**Predicting Corn, Wheat and Soybean Yield**

Maura M. Tokay

DATA606 – Delivery-4

Dr. Ergun Şimşek

Problem Definition

Precipitation and temperature are the two main characteristics of weather and climate, they also play an important role on the outcome of crop yield. In this study the goal is to try to predict crop yield for corn, wheat and soybeans using meteorological data using machine learning predictive models like Lasso, Decision Tree Regressor, Random Forest Regressor and Neural Networks. A feature analyses will also be conducted to understand the importance of each meteorological parameters as an outcome of the model. It will be interesting to verify if the feature importance will be the same for each crop or not. Once we can quantify the importance of weather parameters listed below for crop yield we can relate the impact of climate change on food security.

Literature Review

Deep Neural Networks (DNN) was used to predict crop yield in the study conducted by Khaky and Wang [1] in response to the 2018 Syngenta Crop Challenge. DNN outperformed Lasso, Shallow Neural Networks (SNN) and Regression Tree. They also found that environmental factors (weather) had greater effect on crop yield than genotype and soil composition. For this study, genotype and soil composition is not a variable because all measurements were made with the same soil condition and seed variety. The goal is to determine the importance of each weather variable on crop yield outcome. Crane-Droesch [2] also used DNN to predict crop yield and climate change impact assessment in agriculture. In his study he included growing degree days (GDD) as a feature in his model which ended up having a great impact on crop yield prediction. GDD "are used to estimate the growth and development of plants and insects during the growing season. The basic concept is that development will only occur if the temperature exceeds some minimum development threshold, or base temperature (TBASE). The base temperatures are determined experimentally and are different for each organism" [3]. The base temperature is 10°C for corn and soybean and 4.4°C for wheat.

GDD = 1

Where:

= maximum temperature (°C)

= minimum temperature (°C)

= base temperature (°C)

Exploratory Data Analyses

This dataset is part of the Farming System Project (<https://www.ars.usda.gov/northeast-area/beltsville-md-barc/beltsville-agricultural-research-center/sustainable-agricultural-systems-laboratory/docs/farming-systems-project/>) at USDA, Beltsville MD. This data is not available online on the USDA website but can be found on my GitHub (<https://github.com/mmtokay/DATA606/tree/master/datasets>).

The data available for analyses span 20 years, 1996-2016. The meteorological data was measured daily on the same site where crops were cultivated. The data is split in two files, one that contains crop information and other with weather data.

The data is split in two files, one that contains crop information and other with daily weather data. The crop file contains the following information: crop, growing season, system name, grain yield, planting date and harvest date. The weather file contains year, Julian day, month, day, date, average temperature, maximum temperature, minimum temperature, maximum humidity, minimum humidity, average radiation, mean wind and precipitation.

After analyses, these are some finds about the data:

1. There is no crop data for 1999, this year Maryland had a drought and because the project did not use irrigation, crops never matured.
2. Wheat does not have crop data for 1996, 1999, 2003, 2004, 2007, 2010.
3. Data was separated by crops: 390 labeled data for corn, 500 labeled data for soybean and 223 labeled data for wheat.
4. Average radiation will not be used because data is missing for years 2003-2008.

Feature Engineering

Data transformation took place to apply machine learning predictive models, Lasso, Decision Tree Regressor, Random Forest Regressor and Neural Networks. Weather information was used to predict yield, and the weather data was grouped by week. The week duration parameter was created subtracting planting date from harvest date. The minimum number of weeks for each crop was calculated to be used to predict yield. The minimum week duration for corn, soybean and wheat is 16, 15 and 31 weeks, respectively. Because of the high number of features created, it was decided to create other three files for each crop with one to three weeks less than the minimum number of weeks. Four functions were created to group weather data using average, maximum, minimum and summation. A function was also created to calculate growing degree days (Equation 1). Finally, a function was created to build the file with weather features and target data (crop yield).

Model Construction

Supervised learning algorithms were used, more specifically regression algorithms that are used to predict continuous numerical values, however before applying those algorithms a RobustScaler [4] was used to better handle the outliers. To do that the algorithm removes the median and scales the data using the 1st and 3rd quartile for each feature independently. The regression algorithms used were Lasso, Decision Tree Regressor, Random Forest Regressor using Scikit-learn and Neural Networks using TensorFlow.

Lasso regression performs L1 [regularization](https://www.statisticshowto.datasciencecentral.com/regularization/), which adds a penalty equal to the[absolute value](https://www.statisticshowto.datasciencecentral.com/integer/#abs)of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients. Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models [5].

Decision tree regression builds a classification model in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. It results in a tree with decision nodes and leaf nodes. A decision node has two or more branches each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree corresponds to the best predictor called root node.

Random Forest Regressor is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting [6].

Neural Network was used building a Sequential model using seven densely connected hidden layers, and an output layer that returns a single value. The optimizer used was RMSprop algorithm [7].

Model Evaluation

The model evaluation was done using Median Absolute Error (MAE) and Coefficient of Determination (R2). The median absolute error “is calculated by taking the median of all absolute differences between the target and the prediction” [8]. The coefficient of determination “is a statistic that will give some information about the [goodness of fit](https://en.wikipedia.org/wiki/Goodness_of_fit) of a model. In regression, the *R*2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An *R*2 of 1 indicates that the regression predictions perfectly fit the data” [9].

Model Results

To evaluate the best model these study uses a combination of accuracy measured by median absolute error and confidence level measured by coefficient of determination. Taking into consideration this combination, the best model to predict corn was Decision Tree using 13 weeks of weather data with a confidence level of 82.6% (Fig. 1) and for the remaining results the accuracy was between ±715.2 kg/ha (Fig. 2). Looking at figure 1, the coefficient of determination is the same when running decision tree for 13, 14, 15 and 16 weeks of weather data. This shows that the model does not give weight for variables in weeks 14, 15 and 16 and therefore these weather parameters did not contribute for corn growth significantly. The top five features for corn yield prediction using 13 weeks of weather data are maxTemp10, SystemNameType, minHum3, minTemp9 and minTemp7. It is interesting that temperature and humidity has more influence than precipitation on corn yield prediction.

# A screenshot of a cell phone Description automatically generated

Figure 1. Coefficient of determination for corn yield.

A screenshot of a cell phone

Description automatically generated

Figure 2. Median absolute percentage error for corn yield.

The best model for soybean was Random Forest Regressor with 400 estimator 30 depth using 15 weeks of weather data with a confidence level of 83% (Fig. 3) and for the remaining results the accuracy was between ±261kg/ha (Fig. 4). The top five features were minTemp12, maxTemp7, minTemp5, Precip10 and minTemp1. The temperature and precipitation had more influence on soybean yield prediction.

A screenshot of a cell phone

Description automatically generated

Figure 3. Coefficient of determination for soybean yield.

A screenshot of a cell phone

Description automatically generated

Figure 4. Median absolute percentage error for soybean yield.

The best model for wheat was Neural Networks with 7 hidden layers using RMSprop optimizer using 29 weeks of used weather data with a confidence level of 74.5% (Fig. 5) and for the remaining results the accuracy was between ±354.6kg/ha (Fig. 6).

A screenshot of a cell phone

Description automatically generated

Figure 5. Coefficient of determination for wheat yield.

A screenshot of a cell phone

Description automatically generated

Figure 6. Median absolute percentage error for wheat yield.

Conclusion

The overall performance of all the models was rated as good (74% to 83%) but it could be better if more data were available to train the models.

A model that fits all was not true for this study, the best crop yield model for each crop was unique. It is important to test more than one model when using machine learning algorithms.

# Other interesting finding was that relative humidity is more important than precipitation for crop prediction. The assumption that temperature would play an important role on crop prediction was confirmed.

Climate change will have an impact on crop yield. The increase in extreme weather is a well-known signal of climate change. The increase in temperature is often accompanied by drought and flooding conditions. The climate anomalies have significant impact on long-term climate trends. For a future study, a sensitivity analyses can be performed to understand the climate anomalies on yield prediction.

# 

# References

1. Khaki, Saaed, and Lizhi Wang. “Crop Yield Prediction Using Deep Neural Networks.” *Frontiers*, Frontiers, 26 Apr. 2019. Retrieved February 25, 2020 from https://[www.frontiersin.org/articles/10.3389/fpls.2019.00621/full](http://www.frontiersin.org/articles/10.3389/fpls.2019.00621/full).
2. Crane-Droesch, Andrew. “Machine learning methods for crop yield prediction and climate change impact assessment in agriculture.” *Environmental Research Letters, Vol 13, Num 11*, 26 Oct. 2018. Retrieved March 1, 2020 from <https://iopscience.iop.org/article/10.1088/1748-9326/aae159>
3. Explanation of Growing Degree Days, Midwestern Regional Climate Center. Retrieved March 1, 2020 from mrcc.illinois.edu/gismaps/info/gddinfo.htm.
4. Scikit-learn (2019). Sklearn.preprocessing.RobustScaler. Retrieved May 12, 2019 from

<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html>

1. StatisticsHowTo (2019). Lasso Regression: Simple Definition. Retrieved May 12, 2019 from <https://www.statisticshowto.datasciencecentral.com/lasso-regression/>
2. Scikit-learn (2019). Sklearn.ensemble.RandomForestRegressor. Retrieved May 12, 2019 from <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>
3. Tf.keras.optimizers.RMSprop TensorFlow Core v2.2.0. (n.d.). Retrieved May 09, 2020, from https://www.tensorflow.org/api\_docs/python/tf/keras/optimizers/RMSprop
4. 3.3. Metrics and scoring: Quantifying the quality of predictions. (n.d.). Retrieved May 08, 2020, from <https://scikit-learn.org/stable/modules/model_evaluation.html>
5. Coefficient of determination. (2020, April 29). Retrieved May 08, 2020, from https://en.wikipedia.org/wiki/Coefficient\_of\_determination