DAT102x: Predicting Chronic Hunger

Marko Peltojoki

October 2018

Executive Summary

- This report presents analysis of data concerning economic, social, and political factors that are indicative of trends in chronic hunger in countries around the world. The goal was to predict the annual **prevalence of undernourishment** at the country level from the socioeconomic indicators. The prevalence of undernourishment expresses "the probability that a randomly selected individual from the population consumes an amount of calories that is insufficient to cover her/his energy requirement for an active and healthy life" (FAOSTAT). It can be understood as the **percent of the total population that is facing chronic hunger**. Data is compiled from the Food and Agricultural Organization of the United Nations as well as the World Bank.
- There are 45 variables in the dataset. Each row in the dataset represents a country in a given year.
- After exploring the data by creating visualizations of the data, several potential relationships between socioeconomic indicators and prevalence of undernourishment were identified. After further analysis only three variables were selected.

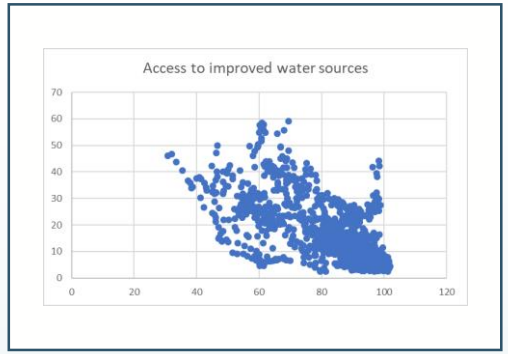
Executive Summary

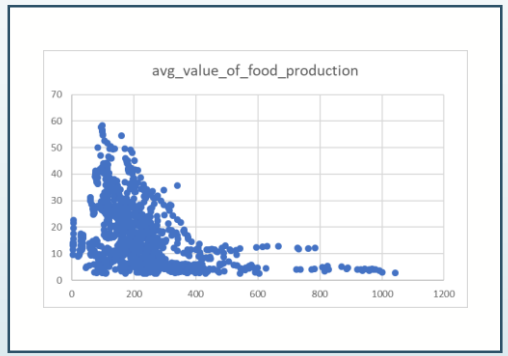
- While many factors can help indicate the prevalence of undernourishment, the most significant features found during the analysis were:

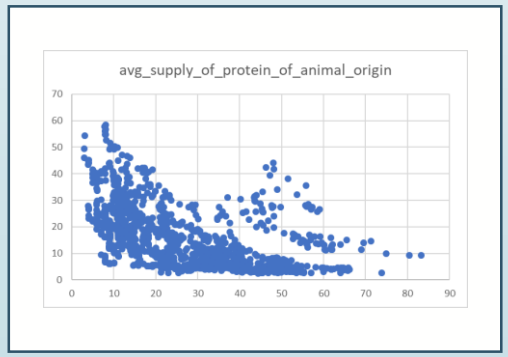
- Average value of food production - estimated food net production value of a country expressed in per capita terms, measured in constant 2004-06 international dollars per person. Lower food production seems to increase the risk of the prevalence of undernourishment.
- Average supply of protein of animal origin - average protein supply expressed in grams per capita per day including protein from meat, milk, eggs, fish, seafood, and other animal products. Lower supply of protein seems to increase the risk of the prevalence of undernourishment.
- Access to improved water sources - percent of the population with reasonable access to an adequate amount of water from an improved source, such as a household connection, public standpipe, borehole, protected well or spring, and rainwater collection. Higher access to water sources seems to decrease the risk of the prevalence of undernourishment.

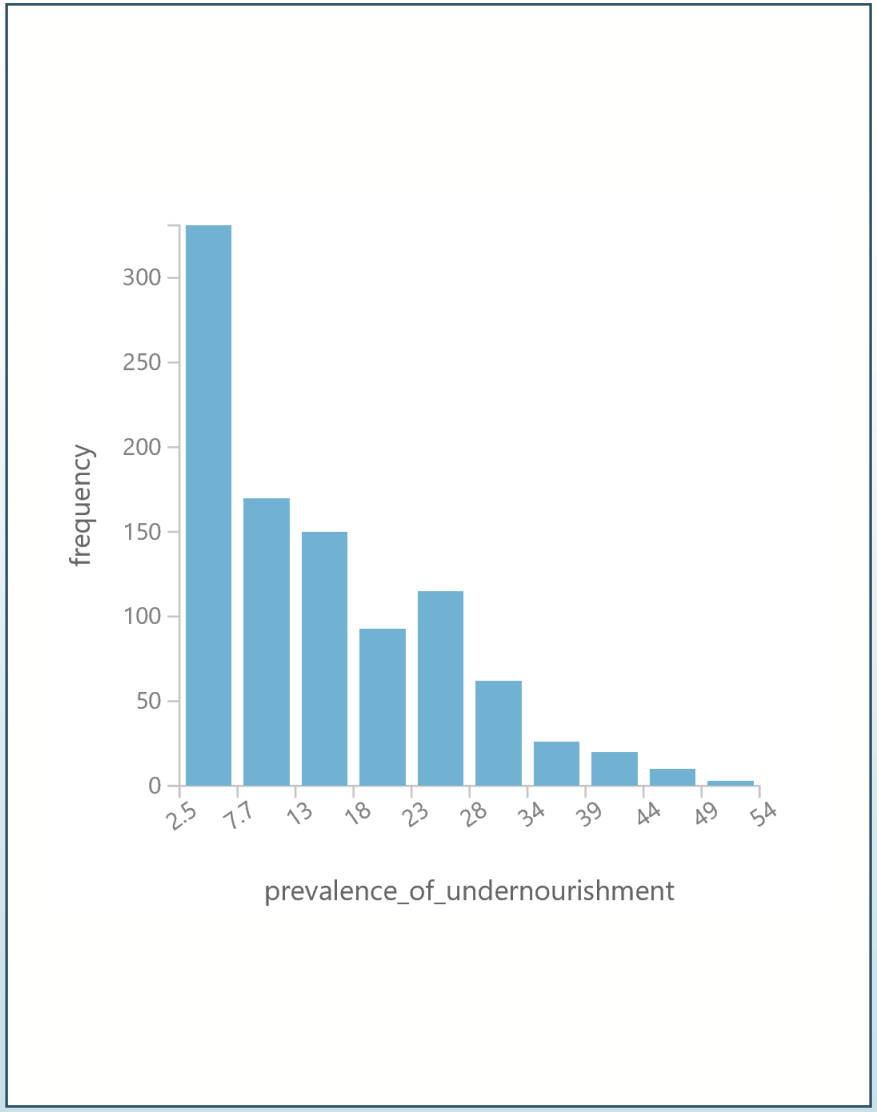
Initial Data Exploration

- The initial exploration of data began with creating scatter plots for each variable to find possible correlations between the prevalence of undernourishment and each variable. The most promising correlations are shown here.
- The histogram of the prevalence of undernourishment shows that it is right skewed, so there is low amount of high risk countries.
- The scatter plots show that the indicative value of each variable predicting the high risk countries has huge variance to predict the prevalence of undernourishment. So it is easier to predict low risk countries than high risk countries. In that perspective some data normalization and manipulation might be required.









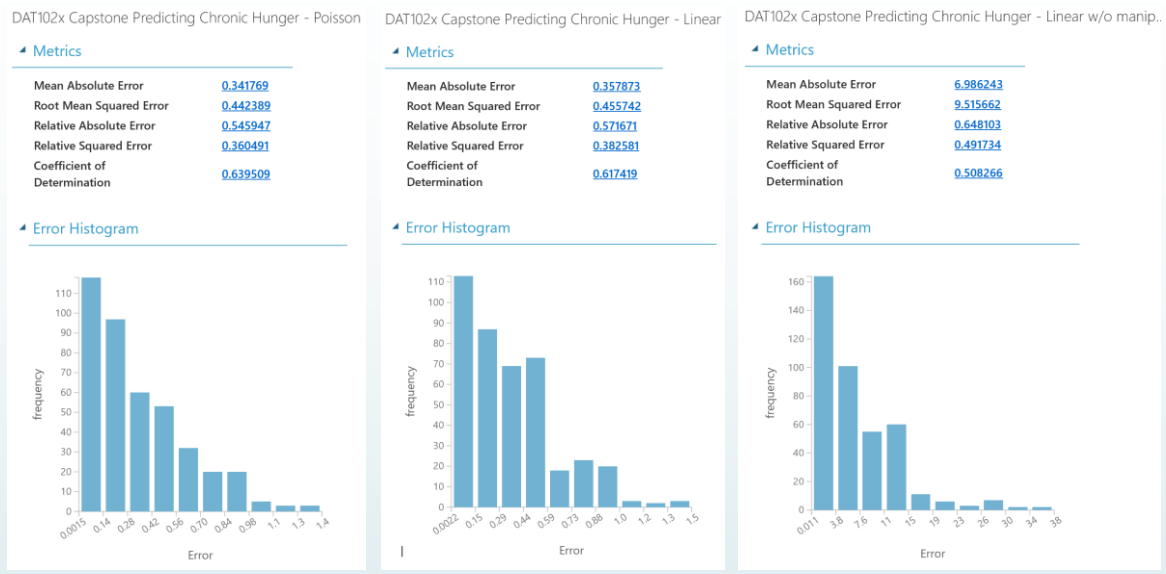
Data Cleancing

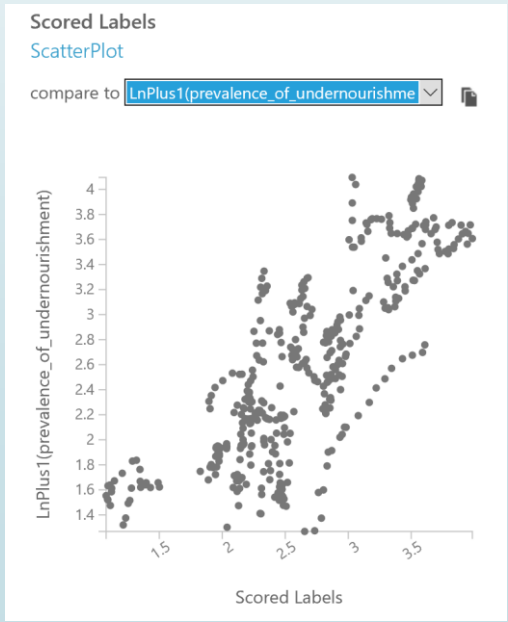
- After selecting the most promising variables it was time to look into the data purity.
- Sorting data by country and by year it was revealed that some variables were missing two or three values per country. As year over year variance was lower than the average value of all the years for spesific country the "nearest neighbour" method was chosen to fill in the missing values for that country, as highlighted in the table.

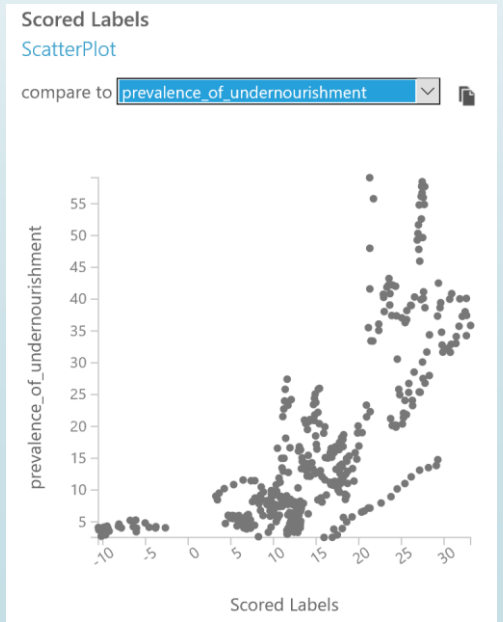
country	year	avg_val	avg_sup	access_
0593aa0	2000	176,2044	37,07125	96,59155
0593aa0	2001	177,4436	40,2555	97,88043
0593aa0	2002	174,2014	43,55124	98,35835
0593aa0	2003	174,0413	44,94913	97,06623
0593aa0	2004	151,4249	47,67435	97,56077
0593aa0	2005	133,3734	48,02044	94,73582
0593aa0	2006	115,6618	49,81632	96,31108
0593aa0	2007	120,8588	47,8116	95,31453
0593aa0	2008	121,8311	45,4513	95,44189
0593aa0	2009	124,0202	45,64414	95,11388
0593aa0	2010	124,5648	43,79474	97,1731
0593aa0	2011	127,5043	40,69305	98,3067
0593aa0	2012	127,0359	41,5754	95,32762
0593aa0	2013	130,1324	41,5754	95,99705
0593aa0	2014	130,1324	41,5754	97,04038
0593aa0	2015	130,1324	41,5754	97,89424
066b021	2000	91,80607	7,920968	58,56932
066b021	2001	98,40674	8,961027	61,55148
066b021	2002	101,5759	8,92678	62,23402
066b021	2003	108,3151	8,964986	63,33412
066b021	2004	112,15	9,095709	62,91919
066b021	2005	113,028	9,814236	64,09073
066b021	2006	112,5337	9,913674	66,8074
066b021	2007	114,3911	9,884698	66,04068
066b021	2008	116,2181	10,91883	66,28963
066b021	2009	119,6054	11,11272	69,81044
066b021	2010	127,2202	11,21947	69,34462
066b021	2011	125,5197	13,97778	69,39612
066b021	2012	134,651	17,06552	72,76308
066b021	2013	139,8811	17,06552	72,06329
066b021	2014	139,8811	17,06552	73,70716
066b021	2015	139,8811	17,06552	72,49132

Regression

- Several regression models like Poisson, Linear and Neural Network regression were tried out with data normalization (MinMax) and manipulation to logarithmic scale (LnPlus1), the most promising results were provided by Poisson regression. The model was trained with 70% of data, tested with 30% of data.
- After that trying without data manipulations the Linear regression model proved to provide best scores in the competition, to my surprise, eventhough the Coefficient of Determination was the lowest one.







Conclusion

- The analysis show that with as few as three variables, such as Average value of food production, Average supply of protein of animal origin and Access to improved water sources, you may predict the prevalence of undernourishment, eventhough the relationship may not be tightly linear and may have some larger variance between predicted and actual values.