Noah Andersen, Lincoln Bay, Eliza Campbell
4/14/2024

Utah Counties Agricultural Census - Predicting Gap Years

1. Introduction

The United States Department of Agriculture's National Agricultural Statistics Service (NASS) performs an "Agricultural Census" to see the state of agricultural production and health in the United States economy (USDA, 2023). It provides critical information that cannot easily be found anywhere else to inform farmers, investors, and policymakers about the effects of weather, climate, and institutional decisions. However, this census is only collected once every five years, and an accompanying survey performed every year frequently fails to collect enough data to publish results for all counties. In fact, if not enough farmers in one county fail to respond, the NASS frequently withholds data from other counties in the state to prevent third parties from calculating the unresponsive county's yield themselves. It's not because there isn't agricultural production there, either; which counties are not included changes year-to-year (Johanns & Thessen, 2020).

We plan to use environmental and economic data to generate a dataset filling in these gaps in non-census years for counties in which surveys produced insufficient responses in the state of Utah. Having additional data such as this one would likely prove helpful to each of the groups previously mentioned, who rely on agricultural data to make important decisions.

2. Data

We obtained our data from several sources. The United States Department of Agriculture's National Agricultural Statistics Service provided us with past agricultural censuses, as well as other information regarding crop yields. The U.S. Census, published by the United States Census Bureau, provided our team with information regarding population growth for each county, and the World Bank provided us with economic information, such as GDP by industry. Lastly, the National Oceanic and Atmospheric Administration has published detailed data regarding weather patterns and measurements that proved valuable for our analyses.

From these sources, we developed several variables to use in our predictions, such as county population, GDP, average temperature, wheat prices, precipitation, among others. A subset of the 82 features used in our analysis can be found in (Table 1)

Table 1

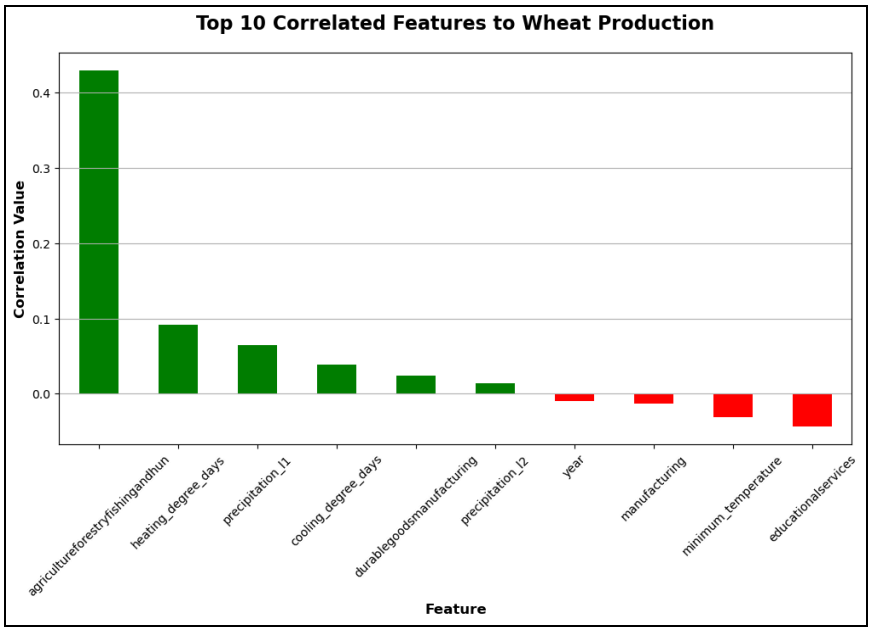
Summary statistics for a subset of the 82 features used in this analysis (numeric features). Structured on a year and county level.

Metric	Year	Wheat Prod. (Bushels)	Population	Wheat Price ($/Bushel)	Precipitation	Avg. Temp.
Count	**70**	**70**	**70**	**70**	**70**	**70**
Mean	**2012**	**370,369**	**179,167**	**7.7**	**14.7**	**49.4**
Std. dev.	**6.7**	**646,548**	**290,928**	**2.0**	**6.6**	**3.2**
Min	**2002**	**120**	**2,124**	**4.7**	**6.5**	**42.0**
Max	**2022**	**2,879,583**	**1,186,257**	**9.6**	**38.2**	**57.5**

In order to understand which features from our dataset were most correlated with the outcome variable of wheat production, a correlation matrix was calculated. Then, the top 10 most correlated features were found, the results of which can be found in (Figure 1). The most correlated feature with wheat production was shown to be the total GDP of that county from agriculture, while the second most correlated feature was found to be the number of heating degree days. These results are what one might expect; there is likely a strong relationship between the production level of a crop and the GDP of that product's industry. Similarly, an increase of heating degree days is likely to positively affect the production level of crops in that area.

Figure 1

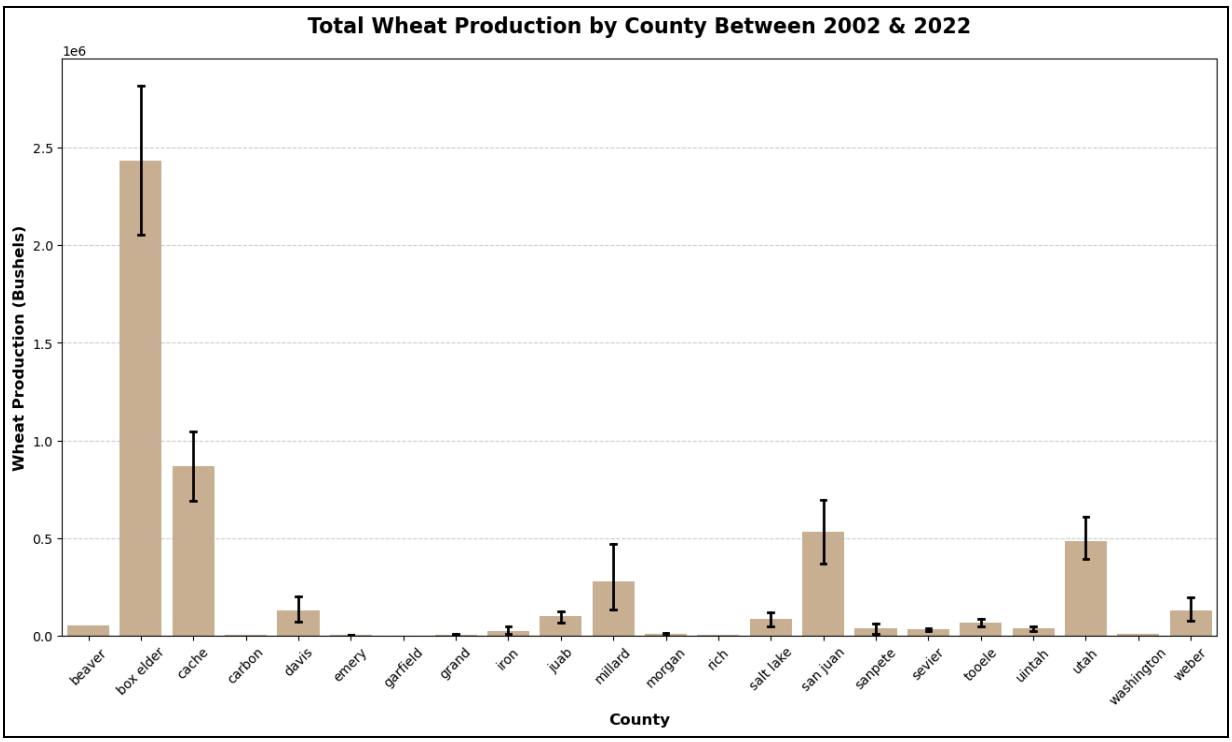
Top 10 most correlated features with wheat production.



The aggregated wheat production for each year by county can be found below in (Figure 2). It is evident that there is a disparity in the amount of wheat produced between different counties in the state of Utah. This disparity is likely due to the many features, such as population or average temperature, that we include in our model to generate our predictions.

Figure 2

Aggregated wheat production per county between 2002 & 2022



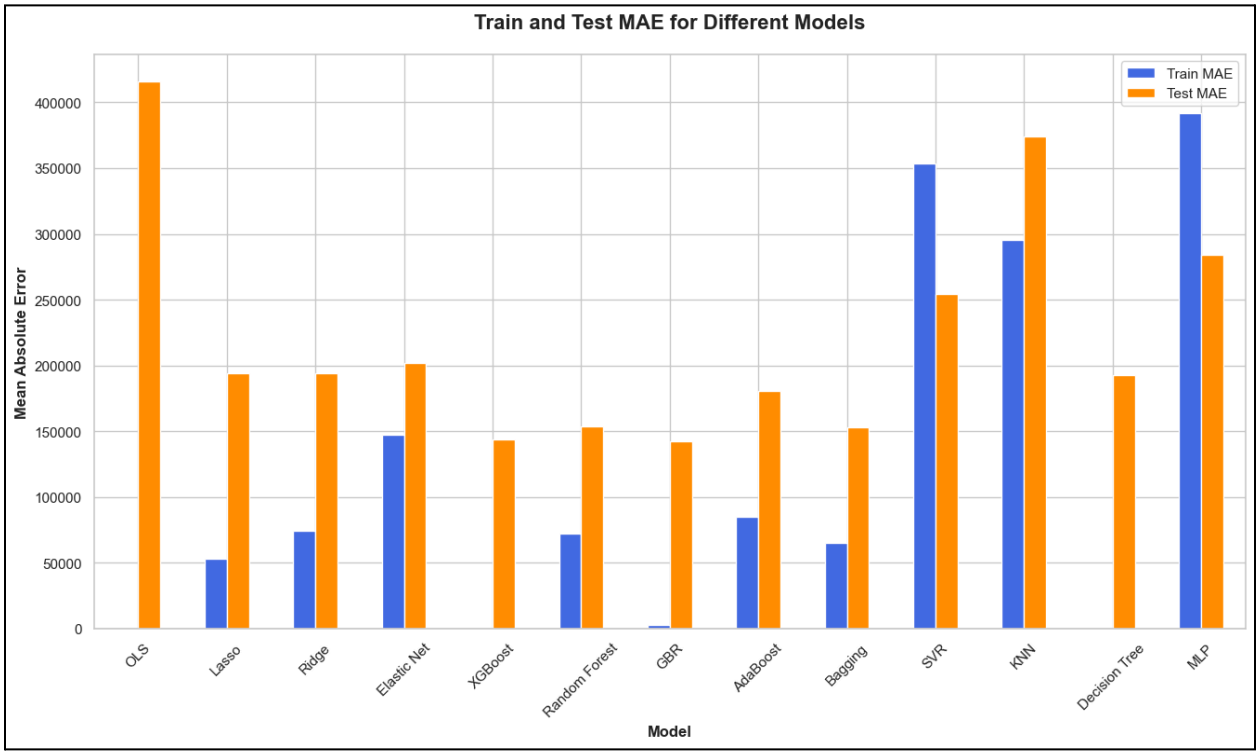
With this data aggregated and prepared for our analyses, we were able to proceed in our prediction process.

3. Methods

To predict the gap years we implemented various regression models. To validate these regression models different metrics such as R^2, mean absolute error, and mean squared error. The data was then partitioned off into a training set and a testing set for validation. We used a 75/25 train test split. Our input features were also normally scaled to mean zero and unit variance. The out of the box performance of these regression models can be found in (Figure 5). From these results, we found that the highest performing model based on the test set mean absolute error was XGBoost (Extreme Gradient Boosting), while a normal GBR (Gradient Boosting Regressor) followed closely behind in terms of test error.

Figure 5

Out of the box performance for various regression models.



For our preliminary results, we performed a Linear Regression with ElasticNet regularization. This form of regularization is a mix of Lasso and Ridge regressions, selecting coefficients that minimize the loss generated by (Equation 1).

Equation 1

Formula for Linear Regression with Elastic Net regularization.

$$\frac{\sum_{i=1}^{n}(y_i - x_i^J \hat{\beta})^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{m} \hat{\beta}_j^2 + \alpha \sum_{j=1}^{m} |\hat{\beta}_j| \right)$$

Following those preliminary results we focused on the highest performing model, XGBoost. To fine tune this model to the data, a 10 folder grid-search cross validation was performed to provide a general idea of what the best hyperparameters were. We leveraged this 10 fold cross validation to attempt to generalize the data between all possible train and test set possibilities, in case we happened to get lucky with an easy test set. Then, after we had obtained a good idea of what the ideal hyper parameters were, we individually fine tuned each parameter to minimize our train and test metric (mean squared error). This fine tuning drastically reduced the loss in our model, improving prediction performance greatly.

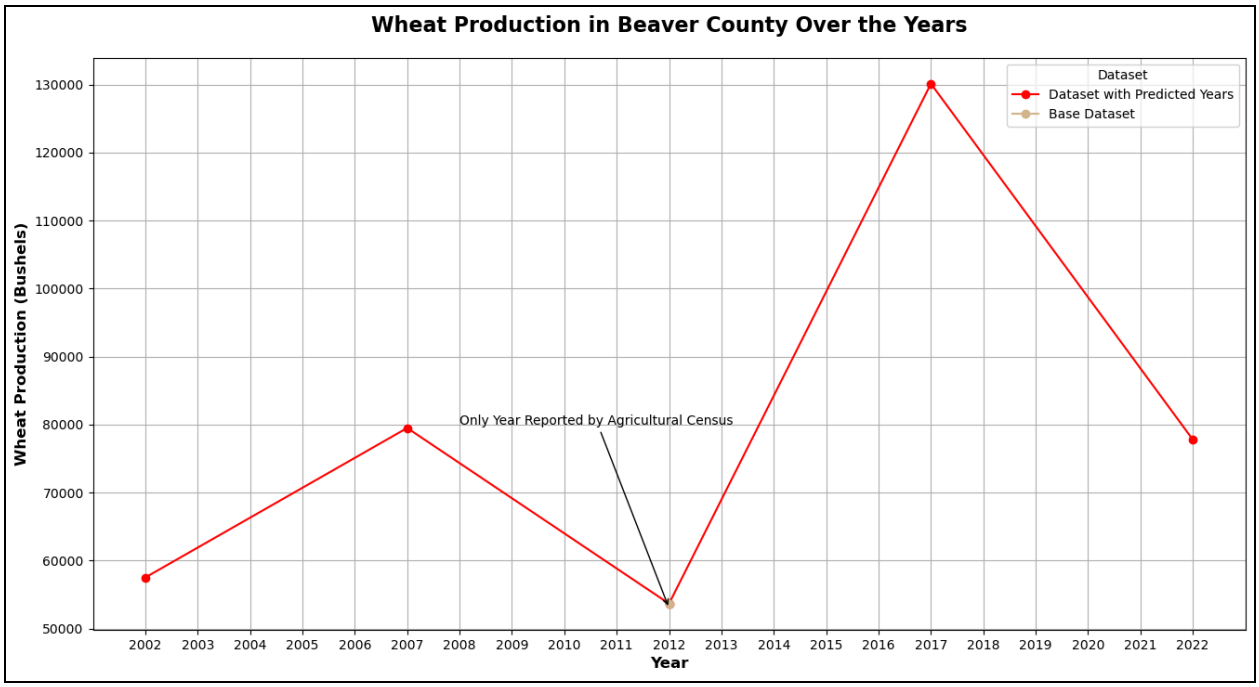
4. Results

Once we had fine-tuned the model to the data, we were then ready to predict the gap years that were not reported by the agricultural census. To complete this, we obtained the data for the missing county and years with the associated input features, scaled them similarly to how we had in training and testing, and sent the data through our model to predict the missing values. This venture was successful; our predictions resulted in an out-of-sample prediction accuracy based on an R^2 score of 0.748.

The results for two counties that were missing years from the census can be found in (Figure 3) and (Figure 4). As is shown in the figures, Beaver county was missing years 2002, 2007, 2017, and 2022 while 2017 was reported. Similarly, we can see that Tooele county was missing the years 2007, 2017, and 2022 with 2002 and 2012 being reported.

Figure 3

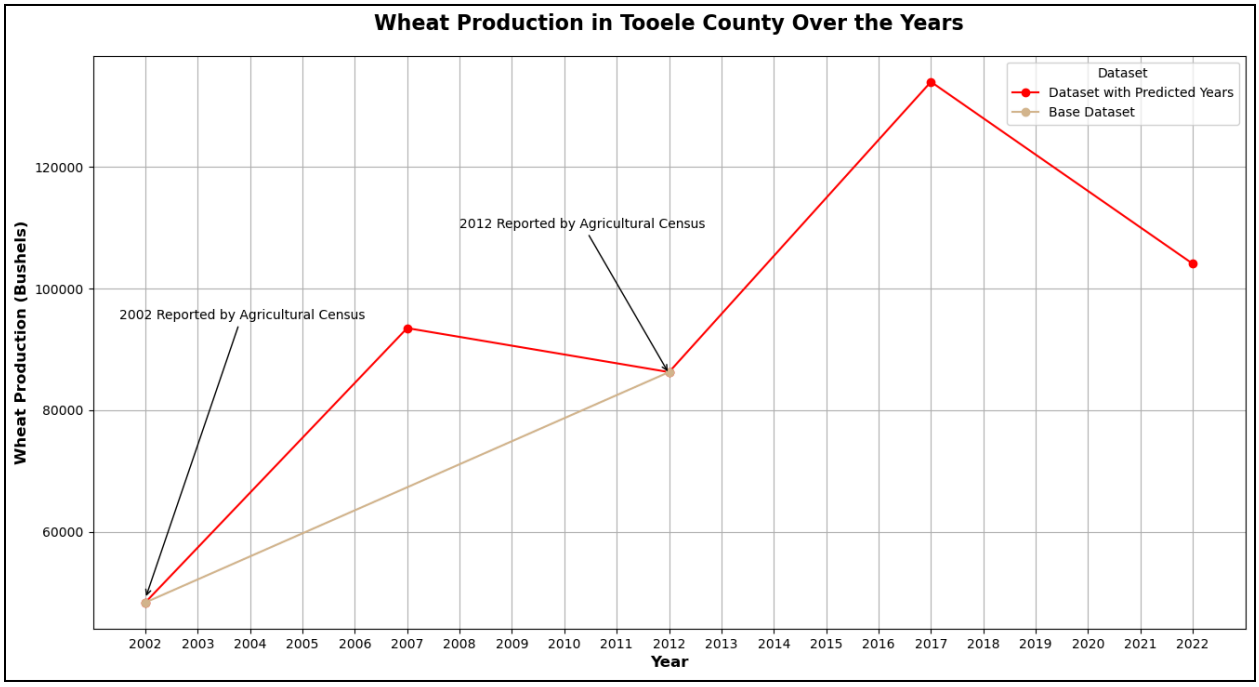
Wheat production in Beaver County with predicted years.



As is evident in the figure above, our predictions bring insight that would not have been otherwise available with the information reported to the public. In fact, those making decisions from the published data may be doing so from insufficient information, and could ultimately suffer the negative consequences of an uninformed purchase or policy.

Figure 4

Wheat production in Tooele County with predicted years.



Similarly, (Figure 4) sheds light on a situation that may not be so transparent to the public from the information reported in the agricultural census. This occurrence highlights the value of our project, and the benefit of using machine learning to predict data that is not readily available.

5. Conclusion

From our analyses, we were able to generate a dataset that effectively filled the gaps left by insufficient responses in the agricultural census. Our approach not only addressed the issue of missing data but also showed the potential of machine learning algorithms to predict agricultural data in a scenario where one might need such data to make important decisions. Specifically, the use of XGBoost highlighted the techniques one can use to optimize predictive power, while achieving favorable train and test error rates.

We believe that this process can be replicated in the future to generate data that is missing in various industries, especially the agricultural industry and those parallel to it. Our results, as well as those of anyone who replicates this project, can be used by policymakers, farmers, and third-party investors as a resource in their respective decision-making processes. Furthermore, as this undertaking becomes more refined and data becomes more available, it is likely that the predictions yielded by this method will become more accurate and applicable.