Udacity Machine Learning Nanodegree - Capstone Project Proposal

Ruslan Kozhuharov

May 26, 2018

Contents

1 Domain Background

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 dataset for Kivas loans from the last 2 years.

2 Problem Statement

In order to address the requirements of Kiva, the poverty score should be derived from external data. Such data should be detailed enough to be able to predict the target metric. In addition, the poverty score itself needs to be defined. Thus, the task of producing a poverty score can be broken down into several sub-tasks:

- Find a local, detailed and reliable datasource with personal metrics for the region with highest significance for Kivas operations (since data for different countries is differently formatted and not always available, we will limit ourselves to one country with reliable data).
- Define the poverty score metric.
- Create a model that predicts the poverty score.
- Join the poverty score to the existing Kiva loans database on borrower gender and region.

3 Datasets and Inputs

In order to build the poverty score, we will focus exclusively on the Philippines as this country receives both the highest number of loans and the highest loan amounts from all areas Kiva operates in. 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 dataset is published on Kaggle and is produced by the Philippines Statistics Authority (PSA). This dataset 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 dataset and join the calculated poverty score by region and gender of the borrower.

4 Solution Statement

In order to derive the poverty score, we will normalize the household income from the FIES dataset and use it as a target variable. We will then develop a model that predicts the household income from the other available features in the dataset. We will define the poverty score as the inverse of the normalized income (i.e. 1 - predicted normalized income). In this case, with a predicted income of 0.15 (15% of the normalized income on a national level), a person would have a poverty score of 0.85 out of 1. We will then calculate the poverty score for all individuals in the FIES dataset. Having done that, we will aggregate the results (taking the mean value) by region (administrative region in the Philippines) and household head gender. We will then join these aggregated results to every Kiva loan record by borrower gender and region. Finally, we will produce plots that visualize the poverty by region and gender.

5 Benchmark Model

Since the normalized values for the household income have a range between 0 and 1, we will use 3 dummy predictors to evaluate the performance of our actual model against:

- A dummy predictor that always predicts household income 0
- A dummy predictor that always predicts household income 1
- A dummy predictor that always predicts household income 0.5

We will evaluate these regressors based on our selected evaluation metrics and compare them to our actual model under development.

6 Evaluation 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:

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

- Mean squared logarithmic error:

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

- Explained variance score:

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

7 Project Design

The project will be organized in the following phases:

7.1 FIES Data Cleanup

In this phase, we will perform several standard operations on all the columns of the FIES dataset, namely:

1. Where appropriate, we will rename the column to a query-friendly name (e.g. 'Total Household Income' will be renamed to 'household_income').

2. Where appropriate, we will map the discrete values of a column to query-friendly names and one hot encode them into several columns (e.g. the 'Household Head Sex' column contains the values 'Male' and 'Female', we can map these values to 0 and 1 respectively and rename the column to 'household_head_f').

3. Where appropriate, we will reduce the distinct values of a column to few categories (e.g. the 'Household Head Education' column contains various values for every grade completed and every different university degree - we can map these values to: 'elementary', 'primary', 'secondary' and 'tertiary' education levels).

7.2 Kiva Loans Region Matching

The Kiva loans region column contains unstructured string data. Sometimes the value of the column represents a city name, sometimes a pattern of ¡city¿, ¡province¿ and sometimes the name of a non-administrative population center. We will attempt to match such locations to a specific region in the Philippines. To do that, we will build a small table reflecting the hierarchy of all major administrative regions and centers in the Philippines as follows:

- Island Group
- Region
- Province
- Province capital city

We will use this table and attempt to match various entities from it the strings in the Kiva loans data set. If we find matches, we will assign the corresponding region value from the table to the respective Kiva loan record.

7.3 Data Selection

We will use the processed data from phase 1 and inspect the correlation patterns between the various features. We will remove features that highly correlate with the target variable (both in semantics and in correlation levels).

> Example: the FIES data contains the columns Total Household Income, Income from Wages and Income from Entrepreneurial Activities. If we are to predict the Total Household Income, we need to dismiss the Income from Wages and Income from Entrepreneurial Activities as these fields have the same meaning and therefore cannot be predictors of the Total Household Income.

We will also make an initial assessment on which features are most appropriate based both on correlation levels and common sense.

7.4 Initial Model Selection

In this phase, we will select a model based on a choice of several out-of-the-box solutions and the dummy scorers from the Benchmark Mbenchodel chapter. To select the model, we will compare the performance of all the initial models on the training and validation sets and on the selected evaluation metrics detailed in the Evaluation Metrics chapter.

7.5 Final Model Selection and Feature Set Reduction

Out of all the models from the Initial Model Selection, we will choose the best two and inspect the features with highest weights. We will then select the model that has:

- The most common sense features (e.g. 'occupation_farmer', 'no_mobile_phone', etc.). The purpose of this criterion is to present Kiva with a model that could be converted to questions easily asked in the field. This will allow Kiva and Kiva partners to quickly assess a persons poverty levels based on simple and straightforward input from the potential borrowers.

- The potential to reduce the number of features to 5 and below. Again the practicality of Kiva's work imposes on us that we are able to have few, clear and crisp input to the model.
- The best performance on the Evaluation Metrics.

7.6 Poverty Score Calculation and Application

In this phase, we will calculate the poverty score based on the selected model. We will then calculate the mean poverty score by region and gender in the Philippines. We will use these results and join them to Kiva's original loan dataset (again by region and gender) and visualize the results.