Crop yield prediction using machine learning methods and satellite data

Ilya Chernyaev

Abstract

Retrieving accurate crop prediction data is an important matter of current interest both for government and business planning. Due to recent development in the realm of remote sensing and rapid advancements in the field of machine learning it has become possible to tackle this problem without expensive in-field observations. While the domain of crop prediction is dominated by deep learning approaches that use recurrent neural networks, this paper suggests treating geospatial data as tabular data and apply classic machine learning techniques to solve the problem of crop prediction.

1 Introduction

The problem of accurate crop prediction has been of crucial importance both for governments and business around the world in the recent decades. Knowing precise yield outcomes in advance makes a huge difference for market planning, logistics and food security around the globe. However, until recently this task demanded considerable investment of resources and required expensive in-field monitoring and sampling. Nowadays, fields can be monitored remotely using satellite data which is obtained and processed in a matter of seconds making crop prediction and monitoring much cheaper. A rapid development of free satellite imagery databases as well as latest advancements in machine learning techniques allow to build models that can predict crop yield with stunning accuracy even months before the harvest season.

Nowadays, the field of crop prediction is dominated by deep learning approaches which in most cases treat satellite data as collections of images with different bands and apply computer vision techniques to exploit complex structure of data and extract additional features. Those approaches give satisfying results, but require significant computational resources for image processing and lack the power of interpretation.

Nevertheless, satellite data can be viewed not only as a collection of multi-band images indexed by time and space coordinates. It is also possible to extract distributions of these images' bands and base the analysis on statistical properties of these distributions. Thus, remote sensing data can be converted to tabular format preserving time and spatial characteristics, but simplifying the analysis.

This paper relies heavily on the latter approach and applies the most advanced methods for tabular data analysis in order to obtain yield forecast.

2 Literature review

There exist a variety of methods used for yield forecast. One of the most comprehensive literature review in the field is given by Van Klompenburg et al. (2020) [1].

An approach based on analysing vegetation indices and employing statistical learning techniques are presented by Meng et al. (2019) [2] where high resolution images from MODIS are used to produce a new vegetation index for cotton yiled prediction in California. As for classical machine learning approaches, the work of Rodríguez et al. (2017) [3] take an attempt at predicting citrus orchards yield using data collected directly from the field during the growing season. In the realm of deep learning approaches there is much research that is based predominantly on the usage of classical convolutional neural architectures or LSTM (Long short term memory) neutral networks. The paper by You et al. (2017) [4] is a case in point.

3 Data review

The data selected for this research is *corn* and *soybean* production reports in the US over the time span of 20 years (2001 - 2021). These particular agricultural crops have been selected because of their importance for the agricultural system of the US: namely, they happen to be the largest crops in terms of total production and planting territory. The choice of the time span is due to the availability of appropriate geospatial data. Finally, the states were selected such that most of their crop territories are covered by the chosen crops so that planting and harvest data is fully available for these territories.

3.1 Yield data

The exact yield numbers as well as the size of the planting areas in each state were obtained from USDA (U.S. Department of Agriculture). Yearly records for both corn and soybeans have been obtained via USDA database system.

3.2 State boundaries and crop mask

State boundaries were obtained from The United States Census Bureau TIGER dataset [5] that contains the 2018 boundaries for the primary governmental divisions of the United States (see Fig.1). Crop layer has been obtained from Cropland Data Layer (CDL) [6] created yearly by USDA using moderate resolution satellite imagery and extensive agricultural ground truth (see Fig.2).

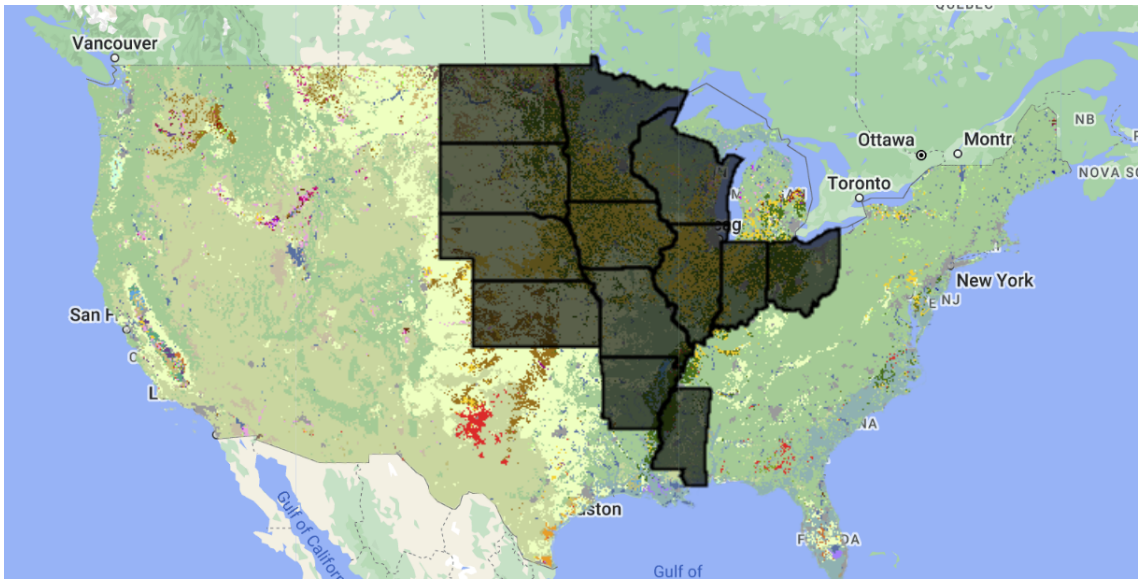

Figure 1: State Boundaries: 13 States that were Selected for Examination

Figure 2: Cropland Data Layer (detailed)

3.3 Vegetation indices and weather data

In this study three main datasets with respective indicators were used, namely MODIS Combined 16-Day NDVI dataset [7] (see the details below), MODIS Terra Land Surface Temperature and Emissivity 8-Day Global 1km dataset [8] and ERA5-Land Monthly Averaged - ECMWF Climate Reanalysis [9] for monthly precipitation data.

Temperature is recorded in Kelvins and total monthly precipitation level is reported in meters. While the meaning of temperature and precipitation data is easy to grasp, NDVI may require some explanation. This acronym stands for Normalized difference vegetation index which is widely used in agricultural applications since it quantifies vegetation by measuring the difference between near-infrared (which vegetation strongly reflects) and red light (which vegetation absorbs). It is calculated according to the following formula (NIR stands for near infra red band):

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

3.4 Data preprocessing

As it can be seen in Fig. 1, corn and soybean crops (yellow and green colors respectively) are predominant in the selected state. Other crops are dispersed over different states which is why a similar analysis for them is associated with a number of problems. Within each given state 500 data-points that lie one the crop mask were sampled (see Fig. 3) in order to reduce computational difficulty. Those data points are used to calculate aggregate indices.

Datasets in Google Earth Engine are represented as Image Collections, where each image captures the data from satellite bands for a specific period. For instance, 16-day aggregate NDVI means that NDVI values for 16 days were aggregated in a singe Image in the Image Collection. Thus, for each Image one can calculate statistics (mean, median etc.) of the band values. Such a statistic represents an aggregation over a period of 16 days and the region of interest. Within one year there are approximately 23 16-day periods which gives times series data with 23 points in each year, for each region and crop. An example of such a time-series data for NDVI values in the state of Iowa can be seen in Fig.5

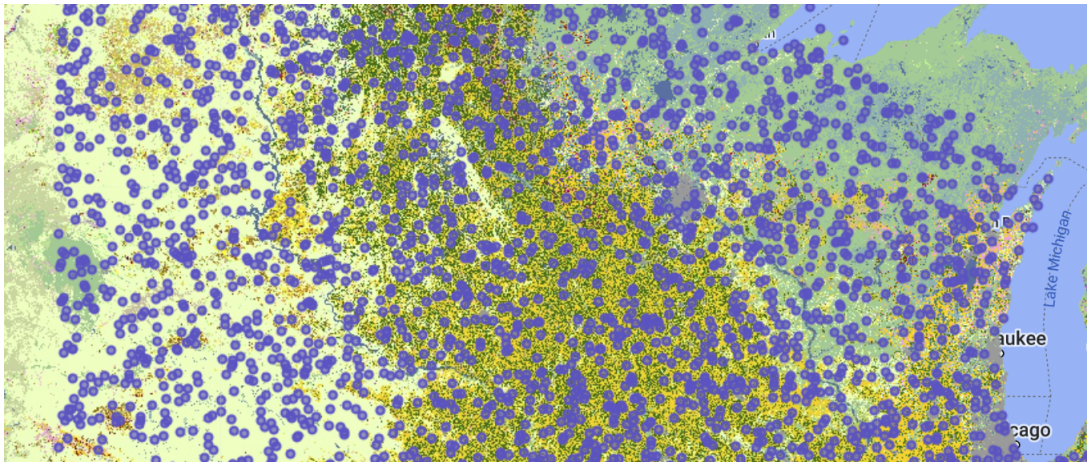

Figure 3: Sampling Crop Fields Within Each State

Now, the structure of the collected dataset becomes more clear. For a single crop yield output we

have the following features:

- Categorical variables of state
- Categorical variable of crop type (corn / soybeans)
- 23 values of mean NDVI (each 16 days)
- 12 values of monthly average precipitation
- 46 values of weekly surface temperature average

In total, the collected dataset contains data for 13 states, 20 years, 2 crops which gives 520 observations and 83 features.

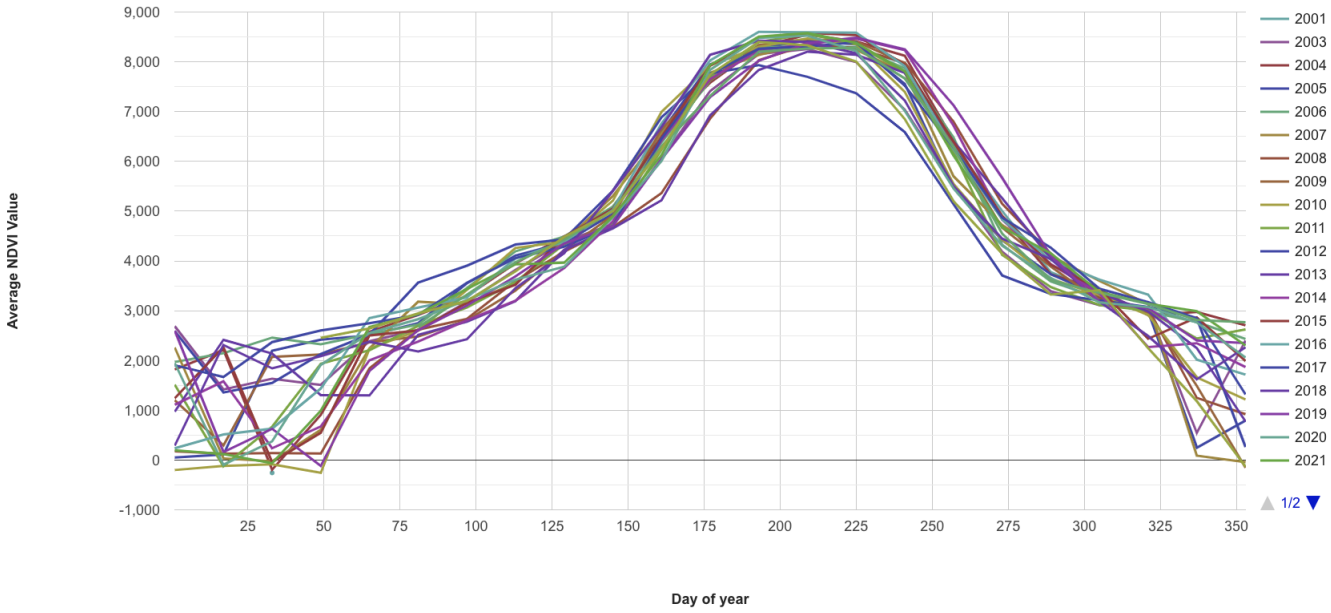

Figure 4: Average NDVI values for the state of Iowa in different years

3.5 Note on the structure of data

It is worth mentioning that data has seasonal character. The crops selected have almost the same planting and harvest seasons which makes them handy for our analysis. However, this means that we cannot use the data for the whole year for our prediction. Hence, we adopt three different time thresholds and build models for each of the thresholds. In the first case we use data only up to March (even before the planting season), in the second one we use data up to June and in the third one up to September (the beginning of harvest season). The dataset is then split for train and test in proportion 80 / 20 leaving last four years for test.

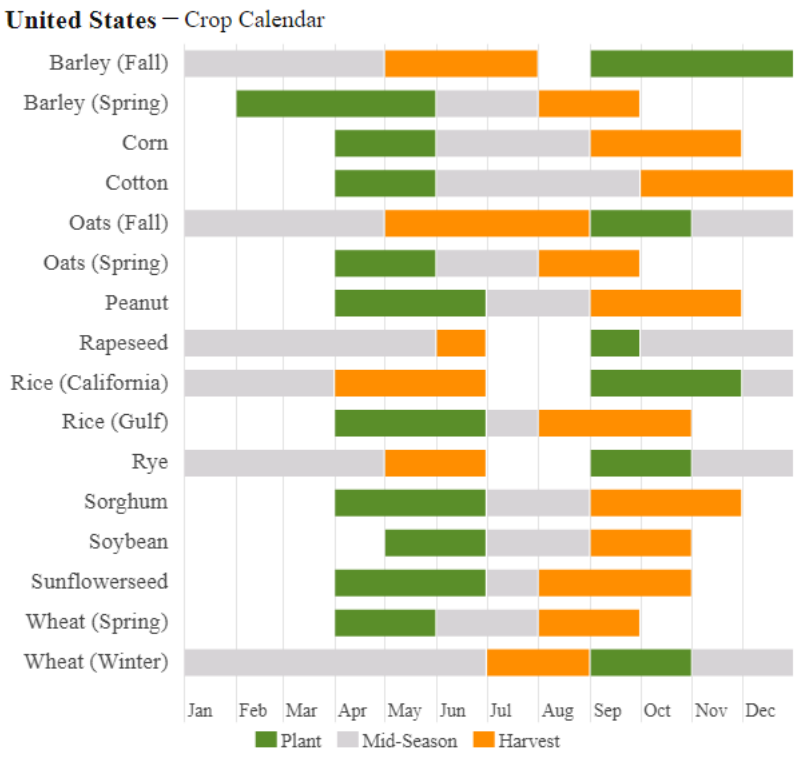

Figure 5: Crop calendar published by USDA for crops in US

4 Machine learning models and metrics

Since the final dataset is presented as a tabular data, machine learning techniques can be applied to tackle the problem. In this research there have been tested three different models:

- Linear regression (Sklearn implementation)
- Decision tree regressor (Sklearn implementation)
- Gradient boosting (CatBoost implementation)

All the models were trained and tested on different time spans of the dataset (see sec. 3.5 for detail).

It should be noted that we do not expect any convincingly good results from linear regression and it was considered purely as the most simple model that can be of use when prediction of continuous variable is in question.

4.1 Hyper-parameters tuning

Decision tree regressor as well as GBDT (Gradient boosting over decision trees) can give decent results using default learning parameters. However, hyper-parameters tuning can increase the perfor-

mance of these algorithms drastically and reduce over-fitting to the training data, especially in the case of decision trees. That is why it was decided to run cross validation grid search over selected parameters for the two later models.

For decision tree regressor the following hyper-parameters were tuned before training and prediction:

- criterion
- min_samples_split
- max_depth
- min_samples_leaf
- max_leaf_nodes

For CatBoost GBDT the following hyper-parameters were tuned:

- iterations
- learning_rate
- depth
- l2_leaf_reg

4.2 Performance Metrics

Since we have the problem of continuous variable prediction two metrics have been chosen for evaluating the quality of the model. To measure the overall error Root Mean Square Error (RMSE) was chosen. It shows by how much on average the prediction deviates from the actual data and gives the error in terms of initial units of target variable which makes it a good metric for increasing interpretability of the outcomes:

$$RMSE(\hat{y}, y) = \sqrt{MSE(\hat{y}, y)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Usually, one may be worried that RMSE is sensitive to outliers, however in this case there is no outliers in the data, hence this metric can be meaningfully used.

Another metrics chosen is coefficient of determination (R^2) which is used to measure the variability in the target variable that is explained by the model:

$$R^2(\hat{y}, y) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

5 Results

In the course of the research three different models have been trained over three different time spans of data. The resulsts are as follows:

	Linear regression		**Decision tree**		**Gradient boosting**	
Time span	*RMSE*	*R2*	*RMSE*	*R2*	*RMSE*	*R2*
Jan - Mar	30.18	0.77	16.28	0.93	12.12	0.96
Jan - June	37.54	0.64	16.96	0.93	11.38	0.97
Jan - Aug	36.38	0.67	13.28	0.96	11.94	0.96

As it can be seen, gradient boosting in all cases gives the best result with highest R^2 equaling 0.97 and the lowest RMSE equaling 11.38. Oddly enough the model that has the longest data time span of all is not the best in terms of performance metrics. For instance, linear regression performs the best with the smallest span of data. One possible explanation lies in the feature importance that can be derived after training. It turns out that the most important features for all the algorithms are crop_type, state and area_planted. It seems to be quite reasonable. These features are available at the beginning of the planting season and have strong correlation with target variable (it would be strange to expect low values of yield in the states with considerable planting territories). Thus, any additional data may hinder linear regression from giving better results.

It is also worth interpreting RMSE values. On average we have a mistake of 11.38 in best case. Initially crop yield data is measured in bushed per acre. On average across the states and years we have 153 bu per acre of corn yield and 44 of soybeans yield. That means that on average we make a 7% mistake when predicting corn yield and a 25% mistake when predicting soybeans yield.

6 Further research

There exist numerous possibilities for further research in the field. Firstly, one could gather more variables for analysis: use soil quality data before planting, wind conditions and atmospheric parameters. There features might improve the prediction quality of the model. Secondly, it is possible to include more territories or crops in the analysis. This approach, however, have some limitations to it, videlicet in this case crop yields will almost surely have different units of measurements which makes it more difficult to interpret the obtained metrics. Thirdly, external validity of the model might be tested so as to find out whether its predictive power holds for different countries and climatic tendencies. This approach will require collecting additional data. Finally, more advanced methods of prediction such as LSTM neural networks could be used with another way of data representation for the problem.

References

[1] Van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture 177, 105709.

[2] Meng, L., Liu, H., Zhang, X., Ren, C., Ustin, S., Qiu, Z., Xu, M., Guo, D., 2019. Assessment of the effectiveness of spatiotemporal fusion of multi- source satellite images for cotton yield estimation. Computers and Electronics in Agriculture 162, 44–52.

[3] Rodríguez, S.I.D., Mazza, S.M., AÍvarez, E.F.C., Giménez, L.I., Gaiad, J.E., 2017. Machine learning applied to the prediction of citrus production. Spanish journal of agricultural research 15, 7.

[4] You, J., Li, X., Low, M., Lobell, D., Ermon, S., 2017. Deep Gaussian process for crop yield prediction based on remote sensing data, in: Thirty-First AAAI conference on artificial intelligence.

[5] U.S. Census Bureau (2018). TIGER: US Census States 2018. Retrieved from [https://developers.google.com/earth-engine/datasets/catalog/TIGER_2018_States]

[6] USDA National Agricultural Statistics Service Cropland Data Layer. 2021. Published crop-specific data layer [Online]. Available at https://nassgeodata.gmu.edu/CropScape/ (accessed 05.01.2023; verified 05.01.2023). USDA-NASS, Washington, DC.

[7] LP DAAC. MODIS Combined 16-Day NDVI. Retrieved from [https://developers.google.com/earth-engine/datasets/catalog/MODIS_MCD43A4_006_NDVI]

[8] LP DAAC. MOD11A2.061 Terra Land Surface Temperature and Emissivity 8-Day Global 1km. Retrieved from [https://developers.google.com/earth-engine/datasets/catalog/MODIS_061_MOD11A2]

[9] Muñoz Sabater, J., (2019): ERA5-Land monthly averaged data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (2023), doi:10.24381/cds.68d2bb30