Predicting Food Health & Security

Victor Vulovic#1

#GalvanizeU, University of New Haven  
44 Tehama St, San Francisco, CA 94105, USA

1hello@victorvulovic.com

Abstract— This document gives an overview of data exploration, project procedure, and experimental results for 5 years worth of data from the Global Food Security Index (GFSI). Efforts on this dataset reflect a basic blueprint for other datasets of interest to “Project 8”, a cross-collaborative initiative of Nielsen, Inc. and The Demand Institute. Project 8 is currently working to create a massive global database dedicated to solving the 2020 Sustainable Development Goals of the U.N.

Keywords— GridSearch, Machine Learning, Global Development, Food Security, Public Policy

1. Introduction

In September of 2015, world leaders at the UN Summit declared 17 Sustainable Development Goals (SDGs) [1] to be achieved by 2030. These goals represent inspiration of all global citizens to collaborate on major world issues such as poverty, hunger, education, environmental protection, etc. Number two on the list of 17 is for Zero Hunger. This is what inspired a collaboration with Nielsen, The Demand Institute, The Conference Board, and the United Nations Foundation to create Project 8 [2].

As part of a Capstone assignment for GalvanizeU, I chose to work with Nielsen and The Demand Institute to attempt to model food security/risk of shortage. Known formally as Project 8, this current initiative aims to alleviate food insecurity through data awareness and analysis by aggregating the world’s key data sources that impact food security. The goal of this project was to create a model for assessing and predicting communities at risk of food insecurity.

Project 8 seeks to become a global data hub for human needs, a resource for policy-makers, development professionals, field-workers, and community leaders alike. As described by leaders at Project 8, there are 3 main attributes hindering progress in pursuit of organizing data on global human needs:

1. Fragmented: there is no shortage of data available, but rather there are very many different sources with many various collection, storage, and sharing methods. They are hosted by different organizations, countries, & communities and are usually not in the same format, language, and/or timeframe. Many countries lack any collection mechanism whatsoever.
2. Variable: despite universality in the nature of basic human needs, there is a litany of perspectives and methodologies for collecting and recording data on said needs. Comparing & contrasting different approaches to data collection is challenging, makes normalization difficult, and reduces consistency when attempting to query databases. Additionally, countless unwritten (and explicit) assumptions impede a universal understanding of data, and thus, how to most appropriately solve problems.
3. Cumbersome: actually accessing this data requires technical (data/software engineering) skills and knowledge, access to computing power, and enormous patience. No one source of data is exactly like the other. Many API’s rely upon differing technology stacks, have strict query limits, and sparse/confusing documentation.

The current approach to solving the issue is to host Project 8 as a social network and shared Salesforce CRM tool. The idea behind this is to have a well-designed and user-friendly interface for people to quickly and clearly get answers to any questions they may have regarding the 17 SGDs. So far, they have launched a beta prototype of the product and received feedback from early members.

Moving forward, the plans are to continue growing the user base, iterating the product & design, and creating value for current and future users. This last part is what my Capstone proposal intended to do for the initiative.

The following section outlines more of the specifics of the Capstone proposal and hypothesis for experiment & research.

1. Overview
2. *Initial Proposal & Hypothesis*

The delivery of this Capstone project to GalvanizeU students was somewhat awkward in the sense that we, as novice Data Scientists, had little- to high-level understanding of the Capstone companies and the intricacies of their company strategies, cultures, and/or true nature of the task at hand. As such, it was our assignment to create a proposal based on a brief presentation that candidate companies presented to the entire student body.

Based on a broad strokes understanding of Project 8’s mission and background, the impression was that there was a very strong pre-existing user community, vast database, and ease of access to data. The problems outlined in the Introduction of this paper were not discovered until after the initial proposal was accepted by Project 8.

With these circumstances in consideration, my initial proposal was to take current data assets and combine them with 1-2 additional sources: climate and macroeconomic data. This combination could be used to see how measures of food security were affected by changes in climate or massive economic/political events.

With the impression that Project 8 already had a large amount of data available and user activity ready to go, this proposal was ambitious, yet achievable. However, the reality was quite different. There was very little user activity and pre-uploaded/cleaned data available. Instead, there were several very small (<50 rows) spreadsheets of unlabelled data.

In light of the shifted situation, and after a couple of weeks of research, communication, and learning, the focus of the proposal became more focused on data engineering than data science.

Eventually, after much trial and error, the hypothesis crystallized into: is there a learnable model in a single, robust data source, and can that procedure be replicated for future data sources?

1. Implementation

As described in the Introduction, most of the work that needs to be done for Project 8’s success at this stage of its development is in the realm of data engineering. However, pure engineering is not adequate for this problem. A solid understanding of the issue of “food security” and its background, history, and current state is absolutely crucial. Only once most of the engineering has been accomplished will there be room for productive data science. The available API’s are simply too highly fragmented, variable, and cumbersome to get enough data for sophisticated machine learning techniques such as deep learning.

1. Research

Food Security is a very complex, poorly understood, and difficult to solve phenomenon. Suffered by developing and developed countries alike, food security means different things in different cultures. Developing a single metric for food security is notoriously difficult because of these difference. Additionally, collecting accurate and timely data from all relevant countries is a massive task and often understaffed. As a result, there are many ways to measure the phenomenon and many techniques for measurement. Public resources who share data also don’t put an enormous effort into making their data easy to access or easy to figure out how to access.

An attempted work-around of this problem was creating a composite score for food security; however, I found that not only does that process create dubious results, it’s a very involved statistical procedure involving rigorous hypothesis testing for feature importance at levels of statistical significance— yet another hurdle at the mercy of not having enough, well-groomed and available data. Also a procedure out of scope for this project.

1. API Exploration & Pruning

Project 8 eventually supplied a list of approximately 40-50 API’s with access to data. I spent a significant amount of time systematically testing these resources. I read through all the documentation, made test calls, and recorded limitations such as query limits.

After some time, it became clear which API’s would be more helpful than others, so I classified each API based on likelihood of utility into four categories of utility: “no use”, “low”, “medium”, “high”.

1. Extract, Transform, Load (ETL)

For the most promising and straightforward API’s I downloaded samples and cleaned and pre-processed the data to make it ready for combination and analysis. After this process, I found that some data was more useful than others. Most downloadable data sets only covered activity for single years, and querying the API’s for all relevant columns was limited by obscure documentation and/or query limits. As such, I decided to focus on resources that provided several years worth of data to have the beginnings of a useful training set for machine learning model testing.

The most robust data source came from the Global Food Security Index (GFSI), which had 5 years worth of data. Nonetheless, total available data only totaled ~550 data points with ~35 features.

1. Analysis, Model Development & Motdel Tuning

In the GFSI datasets, there are 3 “food security” metrics that are collected: 1) Agency to ensure the safety and health of food; 2) Percentage of population with access to potable water; and 3) Presence of formal grocery sector. For my analysis, I chose 1) Agency to ensure the safety and health of food as the dependent variable for two main reasons. First, it sounded to be the most broad capture of food security in the dataset. Second, it was a simple binary classification of 0 or 1 based on a qualitative assessment. This made it more of a straightforward classification problem that could still generate somewhat useful predictions on even as small of a data set as this one.

In pursuit of the most useful model, I chose several classification algorithms and ran on a single year’s worth of data to determine the most performant models.

Once those top models were found, I used the GridSearch class from the Scikit-Learn library to find optimal hyper-parameteres to fine-tune the models. All code is hosted on a public repo in python and Jupiter notebook code [3].

In the next section, the results of the analysis is outlined in greater detail.

1. Model Results

After running baseline models with default parameters, I found the Random Forest Classifier and K-Nearest Neighbor Classifier to be the most effective. After hyper-tuning the parameters for my models, I found the below results on testing data:

Model comparison  
GFSI - Health & Food Safety Prediction

|  |  |  |  |
| --- | --- | --- | --- |
| Mod. | Performance Metrics (1 year window) | | |
| Name | Accuracy | MSE |
| RF | Random Forest Classifer | 91.3% | 0.087 |
| KNN | K-Nearest Neighbors Classifier | 91.3% | 0.087 |
| SVC | Support Vector Classifier | 87.0% | 0.130 |
| LR | Logistic Regression | 82.6% | 0.174 |

It’s likely that more data points would increase accuracy, which again goes back to the original challenge for Project 8: data engineering.

These accuracy and error results stayed consistent for up to a 2 year window. Then the model became less performant, as this is time-series data. As a result, these two models would be most useful for policy-makers and Project 8 users, who only need to have accurate forecasts for 1-2 years in the future.

For a more long-term forecast, we should use algorithms that are more sensitive to sequential patterns and/or conditional probabilities. More discussion on these approaches are in the following section.

1. Future Work & Conclusions

Food Security is inherently a long-term issue. Hopefully, it is something realistically addressable within the 15 year time window that the UN has given the worlds as a target with their 17 SDGs. As such, having a model that makes great predictions in 1-2 years could be useful for addressing short-term symptomatic issues related to food security. However, in order for individuals to receive real benefit to solving a long-term problem like food security, then they long-term solutions and models that can deliver reliable results into the future.

The best ways to do this are with Hidden Markov Models (HMM) and/or a combination of Recurrent Neural Networks (specifically Long-Short Term Memory Units). Employing a HMM seems most plausible given the smaller amount of data available, however, in order for the the algorithm to work, we need to establish probabilities of countries moving from one state to another state. We might be able to do this with the small amount of data available.

A more compelling strategy would be to create an RNN/LSTM that would learn the sequential, long-term pattern of the time-series. Those neural nets are designed to recognize sequential patterns, so they would be a great application for this problem. However, they are very data-hungry, and ~550 data points won’t suffice— we need orders of magnitude more to adequately train a neural net.

Furthermore, assuming we could get more data, it would be very interesting to see if we could create a reinforcement learning algorithm to predict future states of food security for a given country.

Additionally, future work could include transfer learning from pre-trained image neural nets on geospatial data to determine which regions are at higher risk based on current physical state of the land.

In sum, food security is an inherently difficult problem to solve and understand in and of itself. Adding in the task of organizing, cleaning, and analyzing data on human needs from a litany of sources is a daunting, yet vital, endeavor.

A beneficial initiative not only for data engineers but for all community development practitioners would be a centralized (or open-sourced?) organization/procedure for collecting data in mutually agreed upon way. As such, we could avoid many of the problems that initiatives such as Project 8 are seeking to alleviate.

References

1. (2017) The United Nations Website. [Online]. Available: <http://www.un.org/sustainabledevelopment/sustainable-development-goals/>
2. (2017) The Project 8 Website. [Online]. Available: <http://demandinstitute.org/projects/project-8/>
3. (2017 Victor Vulovic’s Github Capstone Repo. [Online]. Available: https://github.com/vsquared10/capstone