Regression Models for Predicting Commodity Sales Prices

Daniel Mortensen

17-Nov-21

Introduction

Agricultural demand is expected to increase by about 60% by 2050 as a result of population growth. Corn is one of the most important cereal grains in the United States. *Grain* corn is the type of corn we are familiar with in our day-to-day lives, as opposed to *silage* corn, which is primarily used for livestock feed. In 2020, US grain corn sales generated over $60 billion dollars in revenue. However, as can be seen in Figure 1, the sales price of grain corn can fluctuate significantly. Within the last 20 years alone, the price per bushel has varied from as low as $1.52 to as high as $7.63. The current price per bushel is sitting around $4. This instability creates significant investment risk for grain corn production. Here, a model is presented that can be used to calculate both annual and monthly average sales prices for grain corn, with the aim of reducing investment risk.

Data Collection

Corn data was downloaded from quickstats.nass.usda.gov, a data repository for US crops managed by the US Department of Agriculture. All data related to grain corn that was collected on an annual and monthly basis, except those with additional qualifiers such as "IRRIGATED" or "ORGANIC", were downloaded for the years 1950 to 2020. This data was then filtered for features that were recorded at least 90% of the time. During this process, the baseline was also defined as the average sales price in the previous three years. For the monthly data, a three-month lag was simulated to represent real world data-update lag times. A plot of the baseline data along with the true monthly sales prices are shown in as a function of time Figure 2. Note that the baseline predictions are essentially a smoothed out offset

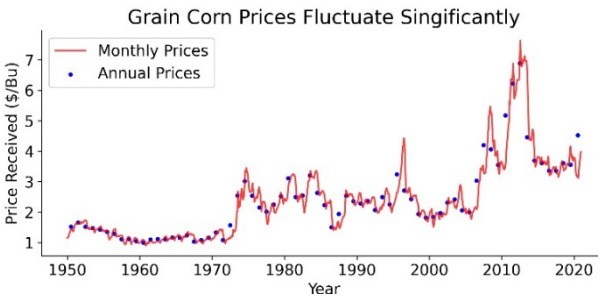

Figure 1. Average annual (blue dots) and monthly (red line) sales price per bushel of corn in USD between 1950 and 2020.

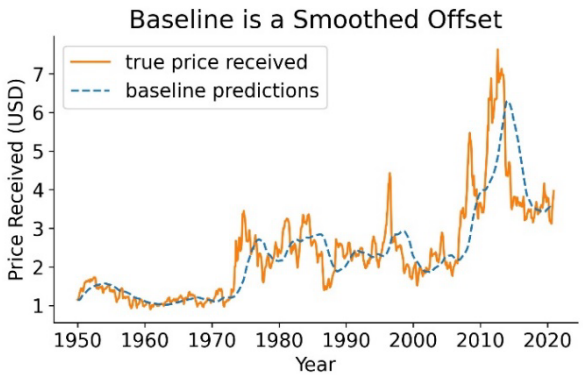

Figure 2. Average monthly sales price (orange solid line) and baseline predicted sales price (blue dashed line) per bushel of corn in USD.

version of the true values. Thus, the baseline does not respond well when there is a rapid change in price. The baseline predictions correlate with the true sales prices with R^2 scores of 0.537 and 0.685 for the annual and monthly data, respectively, meaning that less than 70% of the variation in the pricing of corn is captured by the baseline.

To model climate effects on sales prices, data for ten different climate metrics, such as average temperature and total rainfall, were downloaded for the years 1950 to 2020 from ncdc.noaa.gov, a data repository managed by the United States National Oceanic and Atmospheric Administration. I also added the annual US population (downloaded from multpl.com) to model market size, and the annual inflation rate of the US dollar (thebalance.com) and the total US GDP and percent change in GDP (fred.stlouisfed.org) to model inflation and market health, respectively. Data were analyzed using Python 3 in the Jupyter notebooks associated with this report. Parameters for the kernels required to run these notebooks are also included in the "kernel requirements" subfolder. Some data were not collected monthly and were, therefore, projected onto a monthly time frame using various methods of extrapolation, as discussed in the data scrubbing notebook.

Trends in the Data

Several interesting trends were observed in the data. First, the number of acres of corn harvested each year has not changed significantly in the last seventy years. As shown in Figure 3, there has been an average of about 70 million acres harvested each year, with a minimum of just over 50 million acres in the early 1980's and a maximum of just under 90 million acres around 2010. In contrast, the number of bushels of corn produced over the last seventy years (Figure 4, green line) has

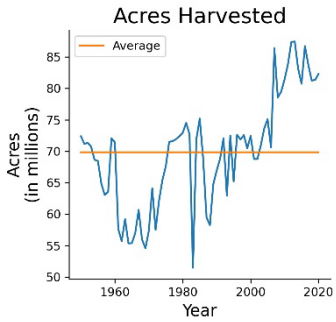

Figure 3. Millions of acres harvested per year. The orange line is a plot of the average acres harvested over all years, used to illustrate the relative consistency of the data.

increased from about 2 billion bushels per year to nearly 15 billion bushels per year.

It is shocking that so much more corn can be produced without harvesting more acres, but over this same period, farming techniques have improved to allow more corn to be harvested from the same amount of land. As can be seen from the blue trace in Figure 4, in 1950 farmers were only able to produce on average about 40 bushels of corn per acre. That amount has steadily increased with time, to where in 2020 farmers were able to produce on average about 170 bushels per acre, a more than 4-fold increase in per acre production.

While farming practices have made it possible to produce more corn per acre, they have also made the production of grain corn more cost effective. This has resulted in a steady decrease in the value of corn (Figure 5). When adjusted for

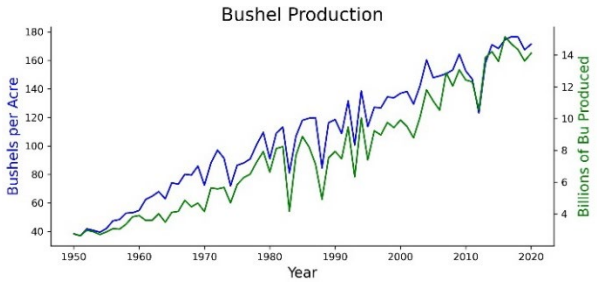

Figure 4. Bushels produced per acre (blue line, left axis) and billions of bushels produced (green line, right axis) over time.

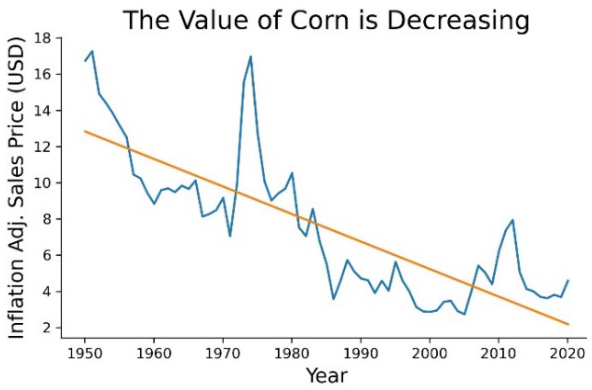

Figure 5. Inflation adjusted sales price per bushel of corn over time. The orange line is a best fit to the data, used to highlight the downward trend.

inflation, the price of corn has dropped substantially over the last seventy years. In 1950, a bushel of corn sold for an equivalent of $17 in today's dollars, whereas now it is selling for an equivalent of just over $4. There have been some interesting spikes in the value of corn, most notably in the early 1970's and around 2010. Interestingly, both periods correspond to significant market recessions. Therefore, the sales price of corn could possibly be a good metric for tracking market health.

The last trend that will be discussed is month-to-month sales price fluctuations. Time series analysis of the monthly sales data (Figure 6) shows that monthly prices are generally higher in summer months, going more than over 3% above the average annual price in May through July. This is likely because the warmer weather in these summer months makes it easier for the corn to grow, resulting in larger, healthier cobs. Prices are lowest in October and November when the temperature starts to drop and changes in humidity and precipitation patterns occur.

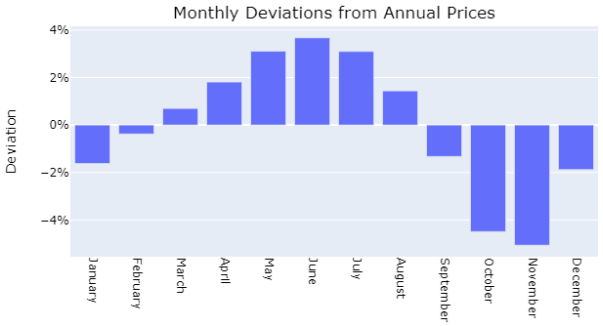

Figure 6. Average monthly deviations from annual sales prices.

Modeling

The average annual and monthly sales prices per bushel of corn (Figure 1) were modeled using eight different machine learning regression models, including linear, ridge, lasso, KNeighbors, a basic decision tree, random forest, support vector regression, and XGBoost regression. An F-test was used to determine if any statistically significant improvements were made by the model relative to the baseline. An F-test is appropriate here because I am comparing the distribution in errors resulting from the model predictions relative to the distribution of errors resulting from the baseline predictions, and the F-test is used to compare breadths of distributions (specifically, standard deviations).

Four models produced statistically better results than the baseline predictions, namely KNeighborsRegressor, DecisionTreeRegressor, RandomForestRegressor, and XGBRegressor. These four models were then fed into a grid search with 3-fold cross validation to optimize for several hyperparameters, resulting in nearly 1,200 total model variants being tested. The best predictor of the corn sales price was XGBRegressor for both the annual and monthly data. These models both used similar hyperparameters. Specifically, the booster method used for both models was "gbtree", the max tree depth was 5, and the minimum child weight was 2. The only difference between the models were the learning rates (0.5 and 0.4 for the annual and monthly models, respectively) and the training objective ("reg:squarederror" and "reg:tweedie" for annual and monthly, respectively).

An R^2 value of 0.950 was obtained for the annual data, meaning that all but about 5% of the variability in the annual sales price can be predicted. For the monthly sales price, an R^2 value of 0.986 was obtained, meaning that only about 1.4% of the variability in the data was unaccounted for. The model for the monthly data only used five features to produce the R^2=0.986. However, when a similar number of features were used to try to model the annual data, the R^2 value was always less than 0.8. Therefore, a total of sixteen features were used to model the annual data. More features were likely necessary for modeling the annual data because there are fewer data points in the annual data, making it more difficult to build as robust of a predictive model.

Various neural networks constructed using tensorflow were also used to try to model the corn sales prices. Decent results were obtained using these networks (best R^2 values obtained were 0.945 and 0.892 for the annual and monthly data, respectively). But the neural network models were not able to outperform the XGBoost models.

Conclusion

XGBRegressor, an XGBoost model, was used to predict the average annual and monthly sales price per bushel of grain corn. The models generated here can account for more than 95% of the uncertainty in these sales prices, resulting in a highly significant reduction in investment risk for this commodity.