Journal of
*Risk and Financial Management*

Article

# A Random Forests Approach to Predicting Clean Energy Stock Prices

**Perry Sadorsky**

Schulich School of Business, York University, Toronto, ON M3J 1P3, Canada; psadorsky@schulich.yorku.ca

**Abstract:** Climate change, green consumers, energy security, fossil fuel divestment, and technological innovation are powerful forces shaping an increased interest towards investing in companies that specialize in clean energy. Well informed investors need reliable methods for predicting the stock prices of clean energy companies. While the existing literature on forecasting stock prices shows how difficult it is to predict stock prices, there is evidence that predicting stock price direction is more successful than predicting actual stock prices. This paper uses the machine learning method of random forests to predict the stock price direction of clean energy exchange traded funds. Some well-known technical indicators are used as features. Decision tree bagging and random forests predictions of stock price direction are more accurate than those obtained from logit models. For a 20-day forecast horizon, tree bagging and random forests methods produce accuracy rates of between 85% and 90% while logit models produce accuracy rates of between 55% and 60%. Tree bagging and random forests are easy to understand and estimate and are useful methods for forecasting the stock price direction of clean energy stocks.

**Keywords:** clean energy stock prices; forecasting; machine learning; random forests

check for updates

**Citation:** Sadorsky, Perry. 2021. A Random Forests Approach to Predicting Clean Energy Stock Prices. *Journal of Risk and Financial Management* 14: 48. https://doi.org/10.3390/jrfm14020048

Academic Editor: Jong-Min Kim

Received: 31 December 2020
Accepted: 21 January 2021
Published: 24 January 2021

## 1. Introduction

Climate change, green consumers, energy security, fossil fuel divestment, and technological innovation are powerful forces shaping an increased interest towards investing in companies that specialize in clean energy, broadly defined as energy produced from renewable energy sources like biomass, geothermal, hydro, wind, wave, and solar. Through technological innovation, the levelized cost of electricity is falling for renewables and onshore wind is now less costly than coal (The Economist 2020). Investment in clean energy equities totaled $6.6 billion in 2019. While this number was below the record high of $19.7 billion in 2017, the compound annual growth rate between 2004 and 2019 of 24% was above that of clean energy private equity or venture capital funding (Frankfurt School-UNEP Centre/BNEF 2020).

Well informed investors need reliable methods for predicting the stock prices of clean energy companies. There is, however, a noticeable lack of information on the prediction of clean energy stock prices. This is the gap in the literature that this paper fills. There are, however, two complicating issues. First, predicting stock prices is fraught with difficulty and the prevailing wisdom in most academic circles, consistent with the efficient markets hypothesis, has generally been that stock prices are unpredictable (Malkiel 2003). More recently, momentum and economic or psychology behavior factors have been identified as possible sources of stock price predictability (Gray and Vogel 2016; Lo et al. 2000; Christoffersen and Diebold 2006; Moskowitz et al. 2012). In addition, the existing literature on stock price predictability shows that predicting stock price direction is more successful than predicting actual stock prices (Basak et al. 2019; Leung et al. 2000; Nyberg 2011; Nyberg and Pönkä 2016; Pönkä 2016; Ballings et al. 2015; Lohrmann and Luukka 2019).

Second, regression based approaches for predicting stock prices or approaches relying solely on technical indicators provides mixed results (Park and Irwin 2007) but machine learning (ML) methods appear to offer better accuracy (Shah et al. 2019; Ghoddusi et al. 2019; Khan et al. 2020; Atsalakis and Valavanis 2009; Henrique et al. 2019). There are many different types of ML methods but decision tree bagging and random forests (RFs) are easy to understand and motivate and they also tend to have good performance for predicting stock prices (Basak et al. 2019; Khan et al. 2020; Lohrmann and Luukka 2019). Decision trees are a nonparametric supervised leaning method for classification and regression. A decision tree classifier works by using decision rules on the data set features (predictors or explanatory variables) to predict a target variable (James et al. 2013). Decision trees are easy to understand and visualize and work on numerical and categorical data (Mullainathan and Spiess 2017). Decision tree learning can, however, create complicated trees the results of which are susceptible to small changes in data. Bootstrap aggregation, or bagging as it is commonly known as, is one way to reduce the variance of decision trees. Bootstrap replication is used to create many bootstrap training data sets. Even though each tree is grown deep and has high variance, averaging the predictions from these bootstrap trees reduces variance. RFs are ensembles of decision trees and work by introducing decorrelation between the trees by randomly selecting a small set of predictors at each split of the tree (James et al. 2013).

Here are some recent examples of papers that use RFs for stock price prediction. Ampomah et al. (2020) compare the performance of several tree-based ensemble methods (RFs, XGBoost, Bagging, AdaBoost, Extra Trees, and Voting Classifier) in predicting the direction of stock price change for data from three US stock exchanges. The accuracy for each model was good and ranged between 82% and 90%. The Extra Trees method produced the highest accuracy on average. Ballings et al. (2015) point out that among papers that predict stock price direction with machine learning methods, artificial neural networks (ANNs) and support vector machines (SVMs) are more popular than RFs. Only 3 out of the 33 papers that they survey used RFs. It is not clear why more complicated methods like ANN and SVM are preferred over simpler methods like RFs. Using data on 5767 European listed companies, they compare the stock price direction predictive performance of RFs, SVMs, AdaBoost, ANNs, K-nearest neighbor and logistic regression. Feature selection is based on company specific fundamental data. They find strong evidence that ensemble methods like RFs have greater prediction accuracy over a one-year prediction period. Basak et al. (2019) use RFs to predict stock price direction for 10 companies, most of which are technology or social media oriented (AAPL, AMZN, FB, MSFT, TWTR). Feature selection is based on technical indicators. They find the predictive accuracy of RFs and XGBoost to be higher than that of artificial neural networks, support vector machines, and logit models. Khan et al. (2020) use 12 machine learning methods applied to social media and financial data to predict stock prices (3 stock exchanges and 8 US technology companies). The RFs method is consistently ranked as one of the best methods. Lohrmann and Luukka (2019) use RFs to predict the classification of S & P 500 stocks. Stock price direction is based on a four-class structure that depends upon the difference between the open and close stock prices. Feature selection is based on technical indicators. They find that the RFs classifier produces better trading strategies than a buy and hold strategy. Mallqui and Fernandes (2019) find that a combination of recurrent neural networks and a tree classifier to be better at predicting Bitcoin price direction than SVM. Mokoaleli-Mokoteli et al. (2019) study how several ML ensemble methods like boosted, bagged, RUS-boosted, subspace disc, and subspace k-nearest neighbor (KNN) compare in predicting the stock price direction of the Johannesburg Stock Exchange. They find that Boosted methods outperform KNN, logistic regression, and SVM. Nti et al. (2020) also find that RFs tend to be underutilized in studies that focus on stock price prediction. Weng et al. (2018) use machine learning methods (boosted regression tree, RFs, ANN, SVM) combined with social media type data like web page views, financial news sentiment, and search trends to predict stock prices. Twenty large US companies are studied. RFs and boosted regression trees outperform ANN or SVM. The main message from these papers is that RFs have high accuracy when predicting stock price direction but are underrepresented in the literature compared to other machine learning methods.

The purpose of this paper is to predict clean energy stock price direction using random forests. Directional stock price forecasts are constructed from one day to twenty days in the future. A multi-step forecast horizon is used in order to gain an understanding of how forecast accuracy changes across time (Basak et al. 2019; Khan et al. 2020). A five-day forecast horizon corresponds to one week of trading days, a 10-day forecast horizon corresponds to two weeks of trading days and a twenty-day forecast horizon corresponds to approximately one month of trading days. Forecasting stock price direction over a multi-day horizon provides a more challenging environment to compare models. Clean energy stock prices are measured using several well-known and actively traded exchange traded funds (ETFs). Forecasts are constructed using logit models, bagging decision trees, and RFs. A comparison is made between the forecasting accuracy of these models. Feature selection is based on several well-known technical indicators like moving average, stochastic oscillator, rate of price change, MACD, RSI, and advance decline line (Bustos and Pomares-Quimbaya 2020).

The analysis from this research provides some interesting results. RFs and tree bagging show much better stock price prediction accuracy than logit or step-wise logit. The prediction accuracy from bagging and RFs is very similar indicating that either method is very useful for predicting the stock price direction of clean energy ETFs. The prediction accuracy for RF and tree bagging models is over 80% for forecast horizons of 10 days or more. For a 20-day forecast horizon, tree bagging and random forests methods produce accuracy rates of between 85% and 90% while logit models produce accuracy rates of between 55% and 60%.

This paper is organized as follows. The next section sets out the methods and data. This is followed by the results and a discussion. The last section of the paper provides some conclusions and suggestions for future research.

## 2. Methods and Data

### 2.1. The Logit Method for Prediction

The objective of this paper is to predict the stock price direction of clean energy stocks. Stock price direction can be classified as either up (stock price change from one period to the next is positive) or down (stock price change from one period to the next is non-positive). This is a standard classification problem where the variable of interest can take on one of two values (up, down) and is easily coded as a binary variable. One approach to modelling and forecasting the direction of stock prices is to use logit models. Explanatory variables deemed relevant to predicting stock price direction can be used as features. Logit models are widely used and easy to estimate.

$$y_{t+h} = \alpha + \beta X_t + \varepsilon_t, \tag{1}$$

In Equation (1), $y_{t+h} = p_{t+h} - p_t$ is a binary variable that takes on the value of "up" if positive or "down" if non-positive and X is a vector of features. The variable $p_t$ represents the adjusted closing stock price on day t. The random error term is $\varepsilon$. The value of $h = 1, 2, 3, \ldots, 20$ indicates the number of time periods into the future to predict. A multistep forecast horizon is used in order to see how forecast accuracy changes across the forecast horizon. A 20-day forecast horizon is used in this paper since this is consistent with the average number of trading days in a month. The features include well-known technical indicators like the relative strength indicator (RSI), stochastic oscillator (slow, fast), advance-decline line (ADX), moving average cross-over divergence (MACD), price rate of change (ROC), on balance volume (OBV), and the 200-day moving average. While there are many different technical indicators the ones chosen in this paper are widely used in academics and practice (Yin and Yang 2016; Yin et al. 2017; Neely et al. 2014; Wang et al. 2020; Bustos and Pomares-Quimbaya 2020).

### *2.2. The Random Forests Method for Prediction*

Logit regression classifies the dependent (response) variable based on a linear boundary and this can be limiting in situations where there is a nonlinear relationship between the response and the features. In such situations, a decision tree approach may be more useful. Decision trees bisect the predictor space into smaller and smaller non-overlapping regions and are better able to capture the classification between the response and the features in nonlinear situations. The rules used to split the predictor space can be summarized in a tree diagram, and this approach is known as a decision tree method. Tree based methods are easy to interpret but are not as competitive with other methods like bagging or random forests. A brief discussion of decision trees, bagging, and random forest methods is presented here but the reader is referred to James et al. (2013) for a more complete treatment.

A classification tree is used to predict a qualitative response rather than a quantitative one. A classification tree predicts that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs. A majority voting rule is used for classification. The basic steps in building a classification tree can be described as follows.

1. Divide the predictor space (all possible values for $X_1, \ldots, X_P$) into J distinctive and non-overlapping regions, $R_1, \ldots, R_J$.
2. For every observation that falls into the region Rj, the same prediction is made. This prediction is that each observation belongs to the most commonly occurring class of training observations to which it belongs.

The regions $R_1, \ldots, R_J$ can be constructed as follows. Recursive binary splitting is used to grow the tree and splitting rules are determined by a classification error rate. The classification error rate, *E*, is the fraction of training observations in a region that do not belong to the most common class.

$$E = 1 - \frac{\max(\hat{p}_{mk})}{k} \tag{2}$$

In Equation (2), $\hat{p}_{mk}$ is the proportion of training observations in the $m$th region that are from the $k$th class. The classification error is not very sensitive to the growing of trees so in practice either the Gini index ($G$) or entropy ($D$) is used to classify splits.

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{3}$$

The Gini index measures total variance across the $K$ classes. In this paper, there are only two classes (stock price direction positive, or not) so $K = 2$. For small values of $\hat{p}_{mk}$ the Gini index takes on a small value. For this reason, $G$ is often referred to as a measure of node impurity since a small $G$ value shows that a node mostly contains observations from a single class. The root node at the top of a decision tree can be found by trying every possible split of the predictor space and choosing the split that reduces the impurity as much as possible (has the highest gain in the Gini index). Successive nodes can be found using the same process and this is how recursive binary splitting is evaluated.

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} log(\hat{p}_{mk}) \tag{4}$$

The entropy ($D$), like the Gini index, will take on small values if the $m$th node is pure. The Gini index and entropy produce numerically similar values. The analysis in this paper uses the entropy measure.

The outcome of this decision tree building process is typically a very deep and complex tree that may produce good predictions on the training data set but is likely to over fit the data leading to poor performance on unseen data. Decision trees suffer from high variance which means that if the training data set is split into two parts at random and a decision tree fit to both halves, the outcome would be very different. One approach to remedying this is to use bagging. Bootstrap aggregation or bagging is a statistical technique used to reduce the variance of a machine learning method. The idea behind bagging it to take many training sets, build a decision tree on each training data set, and average the predictions to obtain a single low-variance machine learning model. In general, however, the researcher does not have access to many training sets so instead bootstrap replication is used to create many bootstrap training data sets and a decision tree grown for each replication. Each individual tree is constructed by randomly sampling the predictors with replacements. So, for example, if the number of predictors is 6, then each tree has 6 predictors but because sampling is done with replacement some predictors may be sampled more than once. Thus, the number of bootstrap replications is the number of trees. Even though each tree is grown deep and has high variance, averaging the predictions from these bootstrap trees reduces variance. The test error of a bagged model can be easily estimated using the out of bag (OOB) error. In bagging, decision trees are repeatedly fit to bootstrapped subsets of the observations. On average, each bagged tree uses approximately two-thirds of the observations (James et al. 2013). The remaining one-third of the observations not used are referred to as the OOB observations and can be used as a test data set to evaluate prediction accuracy. The OOB test error can be averaged across trees.

Random forests are comprised of a large number of individual decision trees that operate as an ensemble (Breiman 2001). Random forests are non-metric classifiers because no learning parameters need to be set. Random forests are an improvement over bagging trees by introducing decorrelation between the trees. As in the case of bagging, a large number of decision trees are built on bootstrapped training samples. Each individual tree in the random forest produces a prediction for the class and the class with the most votes is the model's prediction. Each time a split in a tree occurs a random sample of predictors is chosen as split candidates from the full set of predictors. Notice how this differs from bagging. In bagging all predictors are used at each split. In random forests, the number of predictors chosen at random is usually calculated as the square root of the total number of predictors (James et al. 2013). While the choice of randomly choosing predictors may seem strange, averaging results from non-correlated trees is much better for reducing variance than averaging trees that are highly correlated. In random forests, trees are trained on different samples due to bagging and also use different features when predicting outcomes.

This paper compares the performance of logit, step-wise logit, bagging decision tree, and random forests for predicting the stock price direction of clean energy ETFs. For the analysis, 80% of the data was used for training and 20% used for testing. Classification prediction is one of the main goals of classification trees and the accuracy of prediction can be obtained from the confusion matrix. The logit model uses all of the features in predicting stock price direction. The step-wise logit uses a backwards step-wise reduction algorithm evaluated using Akaike Information Criteria (AIC) to create a sub-set of influential features. The bagging decision tree model was estimated with 500 trees. The random forecasts were estimated with 500 trees and 3 (the floor of the square root of the number of predictor variables, 10) randomly chosen predictors at each split (Breiman 2001). The results are not sensitive to the number of trees provided a large enough number of trees are chosen. A very large number of trees does not lead to overfitting, but a small number of trees results in high test error. Training control for the random forest was handled with 10-fold cross validation with 10 repeats. All calculations were done in R (R Core Team 2019) and used the random forests machine learning package (Breiman et al. 2018).

### *2.3. The Data*

The data for this study consists of the stock prices of five popular, US listed, and widely traded clean energy ETFs. The Invesco WilderHill Clean Energy ETF (PBW) is the most widely known clean energy ETF and has the longest trading period with an inception date of 3 March 2005. This ETF consists of publicly traded US companies that are in the clean energy business (renewable energy, energy storage, energy conversion, power delivery, greener utilities, cleaner fuels). The iShares Global Clean Energy ETF (ICLN) seeks to track the S&P Global Clean Energy Index. The First Trust NASDAQ Clean Edge Green Energy Index Fund (QCLN) tracks the NASDAQ Clean Edge Energy Index. The Invesco Solar ETF (TAN) tracks the MAC Global Solar Energy Index which focuses on companies that generate a significant amount of their revenue from solar equipment manufacturing or enabling products for the solar power industry. The First Trust Global Wind Energy ETF (FAN) tracks the ISE Clean Edge Global Wind Energy Index and consists of companies throughout the world that are in the wind energy industry. TAN and FAN began trading near the middle of 2008. The daily data set starts on 1 January 2009 and ends on 30 September 2020. The data was collected from Yahoo Finance. Several well-known technical indicators like the relative strength indicator (RSI), stochastic oscillator (slow, fast), advance-decline line (ADX), moving average cross-over divergence (MACD), price rate of change (ROC), on balance volume, and the 200-day moving average, calculated from daily data, are used as features in the logit and RFs prediction models.

The time series pattern of the clean energy ETFs shows that the ETFs move together (Figure 1). There was a double peak formation in early 2009 and 2011 followed by a trough in 2013. This was followed by a peak in 2014 and then a relatively horizontal pattern between 2017 and 2019. In response to the global financial crisis of 2008–2009 some countries, like the US, China, and South Korea, implemented fiscal stimulus packages where the economic stimulus was directed at achieving economic growth and environmental sustainability (Andreoni 2020; Mundaca and Richter 2015). This helped to increase the stock prices of clean energy companies. All of the ETFs have risen sharply since the onset of the World Health Organization's declaration of the COVID19 global pandemic (March 2020).

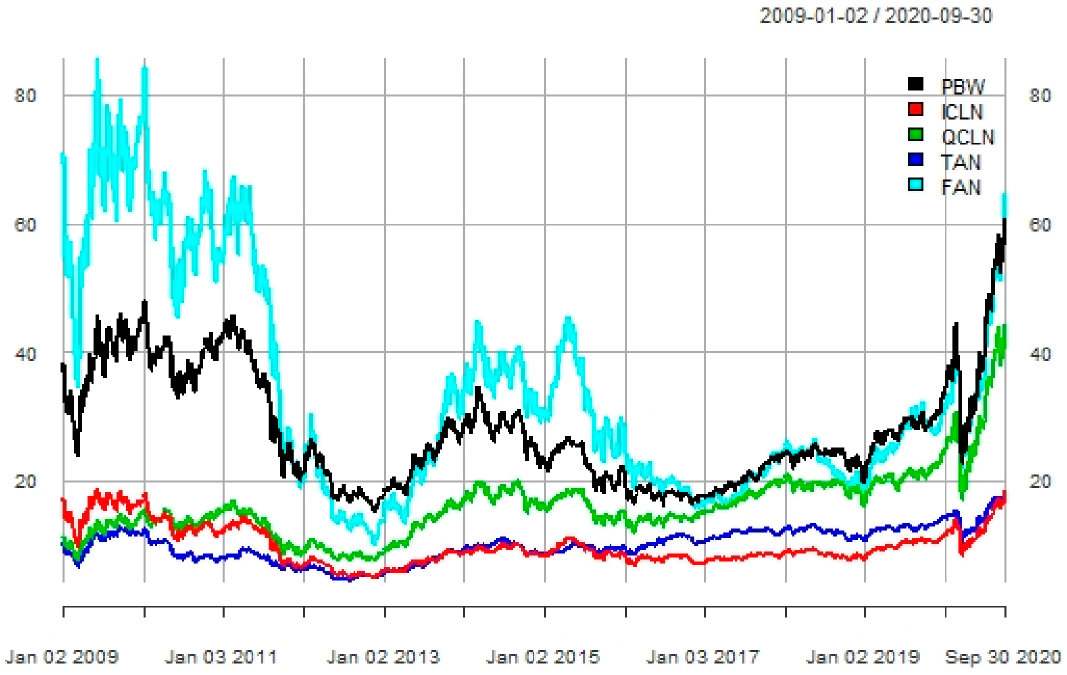

**Figure 1.** This figure shows clean energy ETF stock prices across time. Data sourced from Yahoo Finance.

The histograms for the percentage of up days shows little variation for PBW, ICLN, and TAN (Figure 2). The percentage of up days increases with the number of days for QCLN while for FAN, the pattern increases up to about 7 days after which the percentage of up days shows little variation with longer time periods. Compared to the other clean energy ETFs studied in this paper, QCLN has the strongest trend in the data (Figure 1) and this is consistent with the higher proportion of up days.

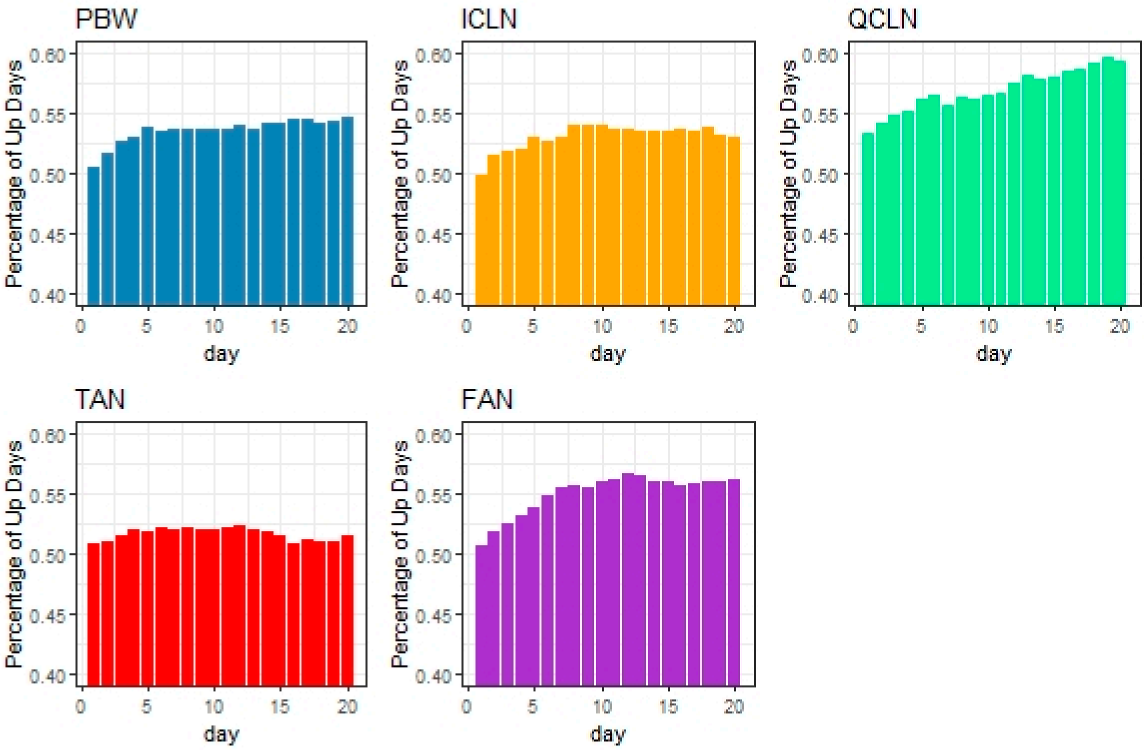

**Figure 2.** This figure shows histograms of clean energy ETF percentage of up days. Data sourced from Yahoo Finance. Author's own calculations.

In order to investigate the impact of the number of trees on the random forests model, Figure 3 shows how the test error relates to the number of trees. The analysis is conducted for a 10-step forecast horizon where 80% of the data is used for training and 20% of the data is used for testing. In each case, the test error declines rapidly as the number of trees increases from 1 to 100. After 300 trees there is very small reduction in the test error. Notice how the test error converges. This shows that random forests do not over fit as the number of trees increases. In Figure 3, out of bag (OOB) test error is reported along with test error for the up and down classification. The results for other forecast horizons are similar to those reported here. Consequently, 500 trees are used in estimating the RFs.

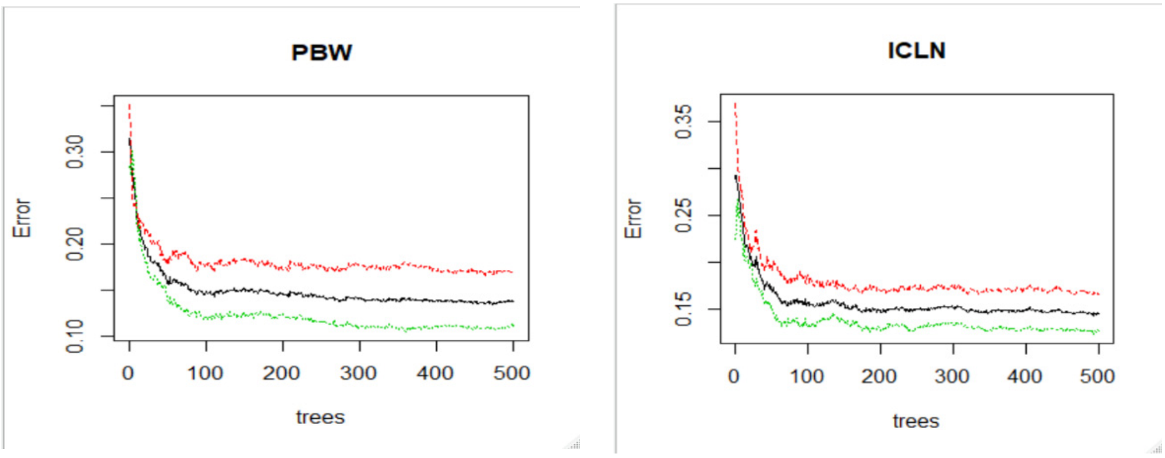

**Figure 3.** *Cont*.

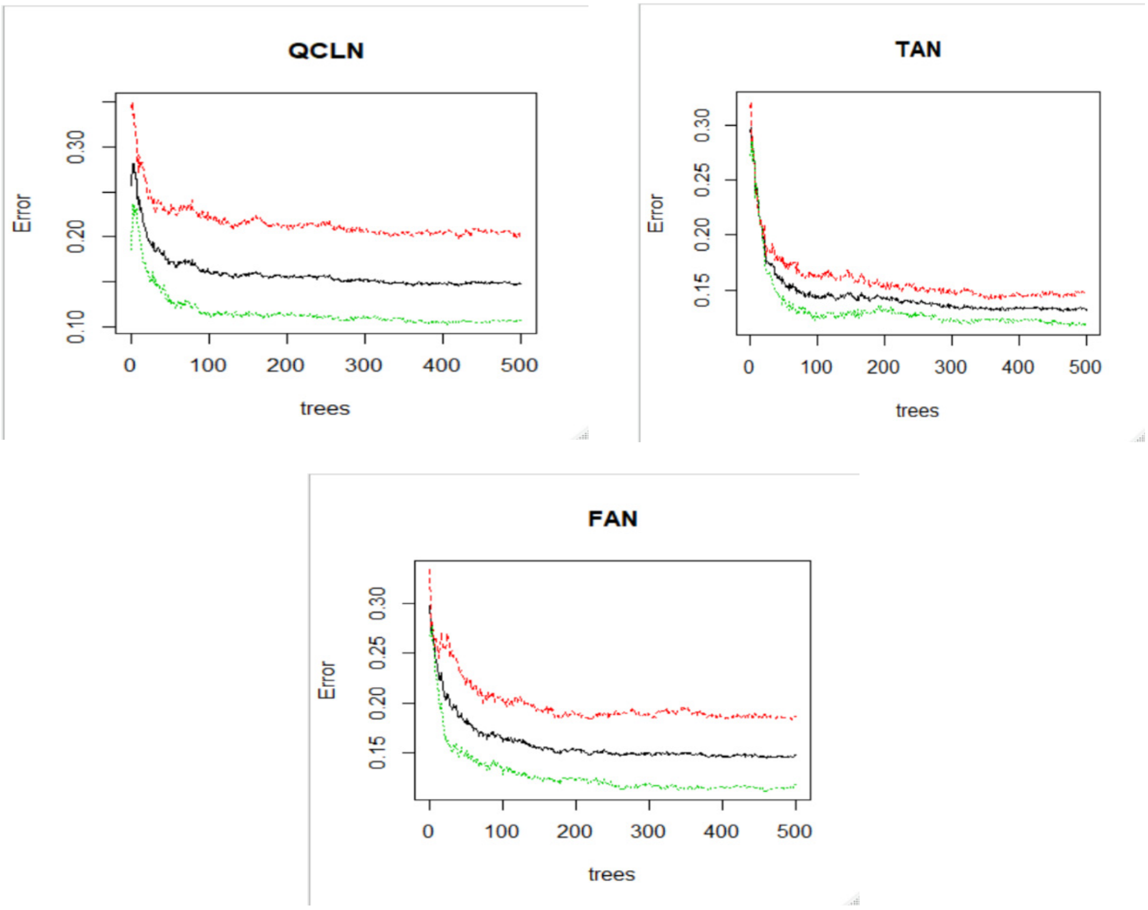

**Figure 3.** This figure shows RFs test error vs. the number of trees. OOB (red), down classification (black), up classification (green). Calculations are done for predicting stock price direction over a 10-step forecast horizon.

## 3. Results

This section reports the results from predicting stock price direction for clean energy ETFs. Since this is a classification problem, the prediction accuracy is probably the single most useful measure of forecast performance. Prediction accuracy is a proportion of the number of true positives and true negatives divided by the total number of predictions. This measure can be obtained from the confusion matrix. Other useful forecast accuracy measures like how well the models predict the up or down classification are also available and are reported since it is interesting to see if the forecast accuracy for predicting the up class is similar or different to that of predicting the down class.

Stock price direction prediction accuracy for PWB (Figure 4) shows large differences between the logit models and RF or logit models and tree bagging. The prediction accuracy for logit and logit stepwise show that while there is some improvement in accuracy between 1 and 5 days ahead, the prediction accuracy never gets above 0.6 (60%). The prediction accuracy of the RFs and tree bagging methods show considerable improvement in accuracy between 1 and 10 days. Prediction accuracy for predicting stock price direction 10 days into the future is over 85%. There is little variation in prediction accuracy for predicting stock price direction between 10 and 20 days into the future. Notice that the prediction accuracy between tree bagging and RF is very similar.

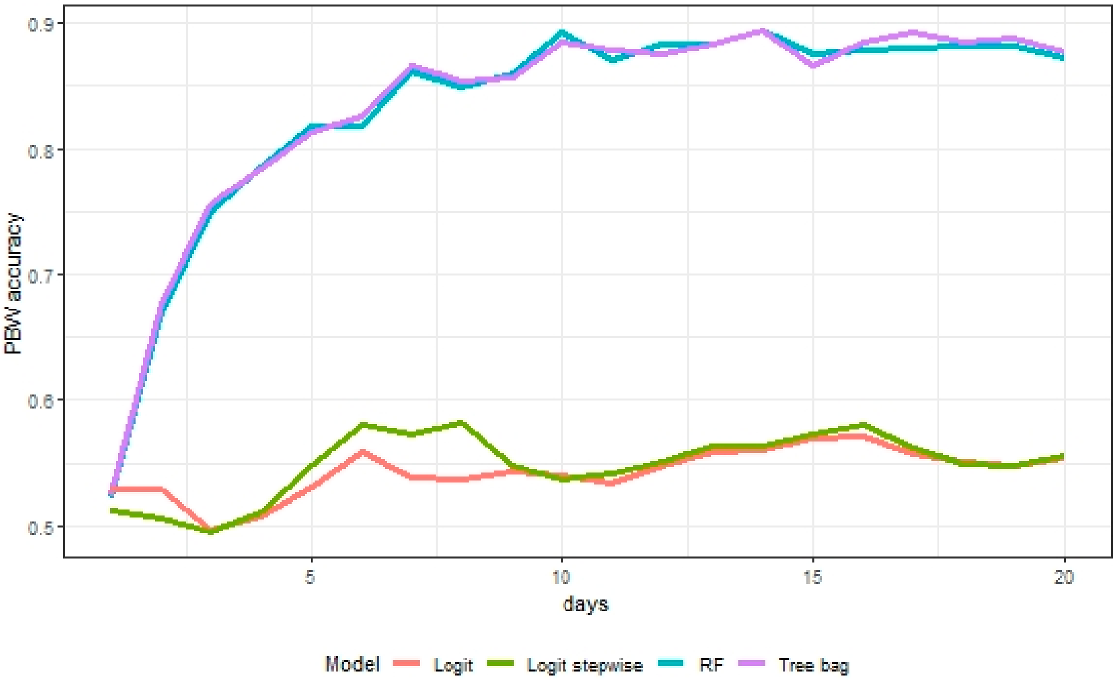

**Figure 4.** This figure shows the multi-period prediction accuracy for PBW stock price direction.

The patterns of prediction accuracy for the other clean energy ETFs are very similar to that which was described for the PBW clean energy ETF (Figures 5–8). For each ETF, the prediction accuracy of RF and bagging trees are very similar and much more accurate than that of the logit models.

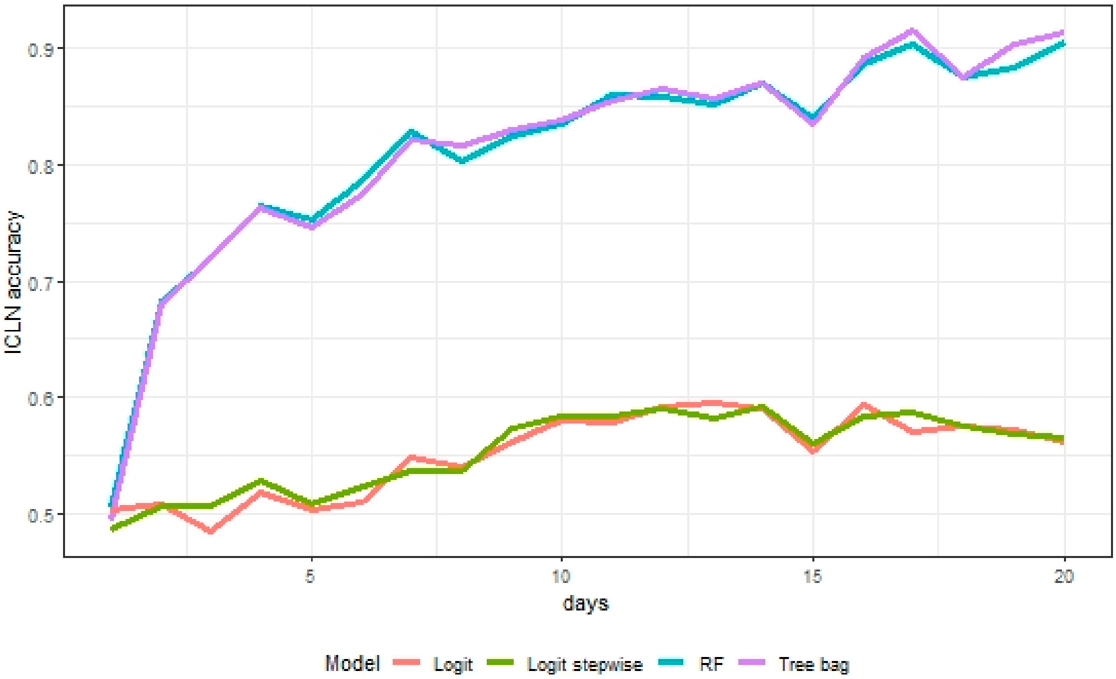

**Figure 5.** This figure shows the multi-period prediction accuracy for ICLN stock price direction.

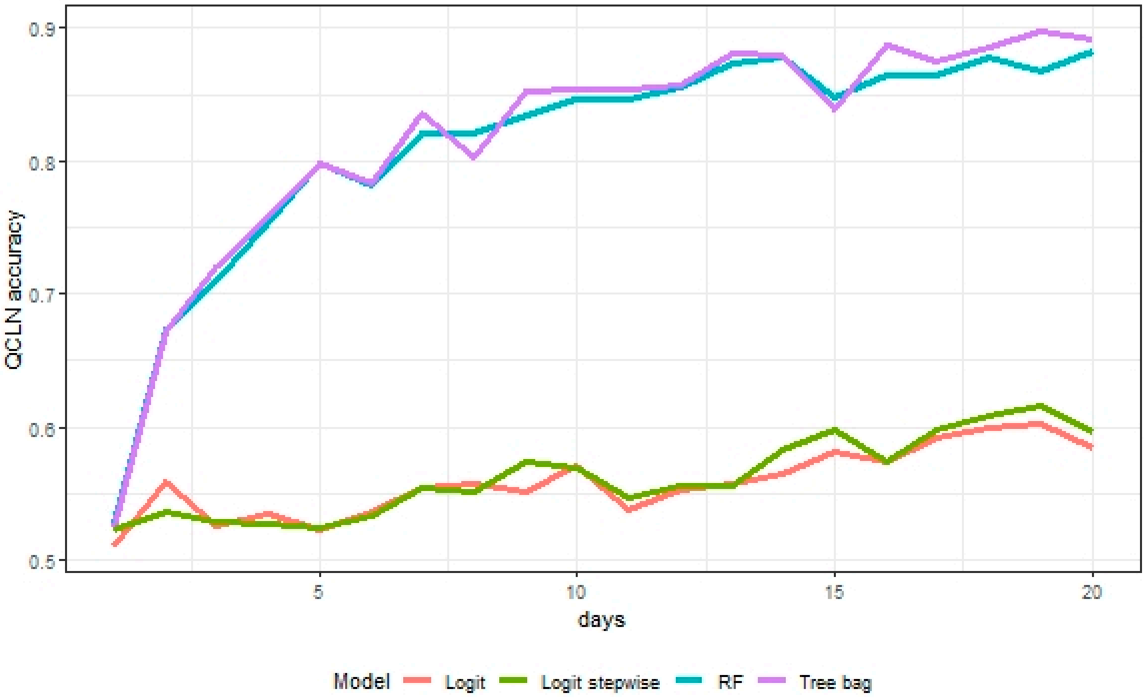

**Figure 6.** This figure shows the multi-period prediction accuracy for QCLN stock price direction.

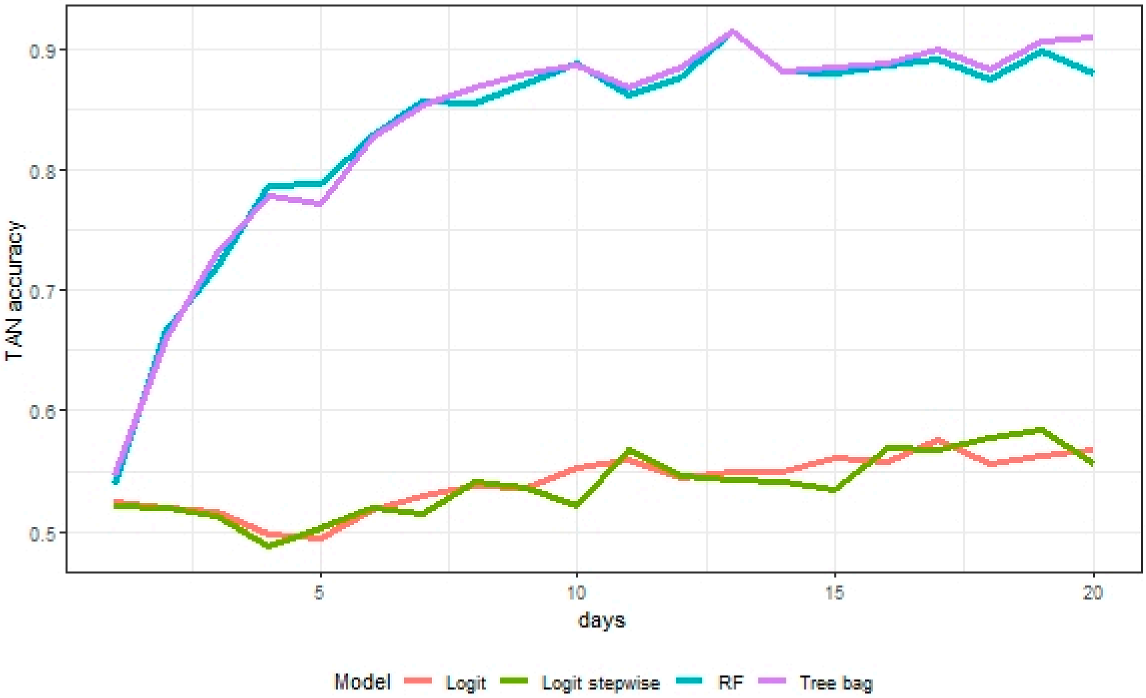

**Figure 7.** This figure shows the multi-period prediction accuracy for TAN stock price direction.

Variable importance is used to determine which variables are most important in the RFs method. The mean decrease in accuracy (MD accuracy) is computed from the OOB data. The mean decrease in Gini (MD Gini) is a measure of node impurity. For each ETF at a 10-period forecast horizon, the OBV and MA200 are the two most important features in classifying clean stock price direction because they have the largest values of MD accuracy and MD Gini (Table 1). Further analysis for other forecasting horizons (not reported) shows that OBV and MA200 are also the two most important features in classifying clean stock price direction for other forecast horizons.

**Table 1.** Variable importance for predicting stock price direction.

| PBW | DOWN | UP | MD Accuracy | MD Gini |
|---|---|---|---|---|
| RSI | 27.25 | 25.08 | 39.08 | 83.52 |
| STOFASTK | 21.72 | 32.02 | 40.34 | 88.16 |
| STOFASTD | 23.57 | 30.88 | 41.60 | 87.75 |
| STOSLOWD | 24.69 | 37.53 | 48.96 | 92.61 |
| ADX | 47.96 | 35.96 | 59.22 | 113.76 |
| MACD | 31.61 | 32.30 | 49.15 | 95.34 |
| MACD SIG | 40.14 | 47.23 | 61.07 | 112.65 |
| ROC | 22.92 | 31.45 | 42.71 | 89.70 |
| OBV | 40.38 | 65.52 | 78.83 | 159.42 |
| MA200 | 55.16 | 59.05 | 76.69 | 163.43 |
| **ICLN** | **DOWN** | **UP** | **MD Accuracy** | **MD Gini** |
| RSI | 25.22 | 26.11 | 43.73 | 84.61 |
| STOFASTK | 24.96 | 28.46 | 38.39 | 90.97 |
| STOFASTD | 24.25 | 29.09 | 40.42 | 90.17 |
| STOSLOWD | 27.16 | 33.68 | 45.18 | 99.41 |
| ADX | 42.63 | 45.21 | 57.23 | 118.19 |
| MACD | 31.44 | 33.40 | 52.24 | 98.33 |
| MACD SIG | 39.21 | 44.27 | 65.39 | 115.14 |
| ROC | 28.61 | 34.54 | 44.93 | 94.22 |
| OBV | 46.48 | 45.74 | 68.72 | 136.85 |
| MA200 | 48.86 | 54.70 | 73.62 | 160.22 |
| **QCLN** | **DOWN** | **UP** | **MD Accuracy** | **MD Gini** |
| RSI | 21.12 | 27.75 | 38.94 | 80.18 |
| STOFASTK | 17.55 | 31.33 | 41.97 | 82.94 |
| STOFASTD | 19.79 | 25.81 | 38.73 | 79.04 |
| STOSLOWD | 22.80 | 28.07 | 39.61 | 84.70 |
| ADX | 48.87 | 39.29 | 57.54 | 121.13 |
| MACD | 26.27 | 36.55 | 49.99 | 100.96 |
| MACD SIG | 33.81 | 40.38 | 55.43 | 106.13 |
| ROC | 20.09 | 28.05 | 37.88 | 83.31 |
| OBV | 38.38 | 49.53 | 70.07 | 157.29 |
| MA200 | 37.90 | 54.03 | 79.05 | 177.52 |
| **TAN** | **DOWN** | **UP** | **MD Accuracy** | **MD Gini** |
| RSI | 26.08 | 26.69 | 40.33 | 82.21 |
| STOFASTK | 24.88 | 30.74 | 40.60 | 89.81 |
| STOFASTD | 25.92 | 26.57 | 40.50 | 88.25 |
| STOSLOWD | 27.52 | 31.09 | 48.16 | 92.26 |
| ADX | 47.65 | 36.85 | 57.51 | 106.17 |
| MACD | 35.33 | 33.96 | 57.22 | 97.09 |
| MACD SIG | 46.82 | 41.11 | 63.02 | 119.30 |
| ROC | 26.14 | 35.20 | 44.29 | 86.07 |
| OBV | 40.31 | 44.17 | 62.35 | 143.25 |
| MA200 | 57.22 | 65.98 | 87.26 | 188.03 |
| **FAN** | **DOWN** | **UP** | **MD Accuracy** | **MD Gini** |
| RSI | 19.60 | 31.69 | 40.69 | 83.58 |
| STOFASTK | 29.17 | 28.86 | 40.30 | 89.74 |
| STOFASTD | 24.42 | 31.22 | 39.79 | 85.90 |
| STOSLOWD | 19.88 | 34.82 | 43.65 | 87.54 |
| ADX | 43.40 | 43.42 | 61.05 | 106.36 |
| MACD | 30.11 | 35.62 | 53.46 | 95.40 |
| MACD SIG | 38.86 | 44.22 | 63.47 | 107.28 |
| ROC | 29.82 | 32.87 | 42.90 | 88.70 |
| OBV | 53.00 | 57.56 | 68.49 | 166.36 |
| MA200 | 48.83 | 65.65 | 83.29 | 169.59 |

This table shows the RFs variable importance of the technical analysis indicators measured using mean decrease in accuracy (MD accuracy) and mean decrease in GINI (MD Gini). Values reported for a 10-period forecast horizon.

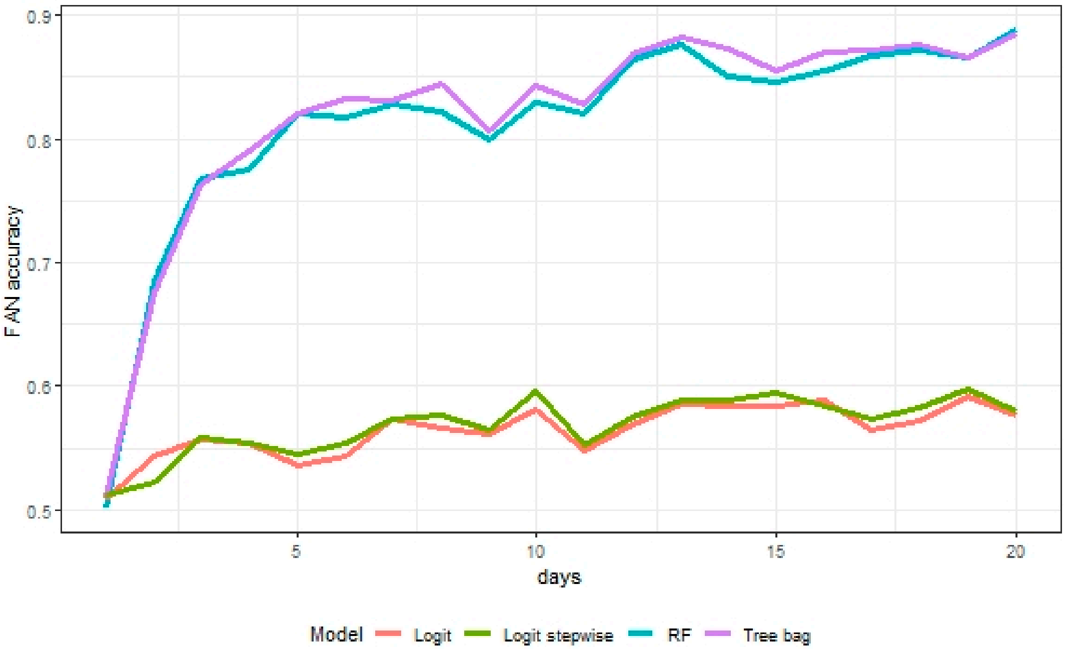

**Figure 8.** This figure shows the multi-period prediction accuracy for FAN stock price direction.

Figures 4–8 show the overall prediction accuracy. Another interesting question to ask is how the prediction accuracy compares between positive prediction values and negative prediction values. Positive predictive value is the proportion of predicted positive cases that are actually positive. An alternative way to think about this is, when a model predicts a positive case, how often is it correct?

Figure 9 reports the positive prediction value for PBW. This plot shows how accurate the models are in prediction the positive price direction. The RFs and tree bagging methods are more accurate than the logit methods. After 5 days, the RFs and tree bagging methods have an accuracy of over 80% while the accuracy of the logit methods never reaches higher than 70%. The pattern of positive predictive value for the other ETFs (Figures 10–13) are similar to what is observed for PBW. For each ETF, after 10 days the positive predictive values for RFs and bagging are above 0.80 and in most cases above 0.85.

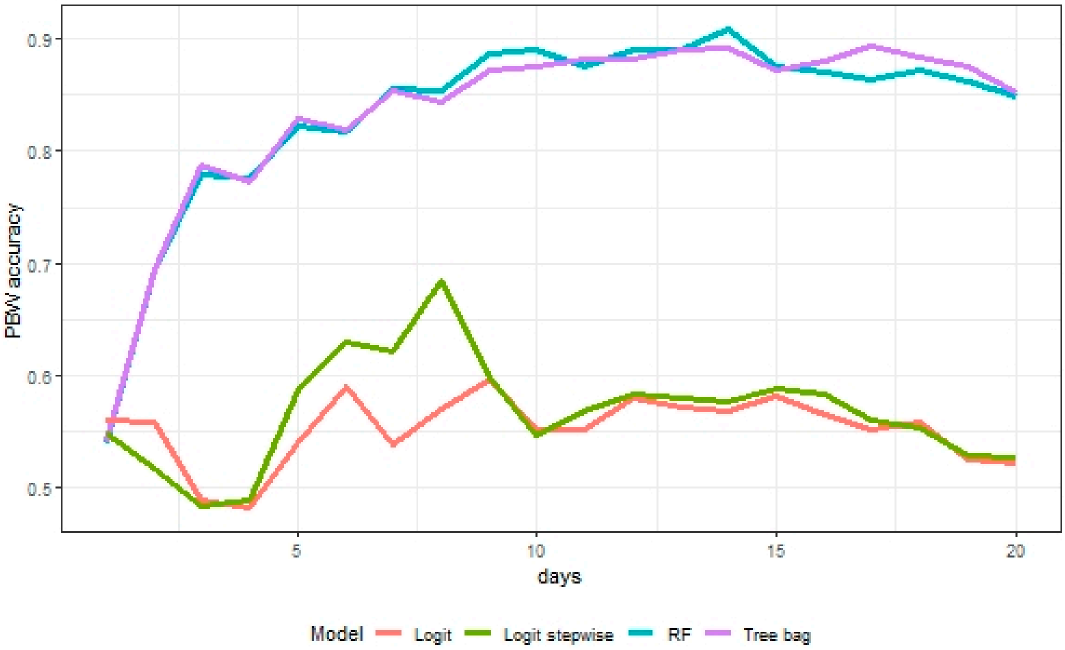

**Figure 9.** This figure shows the multi-period positive predictive values accuracy for PBW stock price direction.

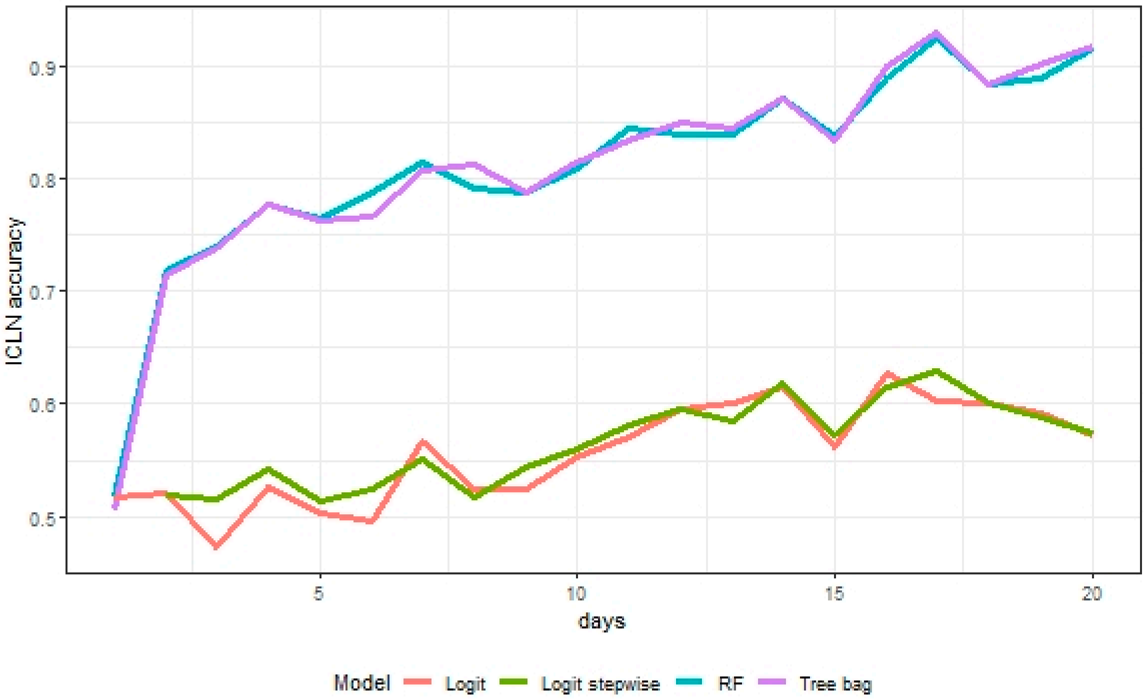

**Figure 10.** This figure shows the multi-period positive predictive values accuracy for ICLN stock price direction.

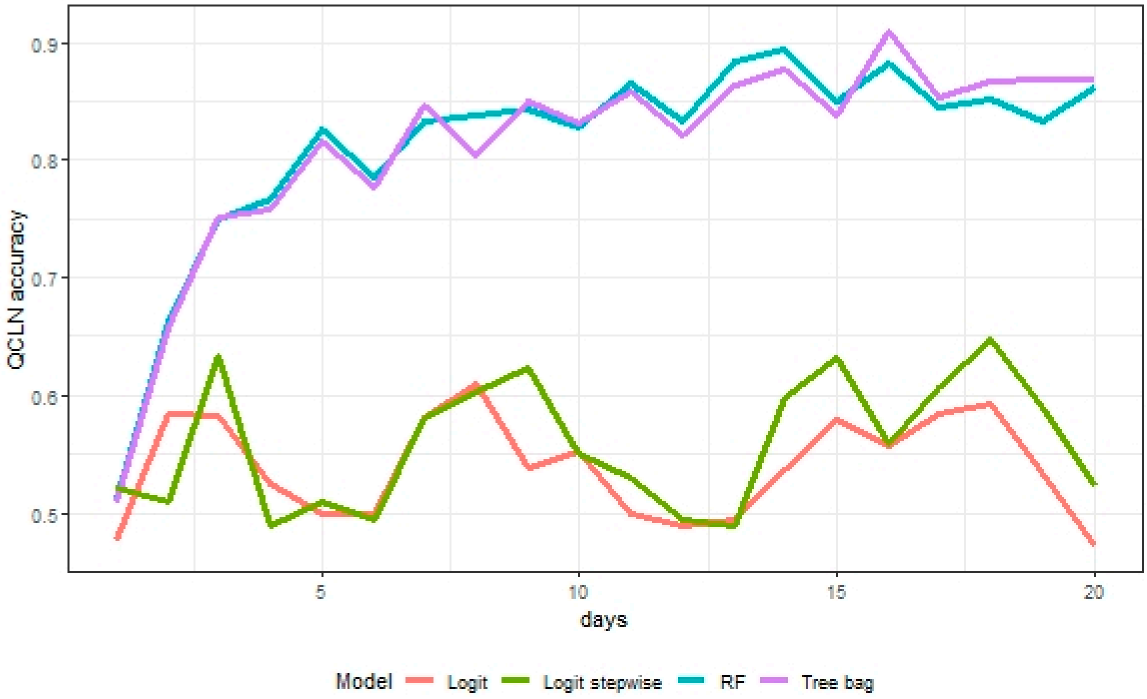

**Figure 11.** This figure shows the multi-period positive predictive values accuracy for QCLN stock price direction.

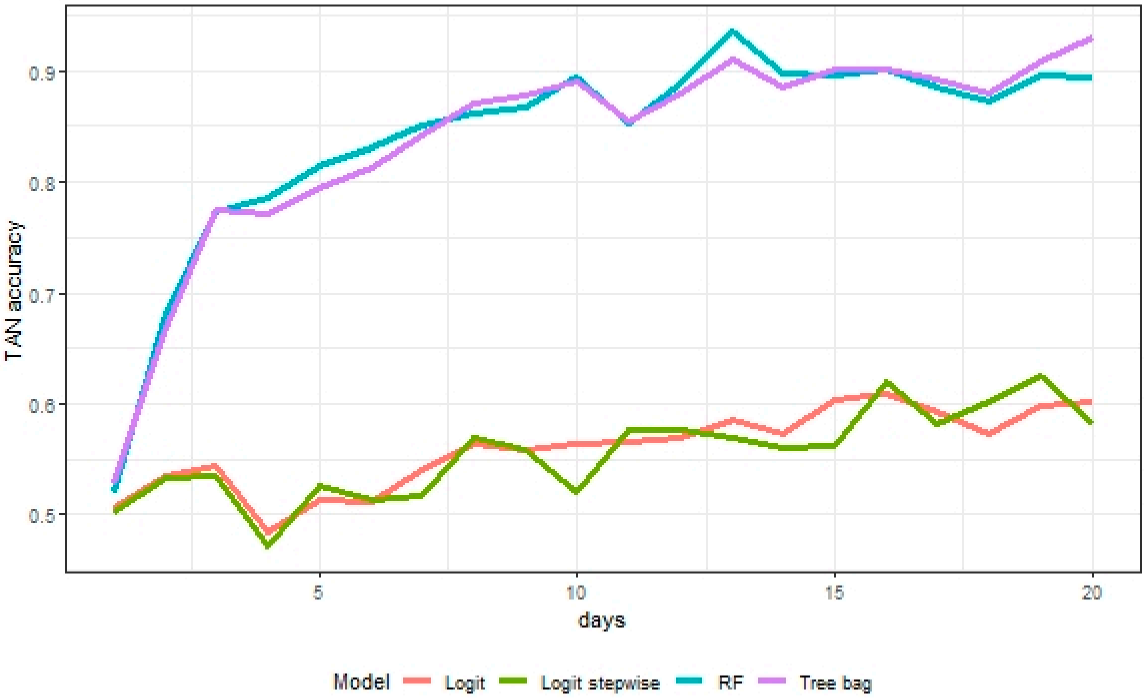

**Figure 12.** This figure shows the multi-period positive predictive values accuracy for TAN stock price direction.

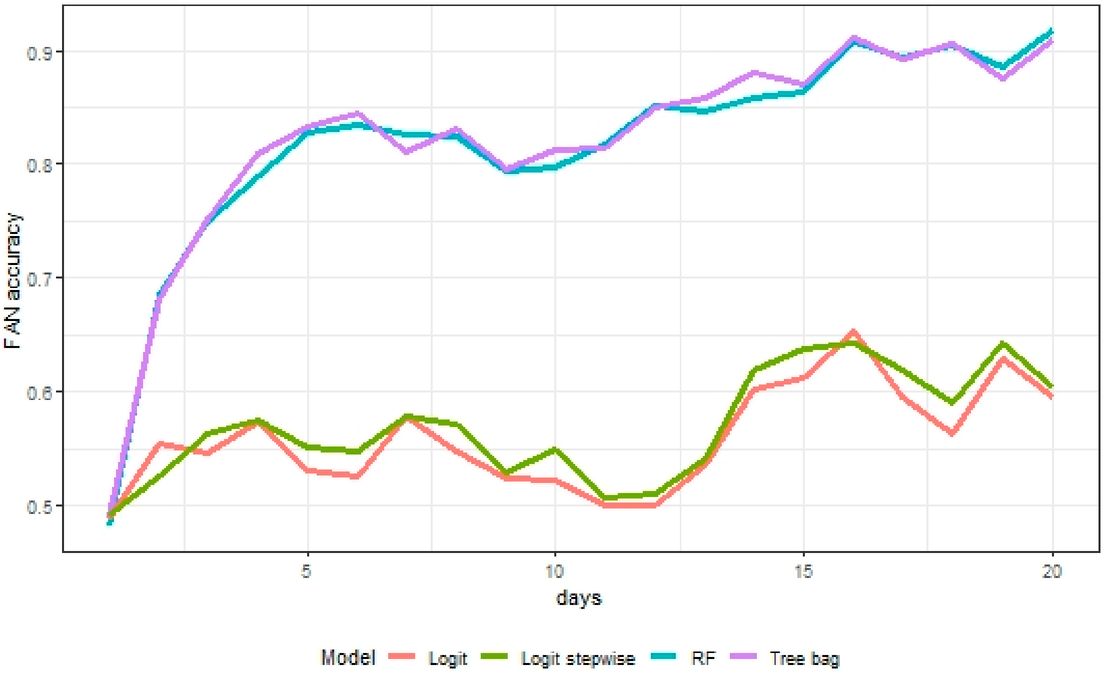

**Figure 13.** This figure shows the multi-period positive predictive values accuracy for FAN stock price direction.

Figures 14–18 show the negative predictive value. The negative predictive value is the proportion of predicted negative cases relative to the actual number of negative cases. Figure 14 reports the negative predictive value for PBW. This plot shows how accurate the models are in predicting the down stock price direction. The RFs and tree bagging methods are more accurate than the logit models. For the RFs and tree bagging models, accuracy increases from 0.5 to 0.8 between 1 and 5 days. After 10 days negative predictive value fluctuates between 0.85 and 0.90. The pattern of negative predictive value for the other ETFs (Figures 15–18) are similar to what is observed for PBW. For each ETF, after

10 days the negative predictive values for RFs and bagging are above 0.80 and in most cases above 0.85.

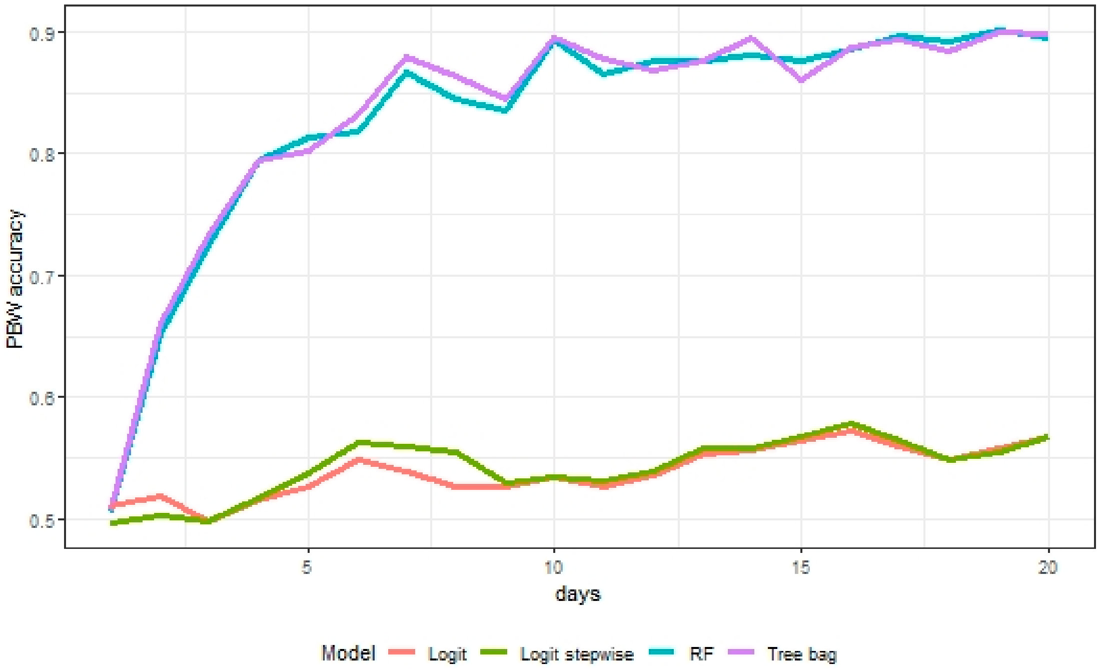

**Figure 14.** This figure shows the multi-period negative predictive values accuracy for PBW stock price direction.

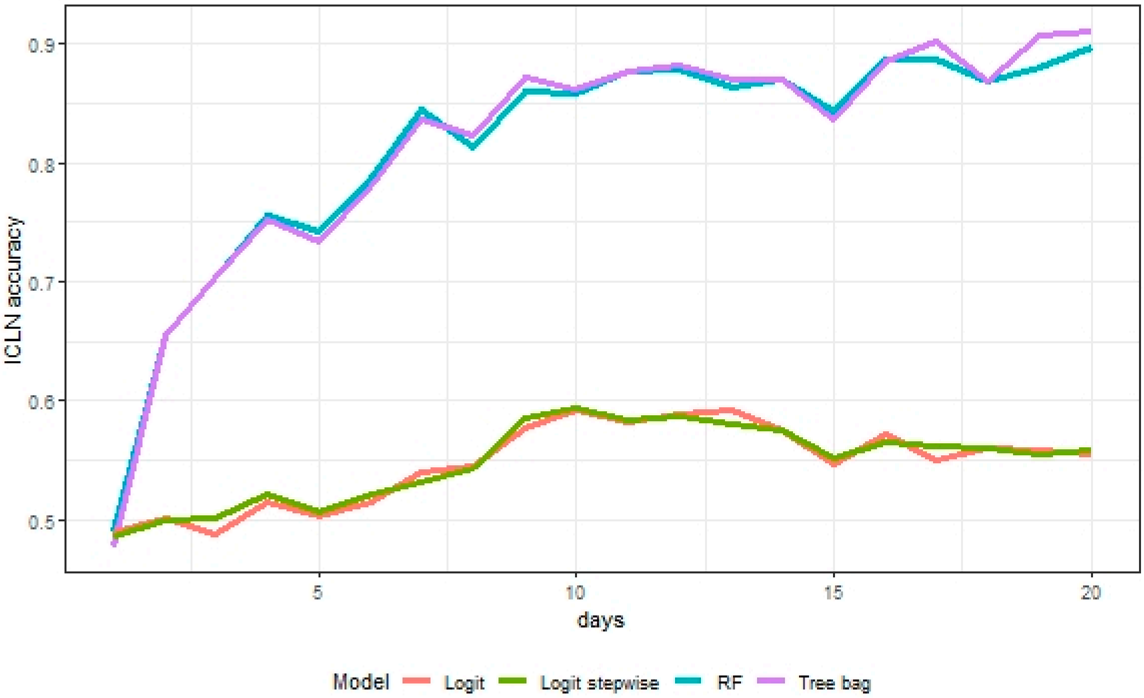

**Figure 15.** This figure shows the multi-period negative predictive values accuracy for ICLN stock price direction.

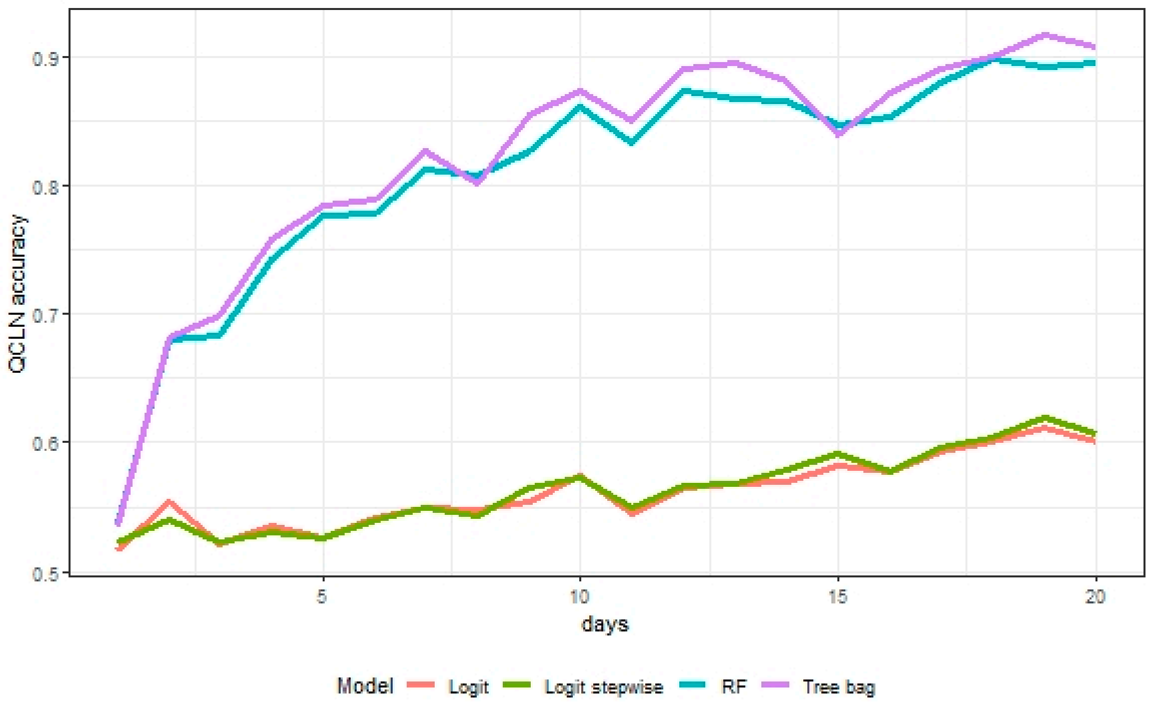

**Figure 16.** This figure shows the multi-period negative predictive values accuracy for QCLN stock price direction.

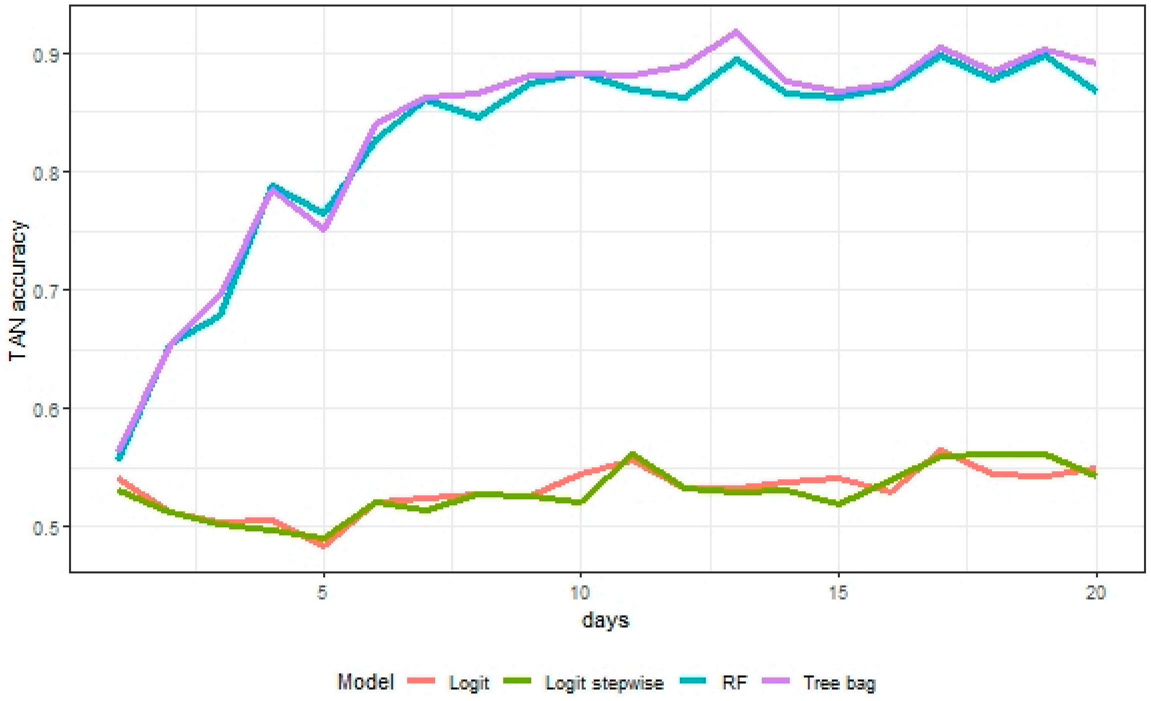

**Figure 17.** This figure shows the multi-period negative predictive values accuracy for TAN stock price direction.

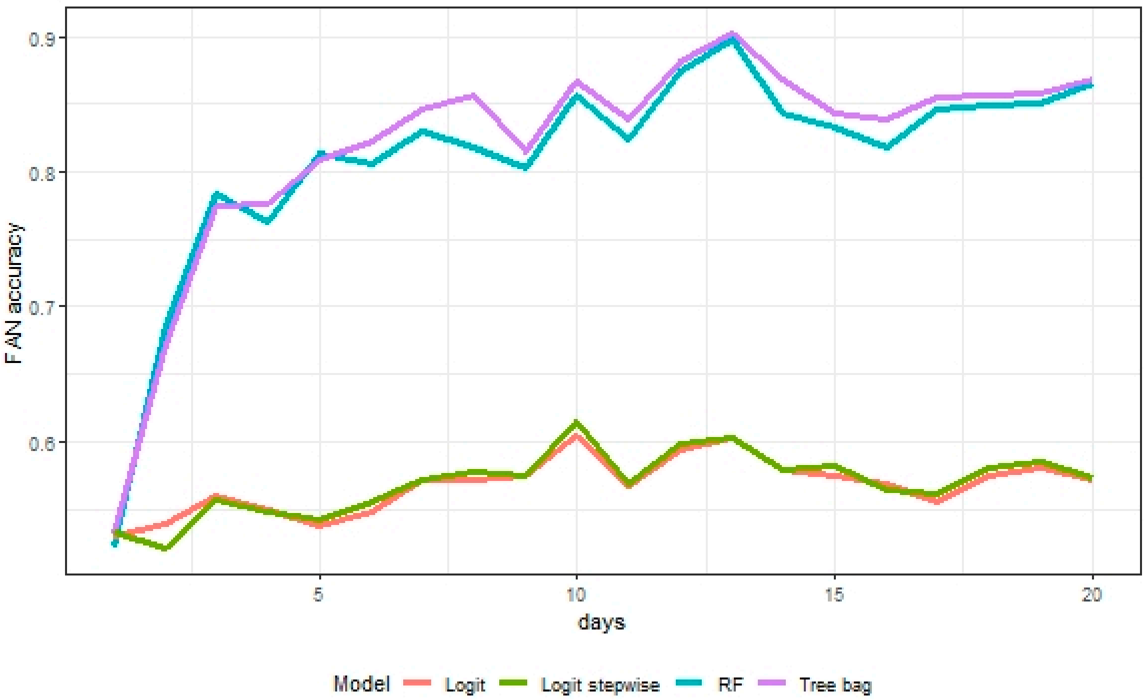

**Figure 18.** This figure shows the multi-period negative predictive values accuracy for FAN stock price direction.

To summarize, the main take-away from this research is that RFs and tree bagging provide much better predicting accuracy then logit or step-wise logit. The prediction accuracy between bagging and RFs is very similar indicating that either method is very useful for predicting the stock price direction of clean energy ETFs. The prediction accuracy for RF and tree bagging models is over 80% for forecast horizons of 10 days or more. The positive predictive values and negative predictive values are similar indicating that there is little asymmetry between the up and down prediction classifications.

## 4. Discussion

The research in this paper shows that RFs produce more accurate clean energy stock price direction forecasts than logit models. These results add to a growing body of research that shows machine learning methods like RFs have considerable stock price direction predictive performance (Ballings et al. 2015; Basak et al. 2019; Lohrmann and Luukka 2019; Weng et al. 2018; Ampomah et al. 2020). None of these studies, however, consider clean energy stock prices. This paper appears to be the first paper to use ML methods to predict clean energy stock price direction.

There is literature on clean energy equity dynamics that largely focuses on the impact of oil prices on clean energy stock returns (Bondia et al. 2016; Dutta 2017; Dutta et al. 2018; Elie et al. 2019; Gupta 2017; Henriques and Sadorsky 2008; Kumar et al. 2012; Maghyereh et al. 2019; Managi and Okimoto 2013; Reboredo 2015; Reboredo et al. 2017b; Reboredo and Ugolini 2018; Uddin et al. 2019; Wen et al. 2014). Popular modelling approaches include multifactor models (Henriques and Sadorsky 2008; Gupta 2017; Reboredo et al. 2017b; Bohl et al. 2013; Sadorsky 2012), vector autoregressions (Kumar et al. 2012; Dutta et al. 2018), or other approaches like wavelets (Maghyereh et al. 2019; Reboredo et al. 2017a), copulas (Reboredo 2015), and quantiles (Uddin et al. 2019; Dawar et al. 2021). While this research is important in establishing that oil prices have a significant impact on clean energy stock prices the focus has not been on forecasting clean energy stock prices.

The results of this present paper could be combined with some of the knowledge discussed in the previous paragraph to expand the feature set used in estimating RFs. It may, for example, be useful to include other variables like oil prices in the set of features used in the estimation of RFs. A comparison could be made between feature sets that are based on technical indicators and feature sets that include oil prices and other macroeconomic variables to see if macroeconomic variables offer additional insight into predicting clean energy stock price direction.

## 5. Conclusions

There is a growing interest in investing in clean energy companies and some of the major drivers behind this interest include climate change, green consumers, energy security, fossil fuel divestment, and technological innovation. Investors in clean energy equities would benefit from a better understanding of how to predict clean energy stock prices. There is, however, a noticeable lack of information on this topic. This is the gap in the literature that this paper fills.

Building on the existing finance literature that shows stock price direction is easier to predict than stock prices and recent developments in machine learning showing that ML techniques offer an improvement in prediction over conventional regression-based approaches, this paper uses RFs and decision tree bagging to predict clean energy equity stock price direction. RFs and decision tree bagging are easier to explain and estimate than other ML techniques like ANNs or SVMs, but RFs appear to be underutilized in the existing literature. Five well known and actively traded clean energy ETFs are chosen for study. For each ETF, prediction accuracy is assessed using a time horizon of one day to twenty days (which is approximately one month of trading days).

RFs and tree bagging show much better stock price prediction accuracy then logit or step-wise logit. The prediction accuracy from bagging and RFs is very similar indicating that either method is very useful for predicting the stock price direction of clean energy ETFs. The prediction accuracy for RF and tree bagging models is over 80% for forecast horizons of 10 days or more. For a 20-day forecast horizon, tree bagging and random forests methods produce accuracy rates of between 85% and 90% while logit models produce accuracy rates of between 55% and 60%. These results are in agreement with other research that shows RFs to have a high stock price predictive accuracy (Ballings et al. 2015; Basak et al. 2019; Lohrmann and Luukka 2019; Weng et al. 2018; Ampomah et al. 2020). The positive predictive values and negative predictive values indicate that there is little asymmetry between the up and down prediction classifications.

There are several different avenues for future research. First, this paper has focused on the comparison between bagging decision trees, RFs, and logit models. A deeper analysis could include other ML methods like boosting, ANN and SVM. Second, this paper used a set of well-known technical indicators for features. The feature space could be expanded to include additional technical indicators or other variables like oil prices or other macroeconomic variables. Third, the analysis in this paper was conducted using ETFs. It may also be of interest to apply machine learning techniques to company specific clean energy stock price prediction.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data used in this study are available from Yahoo Finance at https://finance.yahoo.com/.

**Acknowledgments:** I thank the Schulich School of Business for internal financial support. I thank the reviewers for helpful comments.

**Conflicts of Interest:** The author declares no conflict of interest.

## References

Ampomah, Ernest Kwame, Zhiguang Qin, and Gabriel Nyame. 2020. Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement. *Information* 11: 332. [CrossRef]

Andreoni, Valeria. 2020. The Energy Metabolism of Countries: Energy Efficiency and Use in the Period That Followed the Global Financial Crisis. *Energy Policy* 139: 111304. [CrossRef]

Atsalakis, George S., and Kimon P. Valavanis. 2009. Surveying Stock Market Forecasting Techniques—Part II: Soft Computing Methods. *Expert Systems with Applications* 36, Pt 2: 5932–41. [CrossRef]

Ballings, Michel, Dirk Van den Poel, Nathalie Hespeels, and Ruben Gryp. 2015. Evaluating Multiple Classifiers for Stock Price Direction Prediction. *Expert Systems with Applications* 42: 7046–56. [CrossRef]

Basak, Suryoday, Saibal Kar, Snehanshu Saha, Luckyson Khaidem, and Sudeepa Roy Dey. 2019. Predicting the Direction of Stock Market Prices Using Tree-Based Classifiers. *The North American Journal of Economics and Finance* 47: 552–67. [CrossRef]

Bohl, Martin T., Philipp Kaufmann, and Patrick M. Stephan. 2013. From Hero to Zero: Evidence of Performance Reversal and Speculative Bubbles in German Renewable Energy Stocks. *Energy Economics* 37: 40–51. [CrossRef]

Bondia, Ripsy, Sajal Ghosh, and Kakali Kanjilal. 2016. International Crude Oil Prices and the Stock Prices of Clean Energy and Technology Companies: Evidence from Non-Linear Cointegration Tests with Unknown Structural Breaks. *Energy* 101: 558–65. [CrossRef]

Breiman, Leo. 2001. Random Forests. *Machine Learning* 45: 5–32. [CrossRef]

Breiman, Leo, Adele Cutler, Andy Liaw, and Matthew Wiener. 2018. RandomForest: Breiman and Cutler's Random Forests for Classification and Regression. R Package Version 4.6-14. Available online: https://www.stat.berkeley.edu/~breiman/RandomForests/ (accessed on 25 August 2020).

Bustos, O, and A. Pomares-Quimbaya. 2020. Stock Market Movement Forecast: A Systematic Review. *Expert Systems with Applications* 156: 113464. [CrossRef]

Christoffersen, Peter F., and Francis X. Diebold. 2006. Financial Asset Returns, Direction-of-Change Forecasting, and Volatility Dynamics. *Management Science* 52: 1273–87. [CrossRef]

Dawar, Ishaan, Anupam Dutta, Elie Bouri, and Tareq Saeed. 2021. Crude Oil Prices and Clean Energy Stock Indices: Lagged and Asymmetric Effects with Quantile Regression. *Renewable Energy* 163: 288–99. [CrossRef]

Dutta, Anupam. 2017. Oil Price Uncertainty and Clean Energy Stock Returns: New Evidence from Crude Oil Volatility Index. *Journal of Cleaner Production* 164: 1157–66. [CrossRef]

Dutta, Anupam, Elie Bouri, and Md Hasib Noor. 2018. Return and Volatility Linkages between CO2 Emission and Clean Energy Stock Prices. *Energy* 164: 803–10. [CrossRef]

Elie, Bouri, Jalkh Naji, Anupam Dutta, and Gazi Salah Uddin. 2019. Gold and Crude Oil as Safe-Haven Assets for Clean Energy Stock Indices: Blended Copulas Approach. *Energy* 178: 544–53. [CrossRef]

Frankfurt School-UNEP Centre/BNEF. 2020. *Global Trends in Renewable Energy Investment 2020*. Frankfurt am Main: Frankfurt School of Finance & Management gGmbH.

Ghoddusi, Hamed, Germán G. Creamer, and Nima Rafizadeh. 2019. Machine Learning in Energy Economics and Finance: A Review. *Energy Economics* 81: 709–27. [CrossRef]

Gray, Wesley, and Jack Vogel. 2016. *Quantitative Momentum: A Practitioner's Guide to Building a Momentum-Based Stock Selection System*. Hoboken: John Wiley & Sons.

Gupta, Kartick. 2017. Do Economic and Societal Factors Influence the Financial Performance of Alternative Energy Firms? *Energy Economics* 65: 172–82. [CrossRef]

Henrique, Bruno Miranda, Vinicius Amorim Sobreiro, and Herbert Kimura. 2019. Literature Review: Machine Learning Techniques Applied to Financial Market Prediction. *Expert Systems with Applications* 124: 226–51. [CrossRef]

Henriques, Irene, and Perry Sadorsky. 2008. Oil Prices and the Stock Prices of Alternative Energy Companies. *Energy Economics* 30: 998–1010. [CrossRef]

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. New York: Springer. [CrossRef]

Khan, Wasiat, Mustansar Ali Ghazanfar, Muhammad Awais Azam, Amin Karami, Khaled H. Alyoubi, and Ahmed S. Alfakeeh. 2020. Stock Market Prediction Using Machine Learning Classifiers and Social Media, News. *Journal of Ambient Intelligence and Humanized Computing*. [CrossRef]

Kumar, Surender, Shunsuke Managi, and Akimi Matsuda. 2012. Stock Prices of Clean Energy Firms, Oil and Carbon Markets: A Vector Autoregressive Analysis. *Energy Economics* 34: 215–26. [CrossRef]

Leung, Mark T., Hazem Daouk, and An-Sing Chen. 2000. Forecasting Stock Indices: A Comparison of Classification and Level Estimation Models. *International Journal of Forecasting* 16: 173–90. [CrossRef]

Lo, Andrew W., Harry Mamaysky, and Jiang Wang. 2000. Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation. *The Journal of Finance* 55: 1705–65. [CrossRef]

Lohrmann, Christoph, and Pasi Luukka. 2019. Classification of Intraday S&P500 Returns with a Random Forest. *International Journal of Forecasting* 35: 390–407. [CrossRef]

Maghyereh, Aktham I., Basel Awartani, and Hussein Abdoh. 2019. The Co-Movement between Oil and Clean Energy Stocks: A Wavelet-Based Analysis of Horizon Associations. *Energy* 169: 895–913. [CrossRef]

Malkiel, Burton G. 2003. The Efficient Market Hypothesis and Its Critics. *Journal of Economic Perspectives* 17: 59–82. [CrossRef]

Mallqui, Dennys C. A., and Ricardo A. S. Fernandes. 2019. Predicting the Direction, Maximum, Minimum and Closing Prices of Daily Bitcoin Exchange Rate Using Machine Learning Techniques. *Applied Soft Computing* 75: 596–606. [CrossRef]

Managi, Shunsuke, and Tatsuyoshi Okimoto. 2013. Does the Price of Oil Interact with Clean Energy Prices in the Stock Market? *Japan and the World Economy* 27: 1–9. [CrossRef]

Mokoaleli-Mokoteli, Thabang, Shaun Ramsumar, and Hima Vadapalli. 2019. The Efficiency of Ensemble Classifiers in Predicting the Johannesburg Stock Exchange All-Share Index Direction. *Journal of Financial Management, Markets and Institutions* 7: 1950001. [CrossRef]

Moskowitz, Tobias J., Yao Hua Ooi, and Lasse Heje Pedersen. 2012. Time Series Momentum. *Journal of Financial Economics* 104: 228–50. [CrossRef]

Mullainathan, Sendhil, and Jann Spiess. 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives* 31: 87–106. [CrossRef]

Mundaca, Luis, and Jessika Luth Richter. 2015. Assessing 'Green Energy Economy' Stimulus Packages: Evidence from the U.S. Programs Targeting Renewable Energy. *Renewable and Sustainable Energy Reviews* 42: 1174–86. [CrossRef]

Neely, Christopher J., David E. Rapach, Jun Tu, and Guofu Zhou. 2014. Forecasting the Equity Risk Premium: The Role of Technical Indicators. *Management Science* 60: 1772–91. [CrossRef]

Nti, Isaac Kofi, Adebayo Felix Adekoya, and Benjamin Asubam Weyori. 2020. A Comprehensive Evaluation of Ensemble Learning for Stock-Market Prediction. *Journal of Big Data* 7: 20. [CrossRef]

Nyberg, Henri. 2011. Forecasting the Direction of the US Stock Market with Dynamic Binary Probit Models. *International Journal of Forecasting* 27: 561–78. [CrossRef]

Nyberg, Henri, and Harri Pönkä. 2016. International Sign Predictability of Stock Returns: The Role of the United States. *Economic Modelling* 58: 323–38. [CrossRef]

Park, Cheol-Ho, and Scott H. Irwin. 2007. What Do We Know About the Profitability of Technical Analysis? *Journal of Economic Surveys* 21: 786–826. [CrossRef]

Pönkä, Harri. 2016. Real Oil Prices and the International Sign Predictability of Stock Returns. *Finance Research Letters* 17: 79–87. [CrossRef]

R Core Team. 2019. *R: A Language and Environment for Statistical Computing (Version R Version 3.6.0 (2019-04-26))*. Vienna: The R Project for Statistical Computing. Available online: https://www.r-project.org/ (accessed on 25 August 2020).

Reboredo, Juan C. 2015. Is There Dependence and Systemic Risk between Oil and Renewable Energy Stock Prices? *Energy Economics* 48: 32–45. [CrossRef]

Reboredo, Juan C., and Andrea Ugolini. 2018. The Impact of Energy Prices on Clean Energy Stock Prices. A Multivariate Quantile Dependence Approach. *Energy Economics* 76: 136–52. [CrossRef]

Reboredo, Juan C., Miguel A. Rivera-Castro, and Andrea Ugolini. 2017a. Wavelet-Based Test of Co-Movement and Causality between Oil and Renewable Energy Stock Prices. *Energy Economics* 61: 241–52. [CrossRef]

Reboredo, Juan C., Miguel Quintela, and Luis A. Otero. 2017b. Do Investors Pay a Premium for Going Green? Evidence from Alternative Energy Mutual Funds. *Renewable and Sustainable Energy Reviews* 73: 512–20. [CrossRef]

Sadorsky, Perry. 2012. Modeling Renewable Energy Company Risk. *Energy Policy* 40: 39–48. [CrossRef]

Shah, Dev, Haruna Isah, and Farhana Zulkernine. 2019. Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. *International Journal of Financial Studies* 7: 26. [CrossRef]

The Economist. 2020. The World's Energy System Must Be Transformed Completely. *The Economist*.

Uddin, Gazi Salah, Md Lutfur Rahman, Axel Hedström, and Ali Ahmed. 2019. Cross-Quantilogram-Based Correlation and Dependence between Renewable Energy Stock and Other Asset Classes. *Energy Economics* 80: 743–59. [CrossRef]

Wang, Yudong, Li Liu, and Chongfeng Wu. 2020. Forecasting Commodity Prices Out-of-Sample: Can Technical Indicators Help? *International Journal of Forecasting* 36: 666–83. [CrossRef]

Wen, Xiaoqian, Yanfeng Guo, Yu Wei, and Dengshi Huang. 2014. How Do the Stock Prices of New Energy and Fossil Fuel Companies Correlate? Evidence from China. *Energy Economics* 41: 63–75. [CrossRef]

Weng, Bin, Lin Lu, Xing Wang, Fadel M. Megahed, and Waldyn Martinez. 2018. Predicting Short-Term Stock Prices Using Ensemble Methods and Online Data Sources. *Expert Systems with Applications* 112: 258–73. [CrossRef]

Yin, Libo, and Qingyuan Yang. 2016. Predicting the Oil Prices: Do Technical Indicators Help? *Energy Economics* 56: 338–50. [CrossRef]

Yin, Libo, Qingyuan Yang, and Zhi Su. 2017. Predictability of Structural Co-Movement in Commodity Prices: The Role of Technical Indicators. *Quantitative Finance* 17: 795–812. [CrossRef]