**Predicting Corn, Wheat and Soybean Yield**

Maura M. Tokay

DATA606 – Delivery-3

Dr. Ergun Simsek

Exploratory Data Analyses

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

Some data transformation took place in order to apply machine learning predictive models, Lasso, Decision Tree Regressor and Random Forest Regressor. 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 week duration for each crop was used to calculate the number of weeks to be used on models to predict yield. The minimum week duration for corn, soybean and wheat are respectively, 16, 15 and 31 weeks. Four functions were created to group weather data using average, maximum, minimum and summatory.

A function was also created to calculate growing degree days (GDD) that "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". [1] The base temperature for corn and soybean is 10°C and for wheat is 4.4°C.

GDD = 1

Where:

= maximum temperature (°C)

= minimum temperature (°C)

= base temperature (°C)

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 [2] 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, and Random Forest Regressor. The Python library used was Scikit-learn.

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 [3].

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 [4].

Tables 1-4 show the preliminary results for corn yield using 16, 15 and 14 weeks of weather data. Tables 5-8 show the preliminary results for soybean yield using 15, 14 and 13 weeks of weather data. Tables 9-12 show the preliminary results for wheat yield using 31, 30 and 29 weeks of weather data. The best result for corn is for Decision Tree Regressor model with customization for 15 weeks (table 4). The best result for soybean is for Random Forest Regressor model (table 7) and for 14 weeks wheat the best model was Lasso for 31 weeks (Table 9).

The next step is to do more work with Feature Engineering working with different number of weeks for each crop. I will also drop features with least weight to check if the this will impact model performance.

**Corn – Lasso**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Weeks** | **Mean Absolute Error** | **Mean Squared Error** | **Median Absolute Error** | **Coefficient of Determination** | **Feature Importance** |
| 16 | 929.647 | 1388279.047 | 801.015 | 0.825 | SystemNameType 2332.06  maxHum1 -2115.08  maxTemp10 -1391.40 |
| 15 | 928.892 | 1386607.393 | 794.308 | 0.825 | SystemNameType 2339.59  minHum11 2129.74  maxTemp10 -2103.13 |
| 14 | 928.687 | 1386386.616 | 782.638 | 0.825 | maxTemp10 -2593.89  SystemNameType 2342.50  minHum11 2330.45 |

Table 1. Corn Lasso regression results for 16, 15 and 14 weeks of data.

**Corn -** **Decision Tree Regressor**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Weeks** | **Mean Absolute Error** | **Mean Squared Error** | **Median Absolute Error** | **Coefficient of Determination** | **Feature Importance** |
| 16 | 930.406 | 1408752.936 | 715.187 | 0.826 | SystemNameType 2332.07  maxHum1 -2115.08  maxTemp10 -1391.40 |
| 15 | 930.406 | 1408752.936 | 715.187 | 0.826 | SystemNameType 2339.59  minHum11 2129.74  maxTemp10 -2103.13 |
| 14 | 930.406 | 1408752.936 | 715.187 | 0.826 | maxTemp10 -2593.90  SystemNameType 2342.50  minHum11 2330.45 |

Table 2. Corn Decision Tree Regressor results for 16, 15 and 14 weeks of data.

**Corn -** **Random Forest Regressor**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Weeks** | **Mean Absolute Error** | **Mean Squared Error** | **Median Absolute Error** | **Coefficient of Determination** | **Feature Importance** |
| 16 | 973.851 | 1515128.491 | 788.234 | 0.809 | maxTemp6 0.31  maxTemp9 0.11  maxTemp8 0.10 |
| 15 | 939.915 | 1431510.446 | 762.162 | 0.810 | maxTemp9 0.25  maxTemp6 0.22  maxTemp11 0.05 |
| 14 | 960.041 | 1488230.538 | 722.295 | 0.816 | maxTemp9 0.39  maxHum13 0.12  maxTemp4 0.06 |

Table 3. Corn Random Forest Regressor results for 16, 15 and 14 weeks of data.

**Corn -** **Random Forest Regressor** (**max\_depth=35, n\_est = 500)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Weeks** | **Mean Absolute Error** | **Mean Squared Error** | **Median Absolute Error** | **Coefficient of Determination** | **Feature Importance** |
| 16 | 928.115 | 1385379.136 | 772.609 | 0.826 | maxTemp9 0.23  maxTemp6 0.21  maxTemp8 0.05 |
| 15 | 925.964 | 1382744.199 | 772.609 | 0.826 | maxTemp9 0.24  maxTemp6 0.21  maxTemp8 0.05 |
| 14 | 948.281 | 1448582.602 | 767.763 | 0.817 | maxTemp9 0.40  maxHum13 0.11  maxTemp4 0.04 |

Table 4. Corn Random Forest Regressor (customized) results for 16, 15 and 14 weeks of data.

**Soybean - Lasso**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Weeks** | **Mean Absolute Error** | **Mean Squared Error** | **Median Absolute Error** | **Coefficient of Determination** | **Feature Importance** |
| 15 | 391.326 | 257187.416 | 288.727 | 0.801 | minHum13 931.61  maxTemp7 -848.34  Precip10 512.88 |
| 14 | 390.899 | 253732.244 | 287.818 | 0.803 | maxHum1 -927.86  minHum13 792.91  maxTemp7 -638.90 |
| 13 | 389.874 | 252001.475 | 294.098 | 0.804 | maxHum1 -1031.32  minHum13 887.65  maxTemp7 -673.73 |

Table 5. Soybean Lasso regression results for 15, 14 and 13 weeks of data.

**Soybean -** **Decision Tree Regressor**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Weeks** | **Mean Absolute Error** | **Mean Squared Error** | **Median Absolute Error** | **Coefficient of Determination** | **Feature Importance** |
| 15 | 362.513 | 226064.051 | 267.078 | 0.832 | minHum13 931.61  maxTemp7 -848.34  Precip10 512.88 |
| 14 | 365.446 | 234428.064 | 262.750 | 0.838 | maxHum1 -927.86  minHum13 792.91  maxTemp7 -638.90 |
| 13 | 361.587 | 225529.121 | 262.750 | 0.832 | maxHum1 -1031.32  minHum13 887.65  maxTemp7 -673.73 |

Table 6. Soybean Decision Tree Regressor results for 15, 14 and 13 weeks of data.

**Soybean -** **Random Forest Regressor**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Weeks** | **Mean Absolute Error** | **Mean Squared Error** | **Median Absolute Error** | **Coefficient of Determination** | **Feature Importance** |
| 15 | 371.249 | 242497.114 | 270.451 | 0.814 | minTemp12 0.47  maxTemp7 0.13  minTemp5 0.06 |
| 14 | 358.551 | 224661.315 | 284.613 | 0.835 | minTemp12 0.47  maxTemp7 0.10  minTemp5 0.06 |
| 13 | 366.674 | 234401.733 | 252.278 | 0.826 | minTemp12 0.48  minTemp5 0.11  maxTemp7 0.11 |

Table 7. Soybean Random Forest Regressor results for 15, 14 and 13 weeks of data.

**Soybean -** **Random Forest Regressor** (**max\_depth=35, n\_est = 500)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Weeks** | **Mean Absolute Error** | **Mean Squared Error** | **Median Absolute Error** | **Coefficient of Determination** | **Feature Importance** |
| 15 | 362.157 | 228130.65 | 263.961 | 0.830 | minTemp12 0.46  maxTemp7 0.10  minTemp5 0.08 |
| 14 | 362.648 | 228766.692 | 263.961 | 0.830 | minTemp12 0.46  maxTemp7 0.10  minTemp5 0.08 |
| 13 | 362.665 | 230049.626 | 263.961 | 0.828 | minTemp12 0.46  maxTemp7 0.11  minTemp5 0.09 |

Table 8. Soybean Random Forest Regressor (customized) results for 15, 14 and 13 weeks of data.

**Wheat - Lasso**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Weeks** | **Mean Absolute Error** | **Mean Squared Error** | **Median Absolute Error** | **Coefficient of Determination** | **Feature Importance** |
| 31 | 437.308 | 315384.795 | 417.106 | 0.715 | minHum7 313.14  minHum8 298.33  SystemNameType 281.49 |
| 30 | 437.323 | 315378.236 | 417.103 | 0.714 | minHum7 317.22  minHum8 298.55  SystemNameType 281.51 |
| 29 | 437.352 | 315416.478 | 416.873 | 0.715 | minHum7 323.74  minHum8 294.89  SystemNameType 281.50 |

Table 9. Wheat Lasso regression results for 31, 30 and 29 weeks of data.

**Wheat -** **Decision Tree Regressor**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Weeks** | **Mean Absolute Error** | **Mean Squared Error** | **Median Absolute Error** | **Coefficient of Determination** | **Feature Importance** |
| 31 | 510.093 | 411037.387 | 441.293 | 0.611 | minHum7 313.14  minHum8 298.33  SystemNameType 281.49 |
| 30 | 510.093 | 411037.387 | 441.293 | 0.611 | minHum7 317.22  minHum8 298.55  SystemNameType 281.51 |
| 29 | 510.093 | 411037.387 | 441.293 | 0.611 | minHum7 323.74  minHum8 294.89  SystemNameType 281.50 |

Table 10. Wheat Decision Tree Regressor results for 31, 30 and 29 weeks of data.

**Wheat -** **Random Forest Regressor**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Weeks** | **Mean Absolute Error** | **Mean Squared Error** | **Median Absolute Error** | **Coefficient of Determination** | **Feature Importance** |
| 31 | 524.17 | 442917.360 | 473.258 | 0.572 | Precip22 0.26  minTemp29 0.11  SystemNameType 0.07 |
| 30 | 516.191 | 415804.481 | 487.619 | 0.595 | minTemp29 0.20  minHum8 0.14  Precip22 0.11 |
| 29 | 520.976 | 423395.328 | 465.151 | 0.545 | minTemp29 0.23  minHum8 0.08  SystemNameType 0.07 |

Table 11. Wheat Random Forest Regressor results for 31, 30 and 29 weeks of data.

**Wheat -** **Random Forest Regressor** (**max\_depth=35, n\_est = 500)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Weeks** | **Mean Absolute Error** | **Mean Squared Error** | **Median Absolute Error** | **Coefficient of Determination** | **Feature Importance** |
| 31 | 511.866 | 405060.257 | 440.921 | 0.589 | minTemp29 0.19  minHum8 0.09  Precip22 0.09 |
| 30 | 511.258 | 404878.603 | 440.789 | 0.591 | minTemp29 0.19  minHum8 0.09  Precip22 0.09 |
| 29 | 511.865 | 404994.336 | 440.921 | 0.588 | minTemp29 0.19  minHum8 0.09  Precip22 0.09 |

Table 12. Wheat Random Forest Regressor (customized) results for 31, 30 and 29 weeks of data.

# References

1. Explanation of Growing Degree Days, Midwestern Regional Climate Center. Retrieved March 1, 2020 from mrcc.illinois.edu/gismaps/info/gddinfo.htm.
2. 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>