Predicting the Market Value of Corn

Daniel Mortensen
25-Sep-21

Introduction

Corn is one of the most important cereals 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 in livestock feed). In 2020, grain corn sales generated over $60 billion dollars in revenue in the United States. However, as can be seen in Figure 1 below, the sales price of grain corn can fluctuate significantly overtime. 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 $3. 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.

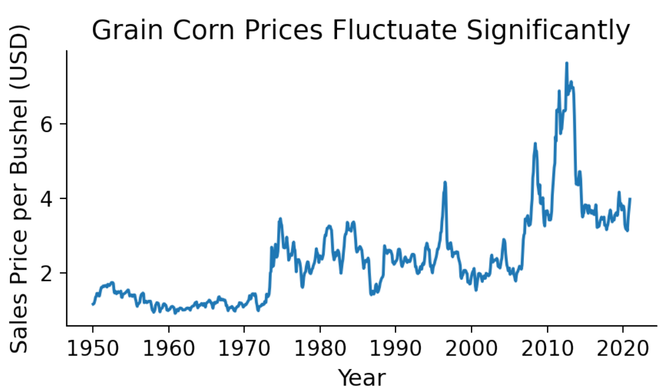

Figure 1. Monthly Sales price per bushel of corn in USD between the years 1950 and 2020.

Data Gathering

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 corn was downloaded between the years 1950 and 2020 in csv format. This data was formatted with multiple columns, including columns for reporting period, geographical level, a column describing the data collected in each row, and a column containing the values of the collected data. This data was then filtered so that only data related specifically to grain corn collected at the national level was retained (as opposed to state or county specific data). Once this was accomplished, a new feature, the price per bushel in the previous year, was engineered and added to this data set, as previous prices generally influence future production, etc.

Annual climate data for these same years was downloaded from ncdc.noaa.gov, a data repository provided by the United States National Oceanic and Atmospheric Administration, in csv format. Data for ten different climate metrics, such as average temperature and total rainfall, were collected. Metadata and other unnecessary data were removed from these data before combining them with the corn data. The yearly US population for this time was obtained from the website multpl.com. Data were analyzed using Python 3 in the Jupyter notebooks associated with this report. Parameters for the programming environments used are provided in the individual notebooks.

Modeling

Annual Prices

The annual sales price of grain corn does not have a direct relationship with time (see Figure 1). However, two other features were found that have more clear correlations with time, namely the total revenue generated per year and the total number of bushels produced. These features are plotted as a function of time in Figure 2. The total profit earned increases exponentially as a combined result of population growth and inflation, whereas the total number of bushels produced increases linearly with time simply because of population growth. The price per bushel can be calculated from these features by dividing the total revenue by the total number of bushels.

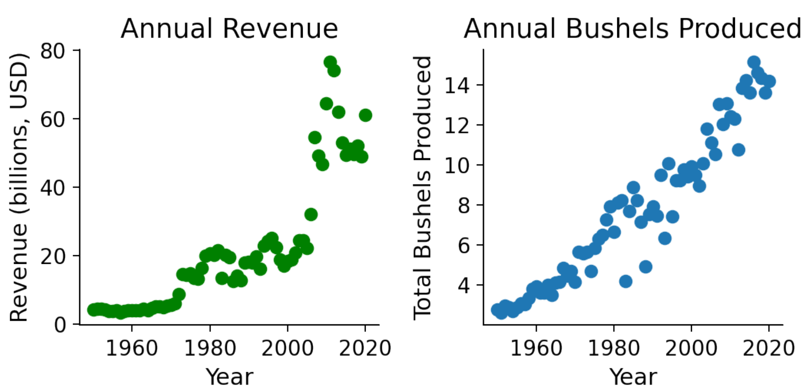

Figure 2. Annual revenues and total crop yields for the years 1950 through 2020.

For modeling purposes, the data was split into train and test sets. 80% of the data was used as training data and 20% was used as testing data. A number of variables, including climate conditions and the US population, were tested to see how well they predicted the features in Figure 2, and six highly correlated variables were discovered: (1) the year, (2) the population of the United States, (3) the average temperature across the country, (4) the minimum annual temperature, (5) the price per bushel in the previous year, and (6) the number of days in the year that were colder than the day preceding it. These six variables were then used, along with the training data, to create a linear regression model for predicting the annual revenues and bushel outputs for the test years. The model was able to accurately predict both features with an accuracy of more than 85%. A K nearest neighbors model was also applied to this data set but resulted in slightly lower accuracy scores.

The price per bushel was then calculated by dividing total revenue by total number of bushels produced. For the train data, the *actual* annual revenue and number of bushels produced were used, whereas for the test data, the *predicted* annual revenue and number of bushels were used. The predicted values for both the train and test data are shown against the actual sales prices in Figure 3 below. The model can predict the price per bushel with an accuracy of about 77%. Thus, any predicted value should to be accurate to within plus or minus 23%. While this may seem like a high error margin, the price per bushel has varied by a little over 500% in the last 20 years. Thus, this model greatly reduces investment risks compared to blind predictions.

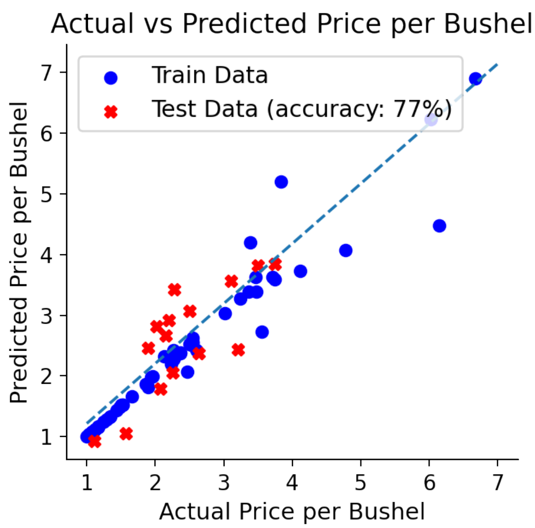

Figure 3. Predicted annual sales prices vs actual sales prices for the train (blue circles) and test (red exes) data. The dashed line is a linear fit to the training data.

Monthly Prices

While predicting annual sales prices is useful, it is clear from Figure 1 that prices vary significantly from month to month. Unfortunately, the total number of bushels produced and total revenue earned are not recorded by the USDA on a month-to-month basis, preventing us from modeling monthly prices the same way annual prices were modeled. Therefore, to project onto the months, the average deviation from the annual price per month was obtained using a time series analysis. The results of this analysis are shown in Figure 4. Monthly prices are generally higher in summer months, going up to a little over 3% above the annual prices 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. It is unclear from this data why that is, but it likely has to do with cooling temperatures and possibly changes in humidity and precipitation patterns.

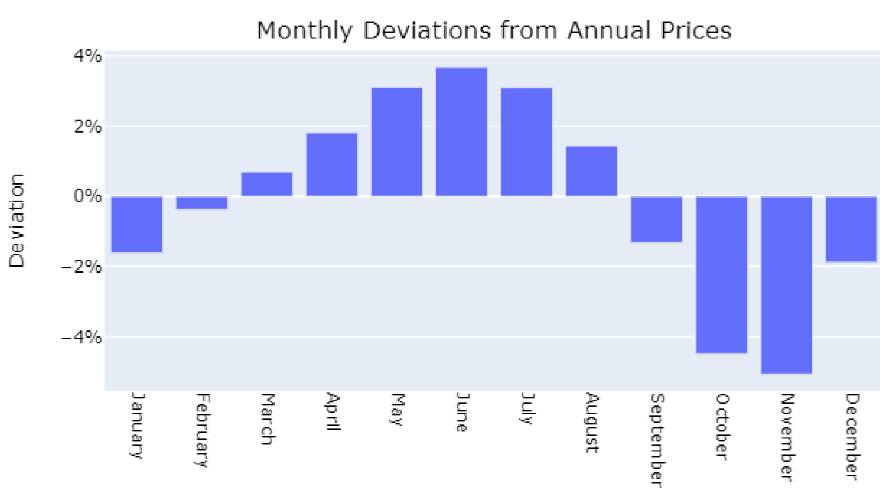

Figure 4. Average deviation from the annual sales price for each month.

Monthly prices were predicted by adjusting the annual prices by the amounts listed in Figure 4 for each month for both the train and test data. A comparison of the actual and predicted monthly prices is shown in Figure 5. There is good agreement between the actual and predicted values for both the train and test data, with the model resulting in an accuracy of 84% for the test data. This means that any monthly price prediction is likely to be accurate to within plus or minus 16% of the actual value. Again, while this may seem like a large prediction gap, the actual sales price can vary by as much as 500%. Thus, this model provides a more than 30-fold reduction in investment risk compared to blind guessing—a significant achievement.

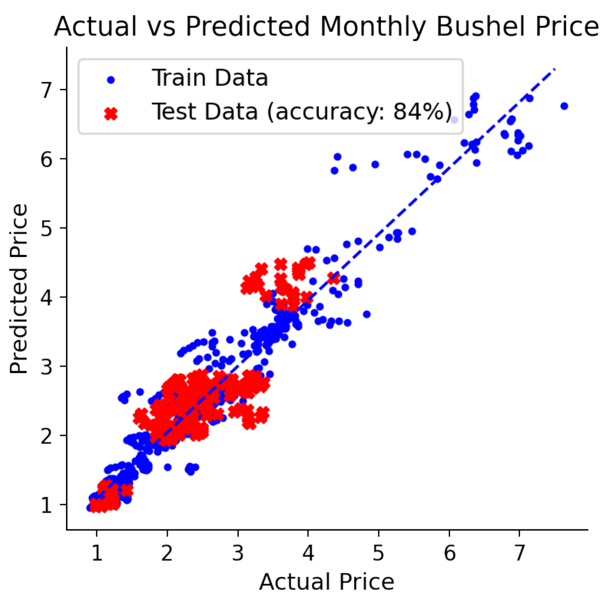

Figure 5. Predicted monthly sales price vs actual monthly sales price for the train (blue circles) and test (red exes) data. The dashed line is a linear fit to the train data.

Conclusion

Agricultural demand is expected to increase dramatically over the next few decades as world populations continue to rise rapidly. Thus, there will be significant growth in the agricultural sector in the decades to come. While the model presented here is simple in nature, it is still relatively powerful and can be used to dramatically reduce risk for investors in grain corn production. I would likely be able to make an even more accurate and precise model with access to higher resolution climate data, such as the proprietary high resolution PRISM data set, which has climate information for the country broken down into squares as small as 800-meter wide, as compared to the nation-wide data used here.