Forecasting Agricultural Commodity Prices through Supervised Learning

Fan Wang, Stanford University, wang420@stanford.edu

ABSTRACT

In this project, we explore the application of supervised learning techniques in predicting the future direction of US corn future prices. We test simple logistic regression, logistic regression with backward feature selection algorithm and support vector machine (SVM). We focus on not only the technical factors of corn future, but also other factors which represent the interrelationship between different commodities. As a result, the testing accuracy of our model reaches more than 75% for 15-day and 20-day returns.

I. INTRODUCTION

Commodity future is an important asset classes in financial markets that have historically demonstrated a high degree of volatility. The Goldman Sachs Commodity Index (an index of 24 of the largest commodity futures) delivered a return of -10.6% p.a. with annual volatility of 23.9% from 2006 to 2015, compared with a 7.3% p.a. return with 15.1% annual volatility for equities (S&P500). Within the commodity future market, agricultural commodities are particularly volatile. This volatility creates challenges for producers and consumers of commodities who aim to hedge price risk, and financial market participants who may seek to diversify multi-asset class portfolios by adding commodities exposure. A statistical approach which can provide insight into the future direction of prices of commodity futures would be of great value to both commercial and financial market participants.

The dataset analyzed in this project is a collection of financial market data: historical time series data of price movements for relevant commodities (corn, crude oil, and soybeans). US corn has the largest agricultural futures market (by number of contract issued), and thus will be the primary focus.

The inputs to our algorithm include various types of technical factors we derive from our dataset. We then use simple logistic regression, logistic regression with backward feature selection algorithm and support vector machine to output the predicted direction (positive or negative) of returns from 5-day to 20-day.

II. RELATED WORK

We begin to study a paper of Ticlavilca, Feuz, and McKee which applies the multivariate Bayesian machine learning regression algorithm in commodity future price forecasting.

They develop the Multivariate Relevance Vector Machine (MVRVM) based multiple-time-ahead (one, two and three month ahead) predictions of monthly agricultural commodity prices. The training sample is the monthly data for cattle, hog and corn prices from 1989 to 2003 and the testing sample is from 2004 to 2009. They use the bootstrapping method to analyze the robustness of the MVRVM and then compare its performance with the performance of Artificial Neural Network (ANN).

Their models show an overall good performance and robustness. The statistical test results also demonstrate the model performs better in one and two month's prediction vs. the three-month prediction.

III. DATASETS, FEATURES AND EXPLORATARY ANALYSIS

The daily price series for 3 commodities - corn, crude oil, and soybeans have been obtained to test if supervised learning techniques can be applied to forecast the price. For each commodity, we have prices for two different future contracts - one is closest to expiry (the "front" month), and the other is expiring in 1 years' time. Table 1 below briefly describes the data.

Table 1: Description of Datasets

Commodity	Contracts	Date[1]
Corn	1-month	1959-07-01 ~ 2016-11-11
	12-month	1968-02-14 ~ 2016-11-11
Crude Oil	1-month	1983-03-30 ~ 2016-11-11
	12-month	1983-03-30 ~ 2016-11-11
Soybeans	1-month	1959-07-01 ~ 2016-11-11
	12-month	1968-12-05 ~ 2016-11-11

The 1-year out (12-month) contract is expressing the market's forecast for where prices are headed and it's expected to show some predictive power of price direction of the 1-month contract. Corn future price and soybeans future

[1] *In order to ensure every price series starts from the same time point, we will use 1983-03-30 as the starting data point to truncate the data.*

price are correlated in so far as they experience similar weather conditions and will have good or bad crop years at the same time. However, farmers also have some choice as to which crop they will plant each year. So, in a year when the price of soybeans has been high relative to the price of corn, it's expected to see some mean reversion the following year as farmers choose to plant more soybeans and less corn given the relative price. Crude oil future price is a good indicator of overall sentiment towards commodities, as well as being an input cost to production of the three grains. Figure 1 below shows the historical charts of the 3 price series: corn, crude oil and soybeans.

Figure 1: Historical Charts of the Price Series[2]

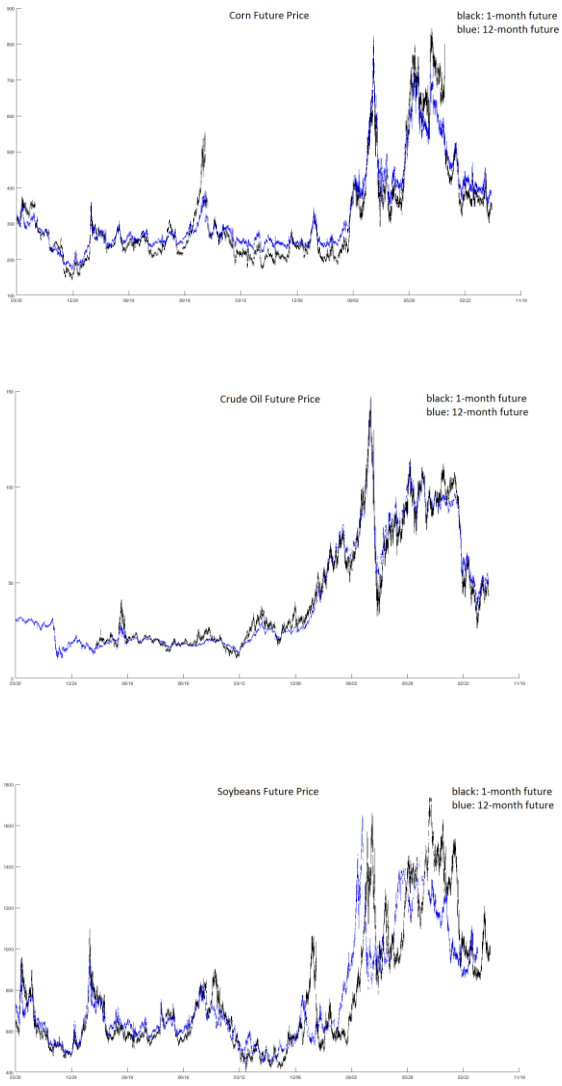


Table 2 shows the Pearson correlation coefficients across all the data samples. We observe the following: 1). 1-month contract and 12-month contract are strongly correlated for the same future; 2). corn is more correlated with soybeans, compared to crude oil; 3). 12-month crude oil contract is slightly more correlated with corn and soybeans (SB), compared to 1-month crude oil contract.

Table 2: Correlations between Different Futures

	Corn 1m	Corn 12m	Oil 1m	Oil 12m	SB 1m	SB 12m
Corn 1m	1.00	0.97	0.77	0.78	0.92	0.94
Corn 12m		1.00	0.86	0.87	0.93	0.97
Oil 1m			1.00	0.99	0.83	0.86
Oil 12m				1.00	0.83	0.87
SB 1m					1.00	0.98
SB 12m						1.00

Focusing on the price of 1-month corn future, we compute the 5-day, 10-day, 15-day, and 20-day positive or negative return (+1 or -1), respectively, as the output(s). In general, we know that the agricultural commodity prices are driven by a wide range of factors such as global economic activity, financial market sentiment, and fundamental factors such as weather, advancements in farming and seed technology, and farmer decision-making. However, since our outputs are short-term based, we decide to limit the feature space to be mainly the technical factors which are computed from the time series dataset.

In order to apply supervised learning techniques, we derive the following several difference types of features:

- % price deviation of 1-month corn future from its 5-day, 10-day, 15-day, and 20-day moving average
- % price difference for 1-month vs. 12-month contract (corn future)
- % price difference for corn vs. soybeans futures
- % price change of crude oil future for 5-day, 10-day, 15-day, and 20-day time window

The reasons of why choose these features and our expectation of the relationship are: 1). if the price deviates too much from moving average, mean reversion tends to happen; 2). 12-month contract tends to lead the direction of 1-month contract; 3). soybeans future may show positive relationship with corn future in short term and negative relationship in long term; 4). crude oil future should have positive relationship with corn future.

IV. METHODS

We now show the definition and computation of model outputs and features. Then we describe the supervised learning techniques applied.

[2] *For crude oil, the 1-month future price and 12-month future price are the same from 1983-03-30 to 1988-12-20. As a result, we will use 1989-12-21 to further truncate the data.*

Computing model outputs[3]

$$\begin{pmatrix} direction_5 \\ direction_{10} \\ direction_{15} \\ direction_{20} \end{pmatrix} = \begin{pmatrix} sign(P_t - P_{(t-5)}) \\ sign(P_t - P_{(t-10)}) \\ sign(P_t - P_{(t-15)}) \\ sign(P_t - P_{(t-20)}) \end{pmatrix}$$

Computing model features

a. The "mean reversion" feature

$$\begin{pmatrix} \%lag\_difference\_MA_5 \\ \%lag\_difference\_MA_{10} \\ \%lag\_difference\_MA_{15} \\ \%lag\_difference\_MA_{20} \end{pmatrix} = \begin{pmatrix} lag_5(\frac{P_t - MA_5}{MA_5}) \\ lag_{10}(\frac{P_t - MA_{10}}{MA_{10}}) \\ lag_{15}(\frac{P_t - MA_{15}}{MA_{15}}) \\ lag_{20}(\frac{P_t - MA_{20}}{MA_{20}}) \end{pmatrix}$$

b. The "1-year out difference" feature

$$\%lag\_k\_difference = lag_k(\frac{P_{t,12\_month\_corn} - P_{t,1\_month\_corn}}{P_{t,1\_month\_corn}})$$

where k =5,10,15 and 20

c. The "corn vs. soybean" feature

$$\%lag\_k\_difference = lag_k(\frac{P_{t,1\_month\_soybeans} - P_{t,1\_month\_corn}}{P_{t,1\_month\_corn}})$$

where k =5,10,15 and 20

d. The "crude oil" feature

$$\%lag\_k\_price\_change = lag_k(\frac{P_{t,1\_month\_crude\_oil} - lag_k(P_{t,1\_month\_crude\_oil})}{lag_k(P_{t,1\_month\_crude\_oil})})$$

where k =5,10,15 and 20

A. Logistic Regression Model

As the most widely used classification technique, logistic regression is our first modeling method.

The hypothesis:

$$h_\theta(x) = \frac{1}{1+e^{\theta^T x}}$$

The cost function:

$$J_\theta = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - h_\theta(x^{(i)}))^2$$

The optimization algorithm:

$$\theta_k := \theta_k - \frac{\partial J}{\partial \theta_k}$$

B. Logistic Regression Model with Backward Selection

The backward selection algorithm can be used together with logistic regression to avoid overfitting. It starts off with the set of all features, and repeatedly deletes features one at a time until only intercept left in the model.

C. Support Vector Machine

Another popular classification method is SVM which solves the optimization problem:

$$\min_{\gamma,\omega,b} \frac{1}{2}\|w\|^2$$
$$s.t.\ y^{(i)}(w^T x^{(i)} + b) \geq 1$$

We apply the RFF kernel in SVM:

$$\exp[-\|x^{(i)} - x^{(j)}\|^2]$$

V. RESULTS AND DISCUSSION

A. Logistic Regression Model

We first train the logistic regression model on *randomly* selected samples from 50% to 90%, and then test the accuracy of prediction on the rest of the sample. Table 3 shows the training and testing accuracy for various size of the sample.

Table 3: Accuracy of Random Sampling

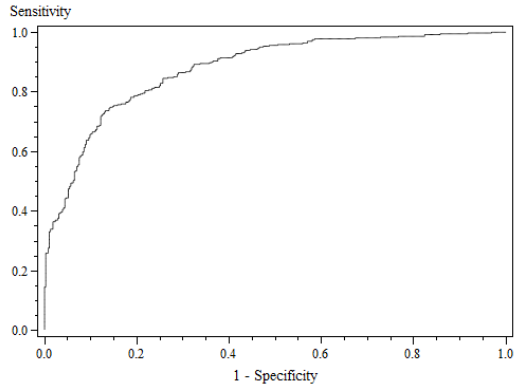
	Size	5-day	10-day	15-day	20-day
Training Set	50%	54.90%	77.10%	85.10%	88.40%
	40%	55.10%	76.90%	85.70%	88.50%
	30%	54.40%	77.20%	85.40%	88.40%
	20%	54.20%	77.50%	85.20%	88.50%
	10%	53.50%	77.30%	85.10%	88.20%
Testing	50%	51.6%	70.6%	76.9%	79.9%

[3] *For the purpose of simplicity, we ignore the "zero" scenario here*.

Set	40%	53.2%	71.0%	77.2%	80.0%
	30%	53.2%	70.8%	77.1%	80.3%
	20%	54.0%	71.6%	78.0%	79.0%
	10%	51.1%	71.4%	77.3%	79.0%

A typical AUC curve with above 75% accuracy is like the following:

Figure 2: AUC of 20-day Return with 90% Training Size



Then we train the model on *sequentially* selected samples from 50% to 90%, and then test the accuracy of prediction on the rest of the sample. Table 4 shows the training and testing accuracy for various size of the sample.

Table 4: Accuracy of Sequentially Sampling

	Size	5-day	10-day	15-day	20-day
Training Set	50%	55.90%	78.10%	85.30%	89.10%
	40%	55.60%	78.40%	85.30%	88.80%
	30%	54.40%	78.30%	85.50%	88.80%
	20%	54.80%	78.20%	85.40%	88.70%
	10%	53.80%	77.80%	85.20%	88.50%
Testing Set	50%	50.1%	76.9%	75.2%	78.5%
	40%	49.2%	69.1%	76.3%	78.9%
	30%	50.2%	69.0%	75.5%	79.3%
	20%	49.0%	67.6%	74.7%	78.7%
	10%	49.9%	66.9%	73.3%	76.5%

We observe our models perform poorly on models of 5-day return. When the accuracy is close to 50% and sometimes less than 50%, it's no better than pure guessing. From the accuracy of training sample, we also see that model built on sequentially selected sample is marginal better than the randomly selected sample. To some extent, this is expected since the market moves in trend. Because of this, we will forgo the randomly selection scheme (and/or cross validation) and use the sequential selection as the only sampling method.

B. Logistic Regression Model with Backward Selection

To avoid overfitting, we apply backward selection algorithm together with logistic regression to control the number of selected features. Table 5 shows the testing accuracy for various size of the sample. While the accuracy is comparable to simple logistic regression, we find the backward feature selection algorithm performs well on models of short-term returns (i.e., the number of selected feature shrink), but performs poorly on long-term return models (i.e., the number of selected features does not shrink).

Table 5: Accuracy of Logistic Regression with Backward Selection and Sequentially Sampling

	Size	5-day	10-day	15-day	20-day
Testing Set	50%	51.04%	68.69%	75.41%	78.64%
	40%	50.15%	68.87%	75.89%	78.56%
	30%	50.45%	68.45%	75.19%	78.92%
	20%	49.81%	68.09%	74.28%	78.18%
	10%	48.42%	67.02%	72.50%	75.72%

C. Support Vector Machine

Our last tried classification technique is SVM. Table 6 shows the testing accuracy for various size of the sample.

Table 6: Accuracy of Support Vector Machine with Sequentially Sampling

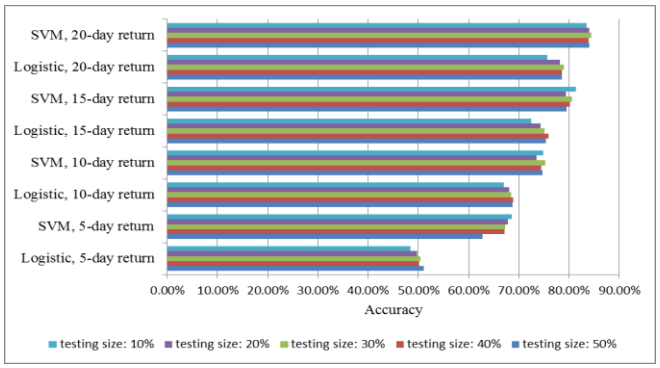
	Size	5-day	10-day	15-day	20-day
Testing Set	50%	62.81%	74.71%	79.56%	83.97%
	40%	67.18%	74.45%	80.22%	83.85%
	30%	67.23%	75.27%	80.57%	84.47%

	20%	67.79%	73.56%	79.33%	83.97%
	10%	68.67%	74.81%	81.41%	83.51%

D. Summary

Figure 3 below summarizes the comparison of performance between logistic regression and SVM.

Figure 3: Accuracy: Logistic Regression vs. SVM



VI. CONCLUSION AND FUTURE WORK

A. Conclusion

Our analysis shows that technical factors of 1-month corn future prices together with other technical factors that represent the interrelationships with related commodities can be a powerful set of predictive features. The accuracy results show an overall good performance of both logistic regression and SVM model. Two noticeable things are: 1). predictions of 20-day's and 15-day's return are more accurate than 10-day's and 5-days', which is in contradiction to the old research paper; 2). SVM models perform better than logistic regression model in every testing size sample.

B. Future Work

Moving forward, the economic or financial relationship (i.e., positive or negative relationship) between corn future return and different features should be taken into consideration when building logistic regression model. Additionally, SVM models with different kernels and ensemble methods should be explored to improve the testing sample accuracy. Moreover, bootstrapping method should be applied to test the stability and robustness of different models.

VII. REFERENCES

[1] A. M. Ticlavilca, D. M. Feuz, and M. McKee, "Forecasting Agricultural Commodity Prices Using Multivariate Bayesian Machine Learning Regression", *Applied Commodity Price Analysis, Forecasting and Market Risk Management*, 2010.

[2] D. Huang, F. Jiang, and J. Tu, "Mean Reversion, Momentum and Return Predictability," 2013, unpublished.

[3] C. A. Kase, "How Well Do Traditional Momentum Indicators Work?" 2006.

[4] C. Zhu, K. He, Y. Zou and K. K. Lai, " Day-Ahead Crude Oil Price Forecasting Using a Novel Morphological Component Analysis Based Model", *The Scientific World Journal*, 2014

[5] S. S. Patil, Prof. K. Patidar and Asst. Prof. M. Jain, "A Survey on Stock Market Prediction Using SVM", *International Journal of Current Trends in Engineering & Technology*, 2016.

[6] R, https://cran.r-project.org/

[7] SAS, http://www.sas.com/en_us/home.html

[8] Scikit Learn, http://scikit-learn.org/