Soybean Price Prediction

Alvian Jonathan Sutrisno, Kirsten Sydney Leong, Louis Evantio Hartono, Vickeisha Lall

{asutrisno3, kleong30, lhartono7, vlall7}@gatech.edu

1. INTRODUCTION

1.1. Background

Soybeans are an important and versatile crop to the United States. The crop is pivotal to the agricultural and food industries as its extracted oils and remaining proteins can be used in food products and animal feeds. Soybeans are economically beneficial to the United States as the country is one of the biggest soybean exporters in the world. The soybean sector averages $115.8 billion per year in total economic impact and $11.6 billion in total wage impact within the country [1]. Given the economic importance of the crop to United States and some of its states, price forecasting for soybeans is very beneficial to help farmers determine when to sell their crops and at what price to mitigate risks and increase revenue.

However, soybean price prediction is difficult as crop price volatility is high [2]. High food price volatility can lower food security for consumers and can also affect producers, especially local farmers, who often lack the adequate tools to manage risk. There have been an increasing number of factors affecting the soybean markets. One of them being world events like the African Swine Fever Outbreak in China that has significantly impacted the soybean markets, as China is a major soybean importer. In addition, factors from the development of soybean bean biotech products in Turkey [3] to weather conditions [4] to trade policies [5] have also affected the soybean industry in recent years. With all these variables affecting the market and the high crop price volatility, there is a huge need to provide a comprehensive view of key factors that influence soybean price. This project may be utilized to better predict the soybean price.

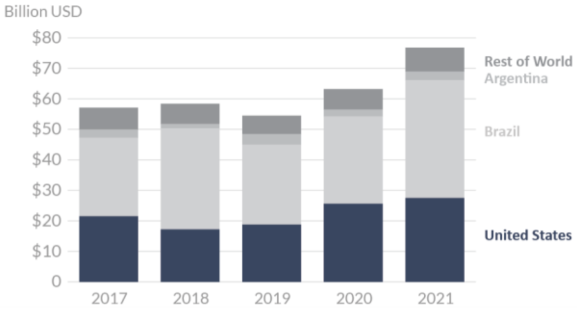


Fig 1. Global Soybean Exports for 2017-2021

1.2. General Approach and Motivation

For this approach, we compiled four datasets of potential predictors for soybean prices based on our literature readings: temperature, global soybean imports and exports, oil and gas prices, United States Consumer Price Index (CPI) and compiled it with a soybean price dataset. We are using five different machine learning algorithms to create time series forecasting models to predict future soybean prices. The models we are exploring are linear regression, lasso regression, ridge regression, elastic net, and long short term-memory models.

With our final model, we strive to answer three research questions:

1. Which factors have the highest influence on the price of soybean?
2. Are crop prices, particularly soybean, independent of inflation?
3. Can a high-performing predictive model be built based on influential factors?

2. METHODS

2.1. Key Variables

Dependent: Price of soybean commodity futures (USD/bushel)

Independent*: (1). Brent Oil stock price, (2). Crude Oil stock price, (3). Natural Gas stock price, (4). Heating Oil stock price, (5). US CPI, (6). US Soybean Production, (7). US Soybean Imports, (8). US Soybean Exports, (9). Argentina Soybean Production, (10). Argentina Soybean Imports, (11). Argentina Soybean Exports, (12). Brazil Soybean Production, (13). Brazil Soybean Imports, (14). Brazil Soybean Exports, (15). China Soybean Production, (16). China Soybean Imports, (17). China Soybean Exports, daily average temperature data of top soybean producing counties ((18). Rapid City, (19). Springfield, (20). Peoria, (21). Sioux City, (22). Little Rock, (23). Jackson, and (24). Fargo)

*****All independent variables are in its lag 1 form to mimic real-world situations in which independent variables at time t are used to predict a dependent variable at time t+1.

2.2. Initial Hypothesis

1. Soybean prices are highly impacted by import and export values of major producing and importing countries.
2. Soybean prices are influenced by other major commodities such as Crude oils and natural gas.
3. Soybean prices are not closely related to the inflation rate.
4. Soybean prices can be predicted by using important factors identified above

3. DATA CLEANING PROCESS

3.1. Datasets

3.1.1 Soybean Price Dataset

The soybean price dataset is the main dataset as it is the focus of this analysis. The dataset is recorded daily and ranges from the year 2000 to 2021. Upon closer inspection, there are 2391 missing values out of 7573 rows. We identify that missing values are due to weekends and holidays at which the information of the price was not available.

Since missing values account for more than 5%, the team decides to perform data imputation. There are several methods that can be used to fill missing values such as using last observed values, linear interpolation, and ARIMA interpolation. The team selects ARIMA interpolation with random walk as it best replicates the uncertainty elements in price change.

3.1.2 Oil and Gas Dataset

We used a Kaggle dataset of daily stock prices for Brent Oil, Crude Oil WTI, natural gas, and heating oil from 2000 to 2022 as these commodities are used throughout the farming process of soybeans. We decided to use only the closing price, so data manipulation was performed so that the cleaned data set had a total of five columns: a date column and four columns of the closing price of the four commodities for each day. Similar to the soybean data set, this data set lacked weekend data as the stock market is not open for trading on these days. Therefore, we used ARIMA interpolation to determine the closing prices of the oil and gases on weekends.

3.1.3 Production, Import & Export Dataset

As shown in Fig. 1, the United States, Brazil, Argentina, and China are the four major countries involved in the soybean market. We used a dataset from the World Agricultural Supply and Demand Estimates (WASDE) report that was made publicly available through Kaggle.

This dataset contains beginning and end stocks, production, import, domestic crush, domestic total, and export data for the world and individual countries. However, we filtered the data to only include the production, import, and export data for the US, Argentina, Brazil, and China, since these are the data points our team is interested in analyzing. Lastly, we performed ARIMA interpolation on this dataset due to some records having 0 values.

3.1.4 Temperature Dataset

We used city temperature data published by the University of Dayton as crop output is affected by temperature. We researched the top 15 soybean-producing counties in the United States and mapped them to the closest city dataset provided in the university's repository. For example, the highest producing county is McLean, Illinois which is located just 45 miles from Peoria, Illinois, one of the cities with recorded daily temperatures. Thus, we selected the daily average temperatures of 7 different cities (Peoria and Springfield, Illinois; Jackson, Mississippi; Fargo, North Dakota; Rapid City, South Dakota, Sioux City, Iowa and Little Rock, Arkansas) and merged it with our other datasets on the date.

3.1.5 US Consumer Price Index (CPI) Dataset

We also used a dataset containing monthly US CPI, which is the average CPI over all US cities. This is a measure of US inflation and contained values from 1913 to 2021. No missing values were found in this dataset.

3.2. Initial Discovery

We carried out seasonal decomposition fitting into soybean price. The intent is to not only understand any underlying trend and/or seasonal pattern, but also to identify outliers. Visually, it can be seen in Fig 4. that the historical soybean price follows a trend and yearly seasonal pattern. However, there is no apparent pattern in the weekly seasonal component.

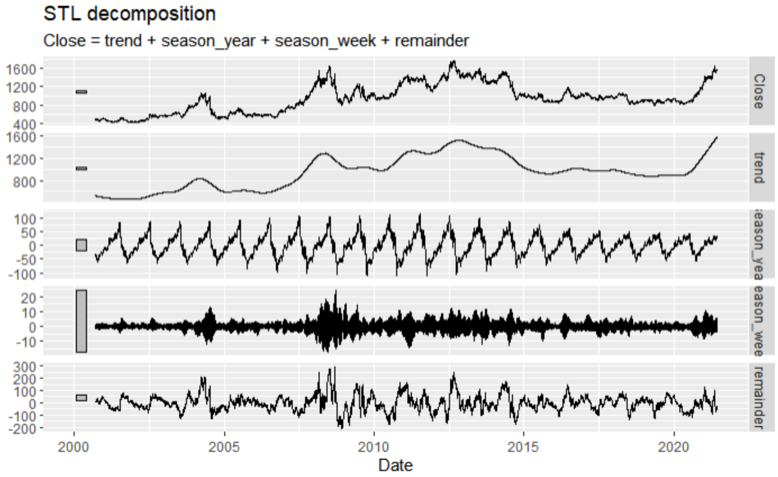


Fig 4. Soybean Price Decomposition

Using the remainder component from the previous analysis, we carry out outlier detection. Outliers are defined as data points that are 3 interquartile ranges (IQRs) from the central 50% of the data. We identify that outliers are present in the dataset and are mostly from the year 2008. We find that it is due to the 2007-2008 world food price crisis and in fact, are natural/true outliers. Hence, no outlier treatment is performed, and we proceed to the next step.

3.3 Joining the Datasets

After cleaning the datasets, our subsequent task was to merge the datasets together. Our response variable, Soybean Price (or soybean close), consists of *daily* observations for the period 15/09/2000 to 09/06/2021. However, two of our predictor variable datasets, namely the Import/Export/Production and CPI datasets, consisted of *monthly* observations for their respective periods. We decided to deal with this by repeating the monthly observations for each day of a given month. The other datasets were the Oil and Gas dataset and the temperature data for the seven counties mentioned above, all of which consisted of daily observations. Once all our datasets had daily values, we performed sequential left joins to the soybean price data (our response variable) on the date column. Prior to doing this, we confirmed that the soybean price data contained all the dates in the time period. The import/export data contained values for the shortest period (2007 to 2020), so the rows before and after this period were removed, and only rows with observations for *all variables* kept. It was then ensured that the sequence of dates in the date column contained all the actual dates for the period (May 2007 – February 2020), and no missing or repeated dates were present. This was done by generating the date sequence for the May 2007 – February 2020 period and comparing it to the date column to ensure they were identical.

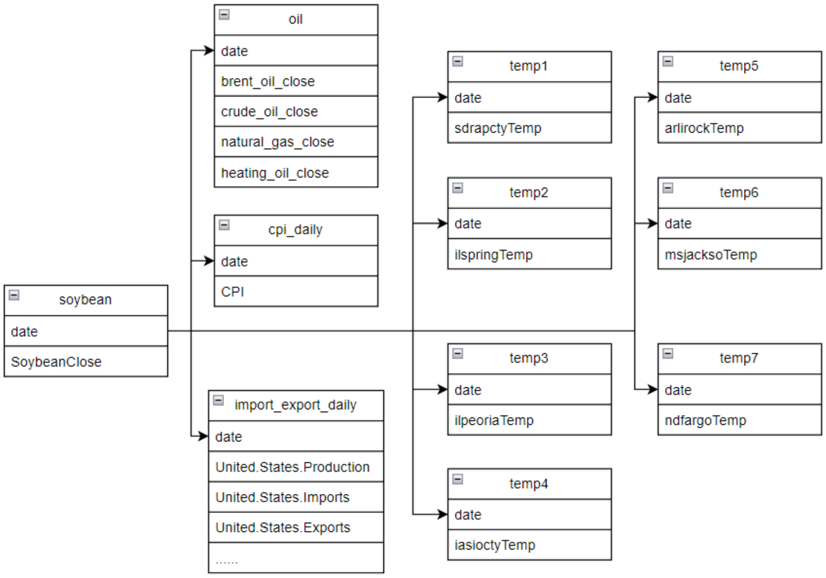


Fig 5. Dataset Relationships

3.4 Relationships Between Variables

Once our datasets were merged, we were interested in the relationship between our response variable, soybean price, and the predictors. We calculated both the Pearson's and Spearman's correlation coefficients for soybean price and each of the predictors to understand any linear (from Pearson's correlation) or overall monotonic relationship (Spearman's correlation). As expected, we found strong positive correlations between soybean price and Brent crude oil, heating oil and crude oil prices, with the Pearson's coefficients being 0.82, 0.80 and 0.78 respectively. The strongest negative relationships were noted between soybean price and Argentina imports and US Production, with Spearman's coefficients of -0.59 and -0.47 respectively. There appeared to be almost no relationship between soybean price and temperature data for the selected counties.

3.5 Principal Component Analysis (PCA)

Principal Component Analysis was also performed to give a snapshot of the potential relationships between the predictors and the response. The following plot shows the variable loadings from the PCA (first 2 Principal Components only (PCs), which explain 56.9% of the variation). This essentially illustrates how much weight each variable has on the PCs shown and can give an indication of the relationships between the variables, based on the position of the variable vectors. The loadings plot below suggests that there may be positive relationships between soybean price (SoybeanClose) and Oil prices (variable vectors close together, small angles between them) and some negative relationships between soybean price and some import/export/production variables (axes opposite directions), with the exception of China Exports. For the most part, these preliminary relationships are what we expected, with higher supply leading to lower demand and consequently lower prices, as well as the strong positive relationships between oil and soybean price. The temperature variables and soybean price appear to have little relationship, as the vectors are roughly at 90° to each other. However, only the first 2 PCs were examined, which explain just 57% of the variation – the models discussed below will give a clearer picture of the relationships that exist from the data.

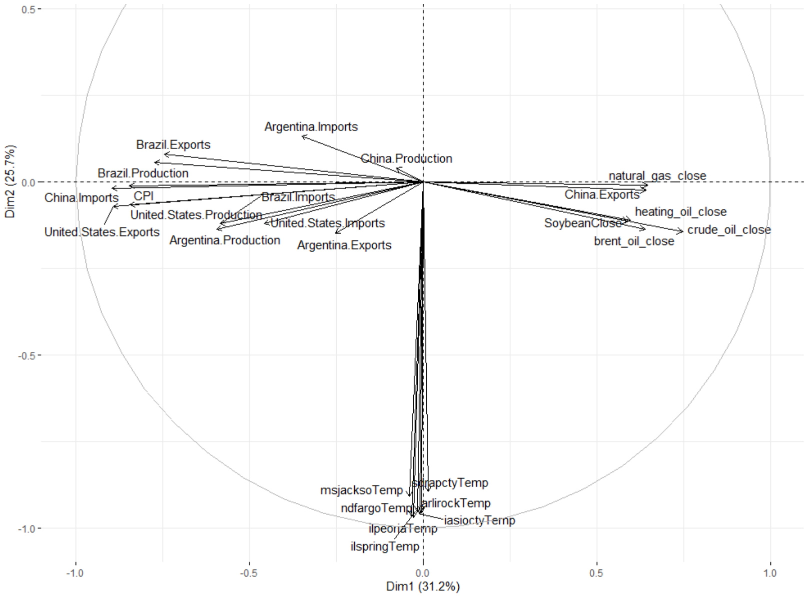


Fig 6. Loadings Plot from PCA

3.6 Dataset splitting

As we have 4650 data points, we decide to split the joined dataset into two: 1. Training and Validation dataset; 2. Testing dataset. When assigning data to the two sets, we decided to forego traditional cross-validation because our data is time-series. We had identified that there is a yearly seasonal pattern in the soybean prices that we wanted to preserve in our models. Therefore, rather than performing random assignment, we designated the last 396 (Jan-2019 to Feb-2020) observations as the testing dataset because the team believes that the number is enough to represent the yearly seasonal pattern. With that said, earlier observations will be used to train the model and carry out evaluation on a rolling forecasting origin which is widely used in Time Series Cross Validation. This method is chosen to preserve the serial correlation of time series data in which the traditional cross validation techniques fail to achieve.

4. RESULTS AND DISCUSSION

4.1 Model Evaluation

4.1.1 Linear Regression

For our basic model, we decided to use the glm() function in R to create a generalized linear model. We discovered that only Crude Oil Prices, United States Exports, and Sioux City Temperature are not statistically significant at the significance level of 0.05.

After creating the model, our next task is to perform a cross-validation to assess the performance of our model. We utilized the cv.glm() function from the boot library in R, to calculate the cross-validation estimate of prediction error for a generalized linear model. From one of the outputs, delta, we obtained a mean squared error (MSE) of 11045.70. Performing a square root of this value, we get 105.1 as our root mean squared error (RMSE)

In addition, our team also wanted to use the R-squared value to compare the five models we have. However, unlike the lm() function, the glm() function does not produce an R-squared value in its output, despite using the Gaussian family. Thus, as a workaround, we have decided to calculate McFadden’s R-Squared instead and obtained an Adjusted R-squared value of 0.82.

4.1.2 Lasso

Since we have several independent variables, a variable selection model such as LASSO could be beneficial by simplifying our final model, making it easier for us to determine which variables are more influential on soybean prices. LASSO regression performs variable selection by pushing the coefficients of the less-significant predictors close to or equal to 0, thereby eliminating unimportant features and preventing overfitting.

In R, we used the cv.glmnet function. The function standardized the features were standardized to have a mean of 0 and standard deviation of 1. We also set parameters so that the function would find the lambda that would minimize the mean square error. This was important because the model's features exist on vastly different scales, and we wanted to ensure that the larger-value features didn't primarily influence the model.

The LASSO Regression model did not actually simplify the model much compared to the linear regression model as it only eliminated China Exports from its model. It did set the coefficients of city temperatures close to zero, reducing their effect on the model indicating that they are not that important to the price predictions. The variables with the highest coefficient values are heating_oil_close, United States Imports, and Brazil Imports. The model achieved an RMSE of 93.64 and an Adjusted R-squared value of 0.83 which is an improvement on the standard linear regression model.

4.1.3 Ridge

Ridge regression is similar to the Lasso model as it will shrink down the coefficients of the predictors towards 0, but it will not eliminate any predictors from the model. Ridge regression has a unique benefit as it penalizes models for having a large sum of squared value of the predictor coefficients. Therefore, the coefficient values will be smaller than Lasso and should be more evenly distributed which reduces the risk of overfitting to the training data set. This model is also particularly helpful in identifying if any of the features are highly correlated and will reduce their impact on the model.

The cv.glment function was used to perform ridge regression in R with 'mse' as the measure to determine the best lambda value and the features were standardized. The function created a model using all predictors. Like the LASSO model, it pushed the coefficients of the temperature variables close to 0, reducing their weight on the model. The predictors with the largest coefficient values are the United States Imports, Brazil Imports, and heating oil close, like the LASSO model but the weight of the coefficients is not as large. The model achieved an RMSE of 103.84 and an Adjusted R-squared value of 0.78 which is a slight RMSE improvement on the linear regression model but not the LASSO model.

4.1.4 Elastic Net

We also opted to use an Elastic Net regression model for regularization. Elastic Net essentially combines the strengths of the Lasso and Ridge methods described above. It works well in instances where there are noise variables that can be removed from the model (like Lasso), as well as in cases where there are mostly useful variables that can be retained in the model (such as in Ridge, which shrinks coefficients, but not to zero).

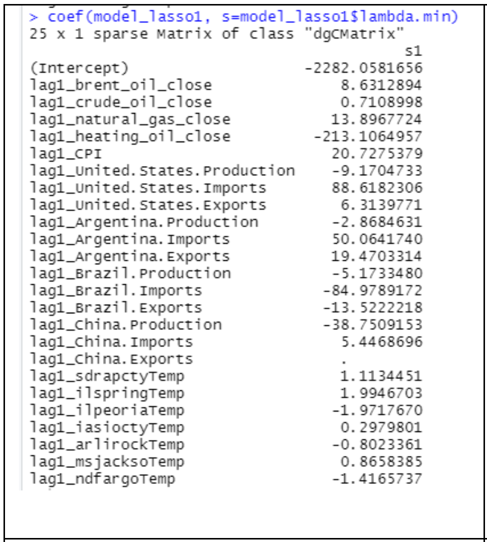
The optimal values of two hyperparameters had to be determined for the Elastic Net model:

- Alpha – It may be useful to view this as the degree of combining the Lasso and Ridge. Alpha can take any value between 0 and 1. The closer the value is to 0, the closer it is to a Ridge regression, and the closer the value is to 1, the closer it is to the Lasso regression.
- Lambda – A penalty coefficient/amount of penalization.

Determining the optimal values for alpha and lambda was done using 5-fold cross validation on different combinations of alpha/lambda, and the values which yielded the lowest Root Mean Square Error (RMSE) were chosen. An alpha of 0.1 (close to Ridge) and a lambda of 2.33 were found to produce the smallest RMSE and were used in the Elastic Net model. This cross validation and model training was done using the train() and trainControl() functions from the caret package in R. The data was centered and scaled (also done for the Lasso and Ridge models) using the 'preProc' option in trainControl().

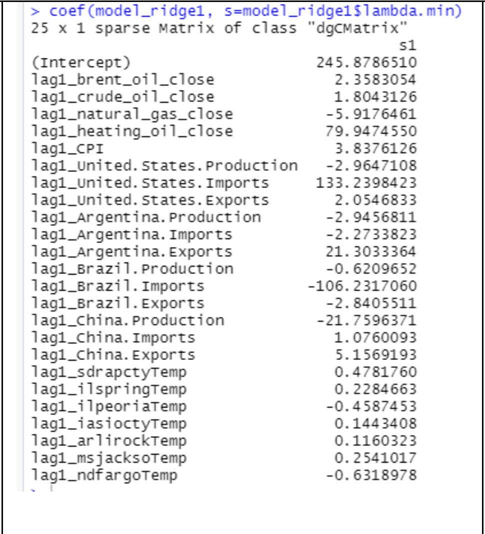
Thirty unique combinations of alpha and lambda were tried when determining the optimal values (‘tuneLength’ option in train() was set to 30).

As can be seen from the output, the model removed the variables China Exports and Little Rock Temperature (coefficients are zero), and the magnitude of all Temperature coefficients are below 1. The largest coefficients from this model are noted to be from United States Imports (positive relationship) and Brazil Imports (negative relationship) - these have the biggest effect on our response variable, Soybean Price. The final model achieved an RMSE of 100.97 and an Adjusted R-squared of 0.80. From the RMSE value, the Elastic Net model outperformed the Ridge and linear regression model, but not the Lasso.



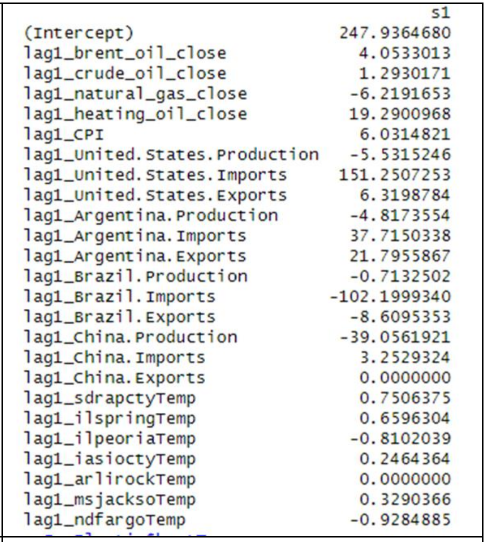
```
> coef(model_lasso1, s=model_lasso1$lambda.min)
25 x 1 sparse Matrix of class "dgCMatrix"
                                          s1
(Intercept)                    -2282.0581656
lag1_brent_oil_close               8.6312894
lag1_crude_oil_close               0.7108998
lag1_natural_gas_close            13.8967724
lag1_heating_oil_close          -213.1064957
lag1_CPI                          20.7275379
lag1_United.States.Production     -9.1704733
lag1_United.States.Imports        88.6182306
lag1_United.States.Exports         6.3139771
lag1_Argentina.Production         -2.8684631
lag1_Argentina.Imports            50.0641740
lag1_Argentina.Exports            19.4703314
lag1_Brazil.Production            -5.1733480
lag1_Brazil.Imports              -84.9789172
lag1_Brazil.Exports              -13.5222218
lag1_China.Production            -38.7509153
lag1_China.Imports                 5.4468696
lag1_China.Exports                 .
lag1_sdrapctyTemp                  1.1134451
lag1_ilspringTemp                  1.9946703
lag1_ilpeoriaTemp                 -1.9717670
lag1_iasioctyTemp                  0.2979801
lag1_arlirockTemp                 -0.8023361
lag1_msjacksoTemp                  0.8658385
lag1_ndfargoTemp                  -1.4165737
```

Fig 7. Lasso Model Summary



```
> coef(model_ridge1, s=model_ridge1$lambda.min)
25 x 1 sparse Matrix of class "dgCMatrix"
                                         s1
(Intercept)                     245.8786510
lag1_brent_oil_close              2.3583054
lag1_crude_oil_close              1.8043126
lag1_natural_gas_close           -5.9176461
lag1_heating_oil_close           79.9474550
lag1_CPI                          3.8376126
lag1_United.States.Production    -2.9647108
lag1_United.States.Imports      133.2398423
lag1_United.States.Exports        2.0546833
lag1_Argentina.Production        -2.9456811
lag1_Argentina.Imports           -2.2733823
lag1_Argentina.Exports           21.3033364
lag1_Brazil.Production           -0.6209652
lag1_Brazil.Imports            -106.2317060
lag1_Brazil.Exports              -2.8405511
lag1_China.Production           -21.7596371
lag1_China.Imports                1.0760093
lag1_China.Exports                5.1569193
lag1_sdrapctyTemp                 0.4781760
lag1_ilspringTemp                 0.2284663
lag1_ilpeoriaTemp                -0.4587453
lag1_iasioctyTemp                 0.1443408
lag1_arlirockTemp                 0.1160323
lag1_msjacksoTemp                 0.2541017
lag1_ndfargoTemp                 -0.6318978
```

Fig 8. Ridge Model Summary

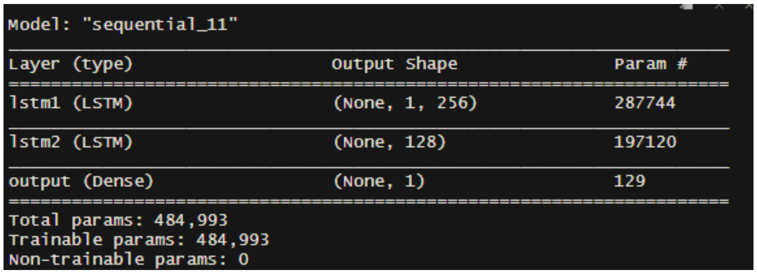


```
                                         s1
(Intercept)                     247.9364680
lag1_brent_oil_close              4.0533013
lag1_crude_oil_close              1.2930171
lag1_natural_gas_close           -6.2191653
lag1_heating_oil_close           19.2900968
lag1_CPI                          6.0314821
lag1_United.States.Production    -5.5315246
lag1_United.States.Imports      151.2507253
lag1_United.States.Exports        6.3198784
lag1_Argentina.Production        -4.8173554
lag1_Argentina.Imports           37.7150338
lag1_Argentina.Exports           21.7955867
lag1_Brazil.Production           -0.7132502
lag1_Brazil.Imports            -102.1999340
lag1_Brazil.Exports              -8.6095353
lag1_China.Production           -39.0561921
lag1_China.Imports                3.2529324
lag1_China.Exports                0.0000000
lag1_sdrapctyTemp                 0.7506375
lag1_ilspringTemp                 0.6596304
lag1_ilpeoriaTemp                -0.8102039
lag1_iasioctyTemp                 0.2464364
lag1_arlirockTemp                 0.0000000
lag1_msjacksoTemp                 0.3290366
lag1_ndfargoTemp                 -0.9284885
```

Fig 9. Elastic Net Model Summary

4.1.5 Long Short Term-Memory (LSTM)

The team is also interested in the performance of machine learning model, Long-Short Term Memory (LSTM) [6], that is widely used in time-series forecasting for its ability to process entire sequence of data. We utilize Keras library and keras_model_sequential() function to build two-layer LSTM model with 256 and 128 number of neurons. The model takes 24 inputs which are all the predictors and outputs one forecast ahead. The full structure of the neural network model can be seen in Fig 10.



```
Model: "sequential_11"
________________________________________________________________
Layer (type)                    Output Shape              Param #
================================================================
lstm1 (LSTM)                    (None, 1, 256)            287744
________________________________________________________________
lstm2 (LSTM)                    (None, 128)               197120
________________________________________________________________
output (Dense)                  (None, 1)                 129
================================================================
Total params: 484,993
Trainable params: 484,993
Non-trainable params: 0
```

Fig 10. LSTM Structure

All variables need to be scaled before they are fed into LSTM model. After that, we apply elbow method to identify that 12 is the optimal number of epochs that best minimizes the validation loss and can be seen in Fig 11. Then, we transform the variable back to calculate the error of the validation set. The LSTM model is able to achieve Root Mean Square Error of 80.15.

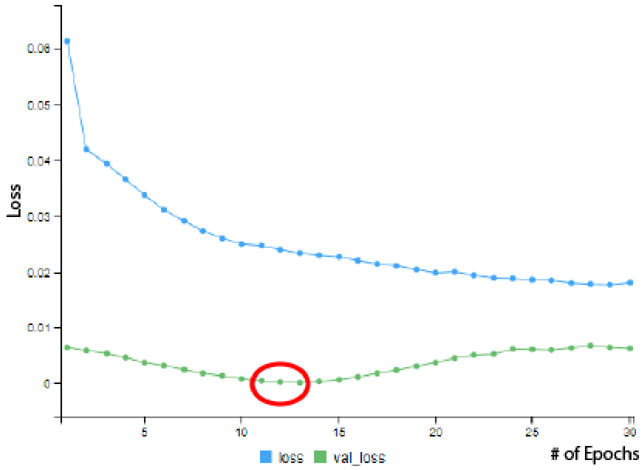


Fig 11. Training and Validation Error

4.2 Results and Significant Variables

Firstly, the team compared features across all models and found that imports, exports, and production values from major countries involved in the soybean market affect soybean prices. While all production variables have negative impact on soybean price, imports and exports values have varying effects. United States and Argentina import-export activities on global soybean trade influence the soybean price positively. However, higher values of soybean import and export from Brazil lead to lower soybean price. Large coefficients for United States Imports and Brazil Imports in particular were noted from the Lasso, Ridge and Elastic Net models. These two variables have a significant and opposite effect on Soybean prices. It is worth noting that Brazil is currently the largest exporter of soybean in the world, followed closely by the United States. Brazil is the main competitor of the US, in terms of soybean export.

Out of four oil and gas price variables, soybean prices move in tandem (that is, in all models) with brent oil prices only. This may be a result of all oil and gas prices having strong correlations to soybean prices, as noted from correlation analysis and the PCA, thus forming multicollinearity between independent variables. In regression analysis, it is often the case that models include only a subset of multicollinear variables.

We also discovered that soybean prices are closely related to the US Consumer Price Index (CPI). This finding fails to support our 3rd hypothesis. A factor that may lead to different outcomes is the time period used in this analysis and the reference. In our analysis, the dataset used is from the year 2007 to 2020, whereas [7] analyzes a much longer period (1965-2022). Having said that, this discovery may signify that in recent years, soybean prices and inflation rates have started to move together.

Regarding temperature variables, 5 out of 7 predictors are identified to have adequate effects on soybean price. However, when compared to other variables, the coefficients are considered small and have diverse impact on soybean prices. The small coefficient of temperature variables further supports the finding from the PCA that there seems to be little relationship between temperature variables and soybean. To summarize, the relationship between the temperature of major counties and the soybean price is inconclusive.

Table 1. Independent Variable Importance

Variables	Model 1 Basic GLM	Model 2 Elastic Net	Model 3 Lasso	Model 4 Ridge
Brent Oil Close*	+	+	+	+
Crude Oil Close	X	+	+	+
Natural Gas Close	+	-	+	-
Heating Oil Close	-	+	-	+
United States CPI*	+	+	+	+
United States Production*	-	-	-	-
United States Imports*	+	+	+	+
United States Exports	X	+	+	+
Argentina Production*	-	-	-	-
Argentina Imports*	+	+	+	+
Argentina Exports	+	+	+	-
Brazil Production*	-	-	-	-
Brazil Imports*	-	-	-	-
Brazil Exports*	-	-	-	-
China Production*	-	-	-	-
China Imports*	+	+	+	+
China Exports	-	X	X	+
Rapid City, SD Temp*	+	+	+	+
Peoria, IL Temp*	-	-	-	-
Springfield, IL Temp*	+	+	+	+
Jackson, MS Temp*	+	+	+	+
Fargo, ND Temp*	-	-	-	-
Little Rock, AR Temp	-	X	-	+
Sioux City, IA Temp	X	+	+	+

+ The variable has a positive coefficient estimate
- The variable has a negative coefficient estimate
X Insignificant variable
* The coefficient estimate of the variable has the same sign across all four models

4.3 Model Performance Comparison

To ultimately evaluate which model is the best at predicting soybean stock prices, we selected the model that yields the lowest Root Mean Squared Error (RMSE). As can be seen in Table 2., LSTM beats other traditional time series regression methods with an RMSE of 80.15 on the validation set. LSTM model can effectively capture underlying patterns, thus producing low-error forecasts. Next, we assessed the LSTM model on the testing dataset. Surprisingly, LSTM performs higher with an RMSE of 69.13 which implies that on average, the prediction can be off +- $69.13 or equivalent to 6.08%. This performance is considered good for time series forecasting.

Table 2. Model's Performance on Validation Set

Metrics	**Model 1 Basic GLM**	**Model 2 Elastic Net**	**Model 3 Lasso**	**Model 4 Ridge**	**Model 5 LSTM**
Root Mean Squared Error (RMSE)	105.1	100.97	93.64	103.84	**80.15**

5. FURTHER RESEARCH

If the team had more time to continue our project, we would have liked to test the models on several sets of training data with different lag values such as 7 days or 30 days. Incorporating other lags may improve model's accuracy as it may capture other underlying patterns such as seasonality. Additionally, it would have been interesting to compare the performance of non-linear models such as random forest or CART models.

6.CONCLUSION

The team has performed a thorough analysis to identify important factors and select the best model in predicting soybean prices. From forming problem statements, collecting relevant datasets, performing exploratory data analysis, treating outliers, to implementing relevant models, all the steps have been carefully discussed and considered with respect to the nature of time-series data.

With that in mind, we successfully identified that key countries' activities in the soybean market, other commodities prices, and inflation rates have major impacts on soybean prices. While different key countries' activities have varying effects on soybean prices, crude oil and inflation rates influence soybean prices positively. Important stakeholders that participate in the soybean industry may take these variables into account in forming macro or micro level decisions. Producers at a local level may make use the information of factors that significantly influence soybean prices to gauge how much revenue (and in which areas) they should invest for production and how profitable it is likely to be.

On top of that, the team was able to build a high performing model in predicting soybean prices by incorporating identified important factors as the predictors and LSTM as the framework. A more macro-level decision maker such as the US government may utilize the model and incorporate these findings in setting domestic soybean price and other policymaking processes to not only maximize export revenue but also encourage food security.

To conclude, this analysis provides a deeper understanding on how different factors influence soybean price. This provides valuable insight to stakeholders at all levels in the US soybean industry.

References

[1] Economic impact of U.S. soybeans &amp; end products on the U.S. economy. NOPA. (2020, April 17). Retrieved April 13, 2023, from https://www.nopa.org/resources/economic-impact-of-u-s-soybeans-end-products-on-the-u-s-economy/

[2] Basting, B. (2022, July 12). Corn and soybean price volatility remains high. Retrieved March 9, 2023, from https://www.farmprogress.com/corn/corn-and-soybean-price-volatility-remains-high

[3] Soybean 2021 export highlights. USDA Foreign Agricultural Service. (n.d.). Retrieved March 8, 2023, from https://www.fas.usda.gov/soybean-2021-export-highlights.

[4] Person, &amp; Braun, K. (2023, February 7). Column: U.S. soybean exports erupt ahead of Brazil's massive, but slower harvest. Reuters. Retrieved March 9, 2023, from https://www.reuters.com/markets/commodities/us-soybean-exports-erupt-ahead-brazils-massive-slower-harvest-2023-02-07/#:~:text=a%20month%20ago-,Column%3A%20U.S.%20soybean%20exports%20erupt%20ahead,Brazil's%20massive%2C%20but%20slower%20harvest&amp;text=NAPERVILLE%2C%20Ill.%2C%20Feb%206,of%20collecting%20its%20record%20harvest.

[5] Major factors affecting global soybean and products trade projections. USDA ERS - Major Factors Affecting Global Soybean and Products Trade Projections. (n.d.). Retrieved March 9, 2023, from https://www.ers.usda.gov/amber-waves/2016/may/major-factors-affecting-global-soybean-and-products-trade-projections/#:~:text=The%20primary%20factors%20driving%20global,major%20agricultural%20importers%20and%20exporters.

[6] R. Ghanbari and K. Borna, "Multivariate Time-Series Prediction Using LSTM Neural Networks," *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, Tehran, Iran, 2021, pp. 1-5, doi: 10.1109/CSICC52343.2021.9420543.

[7] Jiao, H. (2022, June 21). Inflation and commodity prices. farmdoc daily. Retrieved April 14, 2023, from https://farmdocdaily.illinois.edu/2022/06/inflation-and-commodity-prices.html