Predicting the Direction of Exchange-Traded Fund (ETF) Movement

Kai Liu
The University of Texas at Austin

April 2017

1 Summary

In this project, I used machine learning techniques to predict the direction of the following ETFs: XLE, XLU, XLK, XLB, XLP, XLY, XLI, XLV and SPY. In the modeling process, two different types of input predictors are used in predicting the directions. One type of predictors are OHLC prices, volume size and weekday information, which is called pure data-driven method in the report. The other types of predictors are weekday information and some technical indicators chosen from literature, referred as indicator-driven method. In addition, four different models are implemented: logistic regression, Lasso regression, Ridge regression, and artificial neural network. Besides, though I attempted to predict the direction, I also converted the binary dependent variable into other two types to compare which type of dependent variable can achieve the best performance. In conclusion, the accuracy of the best model for each ETF sector ranges from 55% to 60%. And the trading strategy based on my prediction earns less than the all long-only strategy, but it has a lower risk to achieve the earning.

2 Method

2.1 Data

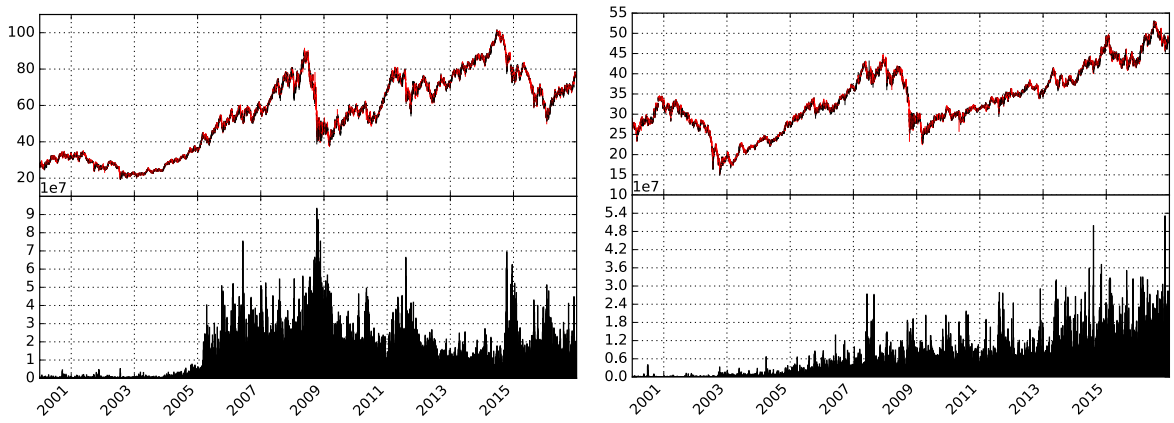
In the project, I attempted to predict the direction of the following ETFs: XLE, XLU, XLK, XLB, XLP, XLY, XLI, XLV, and SPY. The daily OHLC (Open, High, Low, Close) prices and volumes for these ETFs are obtained from Yahoo Finance. The total number of samples is 4,277 trading days, from January 1st 2000 to December 31st 2016. The total 4,277 data points of the daily OHLC prices and volumes of the first two ETF sectors are plotted in Figure 1. I divide each dataset into two parts, 94.1% of the data (January 1st 2000 to December 31st 2015) is used from in-sample training and 5.9% (January 1st 2016 to December 31st 2016) are considered as out-of-sample data. The in-sample data is used to train the models, while out-of-sample is reserved to evaluate performance of models.

2.2 Predictors

In this project, I explored two different types of methods to predict the direction of ETF movement: pure data-driven method, and indicator-driven method.

Pure data-driven method. 1-day lagged version of all the OHLC prices, volume data and weekday information is used to predict the market direction of the second day. For example, if we would like to predict the market direction on 11/30/2010, the predictors are OHLC prices and volume data on 11/29/2010. Weekday information is treated as dummy variables.

Indicator-driven method. Based on previous studies, it is hypothesized that various technical indicators can be used as predictors to construct models to forecast the direction of ETF price. After reviewing prior publications[1, 5, 4, 2, 3], 16 technical indicators are chosen as predictor in constructing prediction models in this project. Table 1 lists selected technical indicators and parameters used. For some technical indicators, i.e. EWMV and Disparity Index, multiple parameters (days) are used to calculate these technical



(a) XLE (b) XLU

Figure 1: Raw data visualization

Table 1: Selected technical indicators and their parameters

Technical indicators	Parameters
Exponential Weighted Moving Average (EWMA)	7, 50, 200 days
Moving Average Convergence/Divergence (MACD)	12&26 days
Relative Strength Index (RSI)	14 days
Average Directional Index (ADX)	14 days
Fast Stochastic Oscillator %K	14 days
Fast Stochastic Oscillator %D	14 days
Slow Stochastic Oscillator %K	3 days
Momentum	4 days
Acceleration	5&34&5 days
William's R	14 days
Accumulation/Distribution	NA
Chaikin Oscillator	3&10 days
William Accumulation/Distribution	NA
On Balance Volume	NA
Disparity Index	5, 10 days
Commodity Channel Index (CCI)	14 days

indicators. Therefore, there are 19 input techinical indicators in total.Weekday information is also included in this method.

2.3 Prediction Models

The purpose of the project is to predict the direction of ETF movement. The direction is determined by Close and Open prices of the same day. The relationship between Close and Open prices is transformed to three different types in order to be predicted in different models:

- **Type I: binary.** If the close price is higher that the open price of the same day, the direction is labeled 1. Otherwise labeled 0. In order to predict binary data, classification models need to be implemented.
- **Type II: Percentage.** The equation to calculate the percentage is $Percentage = \frac{P_{close} - P_{Open}}{P_{Open}}$. It can be predicted using regression models. Then the percentage can be converted to directions: if Percentage is positive, it means the price is going up. Otherwise, it is going down.

- **Type III: Ratio.** The equation to calculate the ratio is $Ratio = \frac{P_{close}}{P_{Open}}$. It can be predicted using regression models as well. Then the ratio can be converted back to directions: if Ratio ¿ 1, it means the price is going up. Otherwise, it is going down.

The following four types of models are implemented in the project based on different data types of the dependent variable.

Logistic Regression. The logistic regression model can be used to predict the direction directly. I consider both L1 and L2 penalties in the model. The difference between L1 and L2 penalty is that L1 penalty can yield a sparse coefficient vector while L2 penalty cannot. However, the limitation for the L1 penalty is that if the predictors have high correlation, the L1 penalty can only select one of them. This can make the model miss useful information and thus has a negative effect on the prediction accuracy.

Hyerparameter search space in this model:

- C(L1 regularization term): {2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100}
- C(L2 regularization term): {2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100}

Lasso Regression. Lasso regression model is a regression model using L1 regularization. It is applied to predict the ratio/percentage.

Hyerparameter search space in this model:

- α(L1 regularization term): {0.1, 1, 5, 10, 30, 50, 100}

Ridge Regression. Ridge regression is a regression model using L2 regularization. It is also applied to predict the ratio/percentage.

Hyerparameter search space in this model:

- α(L1 regularization term): {0.1, 1, 5, 10, 30, 50, 100}

Artificial Neural Network. The artificial neural network implemented here is a multiple perceptron (MLP). It maps sets of input data onto a set of outputs. The MLP model in this project contains an input layer, a hidden layer, and an output layer, each of which is connected to the other in the same sequence as listed above. The input layer corresponds to the input predictors. The hidden layers is used for capturing the nonlinear relationships among predictors. The output layer consists of only one neuron that represents the predicted direction/ratio/percentage.

Hyerparameter search space in the MLP model:

- α(L2 regularization term): {0.1, 1, 5, 10, 30, 50, 100}
- Hidden layers: {(9), (6), (3)}

Figure 2 summarizes the combinations of input predictors, the dependent variable and models. In total, there are 18 combinations.

2.4 Performance Measurement

In the project, I am only interested in forecasting the direction of the ETF movement. Therefore, after doing prediction using the percentage/ratio dependent variable, I transform the predicted results to binary results as well. That is, if the predicted percentage is positive, the direction label is 1; otherwise, it is 0. If the predicted ratio is large than 1, the direction label is 1; otherwise, it is 0.

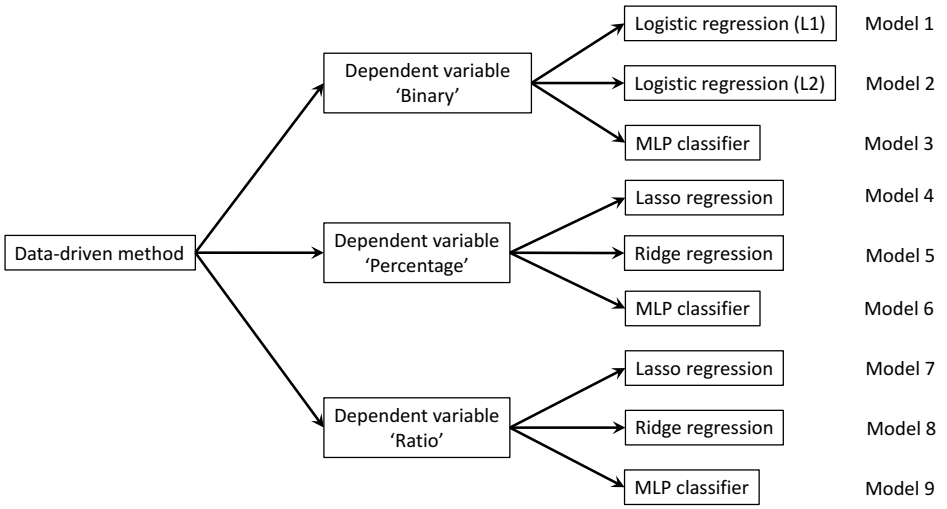
The prediction performance is evaluated by using the following equations:

$$P_t = \begin{cases} 1, y_t - \hat{y_t} = 0 \\ 0, y_t - \hat{y_t} \neq 0 \end{cases} \tag{1}$$

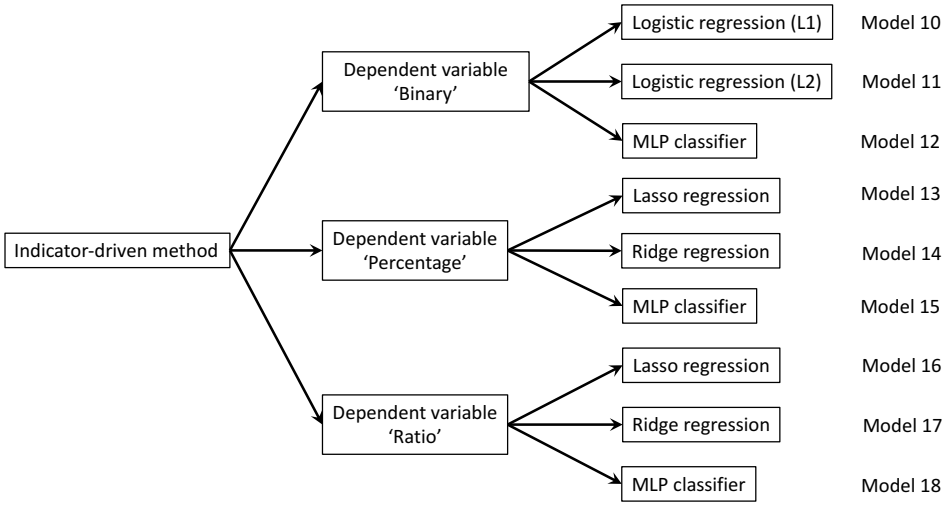
where y_t is the actual direction of the ETF movement for the ith trading day, while $\hat{y_t}$ donates the predicted direction for the ith trading day.

$$Accuracy = \frac{1}{N} \sum_{t=1}^{N} P_t, t = 1, 2, 3, \ldots, n$$

where N is the total number of predicted results.



(a) Data-driven method



(b) Indication-driven method

Figure 2: The combinations of input predictors, the dependent variable and models

2.5 Prediction Process

After retrieving ETF data from Yahoo Finance and calculating input predictors. I plug in the data into the above 18 models to forecast the future direction of ETF, and compare the performance of each model. I conduct the prediction process as follows:

1. Calculate all indicators using all data from January 1st 2000 to December 31st 2016. And discard all samples including NaN.

2. Divide data into training data and test data as described in Data section.

3. Normalize data by removing the mean value of each predictor, and scale it by dividing by its standard deviation. In order to avoid incorporating information from test data into training process, I only use the mean value and standard deviation of training data to normlize all data.

4. Use training data to train models. In order to tune hyerparameters in models and avoid over-fitting, 10-fold crossing validation is implemented in training process. Since the data is time series and indicator

calculation depends on prices and volume size of passing days, I use forward chaining procedure to avoid incorporating test data into training process. The forward chaining procedure is like:

(a) fold 1: training [1], test [2]

(b) fold 2: training [1 2], test [3]

...

(c) fold 9: training [1 2 3 ... 9], test [10]

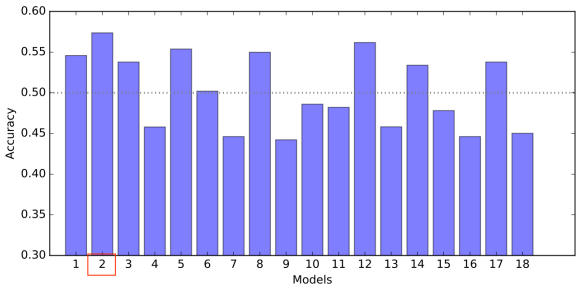
5. Predict the next day's ETF movement using each model with the best hyerparameters, and compare their performance.

3 Experimental Results

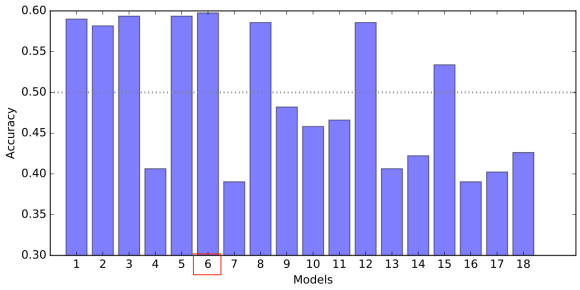
3.1 Prediction Accuracy

To evaluate the performance of each model, I used out-of-sample data which ranges from January 1st 2016 to December 31st 2016. Figure 3 shows the performance of each model on each ETF sector. The model with best accuracy for each dataset is marked with a square.

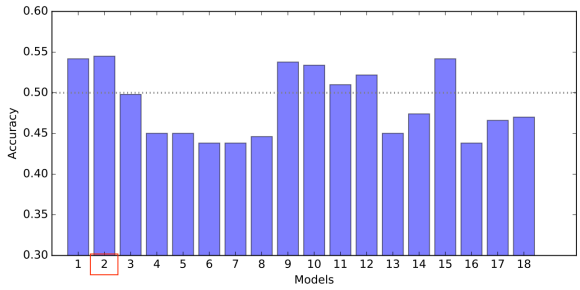
It shows that the best prediction accuracy of each ETF sector ranges from 0.55 to 0.6, which means at least one predictive model can gain some information in forecasting the direction of ETF movement. For some ETF sectors, such as XLE, XLU, XLK, XLY, XLV, the pure data-driven method can achieve a better result, while the indicator-driven method is superior in predicting some other ETF sectors' movement, including XLB, XLP, and SPY. In addition, the performance has no difference in forecasting XLI movement by either data-driven or indicator-driven method. Moreover, predicting either Binary or Percentage/Ratio can achieve good accuracy depending on the dataset. The prediction accuracy has no preference for MLP or regression models.



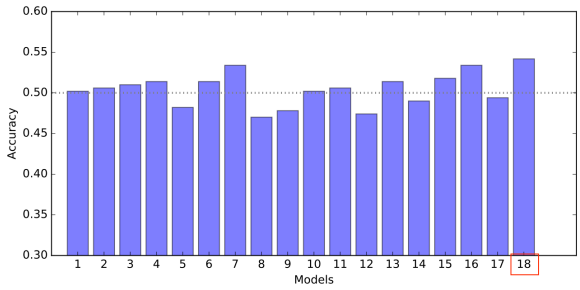
(a) XLE



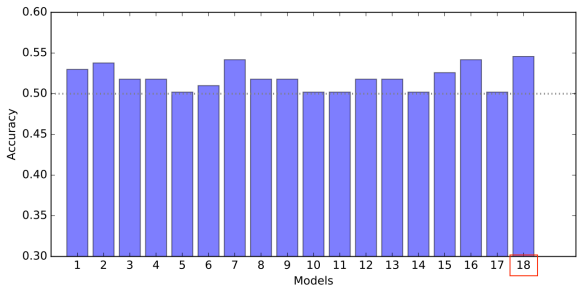
(b) XLU



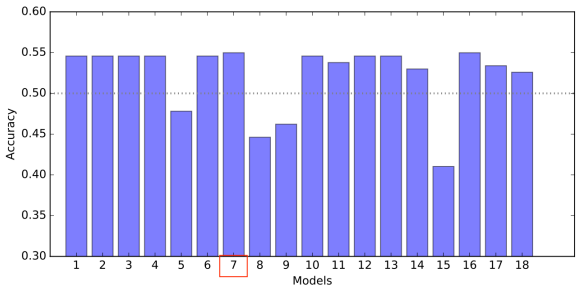
(c) XLK



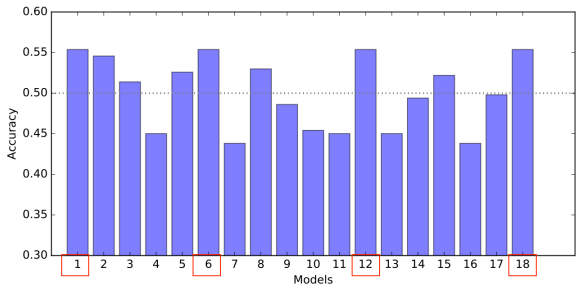
(d) XLB



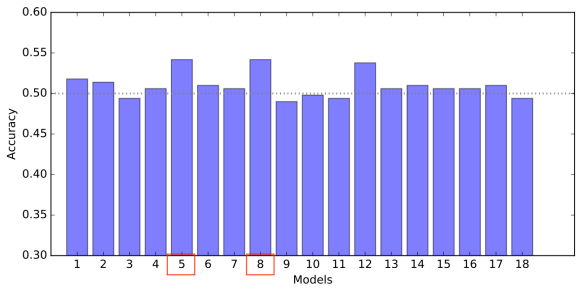
(e) XLP



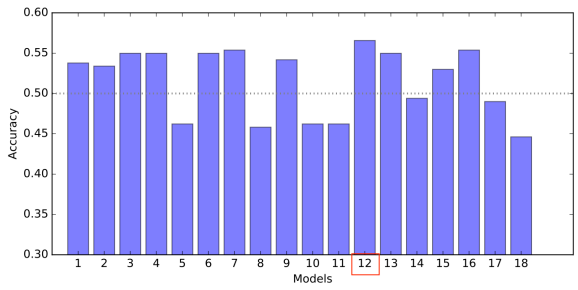
(f) XLY



(g) XLI



(h) XLV



(i) SPY

Figure 3: The performance of each model on ETF forecasting (Continue)

Table 2: The best predicting models and hyerparameter setting

Dataset	Model	Hyerparameters
XLE	Model 2	5
XLU	Model 6	(3), 5
XLK	Model 2	2
XLB	Model 18	(6), 1
XLP	Model 18	(9), 0.1
XLY	Model 7	100
XLI	Model 1	100
XLV	Model 5	0.1
SPY	Model 12	(6), 1

Figure 4: Equity curves for each of the three strategies

3.2 Strategy Evaluation

From the previous analysis, I determined the best predicting model for each ETF dataset. In this section, I use the best models to evaluate different trading strategies. Table 2 lists models used for each ETF sector data, and their hyerparameter setting. Figure 4 shows the cumulative daily portfolio P&L of each strategy. Apparently, the strategy based on prediction is better than SPY long-only strategy, which can earn $0.971 more at the end of 2016. However, it is worse than all long-only strategy, and can earn $0.547 less money at the end of 2016. Table 3 shows the annualized Sharpe ratio for each strategy. The strategy based on my prediction is 3.29, which is higher than that of other two strategies. Based on previous studies, a ratio of 3 or higher is considered excellent. It indicates that the strategy based on my prediction takes much less risk to achieve the return, though the return is a little lower than that of all long-only strategy.

Table 3: The annualized Sharpe ratio for the three strategies

Strategy	Strategy based on prediction	SPY long-only	All long-only
Annualized sharpe ratio	3.29	1.59	2.53

4 Conclusion

In the project, I applied four different models to predict the direction of nine ETF sectors movement. These models used either different types of dependent variables or different input predictors. The best predicting accuracy for each ETF sector ranges from 55% to 60%. In addition, the best models for the nine sectors are different from one another. Therefore, in practice, we should personalize predictive models for each dataset.

The prediction performance of the models in the project may be improved further using three methods. The first approach is to use a subset of these input predictors instead of including all of them in models, especially when using technical indicators as input predictors. Therefore, optimization algorithms, such as forward selection, and randomized Lasso/logistic, can be applied to select predictors with the best prediction power. Besides, other models, including SVM, ensemble method, can be tested for each dataset. At last, in the project, I used the percentage of correctly predicted directions to evaluate the performance of each model. In the future, the trading strategy (earning) based evaluation can be implemented.

References

[1] Can machine learning techniques be used to predict market direction? - the 1,000,000 model test. `http://www.jonathankinlay.com/Articles/ONE%20MILLION%20MODELS.pdf`. [Accessed: 2017-04-14].

[2] Luckyson Khaidem, Snehanshu Saha, S. R. D. Predicting the direction of stock market prices using random forest. *arXiv* (2016), 1605.00003v1.

[3] Masoud, N. Predicting direction of stock prices index movement using artificial neural networks: The case of libyan financial market. *British Journal of Economics, Management & Trade 4* (2014), 597–619.

[4] Mingyue Qiu, Y. S. Predicting the direction of stock market index movement using an optimized artificial neural network model. *PLOSone 11* (2016), e0155133.

[5] Predicting stock markets with neural networks. `https://www.duo.uio.no/bitstream/handle/10852/44765/aamodt-master.pdf?sequence=7`. [Accessed: 2017-04-15].