Udacity Machine Learning Nanodegree - Capstone Project

Ruslan Kozhuharov

May 31, 2018

Contents

1 Definition

1.1 Project Overview

The Kiva organization is a charitable entity that funds underprivileged people around the world through small loans. Such loans are funded on Kivas online platform by one or several donors. The loan amount is then disbursed by Kiva associates to financially excluded individuals. Kiva itself does not collect rent on those loans, but Kiva associates could in order to cover their operating costs. Once a loan is repaid, the individual donors could choose another cause and continue their charitable activity. This helps poor communities by:

- Fostering economic activity
- Providing access to funds for financially excluded individuals
- Encouraging entrepreneurial initiative on the part of the borrower

In order to improve their reach, however, Kiva needs a good assessment of the poverty levels in their areas of operation. To do that, Kiva has requested in a Kaggle challenge that a poverty score is created based on other data and that score is combined with a data set for Kivas loans from the last 2 years.

1.2 Problem Statement

Based on the definition of the Kiva challenge laid out in the Project Overview, the problems we are going to solve are the following:

1. Defining the mathematical expression of poverty
2. Selecting the necessary regional data sets
3. Transforming the data set from step 2 to a form usable for modeling
4. Building the poverty prediction model
5. Joining the predicted poverty to all records of the Kiva loans data set

1.2.1 Defining Poverty

As mentioned above, poverty is a subjective notion that we need to convert to a mathematical measure. In defining that measure we need to take into account the following factors:

- The measure should be applicable on country as well as on global level.
- The measure should be on a scale between 0 to 1 so that it represents a percentage of poverty. This will allow Kiva to report on poverty-affected areas with regards to the measures maximum.
- The measure should be scalable if corresponding data from more countries is added.

With regards to these points, we will define the poverty score as: 1 - predicted normalized total household income. That is, poverty score of 0.8 corresponds to predicted normalized total household income of 0.2. Thus our prediction will always be on scale from 0 to 1, will be applicable to country and global levels and will be scalable as more country profiles are added (the normalized household income will always be from 0 to 1).

1.2.2 Selecting Data Sets

There are two main problems that need to be resolved when selecting data sets useful for the modeling task:

- Data with the widest possible coverage (World Bank, UN statistics, CIA World Factbook) is averaged out for each country and does not present a nuanced enough picture of the poverty levels. That is, in countries with high GINI coefficient, the average income levels do not provide a sufficient information about areas with a significant deviation from the average (poverty affected areas).
- Local country data is presented in different formats. Every country measures different macroeconomic indicators (and sometimes measures the same indicators in different ways). This makes local data, that could eventually be joined together on global level, unreliable. Example: some countries may consider 'unemployment' as the number of working age individuals with no regular employment contract as unemployed. Other countries may consider 'unemployment' as individuals who do not have a regular employment contract AND are looking for employment (i.e. individuals who are not actively looking for jobs are not counted).

To address these issues, we will select the country where Kiva grants the highest number of loans. We will focus exclusively on that country and find a detailed dataset, preferably containing raw data. We will then build an adequate model for the selected country and reduce the number of variables to a few, easily obtainable for other countries. This will make our model, although localized, scalable.

The country receiving the most Kiva loans at this point is the Philippines. Therefore, we will use data from the Philippines Family Income and Expenditure triennial survey (FIES). The data contains information about families incomes, expenses, and living conditions (with detailed information about accommodation types, running water, electricity, communications devices, family size, education, etc.). The data set is published on Kaggle and is produced by the Philippines Statistics Authority (PSA).

This data set will be used to derive a model that can predict poverty out of the available features. The poverty score will then be grouped by region and household head gender. We will also derive the administrative regions from the Kiva data set and join the calculated poverty score by region and gender of the borrower.

1.2.3 Transforming The Data

Since the data set we will utilize will contain raw data, we will need to transform it to numerical or one hot encoded features. We will perform this work separately in a data transformation phase.

1.2.4 Building The Poverty Prediction Model

Once we have transformed the FIES data, we will need to find eventual correlations, remove irrelevant fields and prepare for modeling. We will then proceed and try the effectiveness of several out of the box models (we will refer to them as candidate models) against three dummy predictors.

The dummy predictors will always predict poverty scores of 0, 0.5 and 1, respectively. We will then compare the performance of the candidate models against the dummy predictors based on the evaluation metrics defined in the next chapter. Finally, we will select the best performing model and improve on it.

The improvement will consist of selecting the features with highest predictive potential and discarding the rest. This will allow Kiva to easily integrate the model in their operation (by simply adding the 3-5 new questions to their loan application form).

1.2.5 Joining The Predicted Poverty To Existing Kiva Data

In this phase, we will find an appropriate way to join the results of the poverty prediction model to the existing Kiva loans data set. We will join the poverty scores by borrower gender and region.

1.3 Metrics

We will utilize the following evaluation metrics for the household income prediction:

- Mean squared error:

$$MSE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2 \tag{1}$$

- R2 score:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{samples}-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{samples}-1}(y_i - \bar{y}_i)^2} \tag{2}$$

- Mean squared logarithmic error:

$$MSE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (log_e(1 + y_i) - log_e(1 + \hat{y}_i))^2 \tag{3}$$

- Explained variance score:

$$EV(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}} \tag{4}$$

2 Analysis

2.1 Data Exploration

We are using two main data sets: the Kiva loans dataset and the Philippines Family Income and Expenditure Survey (FIES). Both of the datasets are not immediately suitable for the tasks laid out in the Problem Statement. This means that we will have to transform the data in certain ways. We will now look at each data set and its corresponding data transformation.

2.1.1 Kiva Loans Data Set

The Kiva loans data set contains 671205 records for loans granted by Kiva in the last 2 years. For the purpose of this project we will use the following columns:

- Id: The unique identifier for each loan record. It has a numeric value.
- Country: The name of the country where the loan was granted in camel case. We will use this column to select all records from the Philippines.
- Region: Unstructured representation of an administrative entity that is below the level of a country. The entity could be a city, city and a province name separated by comma, village name, other geographic entities or no value. We will use this column to derive the actual administrative region in the Philippines where the loan was granted. The derived administrative region will be further used to join the poverty score by region.
- Borrower_gernders: The unique categories in this column are male and female. However, the columns values could be any combination of those two that is as long as the number of borrowers (e.g. female, female, male for 3 borrowers of the same loan record). We use this column to derive a borrower gender female one hot encoded column. The new column will be subsequently used to join the poverty score by gender.

Here is a sample of the contents of the above-mentioned columns for 5 records:

id	country	region	borrower_genders
1209550	Philippines	Dipolog -Zamboanga del Norte	female
888025	Nicaragua	Masaya	female, female, female
1171969	Philippines	Roxas Palawan	female
867403	El Salvador	NaN	male
742400	Nigeria	Kaduna	female

As we can see from this short example, the region column could contain a ¡city¿ ¡province¿ tuple (for loan id 1209550, where the city is Dipolog and the province is Zamboanga del Norte), only city (Masays for loan id 888025), municipality name (Roxas Palawan for loan id 1171969) or no valid value (NaN for loan id 867403). We will need to transform this column to a standard value using matching tables and other heuristics.

For the records from the Philippines, the unique values for the borrower_genders column are only: female, male and NaN. Thats why will also transform the borrower_genders column by mapping the female category to 1 and the male category to 0. There are only 80 records with missing values (NaN), so we could do away with them.

2.1.2 Philippines Family Income and Expenditure Survey

The Philippines Family Income and Expenditure Survey (FIES) contains 41544 records of household incomes, expenditure, living conditions and other affluence indicators. The data set contains 59 columns representing information in the following categories:

- Demographics
  - Agricultural Household indicator
  - Household Head Sex (categorical)
  - Household Head Age (categorical)
  - Household Head Marital Status (categorical)
  - Type of Household (categorical)
  - Total Number of Family members
  - Members with age less than 5 year old
  - Members with age 5 - 17 years old
  - Region (categorical)
  - Household Head Highest Grade Completed (categorical)
- Expenses

  - Total Food Expenditure
  - Bread and Cereals Expenditure
  - Total Rice Expenditure
  - Meat Expenditure
  - Total Fish and marine products Expenditure
  - Fruit Expenditure
  - Vegetables Expenditure
  - Restaurant and hotels Expenditure
  - Alcoholic Beverages Expenditure
  - Tobacco Expenditure
  - Clothing, Footwear and Other Wear Expenditure
  - Housing and water Expenditure
  - Medical Care Expenditure
  - Transportation Expenditure
  - Communication Expenditure
  - Education Expenditure
  - Miscellaneous Goods and Services Expenditure
  - Special Occasions Expenditure
  - Crop Farming and Gardening expenses

- Income
  - Main Source of Income (categorical)
  - Total Income from Entrepreneurial Activities
  - Household Head Job or Business Indicator (categorical)
  - Household Head Occupation (categorical)
  - Household Head Class of Worker (categorical)
  - Total number of family members employed

- Living Conditions
  - Imputed House Rental Value
  - Type of Building/House (categorical)
  - Type of Roof (categorical)
  - Type of Walls (categorical)
  - House Floor Area
  - House Age
  - Number of bedrooms
  - Tenure Status (categorical)
  - Toilet Facilities (categorical)
  - Electricity
  - Main Source of Water Supply (categorical)
  - Number of Television
  - Number of CD/VCD/DVD
  - Number of Component/Stereo set
  - Number of Refrigerator/Freezer
  - Number of Washing Machine
  - Number of Airconditioner
  - Number of Car, Jeep, Van
  - Number of Landline/wireless telephones
  - Number of Cellular phone
  - Number of Personal Computer
  - Number of Stove with Oven/Gas Range
  - Number of Motorized Banca
  - Number of Motorcycle/Tricycle

It is worth noting that the fields Household Head Occupation and Household Head Class of Worker contain only 34008 non-NULL values. We should also apply some category reduction to the fields, one hot encoding as well as query-friendly names to these fields. In addition, all the categorical fields (marked as 'categorical' in the bullet list above) will require additional transformations. Besides the categoricall transformations, all column names need to be changed so that they are query-friendly.

It is obvious that this dataset will require a significant number of preprocessing modifications. We can group those in the following preprocessing operations:

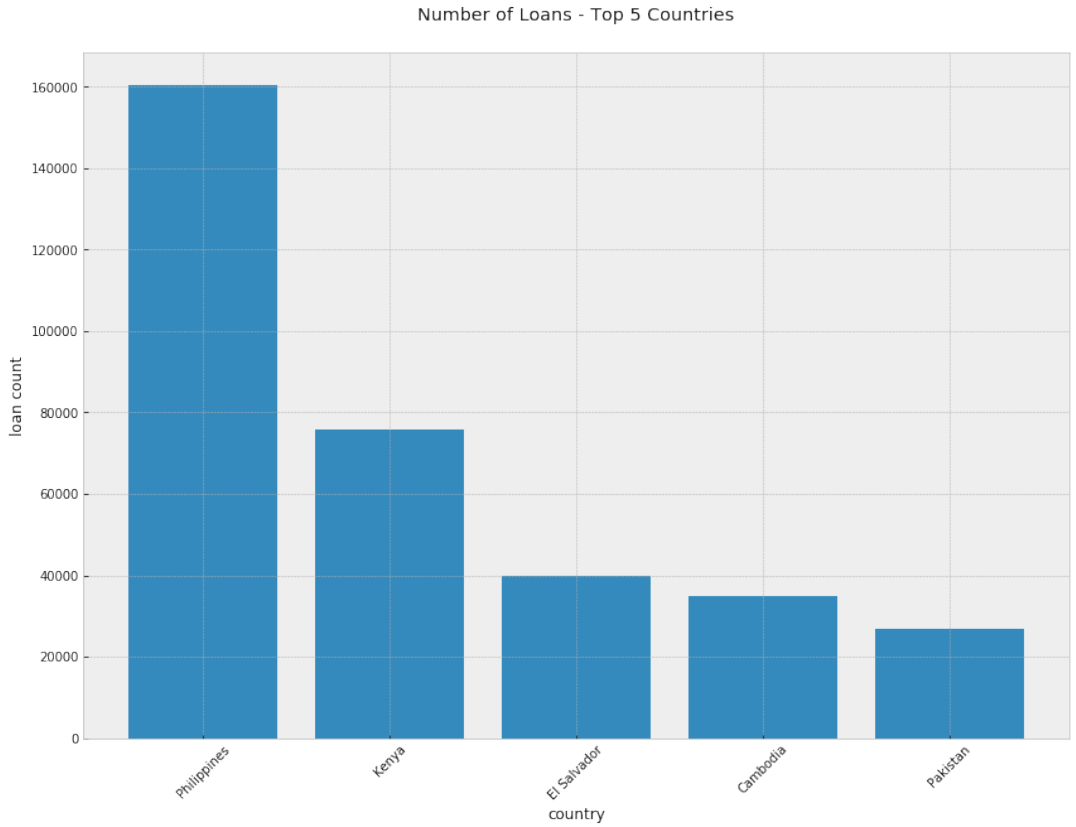
- Mapping column string values to query-friendly names
- One hot encoding
- Columns merging
- Reducing / categorizing unique column values
- Renaming columns to query-friendly names

We will look at all those preprocessing operations in detail in the Data Preprocessing chapter.

2.2 Exploratory Visualization

Let us start with the most important question - is the Philippines the region with the highest count of Kiva loans. To answer it, we will group and count the loans by country, sort the results by country and display the 5 countries with the highest loan counts:

Figure 1: Number of Loans Per Region



As we can see, the Philippines is the region where Kiva mostly focuses their efforts in terms of loan count. Actually, when it comes to loan count, there are twice as many loans granted in the Philippines as in the second country in the list (Kenya).

Now we can focus on the Philippines and the FIES data. We will first derive an estimated family employment feature. It will be defined as: number of employed family members / (total number family members - children). We will also derive no_electricity and no_running_water indicators.

Let's now plot the estimated family employment by region:

3 Methodology

3.1 Data Preprocessing

Figure 2: Estimated Employment Per Region