Prediction Of Crop Extinction Using Environment Stress Data

Technical Summary

Data Science, BrainStation

Roshini Sreenivas

2021-12-05

INTRODUCTION

Problem space: The model forecasts the extinction of crops in nature and the input is environmental stress data. For demonstration purposes, I have chosen three parameters: carbon-dioxide and other greenhouse gas emissions, soil temperature and sea-level rise.

A minimum viable population, ecologists agree, is the point at which we expect these species to go extinct once they've reached these numbers. So, this is when scientists would start harvesting biological matter like sperm samples, oocytes etc to possibly bring the species back in the future. My hypothesis is that applying the principles of data science is a good way to go about an MVP analysis in this case, because a solid prediction can be made using data points like production units (yield), the area harvested, the temperature and air quality; essentially factors that influence how well a crop does in its natural environment.

The goal is to be able to actually predict when a crop goes extinct in the wild so we can pre-emptively stave this off. I also believe that it will help frame better policies around use of crop-land and diversifying the genetic variations of the crop, or even harvesting seeds for a seed bank so we can bring it back under more favorable circumstances.

The data has been collected from multiple sources.

Crop data : <https://www.kaggle.com/raghavramasamy/crop-statistics-fao-all-countries> The crops data is a data set that looks at two production parameters (units and land in hectares) for all the countries that are member countries of the FAO (Food and Agriculture Organisation) from the year 1961 to 2019.

Global surface temperature data : <https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data?ref=hackernoon.com> This is the average global surface temperature dataset, as granular as city level, runs from 1743-2013.

Average sea level rise data: <https://www.kaggle.com/somesh24/sea-level-change> This is the average global sea level rise through the years 1800-2015.

Air Quality data:< https://www.kaggle.com/yoannboyere/co2-ghg-emissionsdata> This dataset looks at air quality data from 1750-2017.

Supplementary air-quality data: <https://www.kaggle.com/unitednations/international-greenhouse-gas-emissions?ref=hackernoon.com> I've only used parts of this dataset for comparison.

PREPROCESSING, EDA AND FEATURE-ENGINEERING

One of the biggest challenges was actually deciding on the granularity of the data set and merging all the data sets into one.

Then, I decided, in the interest of a more efficient process, that I would pick one crop to actually work with: Bananas. The reason I chose bananas was because this particular crop has already taken a few hits in the past. In 1965, The Gros Michel banana officially went extinct, wiped out due to disease (Source) and the same is expected to happen to the Cavendish banana, due to lack of genetic diversity. I wanted to see if the data would reflect this isolated extinction event as well.

And the final reduced table I started working with looked like this:

	Year	Country	Item	Area Harvested	Production	Yield	CO2_GHG	GMSL	AverageTemperature
1440	1961	Algeria	Bananas	0.0	0.0	0.0	6055917.48	-38.091667	11.106
1441	1961	Algeria	Bananas	0.0	0.0	0.0	6055917.48	-38.091667	14.798

I have two target variables, 'Area Harvested' and 'Yield'. Finally, after some basic analyses, I decided to not use Yield as a function of the Area Harvested and just use it as a feature in itself.

I also performed hypothesis testing and built co-relation matrices to determine how the different features are co-related. The co-relation between emissions data and yield was found to be statistically significant ($p<0.05$) and so was the relationship between Sea-level and Yield, but negatively.

For feature engineering, I converted my 'Country' column to One-Hot-Encoded columns. If I had been using the whole data set, I would have clustered them into continents. I can still do that here, but there aren't 190 countries as in the whole dataset and these are mostly all in the temperate zone because that is where Bananas grow best, so that's about three continents. I did most of my EDA in Tableau. I've attached the workbook to my supplementary documents.

The dashboard on Tableau is a great way to visually learn the story of the banana plant.

MODELING

Model	Scores
Time series model with best AIC score	47.75
Cross-validated Logistic Regression Model	0.55

My theory is that the concept makes sense, the model just does not have all the parameters that can be used to predict crop extinction with a high accuracy at the moment. In my opinion as a

subject matter expert as well, the model could use parameters like ambient temperature, humidity and precipitation to be more robust.

The point of building the logistic regression model was just to be able to see the co-efficients and verify the countries in terms of importance. But it also supplements the time series model in its prediction of a good yield.

At this point, I moved on to building a time series model, because ultimately, I want to be able to forecast when Yield is tending to zero. A total yield forecast predicted that bananas will do well (upward trends for yield) for the next three decades.

Finally, I built 3 time series forecast models for the top three banana producing countries in the world picked from EDA and the baseline model co-efficients: Mali, Costa Rica and Honduras. The forecast tends upwards in all cases, but the original yield graph has lots of variation. I did a deep dive into a literature survey and I have summarised all my findings below.

INSIGHTS AND CONCLUSIONS

- The input data is insufficient. The model would learn better with the help of additional environmental inputs like precipitation, humidity, macroecomic inputs like wages and the employment index and others like disease data. Most of these data sets will not be readily available for our use.
- The logistic regression models and the time series model support the other's results. The models point towards the Banana not going extinct for the next three decades at least, as the yield tends upwards.
- Neither model is very accurate because the only parameter they're using as statistically significant is emissions and higher carbon-dioxide in the air is good for plant growth and the model is hence predicting upward trends for Yield (Source).
- Emissions is closely co-related with Yield (with a p value <0.05). And this leads me to believe that other factors like precipitation and humidity will also be closely co-related.
- My intuition for the sharp downward trends in Yields in the country specific forecasts is that its not due to one factor. Each spike can be explained differently. Bananas have a very dark colonial past; the industry was built on the back of slavery. Even to this day, the banana production is affected by outstanding labor and environmental law suits against the giants that monopolise the banana crop market, like Monsanto, for example (Source). Sharp downward spikes can also be explained by isolated events like unfavorable weather or disease outbreaks (Source).

NEXT STEPS

The next steps for this project would involve collecting more input for the model, as mentioned previously. I would need to decide how these parameters would need to look for this particular use-case. And then fine tune the models for better accuracy.