IDS 575
Business Analytic Statistics

SPRING 2018

FINAL PROJECT REPORT

Team Members:
DEVYANI SINGH - 654766729
TARUN KHURANA - 667683595
PRIYANKA SHARMA- 664374611

Introduction

Kiva.org is an online crowdfunding platform and not-for-profit organization which extends loans to financially excluded people, low income entrepreneurs and students around the world. It allows people to lend money at 0% interest rate to those who need financing for education, agriculture, retail, transportation, etc.

Kiva Crowdfunding has provided the data about their active loans on Kaggle. We have further incorporated publicly available data files on Multidimensional Poverty Index (MPI)-rural and urban, Human Development Index (HDI) for countries, Intensity of deprivation-rural and urban for analyzing demographic information and borrower welfare levels. Using this information of borrower welfare and deprivation levels is crucial in understanding the poverty level of each borrower, as it will help Kiva and the lenders in making informed decisions.

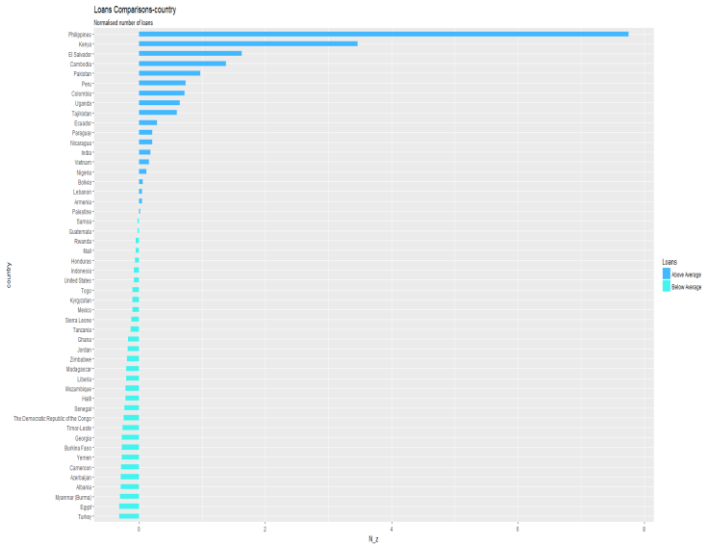
Objective

By using Kiva data files and connecting them with other relevant public data files, we aim to build localized models to estimate the loan amount which Kiva disburses to borrowers and the sectors in which these loans are disbursed. This will help Kiva to set investment priorities and get a better understanding of the target communities.

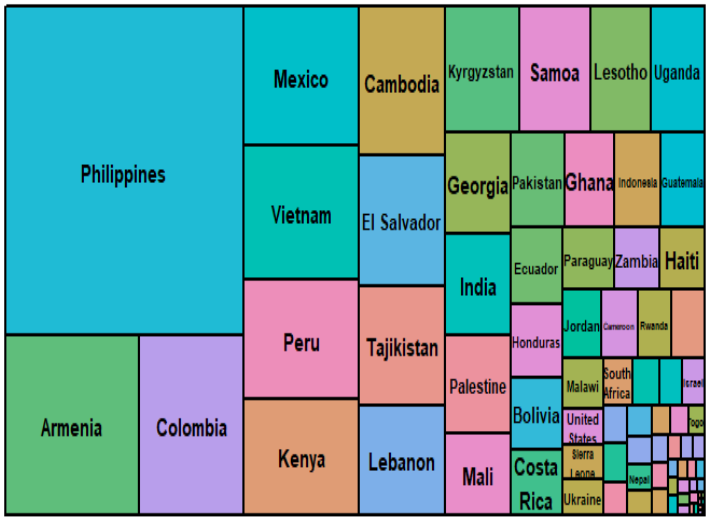
Exploratory Data Analysis

The univariate and bivariate analysis of the variables were done where relationships of important variables w.r.t the loan amount and sector were explored.

The below two plots show that Philippines has the highest average and count of loans

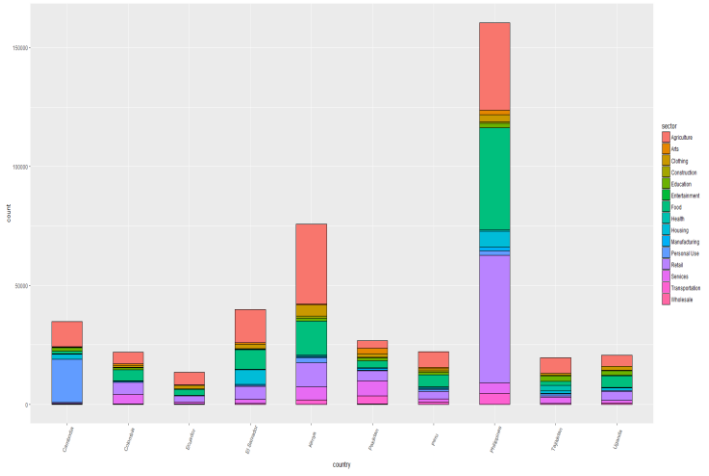


Loan amount per country contrasting with the world

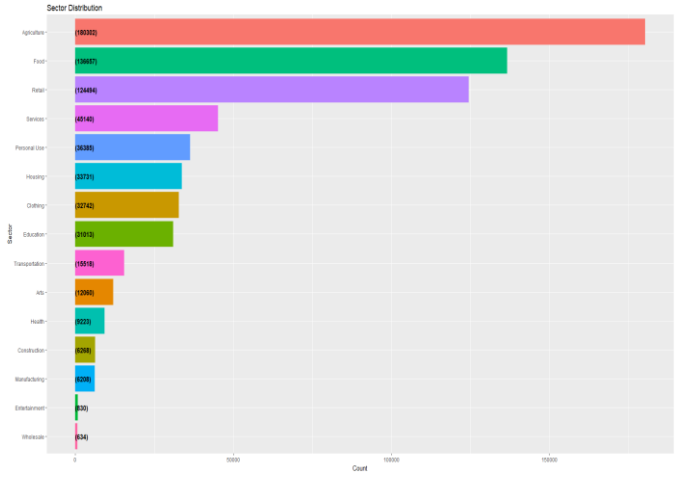


Countries with the highest frequencies of loans

Below are some of other important plots



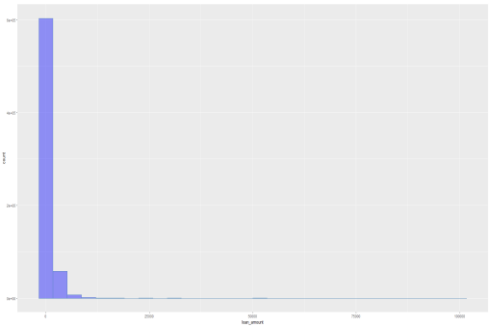
Loan amount distribution with each Sector per Country



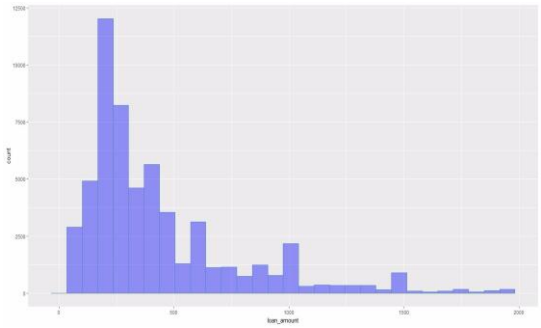
Count of loans in each sector

Data Preprocessing (Wrangling and Cleaning)

1. Merging the data files:
   All the data files were merged together based on an identifier.
2. Renaming the columns:
   The column names were renamed to an easy to use and understandable format.
3. Removal of NAs:
   About 8.2% of the total observations were missing vales and we decided to remove these NAs. Columns with missing values were - Loan Theme Type, HDI, Country Code, Region, MPI_urban.
4. Data Reduction:
   - From the data visualization, we understand that Philippines had the highest number of loans. Also, majority of the loans were from Asian continents. Hence, we scaled down the data and focused our analysis on the Asian continent.
   - Moreover, we aimed to reflect the most recent trends and patterns for our modeling process and hence limited our year to 2017.
5. Data Transformation
   - Dependent variable transformation:
     Loan Amount is the dependent variable which we aim to predict in our first Hypothesis. From its univariate analysis in the above section, we see that it is highly skewed to the right. This is because of the presence of outliers. We therefore removed the outliers above $ 2000 loan amount and we could get a near normal distribution.



Before Transformation



After Transformation

   - Independent variable transformation:
     From our analysis of other independent variables, we see that term in months (duration for which the loan was disbursed) is highly skewed. We removed the outliers for this variable and limited our term to 30 months.
6. Handling categorical variables:
   - Borrower Genders:
     A Kiva loan can be requested by an individual or a group. When the borrowers are part of a group, the genders of all the members of the group are captured. However, this information is not useful in determining any patterns. Hence, we combined the members' gender information to a new level named 'group'.
   - Loan Theme Type:
     We analyzed Loan Theme Types and first shortlisted those which had at least 100 loans in them. We then retained only those loans which were part of these top 100 Loan Theme Types.
7. Dropping unwanted columns:
   We dropped columns such as Partner ID, Country Code, Continent and Year as they contained redundant information.

Hypothesis 1

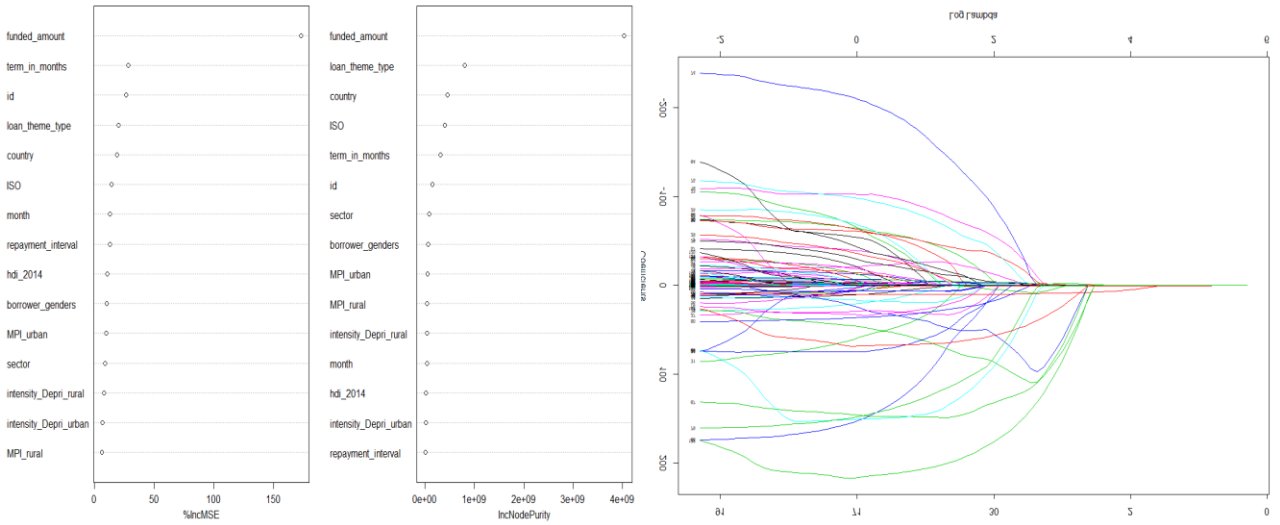
The MPI, regions, loan theme type, term in months, sector and the gender may play a direct role in predicting the loan amount.

Stage 1: Pre-modeling

Before modeling our Hypothesis 1, we first observed the variable importance via Random Forest and LASSO. Then we studied the correlation between loan amount and other predictor variables. Lastly, we releveled some categorical variables to set the best reference level.

These steps are detailed as below:

1. Variable Importance:
   - Random Forest model was built on all the predictor variables.
     Parameters Used: Number of trees = 200, Number of predictors = 6 Include plot/variables
   - LASSO was also used to find importance of variables.



2. Correlation Test:
   Insights:
   - Funded amount was highly correlated with loan amount. This is as per our expectation as there is not a significant difference between these two amounts.
   - Loan amount is positively correlated with Term in months.
   - Loan amount is negatively correlated with Intensity deprivation-rural and urban. This is also as per our expectation because poor and deprived people may require higher loan amounts.
3. Releveling of categorical variables:
   - The reference category of the Loan Theme Type was releveled to the General category. This is because General had higher number of loans in its category.
   - The reference category of Month was also releveled to the 5th.
4. Data splitting: Data was split in the ratio of 70:30 for train and test, respectively.

Stage 2: Modeling

Model 1: Linear Regression

- Loan amount was modeled linearly with the most important variable (term in months).
- Funded amount was not included in building the model as it was highly correlated to loan amount.

Model Performance	Analysis & Insights	What can be improved
Adjusted R-squared = 25.7 %	• Removing the funded amount worked because otherwise it would have led to multicollinearity. • Term in month accounts for approximately 26% of the variation in loan amount.	The prediction accuracy can surely be improved as this was the most basic model with only the most important predictor variable.

Model 2: Multiple Linear Regression

- Loan amount was modeled linearly with all the important variables obtained from Random Forest and LASSO.
- MPI was not included in building the model as it was highly correlated with Intensity of Deprivation.
- Square root transformation was applied on Loan amount and intensity of deprivation to improve their distribution.

Model Performance	Analysis & Insights	What can be improved
Adjusted R-squared = 58 %	• Since MPI and intensity of deprivation were correlated, removing MPI worked because otherwise it would have led to multicollinearity. • This improved model explains approximately 58% of the variation in loan amount.	• The model can be further optimized by implementing regularization techniques. • They can penalize unimportant variables which don't play a role in improving accuracy.

Model 3: Ridge Regression

- Through cross validation, the most optimal value of penalty factor was found.
- This penalty factor is used to introduce the shrinkage penalty to the above model.

Model Performance	Analysis & Insights	What can be improved
Adjusted R-squared = 59.3 %	• RMSE on train and test set reduced considerably. • This regularized model using ridge explains approximately 59% of the variation in loan amount.	The model can be further optimized by pushing some coefficients to zero (that is, by using LASSO regularization technique).

Model 4: LASSO Regression

- Through cross validation, the most optimal value of penalty factor was found.
- This penalty factor is used to introduce the shrinkage penalty to the model.

Model Performance	Analysis & Insights	What can be improved
Adjusted R-squared = 59.9 %	• RMSE on train and test set reduced further. Model is optimized. • This improved model explains approximately 60% of the variation in loan amount.	Non-linear regression techniques can be explored further to improve the prediction of loan amount.

Conclusion for Hypothesis 1

We found the LASSO regression technique to be the ***best tailored statistical model*** giving the best prediction for loan amount. This model is optimized as we used cross validation for finding the penalty factor. There is scope for further improvement of the prediction by using other techniques like non-linear regression and boosting methods.

Hypothesis 2

The loan amount, repayment interval, region, loan theme type, term in months and the MPI may play a direct role in predicting the sectors in which the loans were issued.

Stage 1: Pre-modeling

1. Variable Importance: Before modeling our Hypothesis 2, we first observed the variable importance via Random Forest. It was built on all the predictor variables. Parameters Used: Number of trees = 500, Number of predictors = 6
2. Data splitting: Data was split in the ratio of 70:30 for train and test, respectively.

Stage 2: Modeling

Model 1: Decision Tree

- Decision tree was built for predicting the sectors in which the loans were issued.
- Since loan amount and funded amount are correlated, taking only of them makes sense as they give redundant information. Similarly, MPI and intensity of deprivation are correlated, and we included only intensity of deprivation.

Model Performance	Analysis & Insights	What can be improved
Adjusted R-squared = 48.9 %	• Coming up with optimal values for building Decision Trees is a cumbersome process as it includes multiple trials. • For growing the full tree, Decision Tree computationally takes a lot of time depending on the parameters we choose.	• Complexity parameter, minimum split, minimum bucket and other such parameters can be optimally chosen to improve accuracy. • Ensemble Methods can be implemented to further increase prediction performance.

Model 2: Ensemble Methods (Bagging & Random Forest)

- We built models using Bagging and Random Forest for predicting the sectors.
- Random Forest, in addition to finding the variable importance, proves a better multi-class classifier than Decision Trees and Bagging.

Model Performance	Analysis & Insights	What can be improved
Adjusted R-squared (Bagging) = 51.7 % Adjusted R-squared (Random Forest) = 54.04 %	• Ensemble methods are computationally possible to implement and gives improved accuracy over Decision Trees. • Though better than Decision Trees, Ensemble methods didn't give drastic improvement of prediction accuracy as we had expected.	• Number of trees can be increased to further check the performance of Bagging and Random Forest. • We can explore other classification techniques like Multinomial Logistic Regression to improve the prediction accuracy.

Model 3: Multinomial Logistic Regression

- We built a multinomial logistic regression model to classify the sector variable with multiple levels.

Model Performance	Analysis & Insights	What can be improved
Adjusted R-squared = 52.16 %	• Multinomial Logistic Regression model was easier to build and interpret.	• Prediction performance is still not improved as compared to Random Forest.

Model 4: Naïve Bayes

- We built a preliminary Naïve Bayes model by sampling 70% of each level in the sector variable. This model didn't work well. We then used random sampling and built the model again.

Model Performance	Analysis & Insights	What can be improved
Adjusted R-squared (preliminary model) = 35 % Adjusted R-squared (updated model) = 96.12 %	• Sampling 70% of each level in the sector variable gave low prediction because the size of each stratum obtained after sampling was not proportionate to the size in the original data. • Random sampling gave us better performance accuracy.	• Proportional stratified sampling can be implemented to further improve performance. • Model validation techniques can be applied for assessing this classifier.

Conclusion for Hypothesis 2

We found the Naïve Bayes classifier as the ***best tailored statistical model*** which gives the best prediction for sector. However, there is scope for further improvement of this model by using cross validation, which helps in validating the model effectiveness properly. Sampling technique of proportional stratified sampling can also be explored.

Future Scope/Work

Our work is restricted to the prediction of loan amount and sector only. However, this study can be extended to the prediction of several other variables. One such variable is term in months which explains the duration for which loan amount was disbursed. Another such variable is repayment interval. This will empower Kiva to understand the repayment terms (monthly, irregular, etc.) of a borrower. Moreover, text mining can be performed on the tags variable to understand other crucial borrower characteristics.

References

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