Problem Statement

Can greenhouse gas emissions associated agriculture and land use practices be reduced by two-thirds between 2010 and 2050, by reducing fertilizer use in production (via practices like cover cropping and conservation tilling that promote soil health), while maintaining yields and preserving natural carbon sinks from conversion to agricultural land (“Efficiency”)?

Data

Data for this capstone is compiled from several sources, including USDA, Indiana State Department of Agriculture, HPRCC, Michigan State University, and the NOAA. The data scraped from these sources is very limited, it only spans 28 years. Training machine learning models on such a limited amount of data will be difficult and require techniques to synthetically grow the number of observations in the training set.

Links to scraped data:

* + Precipitation & AVG Temp - <https://github.com/fromo19/Capstone-2/blob/Final/IN%20Climate.csv>
  + IN Corn Acres - <https://github.com/fromo19/Capstone-2/blob/Final/IN%20Corn.csv>
  + IN Corn Yield - <https://github.com/fromo19/Capstone-2/blob/Final/IN%20yield.csv>
  + US Corn Acres - <https://github.com/fromo19/Capstone-2/blob/Final/US%20Corn.csv>
  + Total US Cropland - <https://github.com/fromo19/Capstone-2/blob/Final/US%20Total.csv>
  + Corn GHG Emissions (Michigan State Calculator) - <https://github.com/fromo19/Capstone-2/blob/Final/US%20Cropland%20Greenhouse%20Gas%20Calculator.pdf>
  + IN Corn Historical GDD - <https://github.com/fromo19/Capstone-2/blob/Final/corn_gdd.csv>
  + IN Conservation Tillage Data - <https://github.com/fromo19/Capstone-2/blob/Final/conservation_till.pdf>
  + IN Cover Crop Data - <https://github.com/fromo19/Capstone-2/blob/Final/cover_crop.pdf>
  + IN No-till Data - <https://github.com/fromo19/Capstone-2/blob/Final/no_till.pdf>
  + Fertilizer Usgae - <https://github.com/fromo19/Capstone-2/blob/Final/fertilizeruse.xls>

Method

There are several established machine learning methods to predict the value of continuous dependent variables like our target variable, Efficiency (corn yield per metric ton of NO2 emission). In this capstone I will explore the use of linear based regression techniques like:

1. Regularization Embedded Methods – these methods add regularization constraints to reduce coefficients of less important features and simplify the model.
   * Ridge regression: ridge regression shrinks the coefficients of the least important features closer and closer towards the origin (shrink to 0), without actually setting the coefficient to 0. Ridge regression is most useful when most of the features are useful.
   * Lasso regression: lasso regression is similar to ridge regression; however, lasso regression will completely eliminate unimportant features. Lasso regression is most useful when the model has many useless features.
   * Elastic net: Combines ridge and lasso regression penalties. It groups and shrinks the coefficients of correlated variables, and either removes them completely or leaves them in the equation at a level near zero. Elastic net is great for handling data with many correlated variables, and when it is uncertain the importance of a feature.
2. Ensemble Methods – methods in which multiple weak models (learners) are combined together to generate a singular strong model, this includes bagging and boosting techniques. Each weak model acts as a voter. The votes are then combined to form the basis of the model’s prediction.

* Random Forest Regression: Is a bagging technique which a regression-based trees are trained on a bootstrapped sample of the training data, this process done numerous times in parallel to create an ensemble of weak learners. All learners are independent, which decreases the variance of the model. Ideal for decreasing the overfitting of a model.
* XGBoost Regression: Boosting technique in which many weaker learners are created in sequence. This is done by training a single weak learner, then using it to make predictions on the rest of the training data. Incorrect predictions are given higher weights and then the training data is resampled and a new learner is created. This process repeats for as many learners as the modeler chooses and the all the weak learners are combined into a strong learner through averaging. This process creates learners that are dependent on each other. This decreases model bias and is most beneficial for underfitted models.

Data Cleaning

<https://github.com/fromo19/Capstone-2/blob/Final/Data%20Wrangling%20-%20Final.ipynb>

In order for data to be passed to a regression-based model, there can be no incomplete information for observations, and none of the variables may be categorical. The data was scraped from multiple sources, individually cleaned, and compiled into a single uniform Pandas dataframe prior to analysis.

Main problems wrangling data include:

* Years missing cover crop data - missing values were assumed to be zero, because even though cover cropping has been a farming technique for decades the practice has only recently been widely adopted and acreage data recorded. Years missing data post 2010 were imputed with the average of the year preceding and following said record.
* Limited volume – the total number records in the final scraped dataframe was extremely low at 28, and not suitable for machine learning purposes. The size of the data had to be grown in order to properly train the models. In order to achieve this a combination of SMOTE/ADASYN and Bootstrapping was used to grow the data to 300 records (Performed after initial EDA).

EDA

<https://github.com/fromo19/Capstone-2/blob/Final/EDA%20-%20Final.ipynb>

The data is not missing any values and contains no categorical variables. The dependent variable ‘efficiency’ appears to be normally distributed as verified by the histogram and ECDF. All of the explanatory variables have correlation coefficient signs that are theoretically sound. ‘efficiency’ is positively correlated with ‘yield’ and ‘cover\_crop\_ratio’ and negatively correlated with ‘emissions’ and ‘full\_till\_ratio’. The year 2012 appears to be a particularly large outlier in terms of ‘efficiency’, ‘precipitation’, and ‘yield’. This suggest that this could’ve been a potential drought year. Due to the relatively small size of the data, separate models will be run to explore the effect of this and other outliers on the predictive power of the models. Simple linear regression yields a coefficient of determination of .631, suggesting that the independent features do explain a decent amount of the variance in ‘efficiency’.

Preprocessing

<https://github.com/fromo19/Capstone-2/blob/Final/Preprocessing%20-%20Final.ipynb>

Due to the extremely small nature of the data sample size, synthetic growth techniques (SGTs) and bootstrapping were necessary to perform prior to model training. Techniques like SMOTE and ADASYN can grow sample size without duplicating entries. SGTs create new data points by ‘polling’, the user designated, K\_nearest\_neighbors to each point in the minority class until its size is balanced with that of the majority class.

In the wrangled data there were significantly more years without recorded cover crop information (minority class) than years with cover crop information (majority class). In order to generate more records with cover crop data the 3 nearest neighbors of each cover crop year were used to randomly create new points between the original points. Training and test splits must be done prior to use of these techniques as the newly generated points will contain information from all of its neighbors, thus potentially introducing bias to the models. ADASYN was chosen as the SGT because it generated points more normally distributed than those of SMOTE on the training set. After the implementation of SGTs, the data was then bootstrapped to further increase the sample size to one that is more machine learning ready.

Model Selection & Hyperparameter Tuning

<https://github.com/fromo19/Capstone-2/blob/Final/Model%20Selection%20-%20Final.ipynb>

Data was standardized and fit to the aforementioned regression models using Sklearn pipelines. Outliers in the dataset had significant impacts on model performance. To make a final decision on the most appropriate model, all models were trained on data from which outlier years were removed. RMSE was used to evaluate the performance between different models, the lower the RMSE the better performance of the model.

The models that performed the best (1 from the regularization embedded methods and ensemble methods) were:

1. Regularization - Ridge
2. Ensemble – XGBoost

Both models performed similarly well; However, given the small sample size I worry that the XGBoost model may have been overfitting to the training data. The training R2 score for the XGBoost model was .99, meaning that the model had fit to the training data almost perfectly. This overfitting caused the model performance on the test set to lag with a R2 score of .76. Therefore a ridge model was explored. After tuning alpha, the Ridge model returned a R2 score of .6901 which is just below the R2 score of the Ridge model on the training set .6941. Due to the overfitting witnessed in the XGBoost model, I would recommend using the Ridge regression model to predict efficiency.

Data Issues and Future Improvements

One of the main issues with advanced machine learning models like XGBoost, is the need for large amounts of data. Unfortunately, the data needed to evaluate this capstone problem is very limited. This led to the need to synthetically grow the data prior to model training. Train and test splits must be done before using algorithms like ADASYN or SMOTE, because failing to do so will allow information from the test set to leak into the training set through the generation of new data points. This in turn created the need to split the dataset prior to growth, and further decreased the potential test data size. As a result, I was unable to obtain a large enough test set to suppress the models overfitting to the training set.

To improve this project in the future, I would recommend the collection of more data. Either through the passage of time, or the reduction in collection level (potentially by county or city). This will allow for the more data to become available in the test set, and thus allow for broader evaluation of the models’ ability to generalize.