MinneMUDAC Data Science Challenge

November 9, 2019

Team Name: Women in Math and Stats, G34
Members: Somyi Baek, Cora Brown, Sarah Milstein, Yu Yang
Advisor: Gilad Lerman

The Challenge

- **Objective:** Investigate the factors/characteristics that influence the soybean futures closing prices for 3 different contract months

- **Primary Goal:** Predict soybean closing prices for 5 days: November 4 - 8, and for 3 contract months: March, May, and July 2020

Collected Data

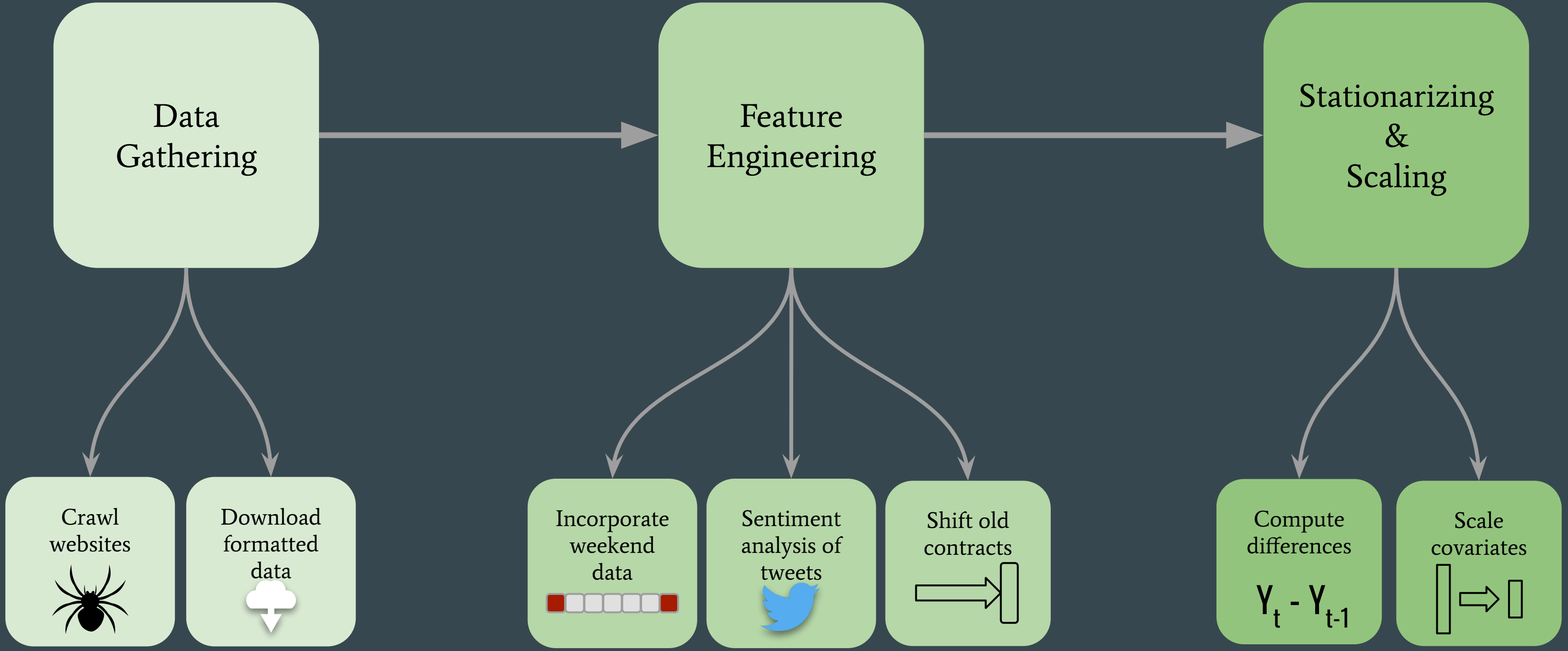
Commodity prices

- Soybean
- Corn
- Canola
- Soybean Meal
- Soybean Oil
- Rice
- Hogs

External features

- Weather
- Interest rates
- Dow Jones
- Tweets
- Tariff
- Fires
- Planting
- Disappearance
- Stocks

Data Pipeline



Data Gathering

Feature Engineering

Stationarizing & Scaling

Crawl websites

Download formatted data

Incorporate weekend data

Sentiment analysis of tweets

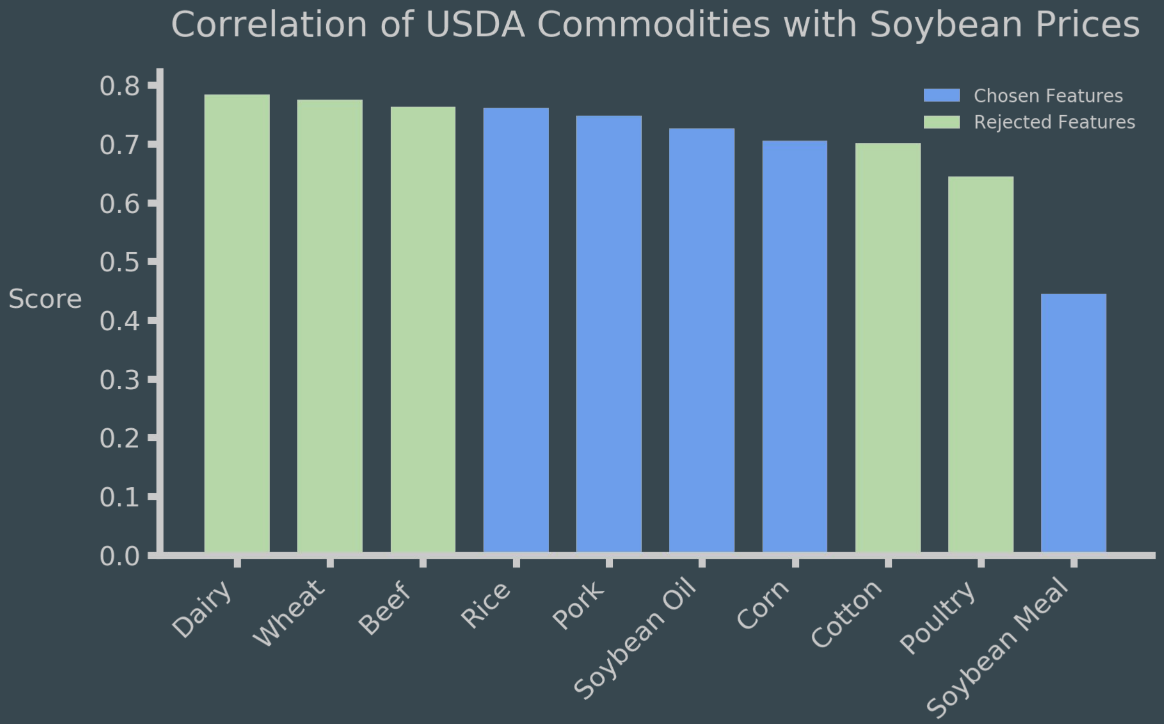
Shift old contracts

Compute differences

$Y_t - Y_{t-1}$

Scale covariates

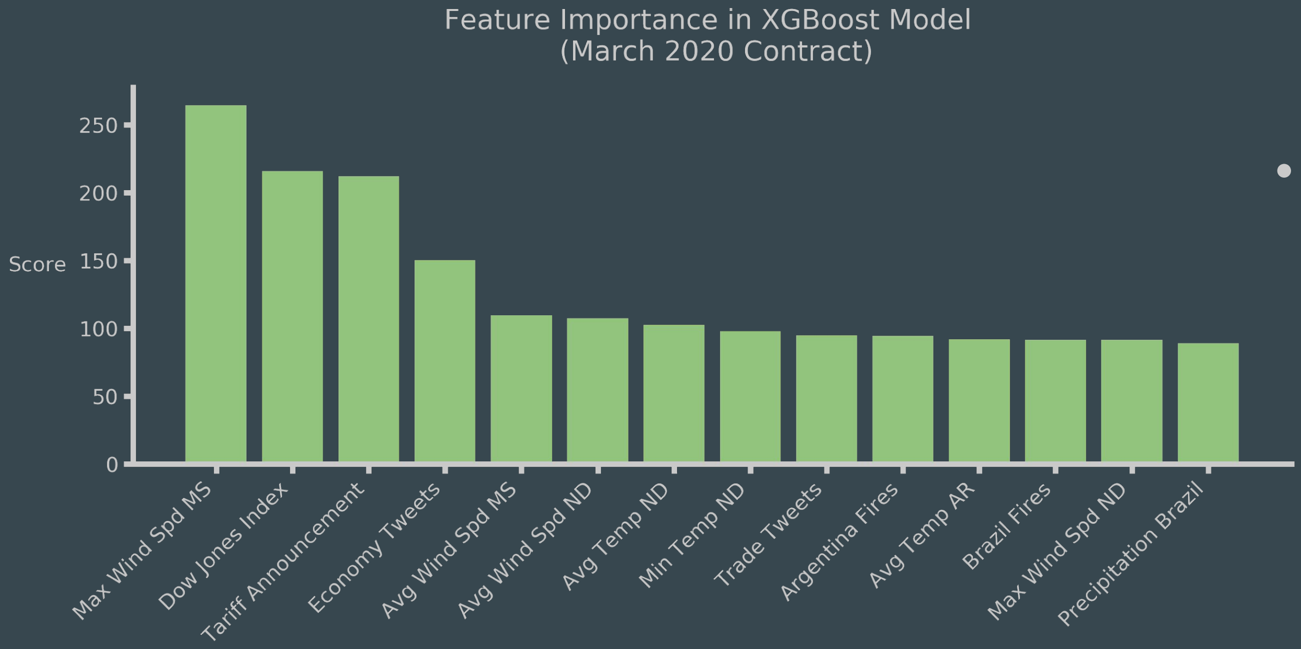
Feature Exploration: Commodities



- Features chosen by Granger causality test
  - Rice
  - Pork
  - Corn

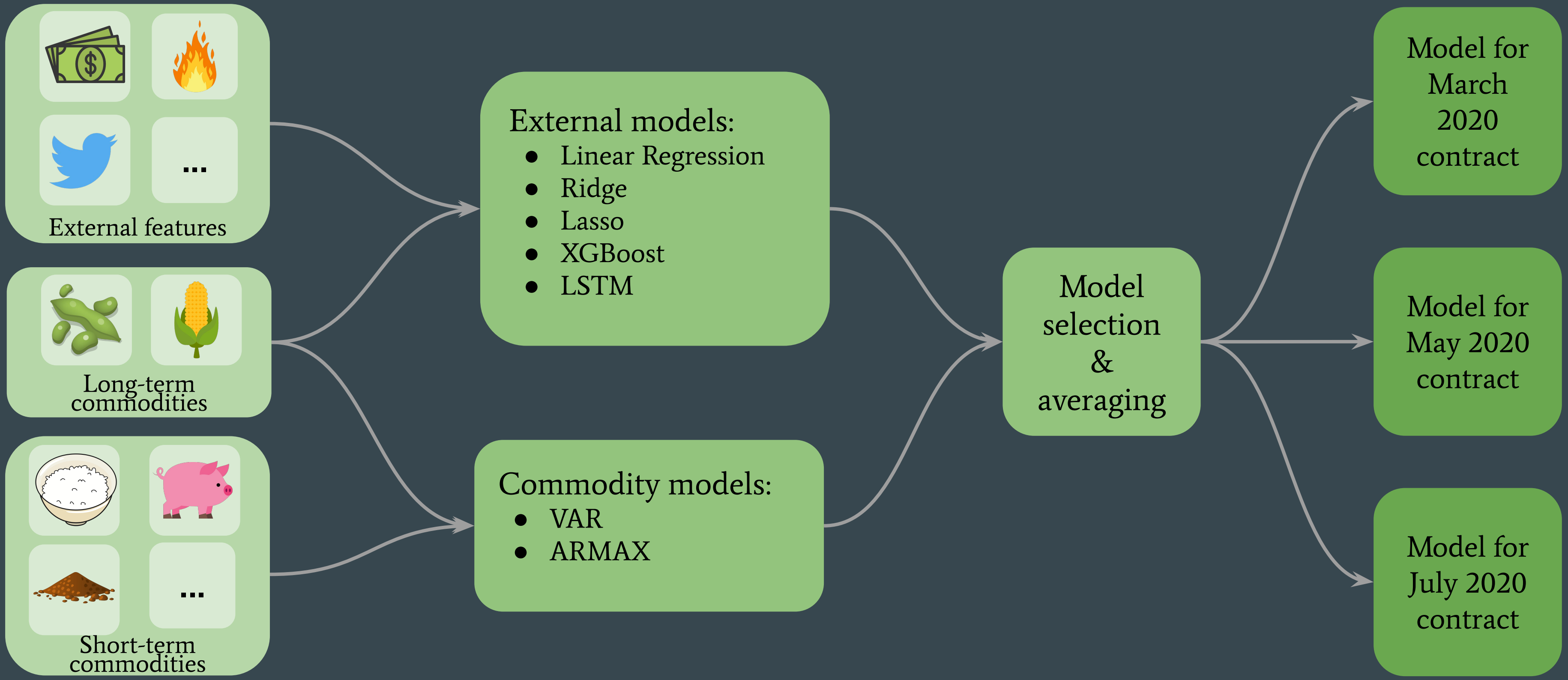
- Features chosen by known interdependencies
  - Soybean oil
  - Soybean meal

Feature Exploration: External Features



- Important features:
  - Tweets on trade and the economy
  - Weather in Brazil, Argentina, Mississippi, North Dakota
  - Fires in Brazil and Argentina

Modeling Strategy



External features

Long-term commodities

Short-term commodities

External models:

- Linear Regression
- Ridge
- Lasso
- XGBoost
- LSTM

Commodity models:

- VAR
- ARMAX

Model selection & averaging

Model for March 2020 contract

Model for May 2020 contract

Model for July 2020 contract

Model Interpretation

March 2020: Nov. 4-8 Prediction

	XGBoost	LSTM	VAR
↑ price	Interest Rate ND Weather Dow Jones	Mar 2017 Soybean May 2017 Soybean Mar 2018 Soybean Argentina Weather	Corn Rice
price ↓	Sunflower Seed Meal Corn Mar 2016 Soybean	MS Weather May 2019 Soybean Jul 2020 Soybean	Soybean Meal Soybean Oil May 2020 Soybean

Conclusion

- Making price predictions is hard.
- Predicting further out days is harder.
- Good indicators of soybean prices:
  - Corn, which has similar uses as soybeans, and whose market size is x3-4 that of soybeans
  - Soybean contracts for different months
  - Soybean oil and soybean meal, which are connected in production processes
  - Macroeconomic indicators, such as Dow Jones Industrial and interest rates
  - Weather in high production areas
- Because our primary goal is 5-day forecast, the predictive power of the related commodities outweighs that of random events, such as tweets and tariffs.

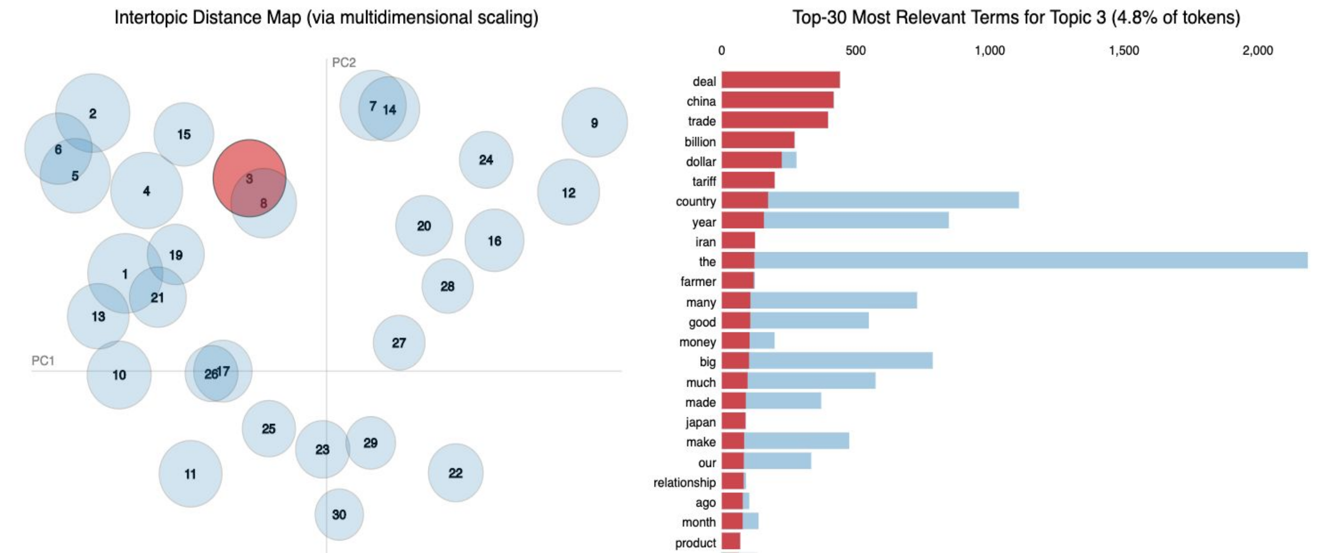
Thank you. Questions?

Data Compromises

- Some commodities futures are only available for a few months (rice, canola)
    - Develop short term and long term models to account for different time scales
- Weekend data from tweets, weather, etc. should affect Monday’s closing price
    - Average values from Saturday, Sunday, and Monday to make features from all three days account for Monday
- Dates for previous contracts (e.g. March 2019) do not overlap with dates for current contracts (e.g. March 2020)
    - Shift dates for previous contracts to roughly overlap with current dates

Feature Engineering of Tweets

- Cluster Trump tweets by topic using LDA model.
- Perform sentiment analysis on trade and economy relevant tweets data.
- Use likes and retweets number as weight to average sentiment score.

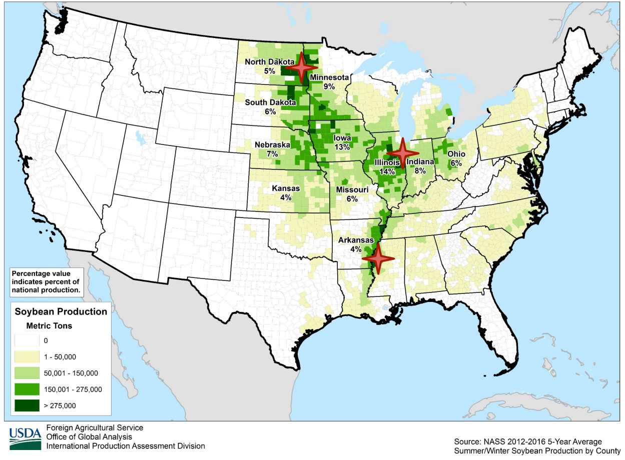


Weather Data

Weather taken from stations located near high soybean production areas.

Source: National Oceanic and Atmospheric Association (NOAA)

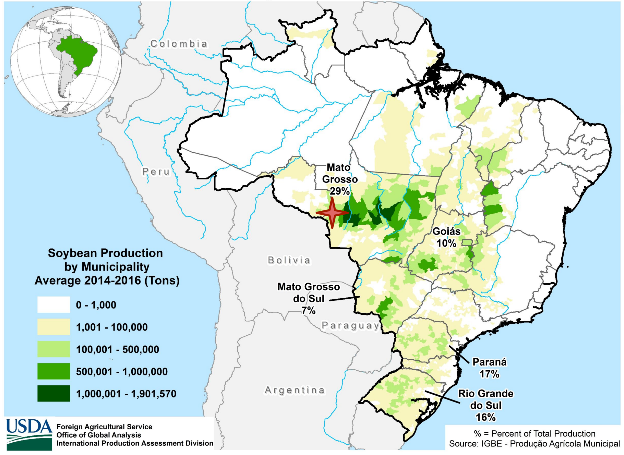
United States: Soybean Production



Foreign Agricultural Service
Office of Global Analysis
International Production Assessment Division

Source: NASS 2012-2016 5-Year Average
Summer/Winter Soybean Production by County

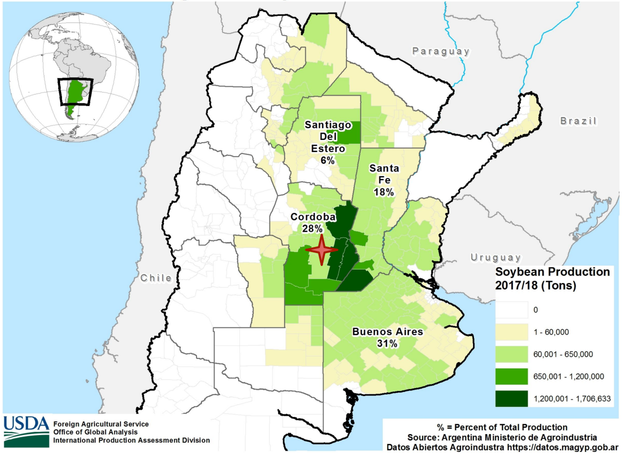
Brazil: Soybean Production



Foreign Agricultural Service
Office of Global Analysis
International Production Assessment Division

% = Percent of Total Production
Source: IGBE - Produção Agrícola Municipal

Argentina: Soybean Production

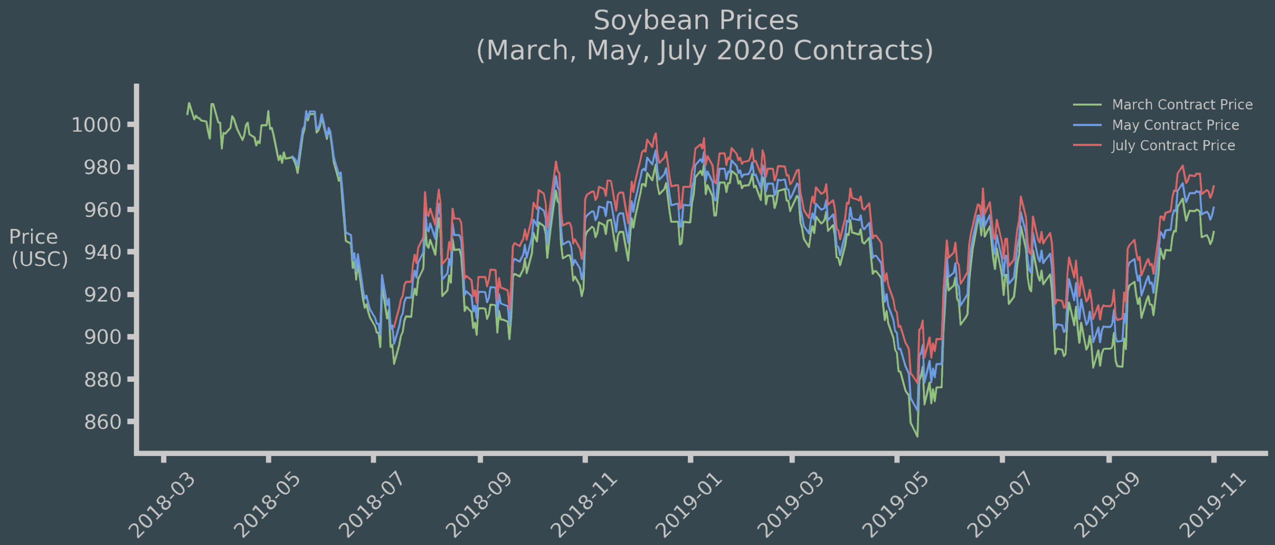


Foreign Agricultural Service
Office of Global Analysis
International Production Assessment Division

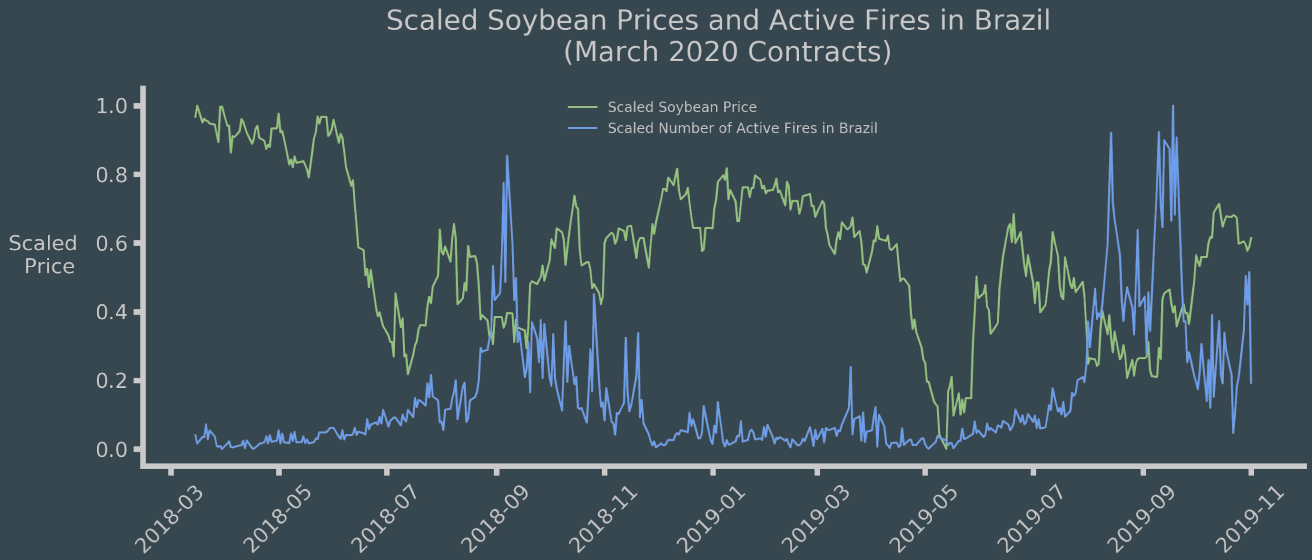
% = Percent of Total Production
Source: Argentina Ministerio de Agroindustria
Datos Abiertos Agroindustra https://datos.magyp.gob.ar

Feature Exploration: Comparison Across Contract Months

- Soybean contracts for different months are highly correlated
- July prices are highest, followed by May and then March
- To capture this correlation, we use May and July prices as features for predicting March prices, etc.

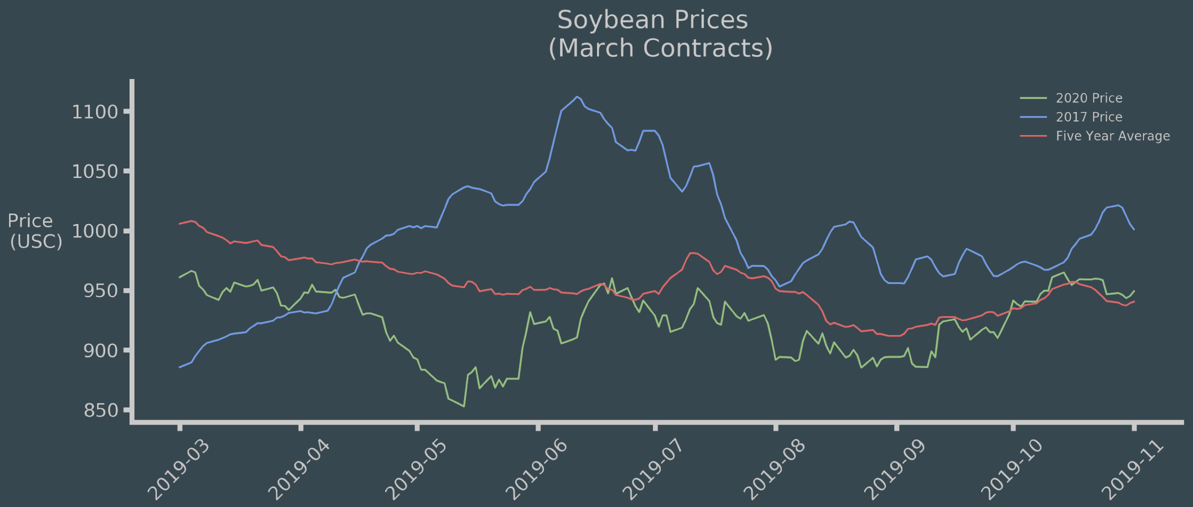


Data Exploration: Fires in Brazil



- The number of fires in Brazil corresponds to jumps in soybean price.

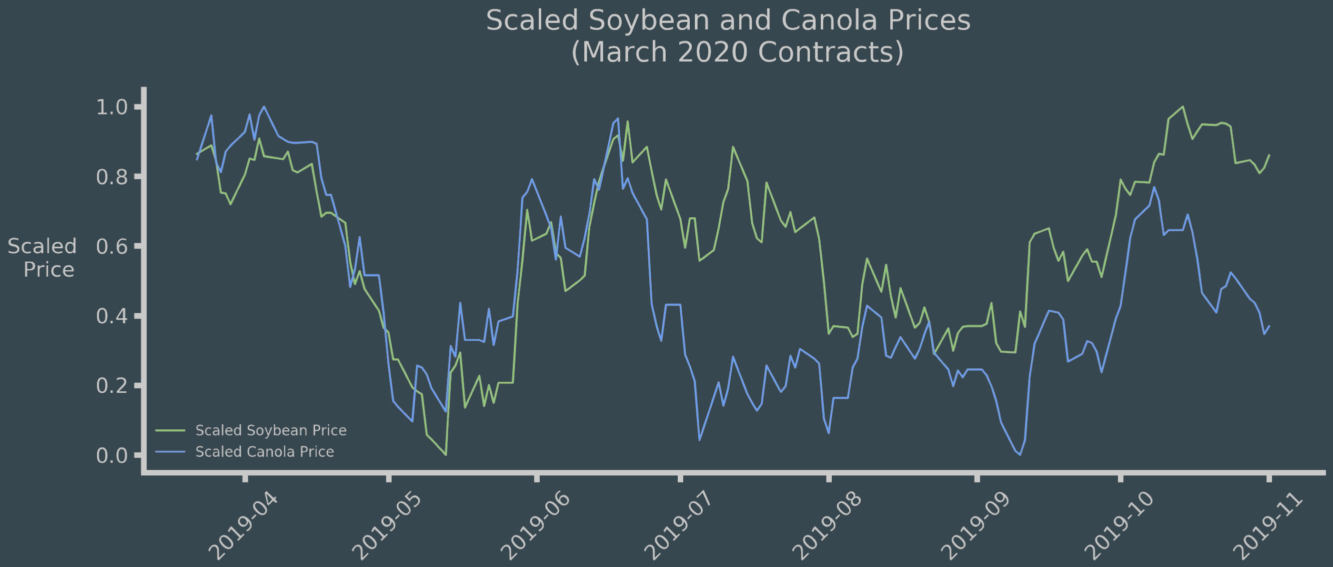
Data Exploration: Historical Soybean Contracts



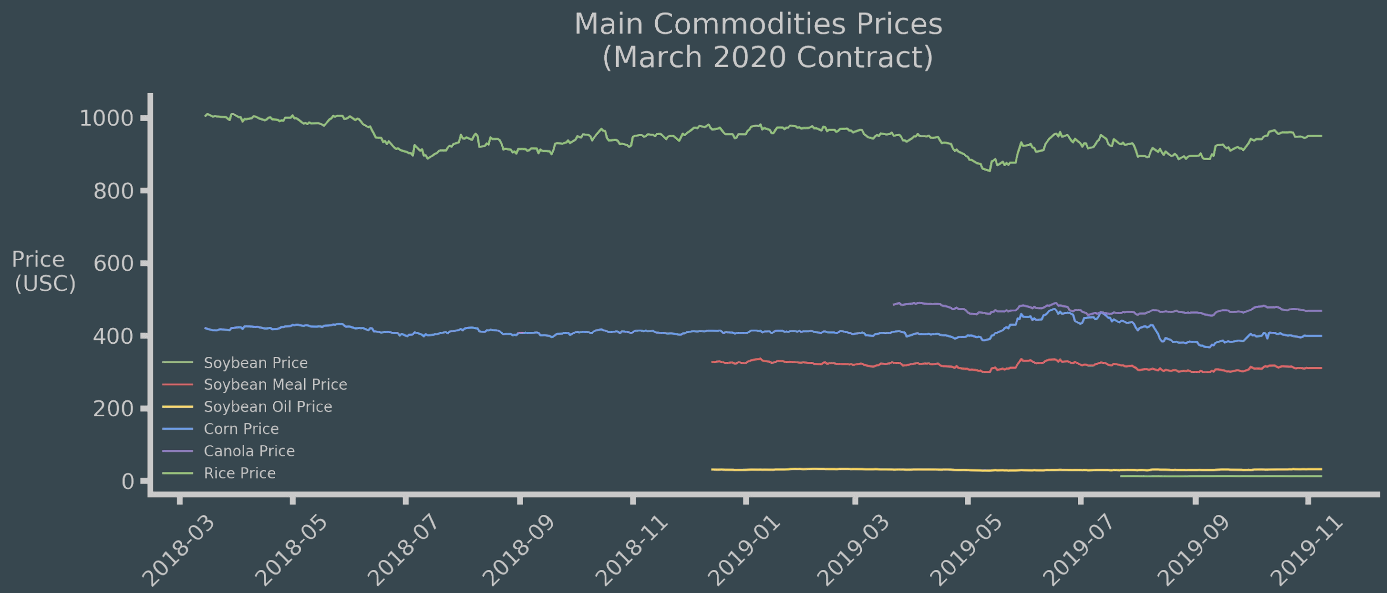
- March 2017 and March 2020 contracts have similar average prices, but the March 2020 price drops after the delayed planting, leading to lower-than-average prices

Data Exploration: Canola and Soybean Prices

- After scaling, we find that canola and soybean markets display similar patterns



Main Commodities Prices

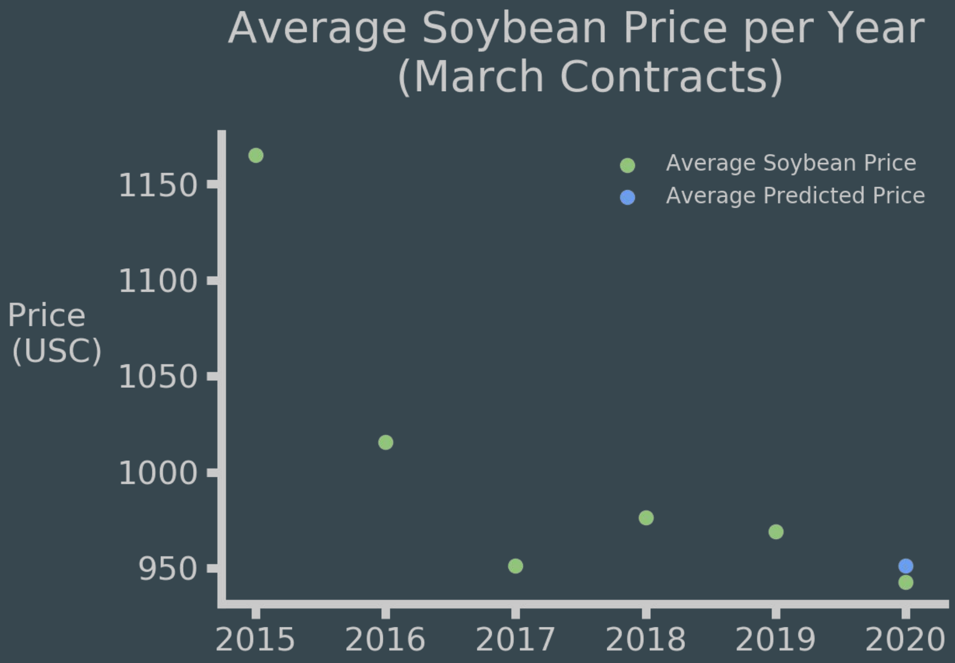


- Corn and Soybean contracts are available for the same amount of time
- Corn and Soybean prices follow similar patterns
- Other contracts are available for much shorter amount of time
- Patterns are more difficult to establish among other commodities

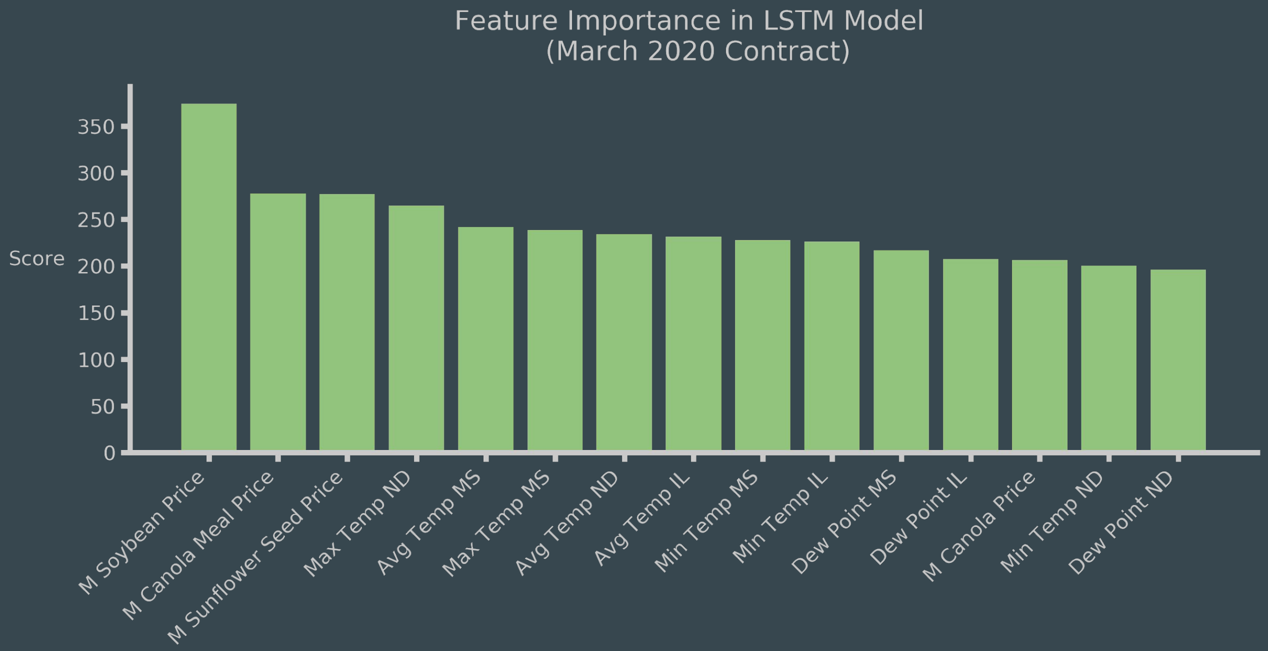
Yearly Average Prices

- Soybean prices on average have declined since 2015

- Our predictions are slightly above the mean for this year, but on trend with the lower prices in recent years



Data Exploration: Feature Importance

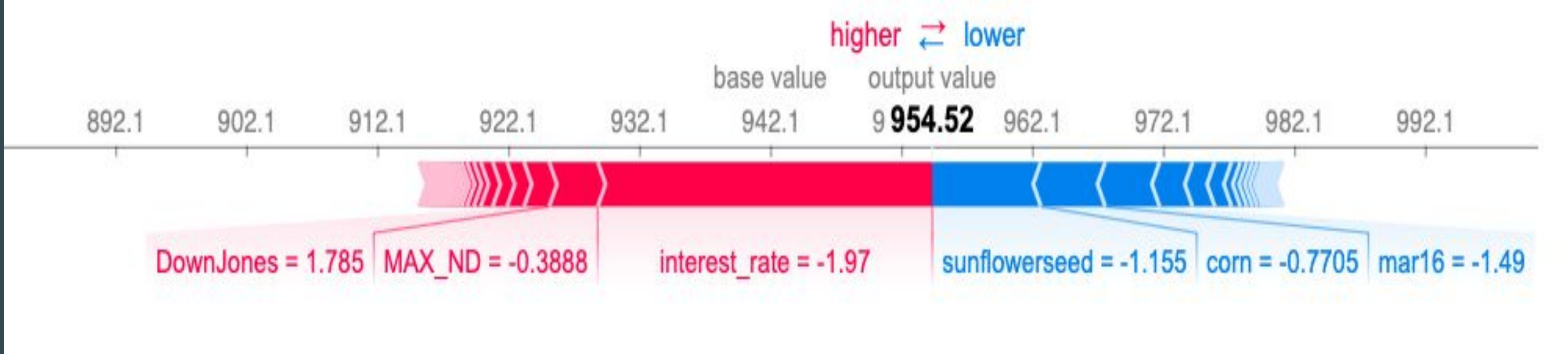


Model Details

- Walk forward cross-validation for model selection
- Grid search for model averaging
- External Model (long term features):
  - Use our external features and long term commodities in models that train farther back in the past
  - Create one model for each day: Monday's model is based on lag 1 exogenous values, Tuesday's model is based on lag 2, etc., and Friday's model is based on lag 5
- Short term model:
  - Use Vector Autoregression to capture the evolution and the interdependencies between multiple economic data
  - Make one prediction of five days so that our predictions capture autocorrelation patterns

Model Interpretation

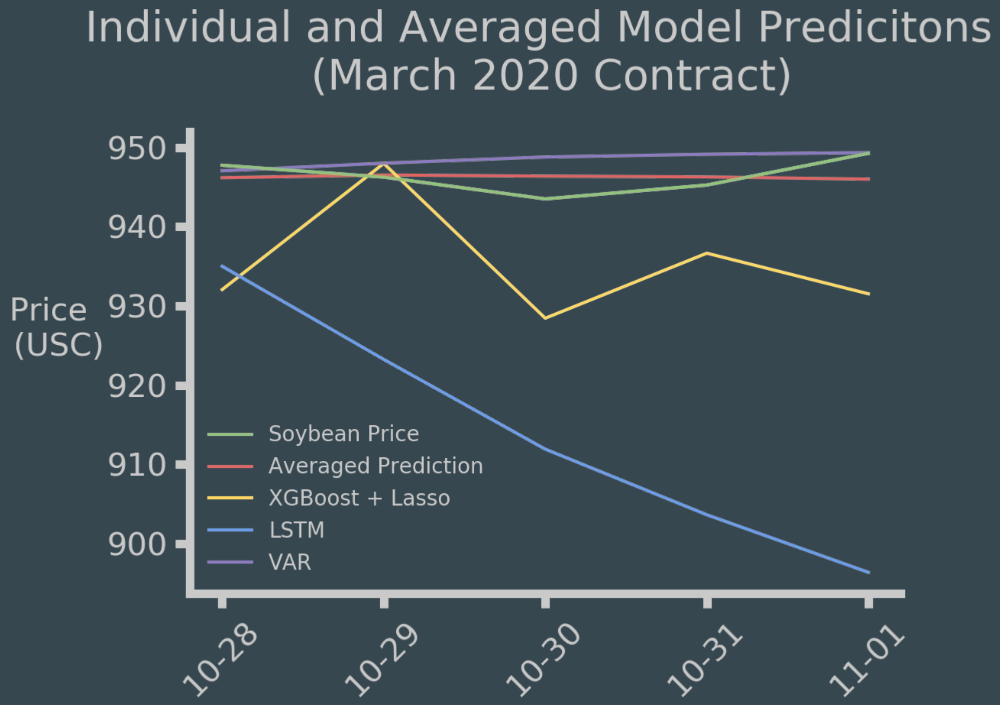
XGBoost for Nov. 5



Model Selection & Weighted Averaging

Approach:

- Use walk-forward cross validation to select models
- Use grid search to find the best combination of models using the week Oct 28 - Nov 1 as validation



Model Results

March 2020			
	Pred.	Actual	Diff.
Nov. 4	950.50	951.25	0.75
Nov. 5	951.00	947.25	-3.75
Nov. 6	951.00	940.75	-10.25
Nov. 7	951.25	948.75	-2.50
Nov. 8	951.75	948.00	-3.75

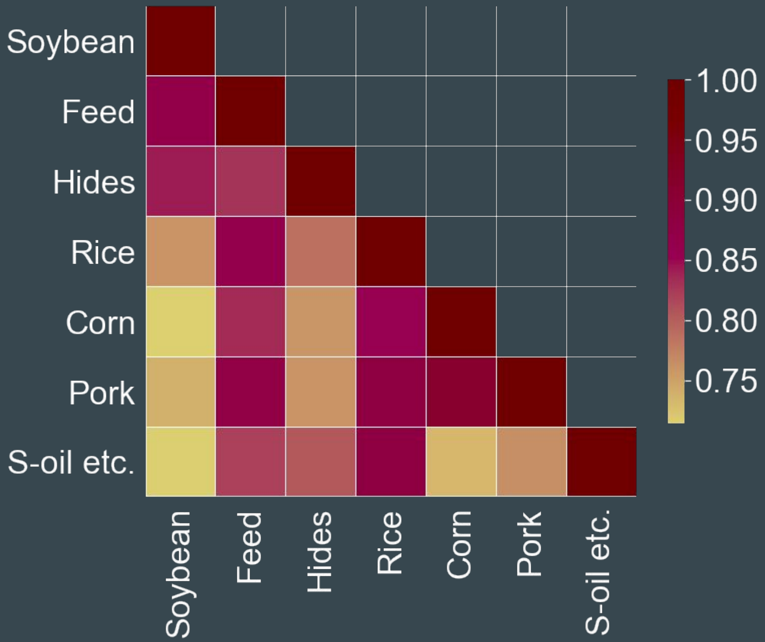
May 2020			
	Pred.	Actual	Diff.
Nov. 4	961.75	963.25	1.50
Nov. 5	962.25	959.00	-3.25
Nov. 6	963.00	952.75	-10.25
Nov. 7	963.50	960.25	-3.25
Nov. 8	964.25	959.50	-4.75

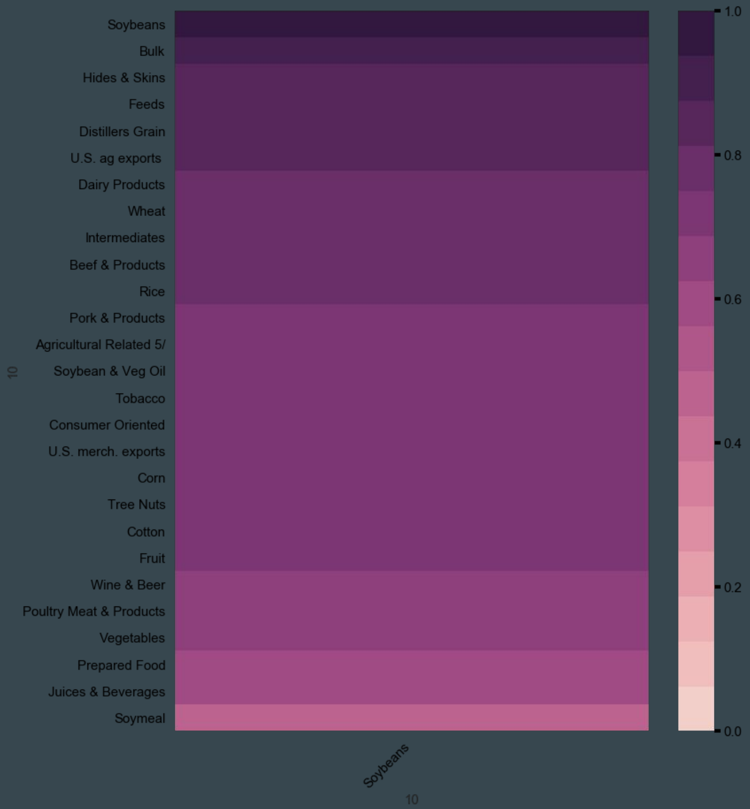
July 2020			
	Pred.	Actual	Diff.
Nov. 4	971.50	973.50	2.00
Nov. 5	972.00	969.25	-2.75
Nov. 6	972.50	963.25	-9.25
Nov. 7	972.75	970.75	-2.00
Nov. 8	973.00	969.75	-3.25

Average error: -3.91

Data Sources

- Data supplied by Farm Femmes
- MRCI's Free Historical Futures Prices: https://www.mrci.com/ohlc/index.php
- Trump Twitter Archive: http://www.trumptwitterarchive.com/archive
- The US-China Trade War: A Timeline: https://www.china-briefing.com/news/the-us-china-trade-war-a-timeline/
- NOAA Weather: https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets
- U.S. Agricultural Trade Data: https://www.ers.usda.gov/data-products/foreign-agricultural-trade-of-the-united-states-fatus/us-agricultural-trade-data-update
- NASA Fire Data: https://firms.modaps.eosdis.nasa.gov/download/


Soybean
Feed
Hides
Rice
Corn
Pork
S-oil etc.
Soybean
Feed
Hides
Rice
Corn
Pork
S-oil etc.
1.00
0.95
0.90
0.85
0.80
0.75


Soybeans
Bulk
Hides & Skins
Feeds
Distillers Grain
U.S. ag exports
Dairy Products
Wheat
Intermediates
Beef & Products
Rice
Pork & Products
Agricultural Related 5/
Soybean & Veg Oil
Tobacco
Consumer Oriented
U.S. merch. exports
Corn
Tree Nuts
Cotton
Fruit
Wine & Beer
Poultry Meat & Products
Vegetables
Prepared Food
Juices & Beverages
Soymeal
10
Soybeans
10
1.0
0.8
0.6
0.4
0.2
0.0

(temp slide for relevant Tim facts)

- Corn is more influential on soybean prices than soybeans, bc corn market 3-4x bigger than soybeans and animal feed is usually corn & soybean mix
- US and China are 2 biggest bulk commodity producers in the world (20% together)