ICDDS Report

Modest Lungociu, Tom Wiley, Dominik Rys

;)' DROP TABLE Teams;–

1 Introduction

This report introduces our approach to the Crop Yield Challenge presented at AI Hack 2021 hosted by Imperial College Data Science Society. The main datasets used for this were extracted from the State of Illinois, USA, and posed a great number of challenges because of the way they were structured. Our main approach was backed by a number of research papers on Recurrent Neural Networks Khaki et al. ((2020)) used in predicting crop yield production for various parts of the US, provided by the organisers of this event.

2 Purpose

Our purpose for this project was to create a consistent prediction of the crop yield for the year of 2020. This would help farmers and the agriculture department of Illinois prevent overspending and adopt a more robust approach when treating the soil after extreme weather events across the state Lobell et al. ((2013)). Followed by this, we were driven to create an LSTM model with Linear Regression to predict the annual yield and create a good visualisation of the evolution of the historical data and import them into a React based web application.

3 Methodology & Pre-Processing

One of the challenges posed by this task and the datasets provided was that for every year and county we had an abundance of features, but we only had access to 1 crop yield label. This would have made regression a very complicated task. However the best approach for this would be to only select a number of features (in our case we only selected soil vegetation and soil moisture) and create a sequential LSTM that would output the predicted yield compared to the actual yield. Two types of weather features in the datasets were important, temperature and accumulated precipitation. Because the maximum temperature and the minimum temperature determines the daily accumulated heat (GDD = (temperatureMax + temperatureMin) /2 - TemperatureBase). This is the key

to deciding how fast the crops will grow. Also, the minimum temperature is important for winter wheat because it needs vernalization (cold temperature) to inducing flowering. During pre-processing several unnecessary columns across the datasets were dropped. Due to data entry is on a monthly base and on a county base, features such as ther latitude, longitude, and time (specific time of day) are not important. In order to deal with the curse of dimensionality of this data and prevent great loss of information, we wanted to approach this by having a kernel function, however due to time constraints we were not able to implement this. When we conducted cross-validation, we encountered issues that were leading to unequal X-train and y-train. The solution was to add additional values to the missed values in that array to form a 2D array which matches all the 1D array dimensionality.

4 Results

We measured our results against the findings of the crop yield of each county in Illinois in 2020. Our accuracy score was 58% and for measuring our model error we used the root means squares error function and obtained a relatively high error of 34%.

5 Finished Product - Crop Yield Predictions.

To present our results we created a basic web-app to show historic crop yield per county, and our predicted 2020 crop yield. The app is built in React, and uses components from the Ant Design framework. The map visualisation was created using with the React Simple Maps package.

6 Conclusion & Findings

We found that the more features included in the model would slightly improve the accuracy. However, the "right" features included in the model can significantly improve predicting accuracy. For example, when we included latitude and longitude in the model, the accuracy score dropped to 0.1 because of the high complexity of understanding the square areas. While when we excluded the latitude and longitude alone, the accuracy score raises to 0.52. Another finding was that the soil type was very important in yield prediction because usually soil type combinations have been proven to be the most important factor which affected the crop yield Viña et al. ((2011)). In conclusion we observed that a more thoughtful approach that would include more features would easily perform better than our current approach and would have a higher success rate at helping the farmers understand their soil fertility better. We also had load of fun :)

References

S. Khaki, L. Wang, and S. V. Archontoulis. A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, 10:1750, 2020. ISSN 1664-462X. doi: 10.3389/fpls.2019.01750. URL https://www.frontiersin.org/article/10.3389/fpls.2019.01750.

D. B. Lobell, G. L. Hammer, G. McLean, C. Messina, M. J. Roberts, and W. Schlenker. The critical role of extreme heat for maize production in the United States. *Nature Climate Change*, 3(5):497–501, May 2013. ISSN 1758-6798. doi: 10.1038/nclimate1832. URL https://www.nature.com/articles/nclimate1832.

A. Viña, A. A. Gitelson, A. L. Nguy-Robertson, and Y. Peng. Comparison of different vegetation indices for the remote assessment of green leaf area index of crops. *Remote Sensing of Environment*, 115(12):3468–3478, 2011. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2011.08.010. URL https://www.sciencedirect.com/science/article/pii/S0034425711002926.