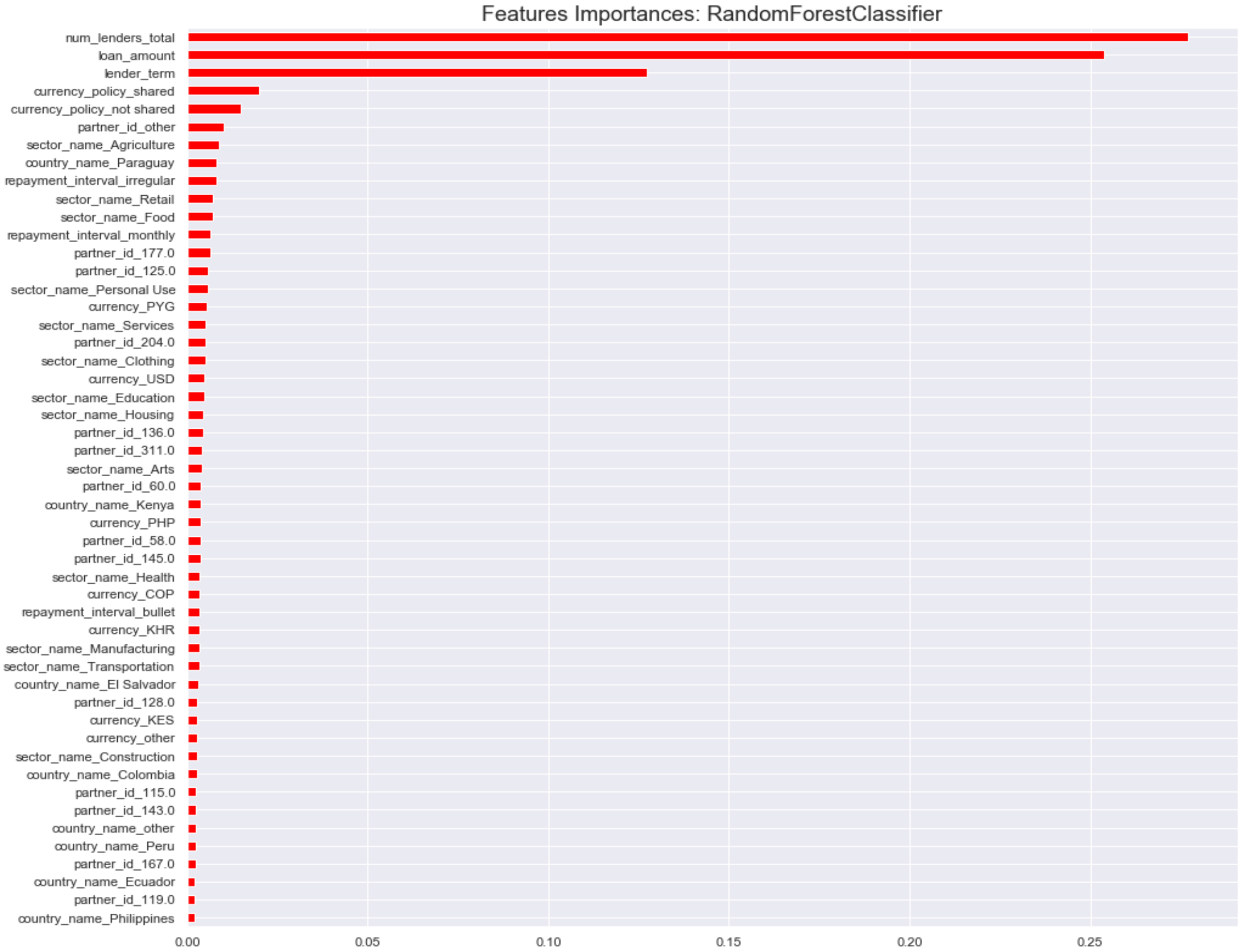
For linear model, took the correlated variables out and scaled the numeric ones. Replaced the categorical variables with dummy variables for both linear and non-linear models.

Models comparison Results:

Random Forest Classifier got the highest scores on all categories as 0.762.



Conclusion:



As seen total number of lenders (‘num_lenders_total’), loan amount (‘loan_amount’), lender term (‘lender_term’) and the currency policy (‘currency_policy_shared’) are the most deciding features if a loan will be delayed or not.