Data Analytics CSCI 4600

Jaimin Vyas

Professor Thilanka Munasinghe

Final Project Report

For this project, we analyze how different variables related to the climate, environment, and other fields affect annual crop yields, specifically wheat yield in the United States. The motivation for this topic initially was the fact that the Earth’s population recently hit 8 billion and continues to rise. With a higher population comes a higher demand for food, and if the production of crops does not rise to meet these higher expectations, then there are real risks of food shortages. There are many hypotheses that can be formulated from this problem, though the one that we will tackle in this report is quantifying which types of climate variables impact the annual yield of wheat in the United States, and to what extent. We will obtain the data from public repositories, perform initial analysis and data munging, then construct two machine learning models to forecast the crop yield using different variables. These models could potentially be used to determine what areas are best to grow wheat in.

We used multiple datasets for this project; for crop yields the dataset was taken from the USDA Production, Supply, and Distribution data portal, for the NDVI we used NASA’s Global Agricultural Monitoring System, for average temperature we used The World Bank, and for precipitation we used the Climate Change Knowledge Portal, which is a subset of The World Bank Group. The first two datasets were given to us by the instructor, and the second two datasets were procured because there were not enough relevant features to test from just from the initial two datasets. All datasets are from trusted governmental or international organizations, and the data was all downloaded from the website. The main variable of interest is from the USDA crop yield data, we try to predict the Value variable, which is an integer variable that represents thousands of megatons. However, there are multiple types of attributes as is seen on the next page in Figure 1 by the Attribute\_Description column, and since we are trying to predict crop yield, we chose Attribute\_Description=Production.

Chart, bar chart

Description automatically generated

Figure 1

As we can see, there are an even number of each Attribute\_Description, so there is no data being lost by choosing only Production. In the NDVI dataset, there are a few values to choose from to represent the NDVI. We can either choose SAMPLE.VALUE, which represents the NDVI at a random day measured within the date range provided, or MEAN.VALUE, which represents the average of all NDVIs from the 8-day sampling period. For this project, we chose to use the MEAN.VALUE, since that is a better metric and represents a real average. The other two datasets do not contain any variables other than the year or date and the mean value of the temperature/precipitation recorded. While analyzing the data, we noticed that the date formats were different for all four datasets, and that the crop yield dataset had two different date columns, one was a calendar year and the other was a market year. We decided to use the market year, since there were very few unique values for the calendar year. The only issue here was that the market year data started from July 1960, which we did not have the NDVI or temperature or precipitation for, so we cut the datasets to start in 2000 and end in 2020. However, the crop production dataset still seems to have fewer unique dates than the NDVI dataset, which might lead to a very small final compiled dataset to test our models on.

We started exploring the statistical aspects of the datasets by examining how the target variable and the features we plan to use behave within their own datasets. Looking at how the NDVI changes over time, we can see that it has very sinusoidal behavior, which is expected as it is a seasonal dependent variable. The graph is shown in Figure 2 below.

A picture containing text, nature

Description automatically generated

Figure 2

We then can look at the wheat production of the United States over time, to see initially whether it is a linear trend or if it has nonlinearity. As can be seen from Figure 3 below, wheat production with relation to time almost seems parabolic, but obviously more variables will need to be added to help describe its behavior more.

Chart, scatter chart

Description automatically generated

Figure 3

We wanted to know how this data was distributed, if it was a Normal distribution or otherwise, and so we used the Shapiro-Wilk test for Normality. First we constructed a quantile-quantile plot of the data and saw that the graph was almost linear. This prompted us to use the test for Normality, since the graph was very close to Normal. According to the test, the distribution had a W value of 0.97633 and a p-value of 0.2646, which means that it is not significantly different from a Normal distribution. The related graphs and outputs are shown below.

Chart, line chart, scatter chart

Description automatically generated

Figure 4

Text

Description automatically generated with medium confidence

Figure 5

As we can see, the Q-Q plot is near linear. Thus we can treat this distribution as Normal, so we likely should center and scale the other variables we use for prediction as well. Then, we merged the NDVI and wheat production datasets on the calendar year and month the observation was made. We then added the temperatures, and the precipitation. With 220 observations of 29 variables, we can now start training our models.

As mentioned earlier, we first center and scale all the data. We used specific variables to predict, so we only picked the Month, Calendar\_Year, NDVI, Temperature, and Precipitation. We chose these variables because the majority of the other variables were either derivatives of one included in this subset, or were of no use to predict the crop yield. The first type of model we use to predict the amount of yield is a linear regression model. Since this is clearly a regression problem, linear regression was an easy first choice to get a benchmark for how the data all behaved together. After using an 80% training and 20% testing split, we trained an initial linear regression model. Our output when we view a summary of the model is as follows

Text

Description automatically generated

Figure 6

As can be seen, the median residual was 1672, which means that the median difference between the predicted value of the training set by the model and the actual value was 1672. The IQR was 17,825, which is actually quite substantial, since the IQR of the Yield value in the training dataframe was about 10,000. Seeing as this is almost double, maybe this model is not that accurate. But, we will test on a test set anyway. We decided not to use a validation set because there were not that many observations after all the merges were done, so if we included a validation set then we would either need to train on less data, which would create a less accurate model, or test on less data, which might not be indicative of the model’s actual performance. Looking back at Figure 6 for the output, we can see that the Adjusted R-squared value is very low, which means that the model does not fit the data well. For this reason, we will discard this model and try again with a different type of regression model. After some munging, we chose to go with a formula that involved the sine of the NDVI, since we noticed its seasonal behavior, and remove the Month and Calendar\_Year variables, since the NDVI was already predictable over the course of a year, leading to multicollinearity. The Temperature variable also seemed to not be helping much, as every time it was added the Adjusted R-squared either barely went up or, as was in most cases, went down. Thus we elected to remove it. The precipitation variable had similar behavior, so we removed that as well. Thus our resulting formula was just the sine value of the NDVI. The output for this model is shown below:

Text

Description automatically generated

Figure 7

As can be seen from this output, the model still does not predict the values well given the low adjusted R-squared value, but the p-value is quite low, so there could still be hope. We will predict using this model and calculate the MSE to determine how well it predicts the testing data. Using the standard MSE formula, we get an error rate of about 30 million. It is hard to imagine what this actually means in the scope of the problem, so we will center and scale the Yield values. Since they are already almost Normal, this should just act as a simple translation. After centering and scaling, then retraining the model, we get an MSE of about 211. Considering that a Normal distribution is centered at 0 with standard deviation 1, this is very, very high. This tells us that a linear regression model is likely not a good way to predict the crop yields. The next model we try and use is a K-Nearest Neighbors model for regression. In this model, we provided every variable to the model to see how it would perform as a baseline, and the model actually was able to fit the data quite well. In the plot below, we can see that over time, the model was actually able to near match the actual observed values for cop Yield over time.

Chart, line chart

Description automatically generated

Figure 8

The vertical line when the Calendar\_Year is 2006 comes from the fact that in the data, there were a lot of observations in that year compares to the other years. When we calculate the MSE for this model’s predictions on the test set, we get a much lower score than the linear regression model got with an MSE of almost 3. This shows that the KNN model is much better at predicting the crop yields than others. This plot was for a model with k=5 neighbors, and if we increase the number of neighbors, the MSE gets higher. This could likely be due to the fact that the Temperature and Precipitation variables throw the predictions of the model off, so the more that are counted when making a prediction, the less accurate it will be. We tried to remove the Temperature variable and the date variables Month and Calendar\_Year so that the model was just left with the NDVI and Precipitation values, but that yielded an MSE of 209. This was odd behavior, especially considering that regardless of whether the number of neighbors was increased or decreased, the MSE increased. We do not have a lot of faith in these models, as their behavior is quite strange, and from the linear models we already know that there is little to no correlation between the variables used to predict crop yield. Even if the NDVI values were taken out when training, the Temperature and Precipitation do not do a good job of telling what the crop yield measure will be.

In conclusion, crop yields are an extremely difficult thing to predict, even if they are narrowed down to one crop and one country. A K-Nearest Neighbors algorithm seemed to work best with the data we had, but it is likely that it was somewhat up to chance, and the model has some strange behaviors when we try to tune its hyperparameters. In the initial stage of this project, we assumed that any sort of climate variables would help predict the future yield of a given crop, because crops depend on climate conditions to thrive. However, after attempting to train a few models, we can see that is not the case. After the initial linear regression model, we also tried a regression tree, but that ended up simply trying to predict the crop yield using one variable, and had a very low accuracy with a very high MSE, comparable to the first linear model we made. If we had more time, we would likely try to get a different set of data to use, and maybe try to go even more local by selecting a specific state in the US that grows wheat. If we were able to do that, then maybe the crop yield variability would be more predictable.

References

Data Sources:

<https://www.worldometers.info/world-population/>

<https://data.worldbank.org/country/united-states?view=chart>

<https://climateknowledgeportal.worldbank.org/country/united-states/trends-variability-historical>

<https://apps.fas.usda.gov/psdonline/app/index.html#/app/home>

<https://glam1.gsfc.nasa.gov/>