2023

Would AI Stocks Estimate Be as Surprised to USDA Stocks Reports As Private Market Analysts?

Asif Mahmud Chowdhury

WOULD AI STOCKS ESTIMATE BE AS SURPRISED TO USDA STOCKS REPORTS AS PRIVATE MARKET ANALYSTS?

BY

ASIF MAHMUD CHOWDHURY

A dissertation submitted in partial fulfillment of the requirements for the

Master of Science

Major in Economics

South Dakota State University

2023

THESIS ACCEPTANCE PAGE

Asif Mahmud Chowdhury

This thesis is approved as a creditable and independent investigation by a candidate for the master's degree and is acceptable for meeting the thesis requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Matthew Elliott

Advisor Date

Joseph Santos

Director Date

Nicole Lounsbery, PhD

Director, Graduate School Date

This thesis is dedicated to my family.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis advisor, Professor Matthew Elliott, for his invaluable guidance, encouragement, and support throughout my research. His expertise, patience, and insightful feedback have been instrumental in shaping the direction of my work and helping me to overcome various challenges along the way.

I would also like to thank the faculty and staff in the Ness School of Management & Economics] for their contributions to my education and the resources they have provided. Additionally, I am grateful to my classmates for their encouragement and camaraderie during this journey.

I am indebted to my family and friends for their unwavering support and encouragement, and for believing in me even during the toughest times.

Finally, I would like to express my gratitude to the participants who generously gave their time and shared their experiences and insights for this study. Without their cooperation, this research would not have been possible.

Thank you all for your contributions to my success. Any errors or omissions in this work are my own.

CONTENTS

ABBREVIATIONS VI
LIST OF FIGURES VII
LIST OF TABLES VIII
ABSTRACT IX
INTRODUCTION 1
BACKGROUND 2
CONCEPTUAL FRAMEWORK 11
LITERATURE REVIEW 3
METHODOLOGY 21
METHODS & PROCEDURE 25
EMPIRICAL ANALYSIS 30
SUMMARY & DISCUSSIONS 54
CONCLUSION 62
LITERATURE CITED 63

ABBREVIATIONS

EMH Efficient Market Hypothesis
MAE Mean Absolute Error
MSE Mean Squared Error
RMSE. Root Mean Squared Error
QAS Quarterly Agricultural Stocks

LIST OF FIGURES

FIGURE 1: SURPRISE EFFECT ON CORN, SOYBEAN & WHEAT MARKET ON 1-DAY & 7-DAY AFTER ANNOUNCEMENTS DAYS 2
FIGURE 2 CUMULATIVE LEVELS OF MARKET EFFICIENCY AND THE INFORMATION ASSOCIATED WITH EACH LEVEL (JONES, 1993:628) 13
FIGURE 3: IMPORTANT FEATURES PLOT FOR CORN & QUARTER 3 30
FIGURE 4: PARTIAL SUMMARY OF THE IMPORTANT FEATURES 31
FIGURE 5 PARTIAL OF CARRY_CHG_YOY FOR CORN & Q3 32
FIGURE 6: PARTIAL OF PROD_CHG_YOY FOR CORN & Q3 33
FIGURE 7: PARTIAL OF STOCK_CHG_YOY_LAG1 FOR CORN & Q3 33
FIGURE 8: IMPORTANT FEATURES PLOT FOR CORN & QUARTER 1,2 AND 4 34
FIGURE 9: PARTIAL SUMMARY FOR CORN & QUARTER 1,2 AND 4 35
FIGURE 10 PARTIAL OF PROD_CHG_YOY FOR CORN & Q1Q2Q4 36
FIGURE 11: PARTIAL OF STOCK_CHG_YOYLAG1 FOR CORN & Q1Q2Q4 37
FIGURE 12: PARTIAL OF STOCK_CHG_YOY_LAG2 FOR CORN & Q1Q2Q4 37
FIGURE 13: IMPORTANT FEATURES PLOT FOR SOYBEAN & QUARTER 3 38
FIGURE 14: PARTIAL SUMMARY FOR SOYBEAN & QUARTER 3 39
FIGURE 15 PARTIAL OF CARRY_CHG_YOY FOR SOYBEAN & Q3 40
FIGURE 16: PARTIAL OF PROD_CHG_YOY FOR SOYBEAN & Q3 41
FIGURE 17: PARTIAL OF STOCK_CHG_YOY_LAG1 FOR SOYBEAN & Q3 41
FIGURE 18: IMPORTANCE FOR SOYBEAN & Q1Q2Q4 42
FIGURE 19: PARTIAL SUMMARY FOR SOYBEAN & QUARTER 1,2 AND 4 43
FIGURE 20 PARTIAL OF STOCK_CHG_YOY_LAG1 FOR SOYBEAN & Q1Q2Q4 44
FIGURE 21: PARTIAL OF PROD_CHG_YOY FOR SOYBEAN & Q1Q2Q4 45
FIGURE 22: PARTIAL OF ETHANOL_CHG_YOY FOR SOYBEAN & Q1Q2Q4 45
FIGURE 23: IMPORTANT FEATURES PLOT FOR WHEAT & Q3 46
FIGURE 24: PARTIAL SUMMARY OF THE IMPORTANT FEATURES 47
FIGURE 25 PARTIAL OF STOCK_CHG_YOY_LAG1 FOR WHEAT & Q3 48
FIGURE 26: PARTIAL OF PROD_CHG_YOY FOR WHEAT & Q3 49
FIGURE 27: PARTIAL OF STOCK_CHG_YOY_LAG2 FOR WHEAT & Q3 49
FIGURE 28: IMPORTANT FEATURES PLOT FOR WHEAT & Q1Q2Q4 50
FIGURE 29: PARTIAL SUMMARY FOR WHEAT & QUARTER 1,2 AND 4 51
FIGURE 30 PARTIAL OF STOCK_CHG_YOY_LAG1 FOR WHEAT & Q1Q2Q4 52
FIGURE 31: PARTIAL OF STOCK_CHG_YOY_LAG2 FOR WHEAT & Q1Q2Q4 53
FIGURE 32: PARTIAL OF PROD_CHG_YOY FOR WHEAT & Q1Q2Q4 54
FIGURE 33: COMPARISON PLOT FOR CORN & Q3 54
FIGURE 34: COMPARISON PLOT FOR CORN & Q1Q2Q4 55
FIGURE 35: COMPARISON PLOT FOR SOYBEAN & Q3 56
FIGURE 36: COMPARISON PLOT FOR SOYBEAN & Q1Q2Q4 57
FIGURE 37: COMPARISON PLOT FOR WHEAT & Q3 58
FIGURE 38: COMPARISON PLOT FOR WHEAT & Q1Q2Q4 59

LIST OF TABLES

TABLE 1: VARIABLES AND THEIR SOURCES, TIMELINE 17
TABLE 2: INDEPENDENT VARIABLES (USA) 19
TABLE 3: INDEPENDENT VARIABLES (SOUTH DAKOTA) 20
TABLE 4: LIST OF FEATURES USE IN XGBOOST 27
TABLE 5: MODELS' PERFORMANCE LEVEL 60

ABSTRACT

WOULD AI STOCKS ESTIMATE BE AS SURPRISED TO USDA STOCKS REPORTS AS PRIVATE MARKET ANALYSTS?

ASIF MAHMUD CHOWDHURY

2023

The USDA survey-based Quarterly Agriculture Stocks (QAS) reports are the primary source of information regarding the relative supply of U.S. corn, soybeans, and wheat for the last fifty years. Research has examined USDA stock reports and their relevancy to the market (e.g., Isengildina-Massa et al., 2021). In addition, private industry analysts estimate expected quarterly grain stock reports before USDA releases them. Market information firms such as Bloomberg and Reuters publish a subset of these estimates a few days before the USDA reports. Previous research has found that when industry analysts have significant differences in stock expectations compared to what the USDA releases for grain stocks, market prices adjust rapidly to what the USDA found in their survey. Many media outlets and previous research attribute the differences in expectations and changes in market prices to a "market surprise" (e.g., Karali et al. (2020)).

Market analysts, USDA officials, and researchers have offered four reasons for market surprises in the grain stocks reports. First, USDA surveys may need to account for grain in transit when surveying stocks. Second, the market often uses weight (e.g., 60 lbs per bushel) to determine supply, while survey estimates ask how much volume (e.g., bushels) is on the farm or in commercial storage. When there is a deviation in the average weight of a commodity for a season, there could be

discrepancies between surveyed stocks and actual stocks by weight. Third, errors in estimating what portion of existing stocks are from old or new crop production may cause surprises in the final annual report before a change in the marketing year. For example, USDA asks in their survey how much old crop corn is on hand on September 1st, although some crops taken in by grain wholesalers can be new crops by this date. There can be discrepancies when the survey respondent must accurately segregate the new and old crop amounts. Fourth, USDA survey-based stock reports contain survey noise. Market analysts may need to account for survey noise in sequential estimates. This paper seeks to use AI methods and large datasets on grain movement to understand the primary reason market analysts are frequently surprised by USDA QAS reports. Given the recent surge in grain movement data, available grain quality data, and data on the output of significant demand sources of grain, particularly at a state level, it is possible to use advances in analyzing high dimensional data (e.g., random forest, gradient boosting) to develop an objective artificial intelligent (AI) market analyst. This paper aims to explore additional public data sources related to commodity demand and supply in the corn, wheat, and soybean markets and apply AI techniques to determine whether data analytics improves the prediction of QAS reports released by USDA for corn, soybeans, and wheat compared to market analysts estimates. Our primary research objective is to determine if AI can more accurately predict QAS estimates from USDA than the survey of Market analysts that Bloomberg and Reuters have historically provided. Our secondary objective is to decompose the surprise by the source of the surprise.

In this effort, we use the Extreme Gradient Boosting ML model to predict the stock estimate of the three major commodities (Corn, Soybean, and Wheat). We used grain stocks and production by state, carry-over stock from the previous year, weekly grain loaded on trains and barges, weekly ethanol production, monthly ethanol crushed, and weekly accumulated exports, market analysts' estimates from Bloomberg and Reuters from the year 2007 to the 4th quarter of 2022. We aggregated all these features every quarter to understand the estimate of stock. After accumulating all the features, we cross-checked the values with the national report of these particular years we found consistency among them. This means the features show actual values from each quarter to understand the accurate estimate of the stock. We also grouped each feature according to 10 Agricultural Regions. We found through our machine learning algorithm that production is the most important one to estimate the quarterly stock, with carry-over and accumulated exports in 2nd and 3rd most essential features of the model. We also found that ethanol production and grain exports have an inverse relation with the grain stock every quarter.

INTRODUCTION

Recently, there has been a sizable difference in grain stock estimates between the USDA and private market analysts. In the last five years, agricultural commodities prices have increased more on announcement days due to the surprise compared to the earlier five years of the last ten years. The prices in the earlier five years were decreasing on the announcement days of the Quarterly Grain Stocks reports. For example, in the third quarter of 2022, the mean estimate for ending stocks for corn in a pre-report Bloomberg survey was 1.4955 billion Bushels, while USDA reported 1.377 billion Bushels. Market report surprises can cause unexpected market volatility and added risk. Thus, market participants started using novel quantitative methods to anticipate market surprises better that can complement market analyst surveys.

The accuracy and effectiveness of USDA price and production projections have been the subject of numerous studies in agricultural forecasting (e.g., Irwin, Gerlow, and Liu, 1994; Bailey & Brorsen, 1998; Sanders and Manfredo, 2002, 2003; Isengildina, Irwin, and Good 2004, 2006; πIrwin, Sanders, and Good 2014). In contrast, less emphasis has been paid to determining the efficacy of private forecasts, changing from survey-based approaches to more advanced data analytics with new technologies such as remote sensing, machine learning (ML), and big data. For example, Reuters has also been predicting QAS estimates using proprietary ML models and marketing them as "smart estimates" in addition to their survey of Market Analysts.

Thus, this paper aims to explore additional public data sources related to commodity demand and supply in the corn, wheat, and soybean markets and apply machine learning techniques to determine whether data analytics improves the prediction of QAS reports

released by USDA for corn, soybeans, and wheat. Our primary research objective is to determine if ML can more accurately predict QAS estimates from USDA than the survey of Market analysts that Bloomberg and Reuters have historically provided. Our secondary objective is to attempt to decompose the surprise into by source of surprise. That is, what is the primary reason that market analysts are frequently surprised by USDA QAS reports?

BACKGROUND

The USDA's quarterly stocks report (QAS) is published four times a year. Data are collected to estimate on and off-farm stocks for each state in the U.S. The USDA uses a stratified sample and survey method to collect on-farm stocks. To collect off-farm stocks, the USDA attempts to contact and survey most buyers and wholesalers of commodities. The off-farm stocks survey counts the amount of grain in all known commercial grain storage facilities. In contrast, the on-farm survey is a probability survey of farm operators. USDA also publishes supply and demand estimates (WASDE) reports monthly. These

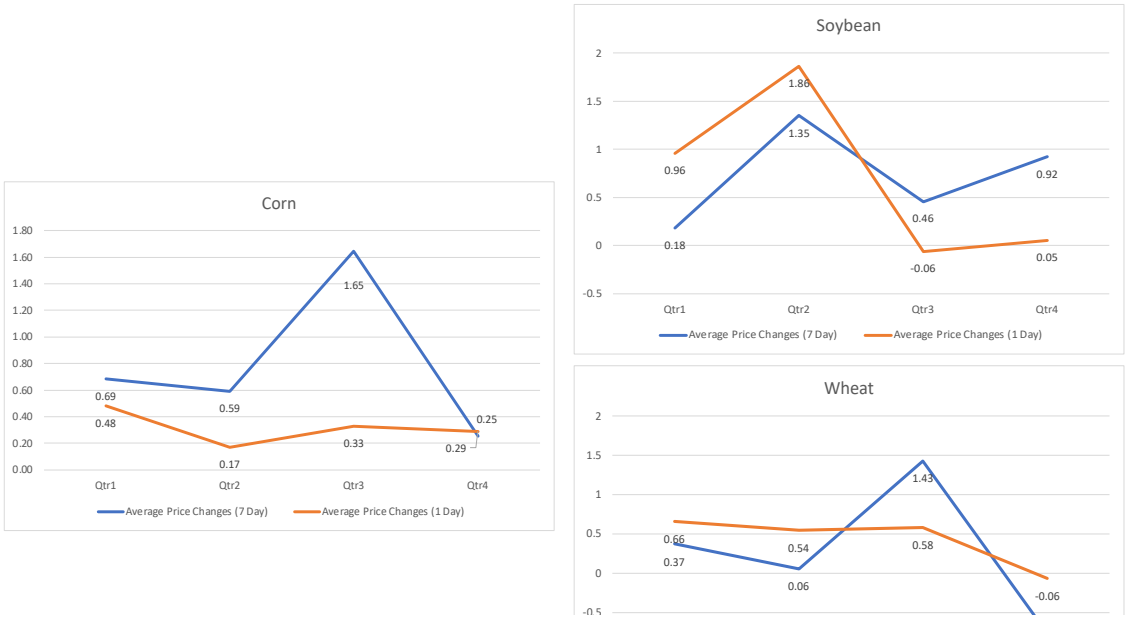


Figure 1: Surprise Effect on Corn, Soybean & Wheat Market on 1-day & 7-day after Announcements Days

Figure 1 shows how the unit price of corn, soybean, and wheat fluctuated after the USDA announcement days. The price fluctuation is also noteworthy after a week of the QAS report from the USDA. These fluctuations are happening because the market has been prepared by the private market analysts' report of next quarter's grains stocks which happened to be different from what USDA has been publishing. In quarter 3, for the corn market, the price changes on the announcement days, on average, by $1.65 for each bushel. In contrast, the soybean market faces an average price raise of $1.86 for quarter 2.

In addition to USDA reports, private industry analysts forecast quarterly grain stock data. Market information firms such as Bloomberg and Reuters publish a survey of private forecasts a few days before the USDA announces. Private industry analysts' estimations have been utilized to proxy market expectations of government data by market participants (e.g., Colling and Irwin, 1990; Grunewald, McNulty, and Biere, 1993; Garcia et al., 1997; Egelkraut et al., 2003).

Private analysts create QAS estimates to aid their clients with agricultural business opportunities. As crop harvests generate production projections, later forecasts primarily reflect demand uncertainty. In both circumstances, providing consistent, accurate forecasts is crucial because forecasts describe present and predicted fundamental supply and demand conditions and outline the risks that market participants and policymakers confront.

LITERATURE REVIEW

The agricultural commodities market is integral to all the farming activities that contribute to the macroeconomic level. The farmers, retailers, producers, suppliers, buyers,

and even the distributors of the food channel are directly or indirectly dependent on these markets. The agricultural commodities market is a part that connects all the agents of the microeconomic and macroeconomic levels to complete their supply chain. Researchers have worked with these markets since the early twentieth century to understand their behavior and impact on various levels. Besides the USDA, several private analysts provide their clients with estimates of following quarter-ending stocks to aid them in making the economic decision. These reports are essential to the agricultural commodities market and all associated agents.

The academic literature has examined commodity futures markets to provide statistical evidence for forecasting crop-ending stocks and elaborating on the sources of market surprises in the agricultural commodities market. Informational impact, the very least as far back as Stigler's fundamental work, the significance of knowledge for the efficient operation of markets has been a primary focus of economic theory (1961). Jensen (2007) analyzes how mobile phone use has changed fishermen's livelihood in Kerala, India. He concludes that the spread of mobile phones facilitated better arbitrage opportunities and decreased waste and price dispersion across geographic markets by giving fishermen and dealers more access to market information, which gave us tremendous insights into how information flow or the information can do to any form of market. Contract theory considerations generally indicate that knowledge asymmetry can have redistributive and efficiency effects. To add insult to injury, when farmers do not know what the market will bear, they might make a less-than-ideal decision about what to produce (and sell) because the relative prices differ from the marginal transformation rate. To the degree farmers sell directly on the market, Jensen (2007) demonstrates how asymmetric information can lead

to significant price variations between regions. Likewise, the usefulness and influence of public information in agricultural markets have been investigated extensively by empirical studies. Many of these studies have primarily relied on USDA reports, such as those on corn and soybean production (Fortenbery & Summer, 1993), harvest predictions (Garcia et al., 1997), WASDE reports on corn and soybean prices (Isengildina-Massa et al., 2008), as well as cold storage reports on cattle and hog prices (Colling and Irwin. 1990). (Isengildina, Irwin, and Good, 2006). After a study indicated that USDA reports have significant market repercussions, suggesting that the public information published by the USDA creates economic welfare benefits (Falk & Orazem, 1985). Public agriculture information initiatives have been questioned despite this research despite its conclusions. Increased private sector access to agricultural market information and analysis (Egelkraut et al., 2003; Good and Irwin, 2006. McKenzie, 2008) questions the utility of publicly available data (Just, 1983; Salin et al., 1998).

Federal budget constraints in recent years have also highlighted concerns regarding the value of public information projects in agriculture. Finally, research has shown that publishing USDA reports have resulted in adverse market reactions (Fortenbery & Summer, 1993; Marone, 2008). There needs to be more scholarly research into the value of Crop Production reports, although market analysts extensively follow them. All event studies contain C.P. reports except for Karali (2012) 's research on USDA reports and conditional return variances and covariances in relevant agricultural futures markets. The application of crop-condition information, for example, in predicting agrarian production, has only been formally studied in a few academic papers (Dixon et al., 1994; Kruse & Smith, 1994; Fackler & Norwood, 1999; Irwin, Good, and Tannura, 2009). Because of this

lack of research, a comprehensive scholarly analysis of the market reaction to the USDA's Crop Production reports needs to be conducted. Because market prices reflect what is already known through private sector analysis and weather data, the Crop Production reports may not impact the markets. If this holds, then the data in the Crop Production reports are redundant.

Furthermore, Fernández-Perez et al. 2018 contribute to the literature in numerous ways. First, they decompose the bid-ask spread for agricultural commodity futures using a spread decomposition technique and monitor changes in bid-ask spread components in reaction to USDA announcements during the trading period. Their research has emerged a new understanding of agricultural items' bid-ask spread (BAS). Secondly, they examine why USDA pronouncements have become more information asymmetric using variables like news surprise and analyst dispersion. According to both assessments of the informational environment, informational traders can be seen around USDA announcement periods. To better understand why USDA declarations have become more information asymmetric (McNew & Espinosa, 1994). There will be more activity from traders accessing private information when there is much speculation regarding the nature of the release. McNichols and Trueman (1994) and Riordan et al. (1995) have found similar results (2013). People are more likely to acquire private information when there is a great degree of uncertainty about the nature of the news before its distribution.

USDA Grain ReportsThe accuracy and dependability of USDA crop production projections are critical, given the importance of these forecasts and their broad impact on the agricultural economy. According to Egelkrautet al. (2003), several studies have assessed the accuracy of USDA agricultural output estimates and their market implications

(e.g., Sumner and Mueller 1989). One aspect of production forecasts that is often forgotten is the method utilized to update forecasts as they progress through a forecasting cycle. There usually are five estimates of yearly corn and soybean production for a particular marketing year from the USDA National Agricultural Statistics Service (NASS), which begin in August and end in January of the marketing year.

To summarize the pricing implications of the U.S. corn balance sheet, Good & Irwin (2014) suggested that the size of marketing year closing stocks may be the essential factor to consider. One price component is limited in explaining price movements in a market where the price is driven by several supply and demand factors. However, the corn market responds to USDA forecasts of year-end marketing inventories in monthly WASDE reports. Because of the significance of these forecasts, it is worthwhile to check their correctness. A great deal of debate has also surrounded the May 2014 WASDE report's projections for the amount of maize stockpiled, with many experts claiming that the old crop estimate was too low. Others claimed that the new crop estimate was too high.

On the contrary, Xiao et al. 2017 contributed significantly to the body of knowledge already available. The USDA's different reports have been cited in several literary works as a source of price volatility and shocks in the agricultural commodities market. USDA's price and production projections have been scrutinized in various publications for their precision and efficiency. For example, Irwin et al. 1994; Bailey and Brorsen 1998; Sanders and Manfredo 2002, 2003; Isengildina, Irwin, and Good 2004, 2006; Irwin, Sanders, and Good 2014; Sanders et al. In addition, Karali et al. (2019) found that for corn and soybeans, the most significant market surprises are associated with Grain Stocks surprises, particularly with September surprises that are several orders of magnitude greater than the

surprises for any other report for these commodities. They also argued that It is noteworthy that the QAS reports are the least studied USDA reports in the literature. The magnitude of surprises for the Grain Stocks report is typically more extensive than those for the other reports.

Also, there is some indication of increasing surprises in corn grain stocks and winter wheat crop production. Specifically, the June Crop Production reports revealed a substantial 0.87 percentage point gain for winter wheat. The June and September Grain Stocks report revealed huge increases of 1.90 and 6.83 percentage points for maize, respectively. (Karali et al., 2019)In addition, a tiny amount of attention was devoted to the Grain ending stocks reports. We know only two studies that evaluate whether data analysis and techniques can better estimate stock forecasts: Botto et al. (2006) and Isengildina-Massa, Karali, and Irwin (2013). Botto et al. analyzed the WASDE balance sheets for U.S. corn and soybean from 1980/1981 to 2003/2004 to calculate the percent forecast error using a two-equation model (2006). The first equation relates the forecast horizon, marketing year, and interaction term to the percentage errors for a specific category and crop. The estimated squared residuals from the first equation are employed as a stand-in for the error variance. The estimated squared residuals' natural log is regressed on the same explanatory factors in the second equation. As a result, the first equation in the framework assesses forecast bias, whereas the second equation assesses trends in the variability of errors over the forecast horizon and throughout the sample period. OLS is used to estimate the equation's parameters, but a panel White estimator that considers period heteroskedasticity and autocorrelation are used to correct the estimates' standard errors. They found that soybean ending stocks were biased toward overestimation during the last years of the study

period, particularly early in the forecast cycle. They also revealed that soybean average price forecasts were biased toward underestimation during the last years of the study period; only during early reports were soybean production and export percentage errors significantly related to forecasting errors in average price. USDA performed reasonably well in generating supply and demand estimates for U.S. corn and soybeans. However, soybean ending stocks forecasts errors have significantly increased in absolute size during recent years. A tendency to overestimate soybean ending stocks was observed during their study timeline. Isengildina-Massa, Karali, and Irwin studied all U.S. corn, soybeans, and wheat categories published within the World Agricultural Supply and Demand Estimates (WASDE) report over 1987/88 through 2009/10 marketing years (2013). To analyze this dataset, they used absolute percent forecast errors as dependent variables in regression analysis to measure the impact of considered factors on the size of the error, and percent errors are used to investigate sources of bias or direction in forecast errors. They concluded that corn, soybean, and wheat forecast errors also grew during economic growth and with changes in exchange rates. At the same time, inflation and oil prices had a much smaller impact. The impact of economic growth was overstated in corn and wheat forecasts and soybean forecasts.

The USDA's forecast of the year's ending stocks is a yearly event. Several studies have evaluated such constants as inflation, national GDP, real GDP, and the whole list. This vacuum in literature was a pressing issue that needed to be addressed. However, it covered the forecast efficiency but did not compare it with the private analysts' performance. Private Analysts & USDAResearchers believe significant discrepancies between consensus analyst projections and government agency forecasts frequently lead to

market volatility. Several studies have indicated a decrease in the informational usefulness of USDA agricultural output estimates (Garcia et al. (1997)), consistent with the rise in the availability of private forecasts. Furthermore, Fortenberry and Sumner (1993) found that the market no longer responds to USDA production projections, although McKenzie (2008) concluded that the market does respond to USDA production forecasts. Several studies on price and production forecasts have either assumed implicitly that private analysts directly forecast the target outcome (to compare their predictions to those of government agencies) (e.g., Garcia, Egelkraut, and colleagues (1997)) or assumed that private forecasts are objective estimates of government forecasts (McKenzie, 2008). Studies like this cannot be used to evaluate the accuracy of end-of-season stock forecasts because the connections between private and government estimates and actual ending stocks have yet to be thoroughly examined. On the contrary, other than (October and November) corn, early season (August, September) and November soybeans, June winter wheat, and July spring wheat, the influence of the remaining Crop Production reports remained largely steady. These indicate that competition from private information sources did not diminish the usefulness of USDA data in crop markets. For decades, USDA reports have significantly impacted market prices, according to earlier assessments evaluating the utility of USDA information. According to Garcia et al. (1997), the USDA's August corn output projection is their study's most valuable data source. Private market analysts' predicting abilities have vastly outperformed the USDA's throughout the study period. As a result of advances in computers, communications equipment, remote-sensing satellites, etc., the cost of information has decreased significantly. However, Karali et al. argued that as the USDA grain stock reports depend on survey estimates of on-and off-farm storage,

which are not observable through precision agriculture or remote sensing technology, Grain Stock reports likely face the slightest degree of competition from private information sources. The on-farm survey is a probability poll of farm operators, whereas the off-farm stocks survey enumerates the grain quantity in all commercial grain storage facilities. (2019) At the same time, predictive forecasts are becoming more important to farmers and dealers as new technologies arise. However, more evidence is needed to show that private analysts can generate more meaningful insights than the USDA using this current technical finding.

CONCEPTUAL FRAMEWORK

The theoretical notion that markets would be surprised and change prices as a result of public reports were developed by Eugene Fama. Fama (1970) reviewed theoretical and empirical research in which he explained the market's characteristics and how prices changed over time based on various factors, including information. Farma defined the Strong EMH as a market that represents all the information available in that market. With a strong EMH assumption, the price reflects all the information obtained from the reports, word of mouth, or other resources. The prices on the market allow the information to one another and respond according to it. He also noted that the market's pricing is subject to real-time information, which enables buyers and suppliers to make their necessary economic judgments. He said this in a separate statement. This theory is known as the Efficient Market Hypothesis. The efficient markets hypothesis (EMH) asserts that markets are efficient and that there is no room for investors to generate extra gains because everything is priced reasonably and precisely. This suggests that there is little chance of outperforming the market.

However, some markets are less efficient than others. A market is only efficient if the prices adequately reflect their actual value. Market inefficiency can occur for various reasons, including

- Asymmetries in information,
- A lack of buyers and sellers (i.e., poor liquidity),
- High transaction costs or delays,
- Market psychology,
- Human emotion.

Market inefficiency results in deadweight losses. Most markets exhibit some inefficiency; in extreme cases, an inefficient market can illustrate market failure.

Any price reaction on the announcement dates of the USDA's report is a prominent example of how the information is not truly reflected on the price of the significant commodities. Indeed, several event studies in the agricultural commodities market have concluded that government agents like USDA announcements affect current and future prices (Milonas, 1987). Suppose the market is truly an efficient one. In that case, we should avoid the surprises of these commodities in this market as these surprises arise mainly from the informational aspect of the market.

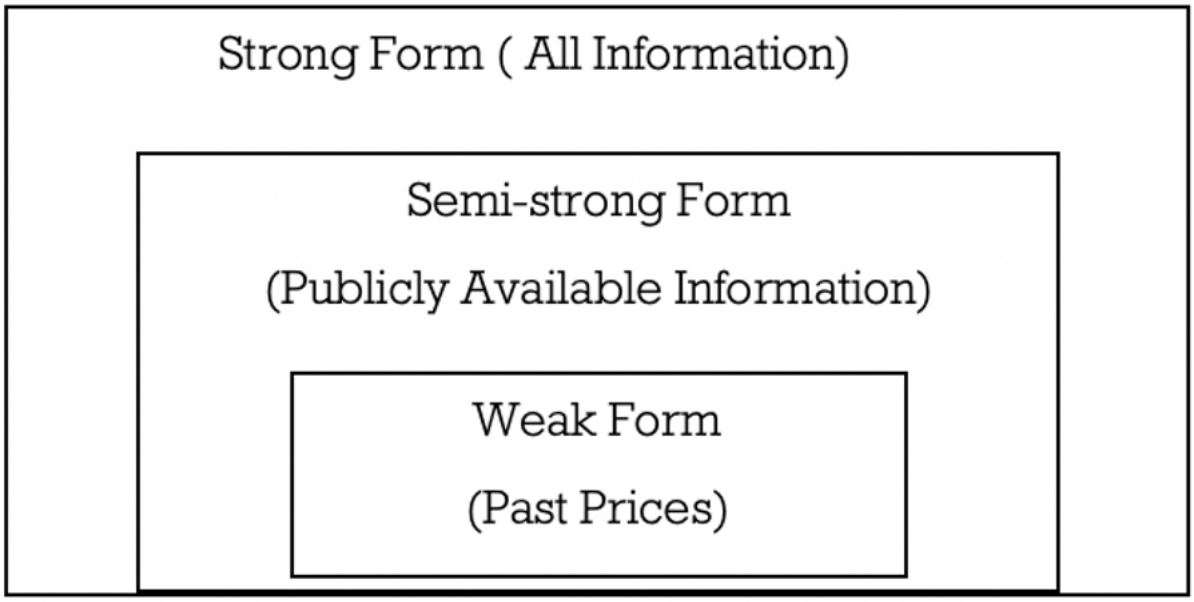


Figure 2 Cumulative levels of market efficiency and the information associated with each level (Jones, 1993:628)

Accepting the EMH in its purest (strong) form may be difficult as it states that all information in a market, whether public or private, is accounted for in a commodity's price. A strong EMH suggests that all publicly available information is incorporated into the crop's current price. No technical or fundamental analysis could reliably capture gain in a market. The strongest form of market efficiencies states that no one can gain or capture profit out of the market by using technical or fundamental analysis. Technical analysis means to study the historical attributes of the targeted stock or commodities, and fundamental analysis is the study of valuations, reports, etc., from the stocks or commodities' end. It is not easy to see any market that portrays all the characteristics of strong market efficiencies. Hence, the modifications of EMH reflect the degree to which it can be applied to markets: the Semi-strong Efficient Market Hypothesis & Weakly Efficient Market Hypothesis. Figure 2 shows the form of market efficiencies with information associated with each level. Neither technical analysis, based on past patterns

of return, nor fundamental analysis, which considers present information, can assist in forecasting future price movements when a market is a semi-strong form efficient. Non-public information can be leveraged to provide returns that are above average. Examining how prices and volumes react to particular occurrences is a common method for testing semi-strong forms of efficiency. Markets are semi-strong form efficient if new information is swiftly reflected in pricing. Such events include special dividends, stock splits, legal proceedings, mergers and acquisitions, tax adjustments, etc. Evidence suggests that developed markets might be semi-strong efficient while developing markets are not. Semi-strong of EMH suggests that all public information (but not non-public information) is factored into the current share price of a stock. The fundamental analysis could yield higher returns, but the technical analysis will not work in this market.

According to the weak form of market efficiency, no rule formed from the analysis of historical trends can be employed to generate excess return because past market dates are fully reflected in current market prices. The weakest version of the efficient market hypothesis is weak-form market efficiency (EMH). The weak form of market efficiency suggests that future price trends cannot be predicted via technical analysis. Weak efficiency, this type of EMH asserts that all primary commodities prices are represented in the current commodities price. Therefore, technical analysis can be utilized for market forecasting and advantage trading. It also suggests that buyers can use fundamental analysis to find undervalued and overvalued assets because the market lacks information or computation abilities. With research on the previous data and information, one can get the desired value in the weak form of the market. Using past market movements to forecast future price movements is known as technical analysis. However, fundamental analysis

and insider knowledge can be applied to generate an excess return in a weak market efficiency.

The efficient market hypothesis (EMH) is significant because it suggests that free markets can optimally allocate and distribute products, services, capital, or labor (depending on the type of market) without central planning, supervision, or government control. According to the EMH, prices reflect all available information and reflect a state of equilibrium between supply (sellers/producers) and demand (buyers/consumers). A significant implication is that "beating the market" is impossible, as there are no anomalous profit opportunities in an efficient market. With the anticipated advent of machine learning technology, we will lessen the market surprises in the agricultural commodity market. The market will absorb all the information it will receive in the price without creating surprises due to the significant disparities in the reports from the private analysts with the USDA.

According to this notion, we expect the agricultural commodities market to be efficient, where none can gain more with technical or fundamental analysis. However, we observe signs of inefficiencies when there are persistent market surprises to public information that would be theoretically incorporated into prices if a strong EMH was assumed.

Suppose we could not predict the market movement by considering all the publicly available information. Then the market is in strong form according to the efficient market hypothesis, as all the available information is already reflected in the price. However, if we could predict the expected stock estimates and the price change expectations after the information is released. In that case, we have to conclude that maybe the market is in a weak form of efficiency because people are now reacting to the information changing price

in the direction we expect them to. If we can take all the publicly available information out there and predict what the market reports will be reasonably accurate.

Furthermore, we have the expectation of what price direction should occur because we think we can predict this in a better way. Then we can determine that this information is vital to the market and the prices. We can predict better with all the publicly available information. So, the theoretical framework presented above can support our hypothesis that the agricultural commodities market is a weak form of efficiency.

DATA AND COVARIATES

The public data we include to estimate the quarterly stocks of corn, soybean, and wheat is the weekly amount of grain loaded on the rail by state, on the weekly amount of corn, soybeans, and wheat that has passed through locks on the Ohio, Mississippi, and Arkansas river on barges. The weekly amount of corn used for ethanol production, The annual amount of corn, soybeans, and wheat produced in each state, and the monthly amount of soybean seeds crushed by the National Oilseeds Processor Association. We look for variations in the probability of yearly changes of these variables in all the states and then aggregate this to the national level.

Table 1: Variables and their sources, timeline

Data Source	Data	Variable Name	Timeline
USDA-NASS	Grain Stocks and Production by State	commodity, state_name, bushels_stocks, bushes_production, carry_over	Aggregated Quarterly
US Surface Transportation Board	Weekly Grain Loaded on Trains by State	rail_bushels	Aggregated Quarterly
US Army Corp of Engineers	Weekly Grain Barges that moved through locks on the Mississippi, Arkansas & Ohio River	barge_bushels	Aggregated Quarterly
Energy Information Adminstration	Weekly Ethanol Production in PADD2 (Midwest)	ethanol_crush	Aggregated Quarterly
National Oilseed Processors Assoc	Monthly Oilseed Crush	oilseed_crush	Aggregated Quarterly
USDA's Foreign Agriculutaral Service	Weekly Grain Exports	weekly exports	Aggregated Quarterly
Reuters	Market Analysts Survey	Average, Actual, Market_analyst_Reuters, surprise	Aggregated Quarterly
Bloomberg	Market Analysts Survey	Average, Actual, Market_analyst_Bloomberg, surprise	Aggregated Quarterly

To that, we added data from weekly grain loaded on trains by state by the U.S. Surface Transportation Board, weekly grain barges that moved through locks on the Mississippi, Arkansas & Ohio river from U.S. Army Crop of Engineers, Weekly Ethanol Production in PADD2 (Midwest) from Energy Information Association, monthly oilseed crash by National Oilseed Processor Association, weekly grain exports by USDA's Foreign Agricultural Service.

In addition, we compiled the data from two sets of quarterly stock information provided by private experts. The first one comes from Bloomberg, and the second one comes from the database maintained by Reuters. While gathering the information from the private analysts, we examined Reuters and Bloomberg for duplicates. We found 205 unique analysts covering corn, 191 covering soybeans, and 184 covering wheat. In addition, as mentioned above, each analyst on the list has been providing their reports of quarterly stockpiles of important crops before the publication of the quarterly reports produced by the USDA. In order to do this study, we combined the quarterly stock data with the historical price data. In the table 2 we can see the our important features for our model aggregated in the national level that is USA. As we don't remove the NA values from our historic dataset it's difficult to obtain summary statistics using that. So we exclude the NA values to find the summary stat table for both USA and South Dakota. Observing the tables and the numbers we can see that the numbers are representing the true values from each yaar starting quarter-4, 2014 to quarter-4, 2022. In addition, we developed average projections based on the data provided by the analysts for the three primary crops to determine how far the quarterly stock predictions deviated from the reports provided by the USDA.

Table 2: Independent Variables (USA)

	bushels_stocks	carry_over	bushels_production	rail_bushels	barge_bushels	oilseed_crush	ethanol_crush	accumulatedExports
2014	**84411.60**	**15730.03**	**160522.55**	**23132.15**	**42558.20**	**66.87**	**634227.30**	**62415.33**
CORN	63239.54	10344.26	116697.60	7770.86	20034.04	23.26	220600.80	15311.13
SOYBEAN	13031.90	700.51	27853.20	7427.89	21608.37	15.99	151663.05	21195.91
WHEAT	8140.16	4685.27	15971.75	7933.40	915.80	27.62	261963.45	25908.29
2015	**195265.16**	**71125.26**	**623459.66**	**78662.35**	**131181.22**	**228.60**	**2932394.85**	**438710.29**
CORN	142331.08	50026.28	458640.54	29079.95	78075.44	86.10	1234805.04	162741.51
SOYBEAN	25984.65	4514.41	115330.32	28290.41	45539.49	62.78	882720.09	204209.29
WHEAT	26949.43	16584.56	49488.80	21291.98	7566.30	79.73	814869.72	71759.49
2016	**213003.24**	**86776.22**	**653016.51**	**85071.51**	**153595.54**	**230.04**	**3171164.22**	**438165.41**
CORN	150477.94	58787.60	474346.08	31238.17	95856.82	85.90	1301015.52	157700.19
SOYBEAN	29539.48	6070.27	123587.91	30485.98	49311.25	64.43	975761.64	203747.78
WHEAT	32985.82	21918.35	55082.52	23347.36	8427.47	79.71	894387.06	76717.44
2017	**227974.16**	**104041.41**	**668144.89**	**84006.28**	**146957.70**	**230.03**	**3358178.46**	**511885.13**
CORN	162793.42	68782.18	489744.18	31156.21	91133.70	86.43	1390495.68	204385.23
SOYBEAN	33290.29	7770.60	130652.10	30315.76	46611.04	64.82	1042871.76	225386.23
WHEAT	31890.46	27488.64	47748.61	22534.31	9212.97	78.78	924811.02	82113.66
2018	**233167.94**	**115521.89**	**656945.57**	**83774.12**	**144022.69**	**250.33**	**3322763.64**	**480241.16**
CORN	160295.39	75715.24	479208.42	31333.55	99054.64	94.61	1388348.64	204902.76
SOYBEAN	40474.37	11986.69	131977.98	30598.06	37928.96	70.95	1041261.48	206104.29
WHEAT	32398.19	27819.97	45759.17	21842.51	7039.10	84.77	893153.52	69234.11
2019	**233338.12**	**123169.27**	**628400.65**	**80013.47**	**99829.89**	**250.44**	**3230454.15**	**438486.98**
CORN	154179.44	74100.74	459739.98	29973.60	51784.41	94.25	1351380.24	192651.35
SOYBEAN	47073.65	22217.49	120306.60	29124.68	40631.41	70.69	1013535.18	172641.72
WHEAT	32085.03	26851.03	48354.07	20915.18	7414.07	85.50	865538.73	73193.91
2020	**215013.25**	**119849.61**	**617179.66**	**88309.84**	**141582.17**	**267.71**	**3107384.91**	**419926.22**
CORN	147046.83	70106.94	453168.36	32255.59	78386.22	99.93	1279817.28	150023.65
SOYBEAN	38554.61	23663.81	116955.81	31555.09	55545.94	74.95	959862.96	185584.54
WHEAT	29411.81	26078.86	47055.49	24499.16	7650.01	92.83	867704.67	84318.03
2021	**194503.66**	**87574.72**	**650460.30**	**94925.23**	**148052.35**	**258.04**	**3165423.03**	**560934.72**
CORN	139254.32	52212.01	478067.31	35042.71	105156.39	96.56	1310919.12	238844.86
SOYBEAN	30760.75	12777.05	130418.64	34436.24	36183.69	72.42	983189.34	251832.11
WHEAT	24488.60	22585.66	41974.35	25446.28	6712.28	89.06	871314.57	70257.75
2022	**136526.42**	**56449.57**	**539057.15**	**69091.35**	**98677.11**	**214.01**	**2849213.43**	**458321.77**
CORN	96189.99	34798.43	398290.47	26541.30	61274.98	82.25	1220032.80	208705.80
SOYBEAN	22647.33	6972.47	109693.68	26071.52	31353.93	61.69	915024.60	201174.96
WHEAT	17689.10	14678.67	31073.00	16478.53	6048.20	70.07	714156.03	48441.01

Table 3: Independent Variables (South Dakota)

	bushels_stocks	carry_over	bushels_production	rail_bushels	barge_bushels	oilseed_crush	ethanol_crush	accumulatedExports
2014	**817.92**	**101.22**	**1148.57**	**199.95**	**362.75**	**0.47**	**4595.85**	**471.94**
CORN	607.50	75.90	787.36	66.65	139.13	0.16	1531.95	106.33
SOYBEAN	132.80	3.44	229.95	66.65	218.27	0.16	1531.95	214.10
WHEAT	77.62	21.88	131.26	66.65	5.36	0.16	1531.95	151.51
2015	**1873.98**	**444.29**	**4443.27**	**722.05**	**1015.42**	**1.70**	**21813.44**	**3499.60**
CORN	1389.95	336.53	3174.26	262.75	542.19	0.62	8575.04	1130.15
SOYBEAN	249.79	22.61	930.94	262.75	430.73	0.62	8575.04	1960.07
WHEAT	234.23	85.15	338.07	196.55	42.50	0.47	4663.37	409.38
2016	**2095.63**	**553.39**	**4560.24**	**686.93**	**1169.44**	**1.67**	**23096.64**	**3413.41**
CORN	1523.60	392.56	3251.40	251.27	665.67	0.61	9034.83	1095.14
SOYBEAN	275.49	34.63	982.87	251.27	456.59	0.61	9034.83	1886.55
WHEAT	296.54	126.20	325.97	184.39	47.18	0.46	5026.98	431.71
2017	**2106.74**	**685.40**	**4313.99**	**658.26**	**1118.06**	**1.69**	**24667.23**	**3982.29**
CORN	1562.17	481.93	3125.06	250.93	632.87	0.61	9656.22	1419.34
SOYBEAN	324.77	54.35	994.29	250.93	431.58	0.61	9656.22	2086.91
WHEAT	219.80	149.12	194.64	156.39	53.60	0.47	5354.79	476.04
2018	**2125.06**	**689.26**	**4199.33**	**733.96**	**1080.24**	**1.83**	**24505.74**	**3736.19**
CORN	1527.06	508.56	3028.40	275.09	687.88	0.66	9641.31	1422.94
SOYBEAN	392.33	77.64	984.66	275.09	351.19	0.66	9641.31	1908.37
WHEAT	205.67	103.05	186.27	183.79	41.16	0.50	5223.12	404.88
2019	**2040.70**	**755.96**	**3667.47**	**529.90**	**779.19**	**1.87**	**23830.80**	**3364.43**
CORN	1374.28	492.91	2669.76	196.30	359.61	0.68	9384.59	1337.86
SOYBEAN	464.91	158.84	794.60	196.30	376.22	0.68	9384.59	1598.53
WHEAT	201.51	104.21	203.11	137.31	43.36	0.52	5061.63	428.03
2020	**1962.71**	**816.35**	**3507.38**	**619.09**	**1101.82**	**1.95**	**22745.73**	**3240.29**
CORN	1365.72	484.13	2556.36	223.12	544.35	0.71	8887.62	1041.83
SOYBEAN	386.42	201.60	745.04	223.12	514.31	0.71	8887.62	1718.38
WHEAT	210.57	130.62	205.98	172.85	43.16	0.53	4970.49	480.08
2021	**1744.79**	**590.98**	**3952.17**	**711.06**	**1103.86**	**1.89**	**23214.77**	**4394.20**
CORN	1297.38	353.02	2910.44	260.02	730.25	0.68	9103.61	1658.64
SOYBEAN	293.75	119.37	883.84	260.02	335.03	0.68	9103.61	2331.78
WHEAT	153.66	118.59	157.89	191.01	38.58	0.52	5007.56	403.78
2022	**1692.77**	**426.01**	**3794.72**	**582.58**	**952.18**	**1.95**	**24555.83**	**3874.79**
CORN	1260.79	269.53	2791.28	216.14	500.30	0.71	9641.31	1499.71
SOYBEAN	282.24	71.39	816.52	216.14	414.32	0.71	9641.31	2013.19
WHEAT	149.73	85.09	186.92	150.31	37.55	0.54	5273.21	361.89

METHODOLOGY

Machine learning methods have become increasingly popular in many industries, including agriculture and commodities markets. In agricultural commodities markets, machine learning algorithms can analyze large amounts of data, such as weather patterns, crop yields, and commodity prices, to identify patterns and predict future trends. Analyzing large amounts of data and identifying risks to changes in market information that can affect prices can help traders make more informed decisions and manage price risk.

A machine learning method that is commonly used in agricultural commodities markets is decision trees. Decision trees are algorithms that use a tree-like model of decisions and their possible consequences. Decision trees can help traders decide when to buy or sell commodities based on weather patterns, crop yields, and market trends in agricultural commodities markets. XGBoost, Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library.

The XGBoost algorithm is an ensemble method that makes accurate forecasts for a given target variable by combining the predictions of many decision trees. (Fan, R. E. et al., 2008) It functions by adding decision trees to the model iteratively, with each new tree correcting the faults of the preceding trees. Optimizing the procedure to minimize the sum of the loss function and a regularization term prevents overfitting and enhances generalization performance. Unlike conventional gradient boosting algorithms, XGBoost incorporates several advances that improve its precision, scalability, and performance. XGBoost, for instance, employs a weighted quantile sketch to efficiently determine the

ideal split points, hence reducing the computational expense of tree construction. It also uses a sparsity-aware approach to efficiently handle missing values and sparse data, which we have with the dataset we gathered from various sources with different periods of start points.

The capability of the XGBoost algorithm to handle both sparse and dense data is one of its fundamental achievements. It accomplishes this by employing a block-wise algorithmic strategy that exploits the sparsity of the data to reduce computation time and memory consumption. (Chen, T., et al., 2013) The approach is also highly scalable and can effectively handle enormous datasets with millions or even billions of rows and features.

Scalability is one of the significant advantages of XGBoost. It accomplishes this by utilizing a framework for distributed computing that can take advantage of several cores and numerous machines to analyze massive datasets efficiently. The XGBoost algorithm relies on the success of gradient boosting, a well-known machine-learning technique that combines the outputs of multiple weak classifiers to create a single robust classifier. (Chen, T., & Guestrin, C., 2016) In addition, the algorithm employs a cache-aware parallelism technique that reduces the overhead of data transmission between CPU and Memory, enhancing its performance. Chen T. et al. (2015) discussed that Its accuracy and scalability have made it one of the most popular machine learning algorithms among data scientists and machine learning practitioners.

The XGBoost algorithm's capacity to manage missing data is an additional essential characteristic. It accomplishes this by assigning each missing value to one of the branches during the tree creation process based on the optimal split derived using the existing data.

In numerous applications, including predictive modeling, natural language processing, computer vision, and recommendation systems, the XGBoost algorithm has attained state-of-the-art performance.

In addition, XGBoost includes a regularization strategy that prevents overfitting and improves the model's generalization performance. The regularization term penalizes complex models, encouraging the computer to choose simpler models that generalize to new data more effectively. XGBoost has gained popularity because of its user-friendliness and the availability of user-friendly libraries in several programming languages, including Python, R, Java, and C++. (Bengio et al. 2013) The approach has been a mainstay in the machine learning toolkits of many data scientists. It has been implemented in various applications, such as predictive modeling, anomaly detection, and recommendation systems.

It's not necessarily accurate to say that XGBoost is universally "better" than other machine learning algorithms for predicting bushels stock of grain commodities in the USA. The performance of different algorithms will depend on various factors, including the specific data set, the features used in the model, and the hyperparameters are chosen.There are some reasons why XGBoost may outperform other algorithms in specific contexts. For example, XGBoost has been shown to perform well in cases with many input variables and data points, often in predicting. XGBoost is particularly good at handling high-dimensional data and can control numeric and categorical variables, which is useful when dealing with complex data sets like those in the agricultural industry.

Additionally, XGBoost's regularization technique can help prevent overfitting, a common problem when building models with many features. (Chen & Guestrin, 2016)

Overfitting occurs when the model fits too closely to the training data and does not generalize well to new data. The regularization term in XGBoost helps reduce the model's complexity and encourages it to select simpler models that generalize better to new data.In contrast, neural networks can be mighty in cases where the data set is extensive and complex, but they may require significant computational resources and expertise to train appropriately. (Goodfellow et al. 2016) Similarly, random forests, LDA, and QDA can also be effective in specific contexts. Still, they may not perform well when the data is high-dimensional or has many input variables. (Hastie et al. 2009)

Ultimately, the choice of which algorithm to use will depend on the specific requirements of the problem, as well as the available data and computational resources. It is essential to consider each algorithm's strengths and limitations carefully and experiment with different approaches to find the best solution for the given problem.

Another advantage of XGBoost is its ability to handle missing data effectively. In many real-world datasets, missing values are common, and dealing with missing data can be challenging. XGBoost is designed to handle missing values automatically by treating them as a separate category and allowing the algorithm to determine the optimal imputation strategy.

Furthermore, XGBoost's scalability and efficiency suit big data problems well. The algorithm is designed to work efficiently on distributed computing platforms, making it a good choice for large-scale datasets. (Zhang & Xu, 2018) It also has a built-in parallelization feature that can leverage multi-core CPUs and GPUs, making it a powerful tool for tackling computationally intensive tasks. (Brownlee, 2019) Regarding interpretability, XGBoost provides several tools to help users understand how the model

makes predictions. For example, it can generate feature importance scores, which give insight into which features are most important in making predictions. It also allows users to visualize the structure of the decision trees in the model, which can help them understand how the model is making decisions.

In summary, XGBoost is a powerful and versatile machine-learning algorithm with several advantages over other methods in specific contexts. Its ability to handle high-dimensional data, missing values, and scalability make it well-suited for predicting bushels stock of grain commodities in the USA. However, it's essential to keep in mind that the choice of algorithm will depend on the specific requirements of the problem, and it's always a good idea to experiment with different approaches to find the best solution.

METHODS & PROCEDURE

This study aims to find the grain stock level of the three major commodities (Corn, Soybean & Wheat) by incorporating all the publicly available information on the grains using the XGBoost algorithm. We first performed data cleaning and preprocessing to prepare the data for analysis, which involved addressing missing values, managing outliers, and standardizing the variables, as all the interested features used in this study don't have that much data back from 2008. We grouped our dataset based on the states and the region too. According to the USDA, ten regions cover all the agricultural producing different crops. For a better output of the result, we created dummy variables for each of the ten farming regions and each quarter of the marketing year.

Initially, using the dplyr package in R to group the data in the master1b data frame in order to load our primary dataset containing information from 2008 to December 2022

by various variables, including commodity, state name, regions, quarter, and various binary variables that indicate whether a commodity is present (e.g., commodity CORN, commodity SOYBEANS). The information is then organized by state name, product, quarter, and year. Then, using the mutate() method, multiple new variables were created by dividing the log of existing variables by the lag() function of the same variable. These new variables represent percentage changes from the previous year in stocks, production, ethanol crushing, exports, rail and barge shipments, and oilseed crushing. In addition, two additional variables have been established to reflect the percentage change in stocks relative to the previous quarter (stock chg qtr lag1) and the percentage change in carryover stocks relative to the prior year (carry chg yoy). The same variables then group the data as previously. The mutate() function is utilized again to generate a new variable representing the percentage change in stock relative to the prior quarter (stock chg qtr lag1). Afterward, the ungroup() method is used to ungroup the data. Afterwards, we divide our dataset into train and validate portions. The training dataset covers data prior to and including 2020, whereas the validation dataset contains data after 2020. Then, the training data and model parameters (nrounds, eta, max depth, method, objective, nthread, and verbose) are used to generate an XGBoost model. We constructed the model separately for each crop, including corn, soybeans, and wheat, as well as per quarter. We divide the quarters by commodity and by quarter 3 and quarters 1, 2, and 4 because the marketing year begins in quarter 3 for the majority of the crops with the exception of wheat. Here is the list of features we used in our XGBoost for the prediction of the grain stocks.

Table 4: List of Features use in XGBoost

Name	Description
stock_chg_yoy	Natural Log of year over year percentage change of grain stock
prod_chg_yoy	Natural Log of year over year percentage change of production
ethanol_chg_yoy	Natural Log of year over year percentage change of grain used in ethanol
export_chg_yoy	Natural Log of year over year percentage change of grain exported
rail_chg_yoy	Natural Log of year over year percentage change of grain transit in rail
barge_chg_yoy	Natural Log of year over year percentage change of grain transit in barges
oilseed_chg_yoy	Natural Log of year over year percentage change of grain used in oilseed crush
carry_chg_yoy	Natural Log of year over year percentage change of grains carry over from previous year
stock_chg_yoy_lag1	Year over year percentage change in the stock from one year ago (i.e., a lag of one year)
stock_chg_yoy_lag2	Year over year percentage change in the stock from two years ago (i.e., a lag of two years)

An XGBoost model is a weighted sum of decision trees, where each tree predicts an increase in the target variable given the input variable values. (Chen & Guestrin, 2016) Let X be the matrix of input variables of dimension (nxp), where n is the number of observations and p is the number of features. Let y or $stock - chg - yoy$ be the vector of the target variable of dimension $(nx1)$.We define a decision tree T_j as a function that takes an input vector x_i and returns a prediction $f_j(x_i)$ of the target variable.

Let F be the additive ensemble of decision trees $F(x) = \sum_{j=1}^{m} f_j\,(x)$ that models the relationship between the input variables X and the target variable y. The XGBoost algorithm learns F by minimizing the following regularized objective function:

$$Obj(F) = \sum_{i=1}^{n} L\left(y_i, F(x_i)\right) + \sum_{j=1}^{m} \Omega\left(f_j\right)$$

Where $L\left(y_i, F(x_i)\right)$ is the loss function that measures the difference between the predicted value and the actual value of the target variable for the i-th training example, and $\Omega\left(f_j\right)$ is the regularization term that penalizes the complexity of the j-th decision tree. The regularization term is defined as:

$$\Omega\left(f_j\right) = \gamma T + \frac{1}{2}\lambda \sum_{i=1}^{T_j} w_{i,j}^2$$

Where T is the number of leaf nodes in the j-th decision tree, $w_{i,j}$ is the weight of the i-th leaf node in the j-th decision tree, and γ and λ are hyperparameters that control the strength of the regularization. The XGBoost algorithm minimizes the objective function $Obj(F)$ by gradient boosting, which iteratively fits new decision trees to the negative gradient of the loss function with respect to the current prediction of the model. The gradient boosting step can be written as follows:

$$f_j = \underset{f}{\operatorname{argmin}} \sum_{i=1}^{n} \left[-\nabla_{F(x_i)} L\left(y_i, F(x_i)\right) - \Omega\left(F(x_i)\right)\right]^2$$

Where $\nabla_{F(x_i)} L\big(y_i, F(x_i)\big)$ is the gradient of the loss function with respect to the current prediction of the model for the i-th training example. Finally, the XGBoost model is given by:

$$F(x) = \sum_{j=1}^{m} \eta\, f_j(x)$$

Where η is the learning rate that controls the contribution of each decision tree to the final prediction. In this paper we run the XGBoost regression model with the following hyperparameters:

- nrounds: the number of boosting rounds or iterations, set to 300.
- eta: the model's learning rate or shrinkage rate, set to 0.05.
- max_depth: the maximum depth of each decision tree in the boosting process, set to 5.
- method: the tree construction method, set to "hist" uses histogram-based approximation for faster speed and lower memory usage.
- objective: the loss function to be optimized during training, set to "reg:squarederror" which is mean squared error regression.
- nthread: the number of threads used for parallel computing, set to 2.
- verbose: the verbosity level is set to 0 to suppress any output during training.

This XGBoost model is trained using the input training data xgb_train and is used for regression, i.e., to predict continuous numerical values. The goal of the training process is to minimize the mean squared error between the predicted values and the actual target values.

EMPIRICAL ANALYSIS

At first, we run our XGBoost for the commodity corn and quarter 3 with the features "stock_chg_yoy", "prod_chg_yoy", "ethanol_chg_yoy", "export_chg_yoy," "rail_chg_yoy," "barge_chg_yoy," "oilseed_chg_yoy," "carry_chg_yoy," "stock_chg_yoy_lag1", "stock_chg_yoy_lag2", "commodity_CORN," "quarter_Q3", and "regions_Corn.Belt". Then we created the important features of a plot (figure 3) to understand the XGBoost model and what features it uses to predict, and what features contribute the most.

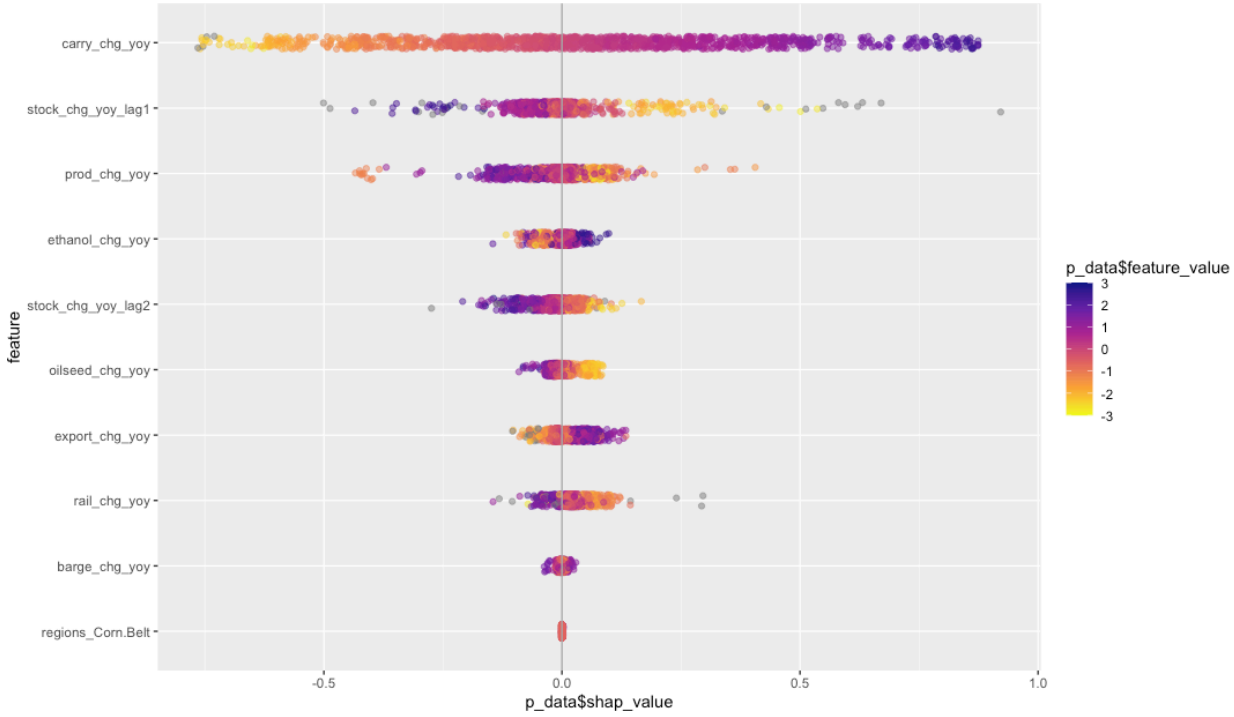


Figure 3: Important Features plot for Commodity Corn & Quarter 3

We also created the partial summary plot (figure 4) of the important features in the below. Above these we will try to interpret the top 3 important features contributing to our model.

For the corn in quarter 3, the natural log of percentage change in carry-over stock

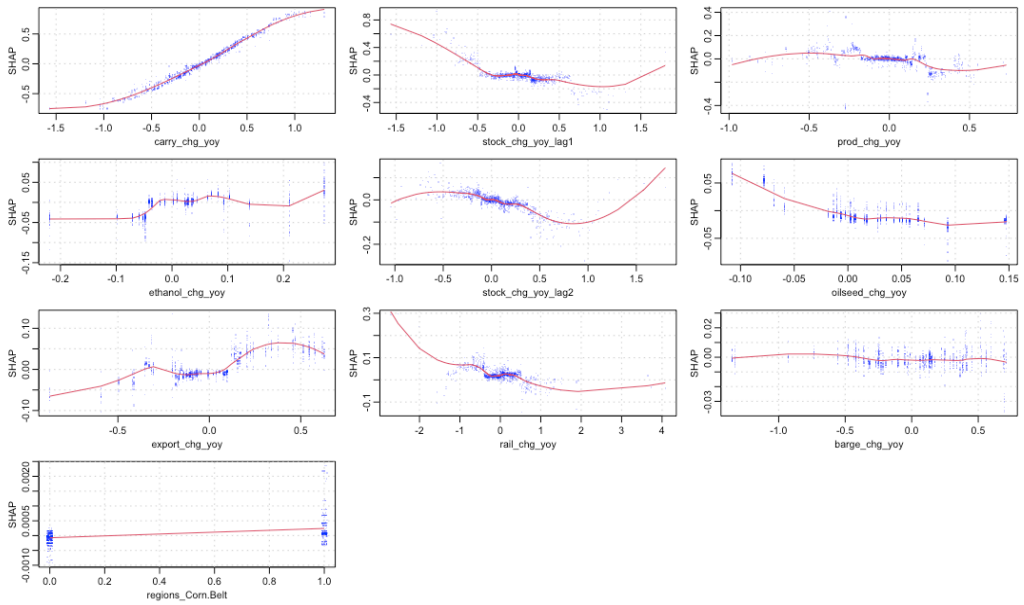


Figure 4: Partial Summary of the important features

contributed as the most important, followed by the percentage change in corn stock in the previous year and the natural log of percentage change in corn production. The rest of the features contribute to the model but are lower than the three above. Now we look individually at these three features and how these are shaping the final prediction of estimating the stock of corn for the following quarters.

The plot (Figure 5) shows the SHAP values for the feature "carry_chg_yoy" in the model. The y-axis shows the SHAP value of the feature for each sample in the test set, while the x-axis shows the value of the feature itself. The color of each dot represents the value of another feature that is highly correlated with "carry_chg_yoy." The blue color indicates a low value, while the red color indicates a high value. The plot shows the relationship between the feature "carry_chg_yoy" and the output of the XGBoost model. The plot shows that higher values of "carry_chg_yoy" are associated with higher values in the model

output, which suggests that increasing "carry_chg_yoy" could increase the model output. When the carry-over of corn from the previous year is increased by 1%, the corn stock for the upcoming quarter will increase by 0.5% approx ceteris paribus.

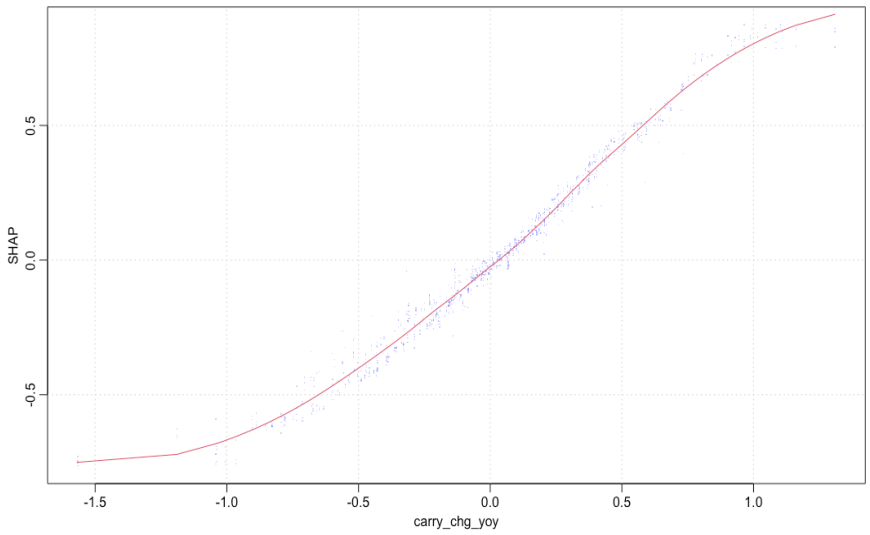


Figure 5 Partial of Carry_chg_yoy for Corn & Q3

The plot (Figure 6) shows the relationship between the feature "prod_chg_yoy" and the output of the XGBoost model. In this case, the slope of the line is close to zero, which indicates that the change in the "prod_chg_yoy" variable is tiny relative to the change in the x variable. Nevertheless, there is a change in the slope; while the percentage change in the production of corn has a positive value, the SHAP values present a negative relation to the model. However, it is essential to note that the absence of a correlation does not necessarily mean that there is no relationship between the variables at all - it may simply be too weak or complex to detect using the particular analytical methods employed.

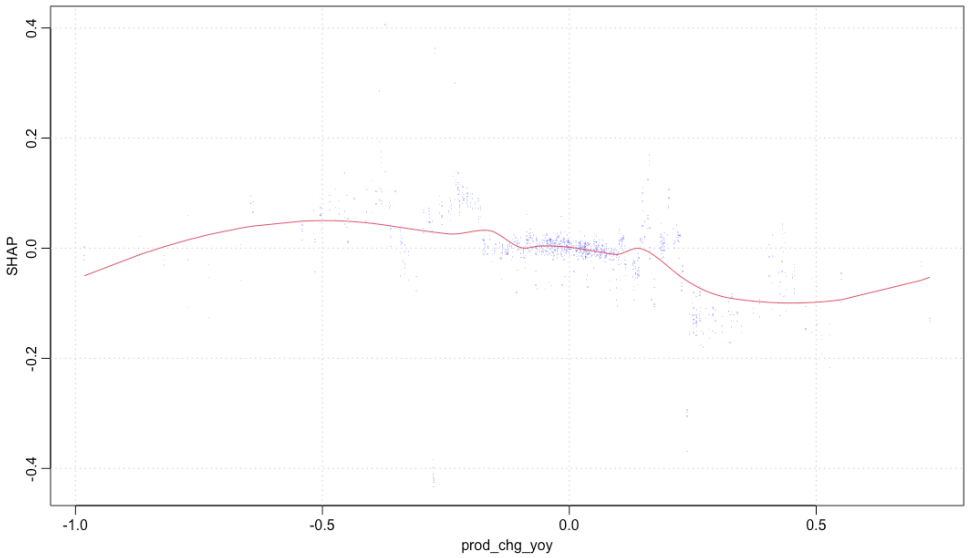


Figure 6: Partial of Prod_chg_yoy for Corn & Q3

The plot (Figure 7) shows the relationship between the feature "stock_chg_yoy_lag1" and the output of the XGBoost model. The plot shows that higher values of "stock_chg_yoy_lag1" are associated with lower values in the model output (vice-versa), which suggests that decreasing "stock_chg_yoy_lag1" could increase the values in the model output. When the year-over-year percentage change in the stock from one year ago

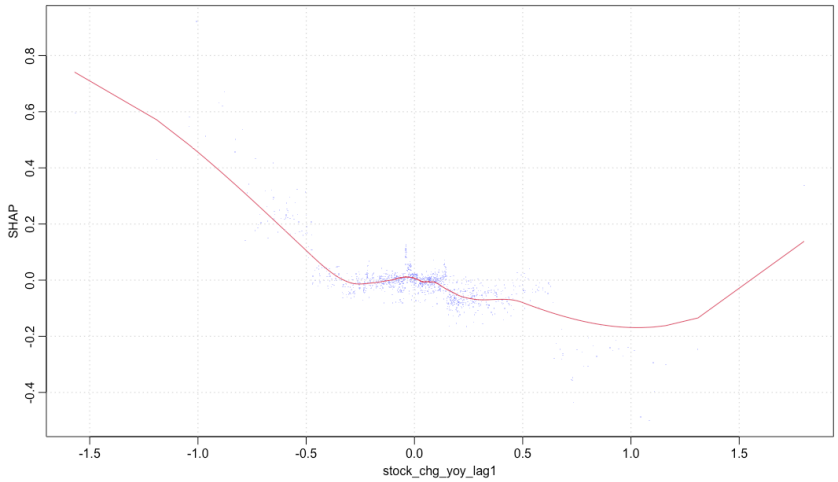


Figure 7: Partial of stock_chg_yoy_lag1 for Corn & Q3

is increased by 1%, the corn stock for the upcoming quarter will decrease by around 0.6% approx. ceteris paribus.

We run our XGBoost for the commodity corn and quarter 1, 2 and 4 with the features "stock_chg_yoy", "prod_chg_yoy", "ethanol_chg_yoy", "export_chg_yoy," "rail_chg_yoy," "barge_chg_yoy," "oilseed_chg_yoy," "carry_chg_yoy," "stock_chg_yoy_lag1", "stock_chg_yoy_lag2", "commodity_CORN," "quarter_Q1," “quarter_Q2,” “quarter_Q4,” and "regions_Corn.Belt". Then we created the important features of a plot (figure 8) to understand the XGBoost model and what features it uses to

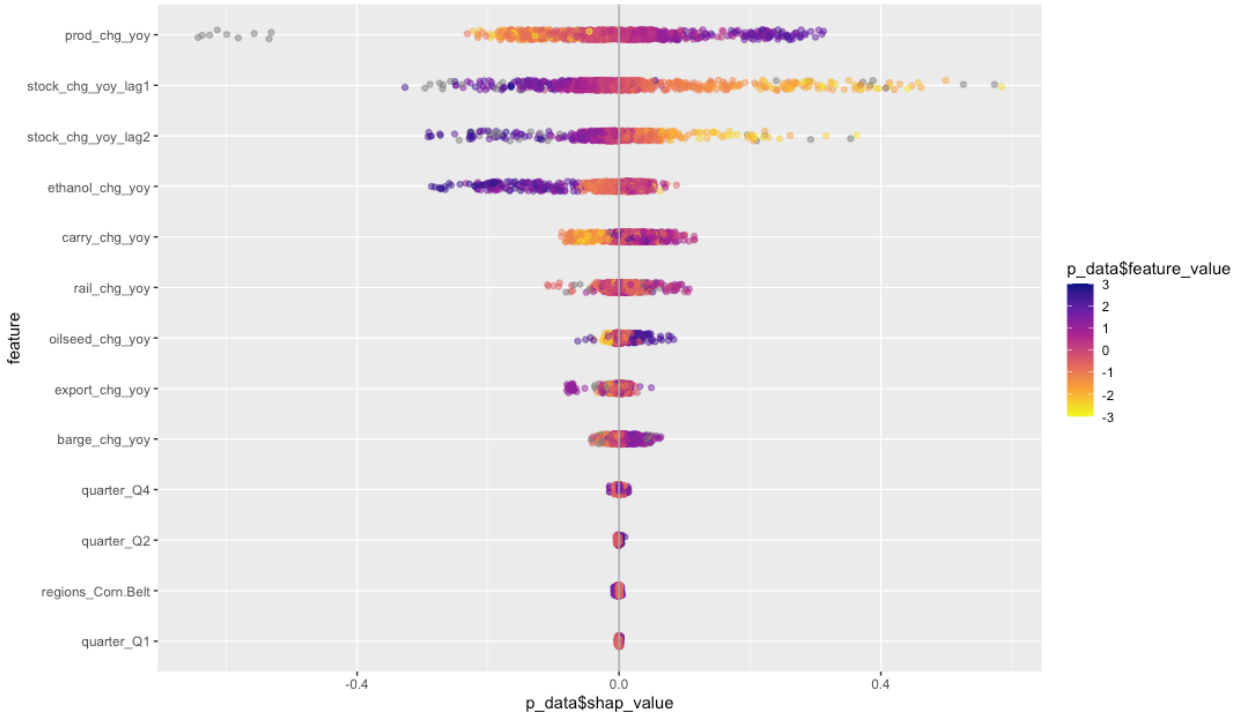


Figure 8: Important Features plot for Corn & Quarter 1,2 and 4

predict, and what features contribute the most.

We also created the partial summary plot (figure 9) of the important features in the below. Above these we will try to interpret the top 3 important features contributing to our model.

For the corn in quarter 1,2 and 4 the natural log of percentage in corn production contributed as the most important, followed by the percentage change in corn stock in the

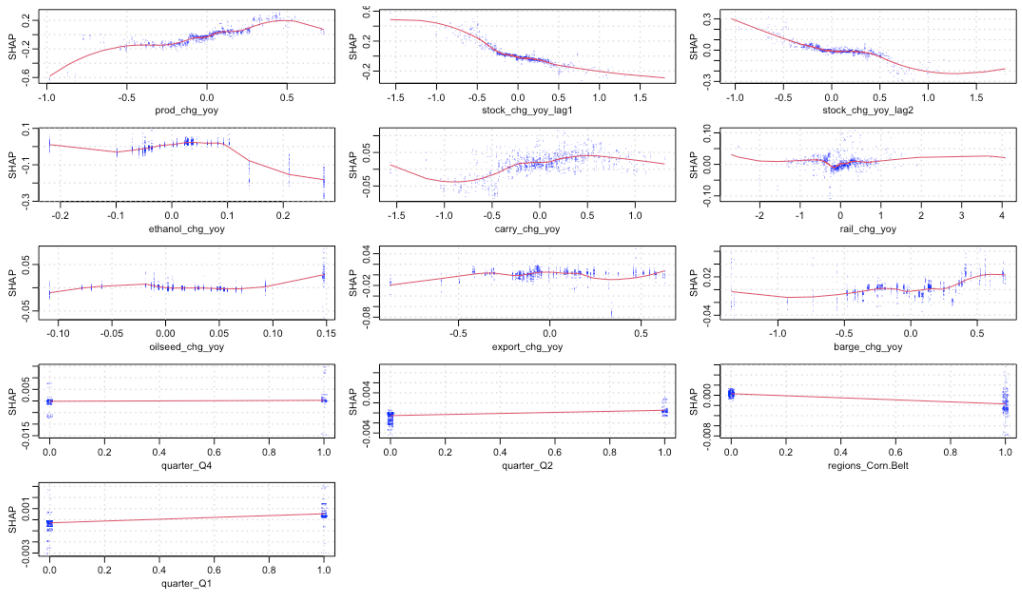


Figure 9: Partial Summary for Corn & Quarter 1,2 and 4

previous year and the natural log of percentage change in stock of corn previous two years. The rest of the features contribute to the model but are lower than the three above. Now we look individually at these three features and how these are shaping the final prediction of estimating the stock of corn for the following quarters.

The plot (Figure 10) shows the SHAP values for the feature "prod_chg_yoy" in the model. The y-axis shows the SHAP value of the feature for each sample in the test set, while the x-axis shows the value of the feature itself. The color of each dot represents the value of another feature that is highly correlated with "prod_chg_yoy." The blue color indicates a low value, while the red color indicates a high value. The plot shows the relationship between the feature "prod_chg_yoy" and the output of the XGBoost model. The plot shows that higher values of "prod_chg_yoy" are associated with high model output, which

suggests that increasing "prod_chg_yoy" could increase the model output. When the production of corn from the previous year is increased by 1%, the corn stock for the upcoming quarter will increase by 0.6% approx. ceteris paribus.

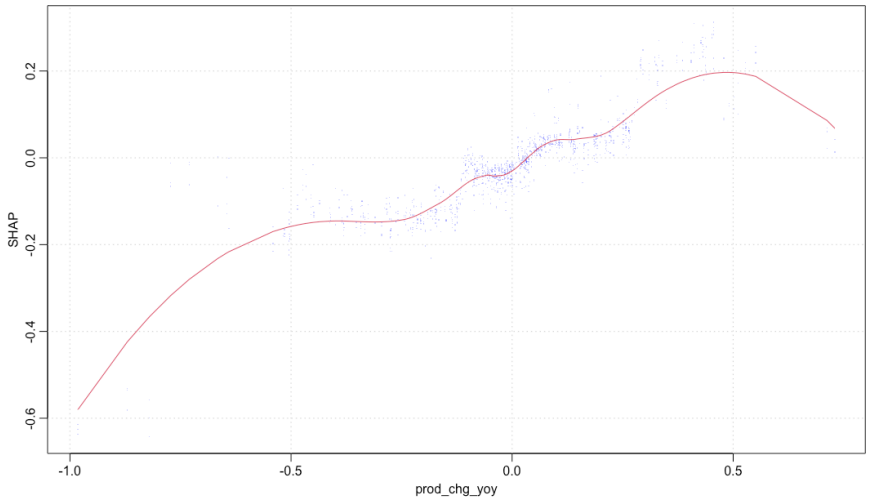


Figure 10 Partial of Prod_chg_yoy for Corn & Q1Q2Q4

The plot (Figure 11) shows the relationship between the feature "stock_chg_yoy_lag1" and the output of the XGBoost model. The plot shows that higher values of "stock_chg_yoy_lag1" are associated with lower values in model output (vice-versa), which suggests that decreasing "stock_chg_yoy_lag1" could increase the values in model output. When the year-over-year percentage change in the stock from one year ago is increased by 1%, the corn stock for the upcoming quarter will decrease by around 0.3% approx. ceteris paribus.

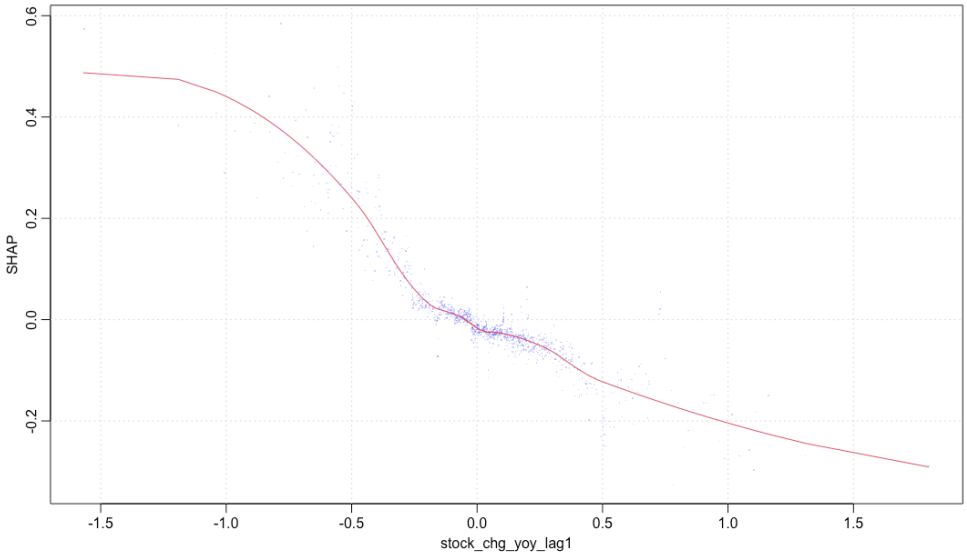


Figure 11: Partial of stock_chg_yoylag1 for Corn & Q1Q2Q4

The plot (Figure 12) shows the relationship between the feature "stock_chg_yoy_lag2" and the output of the XGBoost model. The plot shows that almost higher values of "stock_chg_yoy_lag2" are associated with lower values in model output (vice-versa), which suggests that decreasing "stock_chg_yoy_lag2" could increase the values in model output. When the year-over-year percentage change in the stock from two year ago is

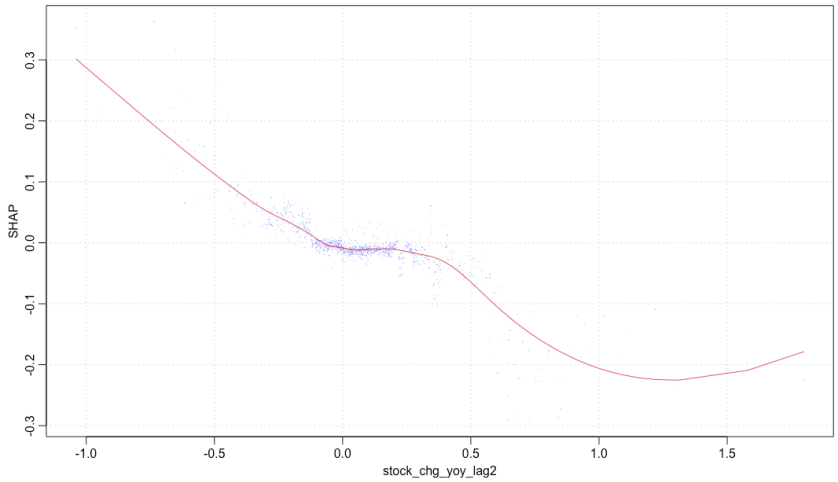


Figure 12: Partial of stock_chg_yoy_lag2 for Corn & Q1Q2Q4

increased by 1%, the corn stock for the upcoming quarter will decrease by around 0.3% approx. ceteris paribus.

We run our XGBoost for the commodity soybean and quarter 3 with the features "stock_chg_yoy", "prod_chg_yoy", "ethanol_chg_yoy", "export_chg_yoy," "rail_chg_yoy," "barge_chg_yoy," "oilseed_chg_yoy," "carry_chg_yoy," "stock_chg_yoy_lag1", "stock_chg_yoy_lag2", "commodity_SOYBEAN," "quarter_Q3", and "regions_Corn.Belt". Then we created the important features of a plot (figure 13) to understand the XGBoost model and what features it uses to predict, and what features

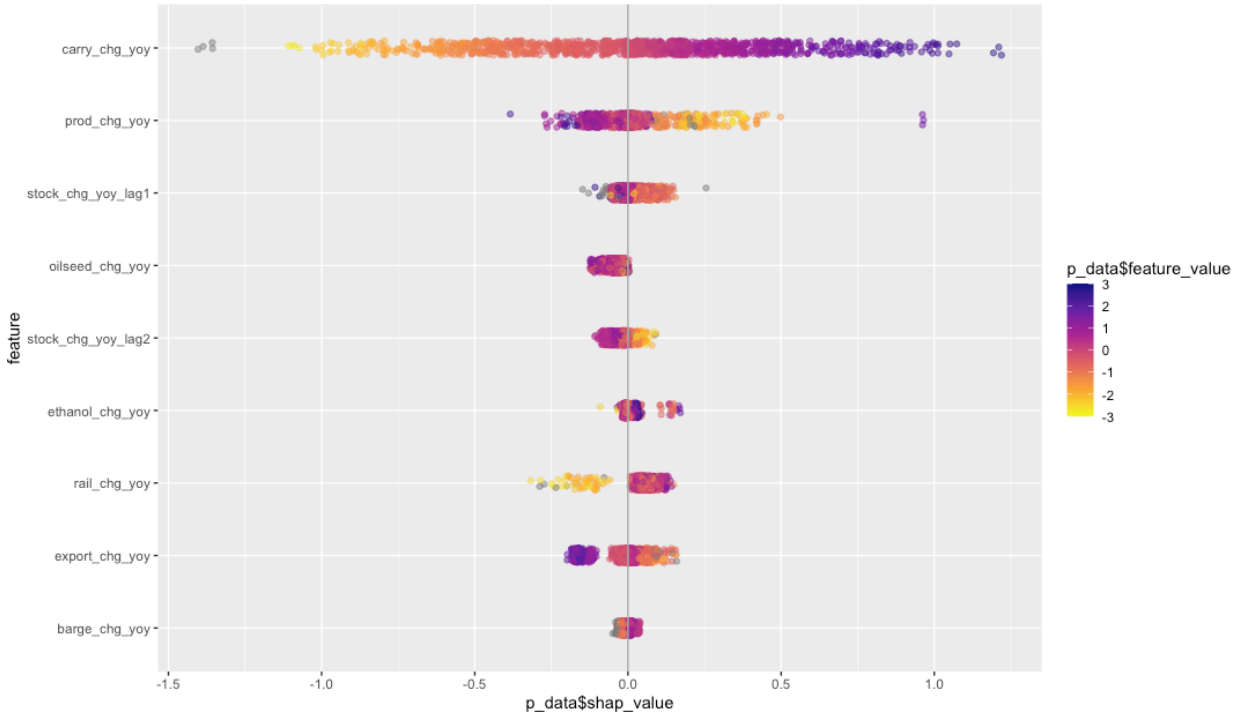


Figure 13: Important Features plot for Soybean & Quarter 3

contribute the most.

We also created the partial summary plot (figure 14) of the important features in the below. Above these we will try to interpret the top 3 important features contributing to our model.

For the soybean in quarter 3 the natural log of percentage in soybean carry-over from previous years contributed as the most important, followed by the percentage change in

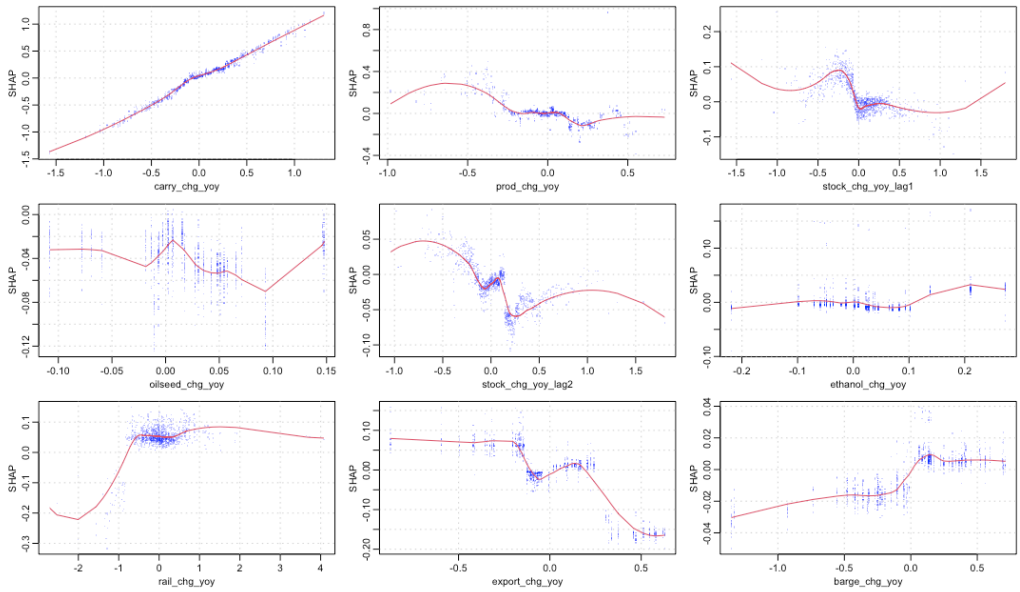


Figure 14: Partial Summary for Soybean & Quarter 3

year over year soybean production and the natural log of percentage change in stock of corn previous one year. The rest of the features contribute to the model but are lower than the three above. Now we look individually at these three features and how these are shaping the final prediction of estimating the stock of corn for the following quarters.

The plot (Figure 15) shows the SHAP values for the feature "carry_chg_yoy" in the model. The y-axis shows the SHAP value of the feature for each sample in the test set, while the x-axis shows the value of the feature itself. The color of each dot represents the value of another feature that is highly correlated with "carry_chg_yoy." The blue color indicates a low value, while the red color indicates a high value. The plot shows the relationship between the feature "carry_chg_yoy" and the output of the XGBoost model. The plot shows that higher values of "carry_chg_yoy" are associated with higher values in model output

(vice-versa), which suggests that increasing "carry_chg_yoy" could increase the values in model output for soybean in quarter 3. When the carry-over of soybean from the previous year is increased by 1%, the corn stock for the upcoming quarter will increase by 1% approx. ceteris paribus.

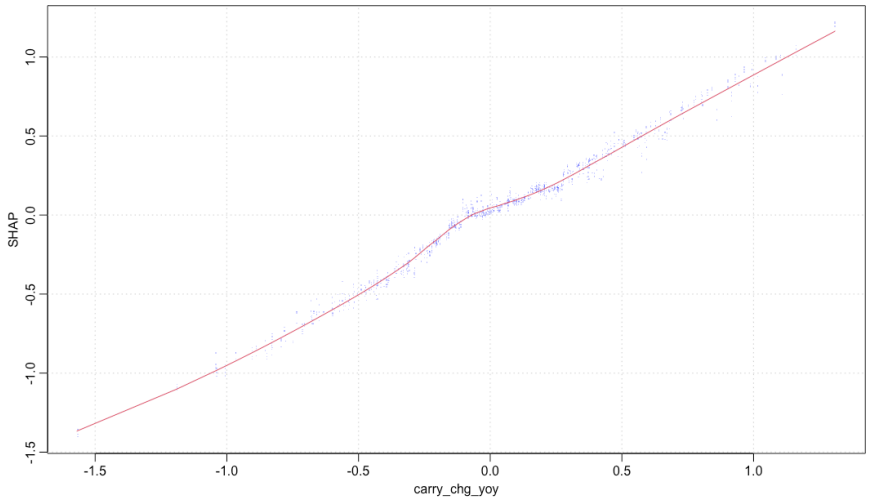


Figure 15 Partial of carry_chg_yoy for Soybean & Q3

The plot (Figure 16) shows the relationship between the feature "prod_chg_yoy" and the output of the XGBoost model. The plot shows that zero values in slope which means of "prod_chg_yoy" are contributing highly but can’t interpreted with a partial plot.

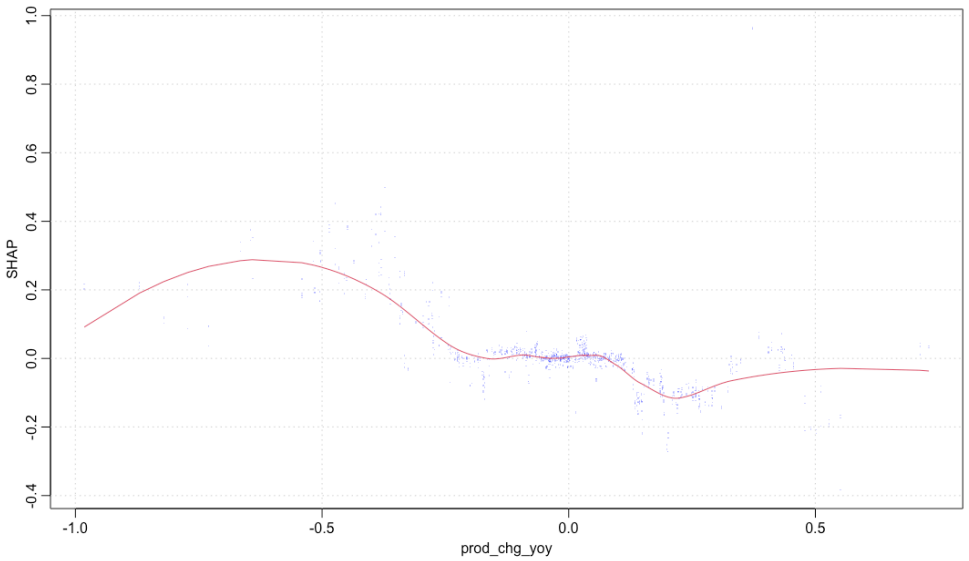


Figure 16: Partial of prod_chg_yoy for Soybean & Q3

The plot (Figure 17) shows the relationship between the feature "stock_chg_yoy_lag1" and the output of the XGBoost model. The plot shows that almost higher values of "stock_chg_yoy_lag1" are associated with lower values in model output (vice-versa) while the feature's value is in between 0% to 0.5%, which suggests that decreasing "stock_chg_yoy_lag1" could increase the values in model output. When the year-over-year

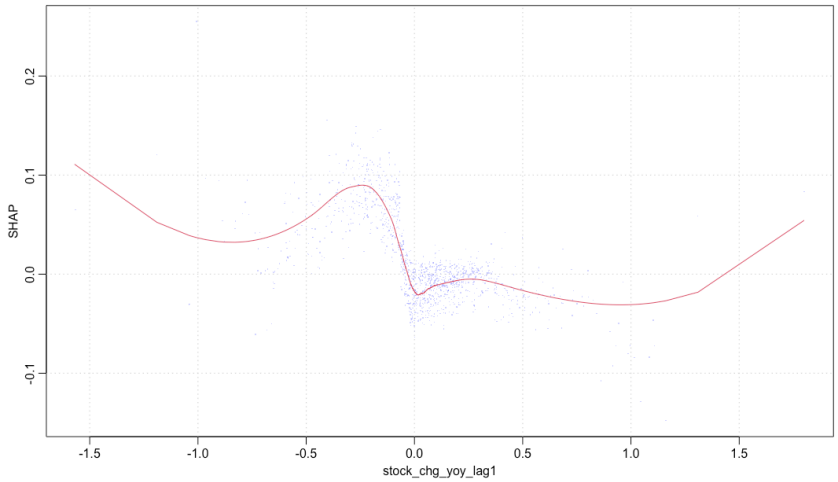


Figure 17: Partial of stock_chg_yoy_lag1 for Soybean & Q3

percentage change in the stock from one year ago is increased by 0.5%, the soybean stock for the upcoming quarter will decrease by around 0.1% approx. ceteris paribus.

We run our XGBoost for the commodity soybean and quarter 1, 2 and 4 with the features "stock_chg_yoy", "prod_chg_yoy", "ethanol_chg_yoy", "export_chg_yoy," "rail_chg_yoy," "barge_chg_yoy," "oilseed_chg_yoy," "carry_chg_yoy," "stock_chg_yoy_lag1", "stock_chg_yoy_lag2", "commodity_SOYBEAN," "quarter_Q1," “quarter_Q2,” “quarter_Q4,” and "regions_Corn.Belt".. Then we created the important features of a plot (figure 18) to understand the XGBoost model and what features it uses to predict, and what features contribute the most.

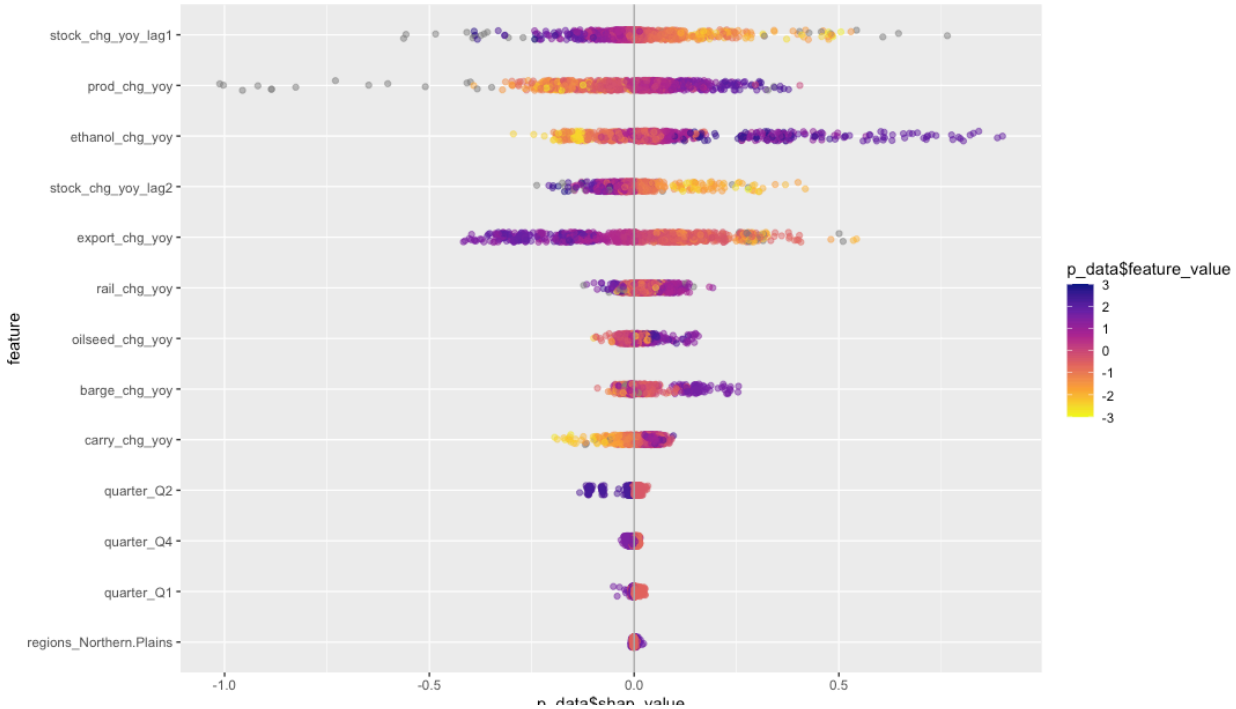


Figure 18: Importance for Soybean & Q1Q2Q4

We also created the partial summary plot (figure 19) of the important features in the below. Above these we will try to interpret the top 3 important features contributing to our model.

For the soybean in quarter 1,2 and 4 the natural log of percentage change in soybean stock for one year ago contributed as the most important, followed by the percentage change in

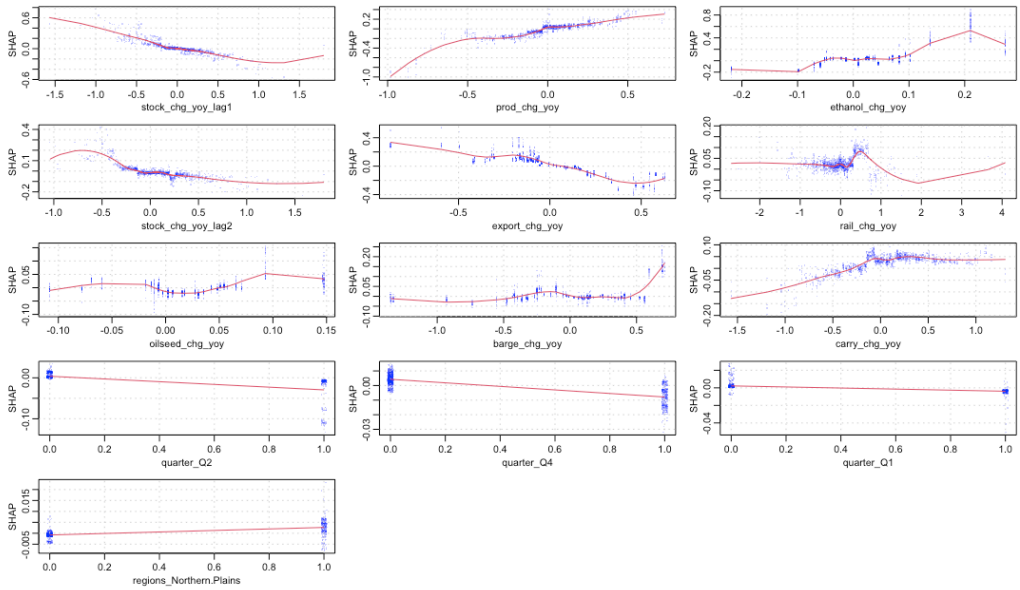


Figure 19: Partial Summary for Soybean & Quarter 1,2 and 4

soybean production and the natural log of percentage change in ethanol production. The rest of the features contribute to the model but are lower than the three above. Now we look individually at these three features and how these are shaping the final prediction of estimating the stock of corn for the following quarters.

The plot (Figure 20) shows the SHAP values for the feature "stock_chg_yoy_lag1" in the model. The y-axis shows the SHAP value of the feature for each sample in the test set, while the x-axis shows the value of the feature itself. The color of each dot represents the value of another feature that is highly correlated with " stock_chg_yoy_lag1." The blue color indicates a low value, while the red color indicates a high value. The plot shows the relationship between the feature "stock_chg_yoy_lag1" and the output of the XGBoost model. The plot shows that higher values of “stock_chg_yoy_lag1" are associated with lower values of the model output (vice-versa), which suggests that increasing

"stock_chg_yoy_lag1" could decrease the values in model output. When the change of stock of soybean from previous year is increased by 1%, the soybean stock for the upcoming quarter will increase by 0.3% approx. ceteris paribus.

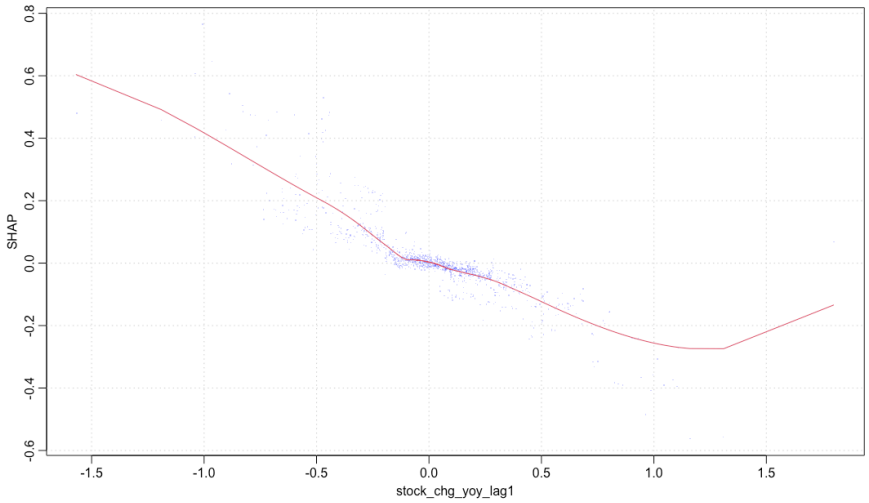


Figure 20 Partial of stock_chg_yoy_lag1 for Soybean & Q1Q2Q4

The plot (Figure 21) shows the relationship between the feature "prod_chg_yoy" and the output of the XGBoost model. The plot shows that almost higher values of "prod_chg_yoy " are associated with higher values in model output, which suggests that increasing " prod_chg_yoy " could increase the model output. When the year-over-year percentage change in the soybean production is increased by 1%, the soybean stock for the upcoming quarter will increase by around 1% approx. ceteris paribus.

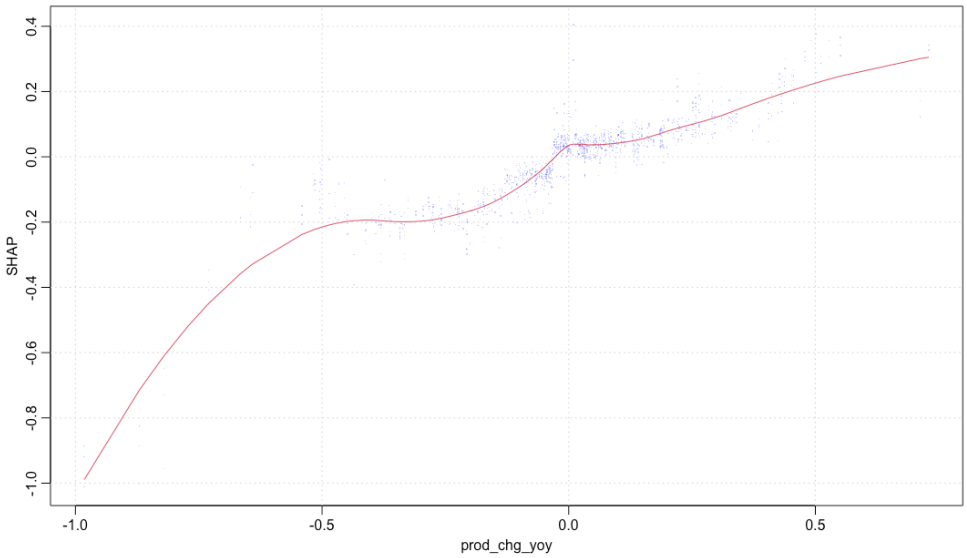


Figure 21: Partial of prod_chg_yoy for Soybean & Q1Q2Q4

The plot (Figure 22) shows the relationship between the feature "ethanol_chg_yoy " and the output of the XGBoost model. The plot shows that almost high positive values of " ethanol_chg_yoy " are associated with higher values in model output, which suggests that increasing "ethanol_chg_yoy " could increase the model output. When the year-over-year percentage change in the ethanol production is increased by 0.2% approx., the soybean

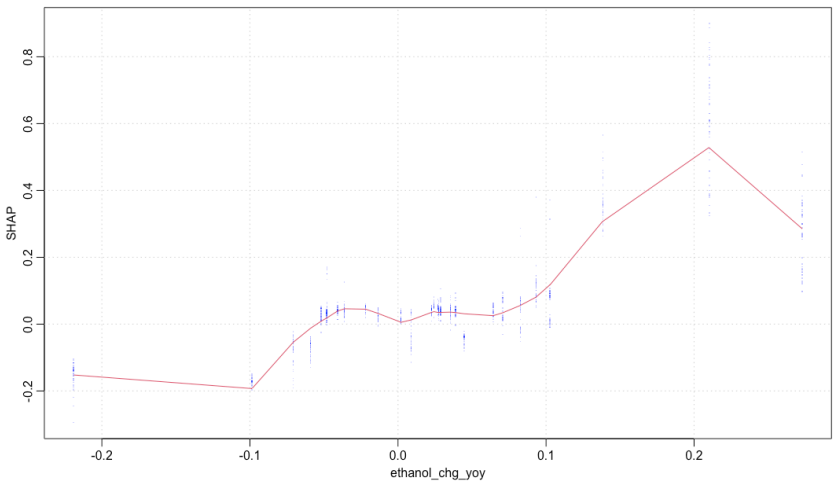


Figure 22: Partial of ethanol_chg_yoy for Soybean & Q1Q2Q4

stock for the upcoming quarter will decrease by around 0.4% ceteris paribus for the model Q1, Q2 and Q4.

We run our XGBoost for the commodity wheat and quarter 3 with the features "stock_chg_yoy", "prod_chg_yoy", "ethanol_chg_yoy", "export_chg_yoy," "rail_chg_yoy," "barge_chg_yoy," "oilseed_chg_yoy," "carry_chg_yoy," "stock_chg_yoy_lag1", "stock_chg_yoy_lag2", "commodity_WHEAT," "quarter_Q3", and "regions_Corn.Belt". Then we created the important features of a plot (figure 23) to understand the XGBoost model and what features it uses to predict, and what features

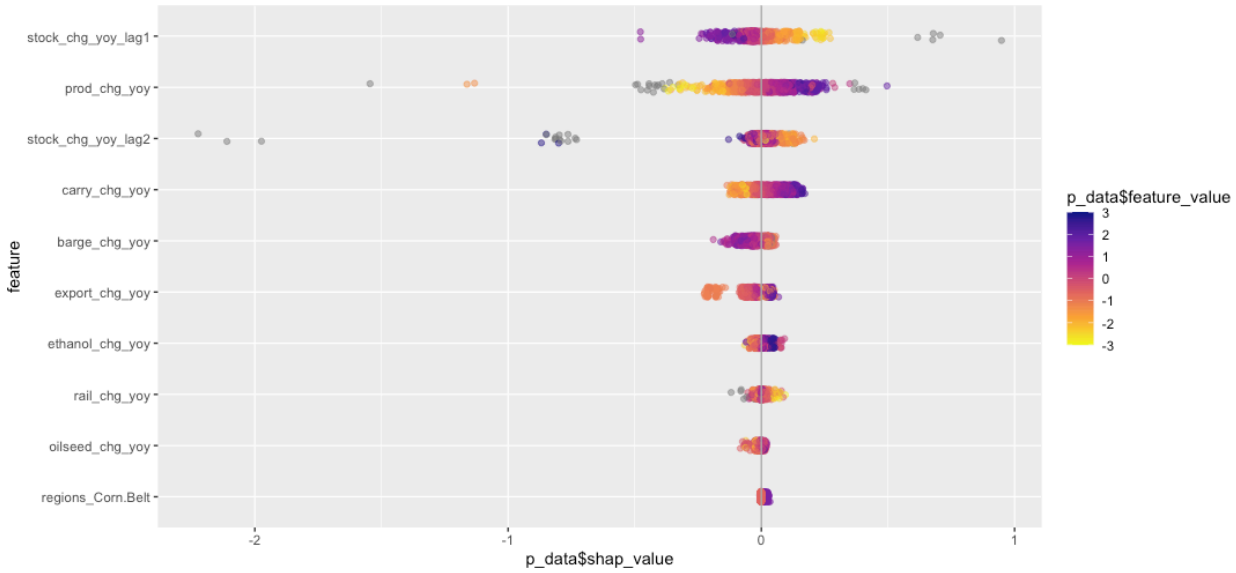


Figure 23: Important Features plot for Wheat & Q3

contribute the most.

We also created the partial summary plot (figure 24) of the important features in the below. Above these we will try to interpret the top 3 important features contributing to our model.

For the wheat in quarter 3, the natural log of percentage change in wheat stock from one year ago contributed as the most important, followed by the percentage change in wheat

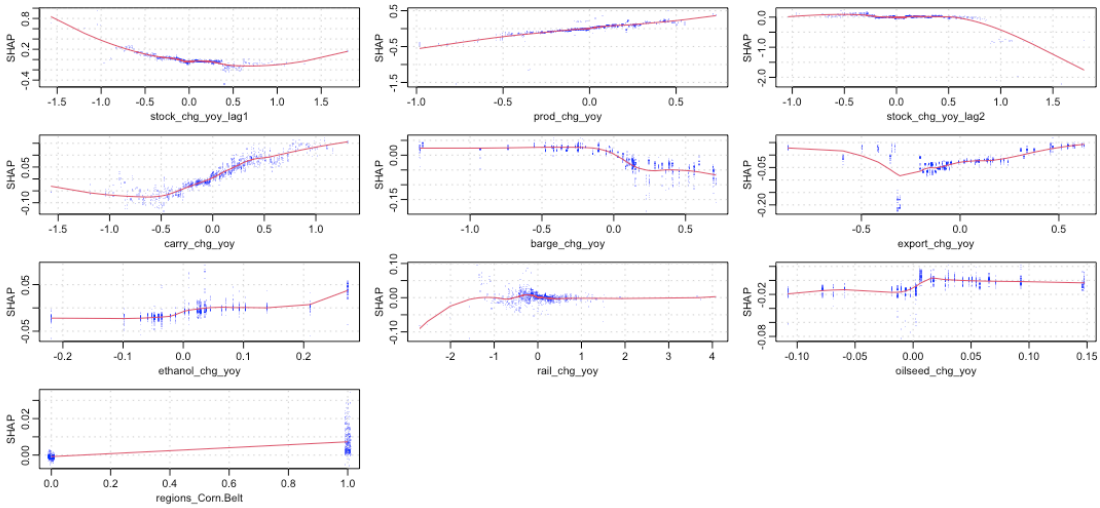


Figure 24: Partial Summary of the important features

production and the natural log of percentage change in wheat stock from two years ago. The rest of the features contribute to the model but are lower than the three above. Now we look individually at these three features and how these are shaping the final prediction of estimating the stock of corn for the following quarters.

The plot (Figure 25) shows the SHAP values for the feature "stock_chg_yoy_lag1" in the model. The y-axis shows the SHAP value of the feature for each sample in the test set, while the x-axis shows the value of the feature itself. The color of each dot represents the value of another feature that is highly correlated with " stock_chg_yoy_lag1." The blue color indicates a low value, while the red color indicates a high value. The plot shows the relationship between the feature "stock_chg_yoy_lag1" and the output of the XGBoost model. The plot shows that a lower value of "stock_chg_yoy_lag1" are associated with higher values in model output, which suggests that increasing "stock_chg_yoy_lag1" could

decrease the model output. When the year over year change in the wheat stock from previous year is increased by 1%, the wheat stock for the upcoming quarter will decrease by 0.4% approx. ceteris paribus.

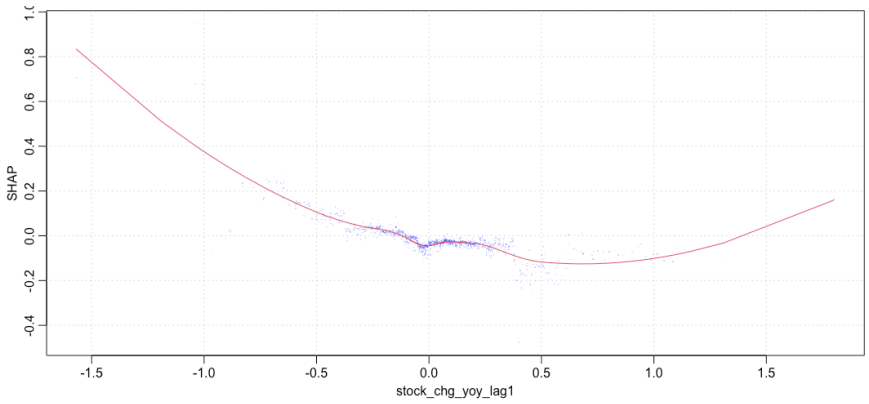


Figure 25 Partial of stock_chg_yoy_lag1 for Wheat & Q3

The plot (Figure 26) shows the relationship between the feature "prod_chg_yoy" and the output of the XGBoost model. In this case, the slope of the line positive, which indicates that the change in the "prod_chg_yoy" variable is positively related to the model output. The plot shows that a lower value of "prod_chg_yoy " are associated with lower values in model output, which suggests that increasing "prod_chg_yoy " could increase the model output. When the year over year change in the wheat production is increased by 1%, the wheat stock for the upcoming quarter will increase by 0.5% approx. ceteris paribus.

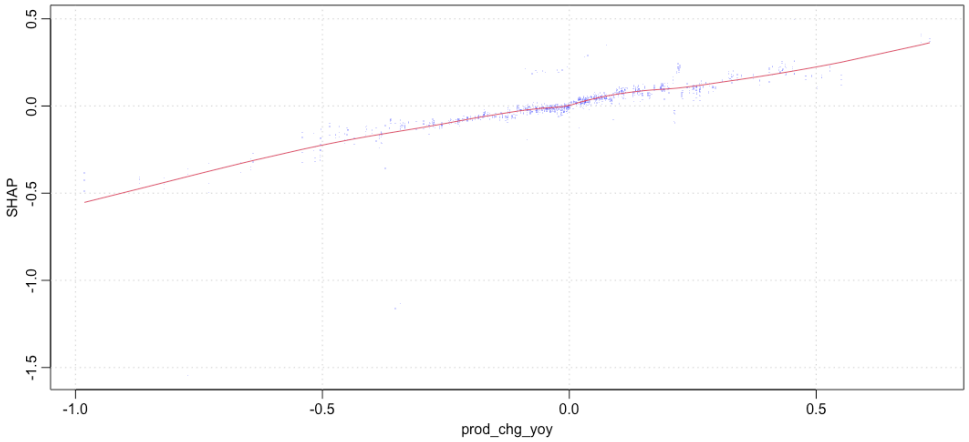


Figure 26: Partial of Prod_chg_yoy for Wheat & Q3

The plot (Figure 27) shows the relationship between the feature "stock_chg_yoy_lag2" and the output of the XGBoost model. The plot shows a flat relationship between the feature and the output of the model until there is a negative relationship once the feature value changes to positive percentages, which suggests that decreasing "stock_chg_yoy_lag2" could increase the values in model output. When the year-over-year percentage change in

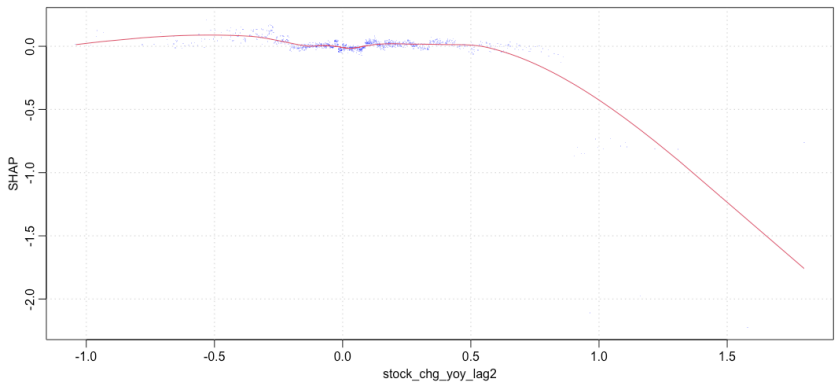


Figure 27: Partial of stock_chg_yoy_lag2 for Wheat & Q3

the wheat stock from two year ago is increased by 0.5%, the wheat stock for the upcoming quarter will decrease by around 0.5% approx. ceteris paribus.

We run our XGBoost for the commodity wheat and quarter 1, 2 and 4 with the features "stock_chg_yoy", "prod_chg_yoy", "ethanol_chg_yoy", "export_chg_yoy," "rail_chg_yoy," "barge_chg_yoy," "oilseed_chg_yoy," "carry_chg_yoy," "stock_chg_yoy_lag1", "stock_chg_yoy_lag2", "commodity_WHEAT," "quarter_Q1," “quarter_Q2,” “quarter_Q4,” and "regions_Corn.Belt".. Then we created the important features of a plot (figure 28) to understand the XGBoost model and what features it uses to predict, and what features contribute the most.

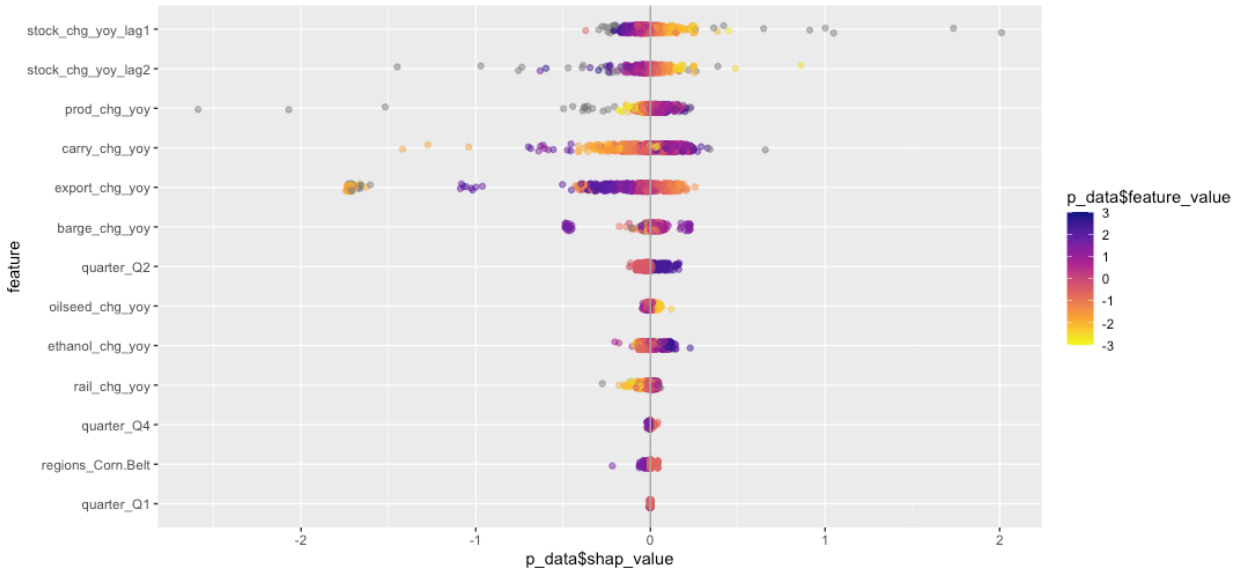


Figure 28: Important Features Plot for Wheat & Q1Q2Q4

We also created the partial summary plot (figure 29) of the important features in the below. Above these we will try to interpret the top 3 important features contributing to our model.

For the wheat in quarter 1,2 and 4 the natural log of percentage change in wheat stock for one year ago contributed as the most important, followed by the percentage change in wheat stock from two years ago and the natural log of percentage change in wheat

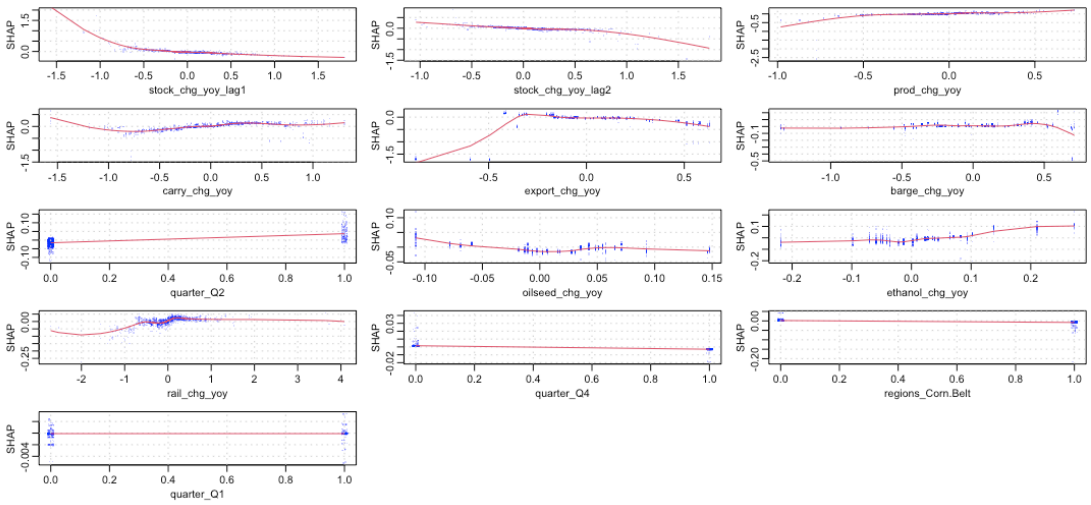


Figure 29: Partial Summary for Wheat & Quarter 1,2 and 4

production. The rest of the features contribute to the model but are lower than the three above. Now we look individually at these three features and how these are shaping the final prediction of estimating the stock of corn for the following quarters.

The plot (Figure 30) shows the SHAP values for the feature "stock_chg_yoy_lag1" in the model. The y-axis shows the SHAP value of the feature for each sample in the test set, while the x-axis shows the value of the feature itself. The color of each dot represents the value of another feature that is highly correlated with " stock_chg_yoy_lag1." The blue color indicates a low value, while the red color indicates a high value. The plot shows the relationship between the feature " stock_chg_yoy_lag1" and the output of the XGBoost model. The plot shows that higher values of " stock_chg_yoy_lag1" are associated with lower values in model output (vice-veresa), which suggests that increasing "stock_chg_yoy_lag1" could decrease the values of model output. When the change of

stock of wheat from previous year is increased by 1%, the wheat stock for the upcoming quarter will decrease by 1.9% approx. ceteris paribus.

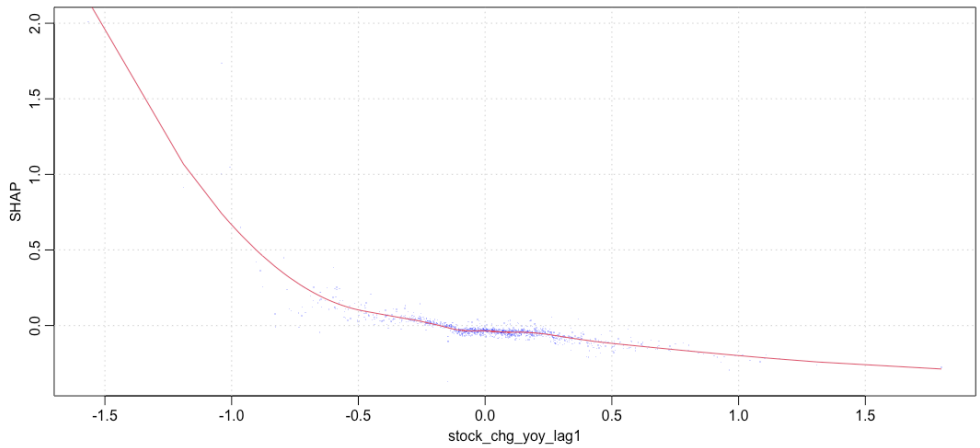


Figure 30 Partial of stock_chg_yoy_lag1 for Wheat & Q1Q2Q4

The plot (Figure 31) shows the relationship between the feature "stcok_chg_yoy_lag2" and the output of the XGBoost model. The plot shows that almost higher values of "stcok_chg_yoy_lag2" are associated with lower values in model output, which suggests that increasing "stcok_chg_yoy_lag2" could decrease the values in model output. When the year-over-year percentage change in the wheat stock from previous two years is increased by 1%, the wheat stock for the upcoming quarter will decrease by around 0.3% approx. ceteris paribus.

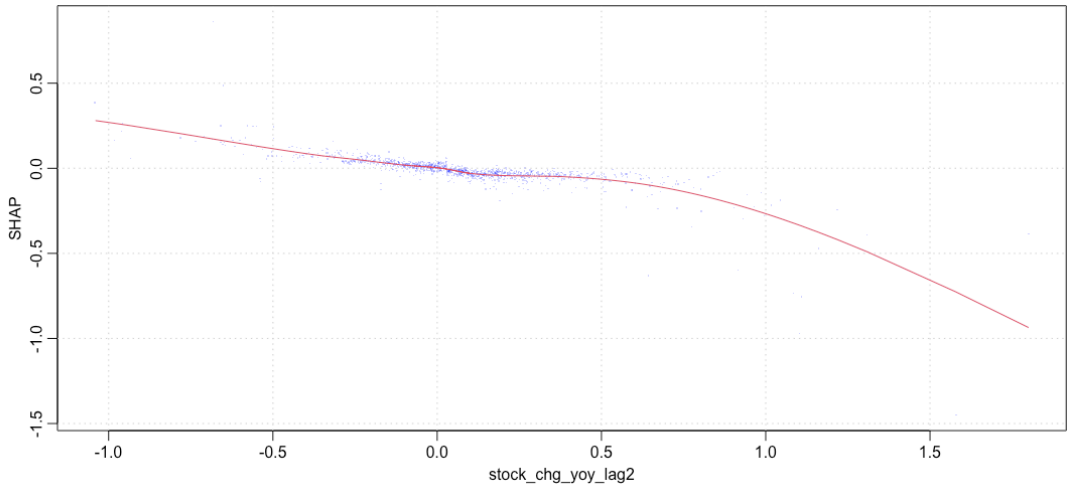
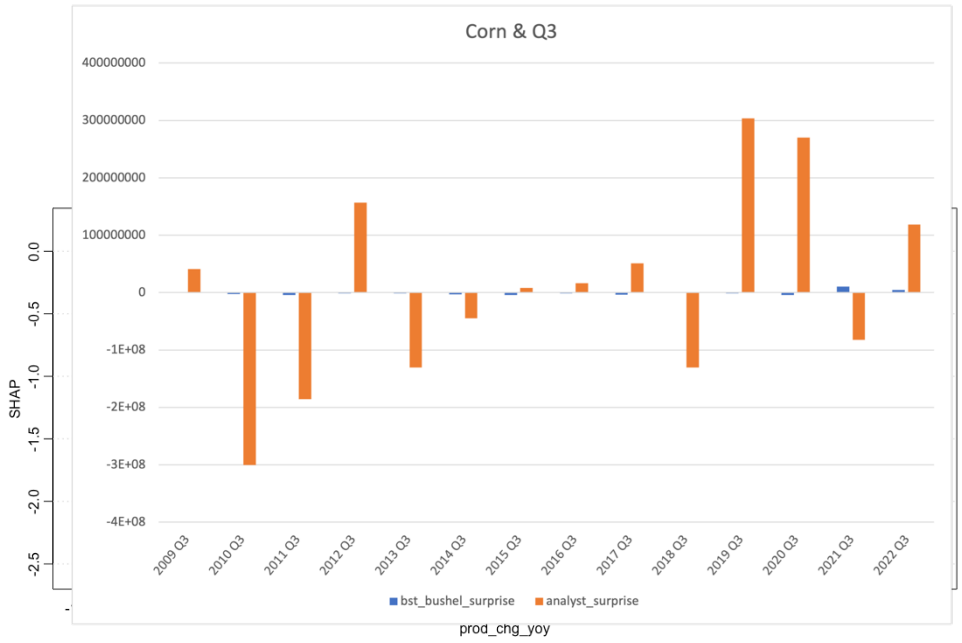


Figure 31: Partial of stock_chg_yoy_lag2 for Wheat & Q1Q2Q4

The plot (Figure 32) shows the relationship between the feature "prod_chg_yoy " and the output of the XGBoost model. The plot shows that higher values of "prod_chg_yoy " are associated with higher values in model output, which suggests that increasing "prod_chg_yoy " could increase the values in model output. When the year-over-year percentage change in the wheat production is increased by 1%, the wheat stock for the upcoming quarter will increase by around 0.8% approx. ceteris paribus for the model Q1, Q2 and Q4.

SUMMARY & DISCUSSIONS

The model we used for our analysis has worked great, as we have found some notable results. The XGBoost machine learning model has worked and accumulated all the possible data and dealt smartly with NA variables, making our model more exceptional as not all the sources have all the data from 2008. We will try to investigate each model's result that we performed in the empirical analysis. After running the model to predict the data for the grain stock by commodities and quarter, we created our dataset. We merged it with the analysts' dataset, taking account of each quarter's actual grain stock level for each commodity and trying to find the error term by subtracting the predicting values from our model and the analysts'. The error term from the model is the difference between the XGBoost prediction value and the actual grain stock in bushels by the USDA, which we named after "bst_bushel_surprise." Furthermore, the error term from the analysts is the



FigFigure 33: Comparison plot for Corn & Q3

difference between the average prediction by all the reports releasing analysts each quarter and the actual grain stock in bushel by USDA, which we named after "analyst_surprise."

Our first model takes account of the commodity corn and quarter 3. Figure 33 shows the comparison of the XGBoost model's surprise and the analysts' surprise. The blue represents the XGBoost, whereas the orange represents the analysts' surprise. We can also see that till 2020 the data was trained for the model, and the rest of the year was used for the out-sample prediction. From the figure, we can see that the machine learning algorithm that we used for this model is doing a pretty great job compared to the private market analysts who have been providing numbers prior to the publication of the USDA quarterly grain stock report. For example, In Quarter 3 of 2019, the analysts overestimated the market

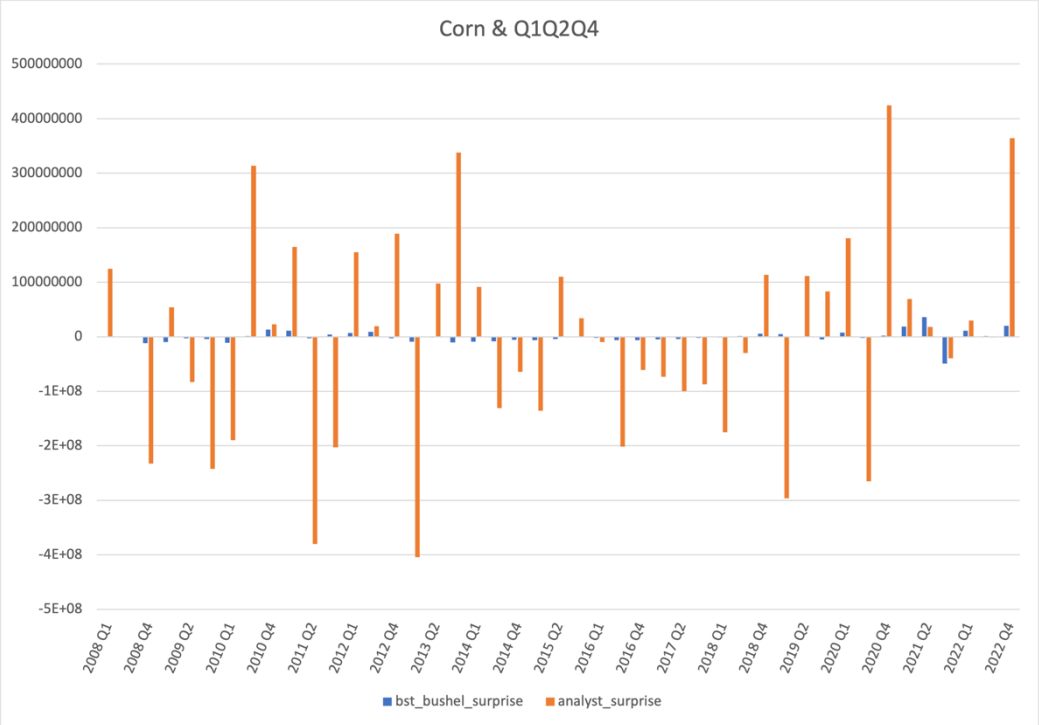


Figure 34: Comparison plot for Corn & Q1Q2Q4

and anticipated much higher than the actual corn stock. The surprise is around 300 million

bushels, whereas our model is really close to what USDA anticipated of corn stock for that quarter.

Figure 34 represents the commodity corn and quarter 1,2 and 4. This compares the XGBoost model's surprise and the analysts' surprise. The blue represents the XGBoost, whereas the orange represents the analysts' surprise. From the figure, we can see that the machine learning algorithm that we used for this model is doing a pretty great job compared to the private market analysts who have been providing numbers prior to the publication

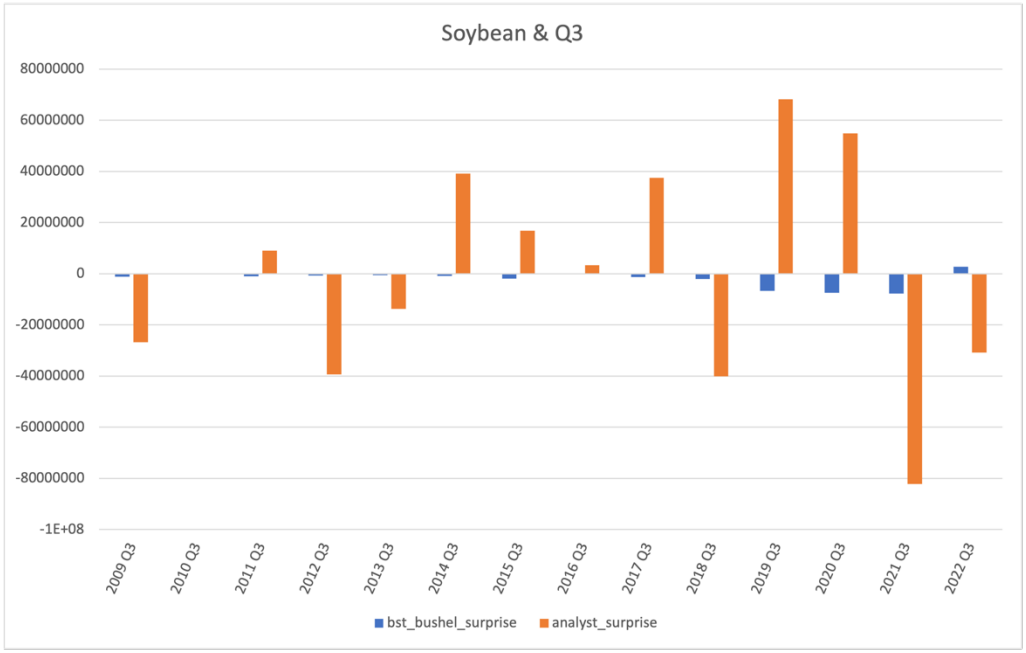


Figure 35: Comparison Plot for Soybean & Q3

of the USDA quarterly grain stock report. For example, In Quarter 4 of 2019, the analysts overestimated the market and anticipated much higher than the actual corn stock. The surprise is more than 400 million bushels, whereas our model is really close to what USDA anticipated of corn stock for that quarter.

Figure 35 represents the commodity soybean and quarter 3. This compares the XGBoost model's surprise and the analysts' surprise. The blue represents the XGBoost, whereas the orange represents the analysts' surprise. From the figure, we can see that the

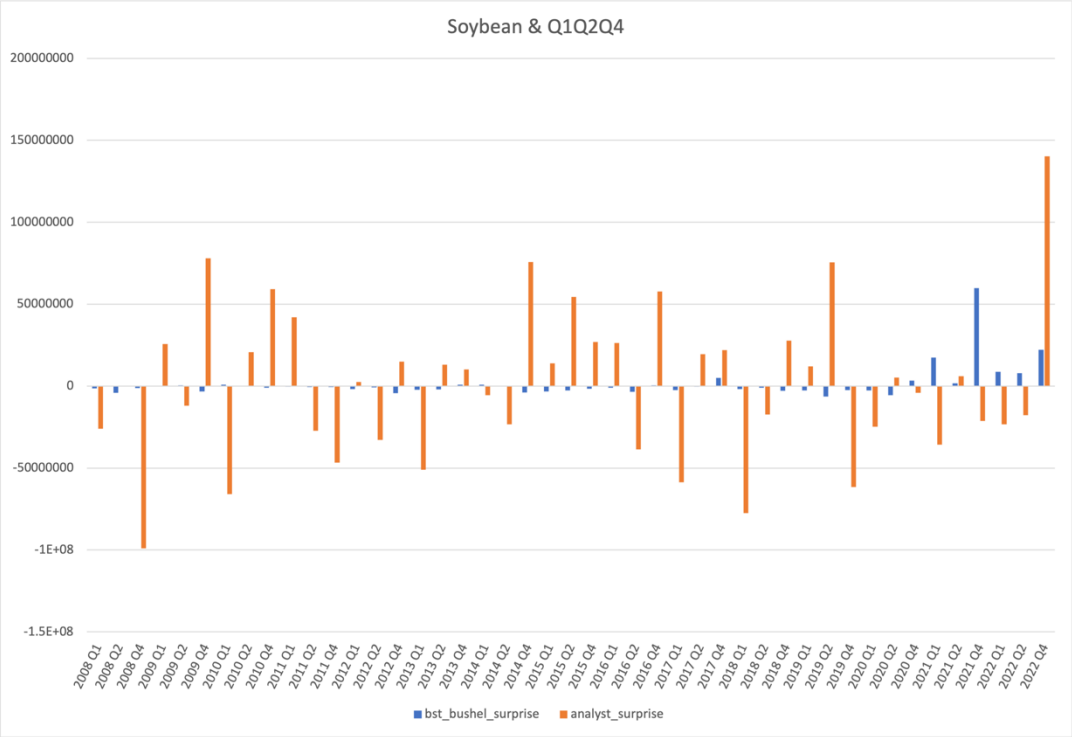


Figure 36: Comparison Plot for Soybean & Q1Q2Q4

machine learning algorithm that we used for this model is doing a pretty great job compared to the private market analysts who have been providing numbers prior to the publication of the USDA quarterly grain stock report. For example, In Quarter 3 of 2021, the analysts underestimated the market by a huge margin and anticipated much lower than the actual soybean stock. The surprise is more than 80 million bushels, whereas our model is really close to what USDA anticipated of soybean stock for that quarter.

The commodity soybean and quarter 1,2 and 4 is represented in Figure 36. This compares the XGBoost model's surprise and the analysts' surprise. The blue represents the XGBoost, whereas the orange represents the analysts' surprise. From the figure, we can see

that the machine learning algorithm that we used for this model is doing a pretty great job compared to the private market analysts who have been providing numbers prior to the publication of the USDA quarterly grain stock report. For example, In Quarter 4 of 2022,

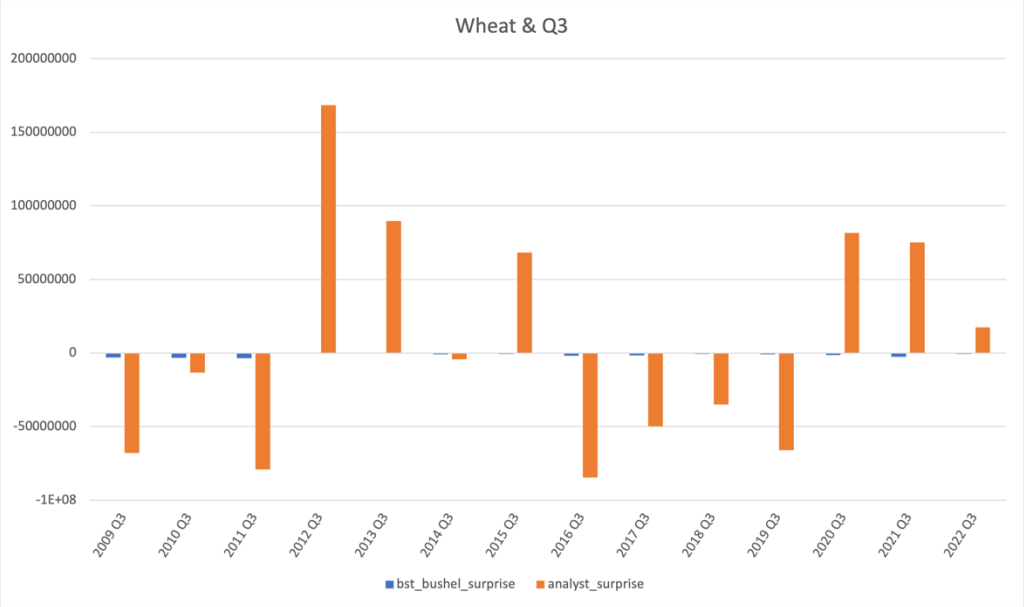


Figure 37: Comparison Plot for Wheat & Q3

the analysts overestimated the market by a huge margin and anticipated much lower than the actual soybean stock. The surprise is more than 100 million bushels, whereas our model is really close to what USDA anticipated of soybean stock for that quarter. However, our model missed one time in its out-sample period, where it did not work as well as it has been compared to the market analysts. In quarter 4 of 2021, the model overestimated the market by more than 50 million soybean bushels, whereas the analysts estimated close to the actual value.

Figure 37 represents the commodity wheat and quarter 3. This compares the XGBoost model's surprise and the analysts' surprise. The blue represents the XGBoost, whereas the orange represents the analysts' surprise. From the figure, we can see that the machine learning algorithm that we used for this model is doing a pretty great job compared

to the private market analysts who have been providing numbers before the publication of the USDA quarterly grain stock report. For example, In Quarter 3 of 2020, the analysts overestimated the market and anticipated much higher than the actual wheat stock. The surprise is more than 60 million bushels, whereas our model is really close to what USDA anticipated of wheat stock for that quarter.

The commodity wheat and quarter 1,2, and 4 are represented in Figure 38. This

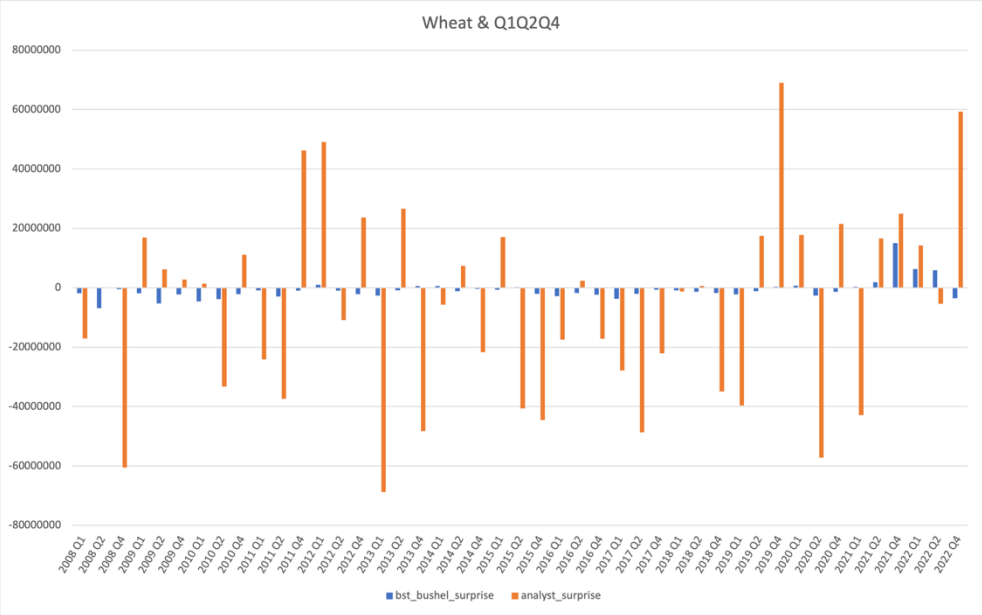


Figure 38: Comparison Plot for Wheat & Q1Q2Q4

compares the XGBoost model's surprise and the analysts' surprise. The blue represents the XGBoost, whereas the orange represents the analysts' surprise. From the figure, we can see that the machine learning algorithm that we used for this model is doing a pretty great job compared to the private market analysts who have been providing numbers before the publication of the USDA quarterly grain stock report. For example, In Quarter 4 of 2019, the analysts overestimated the market by a huge margin and anticipated much lower than the actual wheat stock. The surprise is more than 60 million bushels, whereas our model is really close to what USDA anticipated of wheat stock for that quarter.

After investigating our model's results and comparing them with the current market analysts' information, it is evident that artificial algorithms will not be surprised as the private market analysts. We then perform the model accuracy measurement to understand and test the errors of our model by commodities and quarter so that we can understand this much at least the prediction is working in the right direction. Table 5 shows the Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared values of each model performing the prediction and comparison with the private market analysts. The MSE measures the average squared difference between a model's predicted and actual values. It gives an idea of how well the model fits the data. A lower MSE value indicates a better fit of the model to the data. All the model we used in this paper has an MSE value of less than 6% approx. The lowest MSE is 1.97% for the model with commodity corn in quarters 1,2 and 4, and the highest MSE is 5.99% for the model with commodity soybean in quarter 3. In all cases, the MSE values exhibit that the models are performing well.

Table 5: Models' Performance Level

Commodity	Quarter	MSE	MAE	RMSE	R-squared
Corn	Quarter 3	0.04363	0.10483	0.20888	0.83252
	Quarter 1,2 & 4	0.01973	0.09630	0.14046	0.73201
Soybean	Quarter 3	0.05992	0.07742	0.24478	0.89663
	Quarter 1,2 $ 4	0.07155	0.13904	0.26748	0.66851
Wheat	Quarter 3	0.01308	0.07635	0.11438	0.93079
	Quarter 1,2 & 4	0.04370	0.13988	0.20905	0.92032

The MAE measures the average absolute difference between a model's predicted and actual values. It shows how far off the predictions are from the actual values. A lower MAE value indicates a better fit of the model to the data. All the model we used in this paper has an MAE value of less than 14% approx. The lowest MAE is 7.6% for the model with commodity wheat in quarter 3, and the highest MAE is 13.9% for the model with commodity wheat in quarters 1,2 and 4. In all cases, the lower MAE values exhibit that the models are performing well. The RMSE measures the square root of the average squared difference between the predicted and actual values in a model. It is similar to the MSE but is in the same units as the data. A lower RMSE value indicates a better fit of the model to the data. All the model we used in this paper has an RMSE value of less than 27% approx. The lowest RMSE is 11.4% for the model with commodity wheat in quarter 3, and the highest RMSE is 13.9% for the model with commodity Soybean in quarters 1,2 and 4. In all cases, the lower RMSE values exhibit that the models are performing well. R-squared is a measure of how well the model fits the data. It ranges from 0 to 1, with a higher value indicating a better fit. An R-squared value of 1 indicates a perfect fit, while a value of 0 indicates that the model does not explain any of the variability in the data. All the model we used in this paper has an R-squared value of greater than 67% approx. The lowest R-squared is 66.8% for the model with commodity soybean in quarters 1,2, and 4, and the highest R-squared is 93.07% for the model with commodity wheat in quarter 3. In all cases, the greater R-squared values exhibit that the models are performing well as most of the models perfectly fit, reflecting the r-squared values from Table 5.

CONCLUSION

From the above-detailed analysis, comparison, and performance check of each model, there is evidence that the machine learning algorithm performs better than the existing private market analysts. The partial contributions of each feature aid in predicting better the actual number of grain stock available in the marketing year. We can deduce from the theoretical framework that the agricultural commodities market exhibits weak market efficiency. Using proper machine learning or artificial intelligence, one can trace the market sentiment and gain more than mass participants using reports and public information. The considerable misses in the prediction by the private market analysts of the quarterly grain stock can be accounted for by not including the grain in transit, differences in calculation unit of grains, and not accounting correctly for the leftover grain from the previous marketing year. Furthermore, due to the ignorance of survey noise in the USDA quarterly reports survey. By checking all these measures and running all these into a systematic machine learning model, we can beat the private market analysts. Therefore, Artificial intelligence will not be as surprised as the private market analysts by USDA quarterly grain stock reports.

LITERATURE CITED

1. Adjemian, M. K. (2012). Quantifying the WASDE Announcement Effect. American Journal of Agricultural Economics, 94(1), 238–256. https://doi.org/10.1093/ajae/aar131
2. Allen, P. Geoffrey. (1994). Economic forecasting in agriculture. International Journal of Forecasting, 10(1), 81–135. https://doi.org/10.1016/0169-2070(94)90052-3
3. BAUR, R. F., & ORAZEM, P. F. (1994). The Rationality and Price Effects of U.S. Department of Agriculture Forecasts of Oranges. The Journal of Finance, 49(2), 681–695. https://doi.org/10.1111/j.1540-6261.1994.tb05157.x
4. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828.
5. Botto, A. C., Isengildina, O., Irwin, S. H., & Good, D. L. (2006). Accuracy trends and sources of forecast errors in WASDE balance sheet categories for corn and soybeans (No. 379-2016-21949).
6. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
7. Brownlee, J. (2018). XGBoost With Python Mini-Course. Retrieved from https://machinelearningmastery.com/xgboost-with-python-mini-course/
8. Brownlee, J. (2019). XGBoost for Machine Learning. Machine Learning Mastery.
9. Chen, T. (2018). XGBoost Documentation. Retrieved from https://xgboost.readthedocs.io/en/latest/

10. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
11. Chen, T., Li, H., Yang, Q., & Yu, Y. (2013, February). General functional matrix factorization using gradient boosting. In International Conference on Machine Learning (pp. 436-444). PMLR.
12. Chen, T., Singh, S., Taskar, B., & Guestrin, C. (2015, February). Efficient second-order gradient boosting for conditional random fields. In Artificial Intelligence and Statistics (pp. 147-155). PMLR.
13. Colling, P. L., & Irwin, S. H. (1990). The Reaction of Live Hog Futures Prices to USDA Hogs and Pigs Reports. American Journal of Agricultural Economics, 72(1), 84–94. https://doi.org/10.2307/1243147
14. Dixon, B. L., Hollinger, S. E., Garcia, P., & Tirupattur, V. (1994). Estimating corn yield response models to predict impacts of climate change. Journal of Agricultural and resource economics, 58-68.
15. Egelkraut, T. M., Garcia, P., Irwin, S. H., & Good, D. L. (2003). An Evaluation of Crop Forecast Accuracy for Corn and Soybeans: USDA and Private Information Agencies. Journal of Agricultural and Applied Economics, 35(1), 79–95. https://doi.org/10.1017/s1074070800005952
16. Fackler, P. L., & Norwood, B. (1999). Forecasting crop yields and condition indices.
17. Falk, B., & Orazem, P. F. (1985). A Theory of Future's Market Responses to Government Crop Forecasts (No. 138-2021-2867)

18. Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The journal of Finance, 25(2), 383-417.
19. Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. the Journal of machine Learning research, 9, 1871-1874.
20. Fernandez-Perez, A., Frijns, B., Indria
21. Fortenbery, T. R., & Sumner, D. A. (1990). The effects of USDA reports in futures and options markets.
22. Fortenbery, T. R., & Sumner, D. A. (1993). The effects of USDA reports in futures and options markets. Journal of Futures Markets, 13(2), 157–173. https://doi.org/10.1002/fut.3990130204
23. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189-1232.
24. Garcia, P., Irwin, S. H., Leuthold, R. M., & Yang, L. (1997). The value of public information in commodity futures markets. Journal of Economic Behavior & Organization, 32(4), 559–570. https://doi.org/10.1016/s0167-2681(97)00013-9
25. Good, D. L., & Irwin, S. H. (2006). Understanding USDA corn and soybean production forecasts: Methods, performance and market impacts over 1970-2005.
26. Good, D., & Irwin, S. (2014). Accuracy of USDA Forecasts of Corn Ending Stocks. farmdoc daily, 4(94).
27. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

28. Grunewald, O., McNulty, M. S., & Biere, A. W. (1993). Live Cattle Futures Response to Cattle on Feed Reports. American Journal of Agricultural Economics, 75(1), 131–137. https://doi.org/10.2307/1242961
29. Gunnelson, G., Dobson, W. D., & Pamperin, S. (1972). Analysis of the accuracy of USDA crop forecasts. American Journal of Agricultural Economics, 54(4_Part_1), 639-645.
30. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.
31. Irwin, S. H., Gerlow, M. E., & Liu, T.-R. (1994). The forecasting performance of livestock futures prices: A comparison to USDA expert predictions. Journal of Futures Markets, 14(7), 861–875. https://doi.org/10.1002/fut.3990140707
32. Irwin, S. H., Good, D. L., & Tannura, M. (2009). 2009 Final Corn and Soybean Yield Forecasts (No. 1633-2016-135084).
33. Isengildina-Massa, O., Irwin, S. H., Good, D. L., & Gomez, J. K. (2008). The Impact of Situation and Outlook Information in Corn and Soybean Futures Markets: Evidence from WASDE Reports. Journal of Agricultural and Applied Economics, 40(1), 89–103. https://doi.org/10.1017/s1074070800023488
34. Isengildina-Massa, O., Karali, B., & Irwin, S. H. (2013). When do the USDA forecasters make mistakes?. Applied Economics, 45(36), 5086-5103.
35. Isengildina, O., Irwin, S. H., & Good, D. L. (2004). Evaluation of USDA Interval Forecasts of Corn and Soybean Prices. American Journal of Agricultural Economics, 86(4), 990–1004. https://doi.org/10.1111/j.0002-9092.2004.00648.x

36. Isengildina, O., Irwin, S. H., & Good, D. L. (2006). Are Revisions to USDA Crop Production Forecasts Smoothed? American Journal of Agricultural Economics, 88(4), 1091–1104. https://doi.org/10.1111/j.1467-8276.2006.00918.x
37. Just, R. E. (1983). The impact of less data on the agricultural economy and society. American Journal of Agricultural Economics, 65(5), 872-881.
38. Karali, B. (2012). Do USDA announcements affect comovements across commodity futures returns?. Journal of Agricultural and Resource Economics, 77-97.
39. Karali, B., Isengildina-Massa, O., Irwin, S. H., Adjemian, M. K., & Johansson, R. (2019). Are USDA reports still news to changing crop markets?. Food Policy, 84, 66-76.
40. Kruse, J., & Smith, D. (1994). Yield estimation throughout the growing season.
41. Lehecka, G. V. (2014). The value of USDA crop progress and condition information: Reactions of corn and soybean futures markets. Journal of Agricultural and Resource Economics, 88-105.
42. Liu, Q., & Lu, J. (2019). Anomaly detection in image processing based on XGBoost. In 2019 IEEE International Conference on Mechatronics and Automation (ICMA) (pp. 851-855). IEEE.
43. Marone, H. (2008). How Do Wheat Prices React to USDA Reports?. UNDP/ODS working paper, United Nations Development Programme, Office of Development Studies, New York.

44. McKenzie, A. M. (2008). Pre-harvest price expectations for corn: The information content of USDA reports and new crop futures. American Journal of Agricultural Economics, 90(2), 351-366.
45. McNew, K. P., & Espinosa, J. A. (1994). The informational content of USDA crop reports: Impacts on uncertainty and expectations in grain futures markets. The Journal of Futures Markets (1986-1998), 14(4), 475.
46. McNichols, M., & Trueman, B. (1994). Public disclosure, private information collection, and short-term trading. Journal of Accounting and Economics, 17(1-2), 69-94.
47. Milonas, N. T. (1987). The effects of USDA crop announcements on commodity prices. The Journal of Futures Markets (1986-1998), 7(5), 571.
48. Salin, V., Thurow, A. P., Smith, K. R., & Elmer, N. (1998). Exploring the market for agricultural economics information: Views of private sector analysts. Applied Economic Perspectives and Policy, 20(1), 114-124.
49. Sanders, D. R., & Manfredo, M. R. (2002). USDA production forecasts for pork, beef, and broilers: an evaluation. Journal of Agricultural and Resource Economics, 114-127.
50. Stigler, G. J. (1966). Theory of price.
51. Summer, D. A., & Mueller, R. A. (1989). Are harvest forecasts news? USDA announcements and futures market reactions. American Journal of Agricultural Economics, 71(1), 1-8.

52. Svensson, J., & Yanagizawa, D. (2009). Getting prices right: the impact of the market information service in Uganda. Journal of the European Economic Association, 7(2-3), 435-445.

53. Tianqi, C. (2015). Scalable Machine Learning with XGBoost. Presented at the New York R Conference.

54. Von Bailey, D., & Brorsen, B. W. (1998). Trends in the accuracy of USDA production forecasts for beef and pork. Journal of Agricultural and Resource Economics, 515-525.

55. wan, I., & Tourani-Rad, A. (2018). Surprise and dispersion: informational impact of USDA announcements. Agricultural Economics, 50(1), 113–126. https://doi.org/10.1111/agec.12470

56. Xiao, J., Hart, C. E., & Lence, S. H. (2017). USDA Forecasts Of Crop Ending Stocks: How Well Have They Performed?. Applied Economic Perspectives and Policy, 39(2), 220-241.

57. Zhang, T., & Xu, W. (2018). Distributed XGBoost: A scalable machine learning system for big data. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 4113-4120). IEEE.

58. Zhou, L., Chen, H., & Yan, Z. (2019). A hybrid XGBoost-based recommendation system for smart city applications. Journal of Systems Architecture, 95, 1-11.