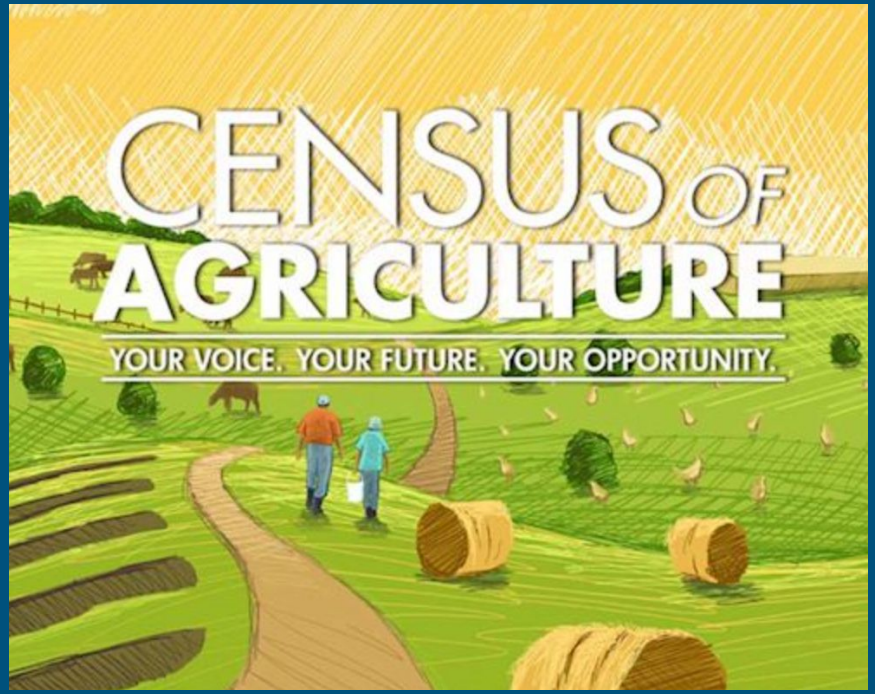
Generating Agricultural Data

Noah Andersen, Lincoln Bay, Eliza Campbell

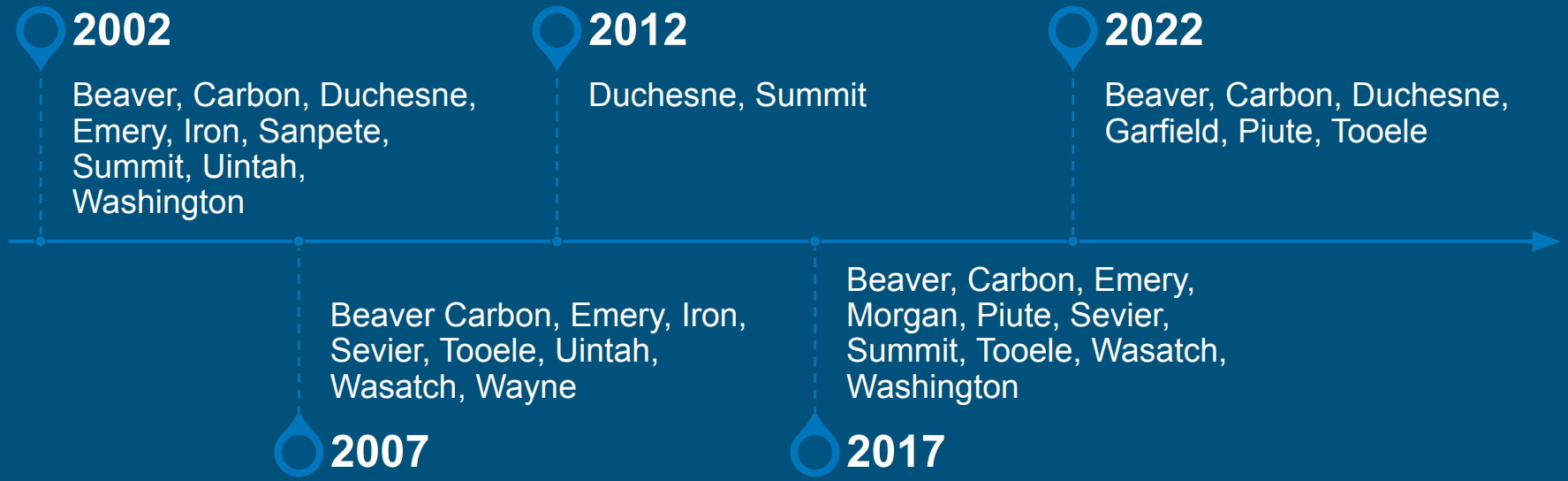
Our task



The Agricultural Census releases data every 5 years. If little enough data is gathered in a county that privacy is violated, that county's (and to prevent deduction, another county's) results are withheld.

We will be using the data from past years to fill in the gaps for counties where the Agricultural Census results were withheld.

County Year Combinations Missing



Motivation

Valuable information for:

- Farmers/Ranchers
- Policymakers
- Investors

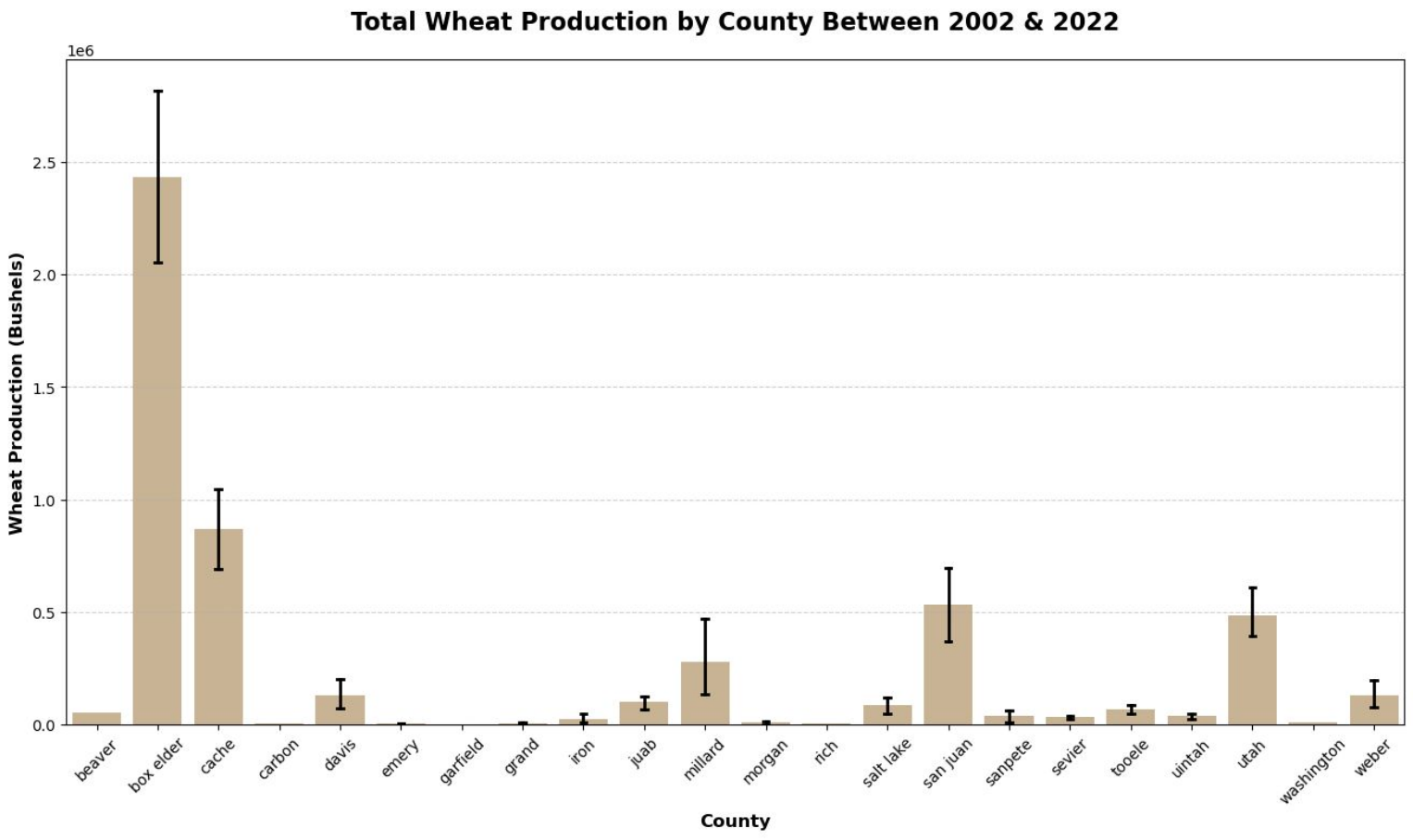
Variables

- County Population
- GDP (by industry & county)
- Industry-specific Covariates
- Wheat Prices
- Wheat Production
- Weather Data

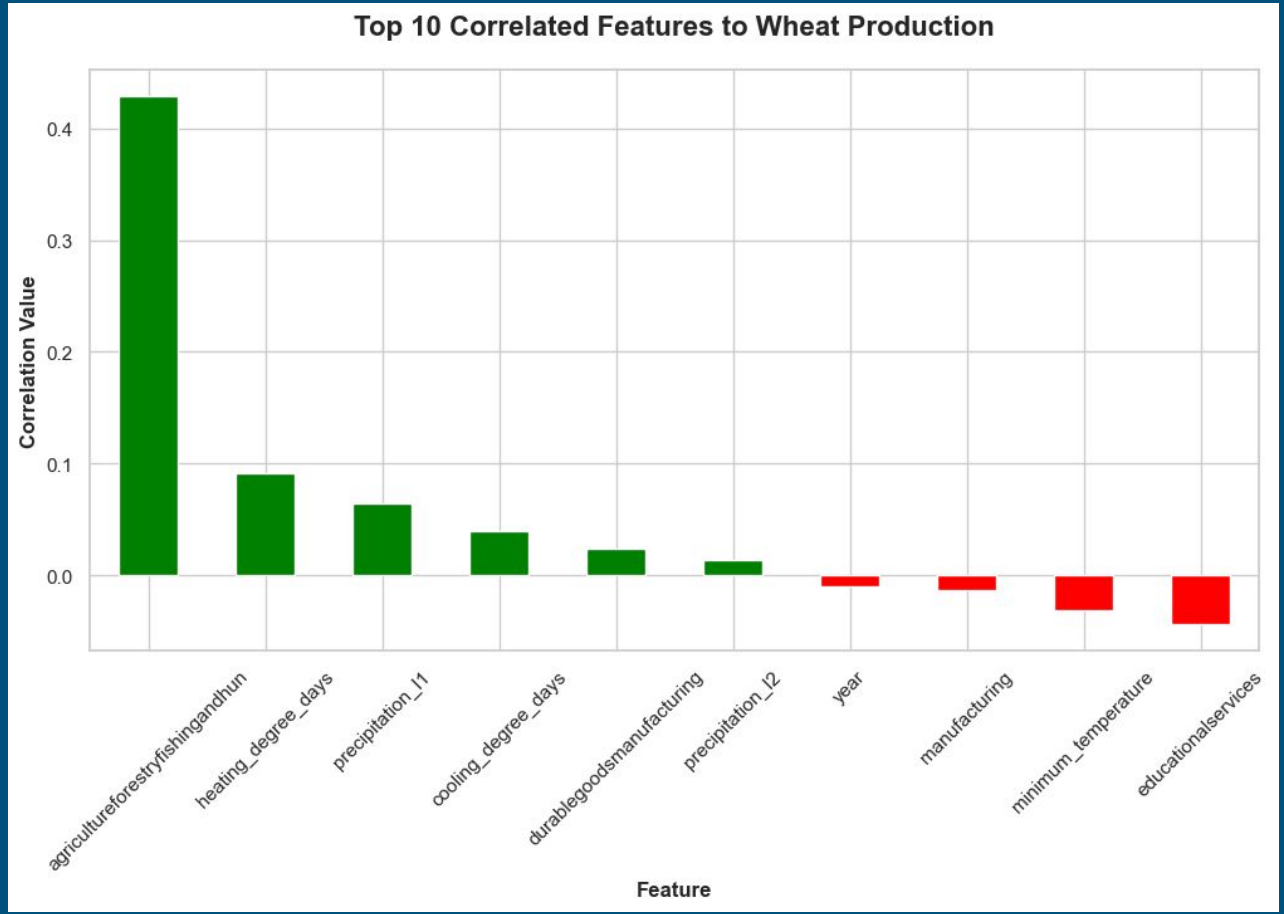
Sources

- US Census, United States Census Bureau
- The World Bank
- National Agricultural Statistics Service (US Dept. of Agriculture)
- NOAA (National Oceanic and Atmospheric Administration)

Wheat Production:



Feature Correlation:



ML methods we are using & how we chose tuning parameters and our out of sample accuracy

- For our preliminary results we performed a Linear Regression with Elastic Net regularization ~ Mixture of Ridge and Lasso regularization.

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

- For our fine tuned results we settled on the highest performing out of the box regression model based on R2 and MAE. The highest scoring model was XGBoost. The hyperparams were then fine tuned based on test set MAE with a 10 Fold CV to confirm hyperparameters.

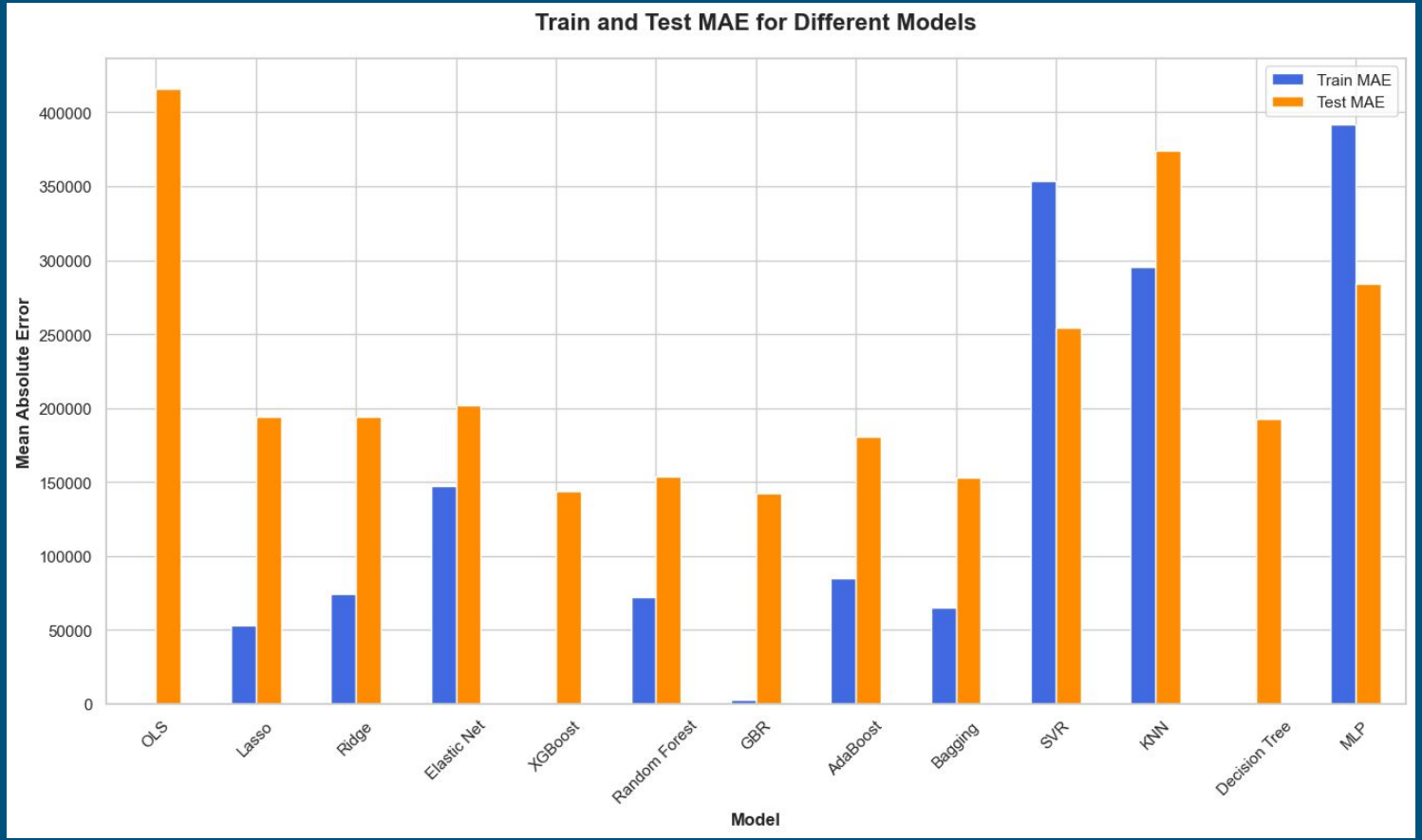
Preliminary Results - Elastic Net

- Parameters obtained via Grid Search cross validation
- Optimal Lasso and Ridge tuning parameters: identically 0.01

Training set accuracy	0.984
Test set accuracy	0.748

- Accuracy defined as the R^2 of predictions with respect to true values

Base Regression Models Performance:

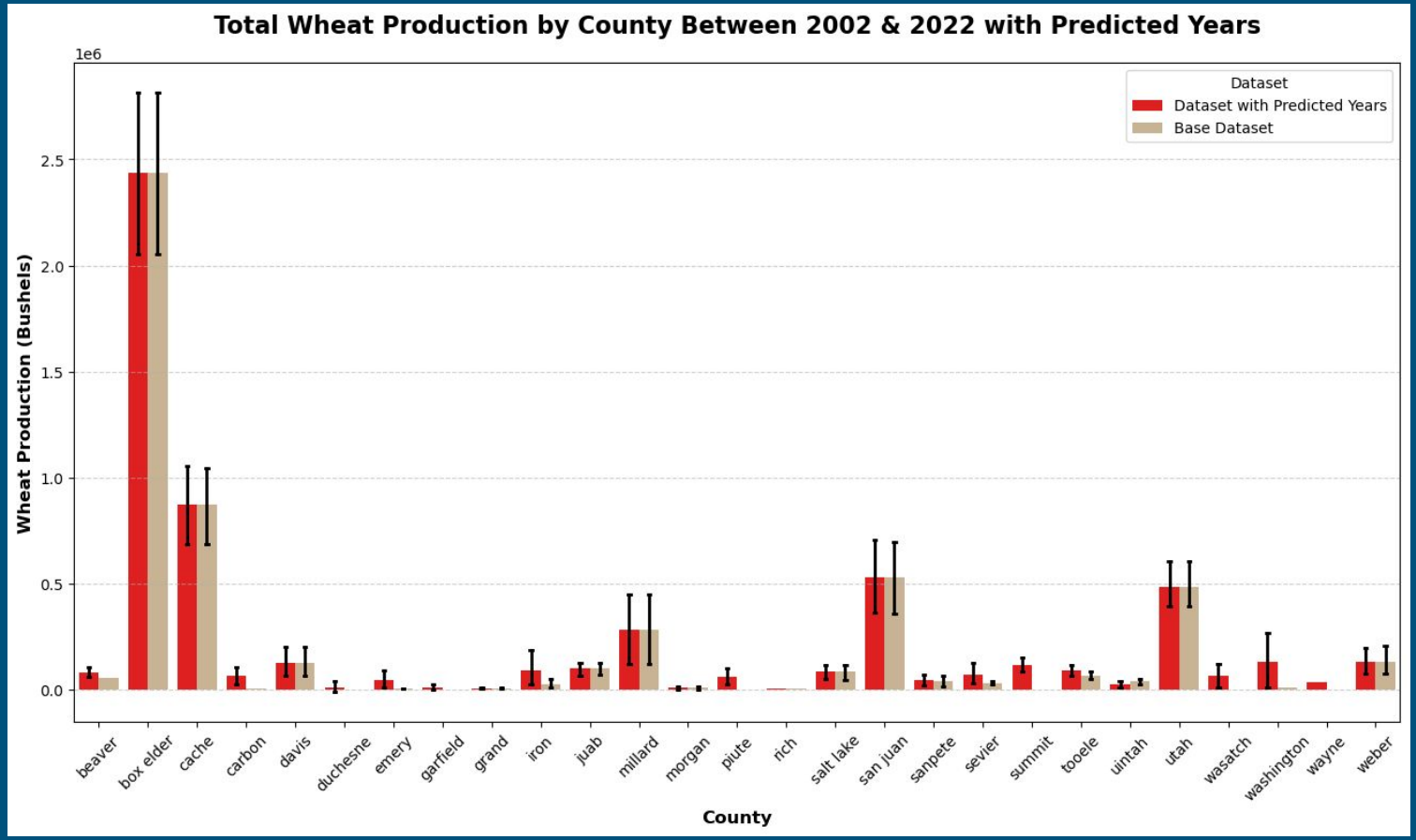


Fine Tuned Results - XGBoost

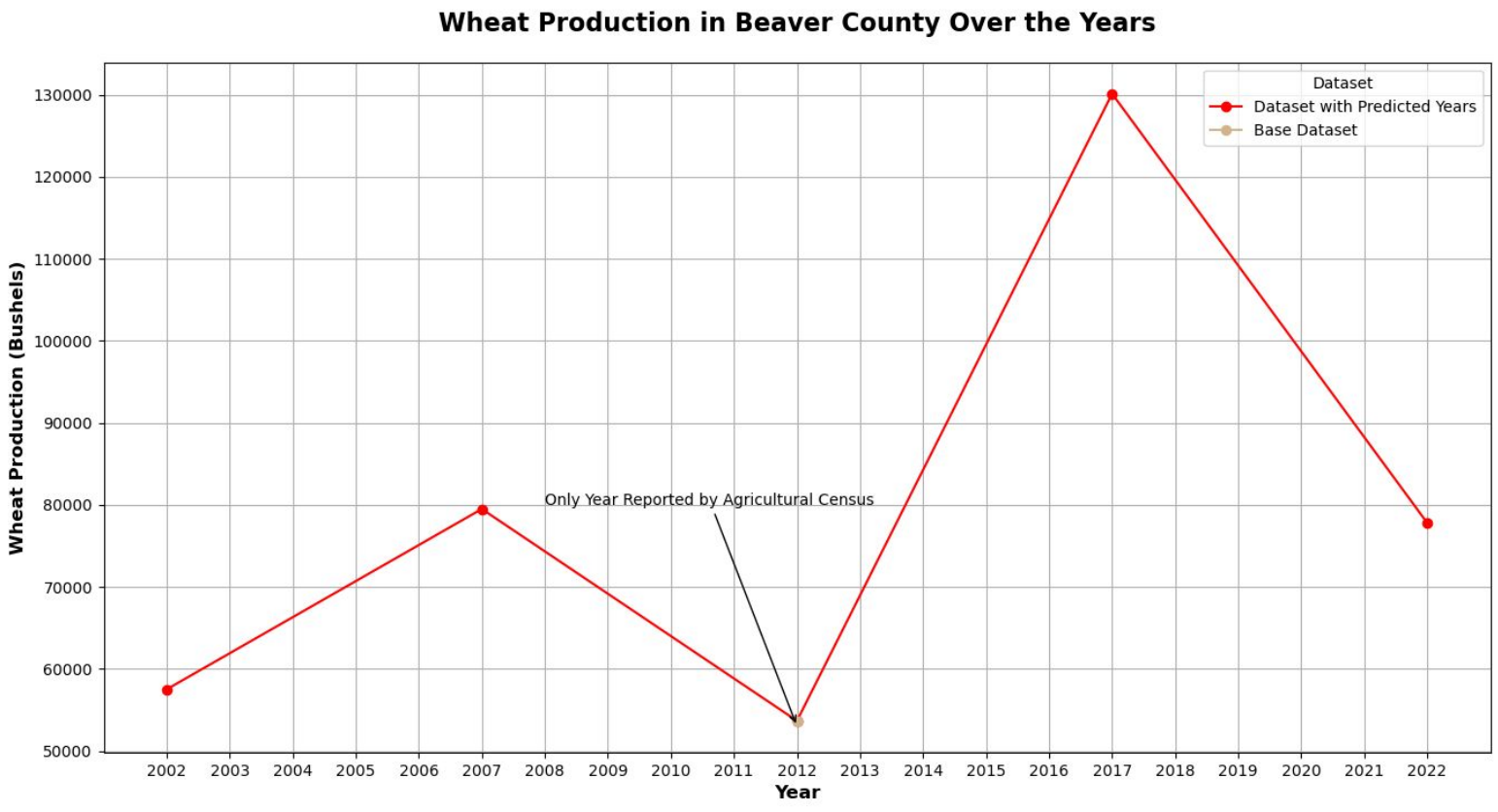
XGBoost was fine tuned on all parameters to minimize test set MAE and then cross validated to confirm fine tuned parameters and prevent overfitting.

Train MAE	Test MAE	10 Fold CV MAE
19.736	134574.296	312229.183

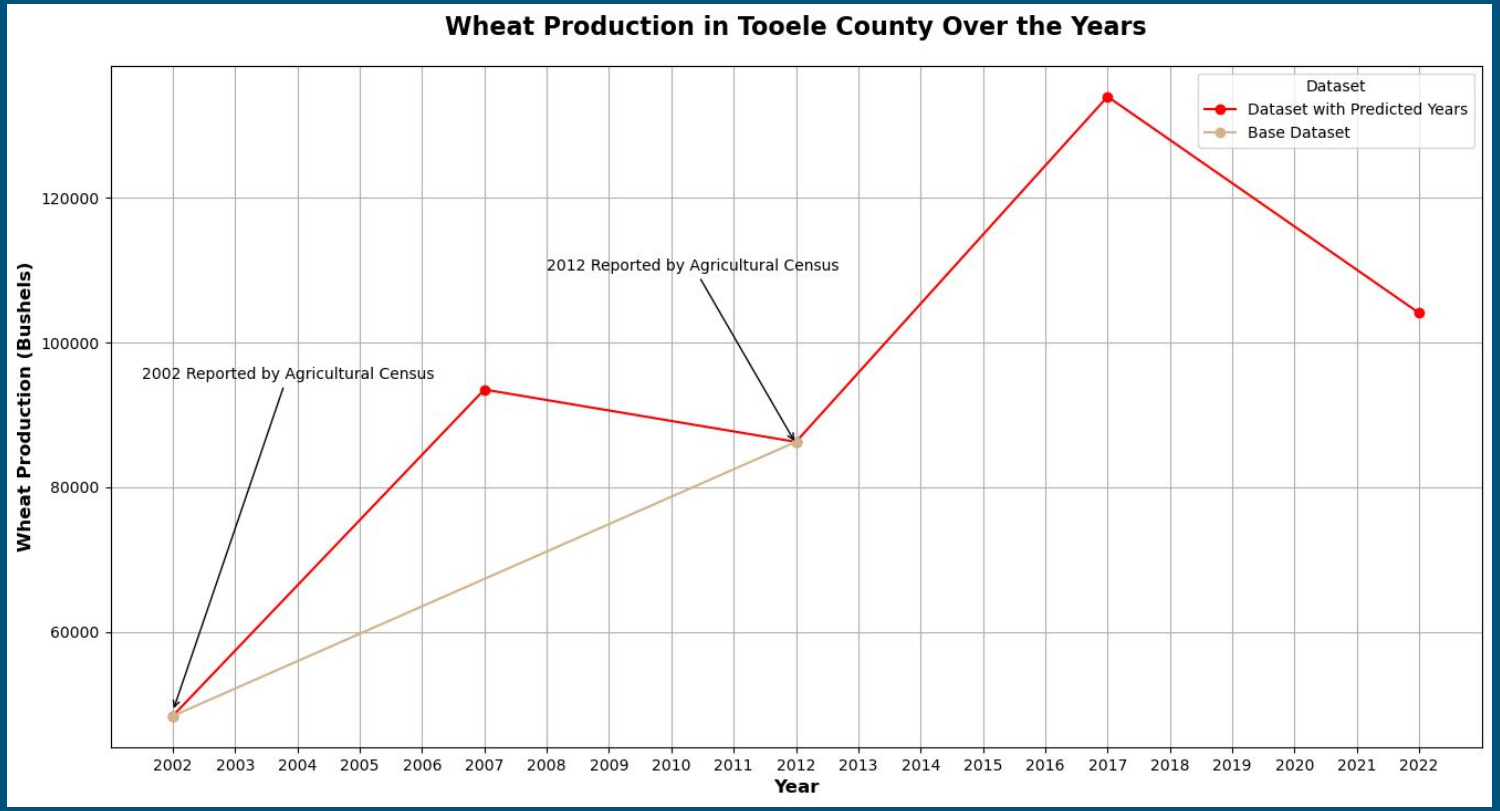
Data After Filled in Gap Years



Example: Beaver County



Example: Tooele County



Conclusion

- In general we saw that most counties experienced an increase in wheat production between 2002 & 2022
- Most counties missing the year 2017 would have seen a large spike in production, followed by a smaller reduction in 2022
- The top three contributing features in obtaining accurate predictions is based on how much the county's GDP relies on agriculture, the temperature, and precipitation level in preceding years