Machine Learning Engineer Nanodegree

Capstone Proposal

Sunil Nayak
Feb 03, 2018

Proposal

I intend to study the yield of crops in some of the countries of the world for which I have sufficient data and find out what factors influence the yield the most. The intent is to build a regression model to predict the yield and find out the factors that influence the yield significantly. The crop yield could vary considerably from region to region even within a country for various reasons. This study does not look at data at a granular level of small regions within countries. The data used with be average for each country and thus cannot be used to estimate the yield in small regions.

Domain Background

United nations study says that "The world needs to produce at least 50% more food to feed 9 billion people by 2050… Unless we change how we grow our food and manage our natural capital, food security—especially for the world's poorest—will be at risk.

http://www.un.org/en/sections/issues-depth/food/

Eradication of poverty and hunger by 203, are sustainable development goals agreed by UN members.
http://www.un.org/sustainabledevelopment/sustainable-development-goals/

Food security can be achieved by, increasing yield of agricultural crops, employing more land for agriculture, increasing the production of meat. It is agreed that meat is not an efficient means of food source as it takes more resources to farm animals for food consumption compared to agriculture.

Problem Statement

Can we build a model that predicts the yield of crops based relevant data like farm size, amount of fertilizer used, water used, labor used etc. Would the model reveal the significant factors that impact the yield of crops? The model should predict the yield with R-squared value of 70% or higher on test set.

Datasets and Inputs

The model should be generic enough to predict the yield based on data that is not region specific. So any notion of geographic region or wealth (e.g. GDP) of the country should not be an input to the model. Some of the inputs like farm size, labor used or farm equipment used may indirectly imply a certain (wealth) affordability of the country.

Food and Agriculture Organization of United Nations provides agricultural data by country:

http://www.fao.org/faostat/en/#data

From this website we will use data for crop yield, water used, fertilizer used, pesticide used, soil erosion and degradation. Most of the data is provided by country, year and per hector or land.

Nation Master is another website that provides data sourced from CIA World Factbook (https://www.cia.gov/library/publications/the-world-factbook/)

http://www.nationmaster.com/country-info/groups/High-income-OECD-countries/Agriculture

From here we use the data for region and income category a country belongs to, farm worker count, tractor count. This data is provided by country, year and per hectare of land.

Solution Statement

In order to predict the yield of crops, I will pick 3 crops of interest, wheat, rice paddy and onion. The models will be built for each of these crops separately. These crops could be used as categorical variables and a single model could be build but in order to understand the model better, I will build one model for each crop. I will use multiple linear regression and random forest regression for this.

Yield for crop is represented in hectograms (100 grams) per hectare in the FAO data. The independent variables (water used, fertilizer used, workers employed, tractors used) are available per hectare of land as well. The model (multiple linear regression or random forest regression) will try to predict the yield for given values of independent variables.

I will use data for most recent 10 years available (2002 – 2015) in order to avoid lot of variability. I could reduce the data further and consider most recent 5 years of data (not decided).

Benchmark Model

There are many studies done on this problem.

This research (http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0156571) studies the yield of wheat and maize globally and in two specific wheat producing regions of USA. This study also uses climate data (temperature, precipitation etc.) as predictors for the yield of wheat and maize. The random forest regression model achieves a Root Mean Squared Error of 4% - 14% of average yield. The multiple linear regression model on the achieved RMSE of 14% to 49%.

There is lack of climate data available from the data sources mentioned above. I will not be using temperature and precipitation data in my study.

Evaluation Metrics

The performance of the model will be evaluated by these methods:

1. *RMSE: Root mean squared error of yield*
   RMSE = square_root[sum((predicted value– observed value)^2) / n]
   Where n is the number of observations
2. Coefficient of determination (R squared value): the ratio of explained variation to total variation of yield:
   R-squared = 1 – [SS_residual / SS_total)
   SS_residual: sum[(predicted value – observed value)^2]
   SS_total: sum[(observed value – mean(observed values))^2]

3. Graphs of estimated yield vs actual yield

Project Design

The data for this project is available in different files and from different sources.

Data consolidation and cleanup (something the course did not cover in detail):

Combine the data from the different sources in csv format.

Evaluate missing data and throw away columns for which we do not have sufficient data.

Eliminate countries for which we do not have sufficient data.

Eliminate outliers

Fill missing data with nearest value or average value (whichever appropriate – not sure yet).

Convert any categorical variable to one-hot encoded values.

Data Visualization:

Frequency distribution of each of the predictors and target variable (yield).

Scatter plot of each of the predictors vs target values to see if there is any correlation.

Scatter plot of pairs of predictors so see if predictors are independent.

Model Building:

Split data into training and test set randomly.

Train the model: Build models MLR and RF for the training data.

Test the model: Generate evaluation matrix values on test data.

Find significant predictors

Find ways to group the countries by similarity in yield and see if the countries are related in any way (by region or economic standing).