Lending Environment Simulator & Lender Evaluation Tool

by Vadim Spirkov, Murlidhar Loka

Abstract In

Background

Objectives

GitHub

Data Analysis

As it has been mentioned KASI Insight does not posess personal financial data that could be used to identify and predict client credit worthiness. In this project we are using the survey data collected by KASI Insight over years from seven African countries. The survey targets people who lend money on regular basis. The survey contains questions pertaining to the borrowing habits of the people the lenders deal with, asks respondents what they feel about the economy in a given country, etc. KASI insight has collected almost **30,000** records over a course of last three years.

Data Dictionary

The survey comprises 38 columns. Majority of them are multi-choice questions. The table below lists the survey columns

ID	Question/Column
0	Timestamp
1	Location ID
2	Has it become more difficult or easier to find a job in your city?
3	Is this a good time for people to make a large purchase such as furniture or electrical appliances, given the economic climate?
4	Compared to the last 6 months, are you able to spend (more, the same or less) money on large purchases over the next 6 months?
5	Will you be able to meet your regular expenses over the next 6 months?
6	How do you expect your household's income to change over the next 6 months?
7	How do you expect general economic conditions in your city to change over the next 6 months?
8	How do you expect general economic conditions in your country to change over the next 6 months?
9	Gender
10	Marital status
11	Age
12	What's your highest level of education?
13	Occupation
14	If you are a student, what level are you currently studying?
15	Race/Ethnicity
16	Country
17	What is the name of the neighborhood where you live?
18	Over the past 3 months, how many times have you lent someone money?
19	On average how much do you lend in general?
20	Who did you lend money to in the past 3 months?
21	When you lend money, when do you usually expect to get it repaid?
22	Do you include either interest or a lending fee when you lend?
23	Do you request guarantees when you lend?
24	Do you receive your money back in time?
25	Assuming that you have lent money at least ten times, how often would you get your money repaid?

ID	Question/Column
26	What's the most common use of the money you lend?
27	Have you ever applied for a bank loan?
28	Are you a tontine / lending club member?
29	What is the most convenient way to get a loan?
30	To what extent do you agree with the following sentences [Access to credit is essential for me to achieve financial freedom]
31	To what extent do you agree with the following sentences [Credit is beneficial only if you have discipline]
32	To what extent do you agree with the following sentences [I would like to have more credit management training]
33	What type of loans are you currently paying of?
34	Do you have a credit score?
35	Do you have a credit card?
36	On average, how much of your total household monthly income do you spend paying off debt each month?
37	If you wanted to take a loan to start a business, how much would you need?

The survey data could be split into three major categories:

- Demographiche Statistics
- Economic Sentiment
- Spending and Borrowing habits

Each question/column should be trated as a catigorical value.

Data Exploration

Let's take a look at the raw survey data. Though the survey multi-choice questions are catergories in nature, the survey answers are stored in alphanumeric format. Overtime some questions have been rephrased. Thus in some cases the answers that pertain to the same category vary. Another problem with the raw data set is the missing values.

TO DO: insert the sample of raw data here

To ractify the problems stated above we have developed a data processing algorithm that normalized and categorised the answers converting them into numeric form. The data processing script also imputes the missing data with the most frequently occuring value for a given category. The clean data set stats are depicted below. **Note**: the column numbers corespond to the question numbers as described in *Data Dictionary* paragraph.

```
#>              2         3         4         5         6         7         8         9        10  \
#> count 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00
#> mean       1.69      2.07      1.90      2.16      1.80      1.82      1.83      1.66      1.99
#> std        0.62      0.61      0.64      0.60      0.64      0.63      0.64      0.48      0.87
#> min        1.00      1.00      1.00      1.00      1.00      1.00      1.00      1.00      1.00
#> 25%        1.00      2.00      1.00      2.00      1.00      1.00      1.00      1.00      1.00
#> 50%        2.00      2.00      2.00      2.00      2.00      2.00      2.00      2.00      2.00
#> 75%        2.00      2.00      2.00      3.00      2.00      2.00      2.00      2.00      2.00
#> max        3.00      3.00      3.00      3.00      3.00      3.00      3.00      2.00      7.00
#>
#>             11        12        13        14        16        18        19        20        21  \
#> count 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00
#> mean       3.68      3.85      2.77      0.24      3.59      2.41      2.07      1.93      2.39
#> std        1.17      1.27      1.20      0.71      1.93      0.90      0.75      0.61      0.97
#> min        1.00      1.00      1.00      0.00      1.00      1.00      1.00      1.00      1.00
#> 25%        3.00      3.00      2.00      0.00      2.00      2.00      2.00      2.00      2.00
#> 50%        4.00      4.00      3.00      0.00      4.00      2.00      2.00      2.00      2.00
#> 75%        4.00      5.00      4.00      0.00      5.00      3.00      2.00      2.00      3.00
#> max        9.00     10.00      9.00      4.00      7.00      4.00      4.00      3.00      6.00
#>
#>             22        23        24        25        26        27        28        29        30  \
#> count 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00
#> mean       3.26      3.28      2.80      2.71      2.61      1.47      1.51      2.66      2.71
```

```
#> std        1.32      1.32      1.16      1.10      1.38      0.50      0.50      1.06      1.10
#> min        1.00      1.00      1.00      1.00      1.00      1.00      1.00      1.00      1.00
#> 25%        2.00      2.00      2.00      2.00      1.00      1.00      1.00      2.00      2.00
#> 50%        3.00      3.00      3.00      3.00      2.00      1.00      2.00      2.00      2.00
#> 75%        5.00      5.00      4.00      3.00      3.00      2.00      2.00      3.00      3.00
#> max        5.00      5.00      5.00      5.00      5.00      2.00      2.00      6.00      5.00
#>
#>              31         32         33         34         35         36         37  credit_score  \
#> count 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00 29,383.00     29,383.00
#> mean       2.84      2.90      0.30      1.25      1.30      7.68      2.19        348.56
#> std        1.13      1.22      1.08      0.44      0.46      1.26      0.58        244.70
#> min        1.00      1.00      0.00      1.00      1.00      1.00      1.00       -720.00
#> 25%        2.00      2.00      0.00      1.00      1.00      8.00      2.00        190.00
#> 50%        2.00      2.00      0.00      1.00      1.00      8.00      2.00        390.00
#> 75%        4.00      4.00      0.00      2.00      2.00      8.00      2.00        540.00
#> max        5.00      5.00      8.00      2.00      2.00      8.00      4.00        970.00
#>
#>        credit_score_category  lender_score  lender_score_category
#> count              29,383.00     29,383.00              29,383.00
#> mean                    1.84        303.80                   1.74
#> std                     0.95        272.73                   0.82
#> min                     1.00       -590.00                   1.00
#> 25%                     1.00        140.00                   1.00
#> 50%                     2.00        380.00                   2.00
#> 75%                     2.00        500.00                   2.00
#> max                     5.00        840.00                   5.00
```

Evidently now all the data is categorized, the missing values imputed. From this point on we will be using the clean data set to do further data exploration, feature engineering and model training.

Demographic Stats

It is useful to understand who took the survey. This knowledge will ultimately give us the answers about the money market participants in Africa.

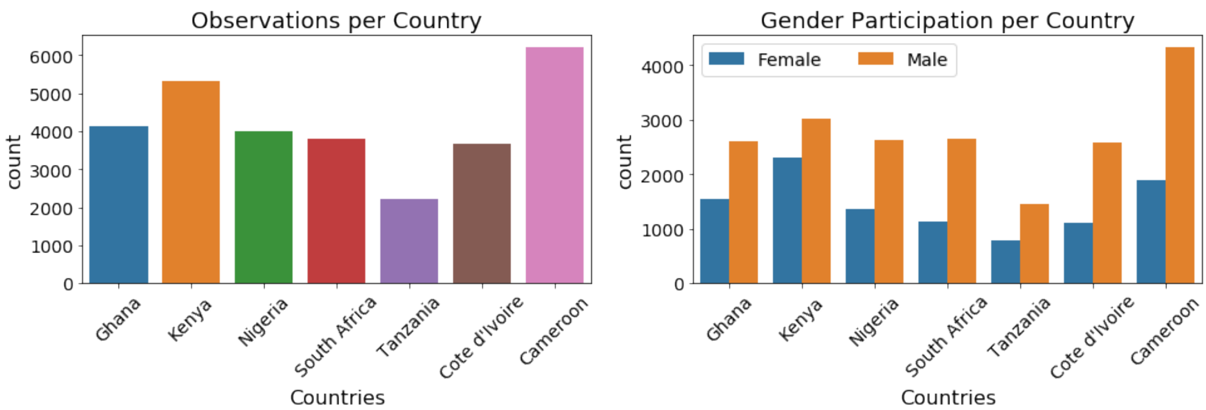

Figure 1: Participation per Country

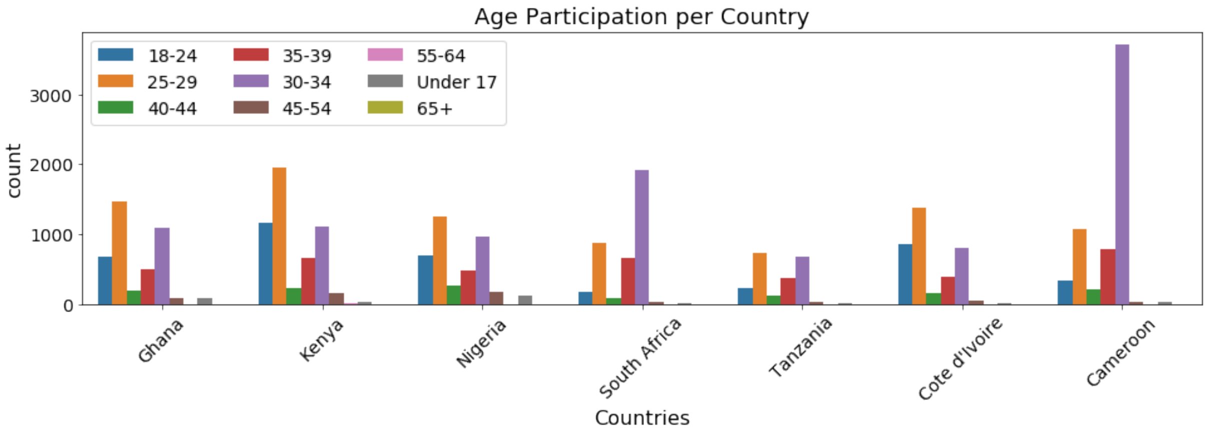

Figure 2: Age of Participans

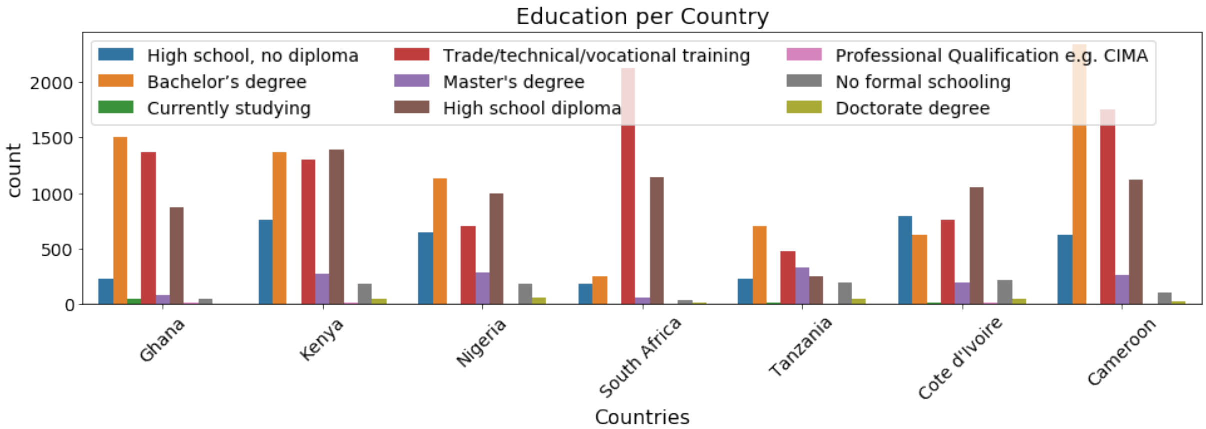

Figure 3: Education of Participants

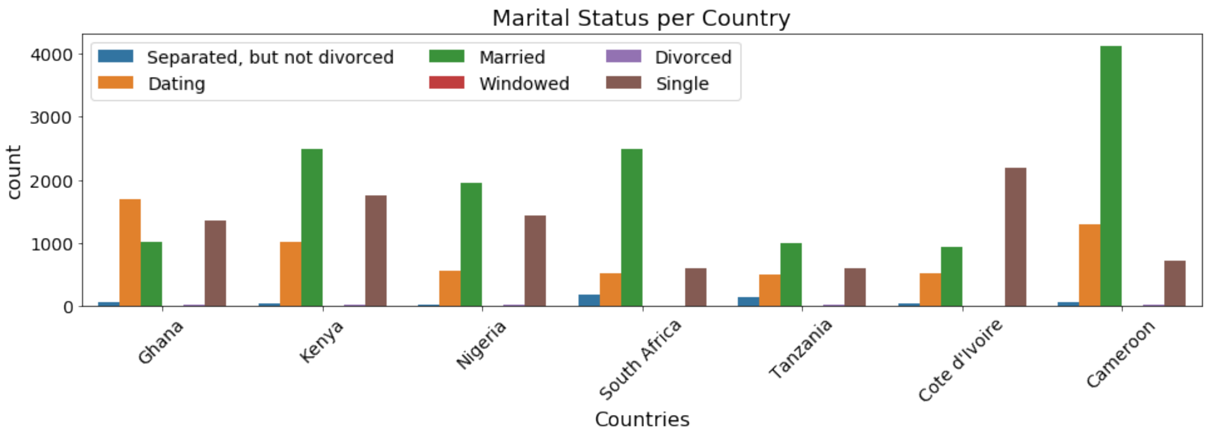

Figure 4: Marital Status of Participants

As per the charts submitted above we can canlude that:

- Camerun has the highest number of obervations and Tansania ihas the smallets representation, where the rest of the countries or more or less equaly represented.
- Males dominate in the money lending business. Kenya though makes an exception where number of femail participants is very close to the male population
- In general people in *30-34* age group are the most active, followed by *25-29* and *18-24* age groups respectively. In Kenya, unlike other countries, the yonger generation is more active.

- Majority of money lenders are either salaried or comission-based employees. Again Kenya makes an exption. The second largest group of the money lenders is the busines owners.
- Education-wise people with the bachelor's degree and skilled trade workers dominate.
- Married people tend to lend money more often. . .

Economic Sentiment

Now let's see what the money lenders think about the state of the economy in ther respective countries. The questions where asked in the six-month perspective in the future from the date of survey.

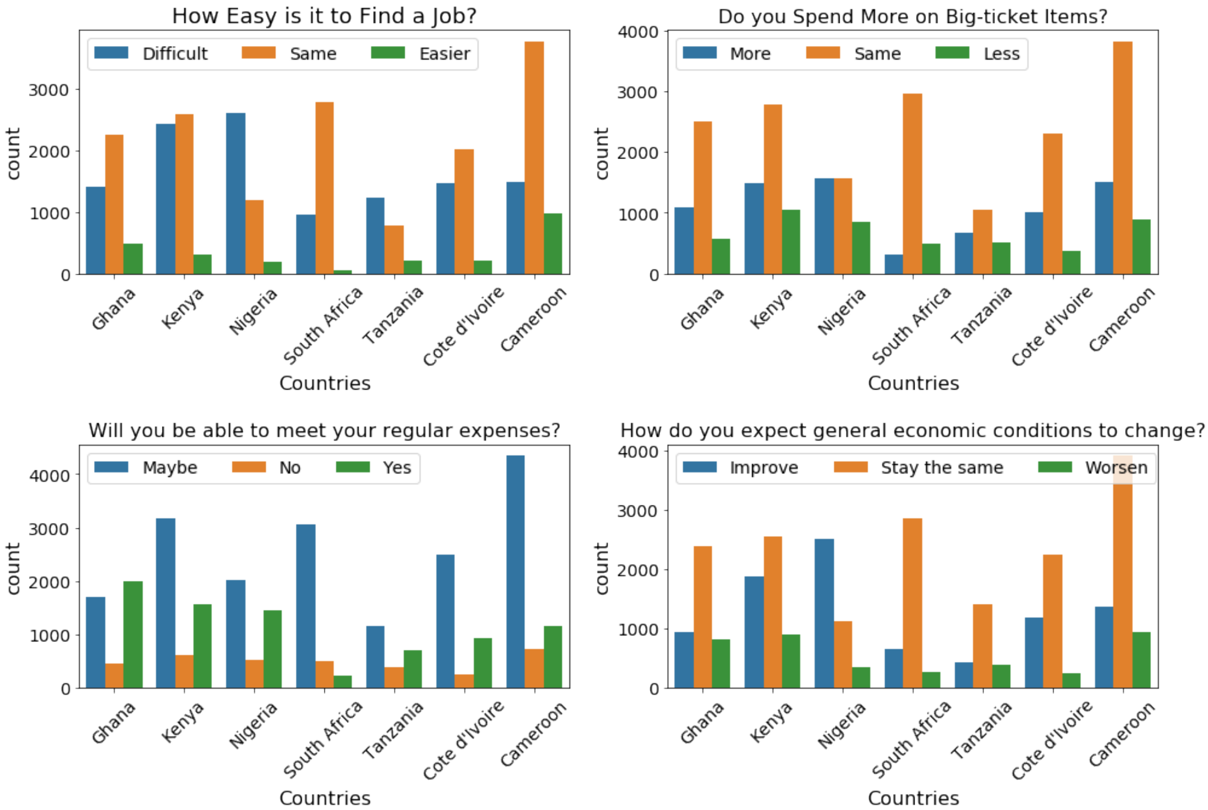

Figure 5: Economic Sentiment

Evidently majority of the survey participants think that the economic situation in their country will be stable over a course of next six months. Many people in Kenya, Nigeria and Ghana find it more diffiult to find a job. Remarkably, despite the fact that people believe that the economic conditions are stable, citisens of all counties are not shure if they are going to meet their reqular expences.

Spending and Borrowing Habits

Spending and borrowing habits is the segment of our particular interest since it affects the most the credit score of the population.

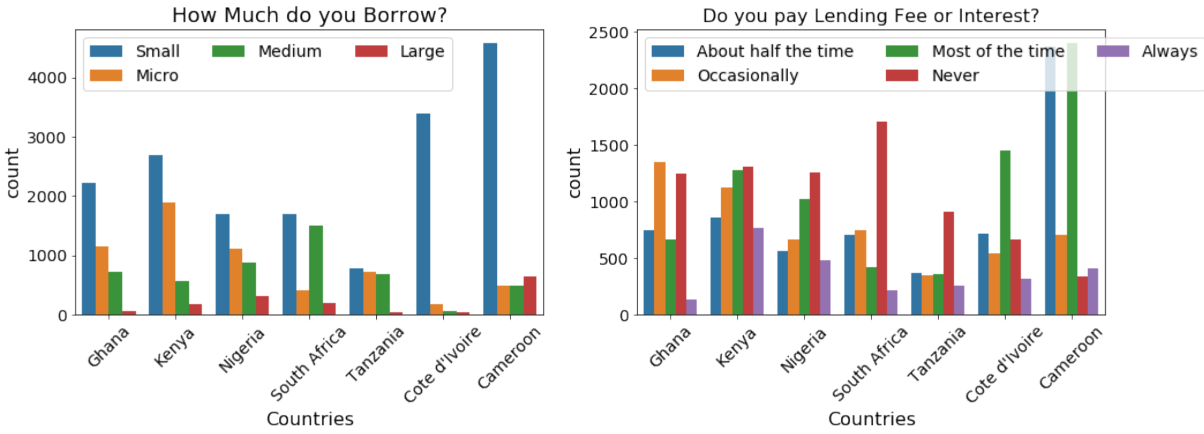

Figure 6: Borrowing Habits

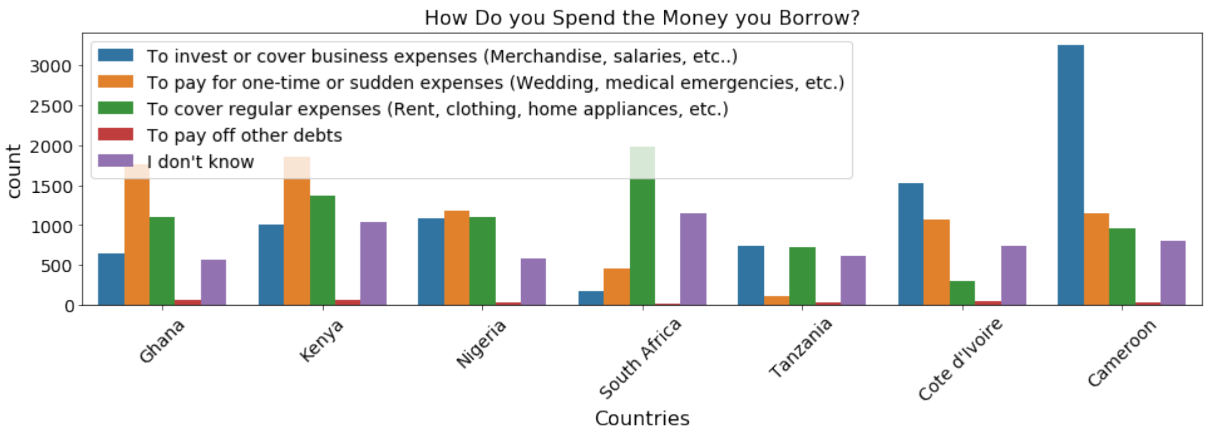

Figure 7: Spending Habits

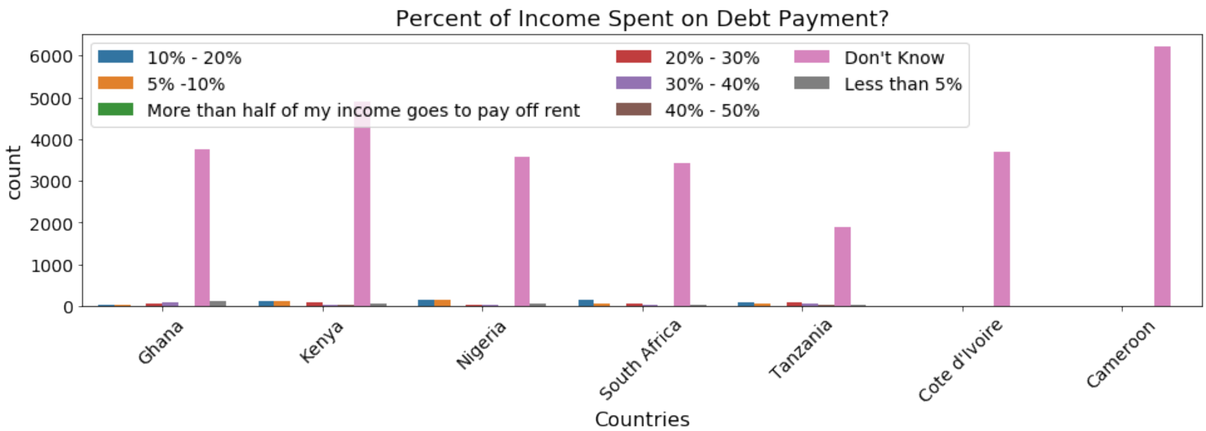

Figure 8: Debt Payment

- Majority of population take either small or micro loans (the exact amounts are country specific).
- It is quite remarkable that the lenders do not charge fees or interest regularly (if at all) more often than not. The Cameroonians make an exception. In opposite the majority of South African lenders never change interest. We have conducted furhter data research that have proved that many people tend to lend to friends and famitly. This fact explains why the fees and interest on loans are waived.
- People in Cameroon, Cote d'Ivoire and Tanzania spend the loans to cover business-related expenses. Citizens of other countries mainly use loan to either cover one-time or or unexpected expences (wedding, medical emergency..) or make ends meet (pay rent, buy clothes, etc.)

- Interestingly people in all countries do not watch how they spend the borrowed money. This fact probably explains why the question *Will you be able to meet your reqular expenses?* genrates uncertain answers (see **Economic Sentiment** paragraph for futher details).

Data Distribution between Categories

There are five credit categories for borrowers and five lender categories. To train the robust classification models we have to ensure that each category has enough observations to support the model traing. Let's review the data distribution between the borrower and lender categories.

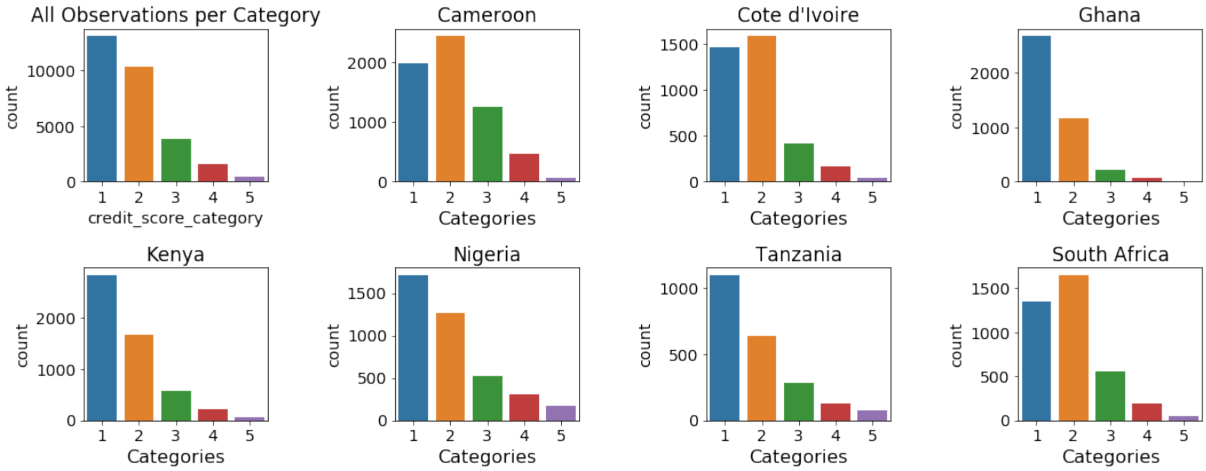

Figure 9: Data Distribution per Credit Categories

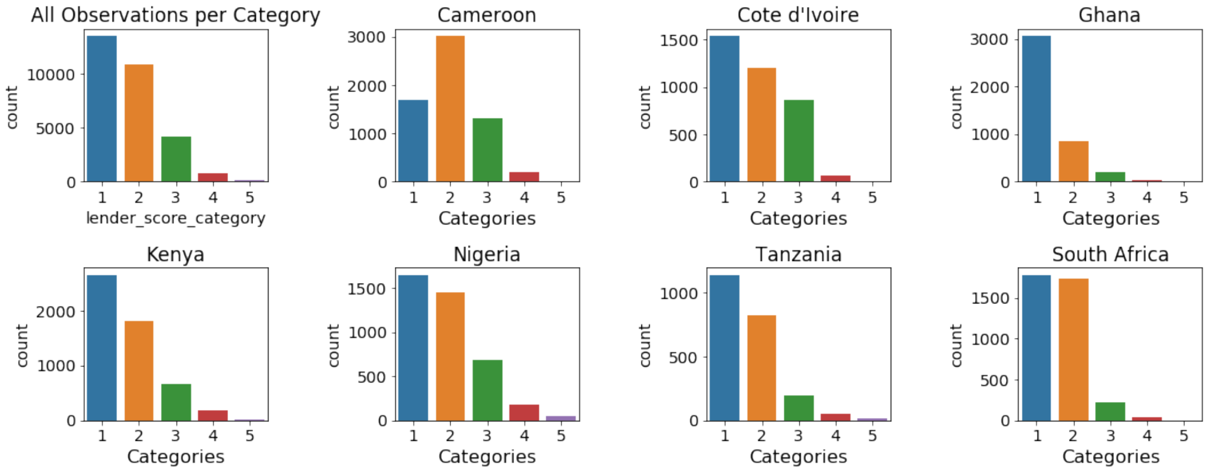

Figure 10: Data Distribution per Lender Categories

As we can observe overall the lending environmnet is not very promising; categories 1 and 2 (*Very Poor* and *Poor*) dominate. The lending climate is visibly better in Cameroon, Cote d'Ivoire and South Africa. It is also worth mentioning that categories 4 and 5 (*Good* and *Very Good*) do not have that much data. The situation is even worse with the lender categories. Thus prior to the model training we would have to upsmple the training data sets to bring all categories to the same level.

Overall looking at the credit and lender scores of the population we observe that the destribution pattern is very similar between all sevent African countries. Thus if KASI Insight adds more countries to the fold there is no need to retrain the models assuming that the newly added countries have the same category distribution...

Feature Selection and Engineering

The data set has 38 columns. We potentially, could employ all of them to fit the models. But this is not the optimal approach. Not all data elements contribute to the category identification equally, some may not contribute at all, so why keep them? Another consideration is that the large and wide data sets make model training much longer, affect the accuracy and speed of the models negatively. Also many input variables add complexity to the user interface making it hard to implement, maintain and use. Thus we have opted to evaluated available data features. The ultimate goal is to undersatand the relationship between the features and the response variables and select the most infulential ones.

Feature Correlation Matrix

Strongly correlated features are redundant thus they could be dropped without impacting the model performance. Figure 11 depicts a correlation heatmap of all 38 data set features. The correlated features would be rendered either in deep black or very light colors. As we can observe none of the features have strong correlation.

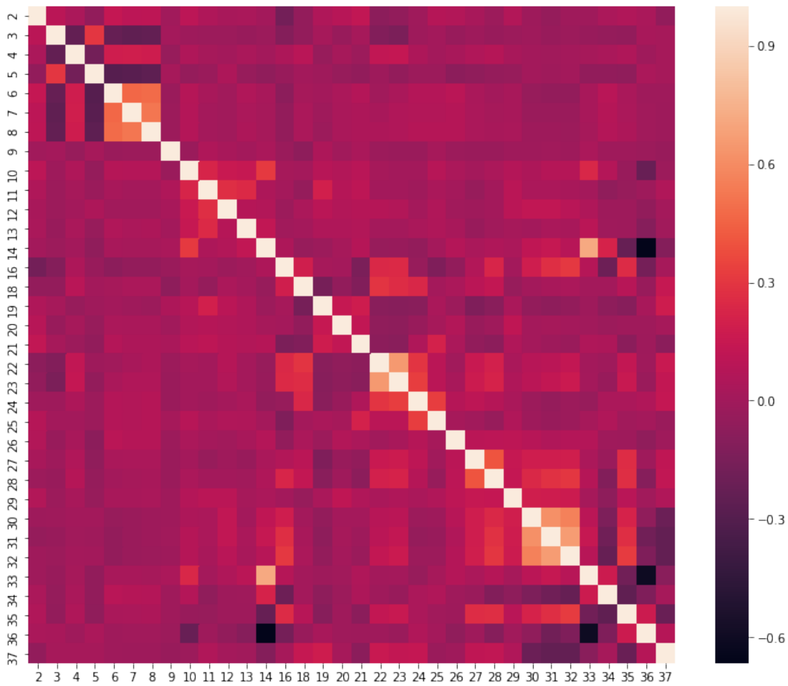

Figure 11: Feature Correlation

Univariate Feature Selection

Univariate feature selection examines each feature individually to determine the strength of the relationship of the feature with the response variable. Next two paragraphs examine relationship between top 20 features and credint and lender categories respectively.

Credit Score Univariate Feature Selection

Num	Feature	Score
24	Do you receive your money back in time?	4749.1
18	Over the past 3 months, how many times have you lent someone money?	1077.69
26	What's the most common use of the money you lend?	727.16
22	Do you include either interest or a lending fee when you lend?	536.27
19	On average how much do you lend in general?	512.08
20	Who did you lend money to in the past 3 months?	441.73
33	What type of loans are you currently paying of?	360.28
23	Do you request guarantees when you lend?	351.02

Num	Feature	Score
14	If you are a student, what level are you currently studying?	301.56
21	When you lend money, when do you usually expect to get it repaid?	218.30
16	Country	197.034
25	Assuming that you have lent money at least ten times, how often would you get your money repaid?	180.84
11	Age	103.30
12	What's your highest level of education?	75.24
29	What is the most convenient way to get a loan?	63.37
2	Has it become more difficult or easier to find a job in your city?	55.15
10	Marital status	39.54
31	To what extent do you agree with the following sentences [Credit is beneficial only if you h…	35.68
30	To what extent do you agree with the following sentences [Access to credit is essential for …	31.48
32	To what extent do you agree with the following sentences [I would like to have more credit m…	25.52

Lender Score Univariate Feature Selection

Num	Feature	Score
24	Do you receive your money back in time?	6667.83
22	Do you include either interest or a lending fee when you lend?	3588.47
23	Do you request guarantees when you lend?	3339.37
18	Over the past 3 months, how many times have you lent someone money?	2014.33
16	Country	595.87
25	Assuming that you have lent money at least ten times, how often would you get your money repaid?	555.59
19	On average how much do you lend in general?	460.15
26	What's the most common use of the money you lend?	444.85
20	Who did you lend money to in the past 3 months?	415.17
14	If you are a student, what level are you currently studying?	174.17
21	When you lend money, when do you usually expect to get it repaid?	112.94
37	If you wanted to take a loan to start a business, how much would you need?	91.9
28	Are you a tontine / lending club member?	78.77
27	Have you ever applied for a bank loan?	76.69
4	Compared to the last 6 months, are you able to spend (more, the same or less) money on large pur…	56.49
11	Age	44.33
8	How do you expect general economic conditions in your country to change over the next 6 months?	42.94
32	To what extent do you agree with the following sentences [I would like to have more credit manag…	39.22
2	Has it become more difficult or easier to find a job in your city?	38.57
35	Do you have a credit card?	37.26

Feature Importance

We measure the importance of a feature by calculating the increase in the model's prediction error after permuting the feature. A feature is "important" if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction. A feature is "unimportant" if shuffling its values leaves the model error unchanged, because in this case the model ignored the feature for the prediction.

Credit Score Feature Impoirtance Evaluation

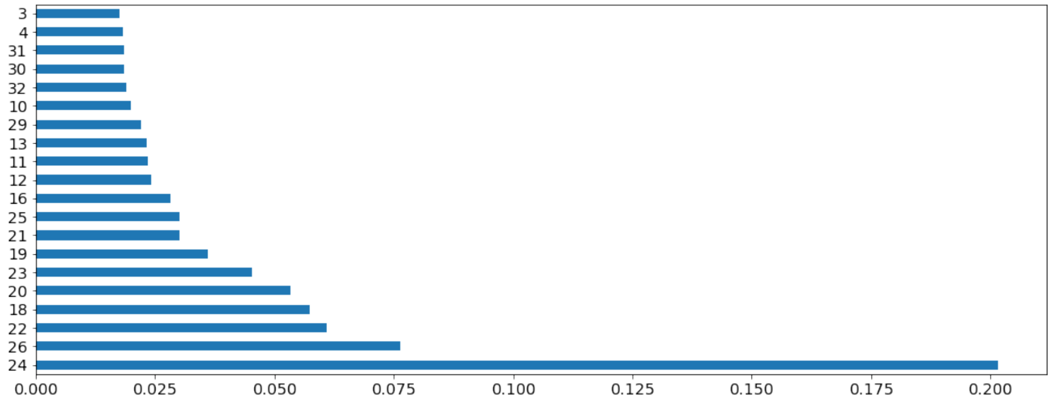

Figure 12: Credit Score Feature Impoirtance

Lender Score Feature Impoirtance Evaluation

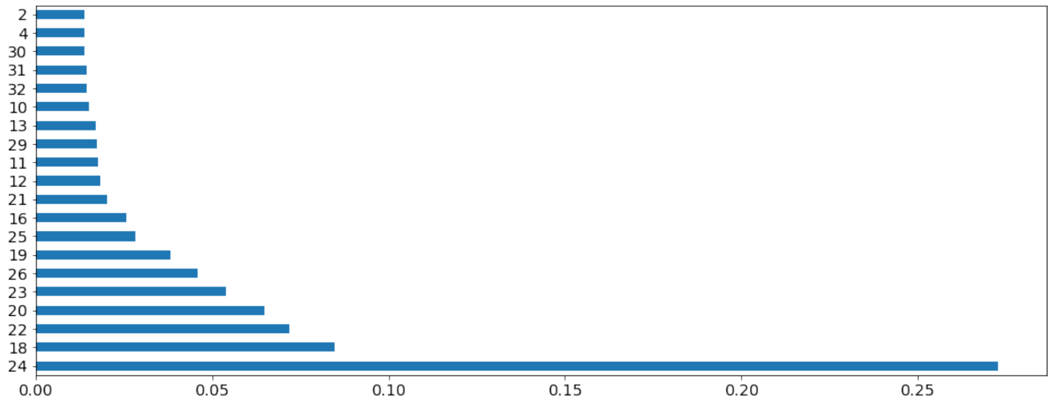

Figure 13: Lender Score Feature Impoirtance

Takeways

We have applied two mathematical algorithms to identify the most significant features for credit score and lender score labels. To no surprise both methods have sucesfully identified the feature that have been used to cacluclate the credit/ lender categories. We have decided to select the top features that have distinctively higher score as a base-line. During the model training and evaluation phase we will increase/decrease the number of features to estimate the effect of the input data dimentionality change the model accuracy.

Top Seven Credit Score Features

Num	Feature
24	Do you receive your money back in time?
26	What's the most common use of the money you lend?
22	Do you include either interest or a lending fee when you lend?
18	Over the past 3 months, how many times have you lent someone money?
20	Who did you lend money to in the past 3 months?
23	Do you request guarantees when you lend?
19	On average how much do you lend in general?

Top Nine Lender Score Features

Num	Feature
24	Do you receive your money back in time?
18	Over the past 3 months, how many times have you lent someone money?
22	Do you include either interest or a lending fee when you lend?
23	Do you request guarantees when you lend?
20	Who did you lend money to in the past 3 months?
26	What's the most common use of the money you lend?
19	On average how much do you lend in general?
25	Assuming that you have lent money at least ten times, how often would you get your money repaid?
16	Country

Model Evaluation and Selection

After we cleaned and normalized the data, labled all observations and gained deep understanding about the features we are ready to start model training and evaluation. To gain the best result possible we will explore and evaluate three algorthim to train the models. They are:

- **Support Vector Machine** (SVM). The greatest strength of SVM is that it has multiple Kernel implementations, that can be tuned to explain multi-dimensional space with high accuracy
- **Random Forest** (RF). Random forest belongs to the class of ensemble models. It has many hyperparameters that could be tuned to achieve high accuracy. The random forest algorithm is not demanding in terms of the data preparation, which makes it the first choice in many real-life scenarios
- **Gradient Boosting Machine** (GBM). GBM is an ensemble model as well. It uses the concept of trees just like the RF model does but applies it differently. GBT builds the trees one at a time, where each new tree helps to correct errors made by previously trained tree. the GBM.

Evaluation Metrics

We believe that the best model has to classify all five categories as accurate as possible. The winning model also would have to identify true positives and true negatives for each category equally well. Thus we choose the muticlass confusion matrix and F1 scores to evalue the models. The higher the F1 score for each category - the better the model performes.

We also take into consideration the model training and inference speed.

Model Training and Evaluation Methodology

- We begin with the splitting the available data into the training (70%) and test (30%) sets.
- We upsample the training data set employing *SMOTE* algorithm.
- We evaluate the three algorthims we have described above. We will be using the deafult algorithm parameters and top features (see *Feature Evaluation* paragraph for more details) to fit the models.
- We select the algorithm that has the best evaluation metrcis.
- Then we evaluate the winning algorithm fitting it with the smaller and larger feature sets.
- If the data dimentionality change makes positive impact on the the winning algorithm we select this feature set for the model.
- We hyper-tune the algorithm parameters in effort to achieve even better model performance

Lending Environment Simulator Model

Following the steps outlined in the previous section we have recieved the following performances stats.

SVM

```
           precision    recall  f1-score   support
     1       0.97      0.95      0.96      3935
```

```
       2       0.92      0.90      0.91      3114
       3       0.81      0.89      0.85      1122
       4       0.81      0.85      0.83       487
       5       0.81      0.89      0.85       157

micro avg      0.92      0.92      0.92      8815
macro avg      0.86      0.90      0.88      8815
weighted avg   0.92      0.92      0.92      8815

Overall algorithm accuracy: 0.9199
```

Random Forest

```
             precision    recall  f1-score   support
       1       0.98      0.96      0.97      3902
       2       0.93      0.93      0.93      3161
       3       0.86      0.90      0.88      1150
       4       0.86      0.84      0.85       464
       5       0.86      0.87      0.87       138

 micro avg     0.94      0.94      0.94      8815
 macro avg     0.90      0.90      0.90      8815
 weighted avg  0.94      0.94      0.94      8815

 Overall algorithm accuracy: 0.9372
```

Gradient Boosting

```
          precision    recall  f1-score   support
       1       0.98      0.92      0.95      3948
       2       0.85      0.87      0.86      3117
       3       0.70      0.67      0.69      1149
       4       0.60      0.82      0.70       471
       5       0.60      0.88      0.71       130

 micro avg     0.86      0.86      0.86      8815
 macro avg     0.75      0.83      0.78      8815
 weighted avg  0.87      0.86      0.87      8815

 Overall algorithm accuracy: 0.8635
```

Winning Model

The **Random Forest** algorithm has come up on top. This model classifies all categories much better then the other two algorithms and demonstrates a nice balance between the recall and presision metrics. The Random forest algorithm is also the fastest to train.

Category	**RF f1-score**	SVM f1-score	GB f1-score
1	**0.97**	0.96	0.95
2	**0.93**	0.91	0.86
3	**0.88**	0.85	0.69
4	**0.85**	0.83	0.70
5	**0.87**	0.85	0.71
Accuracy	**0.9372**	0.9199	0.8635

Dimentionality Change

The winning algorithm performs quite spectacular. It employs the **seven** top features we have identified in the *Feature Selection* section. Let's see how the input data dimentionality change affects the model performance. Firstly we reduce the number of features to **five**.

Top five features

Num	Feature
24	Do you receive your money back in time?
26	What's the most common use of the money you lend?
22	Do you include either interest or a lending fee when you lend?
18	Over the past 3 months, how many times have you lent someone money?
20	Who did you lend money to in the past 3 months?

Model Performance:

```
          precision    recall  f1-score   support
       1       0.94      0.91      0.92      3858
       2       0.84      0.74      0.79      3150
       3       0.58      0.73      0.65      1210
       4       0.55      0.71      0.62       464
       5       0.55      0.89      0.68       133

micro avg      0.81      0.81      0.81      8815
macro avg      0.69      0.80      0.73      8815
weighted avg   0.83      0.81      0.82      8815

Overall algorithm accuracy: 0.8635
```

Evidently the dimentionality reduction caused the model performance deteriorate greatly. Now let's increase the number of features to **nine**.

Top nine features:

Num	Feature
24	Do you receive your money back in time?
26	What's the most common use of the money you lend?
22	Do you include either interest or a lending fee when you lend?
18	Over the past 3 months, how many times have you lent someone money?
20	Who did you lend money to in the past 3 months?
23	Do you request guarantees when you lend?
19	On average how much do you lend in general?
16	Country
21	When you lend money, when do you usually expect to get it repaid?

Model Performance:

0.9547

```
           precision    recall  f1-score   support

       1       0.99      0.98      0.98      3893
       2       0.96      0.96      0.96      3159
       3       0.89      0.92      0.91      1151
       4       0.85      0.87      0.86       482
       5       0.88      0.88      0.88       130

 micro avg     0.95      0.95      0.95      8815
 macro avg     0.91      0.92      0.92      8815
 weighted avg  0.96      0.95      0.95      8815

 Overall algorithm accuracy: 0.9547
```

The dimentionality increase gave us an overall perfromance boost of almost **2%**. It might not seem much. Let see how performance of each category has been affected.

Category	7 Features	9 Features	Gain (%)
1	0.96	0.98	2
2	**0.93**	**0.96**	**3**
3	**0.88**	**0.91**	**3**
4	0.85	0.86	1
5	0.87	0.88	1

Evdently categories 2 (*Poor*) and 3 (*Fair*) have benefited the most form the dimentionality increase. Ultimately it is up to the business to decised if 3% accuracy gain is worth the training time and user interface complexity increase. KASI Insight represantative has opted for higher accuracy.

Hyper-parameter Tuning

Hyper-parameter tuning is ususally the last step in effort to improve the model performance. We will employ *Grid Search* algorithm with **three-fold cross validation** to identify the best model parameters. The parameter grid look as follows:

Patameter	Values
Number of Estimators	200, 300, 400
Minimum Sample Split	5, 10, 20, 30, 40
Maximum Features	'auto', 'sqrt'
Bootstrap:	True, False

The hyper-parameter tuning gave us another **0.5%** performnace gain.

Final Simulator Model Stats

- Number of features: **9**
- Overall algorithm accuracy: **0.9594**

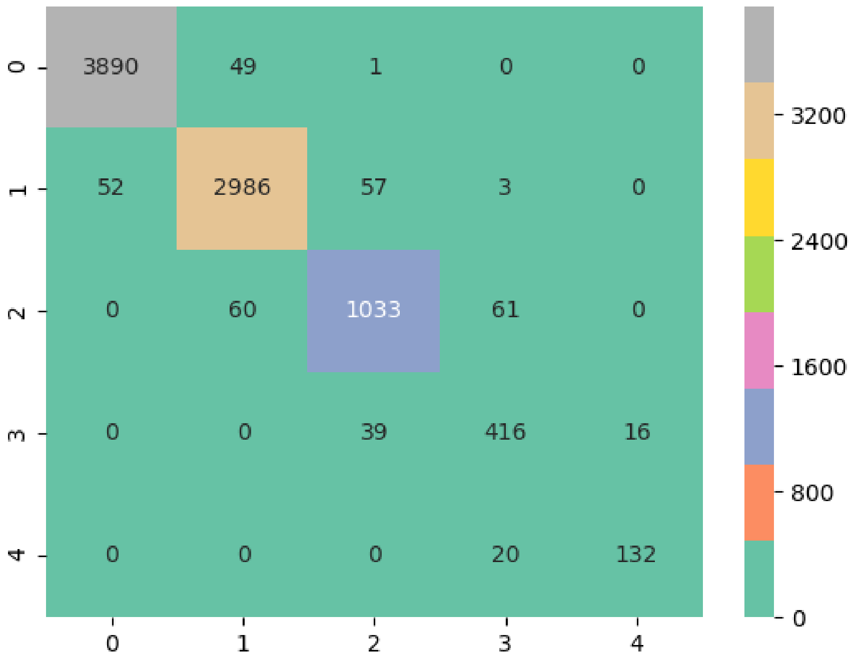

Figure 14: Simulator Model Confusion Matrix

```
              precision    recall  f1-score   support

           1       0.99      0.99      0.99      3940
           2       0.96      0.96      0.96      3098
           3       0.91      0.90      0.90      1154
           4       0.83      0.88      0.86       471
           5       0.89      0.87      0.88       152

   micro avg       0.96      0.96      0.96      8815
   macro avg       0.92      0.92      0.92      8815
weighted avg       0.96      0.96      0.96      8815
```

Lastly we are going to review the model learning and vaidation curves. As per figure 15 the model was learning more about the data as the training size grew. When the training size reached about 30,000 observations the validation curve converged with the training one indicating that the further increase in the training set size will not likely result in better model performance.

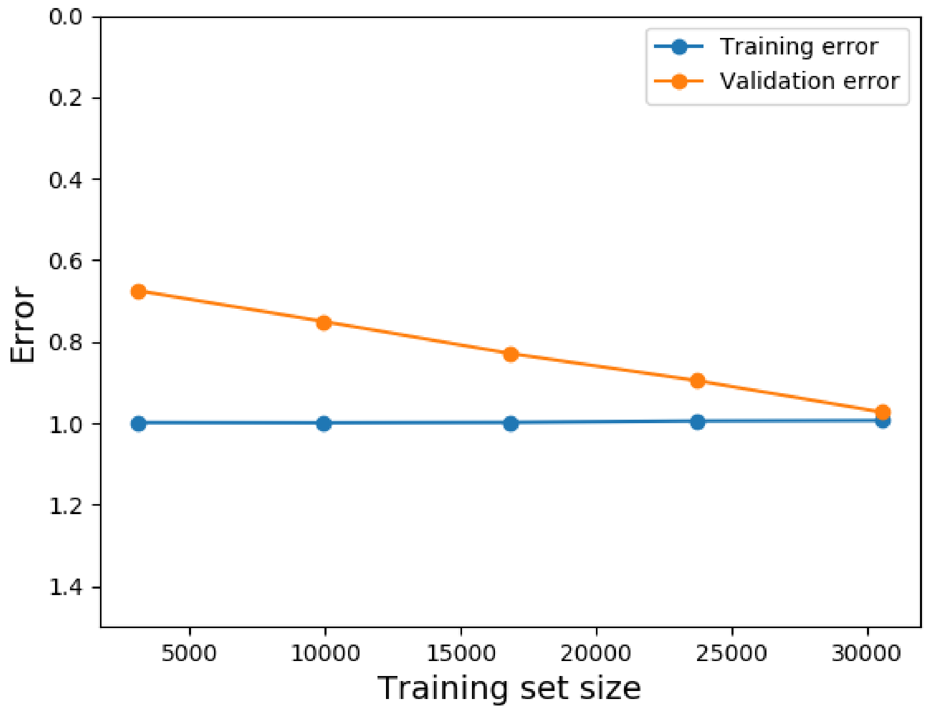

Figure 15: Simulator Model Learning Curves

Lender Evaluator

Noe we are going to repeate all the same steps to find thebest model for the *Lender Evaluator* tool.

Model Deployment

Architecture

Docker

Conclusion

Note from the Authors

This file was generated using *The R Journal* style article template, additional information on how to prepare articles for submission is here - Instructions for Authors. The article itself is an executable R Markdown file that could be downloaded from Github with all the necessary artifacts.

Vadim Spirkov
York University School of Continuing Studies

Murlidhar Loka
York University School of Continuing Studies