Machine Learning Engineer Nanodegree

Capstone Proposal

Rahat Dewan
April 18th, 2018

Domain Background

Kiva Microfunds, (more commonly known as Kiva.org) is an online, non-profit crowdfunding initiative created to extend financial services to the poor and financially excluded around the world - in particular low-income entrepreneurs and students. Over $1 billion in loans have thus far been provided to over 2 million people through the platform, with a repayment rate of between 98 and 99 per cent. Kiva does not collect any interest on the loans it facilitates and Kiva lenders to not make interest on loans.

For the purposes of setting investment priorities, helping to inform lenders and understanding target communities, knowing the level of poverty of each borrower is critical. However, this requires inference based on the very limited data available to us on the characteristics of each borrower.

In Kaggle's first 'Data Science for Good' challenge, Kiva invited the Kaggle community to help them build more localised models for the purposes of estimating the poverty levels of residents in the regions where Kiva has active loans.

Motivation

My personal motivation for undertaking this project in particular stems from my background and prior experience with microcredit and microfinance. My ethnic roots (and, indeed, the country I lived in for two years of my life) lie in Bangladesh, the nation in which Nobel Peace Prize recipient Muhammad Yunus pioneered microfinance. I completed a research project in Dhaka with one of the largest providers of microcredit loans, and while it is not the 'silver bullet' against poverty it was once touted as being, it may still have great potential as a policy tool. I am in particular interested in how adding technology and the latest crowdfunding trends to the traditional microcredit model in order to leverage global goodwill might lead to a reduction in poverty.

Problem Statement

For the locations in which Kiva has active loans, the objective is to estimate the welfare level of borrowers in specific regions, based on shared economic and demographic characteristics, on as granular a level as possible. Ideally the solution will leverage information such as the borrowers gender, average welfare metrics in the area, and borrowing behaviour in order to estimate the borrowers welfare level.

I will also evaluate the possibility of carrying out a performance evaluation of Kiva loans - this will depend heavily on data limitations. At least, I hope to include such an analysis for a certain region which has particularly granular poverty data.

Datasets and Inputs

Kiva has provided a dataset of loans issued over the last two years. This includes characteristics of the loan such as loan amount, activity, sector, location and borrower genders. I will need to pair this data with external data of geographical poverty estimates. Thus far, I have found three relevant datasets that I can map the Kiva-provided one onto in order to provide a solution. These are: * OPHI's MPI * Global Findex * World Bank Living Standards survey

The former two are global but not very granular, whereas the opposite is true for the last.

Any further data I find during the course of the project may also be leveraged to provide more localised predictions, such as: * Means testing data (consumption, household size, various other indicators) * Spatial data including satellite data, environmental and climate data. * Socioeconomic and Demographic data * Conflict data * Demographic and health surveys

Solution Statement

The solution to the problem will be the production and tuning of a model which predicts my chosen dependent variable, Multi-dimensional Poverty Index (MPI), using the input variables from my cleaned and processed data.

I will want my model to be as **localised** as possible, (providing predictions on as granular a level as possible), as **global/extensive** (covering as much of the world or Kiva's target countries as possible) and as **precise** (low standard errors) as possible.

Benchmark Model

The benchmark model for this problem will be Kiva's own current poverty targetting system. Currently they use the OPHI MPI data as follows: * Merge in the MPI with the loan data by region at as granular a level as the data allows. * Take the average across all regions (weighted by volume) for a given loan theme. (This is the Loan Theme MPI Score)

I will need to provide a model that is more localised and precise than this very simple one.

Evaluation Metrics

While I will be exploring other models/classifiers, my default and first avenue of enquiry will be a simple OLS regression on selected characteristics. Thus my evaluation metrics in this case will be R^2, significance of features (p-values) and RMSE.

Project Design

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.

Data Gathering

The first step will be to gather all relevant data from the various sources I have listed above. Throughout the project, if I happen across more useful data, I may choose to include it and see if better predictions can be made.

Data Cleaning and Processing

Once all the data has been gathered, it will need to be cleaned. This will entail removing missing values, outliers or other problematic observations. Then, it will need to be processed such that it can be merged into a complete dataset ready for analysis.

Feature Selection and Transformation

Feature selection will be an integral part of the project, particularly if there are a lot of potential features yielded from the data. While evaluating usefulness and relevance of proposed features, I will consider utilising a PCA procedure or some other feature transformation method. The features may be tweaked and refined gradually as I progress through the project.

First (Benchmark) Model: OLS

My first model will be a simple OLS regression. After tweaking features and specifications to get the most significant and explanatory OLS results I can, I will consider this my personal benchmark model (it is hoped that it will itself be a more successful model than the current Kiva model.) As much as possible, I will take account of traditional econometric considerations such as tests for heteroskedasticity and autocorrelation, and refine my model accordingly.

Experimentation: Different Classifiers & Features

Other classifiers will be considered in the hopes of improving upon the OLS model. The main one under consideration currently is XGBoost. I will choose the most successful classifier after considering evaluation metrics.

In-depth Analysis of a Single Region

If I find appropriate data, I will carry out a more thorough and detailed analysis for a single region or country, in order to see how detailed and precise I can get my model - and also consider carrying out an impact evaluation. This will be useful for Kiva in the future if more data is available, as they will be able expand this more thorough analysis to other regions.

References

Kaggle, Data Science for Good: Kiva Crowdfunding.

Wikipedia, Kiva

Kiva.org

OPHI, Global Multidimensional Poverty Index

The World Bank, Global Findex

The World Bank, Living Standards Measurement Study