Model Evaluation

Kevin Lu and Jevin Maltais

February 7, 2018

1. Introduction

This document summarises a series of analyses performed consisting of feature engineering, modeling, and evaluating model performance. A candidate model is presented and the model's predicted position labels are compared with the ground truth position labels.

2. Feature Engineering

2.1 Source of Ideas

Feature engineering was conducted and ideas were heavily inspired by the following sources:

- The verbal explanation of the trading system
- Standard time-series or absolute momentum-based trading strategy research contained within the academic literature, including The Kinetics of the Stock Markets by Hsu and Lin (2001) and the concept concept of how excess demand affects stock prices
- Technical analysis and indicators, especially ideas related to the Elliot wave principle and other technical indicators in the momentum category

2.2 Chosen Features

In total, approximately 45 types of features were created and examined. All features were created as a function of an individual stock's open, high, low, and close price as well as the stock's volume. After examining the set of 45 types of features, the following 15 feature types were chosen:

1. Geometric return of the closing price
2. Percent drawdown of the closing price
3. Percent drawup of the closing price
4. Number of days with a positive geometric return
5. Volatiity of the one-day geometric return
6. Relative strength index (RSI) indicator
7. Aroon family of indicators consisting of aroon up, aroon down, and aroon oscillator
8. Commodity Channel Index (CCI) indicator
9. Chaikin Volatility indicator
10. Chaikin Money Flow indicator
11. Signal-to-noise indicator
12. William's % R indicator
13. Money Flow Index indicator
14. Chande Momentum Oscillator (CMO) indicator

15. Vertical Horizonal Filter (VHF) indicator

A complete list of the technical indicators that were created and examined can be found in the documentation for the TTR: Technical Trading Rules R package. The documentation also contains links to references that fully describe the calculation of each technical indicator.

For each feature type, 14 individual features were created at the following rolling window lengths: 1 week, 2 weeks, and every month from 1 to 12 months – therefore, each of the feature types described above consist of 14 individual features that are calculated at various rolling window lengths. In total, 240 individual features were created and chosen.

3. Modeling and Cross Validation Framework

This section contains some technical details on regarding the exploratory research that was conducted. You can safely skip this section and move on to the section on evaluating model performance.

Several models using various machine learning algorithms were evaluated to determine which model produced the best predictions. Model predictions were evaluated using a cross validation framework. The cross validation framework used of 26-fold cross validation, where all training observations that belong to an individual stock are assigned to one fold. This cross validation closely replicates how a model would be used in production since the model will be trained on a training dataset consisting of 26 stocks and be used to create predictions on a much wider universe of stocks. The evaluation metric chosen was RMSE.

First, a naive baseline model was created that uses the global mean of the position labels as the prediction for all observations. Next, a collection of models using various machine learning algorithms were trained and the performance was compared to the baseline model. The algorithms examined include linear regression, decision trees, lasso regression, ridge regression, elastic net regression, random forest regression, extra trees, gradient boosted decision trees, and gradient boosted linear models. All models significantly improved upon the baseline model.

An ensemble model using a gradient boosted linear model algorithm implemented in the xgboost library was chosen based on the model's good performance relative to the baseline model and the model's ability to handle missing values.

4. Evaluating Model Performance

4.1 Background

Machine learning engineers often evaluate model performance using a single evaluation metric – often a numerical score that represents how accurate the model's predictions are. In this case, the commonly used root mean squared error (RMSE) was chosen. The interpretation of the RMSE is the average model prediction error in the same units as the ground truth position labels. Lower values indicate more accurate predictions.

To put this evaluation metric in the context of our problem, if we created a naive model that predicted a position label of 0 for all observations, such a model would score a RMSE of approximately 3.0. If we created a better model that predicted the global mean of the position label for all observations, such a model would score a RMSE of approximately 2.8. The current best performing model scores 1.7.

Deciding what values of RMSE are good or bad is difficult because the RMSE is problem specific. Often times it is illustrative to visually compare the ground truth labels with the predicted labels. The remainder of this document contains a series of plots that allow you to do this.

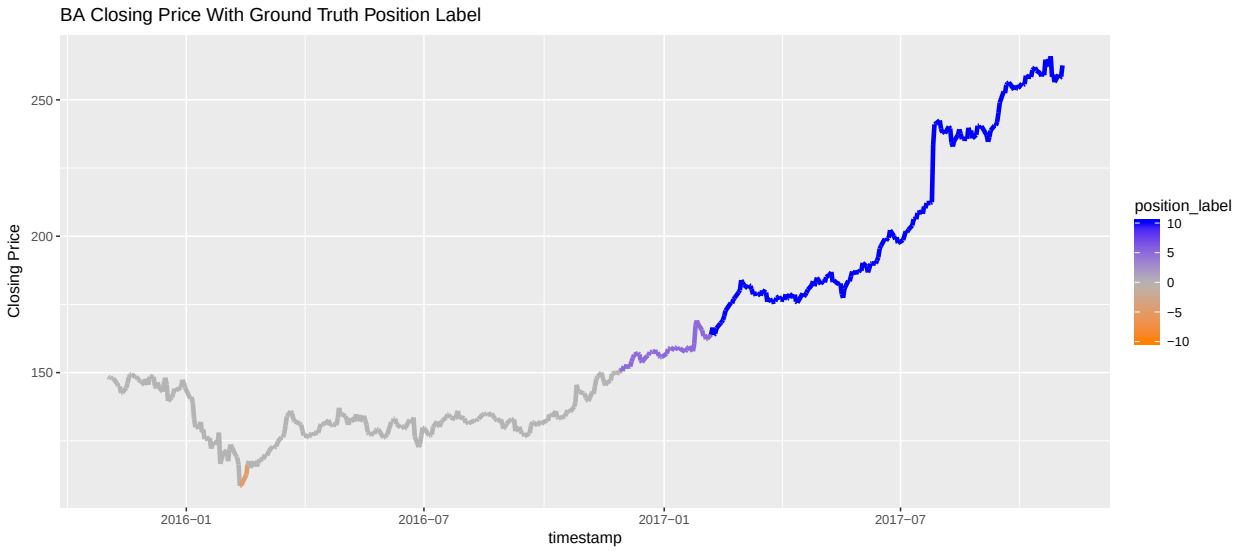
4.2 How to Interpret the Plots

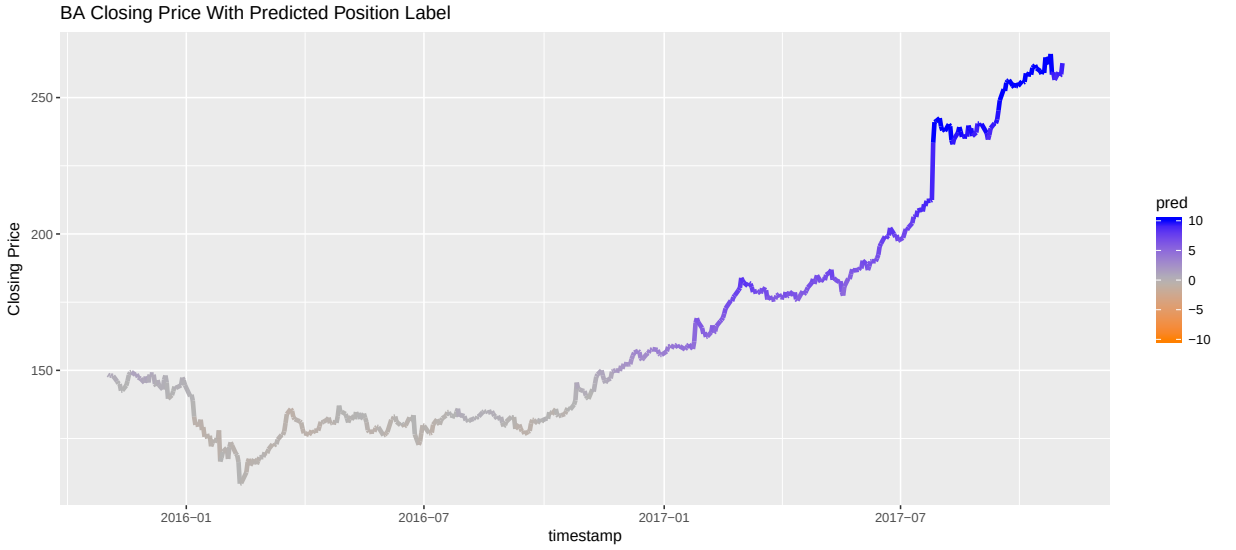
An example plot is included below and this section provides a brief description of how to interpret the plot.

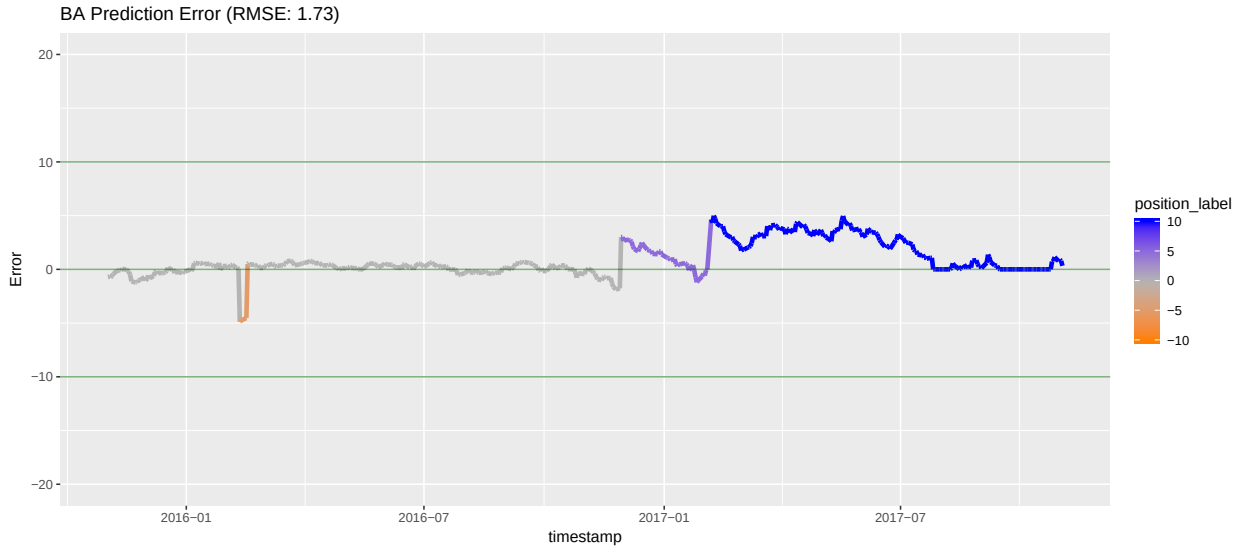
The plot consists of three subplots. All the subplots cover the trainig period – two years of trading history from November 2, 2015 to November 2, 2017. The top subplot contains the stock's closing price with the **ground truth position label** displayed as a color gradient overlayed on top – bright blue indicates a position label of +10, gray indicates a position label of 0, and bright red indicates a position label of -10. Intermediate colors indicate intermediate position labels.

The middle subplot contains the stock's closing price with the **model's predicted position label** displayed as a color gradient overlayed on top. It is important to note that the predicted position labels are out-of-sample predictions – that is, the predictions are generated from a model that is trained on all observations except the observations that belong to the stock being examined. Therefore, the predictions are not overfitted to the stock being examined and accurately represent how accurate the model would perform on truly unknown stocks.

The bottom subplot contains a time series representing the model's prediction error, defined as the ground truth prediction label minus the model's predicted position label. On top of this time series, the ground truth position label is displayed as a color gradient.


BA Closing Price With Ground Truth Position Label
Closing Price
250
200
150
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10


BA Closing Price With Predicted Position Label
Closing Price
250
200
150
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10


BA Prediction Error (RMSE: 1.73)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

5. Examining Model Predictions

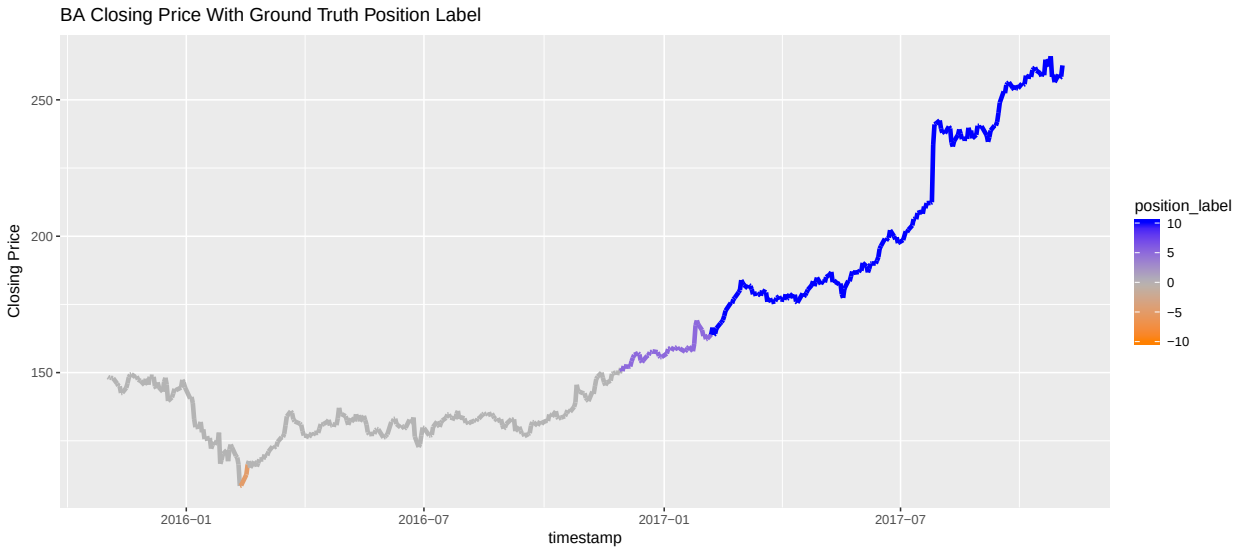
I present the plots grouped in the same clusters identified in the previous update that examined the training data.

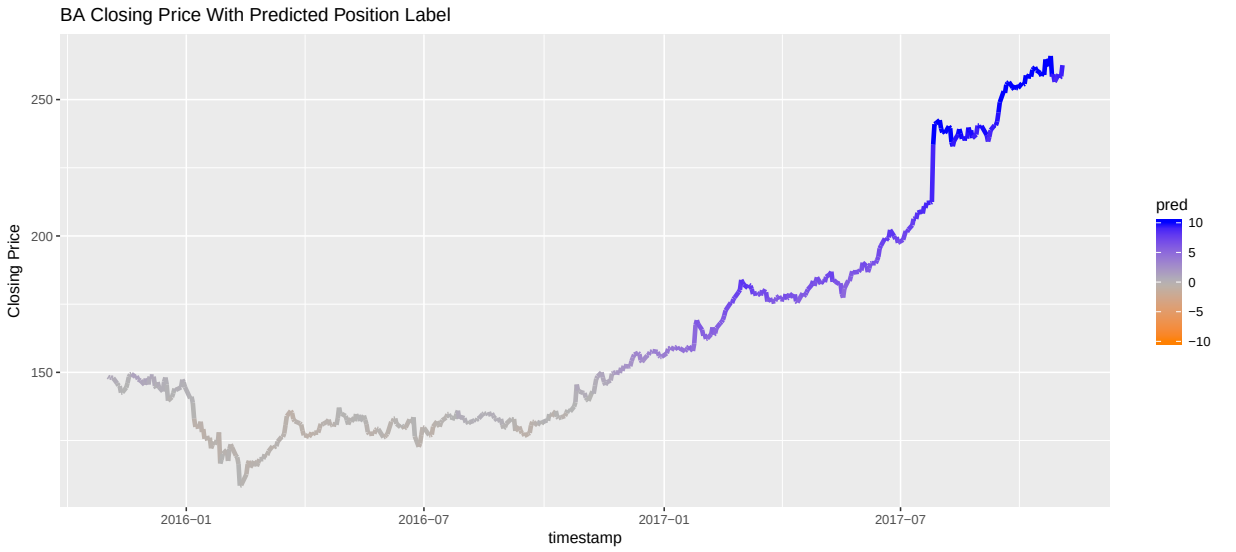
5.1 Strong Upward Trend With Low Volatility

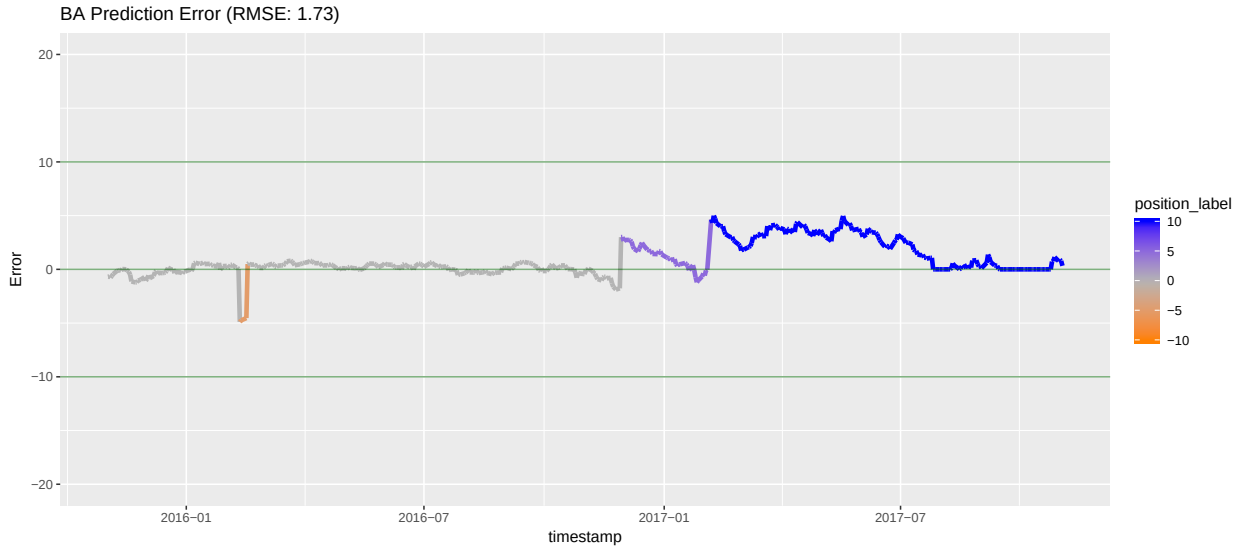
Particular attention is paid to stocks within this cluster that consists of stocks with a strong upward trend with low volatility because these stocks represent the most desireable patterns and trading opportunities as represented by the large number of observations with a +10 ground truth position label.

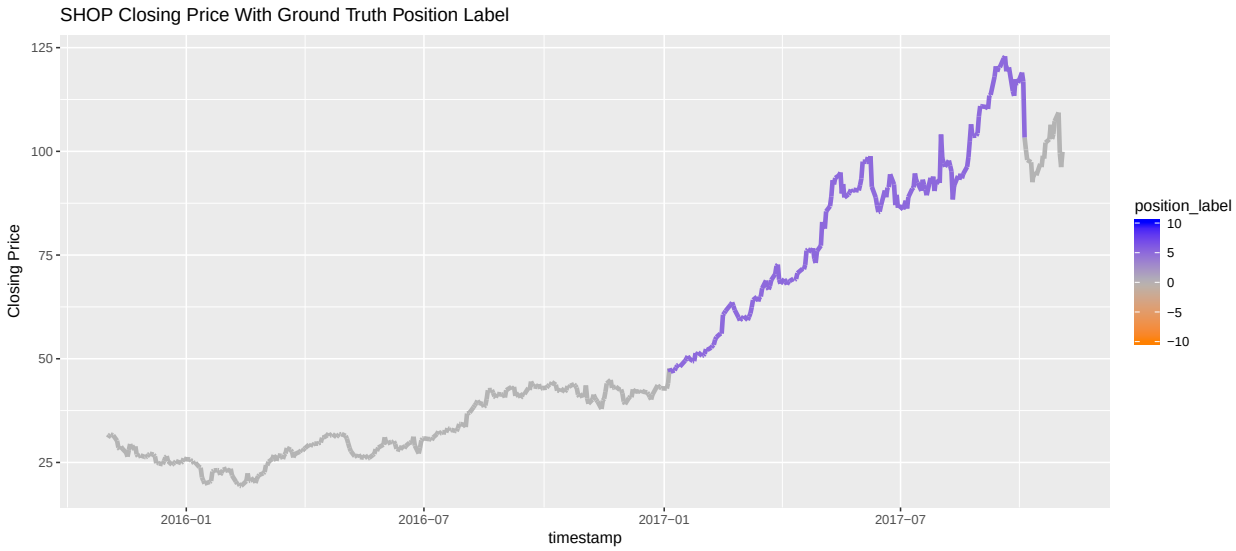
Examination of the plots in this cluster reveal that the model predictions can predict a position label of 0 with almost perfect accuracy. Also directionally, the model predictions track the ground truth position labels quite closely. In particular, the model predictions for SHOP are extremely accurate with a RMSE of 1.38. A closer examination reveals that while the model is almost perfectly directionally accurate (it can predict the direction of the position label with great accuracy), it struggles to differentiate between a ground truth position label of +5 and +10. This can be seen in the BA plot, for example, where the model underpredicted the position label in the first half of 2017 (but still got the direction correct).

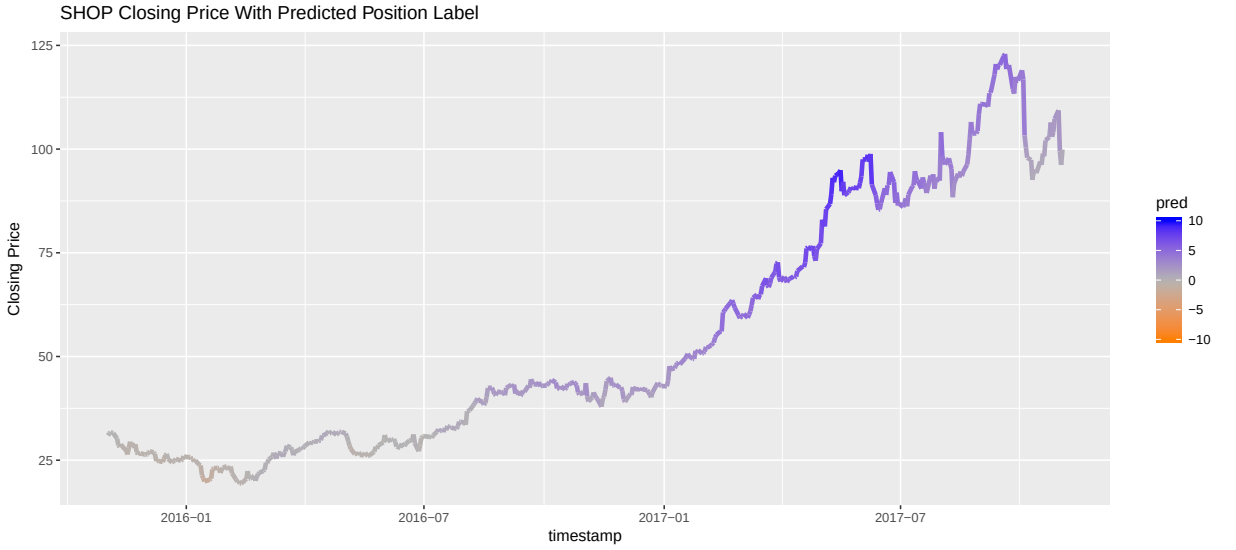
This could be explained because the distinction between a +5 and +10 position label is very subtle and there are insufficient training observations for the model to learn from or the current features are insufficient in this region.

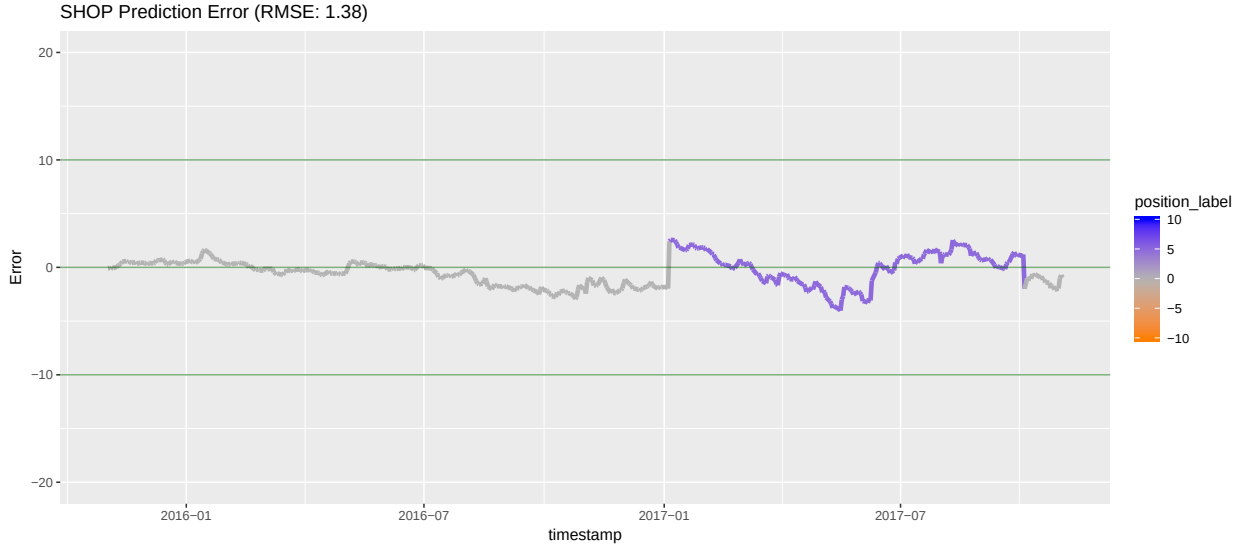

BA Closing Price With Ground Truth Position Label
Closing Price
250
200
150
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

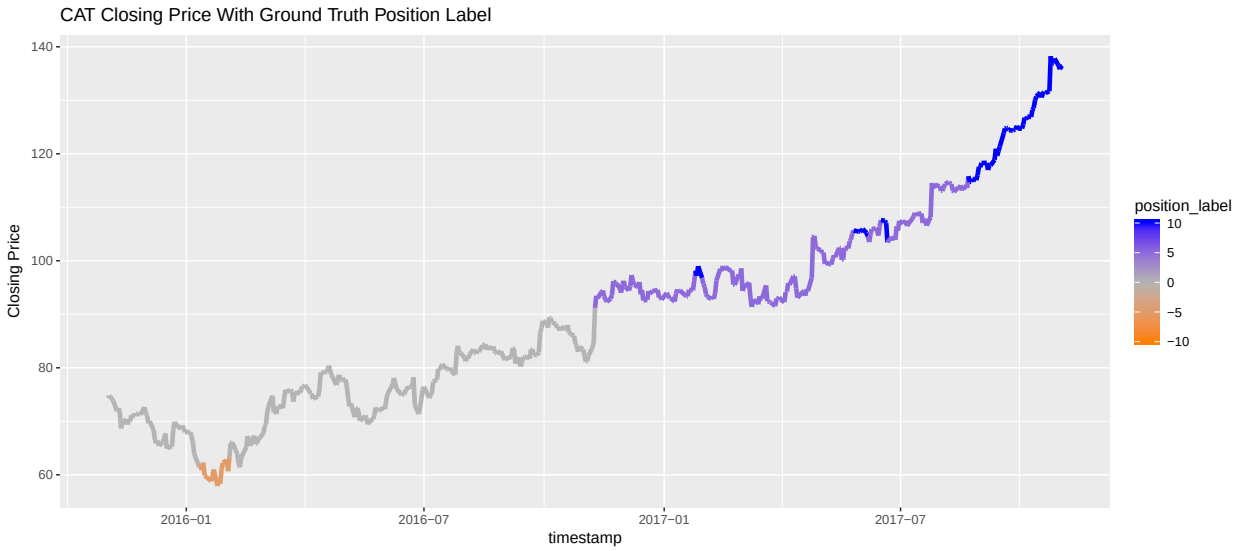

BA Closing Price With Predicted Position Label
Closing Price
250
200
150
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10

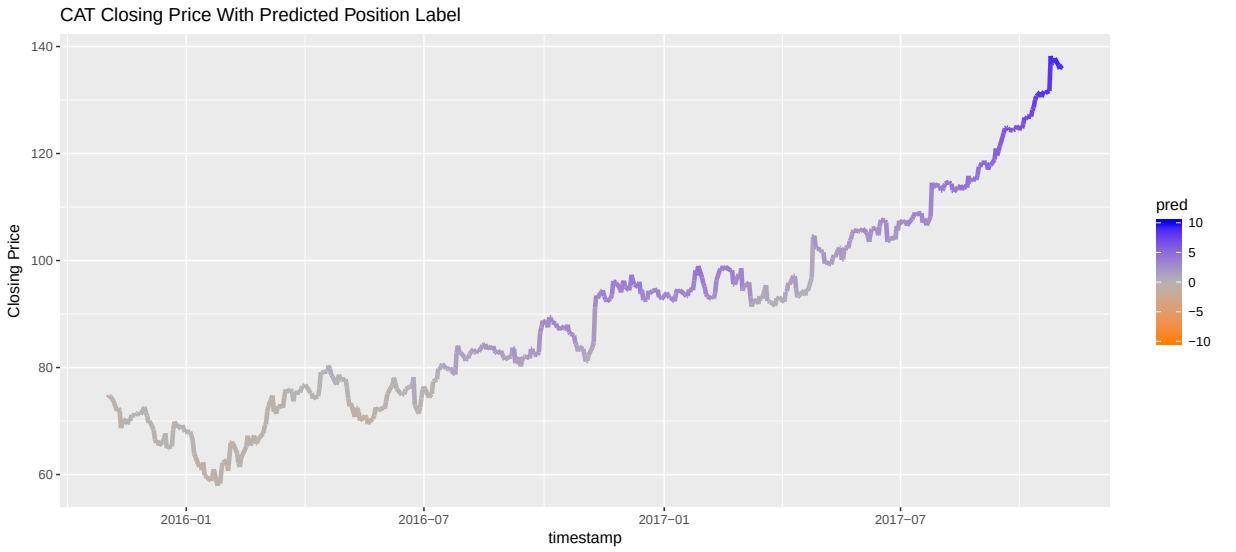

BA Prediction Error (RMSE: 1.73)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

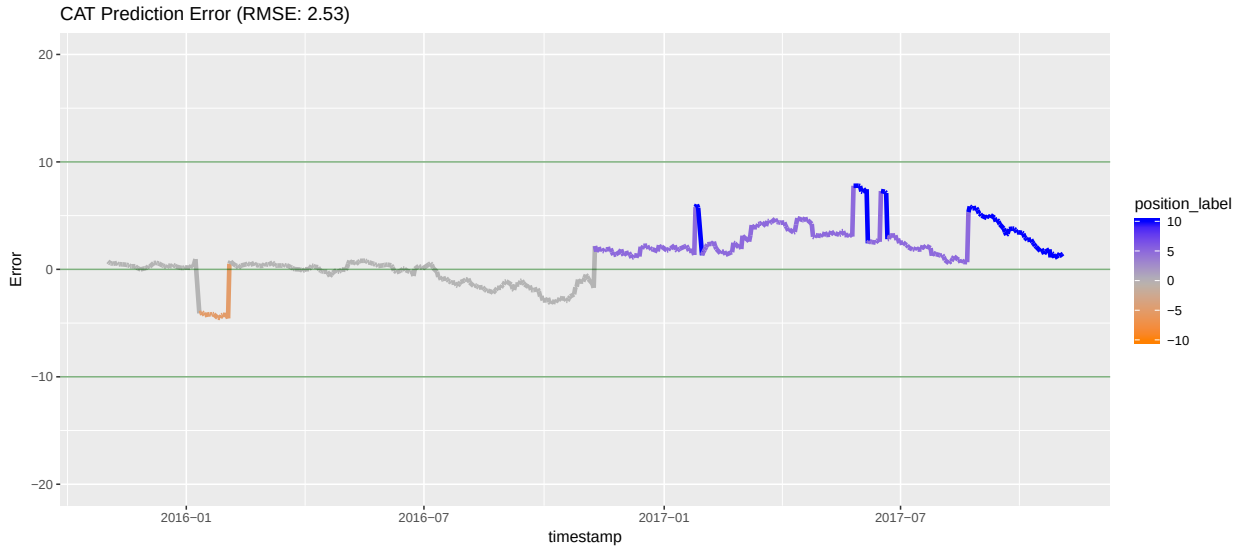

SHOP Closing Price With Ground Truth Position Label
Closing Price
125
100
75
50
25
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10


SHOP Closing Price With Predicted Position Label
Closing Price
125
100
75
50
25
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10

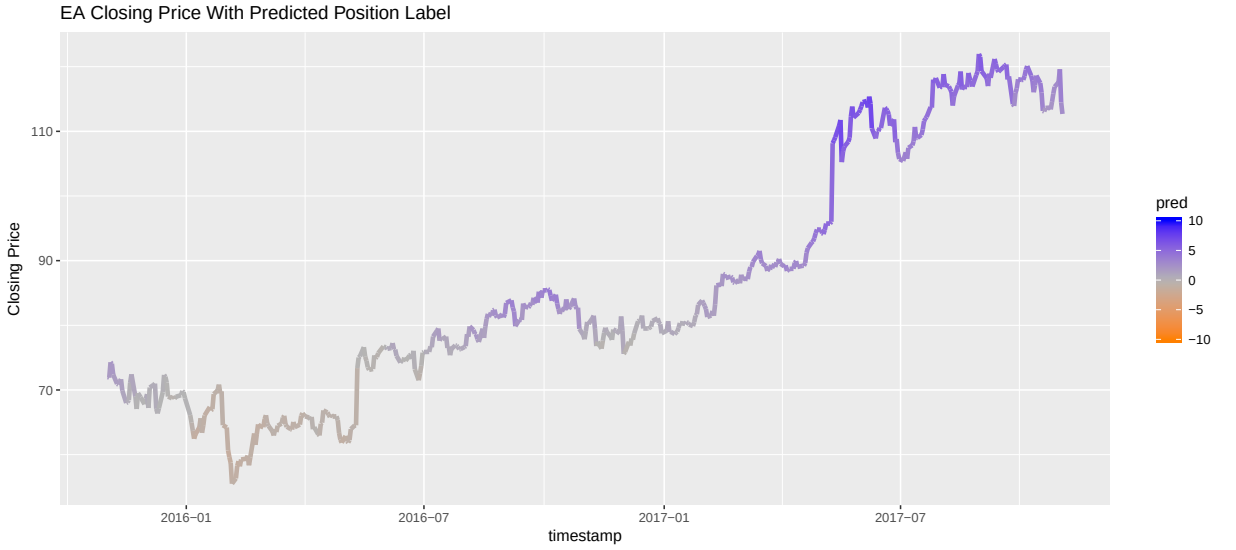

SHOP Prediction Error (RMSE: 1.38)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

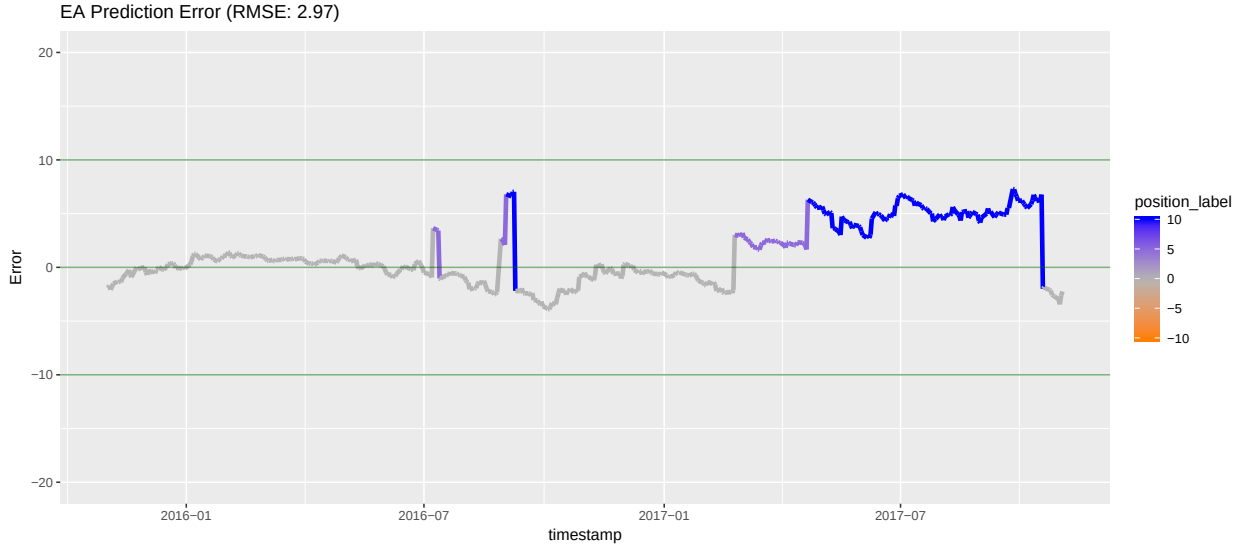

CAT Closing Price With Ground Truth Position Label
Closing Price
140
120
100
80
60
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

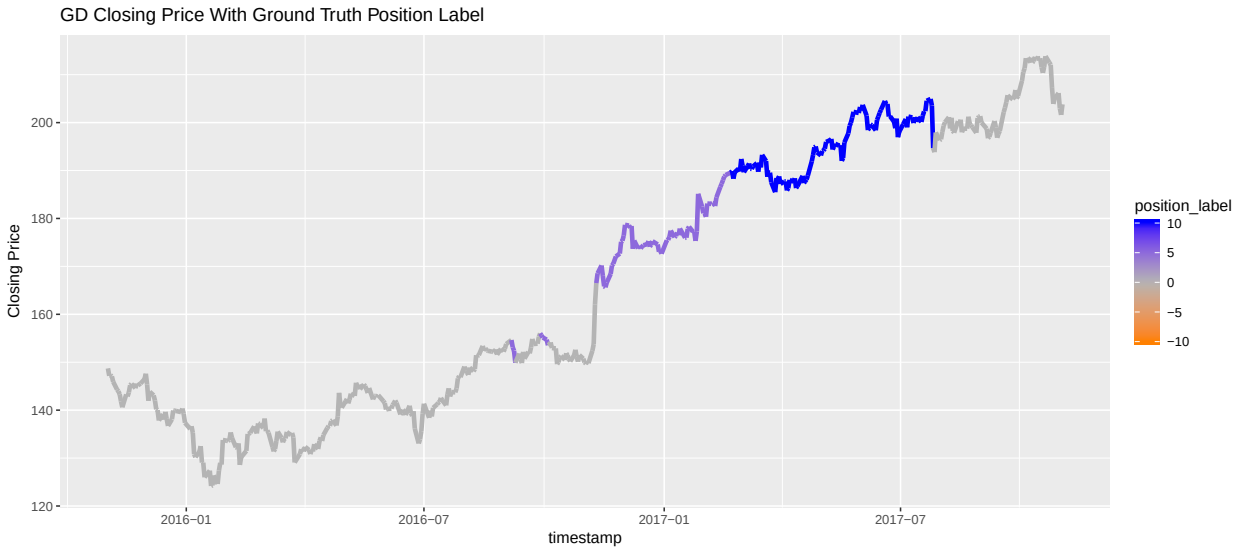

CAT Closing Price With Predicted Position Label
Closing Price
140
120
100
80
60
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10

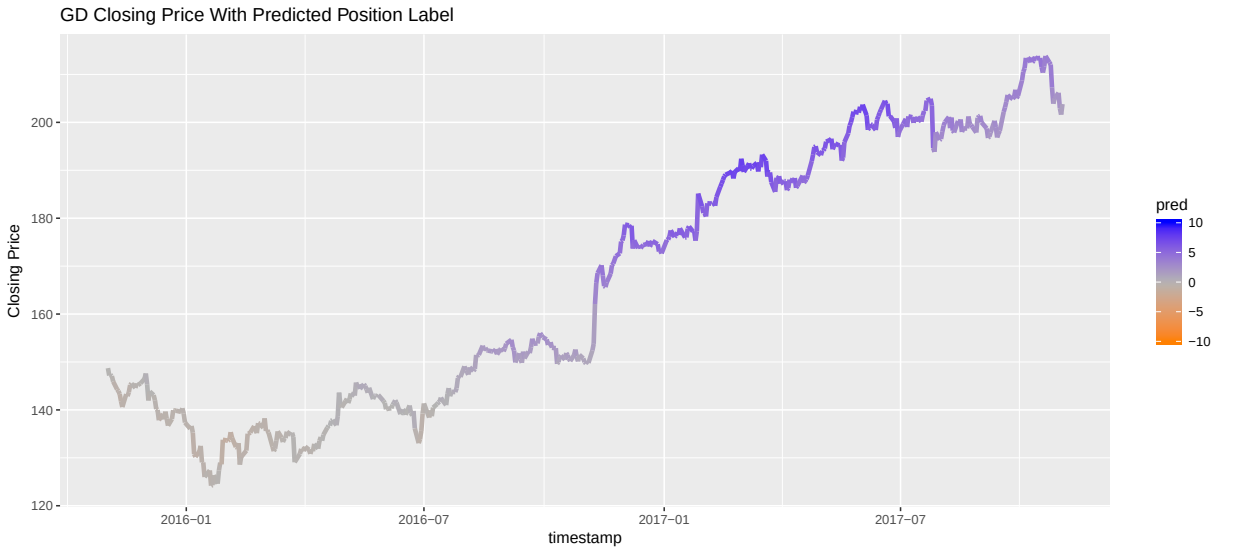

CAT Prediction Error (RMSE: 2.53)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

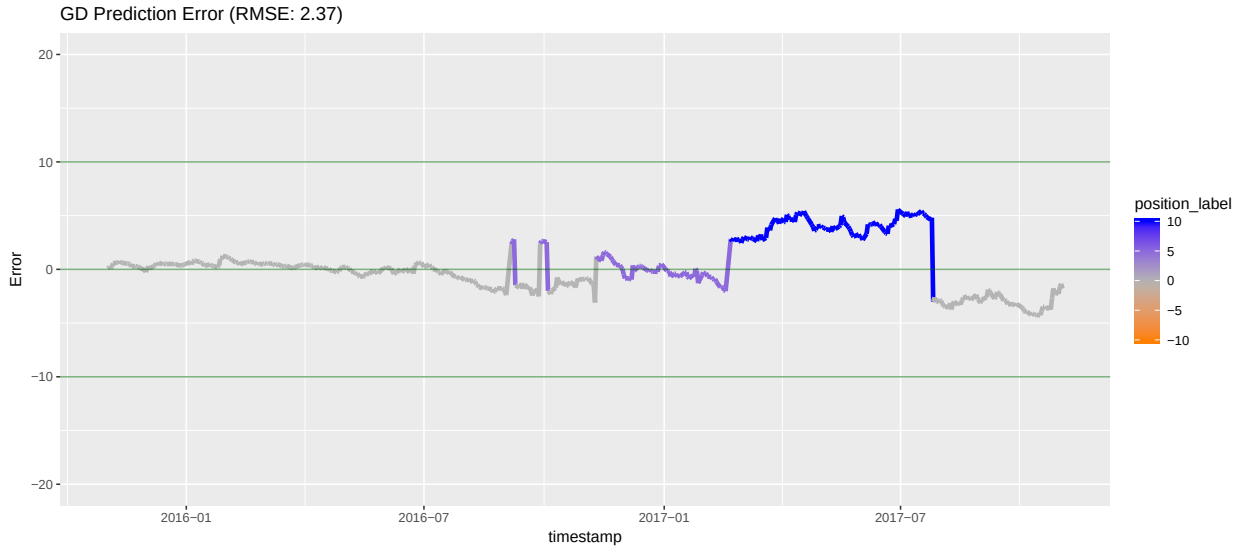

EA Closing Price With Ground Truth Position Label
Closing Price
110
90
70
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

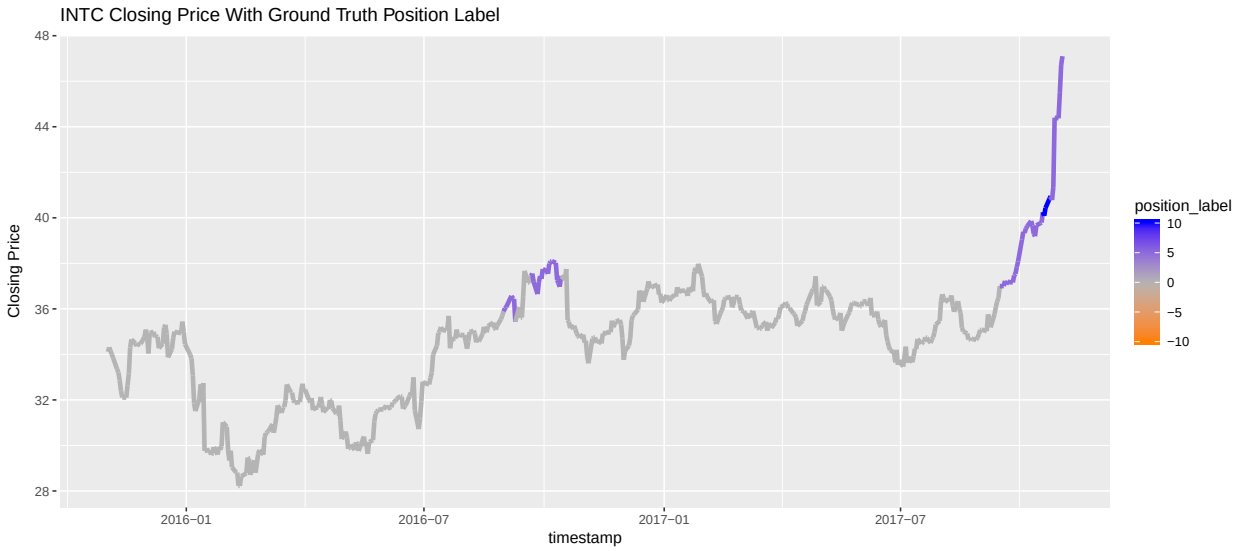

EA Closing Price With Predicted Position Label
Closing Price
110
90
70
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10

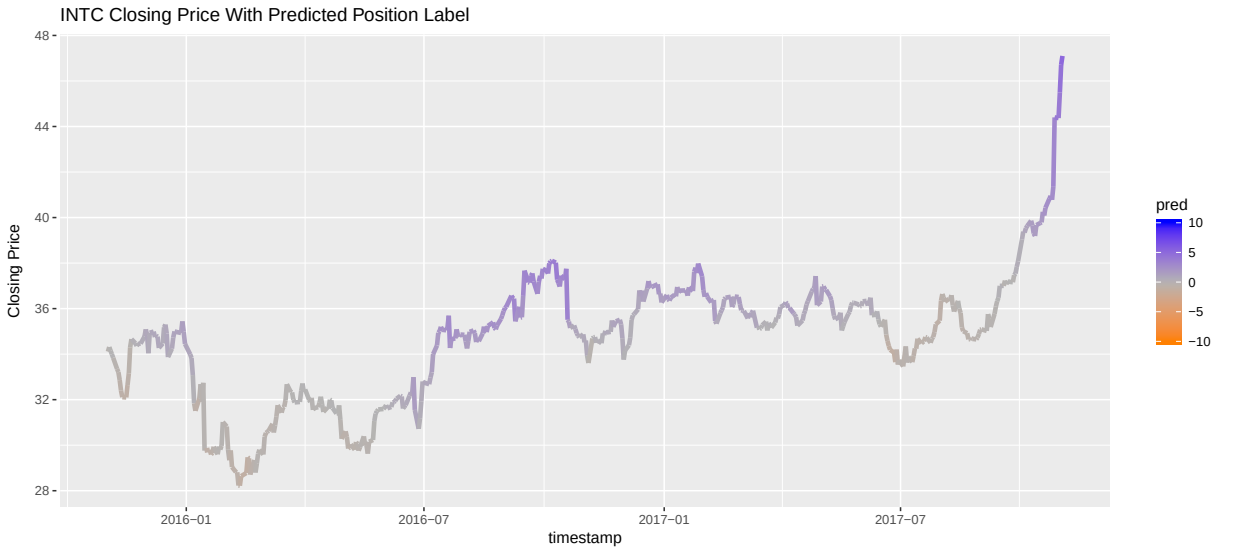

EA Prediction Error (RMSE: 2.97)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

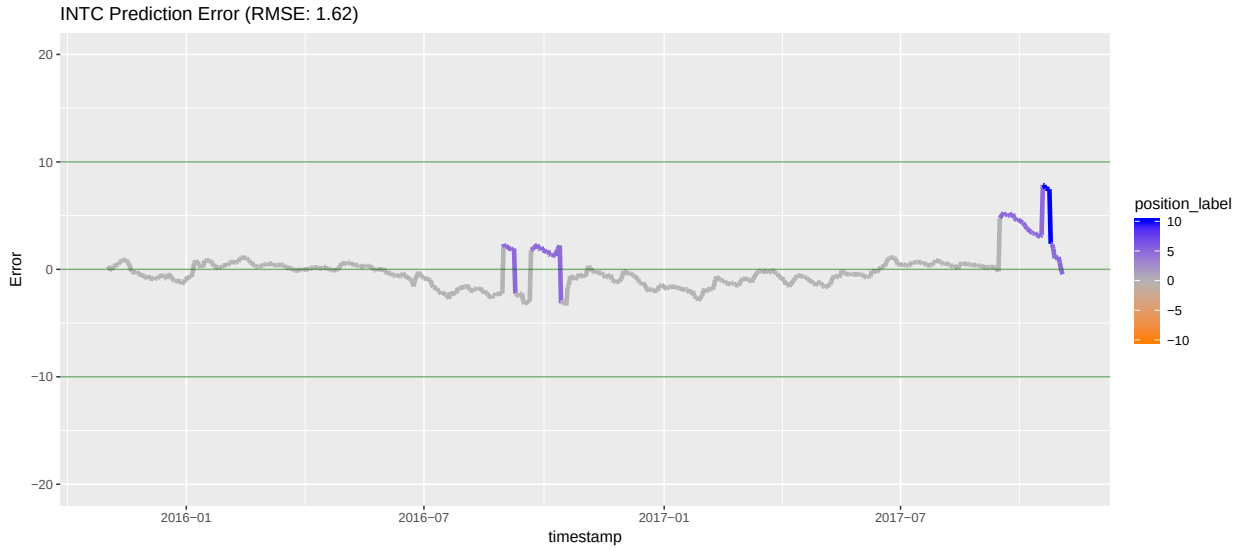

GD Closing Price With Ground Truth Position Label
Closing Price
200
180
160
140
120
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10


GD Closing Price With Predicted Position Label
Closing Price
200
180
160
140
120
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10


GD Prediction Error (RMSE: 2.37)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

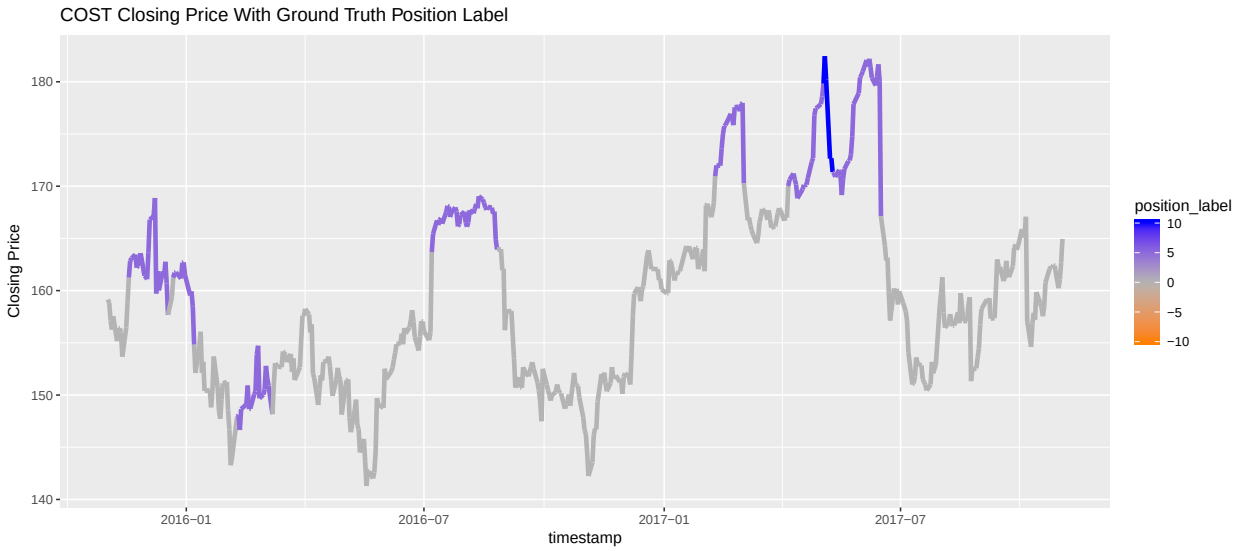

INTC Closing Price With Ground Truth Position Label
Closing Price
48
44
40
36
32
28
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

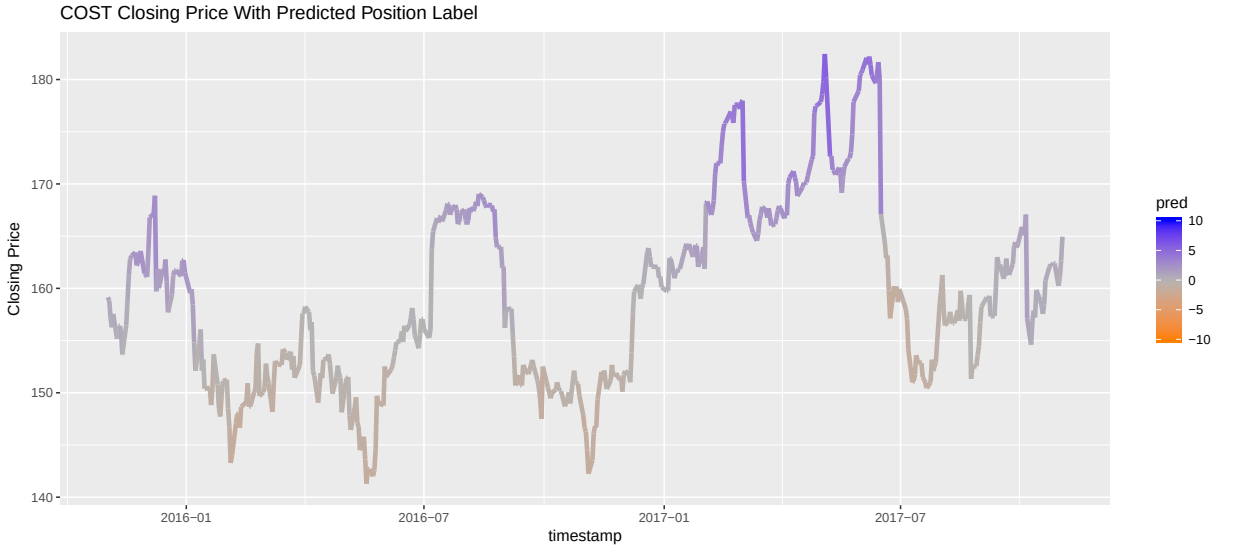

INTC Closing Price With Predicted Position Label
Closing Price
48
44
40
36
32
28
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10

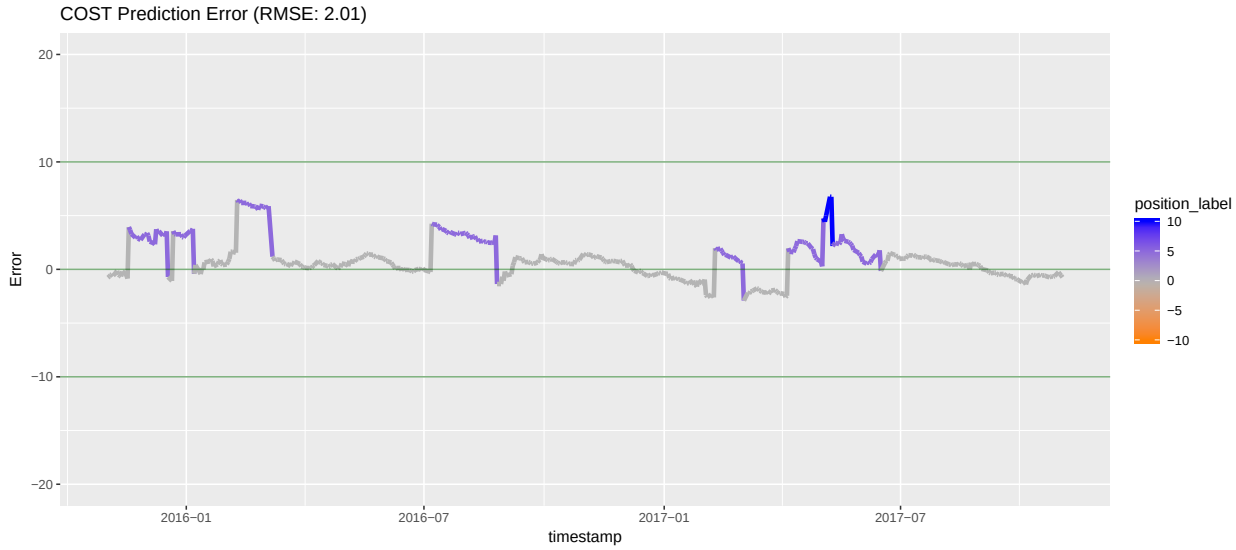

INTC Prediction Error (RMSE: 1.62)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

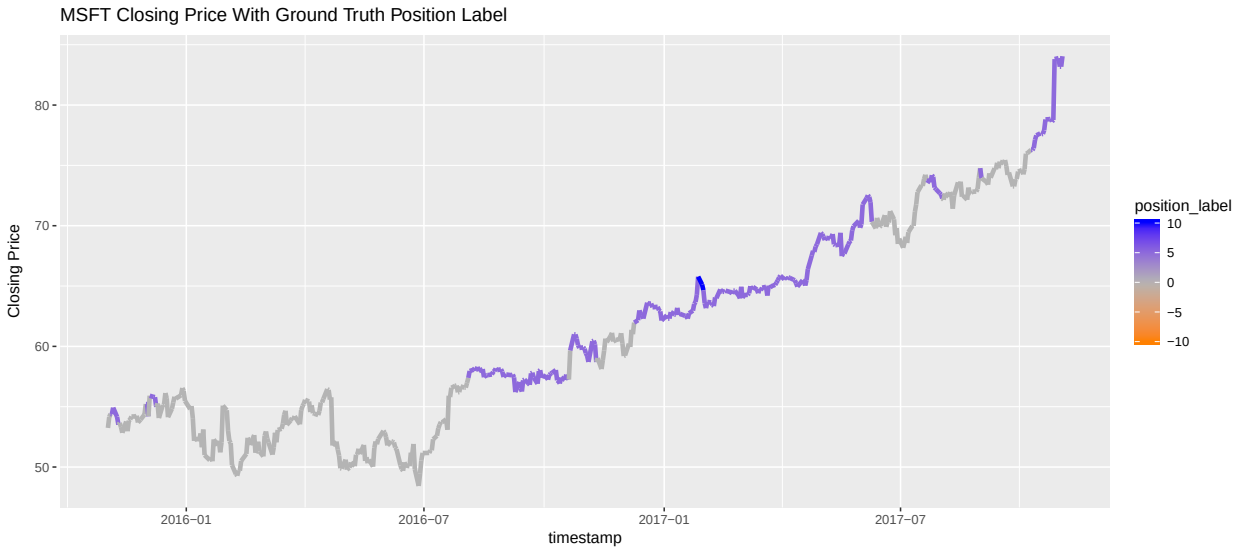
5.2 Upward Trend With High Volatility

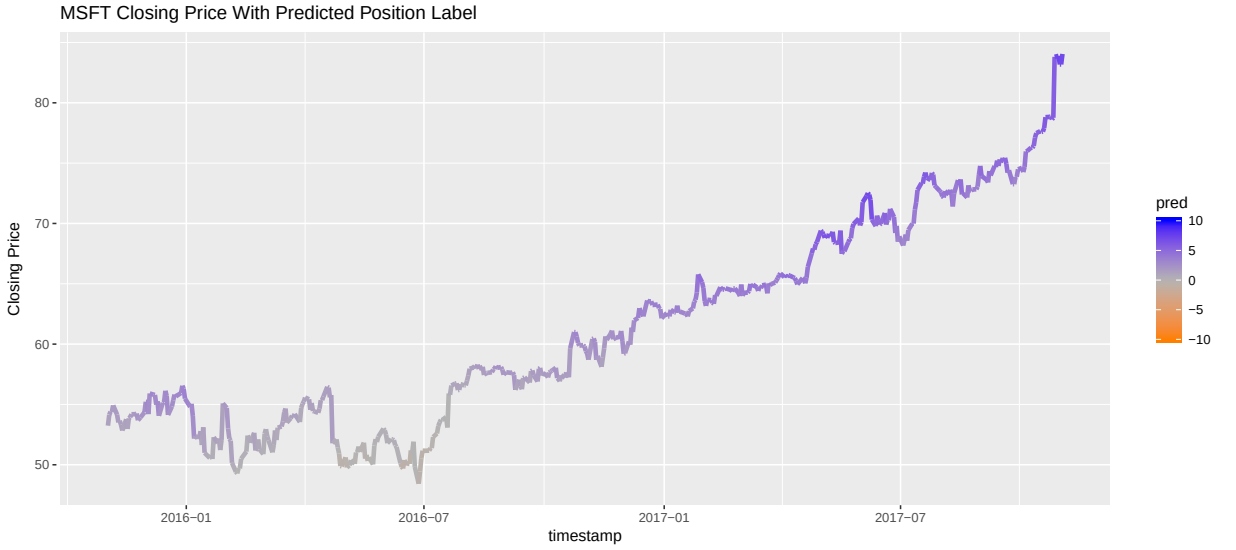
The model performance in this cluster is similar to the previous cluster. Generally, the model can predict the direction of the position label with good accuracy but tends to underpredict many of the +10 position labels in this cluster. There are, however, some isolated instances where the model predicts a positive position label when the ground truth is flat (see MSFT and BAC). For the majority of these instances, the prediction is quite low which reflects the model's uncertainty about the prediction. This should not be a problem when using the model to screen for trading opportunities since only the most confident and extreme predicted positions will be highlighted.

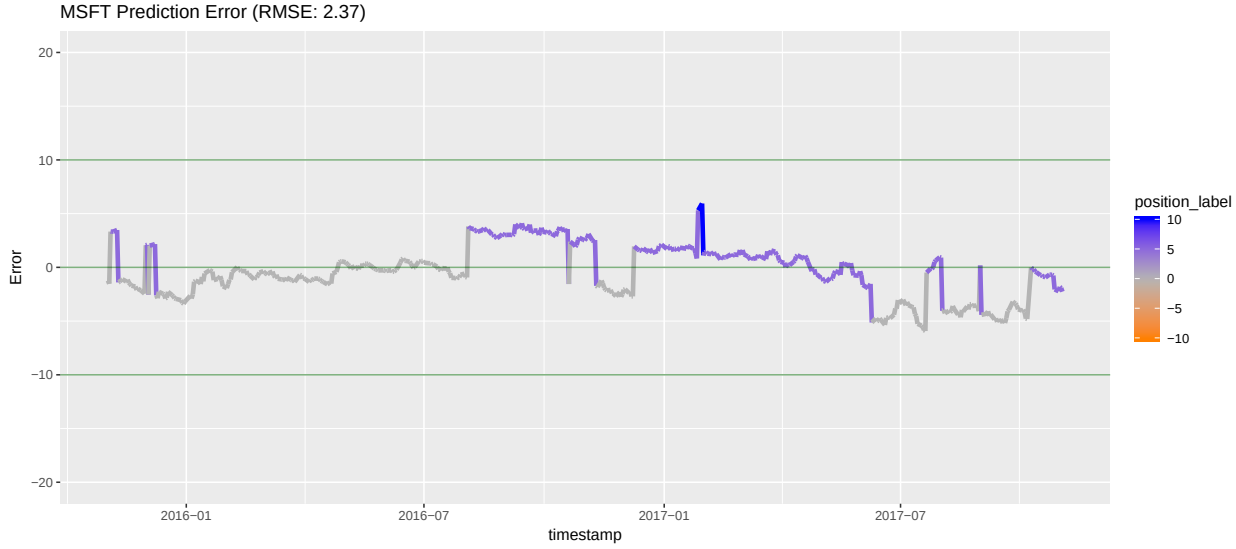

COST Closing Price With Ground Truth Position Label
Closing Price
180
170
160
150
140
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

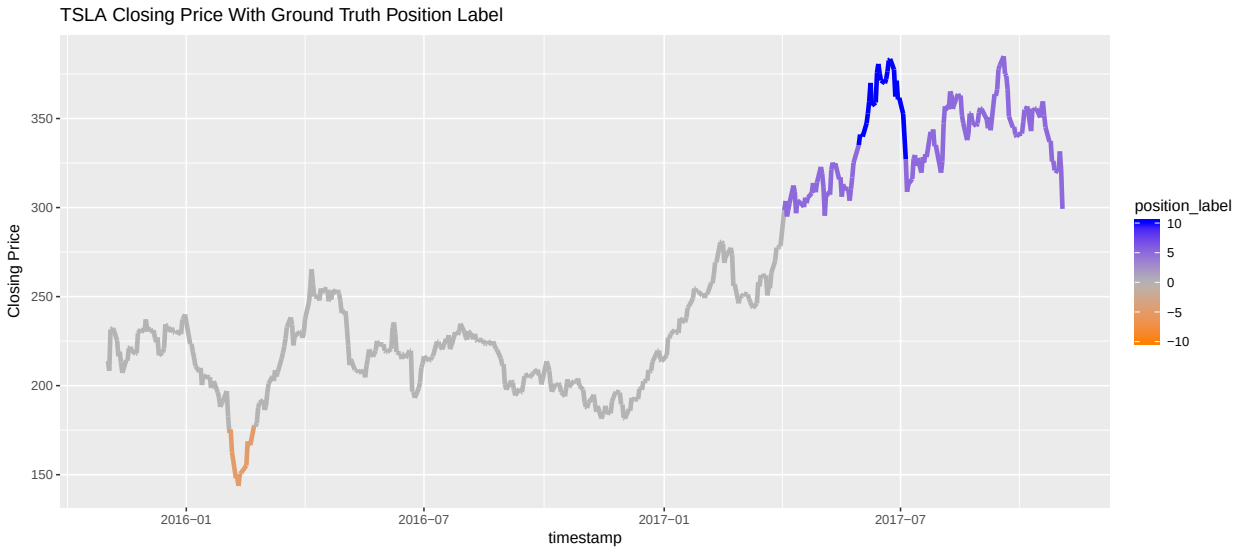

COST Closing Price With Predicted Position Label
Closing Price
180
170
160
150
140
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10

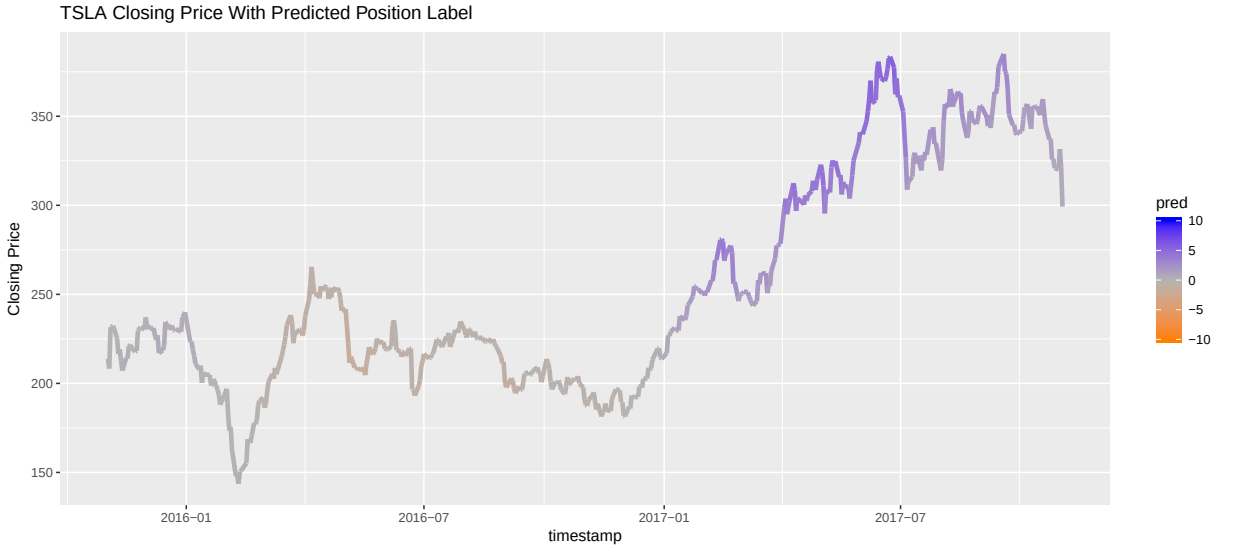

COST Prediction Error (RMSE: 2.01)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

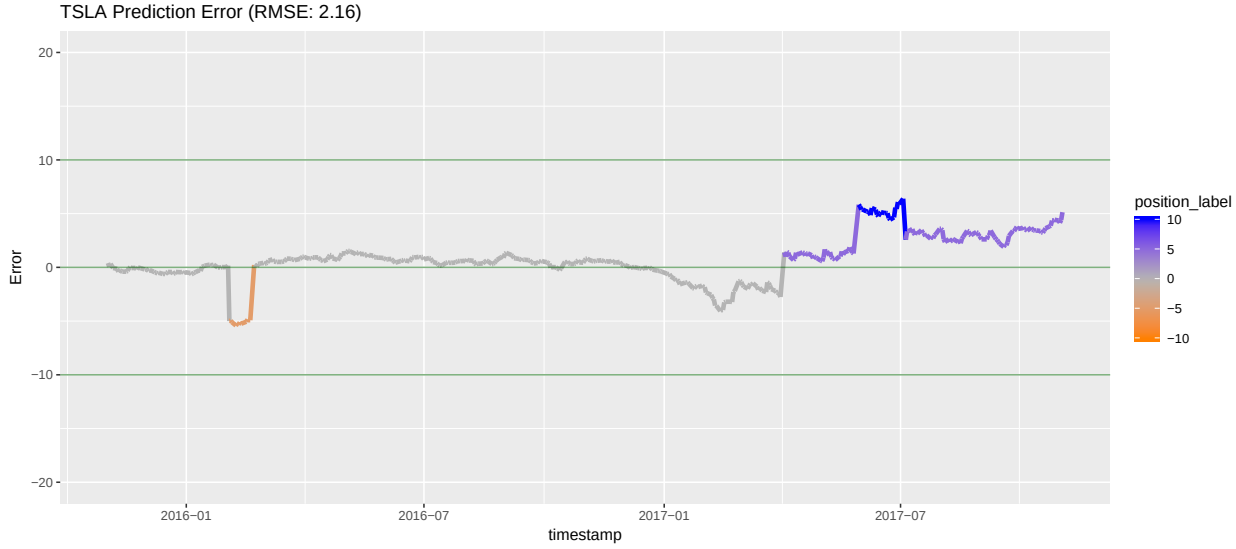

MSFT Closing Price With Ground Truth Position Label
Closing Price
80
70
60
50
2016–01
2016–07
2017–01
2017–07
timestamp
position_label
10
5
0
−5
−10

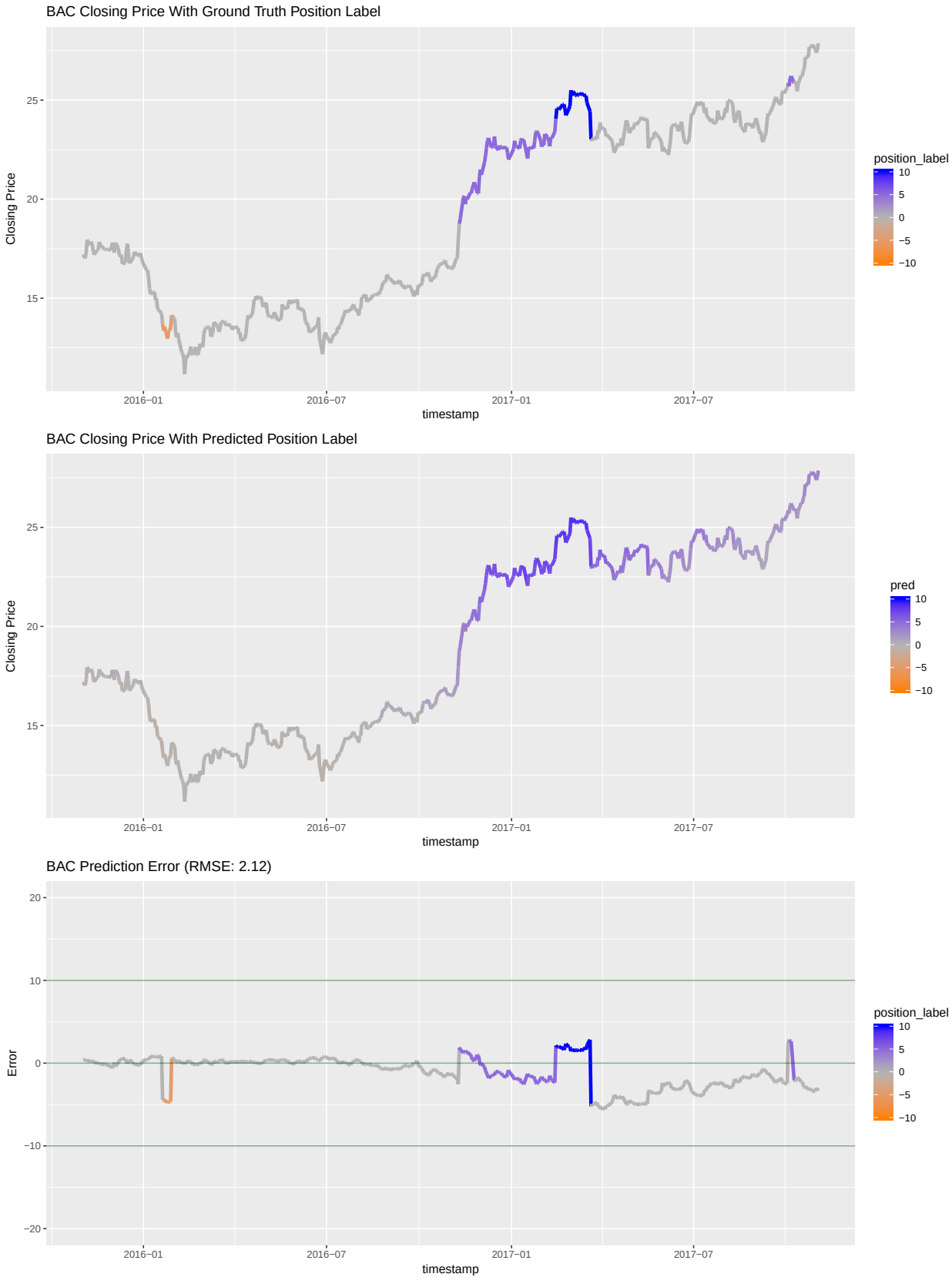

MSFT Closing Price With Predicted Position Label
Closing Price
80
70
60
50
2016–01
2016–07
2017–01
2017–07
timestamp
pred
10
5
0
−5
−10

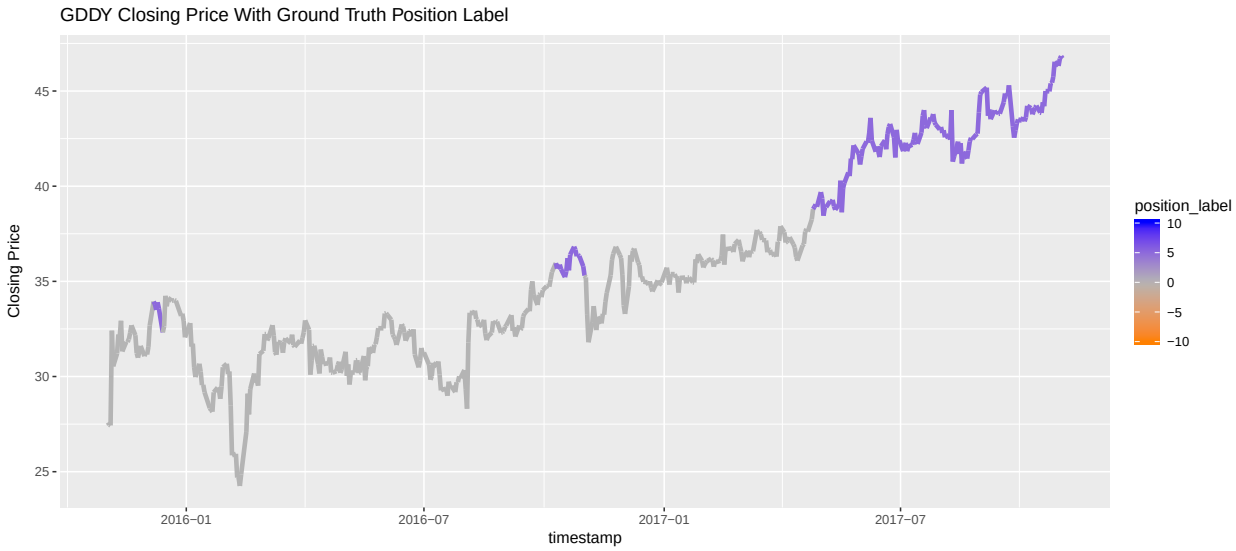

MSFT Prediction Error (RMSE: 2.37)
Error
20
10
0
−10
−20
2016–01
2016–07
2017–01
2017–07
timestamp
position_label
10
5
0
−5
−10

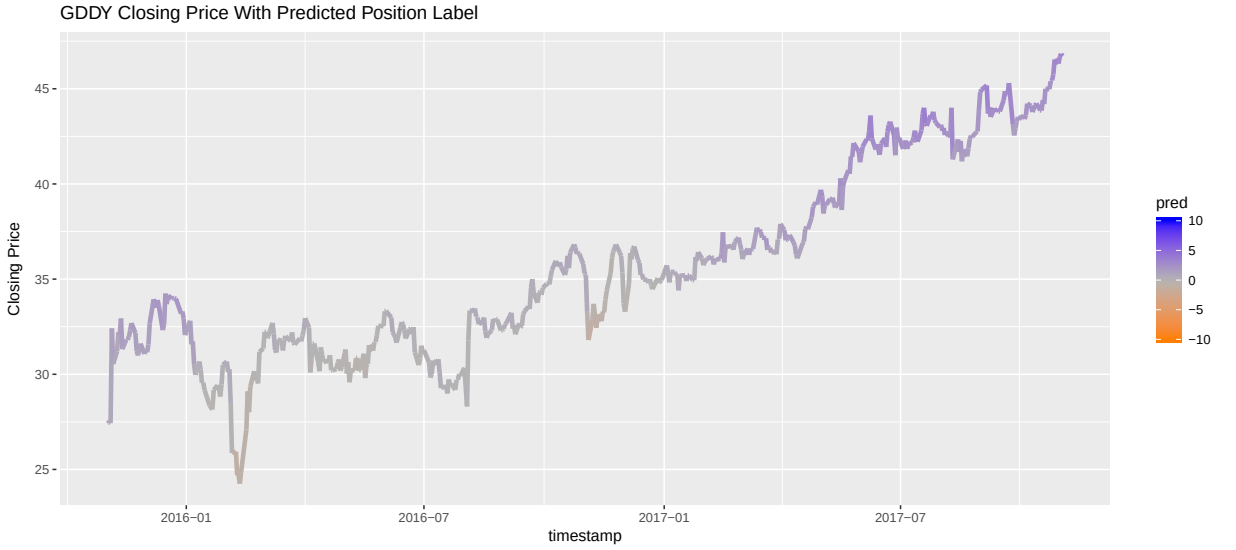

TSLA Closing Price With Ground Truth Position Label
Closing Price
350
300
250
200
150
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

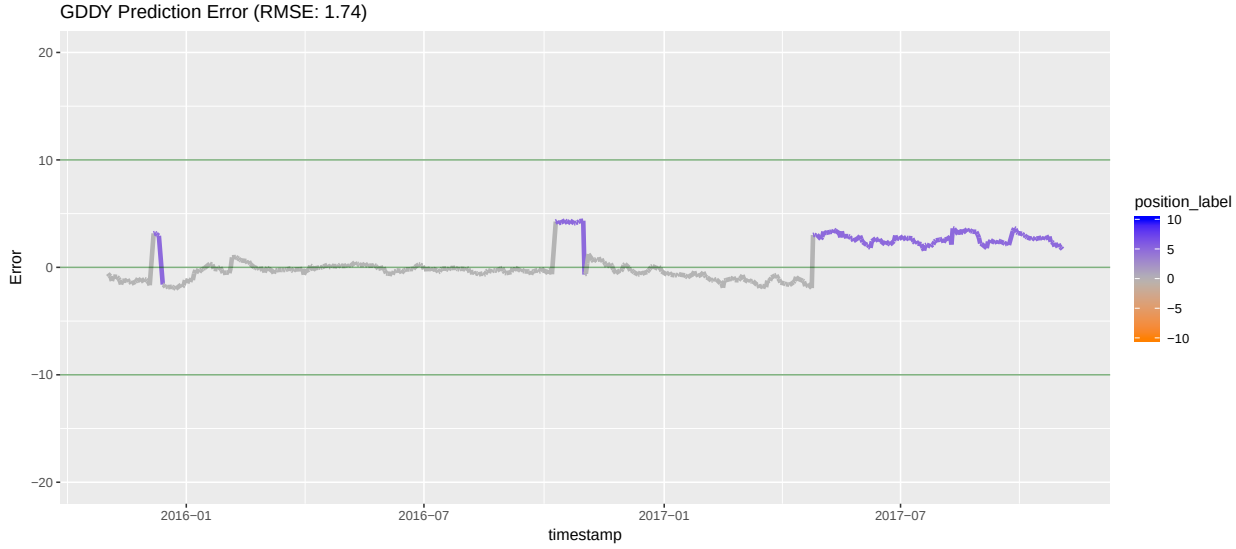

TSLA Closing Price With Predicted Position Label
Closing Price
350
300
250
200
150
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10

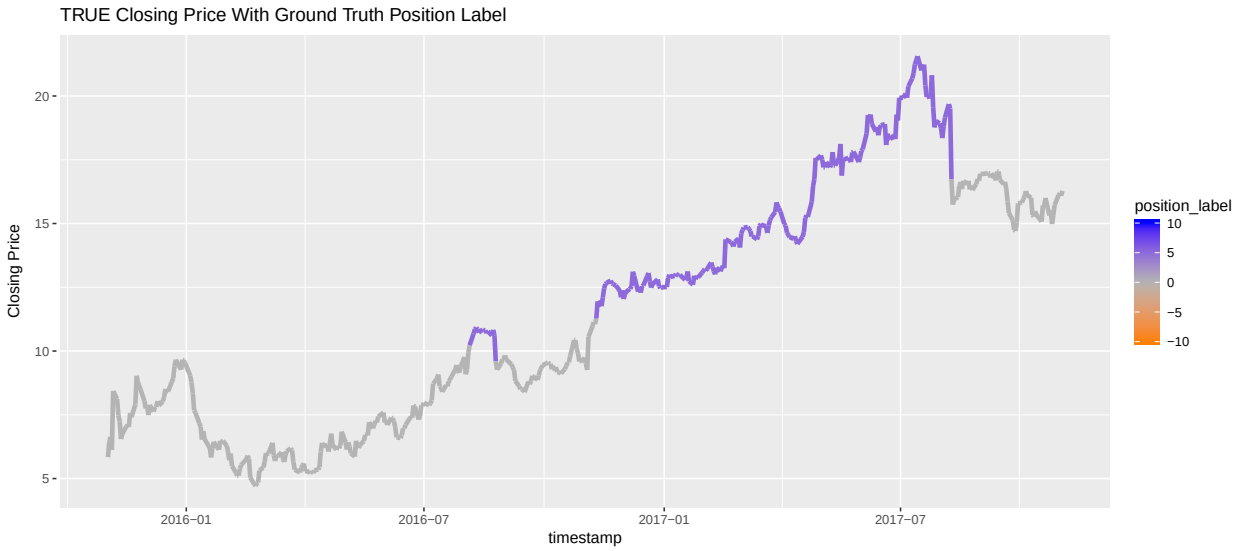

TSLA Prediction Error (RMSE: 2.16)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

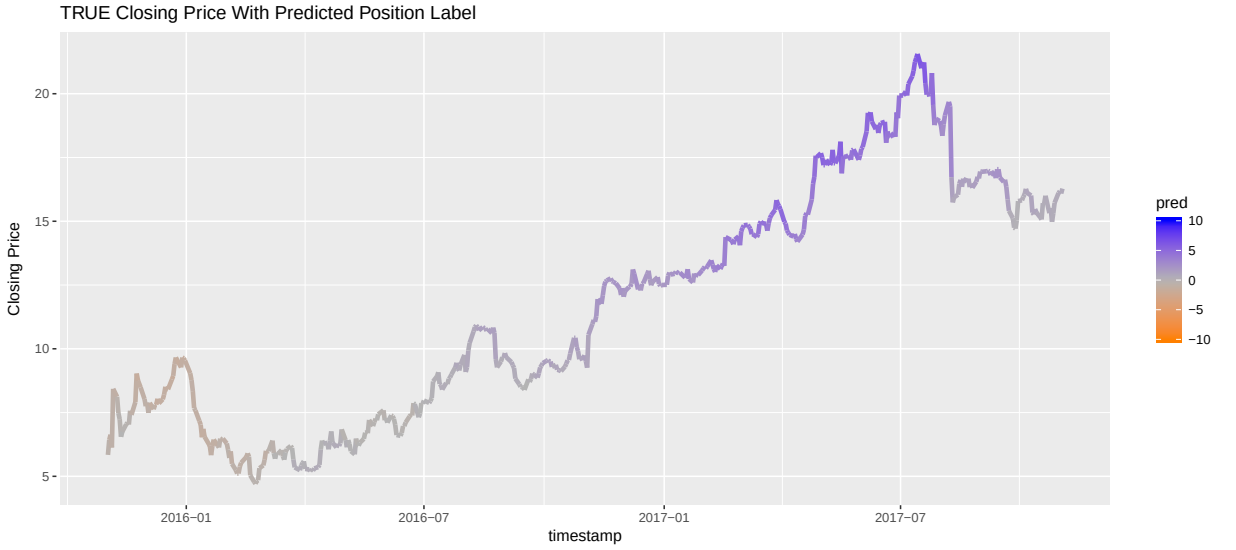

BAC Closing Price With Ground Truth Position Label
Closing Price
25
20
15
2016–01
2016–07
2017–01
2017–07
timestamp
position_label
10
5
0
−5
−10
BAC Closing Price With Predicted Position Label
Closing Price
25
20
15
2016–01
2016–07
2017–01
2017–07
timestamp
pred
10
5
0
−5
−10
BAC Prediction Error (RMSE: 2.12)
Error
20
10
0
−10
−20
2016–01
2016–07
2017–01
2017–07
timestamp
position_label
10
5
0
−5
−10

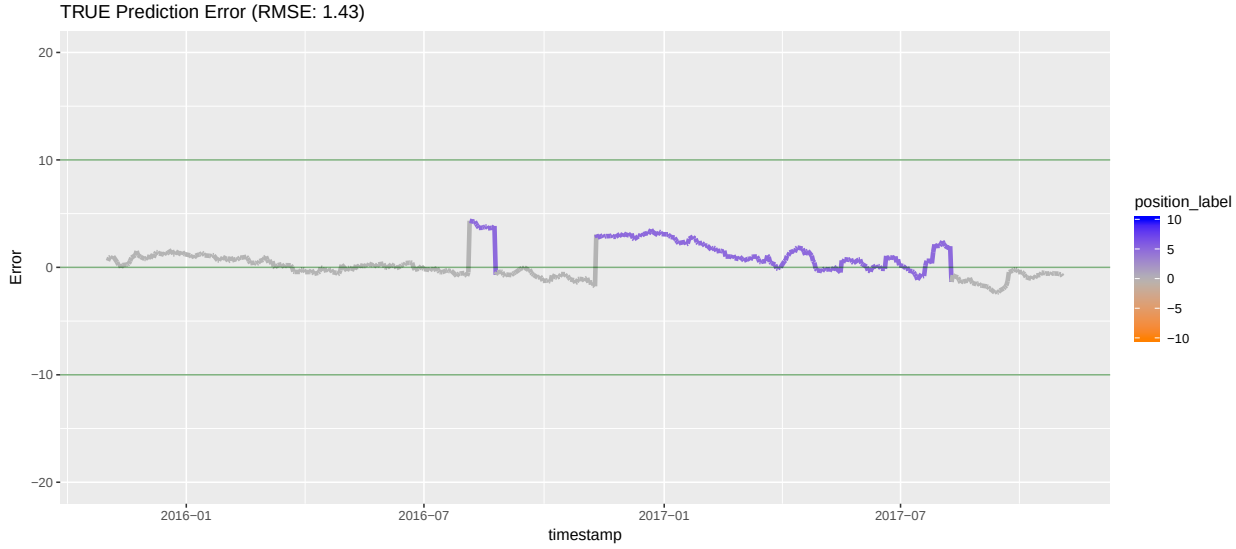

GDDY Closing Price With Ground Truth Position Label
Closing Price
45
40
35
30
25
2016–01
2016–07
2017–01
2017–07
timestamp
position_label
10
5
0
−5
−10


GDDY Closing Price With Predicted Position Label
Closing Price
45
40
35
30
25
2016–01
2016–07
2017–01
2017–07
timestamp
pred
10
5
0
−5
−10


GDDY Prediction Error (RMSE: 1.74)
Error
20
10
0
−10
−20
2016–01
2016–07
2017–01
2017–07
timestamp
position_label
10
5
0
−5
−10

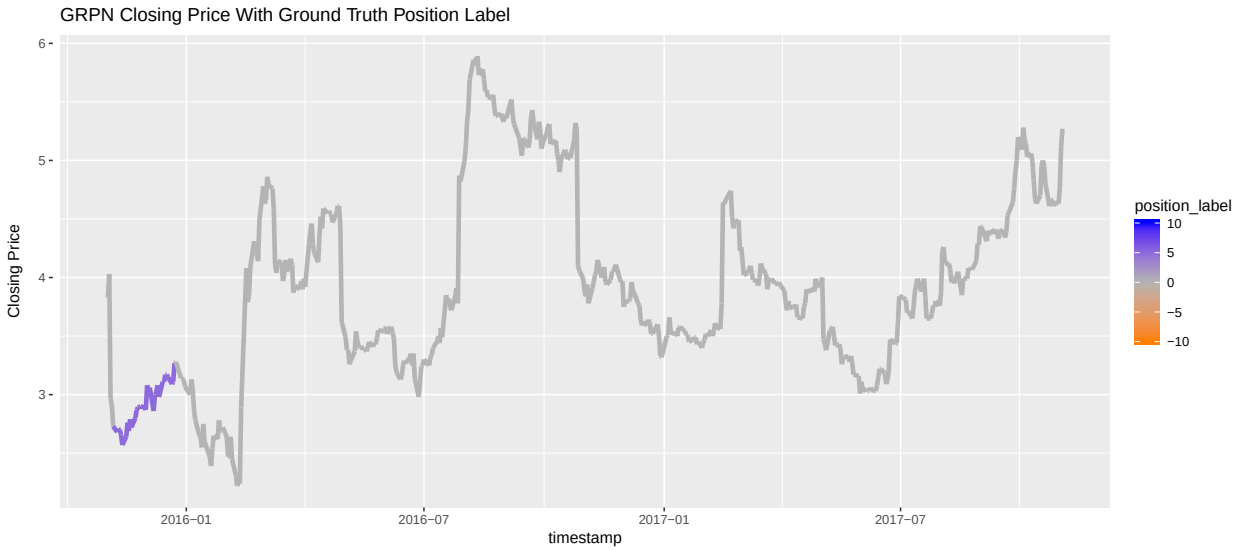

TRUE Closing Price With Ground Truth Position Label
Closing Price
20
15
10
5
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

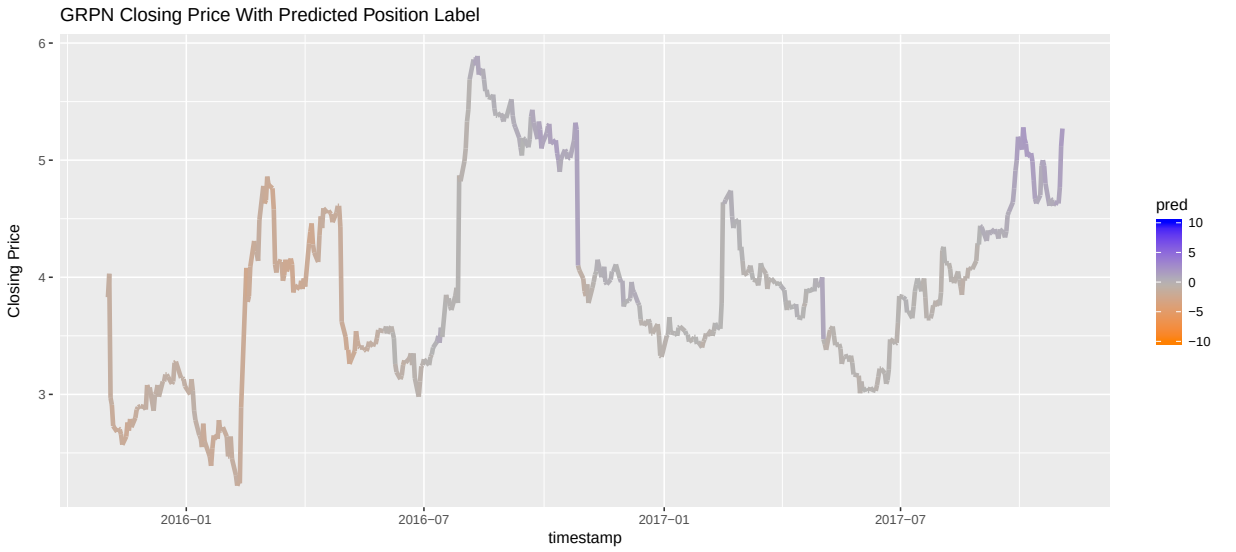

TRUE Closing Price With Predicted Position Label
Closing Price
20
15
10
5
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10

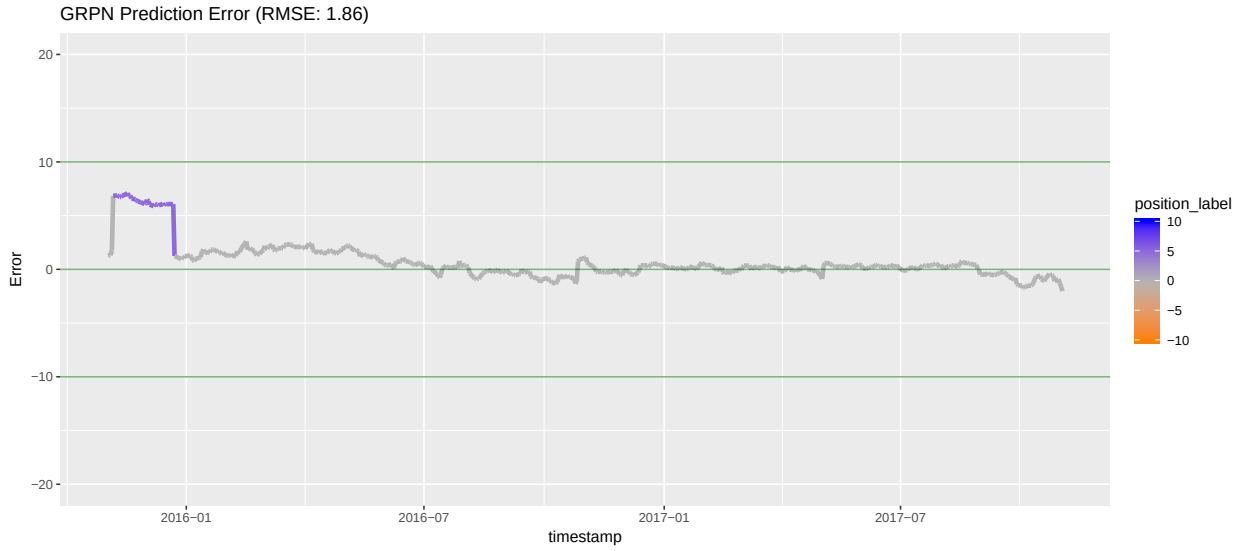

TRUE Prediction Error (RMSE: 1.43)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

5.3 No Trend

Model performance is excellent in this cluster. This result is expected because the majority of the observations in the training data have a ground truth position label of 0 which allows the model to learn very well in this region (see CVE.TO). The RMSE of the predictions in this cluster are the lowest.

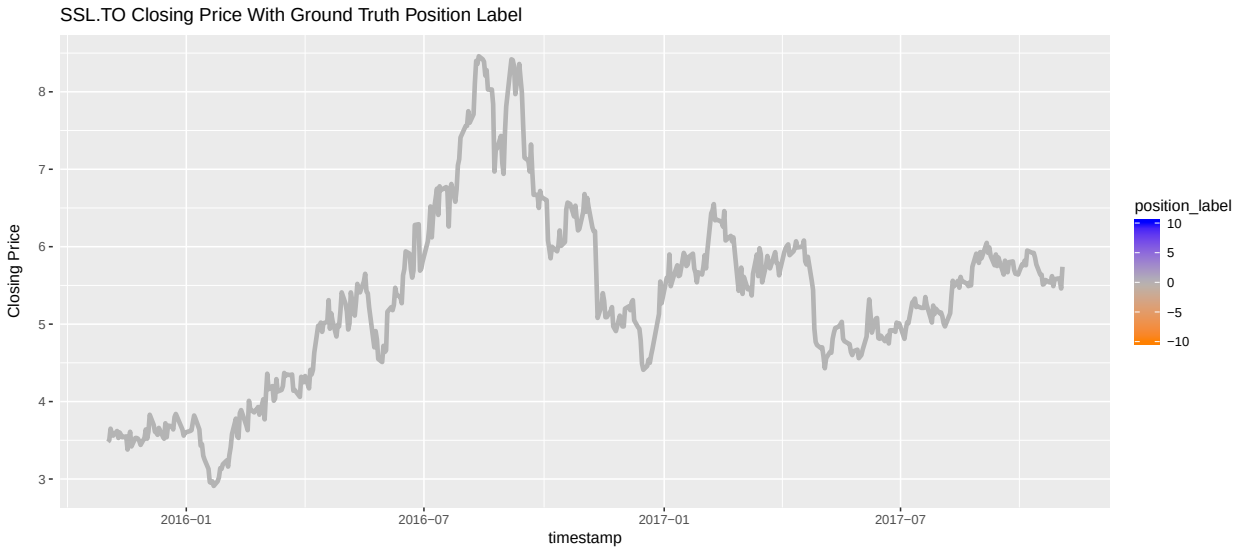

GRPN Closing Price With Ground Truth Position Label
Closing Price
6
5
4
3
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

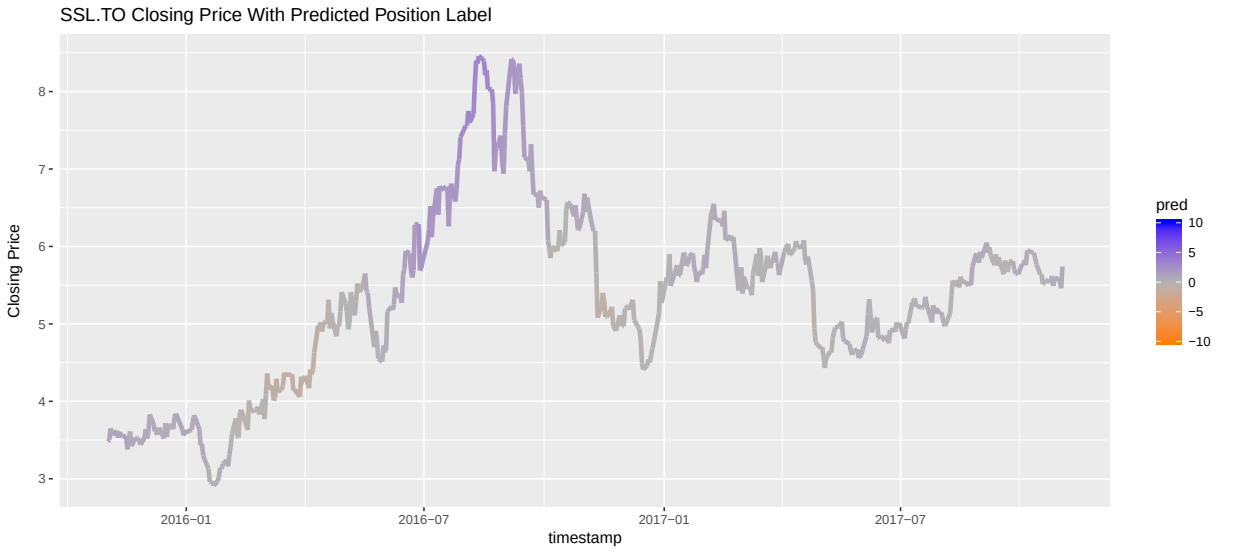

GRPN Closing Price With Predicted Position Label
Closing Price
6
5
4
3
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10

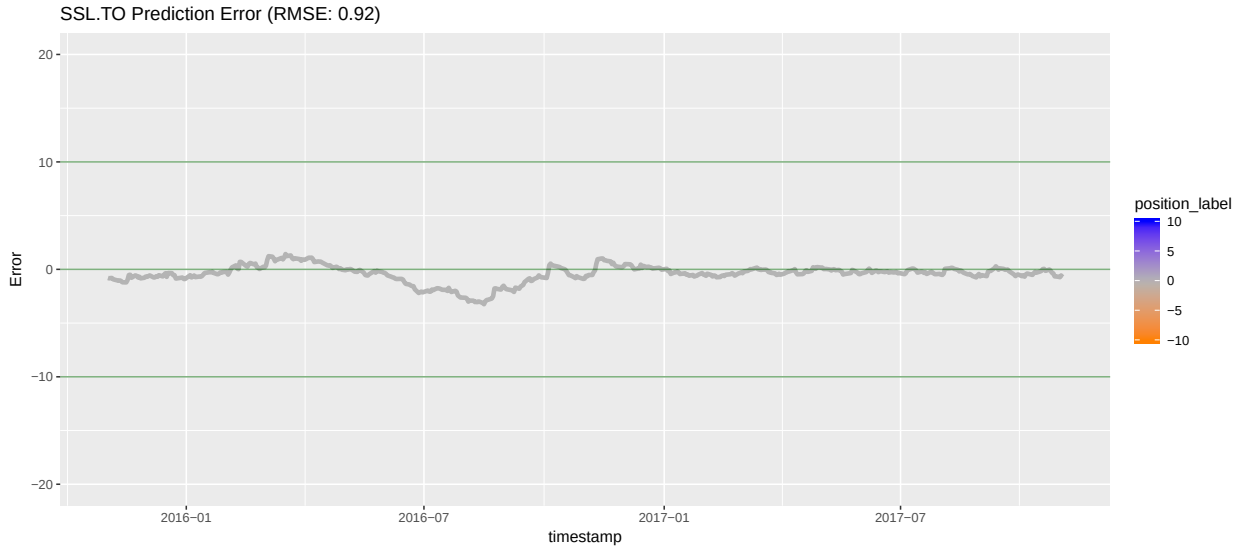

GRPN Prediction Error (RMSE: 1.86)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

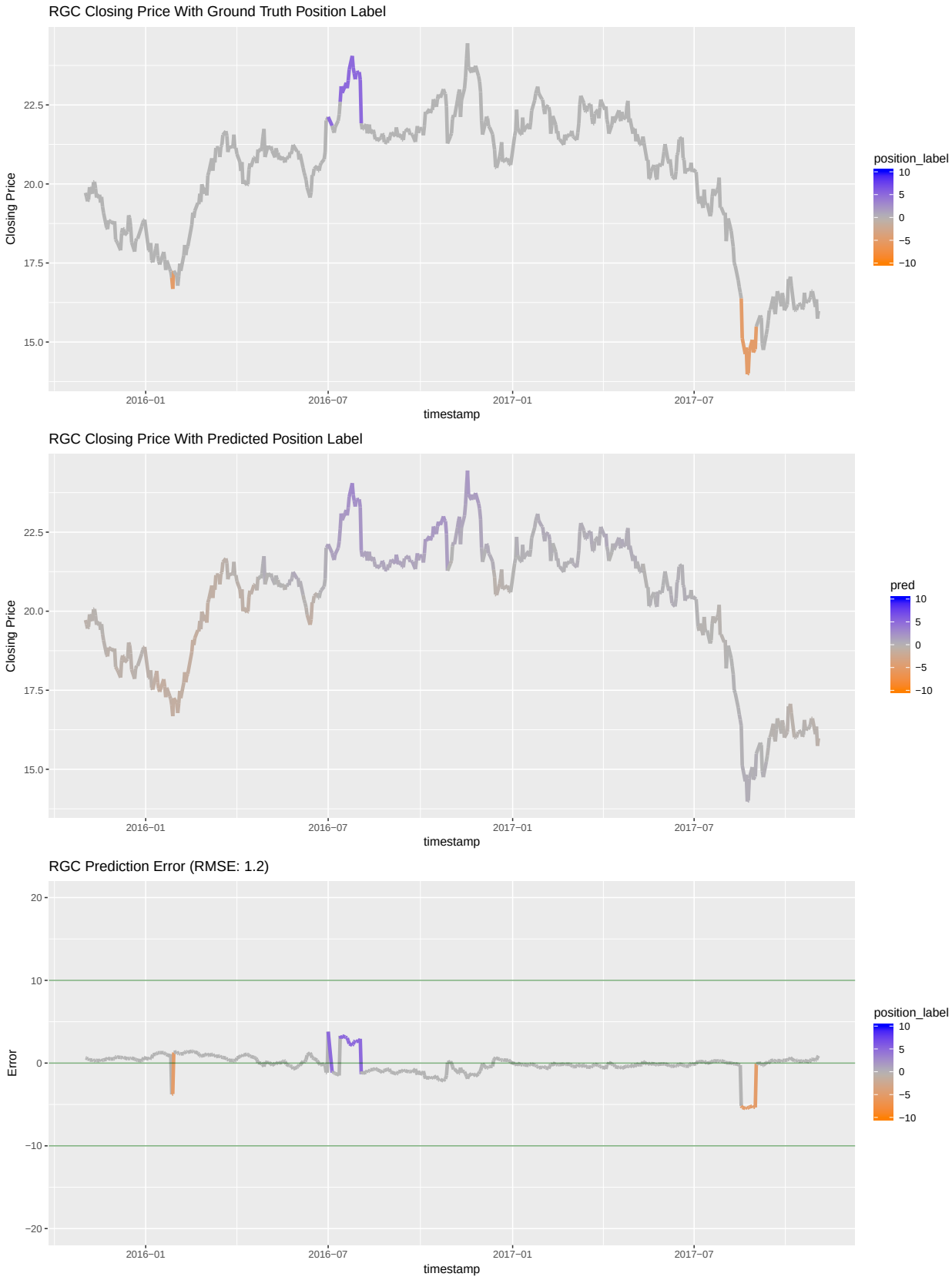

T Closing Price With Ground Truth Position Label
Closing Price
timestamp
position_label
T Closing Price With Predicted Position Label
Closing Price
timestamp
pred
T Prediction Error (RMSE: 1.48)
Error
timestamp
position_label
2016–01
2016–07
2017–01
2017–07

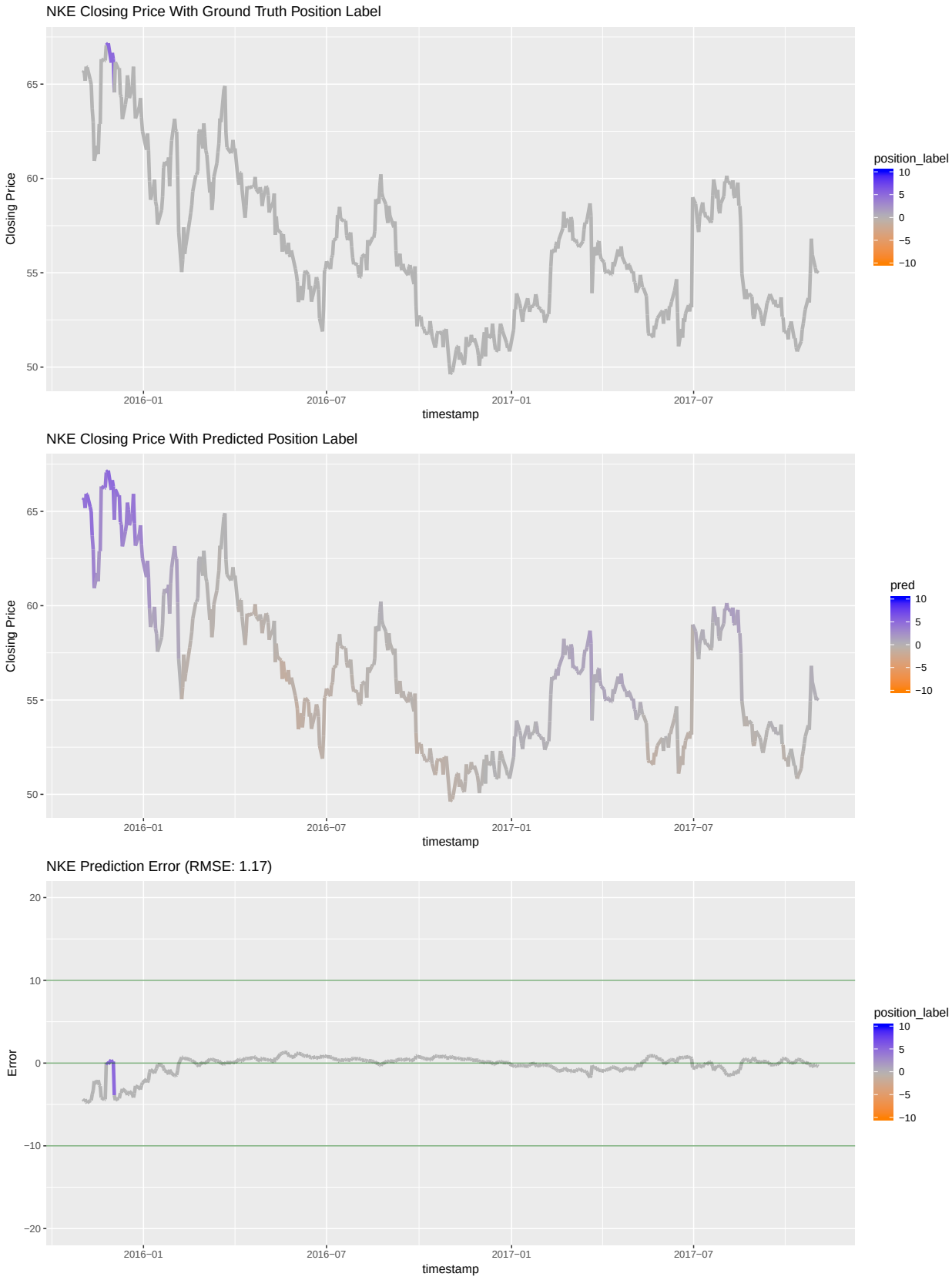

CVE.TO Closing Price With Ground Truth Position Label
Closing Price
20
16
12
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10
CVE.TO Closing Price With Predicted Position Label
Closing Price
20
16
12
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10
CVE.TO Prediction Error (RMSE: 0.32)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

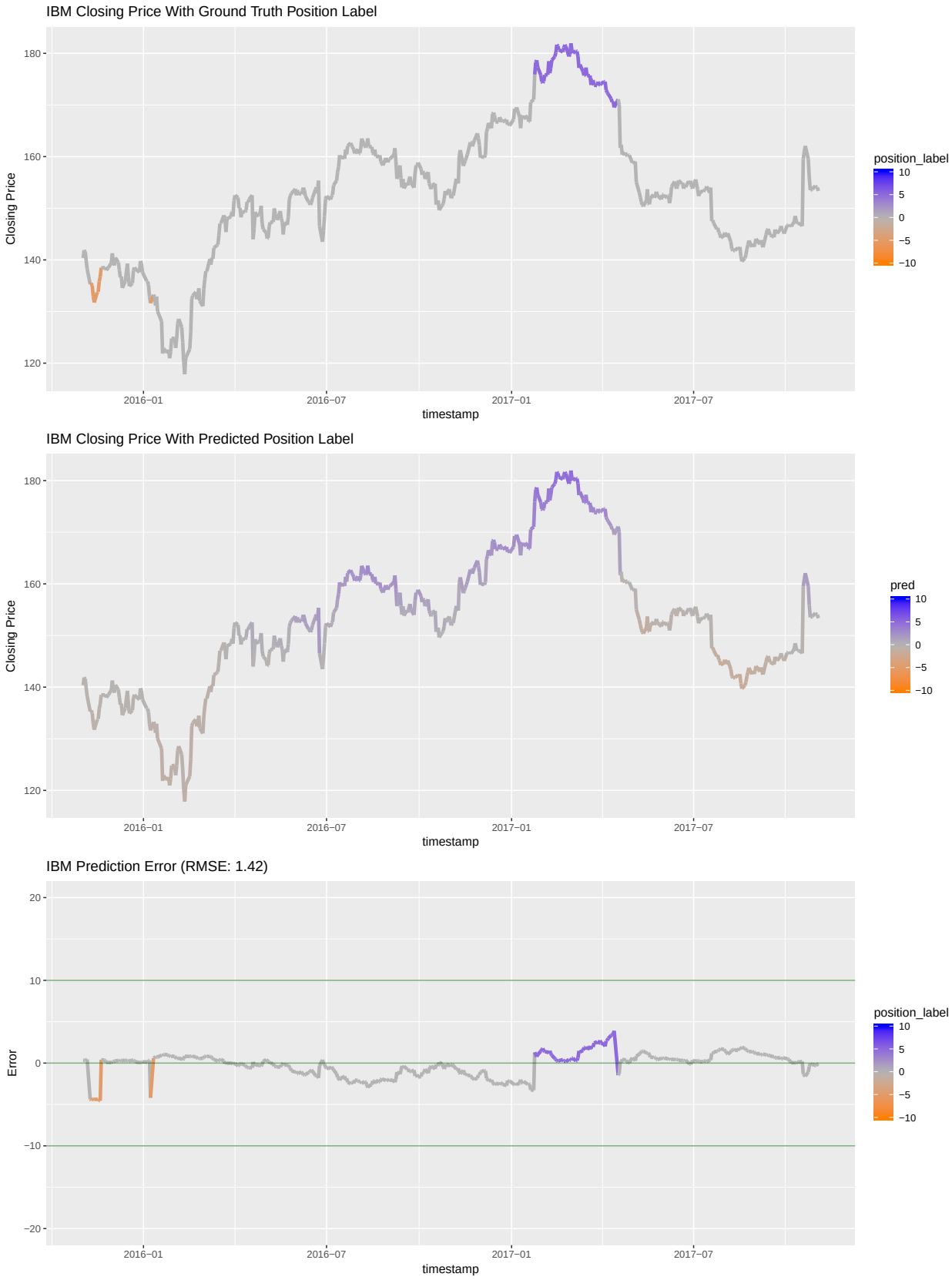

SSL.TO Closing Price With Ground Truth Position Label
Closing Price
8
7
6
5
4
3
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

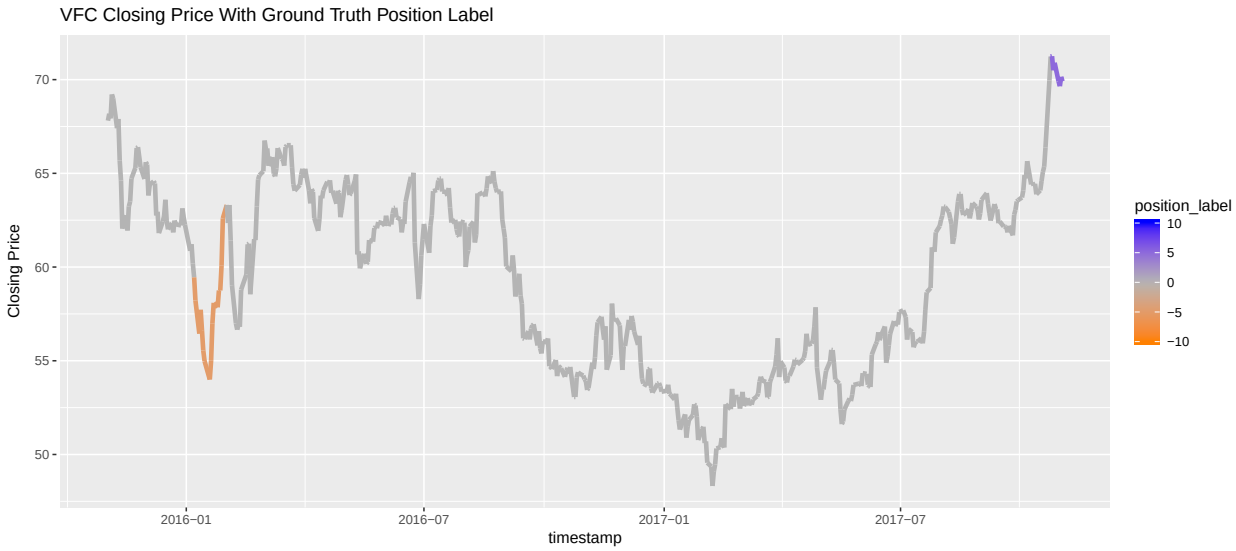

SSL.TO Closing Price With Predicted Position Label
Closing Price
8
7
6
5
4
3
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10

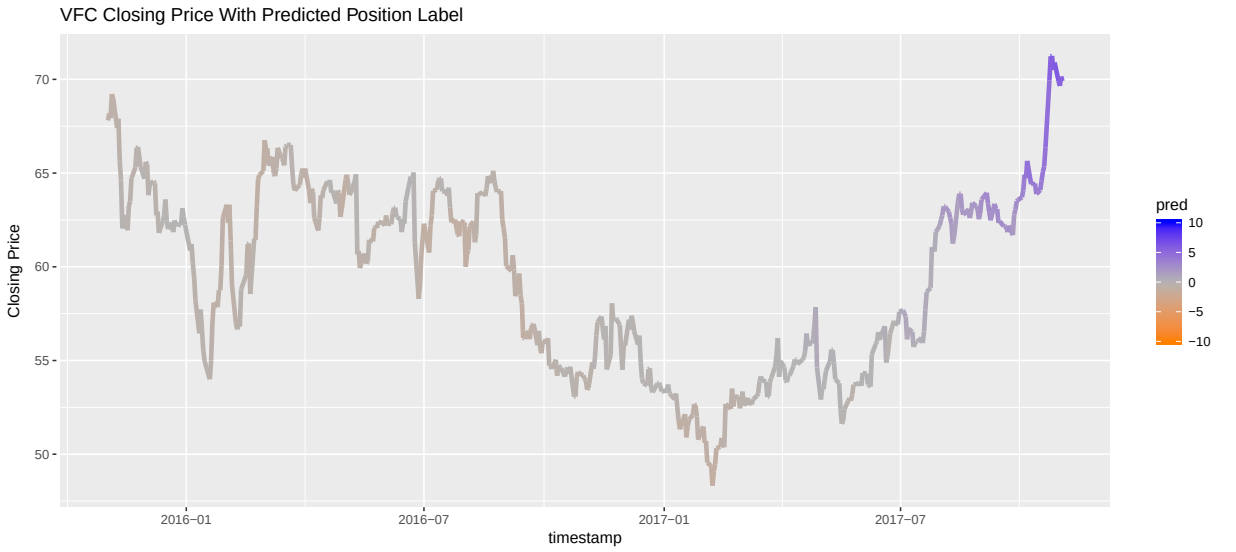

SSL.TO Prediction Error (RMSE: 0.92)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

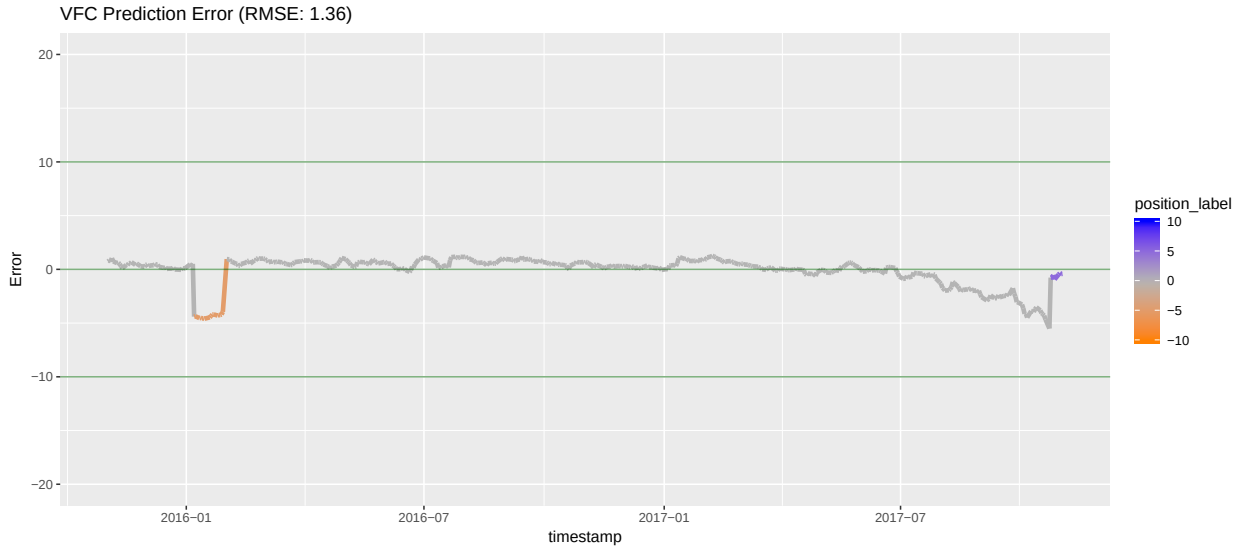

RGC Closing Price With Ground Truth Position Label
Closing Price
22.5
20.0
17.5
15.0
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10
RGC Closing Price With Predicted Position Label
Closing Price
22.5
20.0
17.5
15.0
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10
RGC Prediction Error (RMSE: 1.2)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

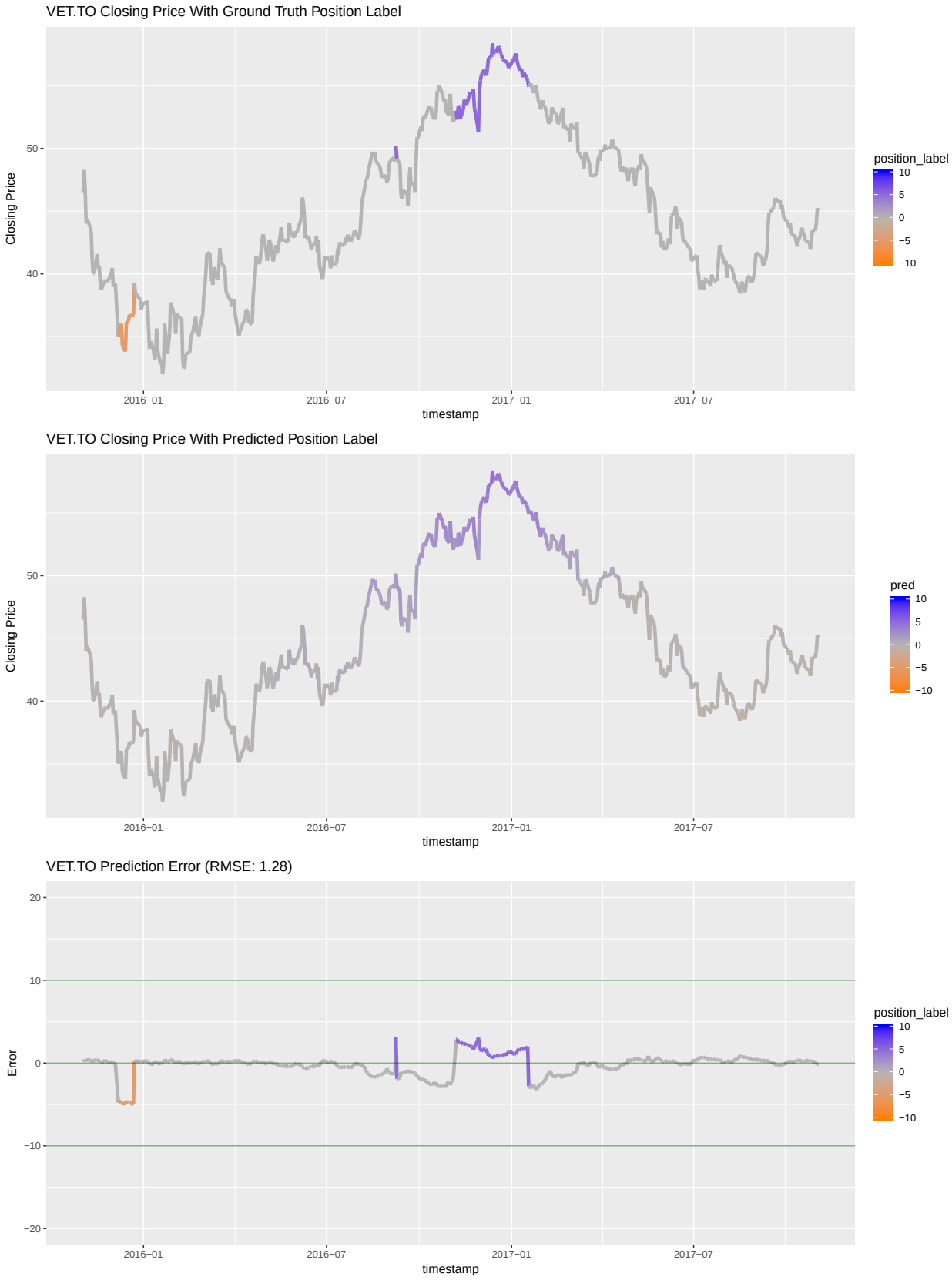

NKE Closing Price With Ground Truth Position Label
Closing Price
65
60
55
50
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10
NKE Closing Price With Predicted Position Label
Closing Price
65
60
55
50
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10
NKE Prediction Error (RMSE: 1.17)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10


IBM Closing Price With Ground Truth Position Label
Closing Price
180
160
140
120
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10
IBM Closing Price With Predicted Position Label
Closing Price
180
160
140
120
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10
IBM Prediction Error (RMSE: 1.42)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10


VFC Closing Price With Ground Truth Position Label
Closing Price
70
65
60
55
50
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

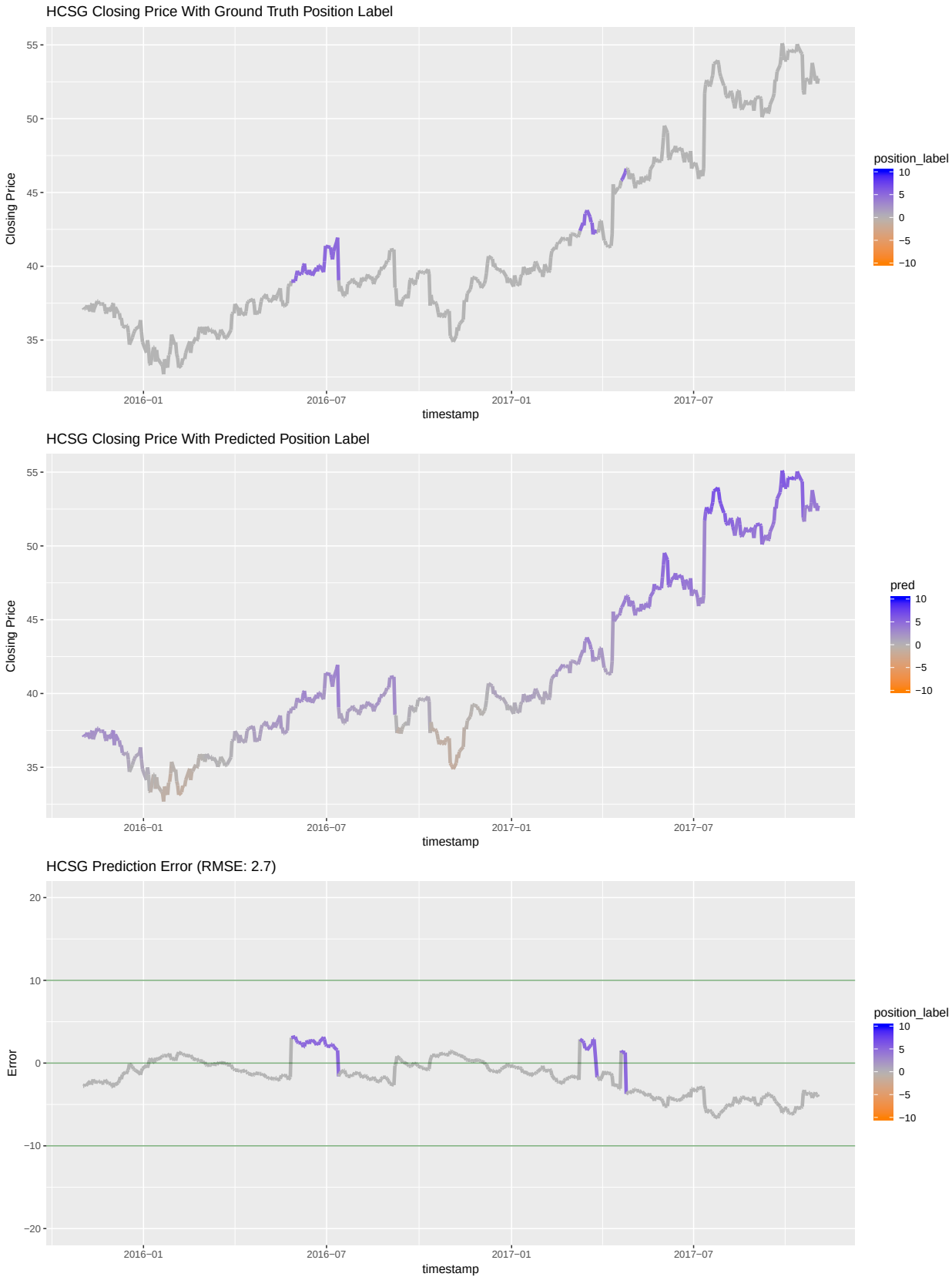

VFC Closing Price With Predicted Position Label
Closing Price
70
65
60
55
50
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10

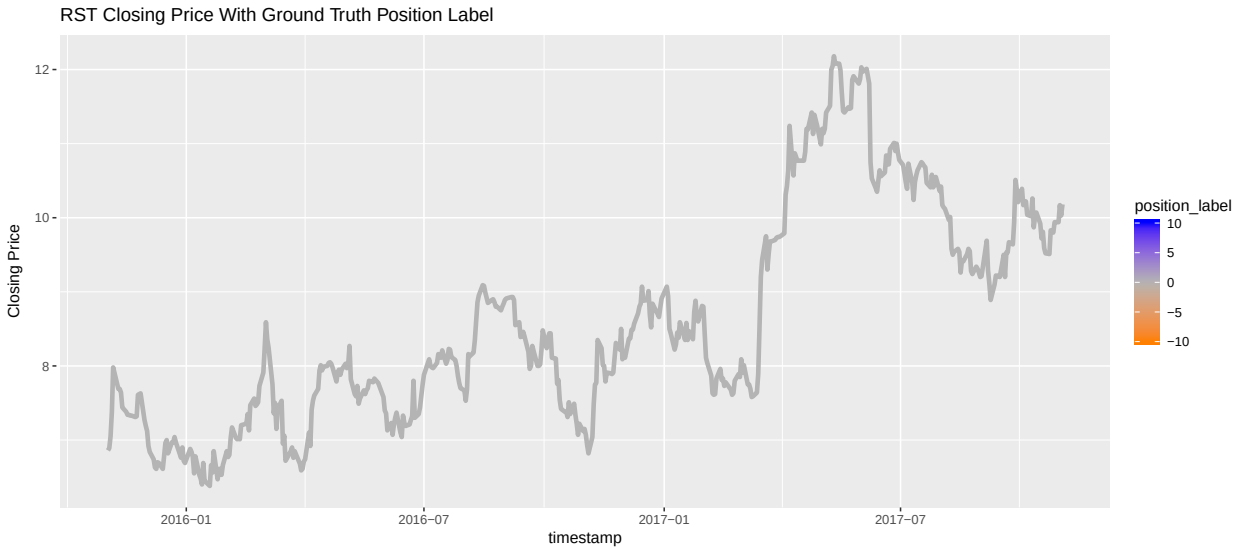

VFC Prediction Error (RMSE: 1.36)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

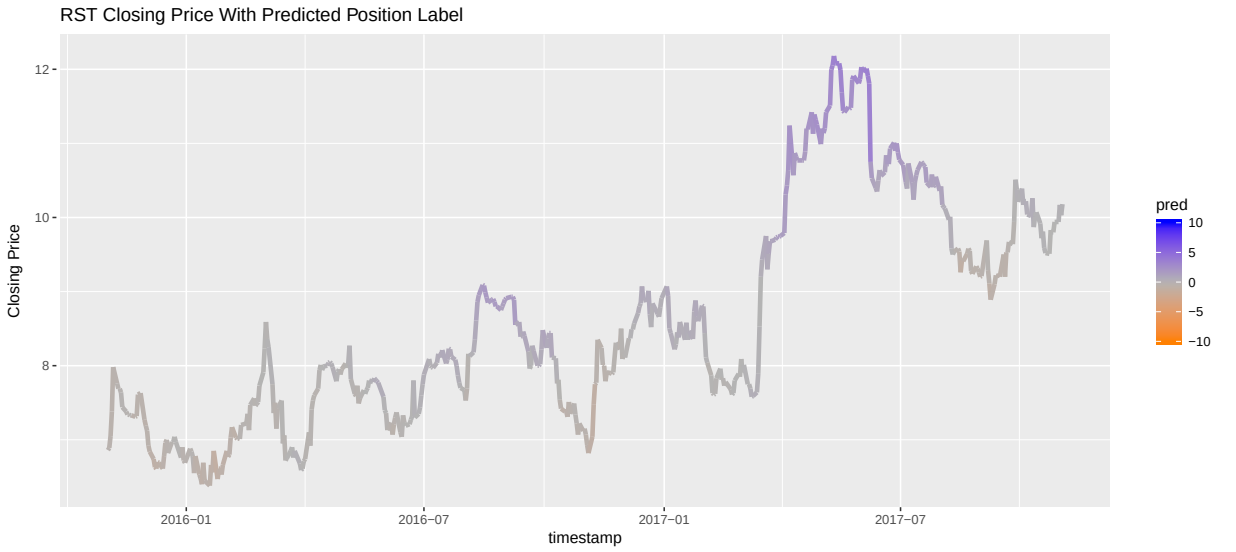

VET.TO Closing Price With Ground Truth Position Label
Closing Price
50
40
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10
VET.TO Closing Price With Predicted Position Label
Closing Price
50
40
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10
VET.TO Prediction Error (RMSE: 1.28)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

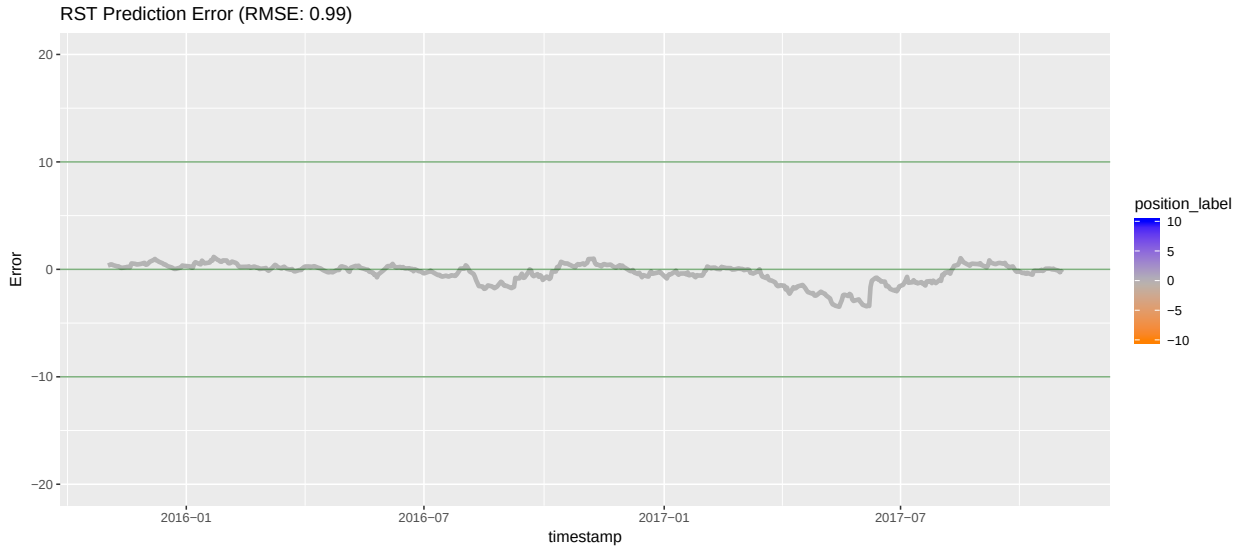
5.4 Too Volatile

This is a small cluster. Model performs poorly for HCSG by overpredicting the position label in the later of 2017. It is important to emphasize that the model predictions are out-of-sample. Therefore, if a specific stock represents a very unique pattern of stock prices that do not appear anywhere else in the training data, then it is very difficult for the model to know what position label should be associated with that pattern of stock prices. Since HCSG is a unique case (strong upward trend but too volatile), the model predictions perform poorly. This shortcoming is somewhat mitigated when the model is used in production because the model will be trained on all observations in the training data.


HCSG Closing Price With Ground Truth Position Label
Closing Price
55
50
45
40
35
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10
HCSG Closing Price With Predicted Position Label
Closing Price
55
50
45
40
35
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10
HCSG Prediction Error (RMSE: 2.7)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

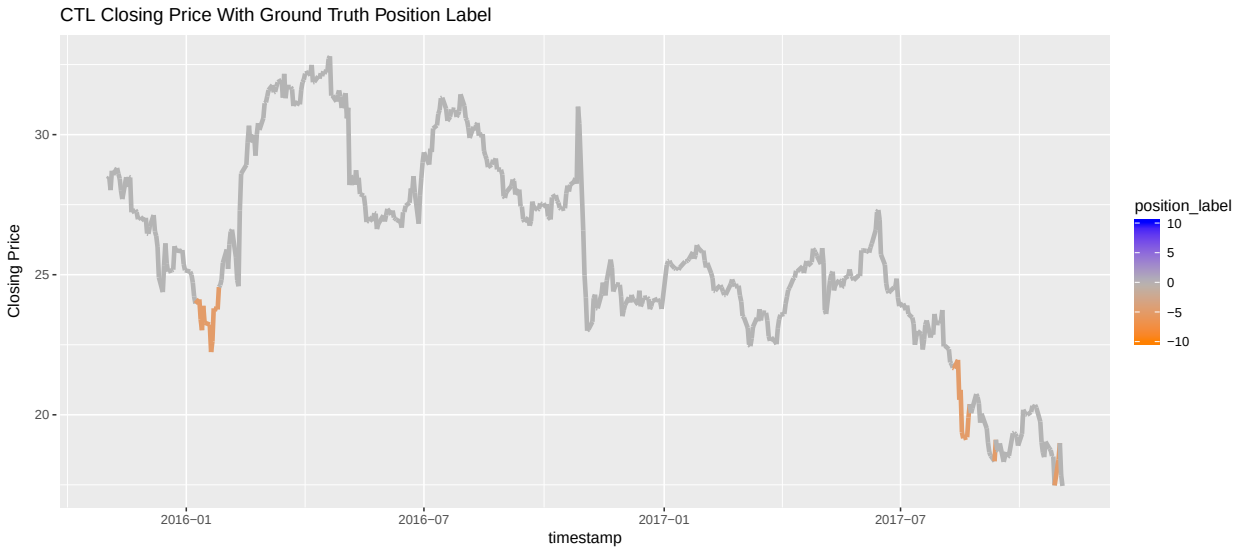

RST Closing Price With Ground Truth Position Label
Closing Price
12
10
8
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

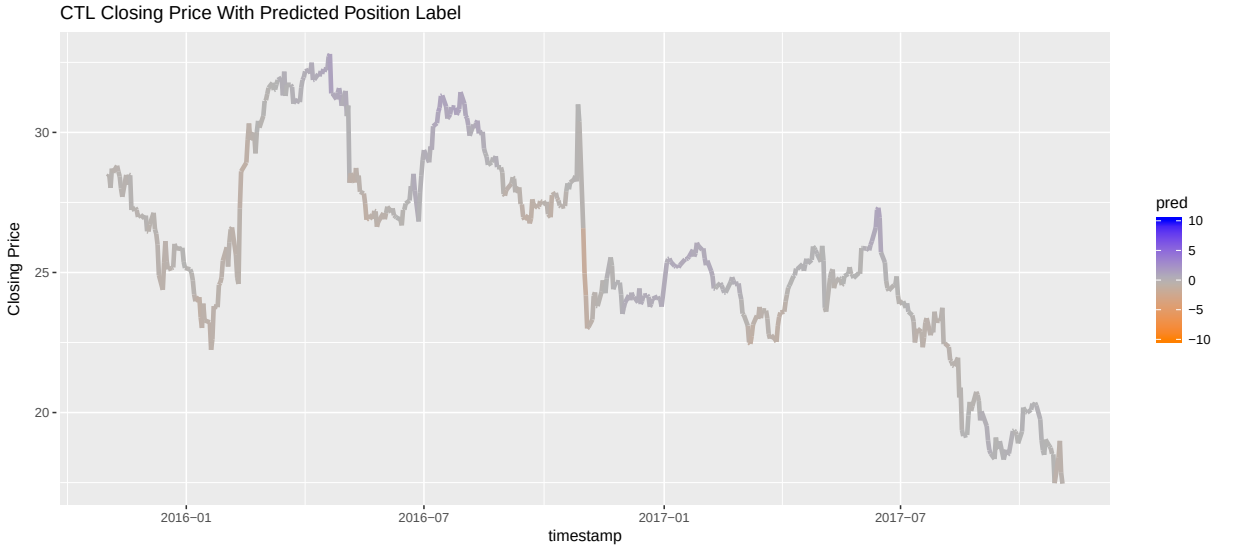

RST Closing Price With Predicted Position Label
Closing Price
12
10
8
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10

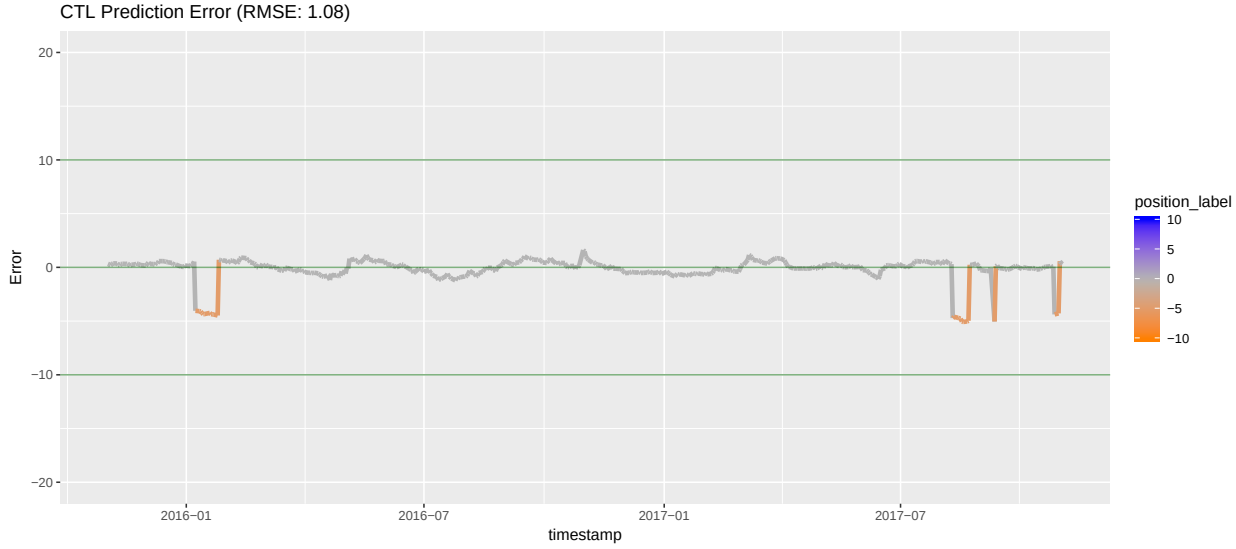

RST Prediction Error (RMSE: 0.99)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10

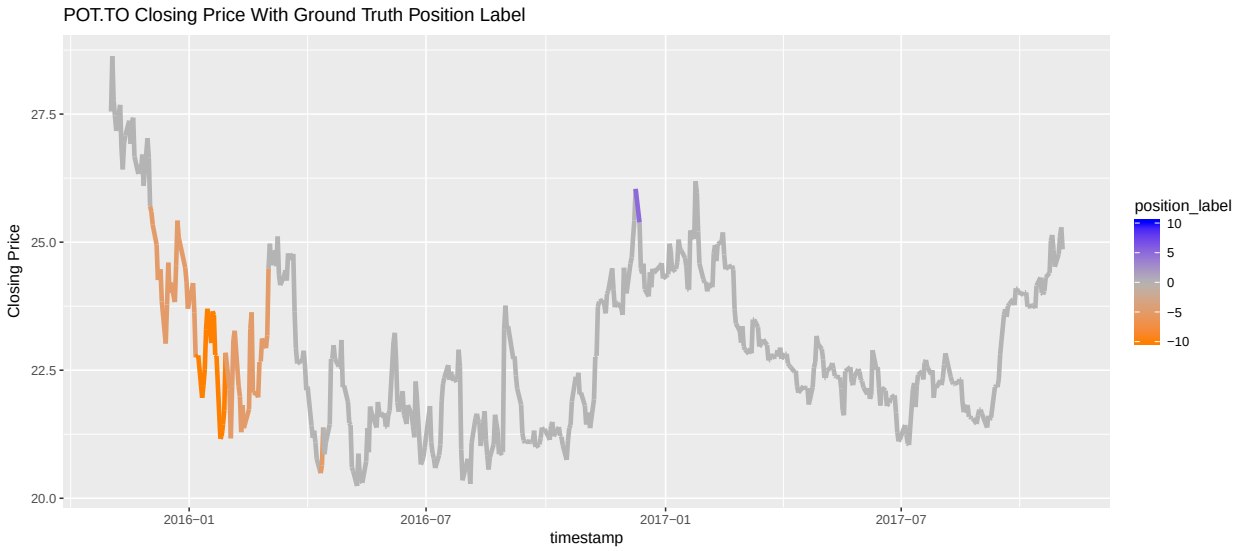
5.5 Downward Trend

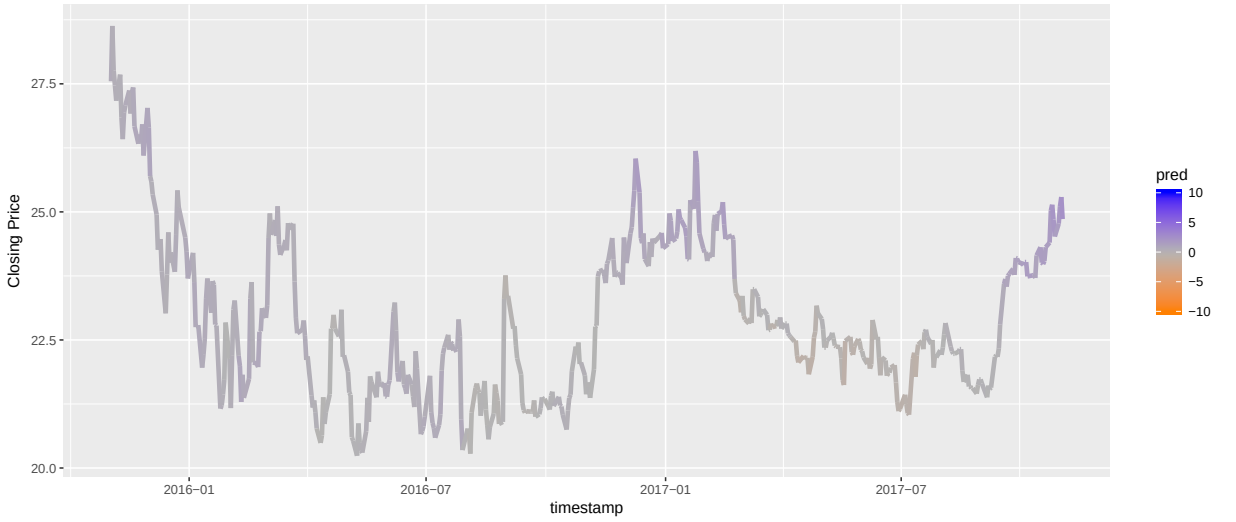
Similar to the previous clusters, the model has extremely high accuracy in predicting a 0 position label. However, the model is seemingly unable to predict very negative position labels, so the model predictions are quite bad when the ground truth position label is negative. The most likely explanation is that there are insufficient training observations with a negative position label for the model to learn from.

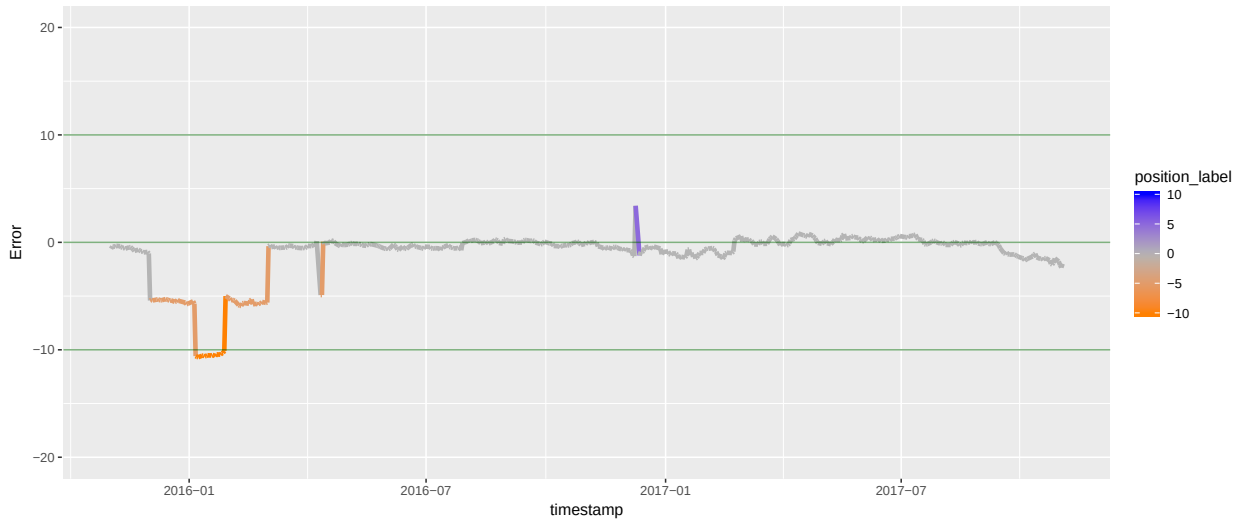

CTL Closing Price With Ground Truth Position Label
Closing Price
30
25
20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10


CTL Closing Price With Predicted Position Label
Closing Price
30
25
20
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10


CTL Prediction Error (RMSE: 1.08)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10


POT.TO Closing Price With Ground Truth Position Label
Closing Price
27.5
25.0
22.5
20.0
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10


POT.TO Closing Price With Predicted Position Label
Closing Price
27.5
25.0
22.5
20.0
2016−01
2016−07
2017−01
2017−07
timestamp
pred
10
5
0
−5
−10


POT.TO Prediction Error (RMSE: 2.62)
Error
20
10
0
−10
−20
2016−01
2016−07
2017−01
2017−07
timestamp
position_label
10
5
0
−5
−10