Machine Learning Models using Remote Sensing for County-Level Corn Yield Prediction in the Midwestern U.S.

Contestants
Gabby Whisler
Jenna Kouba
Micah Robinson
Zach Jannusch*

Dates of Graduation
12 May 2023
18 December 2022*

Advisor
Dr. Zhou Zhang

Student Branch Faculty Advisor
Dr. Brian Luck

Signatures

Zach Jannusch (May 15, 2023)

Jenna M. Kouba
Jenna Kouba (May 15, 2023)

Micah Robinson (May 15, 2023)

Gabby Whisler (May 15, 2023)

Zhou Zhang
Zhou Zhang (May 15, 2023)

Brian D. Luck
Brian Luck (May 15, 2023)

Machine Learning Models using Remote Sensing for County-Level Corn Yield Prediction in the Midwestern U.S.

Abstract

Climate change is causing corn yields to become more unpredictable, heightening the risks of food insecurity and crop shortages. The United States is the largest producer and exporter of corn in the world. The variable environmental conditions can affect year-to-year yield and best management practices for corn farmers in the midwest. The United States Department of Agriculture (USDA) believes there is a need for a free and publicly available yield-predicting model to help producers and other stakeholders make informed decisions. With climate change being a growing concern, it is imperative to have accurate prediction methods that account for variations in climate and other yield factors. The current USDA prediction methodology requires a survey of thousands of farmers, making the predictions labor-intensive, infrequent, and prone to human error. The goal of this project was to design and evaluate a machine learning model that will predict corn yields on a county level for 12 states in the Midwest. Models were tested using the USDA previous year predictions from 2008-2019 to train the models and then evaluated the models' predictive performance for the years 2020 and 2021. The models constructed include Random Forest, Support Vector Regression, Ridge Regression, and Multilayer Perceptron. After evaluating each of these models, Random Forest and Ridge Regression were the most accurate with R^2 values of 0.84 and 0.85, respectively. These results show that machine learning models using remote sensing data have the potential to to rival existing USDA prediction methods with their high accuracies and extremely low cost.

Acknowledgments

This team would like to express gratitude to the Biosystems Engineering Department at the University of Wisconsin-Madison for enabling and supporting this project. Specifically, the team would like to thank their mentor, Dr. Zhou Zhang, for sharing her time and wealth of knowledge with the team over the past year. Thanks to course instructors Kody Habeck for providing the tools needed to accomplish this project and for inspiring the team to participate in this competition, and Andrew Hillman for offering his valuable feedback and guidance. Special thanks to Ph.D. student Yucci Ma for his assistance on this project, especially for his Google Earth Engine coding contributions.

Table of Contents

1. Problem Scope

Varying environmental conditions can affect year-to-year yield and best management practices for corn farmers in the midwest. The USDA believes that corn producers need a free and publicly available yield-predicting model to help them make informed decisions. The goal is to design and evaluate a machine learning model that will predict corn yields on a county level for 12 states in the Midwest, using free and already available yield, soil, and climate data as inputs. The final model and its evaluation should be completed and ready to be presented by December 7th, 2022.

2. Technical Review

2.1 Background

Global warming has resulted in more climate variability, affecting food systems and water resources. Droughts and extreme temperatures have negatively impacted corn production and have resulted in more uncertainty in crop yield prediction (Kang et al., 2009; Crane-Droesch, 2018). Significant storm events bringing excessive precipitation also damage corn yields, another factor contributing to crop yield uncertainty (Li et al., 2019). With a changing climate, it has become increasingly essential to create accurate models which allow farmers and policymakers to make informed decisions on how to best manage their farms to ensure that there is an adequate food supply for the world's population (Becker-Reshef, 2010; Wang et al., 2020; You, n.d.).

The United States is the world's leading producer and exporter of corn, supplying approximately 30% of global maize production (Ma et al., 2021; Li et al., 2019). Crop yield prediction is beneficial for managing the economics of corn. Estimating crop yields can provide information for commodity trading and insurance assessments (Kang et al., 2020; Cai et al., 2017). For producers, crop yield prediction allows them to set goals, evaluate alternative methods, and optimize their management practices (Bocca et al., 2015).

Since the dawn of agriculture, humans have been developing methods to predict crop yields before harvest (Basso and Liu, 2019). Historically, farmers have estimated crop yield based on their field observations and local knowledge. Adages such as "knee high by the fourth of July" served as rudimentary predictors of corn yield (Westfall, 2021). Periodicals like the "The Farmers' Almanac" or "The Old Farmer's Almanac" have been published seasonally since the late 18th century, providing weather predictions to aid farmers in making management decisions (Walsh and Allen, 1981).

Farmers can better manage their fields by utilizing accurate models that provide insight into the link between environmental stresses and crop growth (Guan et al., 2017). Models can be advantageous in predicting weather patterns and their effects on yield, allowing farmers to adjust their budgeting and harvesting plans accordingly (Bocca et al., 2015). Corn is a high-input crop and can contribute to environmental degradation through nitrous oxide emissions and nitrogen and phosphorus-rich runoff. The use of precise model estimates in agriculture can reduce inputs into the soil such as nitrogen and phosphorus fertilizers, which have negative economic and environmental impacts when overused (McNunn et al., 2019).

Machine learning models must include data inputs that provide relevant and valuable information in order to create accurate predictions. The USDA's National Statistics Service conducts two-level surveys that provide state and national yield estimates (Johnson, 2014). These past yield inputs are valuable as they provide a basis for the models to be constructed or trained. Other inputs must be considered as corn yield is highly variable and depends on many environmental factors. Inputs such as temperature and precipitation explain about one-third of crop yield fluctuation (Ray et al., 2015). Another variable utilized for crop prediction is vapor pressure deficit (VPD) which correlates with heat stress and crop water. Other inputs utilized are vegetation indices (VIs), soil moisture, and satellite-based evapotranspiration (Kang et al., 2020).

Several models can predict corn yield with varying levels of success; regression modeling has been used to correlate environmental conditions and crop yields, but the development of machine learning models has improved the accuracy of yield predictions. Machine learning models can handle nonlinear and complex datasets (Khanal, 2018). Machine learning is advantageous as it can estimate yield by creating relationships between the variables and the known yields. Machine learning models produce more accurate yield predictions than traditional regression models (Kaul et al., 2005). However, it can be difficult to quantify the uncertainty of these predictions because machine learning can operate like a black box, where the processes or logic linking data inputs and outputs is not clear (L. V. Jospin, 2022). Examples of existing machine learning models for crop yield prediction include deep neural networks (DNN), Convolutional Neural Networks (CNN), Ridge Regression, and Random Forest. Analysis of the accuracy of these models has been done in previous studies.

2.2 Existing Technology

2.2.1 Machine Learning Model

Regression modeling is an older method of predicting unknown or future values and is often less accurate compared to contemporary models. Most regression models assume linear relationships between inputs and outputs, which is not always the reality. These models include Ordinary Least Square and Least Absolute Shrinkage and Selection Operation (LASSO). A study analyzed these methods against non-linear machine learning methods. The results showed that non-linear models produce significantly more accurate results than linear regression models for corn yield predictions (Wang et al., 2020).

Random Forest (RF) is a supervised learning algorithm that utilizes ensemble learning through the use of many decision trees. Training data is required to teach the model the relationship between features and known outputs. Once trained the model will output predicted values when given feature values. Random forest, as its name suggests, utilizes multiple decision trees that run independently and in parallel to each other to create a “forest” of decision trees. Figure 1 illustrates a forest voting to the final output. To calculate the final output, the values from all decision trees are averaged to get the predicted value (Biau, 2016). RF has been used commonly with remote sensing data. An assumption that this model makes is that the variables will maintain their high/low value of prominence in the dataset. This causes the model to not be as accurate with large complex data sets. One limitation of RF is that it cannot extrapolate, meaning that its output is bounded by the highest and lowest values in the training set (Hengl, 2018). This could be problematic if climate change creates novel weather conditions that result in historically high or low corn yields, which this model will not be able to predict.

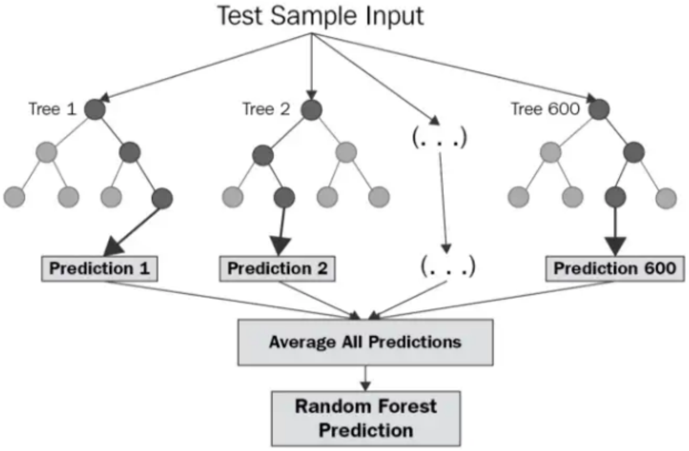


Figure 1: *A forest of trees. The regression RF final output averages all the predictions for the final value. (Chaya, 2020).*

Deep Neural Networks are new and accurate models. They are used because they are the best at predicting the nonlinear relationship between input and output variables. Many relationships between corn yield and environmental factors are nonlinear. Furthermore, certain weather conditions can either help or hurt corn yield depending on the time in the growing season that they occur. Deep Neural Networks work by stacking multiple nonlinear layers of connected nodes, or “neurons”. This model then creates relationships between nodes, forming complex linkages between nodes in each layer. This complexity generally adds to the accuracy of the model but becomes prone to overfitting in deep neural networks with over 20 layers. . However, the development of the outputs is unknown as the relationships are not readily apparent. Recent deep learning crop prediction model was developed by Khaki and Wang in 2019. The study first created a weather prediction network shown in Figure 2. The figure displays one layer of a neural network, known as a shallow neural network. One layer creates many relationships within the input data to find the output variable. This network was created to predict weather to provide the necessary information to determine crop yield. The full crop prediction model in Figure 3 depicts a deep neural network with multiple layers. The model accounts for many different variables for predicting corn yields. This diagram illustrates the increasingly complex relationships between the variables by stacking the layers. This model is currently one of the most accurate for predicting corn yields (Khaki and Wang, 2019).

A study by Kang et al. in 2020 set out to predict corn yield on a county-level for the Midwestern U.S. using six different machine learning models: Lasso, Support Vector Regressor, Random Forest, XGBoost, Long-short term memory (LSTM), and Convolutional Neural Network (CNN). The variables used included weather data, land surface models, soil maps, crop progress reports, and satellite data. The findings supported that the XGBoost was the best performing model in accuracy and stability, whereas LSTM and CNN did not provide useful predictions (Kang et al., 2020).

Although these current models have produced accurate results, further improvements can be incorporated into these predictors. These current models do not provide uncertainties that correspond with their predictions (Ma et al., 2021). Building upon these current models and utilizing the most relevant and accurate inputs, crop prediction can be a valuable resource to combat food insecurity in the wake of the climate crisis.

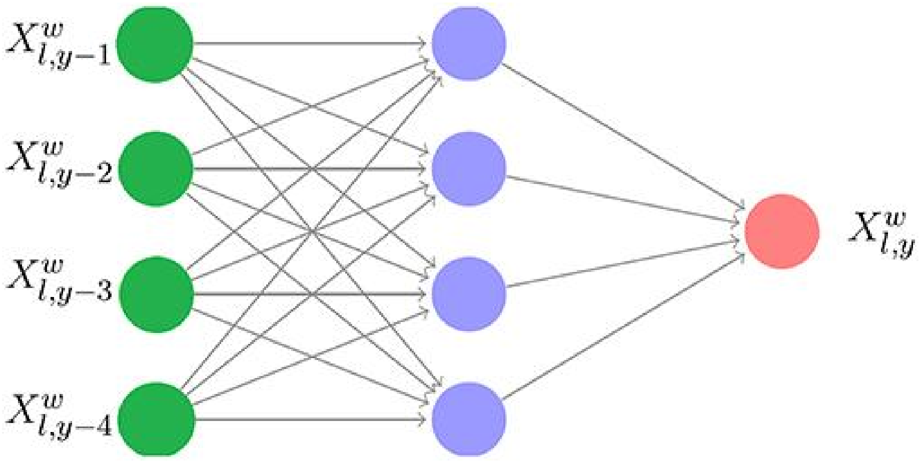


Figure 2: *Shallow Neural Network: weather prediction (w) in a known location (l) of corn yield over four years. Includes one hidden layer to create the output (Khaki and Wang, 2019).*

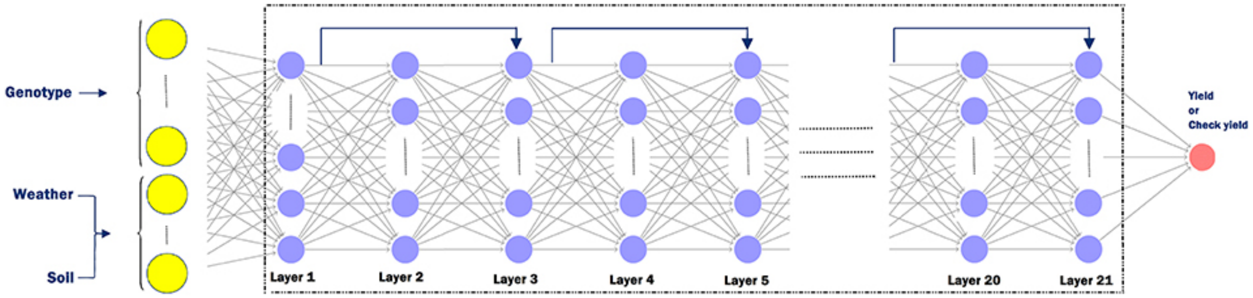


Figure 3: *Deep neural network: Various corn variables over four years that predict crop yield. Includes twenty-one hidden layers to create the output (Khaki and Wang, 2019).*

2.2.2 Remote Sensing

Recent improvements in remote-sensing technology have made it an incredibly valuable resource for large-scale data gathering. Remote sensing data, specifically spectroradiometer data gathered by satellites, can relatively cheaply provide information about vast areas of land (Wang et al., 2018). By analyzing reflected spectra that bounce off of soil or vegetation on the earth's surface, many properties of the soils and plants can be inferred. Various metrics have been developed based on reflected spectra characteristics, such as Normalized Difference Vegetation Index (NDVI), two-band Enhanced Vegetation Index (EVI2), and Normalized Difference Water Index (NDWI), which have significant predictive potential for both maize and soybean yield (Bolton and Friedl, 2013). NWDI is especially effective at characterizing the water content of vegetation and is less sensitive to atmospheric conditions, making it an ideal tool for monitoring crop health (Gao, 1996). Incorporating vegetation indices from remote sending data into predictive models has been found to dramatically increase the model's performance (Becker-Reshef et al., 2010; Johnson, 2014; Guan et al., 2017; Peng et al., 2018).

One existing machine learning model predicted soybean and corn yields based on Normalized Difference Vegetation Index (NDVI) from satellite Moderate Resolution Imaging Spectroradiometer (MODIS), nighttime and daytime land surface temperature from MODIS, and precipitation data from the National Weather Service. By inputting data from 2006-2011 for the US Corn Belt, a regression tree-based model was constructed at the county level with a coefficient of determination (R^2) of 0.93. The model was then used to predict yield for 2012 when a drought occurred. It predicted the yield with an R^2 of 0.77 for corn and 0.71 for soybeans with RMSE values of 1.26 and 0.42 metric tons per hectare, respectively (Johnson et al., 2014).

2.3 Fundamental equations

R^2, RMSE and MAPE are standard performance metrics for evaluating the accuracy of a model's predictions (Chicco D. 2021). These metrics are essential for developing and optimizing a predictive model.

R^2 is the coefficient of determination. R^2 is the proportion of the dependent variable that can be predicted from the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (X_i - Y_i)^2}{\sum_{i=1}^{m} (\overline{Y}_i - Y_i)^2} \qquad (1)$$

RMSE is the root mean square error, which detects the number and extremity of outliers.

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (X_i - Y_i)^2} \qquad (2)$$

Mean Average Percent Error (MAPE) is a valuable metric that expresses the accuracy of measurements. It is most useful when used for data that has relative changes compared to absolute change (De Myttenaere et al., 2016).

$$MAPE = \frac{1}{m} - \sum_{i=1}^{m} \left| \frac{Y_i - X_i}{Y_i} \right| \qquad (3)$$

3. Design Requirements

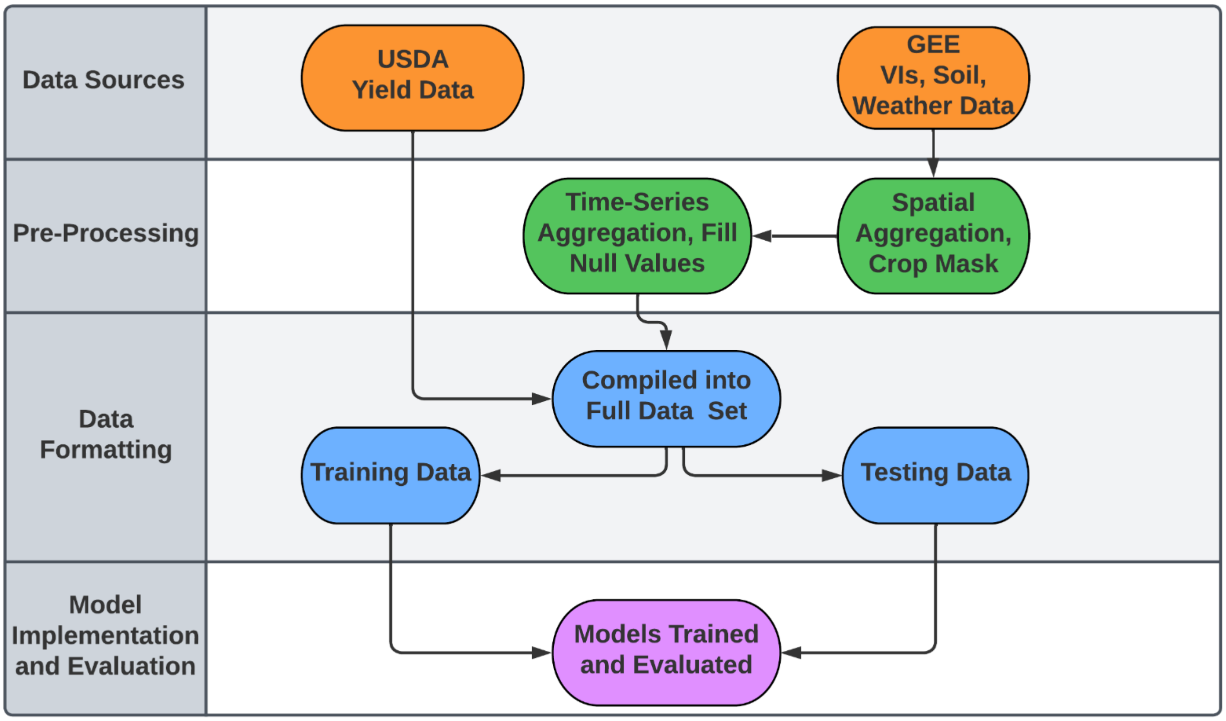
- The final model should be able to create crop yield predictions that are 70% accurate compared to actual crop yields. (client)
    - Percent Accuracy will be evaluated by Mean Absolute Percentage Error across all counties for which predictions are made.
    - Models will be tested against 2020 and 2021 data.
    - Models will be evaluated based on the prediction that they would provide on August 20th of a given year.
- The inputs to the model must contain the following county-level data from the selected 12 midwestern states: soil property data from Web Soil Survey, remote sensing data from Google Earth, temperature data, and precipitation data. (client)
- The model must predict annual corn yields on a county scale for 12 selected states in the midwest. (client)
- All data used for the model must be free and publicly available. (client)
- Final model output yields must be displayed on a map of the included states and counties. This will allow the model output yields to be visually compared to a map of actual yields. (client)

4 Design Description

4.1 Overview

The broad goal of this model was to establish relationships between known measured variables and end-of-season corn yields. The first step was acquiring the necessary data, in this case from the USDA for yield data and Google Earth Engine for Vegetation Indices (VI), soil properties, and weather data. Next, the data was pre-processed and formatted so that it could be understood by machine-learning models. As part of this step, the historical yield for each county was derived from the USDA data, time-series GEE data was aggregated into 16-day time periods, and any missing values were filled in. Finally, all of the data was compiled into a full dataset. This full dataset could then be divided into training and testing data in a variety of ways. Once split into training and testing data, models could be trained and evaluated.

4.2 Detailed Description



Data Sources
USDA Yield Data
GEE VIs, Soil, Weather Data
Pre-Processing
Time-Series Aggregation, Fill Null Values
Spatial Aggregation, Crop Mask
Data Formatting
Compiled into Full Data Set
Training Data
Testing Data
Model Implementation and Evaluation
Models Trained and Evaluated

Figure 4: *Flowchart showing modeling process*

4.2.1 Data Inputs

Google earth engine and historical crop yield provided all the data sourcing for the model. Google earth engine imported weather variables from the Parameter elevation Regressions on Independent Slopes Model (PRISM) dataset, soil variables were uploaded from SSURGO which had Soil Available Water Holding Capacity (AWC), Soil Organic Matter (SOM), and the cation exchange capacity soil property map The vegetation index variables from MODIS NBAR (MCD43A4).

Vegetation indexes use satellite sensors to monitor landscapes. VIs measure the greenness of an area and from those measurements a lot of information can be created. VI's are found to be nearly linearly related to the amount of photosynthesis produced by a plant. This relationship makes vegetation indexes a powerful predictor of crop yield. (Edward P. Glenn, 2008). The VI variables in our model were: EVI, GCI, NDWI, and NDVI.

NDVI is the most widely used vegetation index. It is defined as:

$$\text{NDVI}= \frac{NIR-R}{NIR+R} \qquad (3)$$

NIR is the reflectance values of Red and R is the near Infrared light received from the sensors. It has a strong relation in predicting crop characteristics and chlorophyll content. (Rouse, 1973)

The Enhanced Vegetation Index (EVI) is defined as:

$$EVI = 2.5 \times \frac{(NIR-R)}{(1+NIR+(6\times R-7.5\times Blue)} \qquad (4)$$

EVI Is an improved NDVI and has constants to account for soil and atmosphere influences. The constant 1 accounts for the canopy background and the 2.5 and 7.5 minimize light variations. EVI is better at predicting high biomass areas, photosynthesis, and plant transpiration (Edward P. Glenn,2008), (Rouse, 1973).

$$GCI = \frac{NIR}{Green} - 1 \qquad (5)$$

Green chlorophyll index (GCI) predicts the amount of chlorophyll in leaves. (Gitelson, 2003)

$$NDWI = \frac{\rho(.086\mu m)-\rho(.1.24\mu m)}{\rho(.086\mu m)+\rho(.1.24\mu m)} \qquad (6)$$

The Normalized Difference Water Index (NDWI) is used to measure the amount of liquid water in vegetation from space. ρ represents the radiance in reflectance units. The .86 and 1.24 are reflectances that find the amount of liquid water in a canopy (Gao, 1996).

4.2.2 Data Preprocessing

Historical yield data were calculated for each county and year from the USDA annual yield data. The historical yield was calculated based on data starting in 2001 to until the year before (Historical yield for 2010 would be the average from 2001-2009, for example).

The first step in pre-processing GEE remote sensing data was to aggregate the spatial data over our areas of interest using US county shapefiles. Since the NASS Cropland Data Layer (CDL) became available in 2007 for the regions of this study, and data was only used going back to 2008, the CDL was used as a crop mask to only include land under corn production. Since many of the remote sensing data from GEE were time-series, daily data was aggregated into 13 intervals, each 16 days long. After time-series aggregation, some periods were still missing values and were filled with column averages to ensure that no null values made it through to the models.

Lastly, the many datasets were compiled into one. This full dataset included the following static data: Measured Yield, Year, County ID, Historical Yield, Mean CEC, Mean SOM, and Mean AWC. It also included the following time-series data, with 13 periods for each: Min VPD, Max VPD, Min Temp, Mean Temp, Max Temp, Precipitation, Mean NDVI, Mean NDWI, Mean GCI, and Mean EVI. This resulted in a total of 137 features. This complete dataset could then be split into training/testing data in various ways to create and evaluate predictive models. One way that the data was split was by selecting one year to evaluate the model on, and using all previous years to train the model. This method was able to show how the model performed against the different conditions that each year brought, such as drought in 2012 and heavy precipitation in 2009. This also mimics the way that the model would be used in the real world, predicting one year at a time based on all of the historical training data available.

4.2.3 Model Implementation

Models were constructed in Jupyter Notebook using Python and Scikit-learn libraries. The construction of models followed the following form:

```
model_variable = model_function(parameter_1,... parameter_n)
model_variable.fit(train_x, train_y)
prediction = model_variable.predict(test_x)
```

The first line sets the parameters of the model, if a parameter value is not specified it remains at the default value, which is set by Scikit-learn. Parameters for each model were tuned manually, with model accuracy assessed using RMSE, MAPE, and R^2 for 2021 yield prediction. The second line of code fits the training inputs (Climate data, soil data, and vegetation indices) to the training outputs (known historical yields) using the model. Models were trained with all years between 2008 and the testing year. For example, if the model is predicting 2021's corn yield, all years from 2008 to 2020 would be used to train the model. The third line of code generates the predicted values for the test year using the same input variables as the training data. These predicted values are what are used to evaluate the model.

4.3 Use

This model would be used by a government agency such as USDA to create and disperse crop yield estimates throughout the growing season. These results would likely be displayed on a website in a visual form such as a yield map like the one shown in figure 5. The updating and maintenance of the model, as well as details regarding the dissemination of predictions, are both outside the scope of this project.

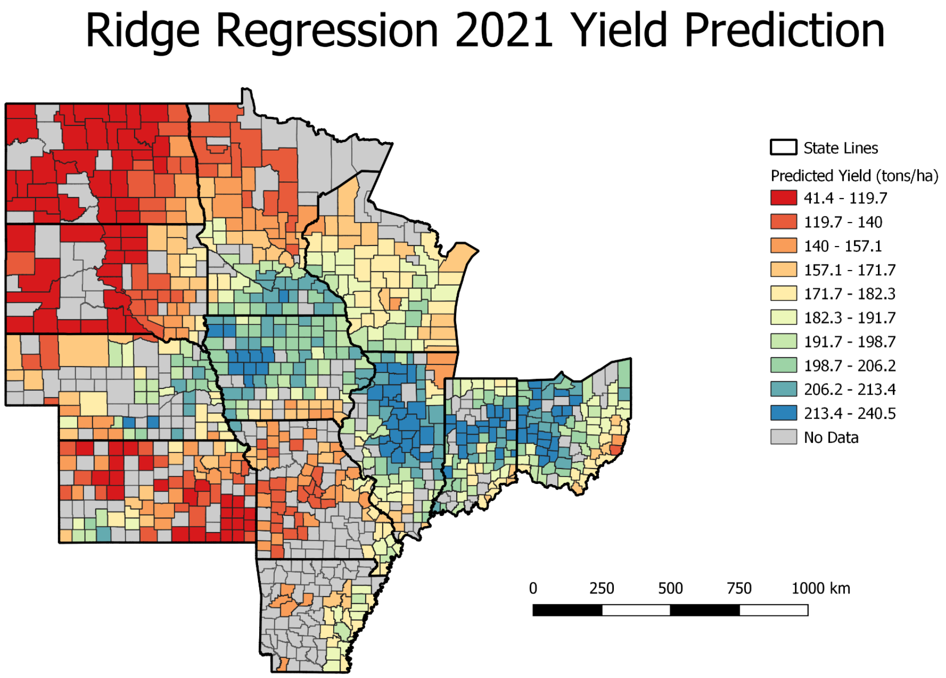


Figure 5: *Map depicting 2021 predicted corn yield by a ridge regression model.*

5. Evaluation

5.1 Overview

The models were evaluated by comparing predicted yields against reported yield values from USDA. Models were evaluated based on R^2, MAPE, and RMSE. Models were initially optimized and evaluated using 2020 and 2021 for testing data and training 2008-2019 for training data. Secondly, each year from 2012 to 2021 was individually evaluated by using it as the test years, and using all years prior (back to 2008) as training data. This evaluation was intended to show how the various models performed over time and in the various conditions that accompanied each year. Lastly, the models were evaluated for intra - season prediction accuracy, by training and testing the models based on only what data would be available at any given point throughout the growing season. Lastly, a variable importance analysis was performed on a selected model to identify which input factors the model weighted heavily in its predictions.

5.2 Testing and Results

The Ridge Regression and Random Forest models generally outperformed the Support Vector Regression and Multi-Layer Perception models by a significant margin. Using the years 2020 and 2021 as testing data and 2008-2019 for training, Ridge and RF achieved R^2 values of around 0.85 while SVR and MLP were near 0.75 (Table 1). These predictions used time-series data available prior to the 277th Day-of-Year, meaning that these predictions would be available on October 4th..

Model Type	R^2	MAPE	RMSE
Ridge	0.86	9.10	15.57
RF	0.84	9.19	16.47
SVR	0.76	12.17	20.16
MLP	0.73	13.10	21.34

Table 1. *The three metrics used to evaluate the corn yield prediction of the four models for 2020 and 2021. The model's outcomes were evaluated using R^2, MAPE, and RMSE.*

Intra-season predictions were calculated using only the data that would be available prior to selected dates throughout the growing season. The 2014-2021 average correlation coefficients of the model predictions throughout the growing season are shown in figure 6. The two models show a similar pattern of increasing R^2 over time until approaching maximum accuracy and plateauing around August 3rd (DOY 215).

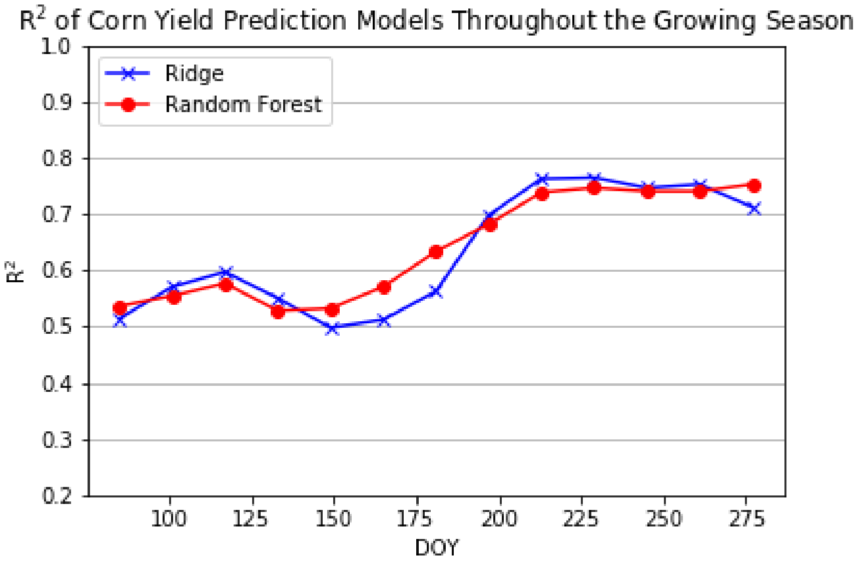


Figure 6. *2014-2021 avg. R^2 values of two models (Ridge and Random Forest) throughout the growing season.*

As part of the evaluation, a variable importance analysis was performed to identify which features affected corn yield the most. As seen in Table 1, NDVI was by far the most powerful predictor of corn yield, with precipitation and historical yield being the second and third best predictors.

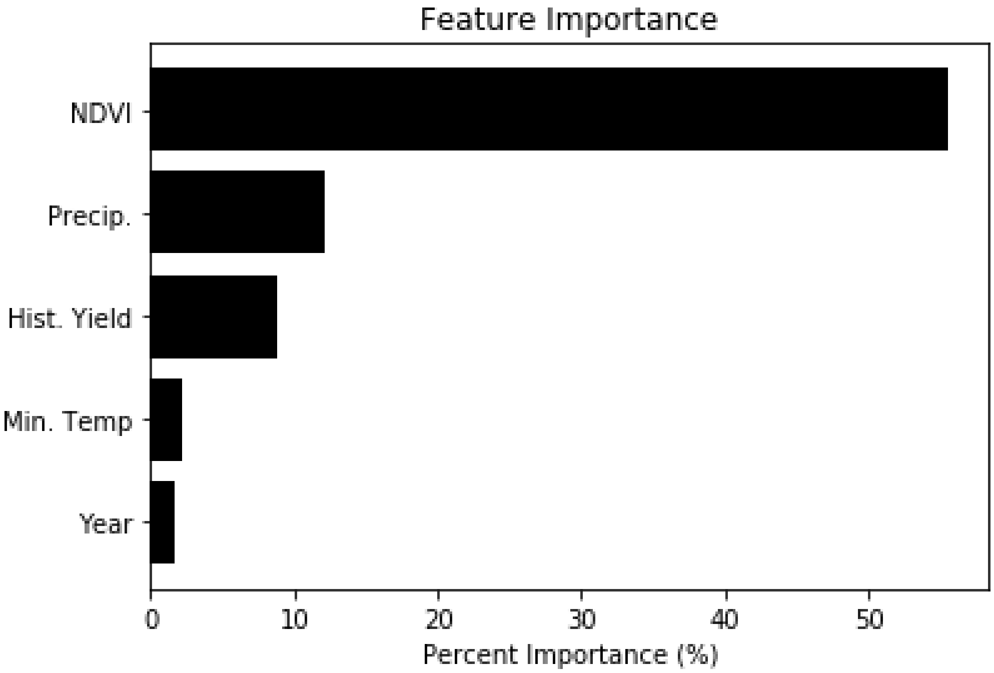


Figure 7: *The top 5 variables that were found to have the most feature importance in the random forest model. NDVI was found to be the most useful by a large margin in predicting crop yield.*

5.3 Sustainability

The use of this model has significant environmental impacts. If the model produces accurate predictions and allows corn farmers and corn ethanol producers to make more educated decisions, their processes could be made more efficient. This could reduce environmental degradation if farmers can use less water and fertilizer inputs to produce similar yields, or if corn ethanol producers can produce more renewable fuels because of more predictable corn markets.

This project can vastly improve economic conditions for many people and industries. For farmers, the model could help them make the best use of their resources by making well-informed decisions. This could result in increased profits for farmers. Additionally, the use of this model could help inform decision-making abilities of industries that depend on corn production, increasing efficiency and boosting profits. Industries that could see significant economic impacts include crop insurers, commodities traders, food manufacturers, and corn ethanol producers.

Lastly, this project will improve social sustainability. The accurate predictions that it provides will help policymakers manage food supply across the globe and give them time to react to any anticipated food shortages. This could improve health outcomes both domestically and around the world. Since the model's predictions will be free and available to everyone, it will reduce inequality by allowing even the smallest of corn farmers access to high-quality predictions. This will help farmers sustain their businesses and way of life.

While this project could improve environmental, economic, and social sustainability with accurate corn yield predictions, inaccurate predictions could cause significant damages. It is important that the inherent risk or relying on predictive models is communicated to users so that the model can be used responsibly.

5.5 Safety Analysis

The two ways the selected corn yield model could fail is through underpredicting or over predicting yield. Underpredicting yield could result in farmers having a lack of resources to manage crops. If more crops survive throughout the growing stage, there might not be enough water, fertilizer and pesticides to properly manage them. If there is insufficient storage for the extra yield, these crops could spoil. Over predicting yield could result in farmers losing profits from harvesting as much less corn than expected. It could result in them purchasing more resources than needed to manage their crops. Farmer's income is dependent on their yield, so it is essential that the model's predictions are accurate. Using MAPE as a metric, the selected model's output was over 70% accurate. Creating models with higher accuracy will prevent these errors in yield prediction. Disclosing that the errors in predictive models will allow farmers to understand the uncertainties in these model outputs.

6. Assessment

6.1 Summary

Crop predictive models promote food security by providing valuable information to producers. With a changing climate, variations in crop yield factors have significant impacts on crop production. Machine learning models have significant predictive power to combat these uncertainties. Multiple models were constructed to predict corn yield on a county level for the midwest including random forest, ridge regression, multilayer Perceptron, and Support Vector Regression. Various yield predictive factors were inputted into the models to provide the models with essential information. Past corn yield data from the National Statistics Service and vegetation index and soil remote sensing data from Google Earth engine were included in the model inputs. The desired model output was 70% accuracy, calculated from MAPE. All four of the models outperformed this accuracy goal. Ridge regression had the greatest accuracy with a 90.9% mean accuracy, an R^2 value of 0.86, and an RMSE of 15.57. It is crucial that the crop predictive models output accurate data to ensure farmers and policymakers can utilize the data to combat the food security crisis.

6.2 Next Steps

While this model was very successful in producing accurate corn yield predictions based on only free and publicly-available data, there is room for models like this to be even better. The model is very accurate in predicting years with fairly standard weather, but the model is not as accurate in predicting outliers such as the 2012 drought year. A higher number of sample years would give the model more data to base predictions on, so the accuracy will increase over time as more training data is made available. Another room for improvement is that null values in the training set could be filled with better methods, such as interpolation, to increase model accuracy. Looking forward, weather predictions could be implemented into the predictive models to account for the impacts of future events on corn yield.

7. References

Basso, B., & Liu, L. (2019). Seasonal crop yield forecast: Methods, applications, and accuracies. *Advances in Agronomy*, *154*, 201–255. https://doi.org/10.1016/bs.agron.2018.11.002

Discusses why crop yield prediction is important, how the information is used, and methods of prediction.

Becker-Reshef, I., Vermote, E., Lindeman, M., & Justice, C. (2010). A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using Modis Data. *Remote Sensing of Environment*, *114*(6), 1312–1323. https://doi.org/10.1016/j.rse.2010.01.010

Describes the construction of a wheat yield prediction model that achieved high accuracies (7-10% error). Advocates for the use of remote-sensing data as a great source of predictive information and describes various methods to do so.

Bocca, F. F., Rodrigues, L. H., & Arraes, N. A. (2015). When do I want to know and why? different demands on sugarcane yield predictions. *Agricultural Systems*, *135*, 48–56. https://doi.org/10.1016/j.agsy.2014.11.008

Discusses how yield forecasting information is used in sugarcane production and processing. This can help us explain how prediction data is used and why it is important.

Bolton, D. K., & Friedl, M. A. (2013). Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology*, *173*, 74–84. https://doi.org/10.1016/j.agrformet.2013.01.007

A forecasting model for soybean and maize yield was created using remote sensing data. They found that EVI2 is a better predictor of crop yield than NDVI for maize

Breiman, L. Random Forests. *Machine Learning* 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

An article that overviews the use of the random forest model with remote sensing data.

Cai, Y., Moore, K., Pellegrini, A., Elhaddad, A., Lessel, J., Townsend, C., ... & Semret, N. (2017, December). Crop yield predictions-high resolution statistical model for intra-season forecasts applied to corn in the US. In 2017 Fall Meeting. Gro Intelligence Inc.

Talks about a predictive model that was built and tested live throughout the 2016 corn growing season, which compared similarly to USDA predictions. Talks about the importance of corn yield predictive models.

Chaya. (2022, April 14). *Random Forest Regression*. Medium. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Used for Random Forest figure 1.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ. Computer science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

Provides information on R^2, RMSE, and MAPE. Describes the functions and their uses.

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters, 13(11), 114003.

Discusses corn yield prediction modeling. They used a Deep Neural Network variation with good results. Talks about global warming and its impact on modeling.

de Myttenaere, A. (2016). Mean Absolute Percent Error for Regression Models. Neurocomputing. https://doi.org/10.1016/j.neucom.2015.12.114Get rights and content

Finger, R. (2010), Revisiting the Evaluation of Robust Regression Techniques for Crop Yield Data Detrending. Amer. J of Ag. Econ., 92: 205-211. https://doi.org/10.1093/ajae/aap021

A summary of the use of regression analysis on crop yield data. Discusses the use of OLS.

Gao, B. C. (1996). NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote sensing of environment, 58(3), 257-266.

Talks about NDWI, what it is, and why it is useful.

Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences. *Atmospheric Environment*, *32*(14), 2627–2636. https://doi.org/10.1016/S1352-2310(97)00447-0. Gives an overview of how the multilayer perceptron model works and its applications.

Gitelson, A. A., Verma, S. B., Viña, A., Rundquist, D. C., Keydan, G., Leavitt, B., Arkebauer, T. J., Burba, G. G., & Suyker, A. E. (2003). Novel technique for remote estimation of CO2 flux in maize. *Geophysical Research Letters*, *30*(9). https://doi.org/10.1029/2002GL016543. Found the GCI equation.

Glenn, E., Huete, A., Nagler, P., and Nelson, S. 2008. “Relationship Between Remotely-Sensed Vegetation Indices, Canopy Attributes and Plant Physiological Processes: What Vegetation Indices Can and Cannot Tell Us About the Landscape.” *Sensors* 8 (4) (March 28): 2136–60. doi:10.3390/s8042136. http://dx.doi.org/10.3390/s8042136. Explains the uses of the NDVI and the EVI.

Guan, K., Wu, J., Kimball, J. S., Anderson, M. C., Frolking, S., Li, B., Hain, C. R., & Lobell, D. B. (2017). The shared and unique values of optical, fluorescence, thermal, and microwave satellite data for estimating large-scale crop yields. *Remote Sensing of Environment*, *199*, 333–349. https://doi.org/10.1016/j.rse.2017.06.043

The conventional approach to using satellite remote sensing data is through Vegetation Index. This paper outlines other, less widely used methods of incorporating remote sensing data into a yield predicting model.

IBM Cloud Education. (n.d.). What is overfitting? IBM. Retrieved May 7, 2022, from https://www.ibm.com/cloud/learn/overfitting#:~:text=Overfitting%20is%20a%20concept%20in,unseen%20data%2C%20defeating%20its%20purpose.

Explains what data overfitting is and how to avoid it.

Johnson, D. M. (2014). An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sensing of Environment*, *141*, 116–128. https://doi.org/10.1016/j.rse.2013.10.027

Corn and soybean yields were modeled using NDVI, precipitation and remote sensing daytime LST and nighttime LST.

Kang, Y., Khan, S., & Ma, X. (2009). Climate change impacts on crop yield, crop water productivity and food security – a review. *Progress in Natural Science*, *19*(12), 1665–1674. https://doi.org/10.1016/j.pnsc.2009.08.001

This paper summarizes the usage of climate models and climate changes impact on water availability, crop yield and food security. It might be useful for describing how the model can be used and why it's important.

Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C., & Anderson, M. (2020). Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US midwest. *Environmental Research Letters*, *15*(6), 064005. https://doi.org/10.1088/1748-9326/ab7df9

Describes and compares various algorithms for predicting corn yield. Will be helpful in developing the models.

Kaul, M., Hill, R. L., & Walthall, C. (2005). Artificial neural networks for corn and soybean yield prediction. *Agricultural Systems*, *85*(1), 1–18. https://doi.org/10.1016/j.agsy.2004.07.009

This paper details the process of developing a neural network yield prediction model for corn and soybeans in Maryland. Their process and findings could be helpful in developing our model.

Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S. (2021). Review on Convolutional Neural Networks (CNN) in vegetation remote sensing*, ISPRS Journal of Photogrammetry and Remote Sensing*, 173, 24-49. https://doi.org/10.1016/j.isprsjprs.2020.12.010

This article discusses the use of Convolutional Neural Networks (CNN) in modeling vegetation remote sensing data. Specifically analyzing the effects different spatial resolutions, sensor types and data types have on the model.

Khaki, S., & Wang, L. (2019). Crop Yield Prediction Using Deep Neural Networks. *Front. Plant Sci., 10*, https://doi.org/10.3389/fpls.2019.00621

Journal provides details on how deep neural networks operate. Images were utilized in the technical review.

Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N., & Shearer, S. (2018). Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. *Computers and Electronics in Agriculture*, *153*, 213–225. https://doi.org/10.1016/j.compag.2018.07.016

Talks about the need for low-cost yield maps, and how remote sensing data and machine learning algorithms can be used quite effectively for this purpose. It details the soil properties that they included in their data, and the different models that they used. Machine learning algorithms outperform linear regression models.

Li, Y., Guan, K., Schnitkey, GD., DeLucia, E., & Peng, B (2019). Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States. *Glob Change Biol*., *25*: 2325– 2337. https://doi.org/10.1111/gcb.14628

This paper details the importance of crop prediction models because of increasing drought conditions from climate change. I cited it in the technical review.

L. V. Jospin, H. Laga, F. Boussaid, W. Buntine and M. Bennamoun, "Hands-On Bayesian Neural Networks—A Tutorial for Deep Learning Users," in IEEE Computational Intelligence Magazine, vol. 17, no. 2, pp. 29-48, May 2022, doi: 10.1109/MCI.2022.3155327.

Background information on Bayesian Neural Networks. Discusses their pros and cons.

Ma, Y., Zhang, Z., Kang, Y., & Özdoğan, M. (2021). Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. *Remote Sensing of Environment*, *259*, 112408. https://doi.org/10.1016/j.rse.2021.112408

This is the article that the crop prediction yield models will be based on. It describes the methods of developing a prediction model and emphasizes the importance of quantifying uncertainty in predictions.

McNunn, G., Heaton, E., Archontoulis, S., Licht, M., & VanLoocke, A. (2019). Using a crop modeling framework for precision cost-benefit analysis of variable seeding and nitrogen application rates. *Frontiers in Sustainable Food Systems*, *3*. https://doi.org/10.3389/fsufs.2019.00108

Discusses the use of publicly available data and modeling to reduce nutrient losses, improve sustainability, and increase the profitability of farming practices. Useful to describe the impact of the models.

NASS (2022). *Surveys.* United States Department of Agriculture National Statistics Service. https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Agricultural_Yield/index.php

Used to provide background for NASS Historical Corn Survey Data

Peng, B., Guan, K., Pan, M., & Li, Y. (2018). Benefits of seasonal climate prediction and satellite data for forecasting U.S. maize yield. *Geophysical Research Letters*, *45*(18), 9662–9671. https://doi.org/10.1029/2018gl079291

Shows that incorporating enhanced vegetation index to a predictive model's inputs can greatly improve performance.

Ray, D. K., Gerber, J. S., MacDonald, G. K., & West, P. C. (2015). Climate variation explains a third of global crop yield variability. *Nature Communications*, *6*(1). https://doi.org/10.1038/ncomms6989

This article quantifies the variables that affect crop yield variation. It will help us decide which data inputs should be included in the models.

Rouse, W., & Haas, R. H. (1973). *MONITORING VEGETATION SYSTEMS IN THE GREAT PLAINS WITH ERTS*. 9.

This article is the first usage of NDVI and EVI.

Tian, H., Wang, P., Tansey, K., Zhang, J., Zhang, S., Li, H. (2021). An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong Plain, PR China, *Agricultural and Forest Meteorology*, 310, https://doi.org/10.1016/j.agrformet.2021.108629.

Discusses the use of the LSTM model with remote sensing data in an agricultural setting.

USDA NASS Cropland Data Layers, Earth Engine Data Catalog. https://developers.google.com/earth-engine/datasets/catalog/USDA_NASS_CDL

Includes background for Google Earth Engine input data

https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Agricultural_Yield/index.php

Used to provide background for NASS Historical Corn Survey Data

Walsh, J. E., & Allen, D. (1981). Testing the Farmer's Almanac. *Weatherwise*, *34*(5), 212-215.

Information regarding farmer’s almanacs and their accuracy.

Wang, A. X., Tran, C., Desai, N., Lobell, D., & Ermon, S. (2018). Deep Transfer Learning for crop yield prediction with Remote Sensing Data. *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, 1–5. https://doi.org/10.1145/3209811.3212707

Describes how remote sensing data can be a valuable input for crop yield prediction.

Wang, Y., Zhang, Z., Feng, L., Du, Q., & Runge, T. (2020). Combining multi-source data and machine learning approaches to predict winter wheat yield in the conterminous United States. *Remote Sensing*, *12*(8), 1232. https://doi.org/10.3390/rs12081232

Compares the performance of a variety of predictive models for winter wheat yield. Concludes that machine-learning models generally perform better than linear regression models.

Westfall, A. T. (2021, July 6). Knee high by the 4th of July. Extension. Retrieved May 7, 2022, from https://extension.purdue.edu/news/county/white/2021/07/Knee-High-by-the-4th-of-July.html

Discusses the origins of the phrase knee high by the 4th of July.

https://naldc.nal.usda.gov/catalog/IND43966364

Yemelyanov, V., Yemelyanova, N., Shved, E., Nedelkin, A. & Fatkulin, A. (2020). Modeling of the Multilayer Perceptrons for Image Recognition of the Steel Microstructures. *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 952-955. 10.1109/EIConRus49466.2020.9038971.

You, J., Li, X., Low, M., Lobell, D. B., & Ermon, S. (n.d.). *COMBINING REMOTE SENSING DATA AND MACHINE LEARNING TO PREDICT CROP YIELD*. Sustainability and artificial intelligence lab. Retrieved April 13, 2022, from http://sustain.stanford.edu/crop-yield-analysis

A project with a similar goal to ours. Details their process, techniques, etc. Describes why a corn yield prediction is important.

8. Appendices

Appendix A - Data Sources

Data from the USDA was utilized to construct the models. The USDA two-level surveys provide state and national yield estimates (Johnson, 2014). To obtain data for these surveys, the USDA conducts telephone interviews with farmers during the row crop growing season from August to November (NASS).

Google Earth Engine provided vegetation index and soil remote sensing data. Crop-specific land cover data is created annually for the continental United States through the Cropland Data Layer (CDL). CDL is composed of satellite imagery and ground truth data collected from the USDA, the National Statistics Service, Research and Development Division, Geospatial Information Branch, and Spatial Analysis Research Section (Google Earth Engine).

Appendix B - Design Alternatives

Convolutional Neural Network (CNN): This type of deep neural network was specially designed to process spatial imagery data, such as remote sensing data (Kang et al., 2020). Neural networks have an input layer, multiple hidden layers, and an output layer. Every input is connected to every node in the hidden layer, where the network learns the strength of each connection. When utilized properly CNN is accurate and its ability to be used on spatial imagery is especially applicable for remote sensing data that will be used in estimating corn yield (Kattenborn et al., 2021). Neural networks require training which makes them more difficult to implement.

Multilayer Perceptron (MLP): Multilayer Perceptron is a simplistic neural network that makes no assumptions of data distribution and can model nonlinear data. MLP has multiple layers of perception and is a feed-forward model. A feed-forward model has data inputted into the input layer and the layer's output is scaled and fed forward as an input to the next layer. (Gardner, 1998). The model works similarly to the shallow neural network displayed in figure 2. MLP is built up of several nodes in the input layer and hidden layer, while one node is in the output layer with connections between the nodes. The connections are weights between the nodes. They scale the outputs of the node to the input of the next node. The number of nodes and connections can vary for each problem but adding too many nodes/connections can cause overfitting while having too few connections/nodes provides insufficient information for the model to create an accurate solution (Ramchoun, 2016). Multilayer Perceptron is used in applications for image and speech recognition (Yemelyanov et al., 2020).

Long-Short Term Memory (LSTM): LSTM is a modified recurrent neural network (RNN). A recurrent neural network loops the output from a hidden layer node back into itself during the next iteration “remember” its previous state. LSTM remembers and updates its memory each iteration, adding new information and “forgetting” what it determines to be irrelevant. LSTM is useful when working with time-related data and non-linear fits (Tian et al., 2021).

Random Forest (RF): Random Forest is a supervised machine learning model that randomly selects a set number of features (Temperature, Precipitation, Soil Data, etc.) from a list to build decision trees. Layers of decision trees are built to create the “Forest” (Breiman, 2001). Training data is required to teach the model the relationship between features and known outputs. Once trained the model will output predicted values when given feature values. The output of each tree is averaged for regression or the most popular designation is chosen for classification. RF has been used commonly with remote sensing data (Breiman, 2001).

Ordinary Least Squares (OLS): OLS is the most well-known linear model, where root sum squared (RSS) is minimized to find the best-fit equation for the data given. This model is particularly simple and easy to implement, however, it is most effective with small samples and few outliers (Finger 2010). For this reason, it is considered the least desirable model for this project.

Ridge Regression: This model is a modified version of OLS regression where it minimizes MSE instead of RSS by adding a second term to RSS that reduces overfitting (Bolton and Friedl, 2013). Ridge regression is superior to linear regression in applications where several variables are similarly predictive of the output.

Support Vector Regression (SVR): SVR is a regression model that sets an “acceptable” deviance from a hyperplane, and the model maximizes the number of points within the set threshold. SVR uses Support Vector Machines to define a hyperplane. A hyperplane is simply a plane one dimension smaller than the actual plane, so the hyperplane of a 3D space is a 2D surface. In the context of SVR, the hyperplane is used to separate objects of different classifications. The optimized hyperplane equation is used to predict the output in this model (Kang et al., 2020).

Appendix B.1 - Selection Methodology

The models selected were chosen based on their accuracy in previous studies and if they were available in Scikit-learn libraries. Multilayer Perceptron, Random Forest, Ridge Regression and Support Vector Regression met these criteria.

Appendix B.2 - Model Type Evaluation Criteria:

Computational Ease: Describes how difficult it is to manipulate the data for use in the model.

Implementational Ease: Defines how difficult it is to understand and run the model. Models that require training data or complex construction are less desirable.

Accuracy: How well the models can predict outcomes, overall. Accuracy will vary with data type and application, however, some models typically perform better than others.

Applicability: How well the function of each model fits the intended use. For example, a model that can handle several parameters is more applicable to the project.

Weights: An accurate predictive model is the primary goal of our project and is the most important consideration in choosing a model. Computational and implementational ease describe the time and effort required to use certain models, which is a minor design consideration in comparison to accuracy. Applicability is also considered lesser than accuracy.

Design Matrix

A higher value is more desirable (1-poor, 10-excellent)

Criteria:	Weight	MLP	LSTM	RF	OLS	Ridge	SVR
Computational ease	0.2	5	6	7	9	8	6
Implementational ease	0.2	4	4	5	9	8	7
Accuracy	0.4	8	8	6	2	4	7
Applicability	0.2	6	6	6	2	5	4
Weighted Sum		6.2	6.4	6	4.8	5.8	6.2

Long-Short Term Memory has the greatest weighted sum score, however, CNN, RF, Ridge, and SVR all had similar scores to LSTM. OLS was the greatest outlier in weight sum, scoring much lower than the other models. Due to the lack of a clear best option, we have decided that utilizing multiple models is the best option. LSTM, CNN, RF, Ridge, and SVR will all be used to predict the corn yield for the 12 midwestern states. The accuracy of each model's prediction will be analyzed, with the strongest one eventually being implemented. The extra effort required to create several yield models is required to ensure the greatest accuracy, which allows the users of the final product to make the best decisions.

There are numerous other decisions that need to be made in the design and creation of a machine-learning model. This includes considerations such as the following:

Types of Input Data - The performance of the models will depend heavily on the input data that is used. Many of the possible data inputs are outlined in the Technical Review and Design Requirements sections of this report. This project requires iteration and optimization of the predictive models, and input data will be included based on the resultant model performance. For that reason, specific data sets are not evaluated in this section and instead will be evaluated in the model itself.

Programming Language - Python has been selected as the language to use for this project, on the recommendation of our client and its widespread use for such projects. Our client has also recommended that scikit-learn Python libraries are used, which will provide useful tools to streamline the implementation, iteration, and evaluation processes.

Visualization Tool - QGIS has been selected as the visualization tool for creating yield maps. It was recommended by the client and provides all of the functions necessary to efficiently generate yield maps that can contrast the results of the various models against each other and actual yield data. Additionally, the application is free to use.

Date Range for Input Data - Growing season-only data will be used for the model. While non-growing season data can provide some information, our client advised that the majority of the predictive power comes from data gathered during the growing season. For example, weather in winter generally has little effect on crop yields.

The most ethical and sustainable model is the most accurate model. As outlined in the technical review, the ability to predict corn yields benefits farmers, food security, the economy, and the environment. A model that will predict future corn yields as accurately as possible will be the most sustainable, safe, and ethical design alternative.