Commodity Price Analysis Through PCA, KNN, and Neural Nets

for Corn and Soybeans

Joe Nunez

Harvey Mudd College

May 25, 2018

Data Set

- The data set I used was taken from Kaggle.com and from Quandl.com, and consisted of monthly economic data concerning corn and soybeans from 2007 to 2014, as well as futures market data for corn and soybean prices over the same period.
- The economic data included 112 different indicators related to corn and 69 different indicators related to soybean data, including figures on imports, exports, total area planted/harvested, regional production figures, and the size of regional reserves.
- The market data consisted of daily measures of market open, high, low, close, volume, and open interest for the futures contracts.

Data Set

- The economic data was gathered from the USDA's monthly World Agricultural Supply and Demand Estimate (WASDE).
- The market data was gathered from Quandl.com

Methods — Labeling

- I took two approaches: one using the market data alone and one using the economic data.
- For both data sets, I used the prices from the market data to determine whether the price of the given commodity (corn or soybeans) went up or down in a given time step. If the price did go up, that data point was labeled true, and if it did not, it was labeled false.
- I also labeled the data with its price at the next timestep for use in regression.
- For market data, the time step was one day, and for WASDE data, the time step was one week.

Methods — Features

- For the economic data, the feature vector was simply all of the entries provided in the WASDE report, after being normalized to all be between 0 and 1 (subtracting by the minimum, then dividing by the difference between min and max).
- For the market data, the feature vector consisted of open, high, low, close, volume, and open interest for each of the previous 22 days.

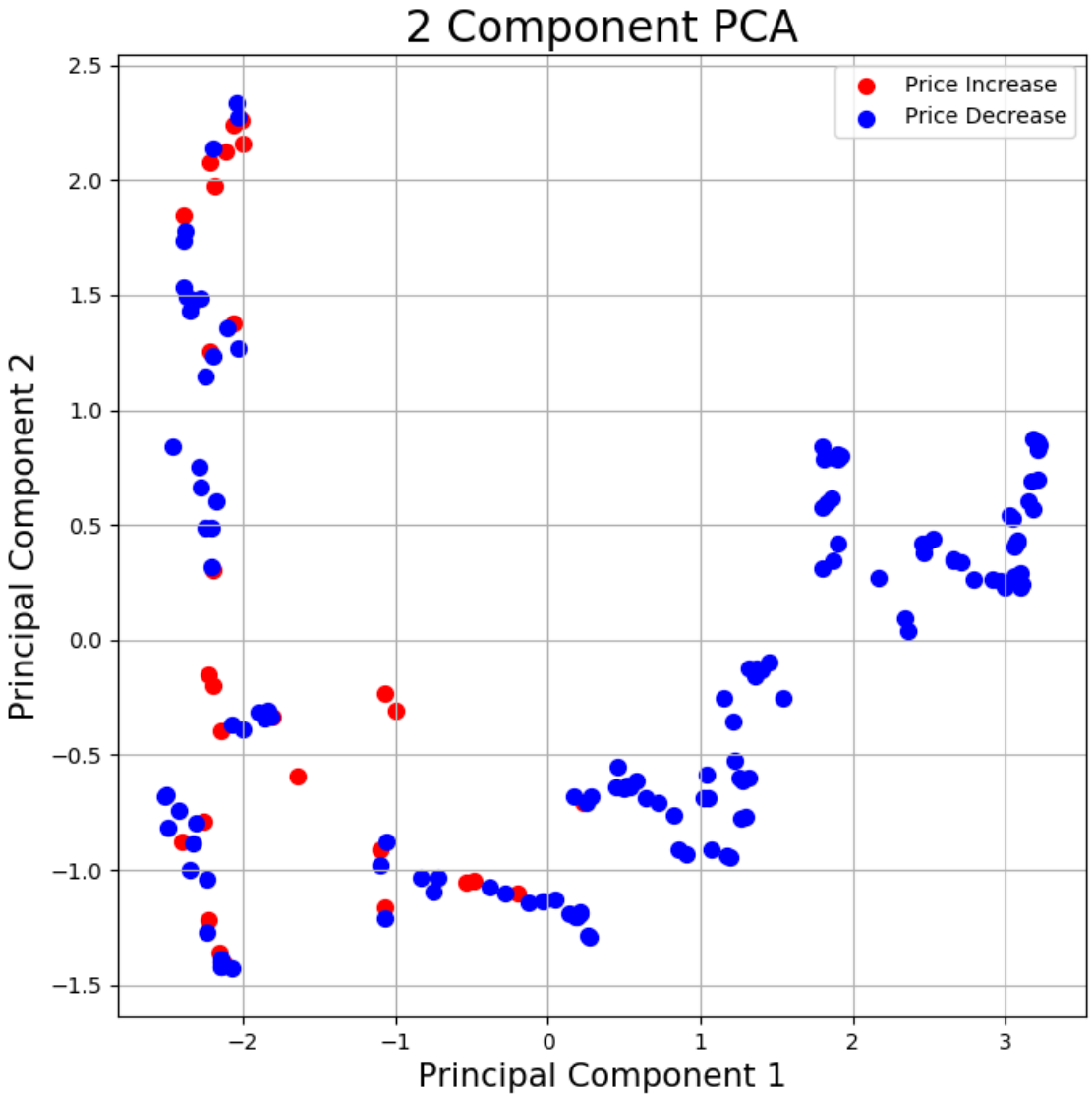
Methods — PCA-KNN

- Principal component analysis was used for all number of features from 1 to 40.
- k nearest neighbors was used to classify the points resulting from PCA, with values of k ranging from 2 to 30.
- Plots will not show higher numbers of components or values of k because they did not show higher performance.
- Sklearn packages were used for both PCA and KNN.

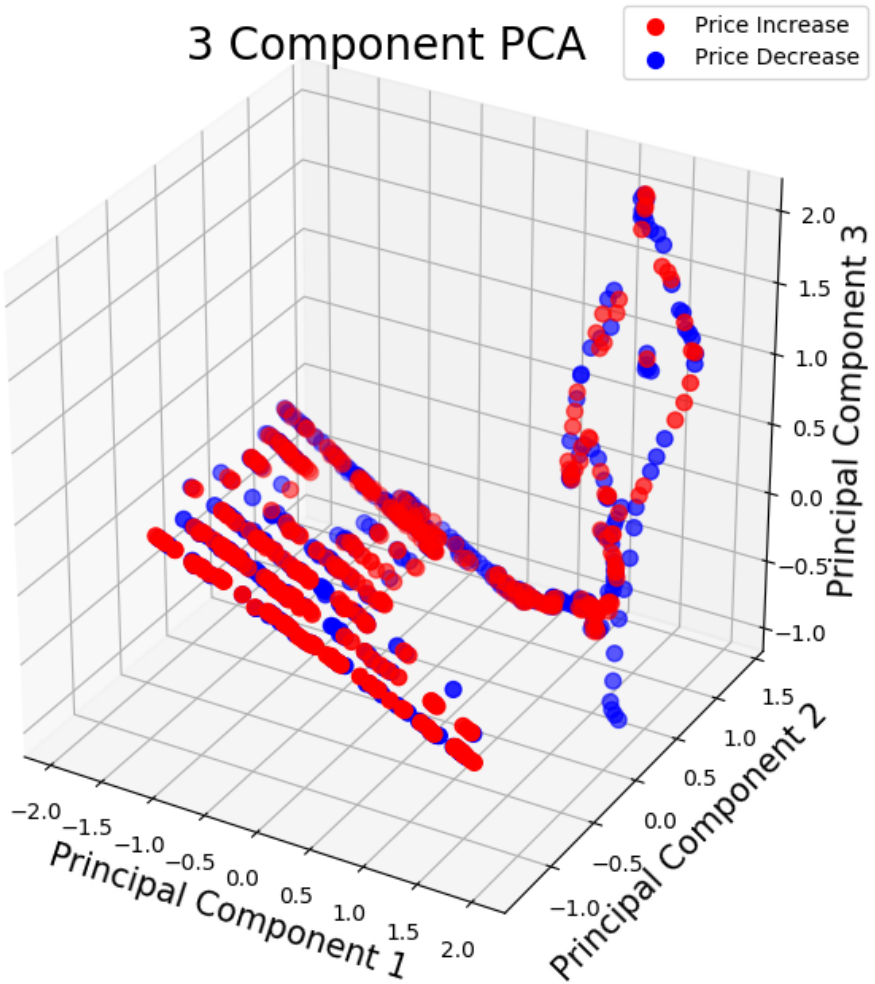
Methods — Neural Network

- Layers of size [256, 256, 32, 1]
- Using sigmoid activation function.
- Used LSTM for market data, where the LSTM layers were the first two layers.
- Used a multi-layer perceptron for WASDE data.
- Neural networks were implemented using Keras.

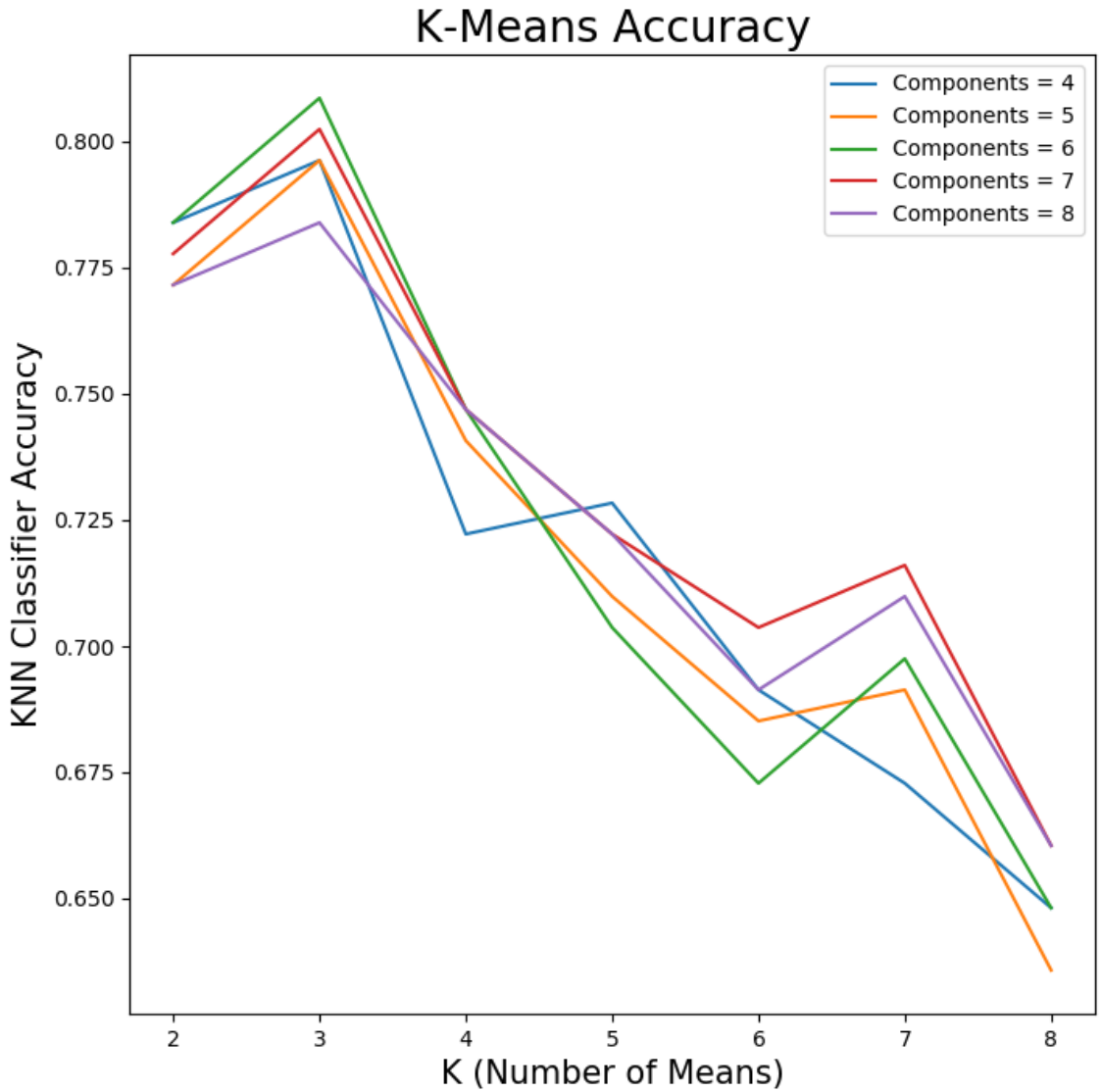
Previously — PCA, Corn, WASDE



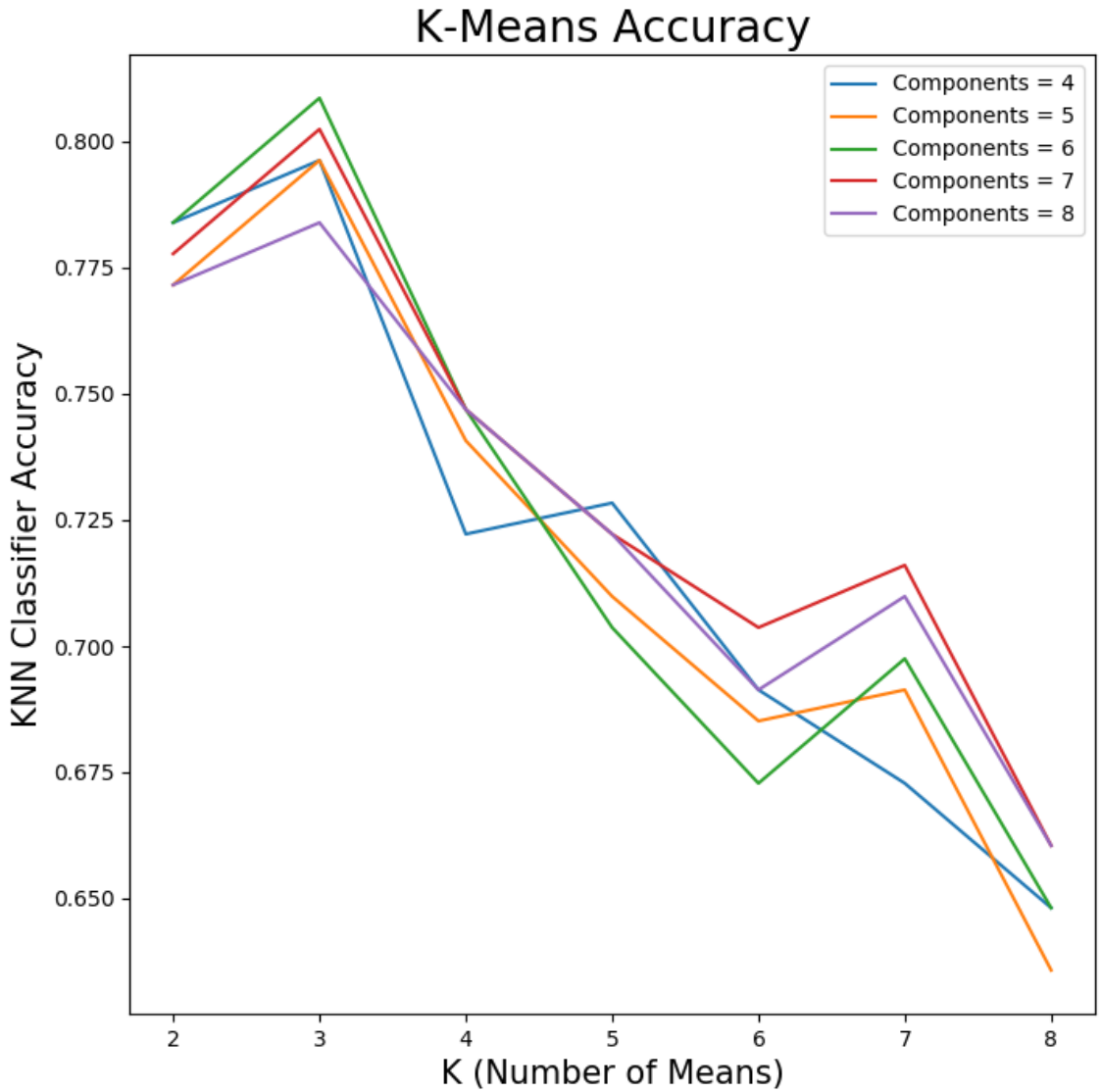
Previously — PCA, Corn, Market



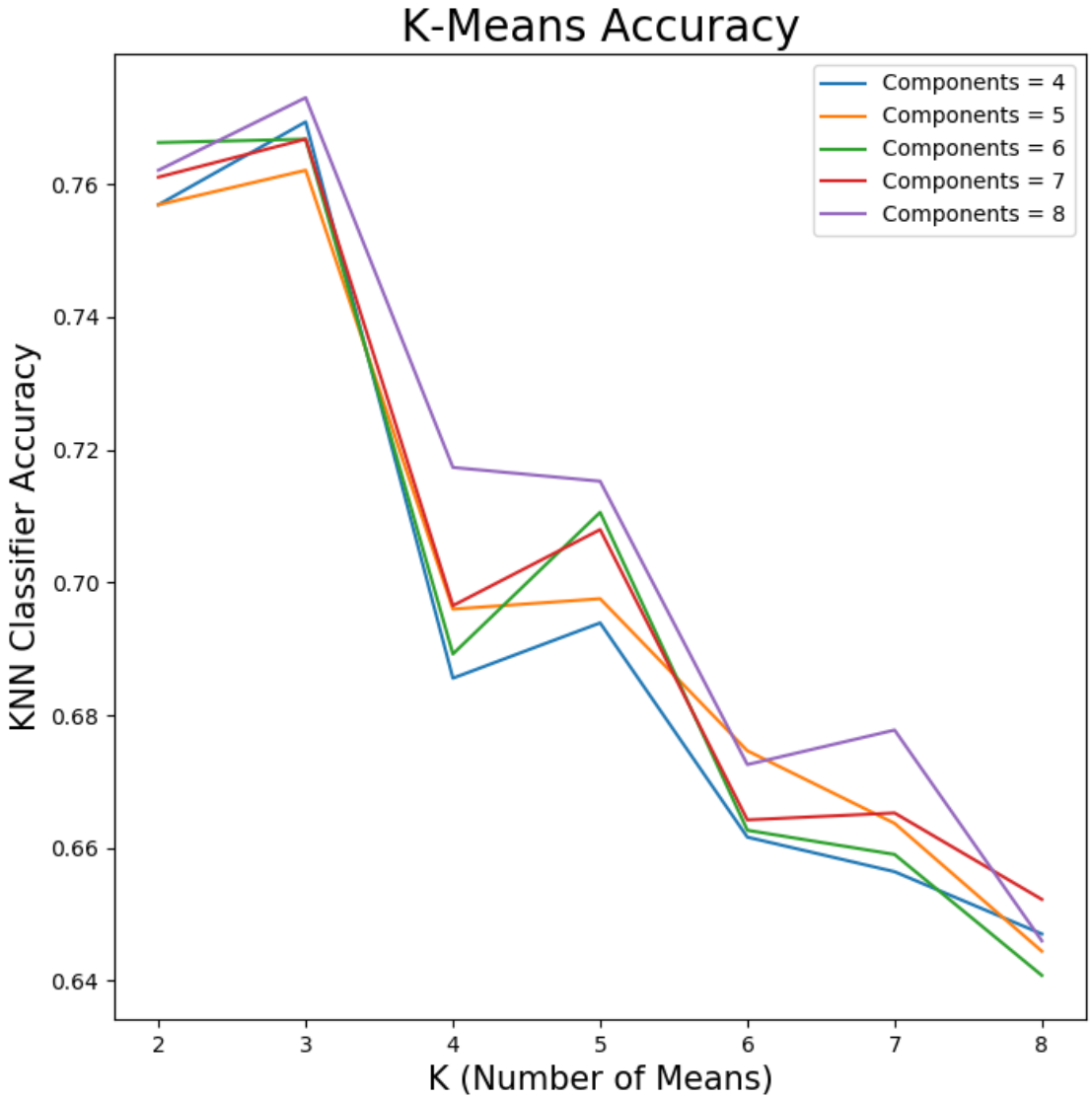
Results — PCA-KNN, Corn, WASDE



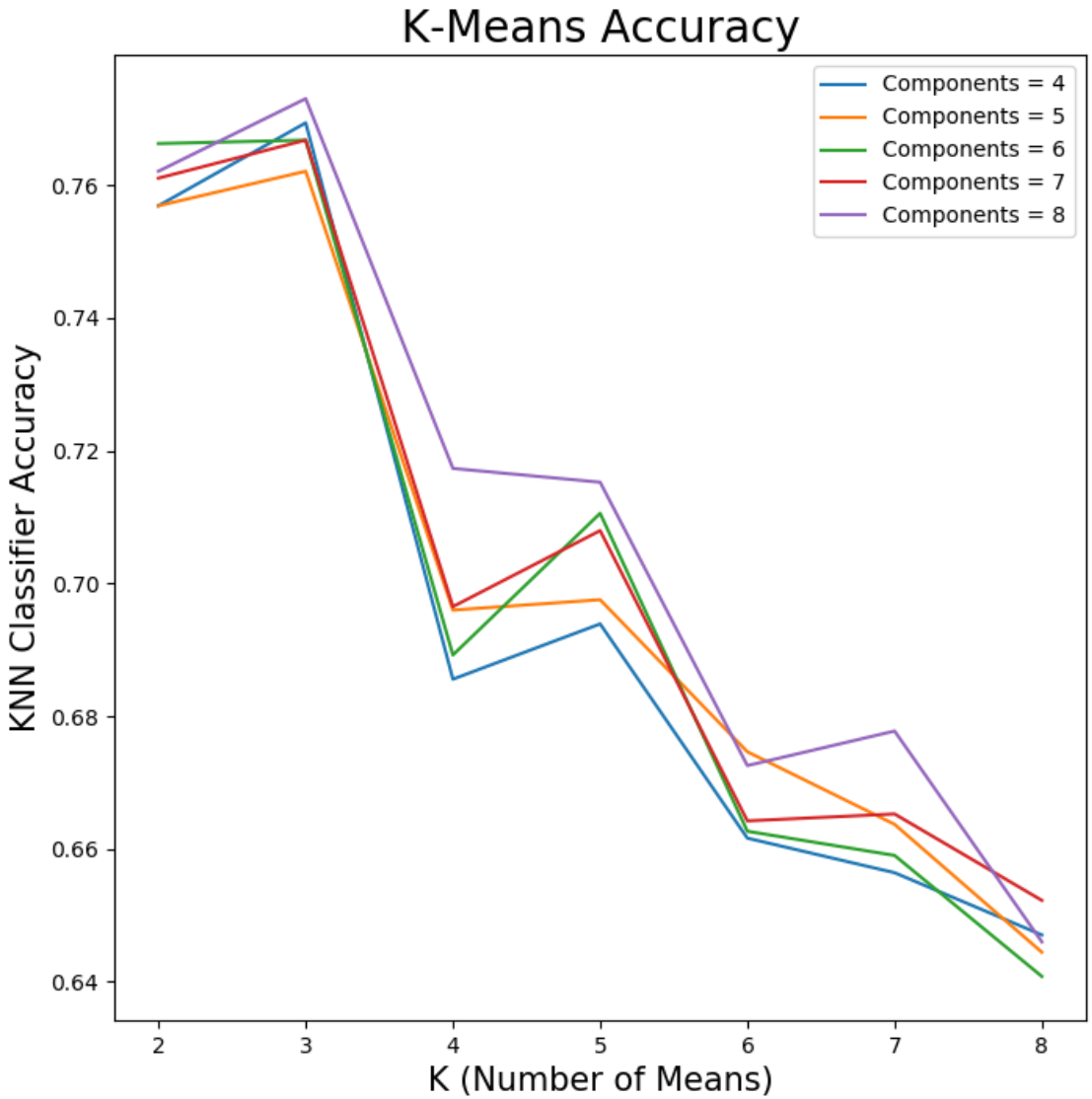
Results — PCA-KNN, Soybeans, WASDE



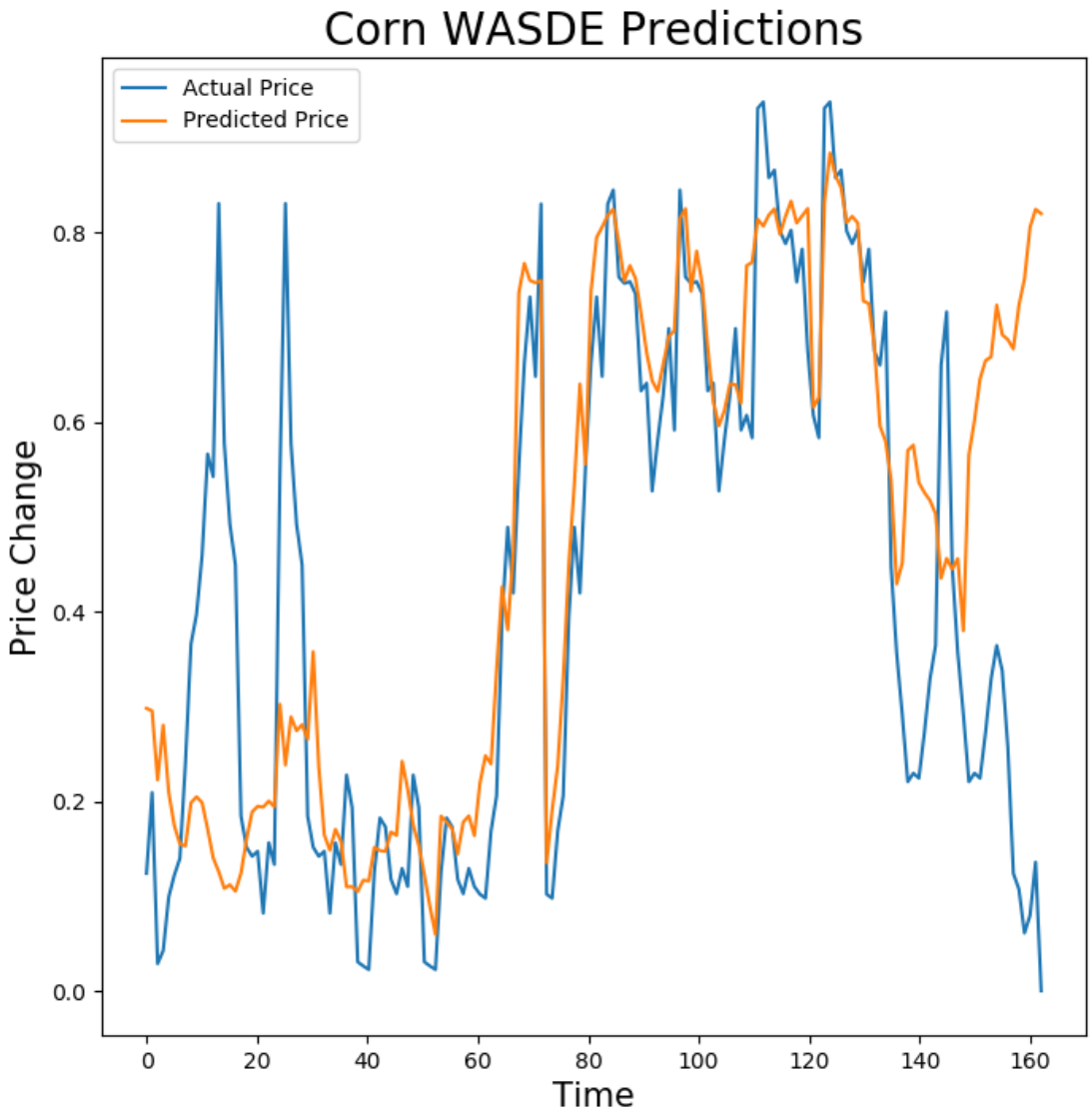
Results — PCA-KNN, Corn, Market



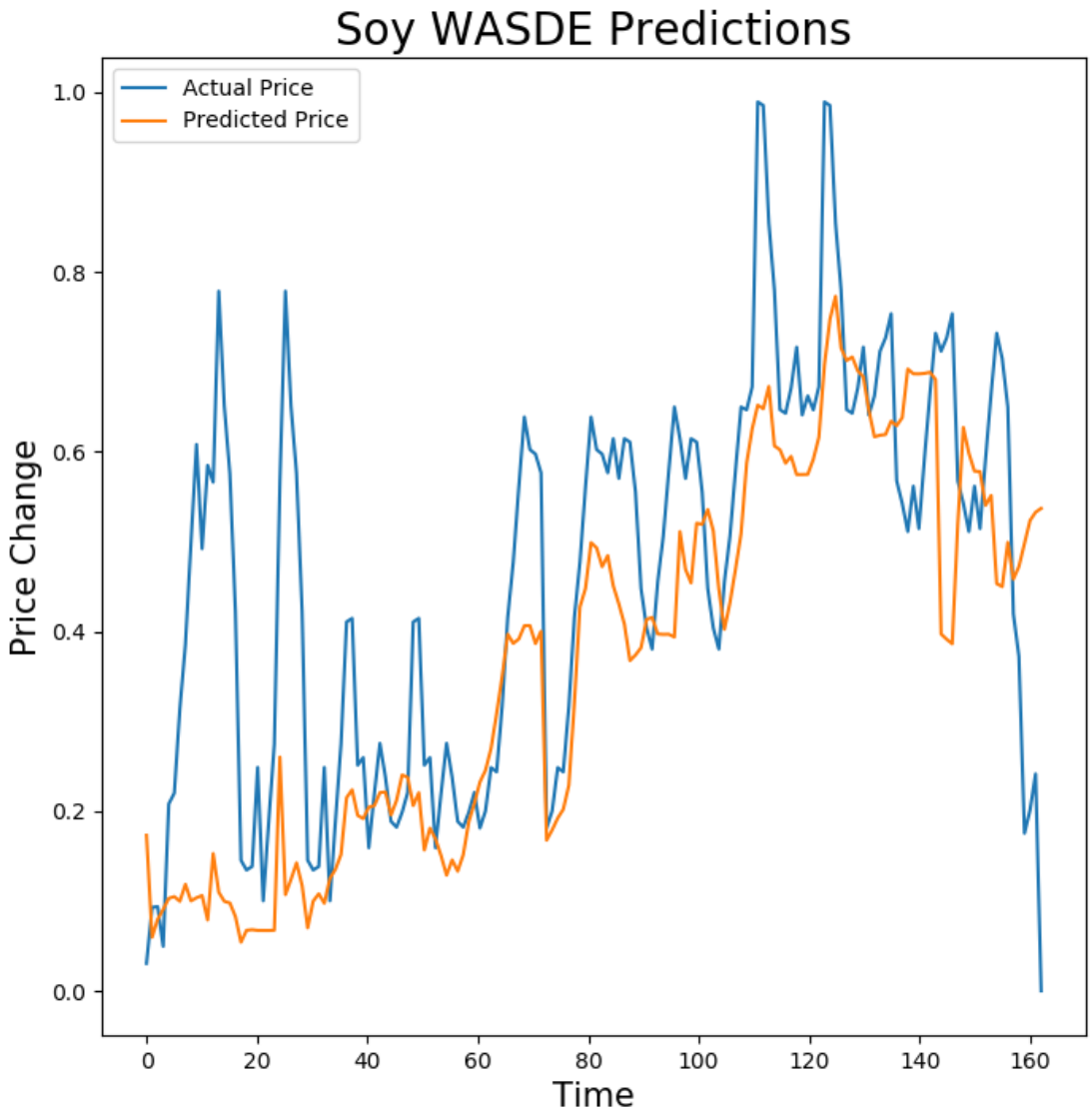
Results — PCA-KNN, Soybeans, Market



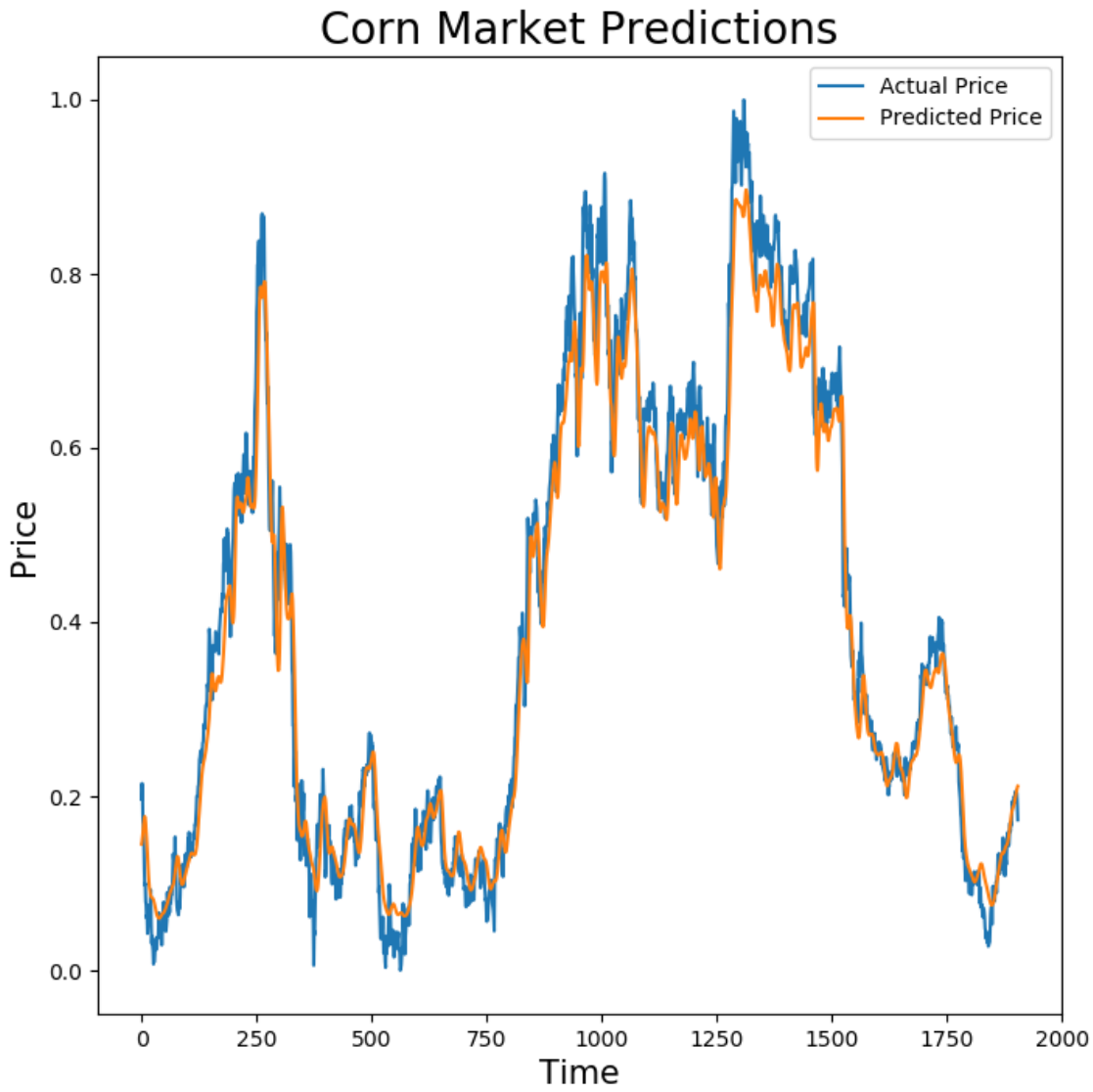
Results — MLP Prediction, Corn, WASDE



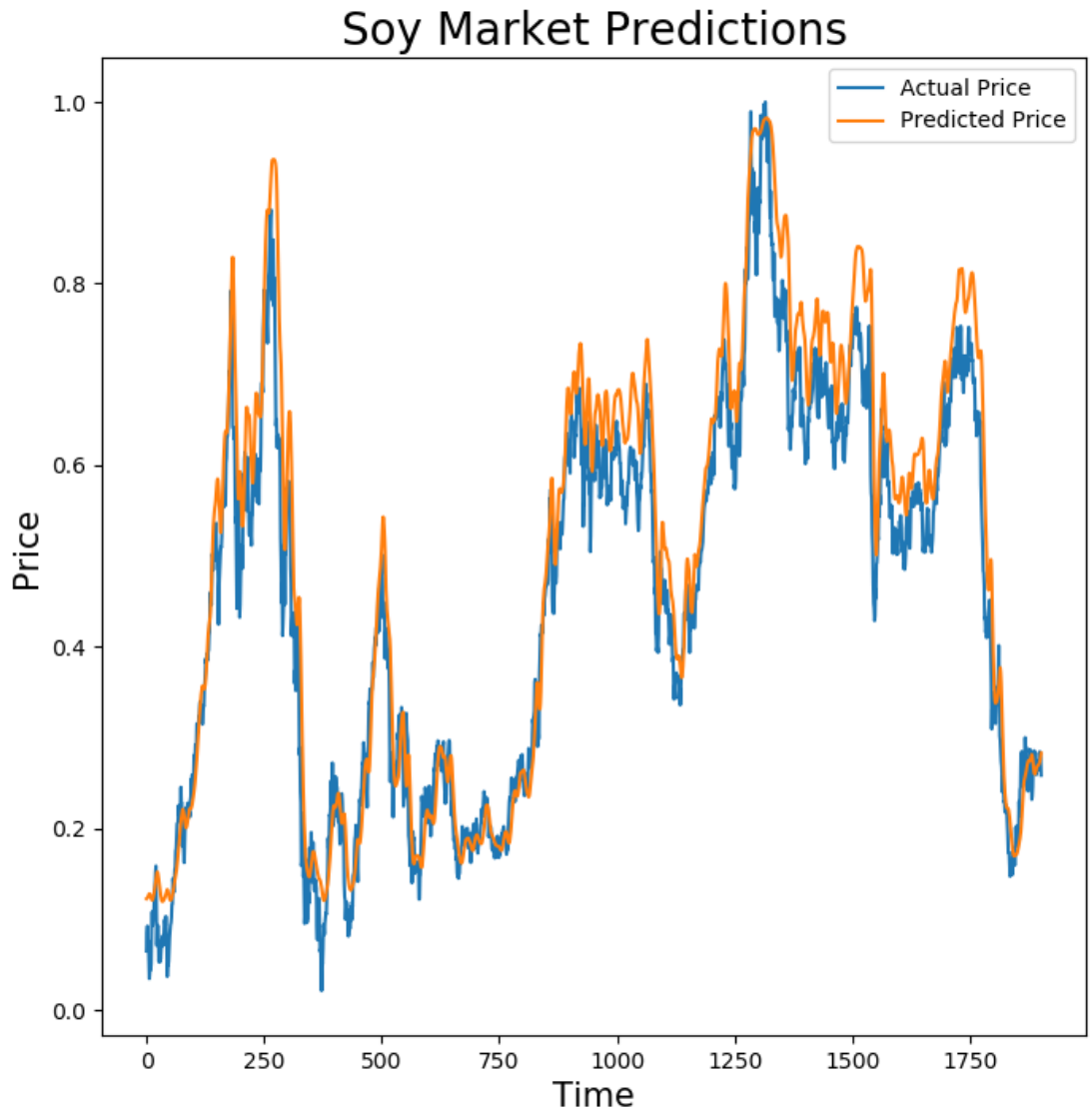
Results — MLP Prediction, Soybeans, WASDE



Results — LSTM Prediction, Corn, Market



Results — LSTM Prediction, Soybeans, Market



Results

- As we can see for the prediction plots, the neural nets are not especially good at guessing the price, but do generally get the trend of the price (up or down) correct.
- If we scan through the predictions, we can see what percentage of the time the predicted price direction matches the actual price direction.

Results

- WASDE Corn MLP Accuracy: 61.3%
- WASDE Soybean MLP Accuracy: 54.8%
- Market Corn LSTM Accuracy: 49.9%
- Market Soybean LSTM Accuracy: 51.2%

Conclusions

- The ability to detect these trends may not be as significant as they appear. The trends may be the result of seasonal trends, which commodities futures contracts often price in, so this knowledge may not be actionable.
- Incorporating futures prices versus spot prices for a contract may serve as a better labeling system, though it would require data I did not have time to gather.
- It is possible that the WASDE data may provide better results under a convolutional model, which will better factor in which elements of the WASDE report are most important for a prediction.

Conclusions

- It is possible that day-long time intervals between market data samples are too long for an LSTM to be able to extract meaningful trends. Markets may simply be too volatile over the course of single day for daily data to provide significant information. It is possible that incorporating intraday pricing will improve performance.
- It is likely that a more sophisticated neural should be able to provide accuracy in the 80% range. In particular, softmax activation functions coupled with convolutional layers may provide comparable performance to PCA-KNN, since convolutional layers have been proved to be equivalent to PCA and softmax may provide a similar categorization effect to KNN.

Questions?

Sources:

Data
https://www.kaggle.com/ainslie/
usda-wasde-monthly-corn-soybean-projections
https://www.quandl.com

Coding
https://github.com/mGalarnyk/Python_Tutorials/blob/master/
Sklearn/PCA/PCA_Data_Visualization_Iris_Dataset_Blog.ipynb