AI Hackathon 2021

Challenge: Crop Yield Challenge
Team: We are Dense

Name: Abu-Al Hassan, Danila Kurganov, Saleh Komies and Helen Li

Date: 21/02/2021

Members Background: Mathematics and Electrical Eng Departments

Crop yield estimation prior to harvesting period with an AI approach.

Abstract

Knowing an estimate of the yield of an agricultural product prior to the harvest period can help governments to introduce policies and farmers to stand a better position in trading with future contracts and prepare supply orders. We further aim to extract meaningful insights the data implicitly carries in with, which is later seen to mimic ‘industry knowledge’ yield info.

The overwhelming largest crop output in Illinois are corn and soybean. These crops have their planting season in April and harvest season in October and November. We aim to predict the yield of crops for a year in advance of the harvest time. Thus we have trained our model only on data between April-September.

Our model uses EVI [*1] (every 16 days) and 2m temperature[*2] (1st, 15th and 28th of every month) from the start of April until the end of September as well as latitude and longitude readings.

Conclusion

The data evaluation was open-ended, hence, considerable time was spent to evaluate which data is useful, and what insights would be meaningful to an end-user. We came to the conclusion that crop yield data would be most useful to an end user if the model could predict yield before the time at which crops were harvested. We therefore trained our model so that in hindsight, it can predict crop yield 2 months into the future. We believe this gives croppers ample times for harvesting logistics. Given our Illinois-centric data, our data analysis was based on crop-type information from this state. In production, our model can be deployed to give yield information, updated over time, from the start of planting to the harvesting phase. Model evaluation shows us little need for EVI data apart from sparse but highly-correlated forecasting in early June. This gives good forecasting in an economical way for the end-user.

Our model is based on a sequential-like ensemble between a NN and Random Forest. This provides robustness for future extreme weather events and yields predictions, as well as minimising errors. For production the ensemble is used.

Interpreting the Random Forest model gave us interesting data insights. We infer IL locations with the best yields, and how variance of temperature on specific days predicts yields. Notable insights, mimicking seed-planting best-practices, were the strong correlation between air temperature at the time when seeds were planted with overall annual yields, and generally stronger yield with higher temperatures (up to drought). One insight, not found in literature, was EVI correlation during specific times with yield. One interpretation is as follows: the greener the crop after 100 days, the better the yield.

Improvements could be done in the model making phase. This includes NN-embeddings, and using more historical data to predict yields. This will encode information such as crop-rotation and other possibly useful information. This would be less useful for the farmer, given their insight, but more useful for end-users unaware of farm-specific data.