Lending Environment Simulation Model & Lender Evaluation Tool

by Vadim Spirkov, Murlidhar Loka

Abstract In 2013, Africa was the world's fastest-growing continent at 5.6% a year, and GDP is expected to rise by an average of over 6% a year between 2013 and 2023. In 2017, the African Development Bank reported Africa to be the world's second-fastest growing economy, and estimates that average growth will rebound to 3.4% in 2017, while growth is expected to increase by 4.3% in 2018. The World Bank expects that most African countries will reach "middle income" status (defined as at least US$1,000 per person a year) by 2025 if current growth rates continue.(Ref: Wikipedia) With growth of the income African and international financial institutions are looking for the opportunities to extend their services to the new geographic areas.

Introduction

According to McKinsey, Africa has the third largest population of unbanked adults (326 millions) representing 80% of the adult's population. According to banking executives in Africa, assessing consumer credit risk is one of the main challenges the banks face, which results in the slowing the access of general public to the credit resources.

KASI Insights is an award-winning consumer and market intelligence firm that provides biometric research at scale in Africa. KASI offers the first crowd sourced credit model. KASI Insights engages, questions and learns in real time from consumers and market participants in over 10 markets in Africa to uncover what they really feel about the products and services the companies consider to offer in the continent before they launch them. (Ref:Insights)

Background

In Africa money lending between the individuals is very common. This practice in fact is a substitution to the small loan franchises, which are wide-spread in Western countries. One of the surveys KASI conducts for years targets the individuals who lend money to other members of the community on regular basis. The firm leverages the wisdom of the crowd to evaluate lending activities within a community over time to compute a community-based score that can be translated into an individual score. Thus KASI's surveys provide risk profile of the population in a given area through the eyes of the lender.

Having many conversations with the KASI Insights CEO, we have concluded that, if properly used, these surveys could be a valuable intel for the banks and other financial institution to evaluate a possibility of establishing branches in suitable communities to conduct small loan business.

The data collected by KASI Insight could also be used by the banks and small lenders alike to evaluate the lending environment in African countries/communities simulating various lending criteria.

In this context we hope that applying Machine learning techniques and the latest advances in the software development we will put the collected data to use and help KASI Insights increase the offering to the interested parties, thus improving the company's position in the field of economic and marketing research in Africa.

Problem Statement

This project has two objectives.

Design and Implement Lending Environment Simulation Model

The first objective is to help KASI Insights to provide a valuable tool to the financial institutions and the small lenders in Africa, which will be used to **predict a credit score class of the target group of people in a given area manipulating multiple lending criteria**.

Approach

The brain of the tool is going to be a **multiclass classification model**. We will be employing a credit score formula provided by KASI Insights and lender's survey data to label each observation. KASI Insights has come up with the following credit score classes:

- **Very Poor** - the credit score less than **350**
- **Poor** - the credit score is between **350 and 550**
- **Fair** - the credit score is between **550 and 650**
- **Good** - the credit score lies between **650 and 750**
- **Very Good** - the credit score is higher tan **750**

We also will provide a Web-based user interface to communicate with the model. We plan to develop the solution employing *Python* and *Angular* JavaScript framework. The model will be hosted in *Flask* Web application server. The code and the supporting software will be shipped to the client as a *Docker* package for the deployment on **AWS** cloud.

Lender Evaluation Tool

Lender evaluation tool will provide means to **estimate the business savviness of an individual who lends money on regular basis**. It is meant to be used by the banks and other financial institutions which are interested in having a proxy in a given community to conduct small loan business in their behalf.

KASI insight has committed to formulate an algorithm that calculates the worthiness of a lender as a business partner in the money lending context. We will be employing the demographic data and the lending habits of the lender contained in the survey data and the lender worthiness score to label all observations.

At the moment of writing we did not know if KASI Insights wanted to predict the lender worthiness score as an absolute number or use a few classes that would define how good/bad the lender as a business partner is. Depending on the final decision we will train either a **regression** or **multiclass classification** model.

Similarly to the first objective deliverable the final product will be a Web-based solution implemented using *Python* and *Angular*. The code and supplementary software will be *dockerized* for easy roll-out on a cloud platform.

Dataset Description

As it has been mentioned previously we will be using KASI lender survey data for the project. The data has been collected surveying people from seven African countries over a course of the last three years. The data is maintained in the Excel spreadsheets in English and French. The survey comprises 38 multi-choice questions. There are almost **30,000** observations. The table submitted below lists the survey questions:

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

ID	Question/Column
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?
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?

Since the survey has a multi-choice format the data could be easily categorized prior to fitting the model. As discussed the labels for each observation are:

- Credit score category based on the credit score. This label will be employed for training of the **Lending Environment Simulation Model**
- Lender business worthiness score or a category (TBD). This label will be used to train the **Lender Evaluation Model**

Data Pollution Challenge & Solution

The survey has been modified a few time since its inception. This lead to data pollution. In the latest version of survey each question offers a few possible answers, 3 - 4 on average, some questions have up to nine possible answers. In realty the preliminary data analysis shows that some questions have dozens, in extreme cases even up to 1000 possible answers...

Solution

To rectify the problem of the data pollution we suggest and have agreed with the business to use the latest version of survey to categorize the answers. Those answers that do not fall into any of the available categories will be categorized as *Others*. Considering the fact the the survey has not been change much for the last couple years we expect that the majority of the data will be correctly classified. On top of it will apply sophisticated algorithm that try to categorize even polluted answers.

Considering racial nature of the question #15 we have reached an agreement with the business to exclude the question from the process.

Data balance

At the moment of writing we did not know if the dataset was balanced. If it is not we would consider the data upsamling techniques.

ML Solution

To meet the project objectives we propose the following pipeline.

Data Preprocessing

1. Convert data from Excel format to *.csv format and encode it in UTF-8 encoding. Save the data into two files: one file will contain the data in English and the other in French.
2. Develop a script to clean and categorize the data using the latest KASI survey to identify the categories correctly
3. Using the formulas provided by KASI Insights label each observations. There will be two labels: one for the credit category class and the other for the lender business worthiness. Save the clean and labeled data into a *.csv file.
4. Document data preprocessing steps so our business partner could reproduce it when the model re-training is required.

Feature Selection and Dimensionality Reduction

The dataset we are dealing with has 37 columns. Apply feature selections techniques such *univariate selection*, *feature importance*, *correlation matrix with heatmap*, etc. to find the features that contribute the most to the prediction. We can also apply PCA technique to reduce the dataset dimensionality. These steps should improve the model performance, speed and reduce overfitting.

Model Selection and Training

1. For each proposed model we will evaluate three algorithms:

- **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.

If time permits we will attempt to evaluate a Neural Network in addition to the aforementioned models.

The listed above models are good for classification and regression tasks. Thus if KASI insights decides to employ regression approach for the second project we are covered.

2. Plot and interpret the model learning curves to detect possible problems such as overfitting, luck of training data, etc.
3. As an evaluation criteria we will employ the multiclass confusion matrix and averaged micro and macro **F1 scores**. The multiclass confusion matrix will clearly tell how well each algorithm detects the categories. We believe that the F1 score is ideal metric for the model evaluation because it provides the **balanced** value of accuracy.
4. Pick the winning model taking into account the following criteria:

- Model prediction power (the higher F1 score the better)
- Model performance. Considering the fact that the model will be used on-line we shall no underestimate this factor; the faster the model the better.
- Training time. For time being, if required, KASI Insight will be re-training the model using PC. Thus re-training the model shall not be taking days…

Project Plan

We have eight weeks to make a dream a reality.

Week 1

Understand the business domain and the data. Discuss the goal the business partner wants to achieve and evaluate the feasibility of the task. Do gap analysis. If the goal the client has in mind is not achievable due to either luck of data, time constraints or misunderstanding of the ML capabilities offer alternative solution, which the client could benefit from. Get all missing pieces of the information from the business partner. Finish and submit project proposal.

Week 2

Clean the data. Do data exploration. Address possible data issues such as unbalanced data. Document the data exploration findings and outcomes.

Weeks 3 - 4

Train and evaluate the models. Pick the winning model using the evaluation criteria discussed earlier. Document the process. Submit the first milestone assignment.

Weeks 5 - 7

Develop turnkey solution:

- Develop Web-client employing *Angular* JavaScript framework.
- Deploy the model to *Flask* Web application service.
- Test the solution end-to-end locally.
- *Dockerize* the code.
- Deploy to AWS cloud.

Submit **Project Solution** assignment.

Week 8

Address possible issues with the code. Polish the project report and the presentation. Educate the client on how to use the product, re-train the model. Advise the client what could be improved and what possible challenges the client might have in future. Do final submission. Ship all project artifacts to the client.

Bibliography

K. Insights. About kasi. URL https://www.kasiinsight.com/about. [p1]

Wikipedia. Economy of africa. URL https://en.wikipedia.org/wiki/Economy_of_Africa. [p1]

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