Forecasting and Trading the Market Open of Highly Volatile U.S. Listed Stocks

Nathan Nechemia Renzoni

MscFE WorldQuant University

nrenzoni1@gmail.com

Abstract

The paper begins by providing motivation for research of intraday directional trading on highly volatile U.S. listed stocks. The research is limited to a subset of stocks typically traded by "day-traders", stocks exhibiting the highest percentage price increase in the overnight trading session. Functions of fundamental company data, as well as price transaction history, are defined. Exploratory data analysis of these variables is performed. These variables are then used as input to a machine-learning model to predict the intraday price percentage change in the regular trading session. Throughout the paper, summary statistics, plots, and tables are provided.

Keywords: *day-trading, intraday stock return forecasting, high volatility stocks, exploratory data analysis of stocks, machine learning*

1. Introduction

Many "day-traders" try to time their entry and exit in highly volatile stocks, with the expectation that they can profit handsomely off the large price swings. In this project, "day-trades" are assumed to be closed on the same day before the end of the regular trading session. Many of these traders trade by instinct, without any or much quantitative analysis on volatile stocks. This project studies a subset of highly volatile stocks, specifically those stocks that have experienced the highest ranked percentage price increase in the prior overnight-trading session. In the regular trading session, many stocks from this universe will tend to have either a large "corrective" price move, i.e., a decline in price, or a further price increase. This project uses exploratory data analysis, multiple predictor variables, and machine learning techniques to

forecast price changes. The motivation for this project is to discover if it is possible to trade these volatile stocks profitably, or if the endeavor is better yet avoided.

2. Literature Review

Gerety, M. et al. (1992) show that expected price volatility in the opening hour of all NYSE stocks between 1933-1988 is influenced by the previous overnight price volatility [4]. These results are confirmed by Chen, C. H. et al. (2012) who show that volatility and price data from after-hours trading improves the conditional volatility forecasts in the next-day regular trading hours over the thirty most actively traded NASDAQ stocks [2]. Since the studies are confined to studying volatility, no analysis is performed on price direction modeling. Furthermore, these studies analyze the same universe of stocks, and no explicit analysis is performed on any dynamic subset of different stocks which are dependent on factors such as the highest overnight price volatility. This project focuses on the subset of U.S listed stocks with the highest overnight upside price volatility and see if this stock universe exhibits any characteristics which can be exploited for profitable trading returns.

3. Methodology

3.1 Data Acquisition

The data set encompasses the period of April 2021 until the end of September 2021. Each day *n* where the NYSE, NASDAQ, and AMEX exchanges are open, every listed stock *S* has an open price $P_{t_{n,1}}$ recorded at $t_{n,1}$ = 9:25 EST, and a price recorded on the close of the prior day *n-1* of $P_{t_{n-1,2}}$, where $t_{n-1,2}$ = 16:00 EST. On each day *n* at time $t_{n,1}$, all stocks are ranked by the highest percent price increase during the overnight trading session, i.e., $\frac{P_{t_{n,1}}}{P_{t_{n-1,2}}} - 1$ for every stock. The 100 stocks with the highest score are sorted in decreasing order, $S_{(1)}$ being the stock with the highest score, and $\{S_{(i)} : 1 \leq i \leq 100\}$ is recorded for each day where trading on the exchanges takes place. The daily list also includes warrants and other non-standard listed instruments. After removal of these securities from the list, approximately 60 stocks remain for each day *n*.

For the stocks in the study on relevant days, the price and size of every reported transaction (i.e., tick data) is retrieved. Unfiltered consolidated tape data is used and is provided by DTN IQ-Feed.

For each stock $S_{(i)}$, company fundamental data is recorded prior to the open of the regular trading session at 9:25. Since the predicted variables (defined in the next section) and trading rules occur after $t_{n,1}$, this data will not result in look-ahead bias in the analysis.

3.2 Predictor Variable Overview

We have a list of 28 different variables which are used to predict stock price changes between 9:30 until an end time between 9:35 and 10:00. Some of the variables are well known in the finance world and do not require explanation. The variables of fundamental market data are recorded at 9:25 on the day of interest for each stock. This includes the stock's price, the respective company's market capitalization, the percent of the total public shares (i.e., the float) which is reported as being shorted, the number of the company's shares outstanding, the value of the cash assets of the company, the last dividend payment amount, and others. Some

variables are created by taking the ratio of two variables, such as last_premarket_price_ratio_to_premarket_high, which divides the highest trade price sampled between 16:30 of the previous market day and 9:30 of the current day by the last recorded traded price before 9:30. For predictor variables calculated on the after-hours price series, outliers of the price series are removed. The outlier removal procedure computes a rolling 30-second mean and standard deviation and removes price values which are outside the range of 2.5 times the rolling standard deviation +/- the rolling mean. The outlier removal procedure is conducted since after-hours reported trades may be carried over trades from the previous regular-hours trade session, and the data may have an incorrect time value. The predictor variable names are listed below:

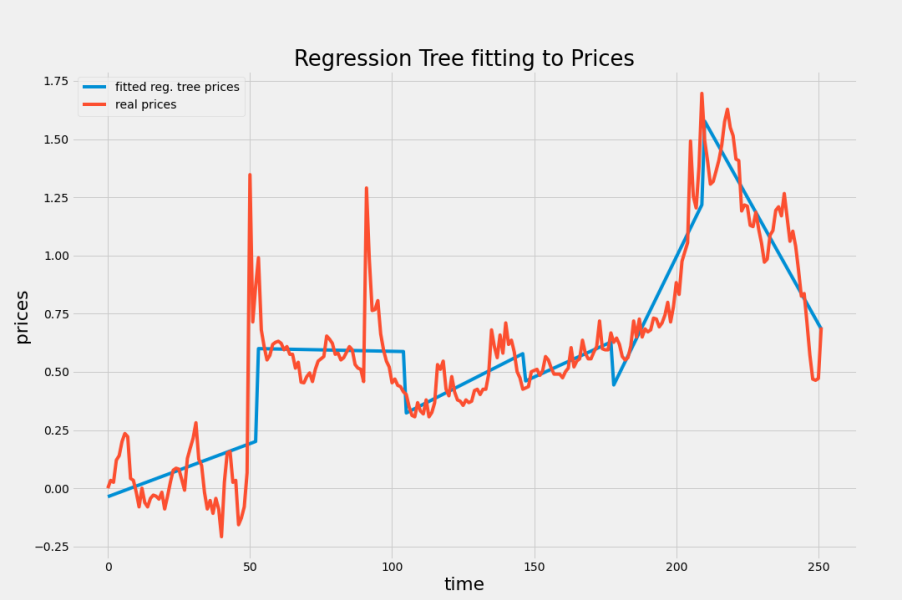
	predictor_var_names
0	stock_price_925
1	market_cap
2	short_float_pct
3	float
4	shares_outstanding
5	avg_daily_volume_10D
6	cash_$
7	cash_to_debt_ratio
8	revenue_$_per_year
9	dividend_$
10	interest_income_$
11	current_assets_$
12	enterprise_value_$
13	first_above_80_pct_after_hours_slot
14	num_fitted_reg_models
15	num_overnight_trades
16	overnight_dollar_volume
17	max_min_price_pct_change
18	momentum_adj_by_r2
19	max_after_hours_price_velocity
20	last_premarket_price_ratio_to_premarket_high
21	last_premarket_price_ratio_to_10_day_high
22	last_premarket_price_ratio_to_10_day_low
23	10_day_avg_volume
24	10_day_price_std_dev
25	overnight_volume_to_10_day_avg_volume
26	year_daily_closes_reg_slope
27	year_daily_closes_reg_fit

Some non-trivial variables are described below:

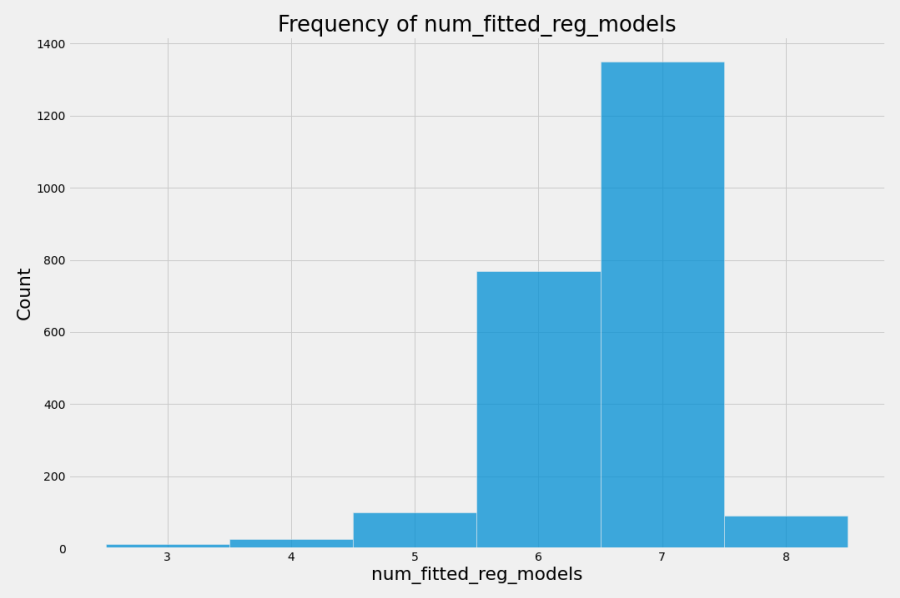
(13) first_above_80_pct_after_hours_slot - this variable describes the time when the trading price of the given stock reaches 80-percent of the price between the last traded price after 16:00

of the previous market day and the highest trade price between 16:00 the previous day and 9:30 the current date.

(14) num_fitted_reg_models - fits an optimal number of linear regression models to the after-hours trading prices, and then counts the number of fitted models. Calculations are performed using the linear-tree python library [6]. Below, we show an example of the fit of the regression tree model to the trade prices.



As can be seen, the fitting of the regression models does not capture extreme price moves. Below, we plot the frequency distribution of the number of fitted regression models per stock (the predictor variable num_fitted_reg_models). The number of fitted trees ranges from a low of 3 to a high of 8. This narrow range is due to the default behavior of the linear-tree python library. A potential improvement for this variable would be to widen the spread of outcome values.



(17) max_min_price_pct_change - the percent change between the maximum after-hours price and the minimum after-hours price.

(18) momentum_adj_by_r2 - To calculate this variable, first an ordinary least squares (OLS) regression model is fitted to the after-hours prices (outliers are removed). The final value is calculated as the slope of the regression multiplied by the r^2 value of the regression (the coefficient of determination). Multiplying by r^2 standardizes large price moves if the move had high volatility [3].

(19) max_after_hours_price_velocity - the maximum of the difference of 1-second interpolated price values. This may serve as a continuous proxy indicative of the volatility of the after-hours trading.

(26) year_daily_closes_reg_slope - an OLS regression model is fitted to the previous 252 days of the stock close prices. The regression slope is recorded.

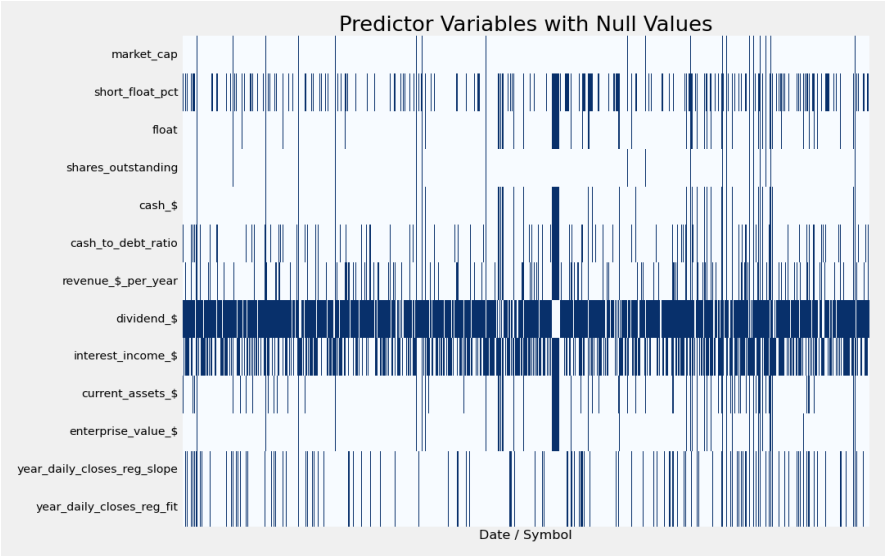
The dates in this study span 125 days, starting from 2021-04-06 until 2021-09-30. The number of unique combinations of dates and stocks in this study is 2,741. For realistic test results, the data set is split into a training set followed sequentially by a test set. The first date of the test set is 2021-08-09.

	dates
total:	125
train:	86
test:	39

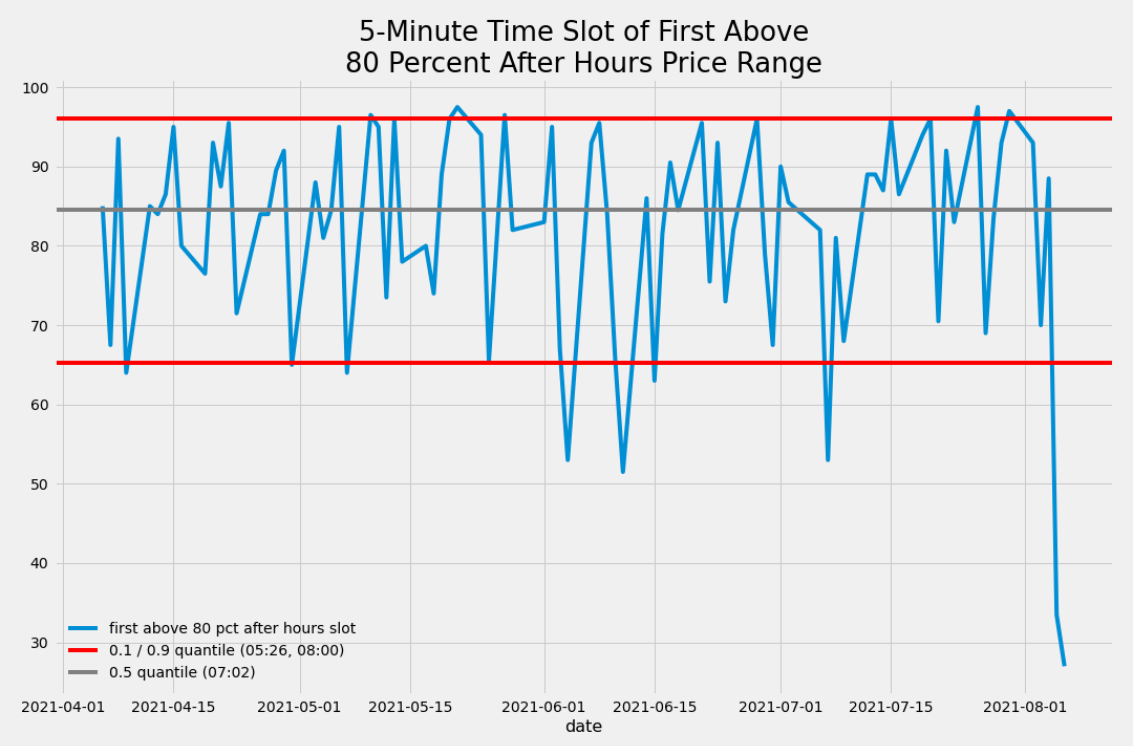
In the following sections, unless noted otherwise, the statistics, plots and tables describe and use only data from the training data set. The test data set is set aside for scoring the predictive models and for and for out-of-sample tests of trading strategies.

Some predictor variables have a high ratio of missing values. The ratios are shown in the below table. The subsequent figure plots the missing values in blue, where the x-axis plots every date/symbol combination in the study. The only variables with missing values are fundamental company data, including the USD amount of a stock company's last issued dividend, the company's interest income, and the percent of the company's public shares currently sold short. When fitting a model to these data, variables with a high ratio of missing values can be dropped. Another strategy is to impute the values; This is performed later using the median value of each missing variable's non-missing values.

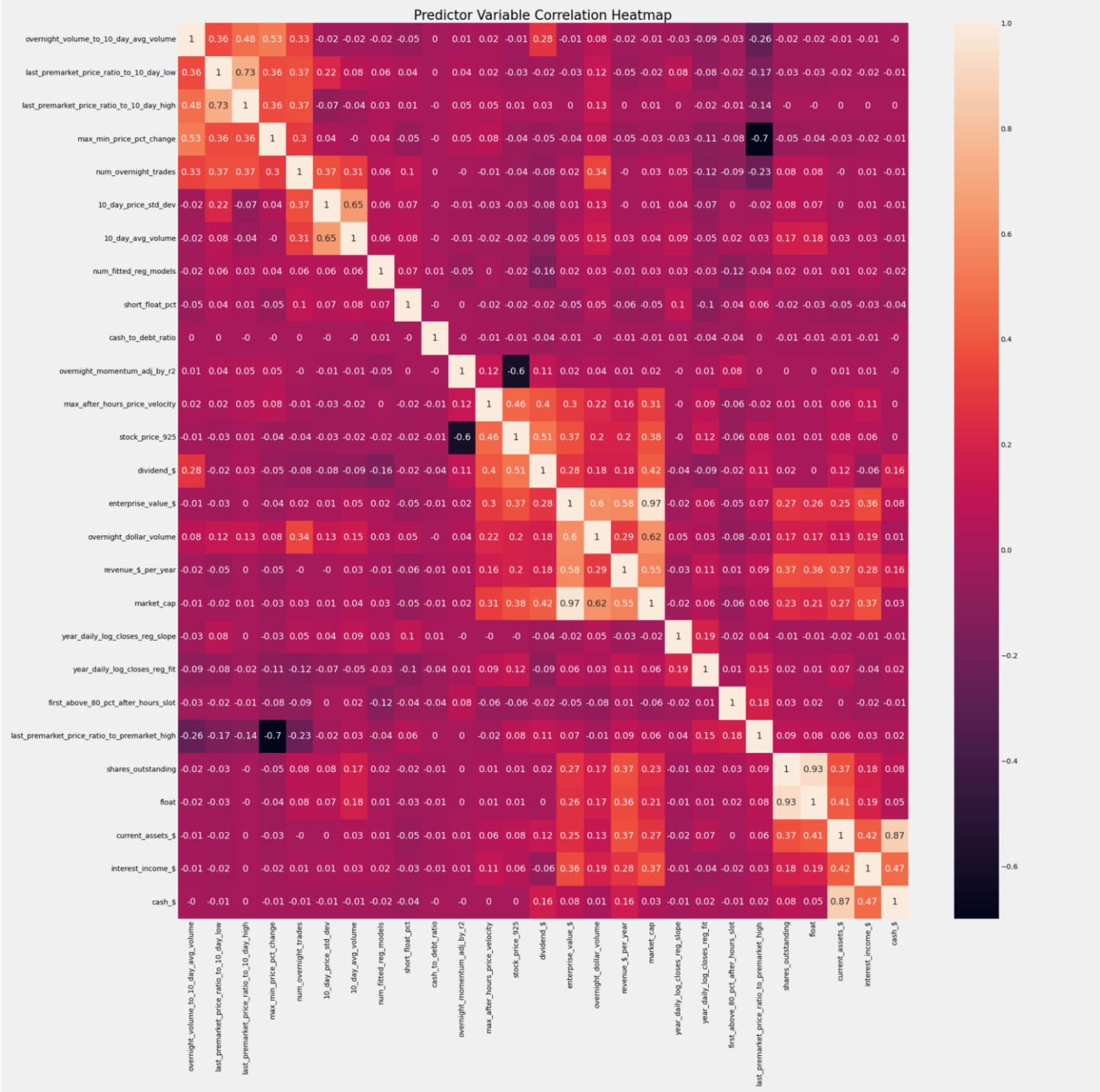
	ratio_na_pred_vars
dividend_$	0.854770
interest_income_$	0.541312
short_float_pct	0.239353
revenue_$_per_year	0.191227
year_daily_closes_reg_slope	0.143101
year_daily_closes_reg_fit	0.143101
cash_to_debt_ratio	0.136712
current_assets_$	0.099233
float	0.076235
cash_$	0.055792
enterprise_value_$	0.055792
shares_outstanding	0.023424
market_cap	0.023424
last_premarket_price_ratio_to_premarket_high	0.000000
overnight_volume_to_10_day_avg_volume	0.000000
10_day_price_std_dev	0.000000
10_day_avg_volume	0.000000
last_premarket_price_ratio_to_10_day_low	0.000000
last_premarket_price_ratio_to_10_day_high	0.000000
stock_price_925	0.000000
max_after_hours_price_velocity	0.000000
momentum_adj_by_r2	0.000000
max_min_price_pct_change	0.000000
overnight_dollar_volume	0.000000
num_overnight_trades	0.000000
first_above_80_pct_after_hours_slot	0.000000
avg_daily_volume_10D	0.000000
num_fitted_reg_models	0.000000



Next, we study the *first_above_80_pct_after_hours_slot* variable. This variable records when prices for a stock go above 80-percent of the range between the final traded price of the stock on the previous day at 16:00 and the highest price which the stock trades at during the after-market session, (between 16:00 of the previous day and 9:30 of the current day). A new time-series is formed from the median of this variable aggregated across all stocks for a given day. The median value of this time-series is around 7:00. Interestingly, the 0.1 and 0.9 quantiles are 5:26 and 8:00, which are on the same day.



Next, a heatmap of the pearson (linear) correlations between all predictor variable is shown. Due to the high number of predictor variables, it is difficult to visually inspect this heatmap plot.

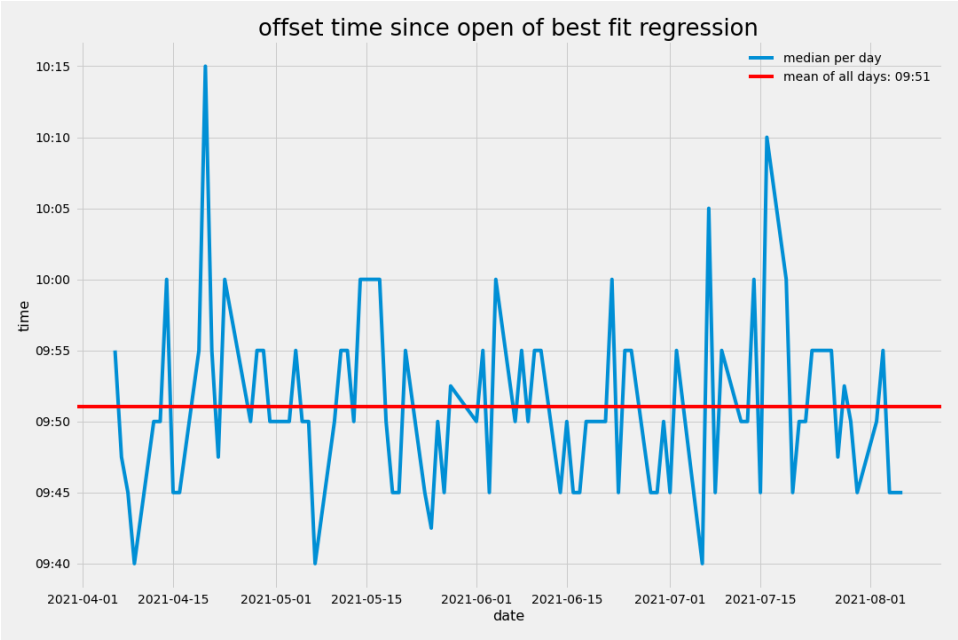


A plot of all pairs of predictor variables with a linear correlation above 0.50 is shown below. Some of the correlations appear to be spurious relationships, e.g., the stock price at 9:25 and the OLS regression slope of the previous year of daily close prices. One way to use this information to improve the predictive modeling performance is to drop the variables with high linear correlations with other variables. This technique was not used in this analysis.

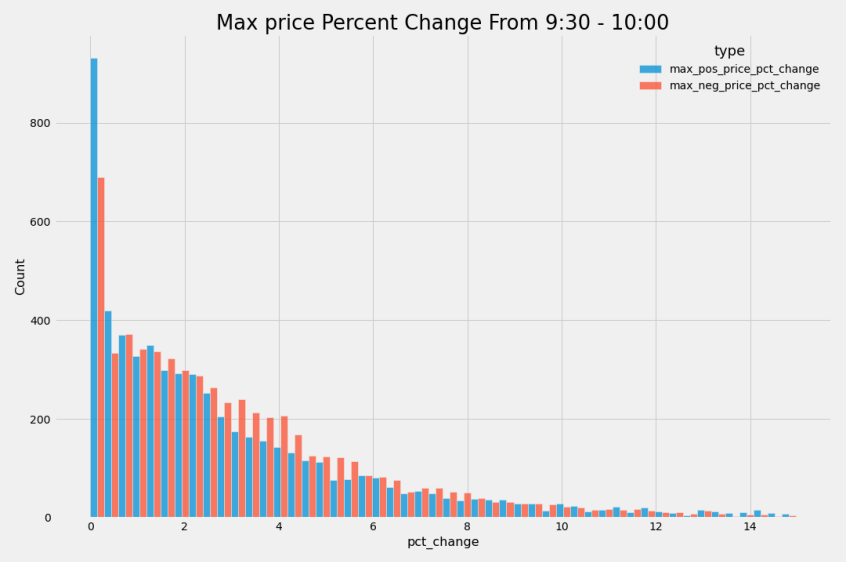
	predict_var_1	predict_var_2	corr
0	market_cap	enterprise_value_$	0.970163
1	float	shares_outstanding	0.934798
2	stock_price_925	year_daily_closes_reg_slope	0.869116
3	cash_$	current_assets_$	0.866381
4	avg_daily_volume_10D	10_day_avg_volume	0.777331
5	last_premarket_price_ratio_to_10_day_high	last_premarket_price_ratio_to_10_day_low	0.733352
6	max_min_price_pct_change	last_premarket_price_ratio_to_premarket_high	0.697779
7	10_day_avg_volume	10_day_price_std_dev	0.678469
8	market_cap	overnight_dollar_volume	0.620837
9	avg_daily_volume_10D	10_day_price_std_dev	0.602997
10	stock_price_925	momentum_adj_by_r2	0.601979
11	enterprise_value_$	overnight_dollar_volume	0.596089
12	revenue_$_per_year	enterprise_value_$	0.577947
13	market_cap	revenue_$_per_year	0.545474
14	max_min_price_pct_change	overnight_volume_to_10_day_avg_volume	0.515717
15	stock_price_925	dividend_$	0.506678
16	momentum_adj_by_r2	year_daily_closes_reg_slope	0.505572

3.3 Analysis and Selection of Target Variable

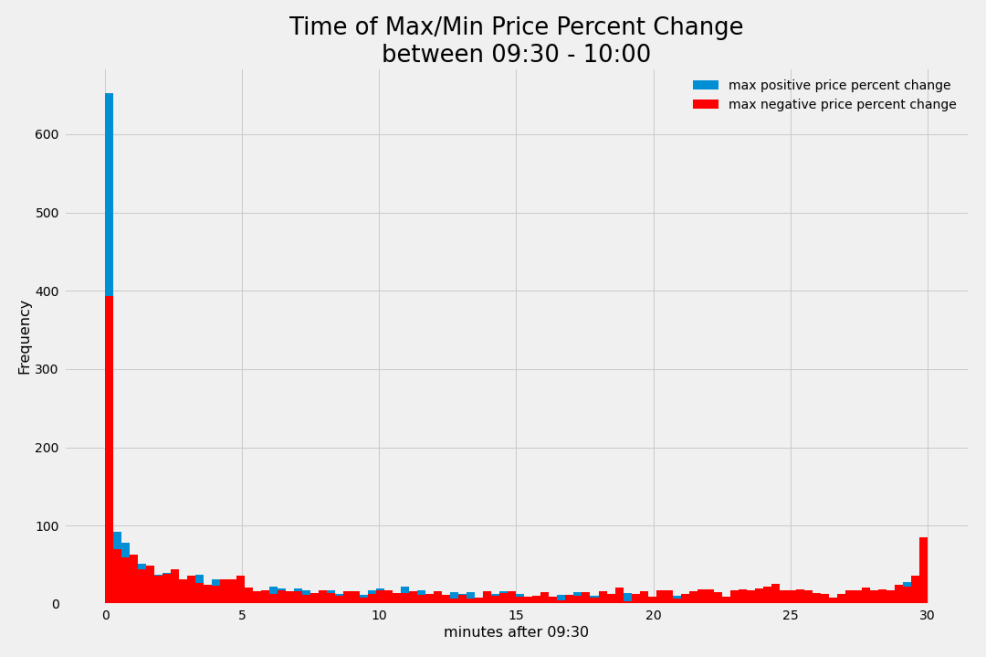
We would like to determine which data to use as the target variable. The ideal end goal is to build a profitable trading strategy. We will limit the analysis to predicting price percent change. We begin by analyzing if there is a consistent offset time every day which has the highest scoring regression fit. For the regression analysis, the time slots of the trade prices are broken up into intervals as follows: [9:30, 9:35], [9:30, 9:40], ..., [9:30, 10:00]. In the plot below, no clear interval is best suited. The median of the best fit regression per day is plotted. The median of these data is around 9:50.



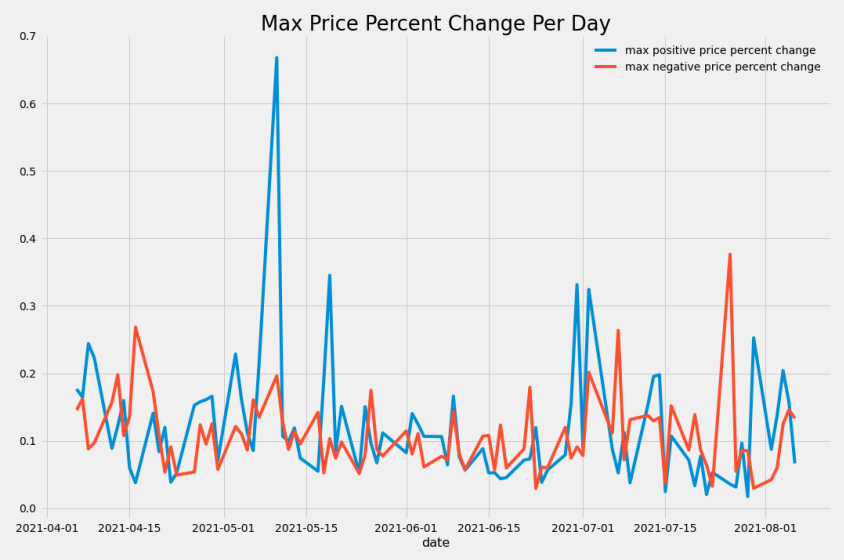
Next, we plot the frequency of the max price percent changes across all the stocks under study. The base price is the first published trade price immediately after the market opens at 9:30. Both the highest positive and highest negative price percent changes are plotted. As can be seen below, the highest frequency of the highest price percent changes for both positive and negative percent changes is closest to 0. Most of the stocks under study do not exhibit large price changes between 9:30 and 10:00.



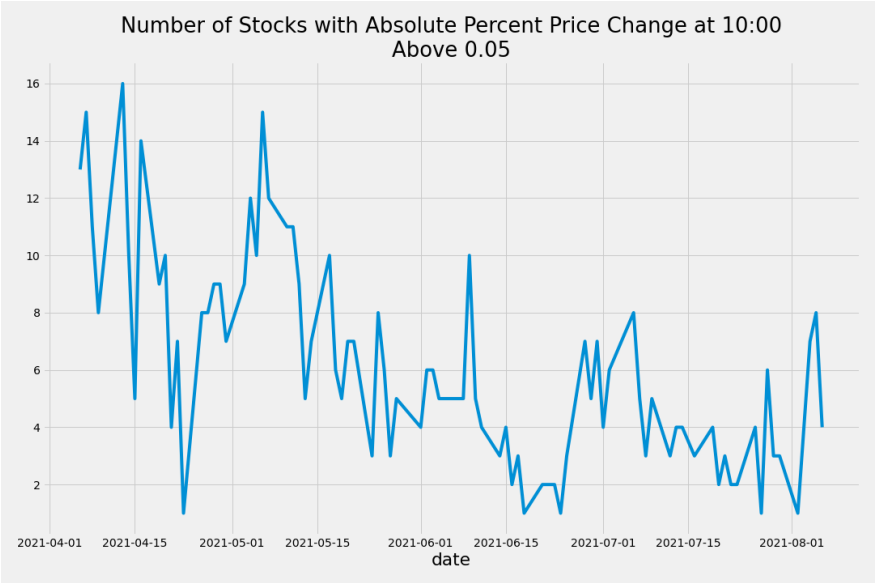
The below plot shows the frequency of the time when the max positive and max negative price percent change since 9:30 occurs. The plot shows that the highest percent price change occurs in the first minute after 9:30. We will keep this result in mind when testing a trading strategy, since the duration of an open trade would ideally be as short as possible to minimize exposure time of holding a stock which can have a black-swan event adversely affecting our position, while also profiting from a price percent change in the traded stock.



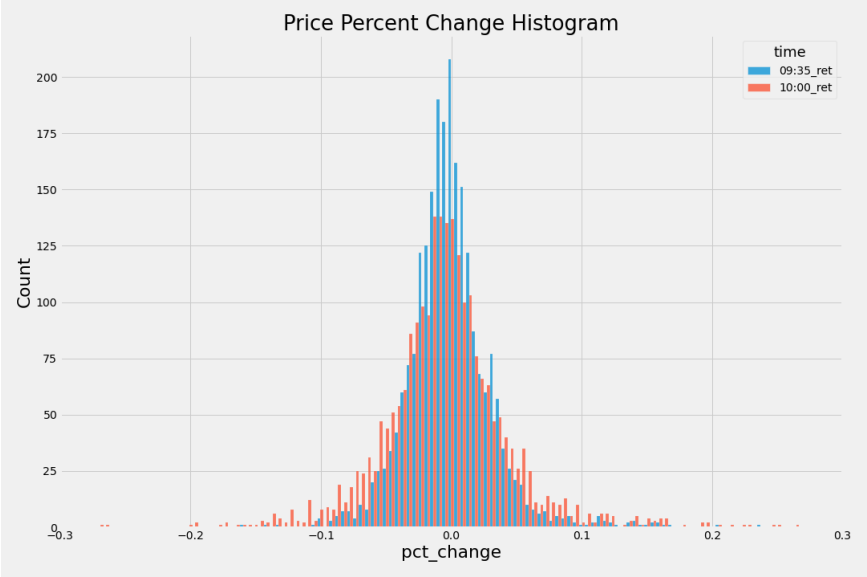
For every day under study, the below plot shows the maximum positive and maximum negative price percent change for all stocks. Occasionally, some stocks exhibit large positive price percentage changes between 9:30 and 10:00. For example, one stock made close to a 70-percent positive price change. However, the largest negative price percent change was only a 40-percent price change.



Below, we plot the daily number of stocks with an absolute price percent change between 9:30 and 10:00 above 5-percent. Over the sampled period, it appears that the number of stocks exhibiting this pattern has been decreasing. This may be because of the COVID-19 pandemic, where in 2020 alone, more than 10 million new brokerage accounts were opened [5]. As a result, U.S. stock market activity has increased. The data set in this study does not go back further in time than April 2021, so we cannot analyze the evolution of this pattern prior to and during the COVID-19 pandemic.



Below, we plot and compare the histograms of the price percent changes occurring during the intervals [9:30, 9:35] and [9:30, 10:00]. From visual inspection, the price percent changes do not appear to be generated from the same process. We confirm this by calculating the first three statistical moments of the empirical data.

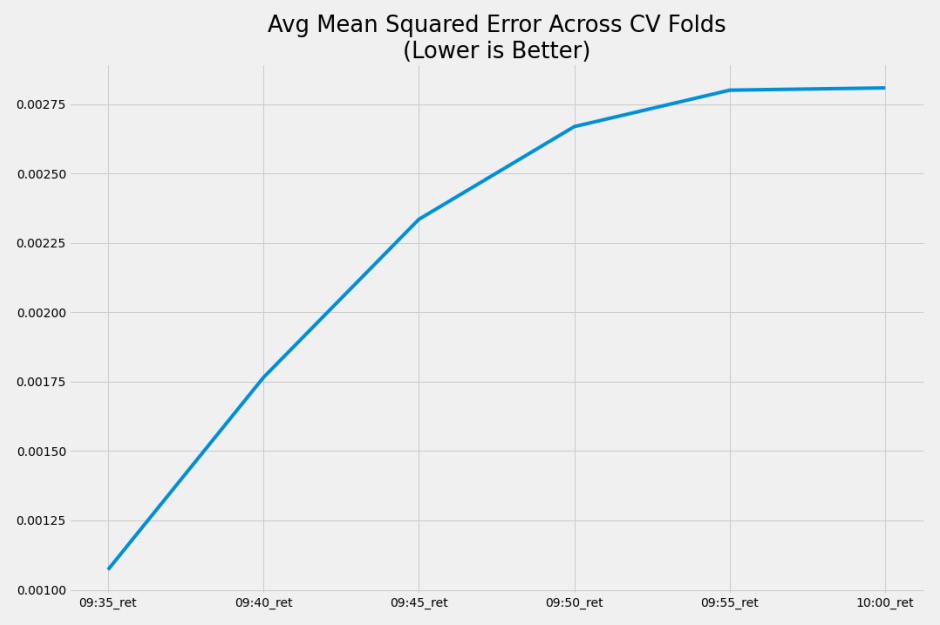


The following table displays the mean, skew, and kurtosis of the price percent changes between the intervals [9:30, 9:35] (a), [9:30, 9:40], ... [9:30, 10:00]. The distribution sampled from the smallest temporal interval (a) is closest to a standard normal distribution of all the distributions. However, it has a kurtosis value of 5.336, which differentiates it from a standard normal distribution which has a kurtosis of zero. As the price percent change is sampled over a wider temporal interval, both the skew and the kurtosis of the resulting distribution increases.

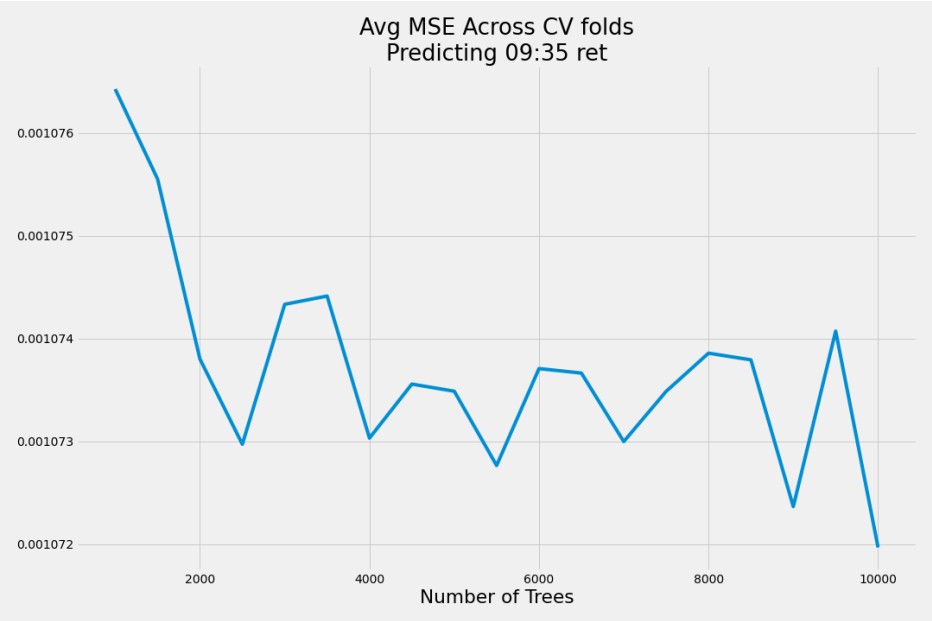
	09:35 ret	09:40 ret	09:45 ret	09:50 ret	09:55 ret	10:00 ret
mean	-0.001	-0.001	-0.002	-0.003	-0.004	-0.004
skew	0.936	1.604	2.278	1.972	1.797	1.368
kurtosis	5.336	11.673	18.036	17.543	19.645	16.664

3.4 Model Fitting and Analysis

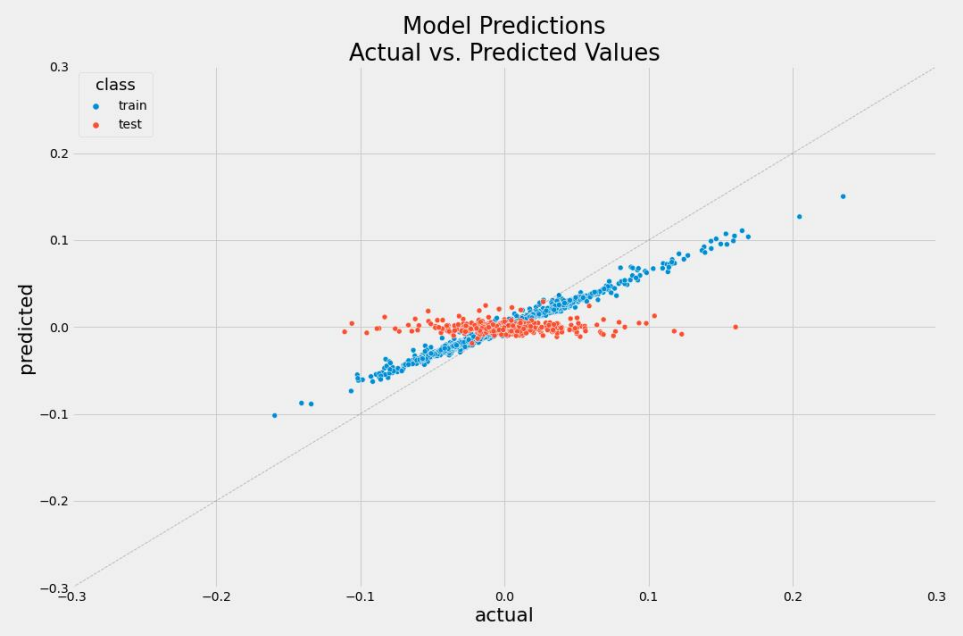
Next, we proceed to predict price percent change values using the predictor variables defined in section 3.1. We use a random forest (RF) model, due to its efficacy in predicting financial price returns [1], with 5,000 base estimators, using standard CART regression trees as the base estimators. Since some of the predictor variables have missing values, we impute the missing values with the median value of the non-missing values of the respective variable. Below, we plot the mean squared error (MSE) of different RF models predicting price percent change between the intervals [9:30, 9:35], [9:30, 9:40], ..., [9:30, 10:00]. To compute the MSE score, a 5-fold cross validation (CV) scheme was used to split the training data into new train / validation sets. The reported MSE value is the average of the MSE on all 5 hold-out CV folds per model. The temporal interval of the best MSE value was the shortest interval [9:30, 9:35]. This result is in line with results from the previous analysis, since the distribution of price percent changes in this temporal interval is closest to the standard normal distribution of all price percentage changes over all temporal intervals under study.



Next, using only the time interval [9:30, 9:35], we vary the number of base estimators in the RF model and record the resulting MSE. The number of base estimators is varied from 1,000 to 10,000 with a step size of 500. The MSE appears to decrease as the number of base estimators increases. However, the decrease of MSE values does not occur smoothly and does not form a plateau, probably since financial data is noisy. For the remainder of the study, 5,000 base estimators are used, since a higher number does not necessarily equate with better predictions due to the high noise in the data.



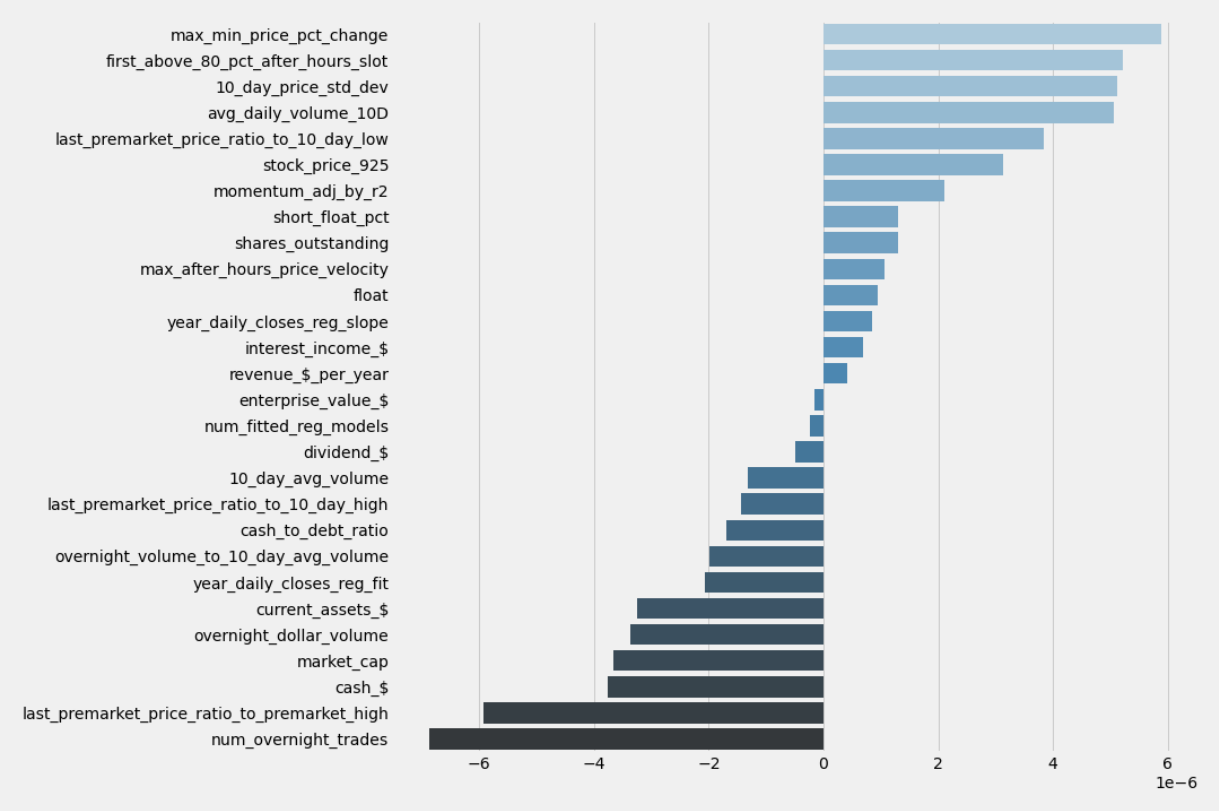
A RF CART regression model with 5,000 base learners fitted to the training data set scored a MSE value of 0.03 on the test data set, which equates with a root mean squared error of 0.173. The below plot of the actual price percent change values against the predicted values demonstrates that the predicted returns do not fit well to the actual returns, both for the training test set and even more so for the test data set. For the test data set, we see that predictions are very poor. Whether the actual price percent change is largely positive or negative, the model predicts values close to zero with a positive skew.



One possible way to quantitatively diagnose this issue is to fit a linear OLS regression to the actual vs. predicted target variable sample points. An OLS regression slope value of 1 and a r^2 close to 1 would indicate that the predicted values match the actual values well. Using these

metrics, we could then perform transformations on the predictor variables, remove predictor variables with low importance scores, etc., and choose the combination with the best resulting metrics. This technique could lead to better predictions (and better trading strategy results).

Next, we use a permutation feature importance test to rank which predictor variables are most predictive of the price percent change between [9:30, 9:35]. This test randomizes a given predictor variable while holding the other variables constant. The MSE is then calculated. An increase in MSE over the MSE of no permutations is interpreted to mean that the variable held constant is predictive and will receive a positively ranked feature importance score. On the flip side, if the MSE value decreases after permuting the variable values, the respective variable will get a negative ranked feature importance score. This diagnostic was performed using the test data set. In the plot below, we see that the variables with the highest scores were *max_min_price_pct_change*, *first_above_80_pct_after_hours_slot*, *10_day_price_std_dev*, *avg_daily_volume_10D*. The variables with the lowest importance scores were: *num_overnight_trades*, *last_premarket_price_ratio_to_premarket_high*, *cash_$*, and *market_cap*.



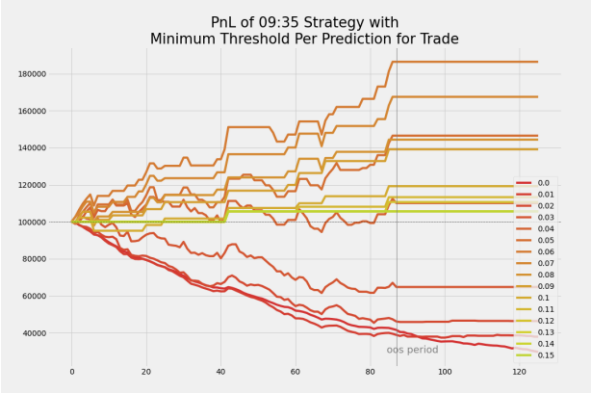
Since some predictor variables have negative permutation importance scores, we could drop these variables from the model and rerun the model training and prediction workflow. However, since the above analysis referenced the testing data set, the results would be positively biased and not indicative of future performance. To perform feature selection properly, only the training data set should be used.

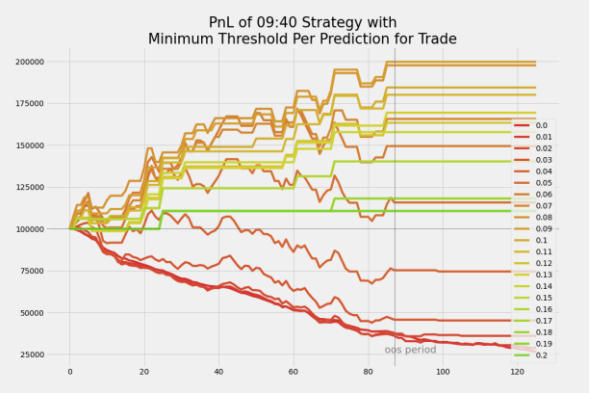
3.5 Trading Strategy Development and Analysis

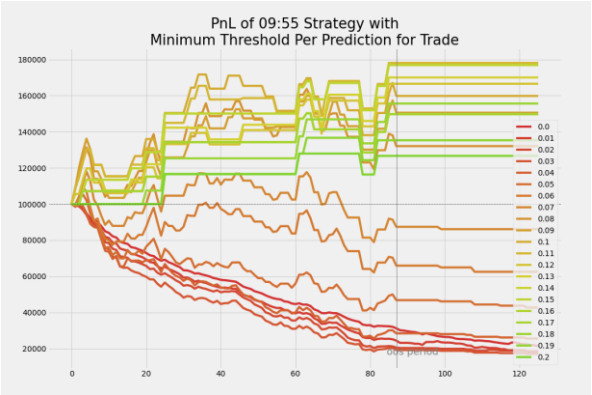
We proceed to develop a trading strategy using the previous results, including the above predictor variables, the predictive model. Trading signals will be generated after predicting the percent price change of each stock under study. Even though MSE is lowest when predicting the price change interval of [9:00, 9:35] as opposed to the interval [9:30, 10:00], financial instrument prices tend to make smaller percentage price moves in shorter time intervals. As a result, even though predictions are more accurate when predicting the shorter time window, potential trading profits might not be sufficient to have a net positive gain due to slippage and commissions costs. Hence, we will test trading strategies over 4 different price change temporal intervals each with its respective predictive model.

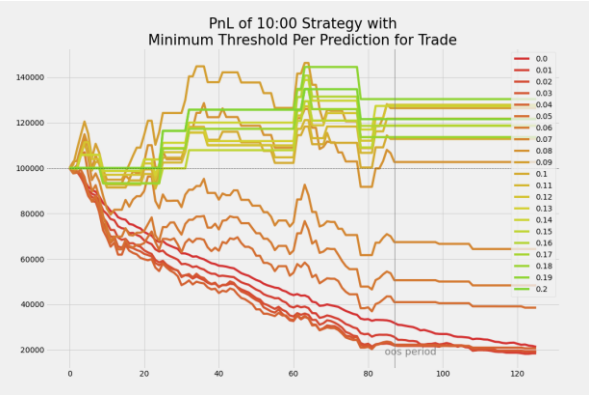
All trading strategy tests will initiate a trade at 9:30 if the predicted price change is above a given threshold. Each trading strategy will exit all positions at the end time boundary of the target prediction price changes. For example, the trading strategy of the model predicting price changes in the interval [9:30, 9:40] will exit all trades at 9:40. The tests will include commission and slippage costs of 1-percent. Also included is a commission of 5-cents per share when borrowing stock to short. Allocation of capital is divided evenly per position for each day. The maximum allocation of capital to a position is limited to a quarter of the total account capital. As an example, if only one trade is to be initiated on a given day and the account capital level is $100,000, only $25,000 will be allocated to this position.

Plots are shown of the Profit/Loss (PnL) of each strategy with different minimum predicted price change thresholds per strategy. The plots display that as the minimum threshold per trade is increased, less trades are made. Most of the thresholds do not permit any trades in the out-of-sample (OOS) period. This result corroborates what was shown above in the actual vs. predicted plot in section 3.3. It was demonstrated that most of the predictions of the OOS (test data set) period had low fluctuation around zero, while the actual values varied more widely around zero. Hence, increasing the minimum absolute threshold of predicted price changes will drastically diminish the number of permitted trades.









Statistics of the trading strategy which exits all positions at 9:35 with a minimum absolute predicted threshold of 0.04 are shown below.

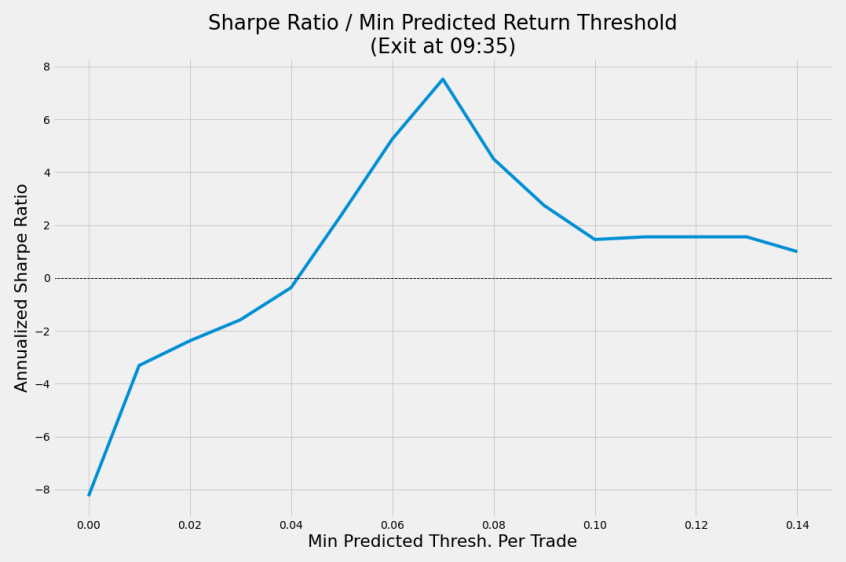
	min trading thresh: 0.04
init_capital	100000
ending_capital	110086.723491
net_return	0.100867
avg_positions_per_day	0.904
slippage_long	0.01
slippage_short	0.01
avg_locate_fee_per_short_per_share	0.05
sharpe	0.437414
max_drawdown	-0.189607
slippage_and_comm	12542.354597
exit_time	09:35:00

For the trading strategy with these parameters, a description of the first 10 trades is shown below. The columns *pos_amount*, *ending_position_capital*, and *slippage_and_comm* are currency units denominated in USD.

		ret	**trade_ret**	**predicted_ret**	**mae**	**direction**	**pos_size**	**pos_amount**	**ending_position_capital**	**slippage_and_comm**
2021-04-06	**SJ**	0.146	-0.092	-0.047	0.025	-1.0	-1330	19790.40	17908.252	66.649
	PHUN	0.009	0.105	0.073	0.023	1.0	7387	15882.05	17553.267	0.022
	ACY	0.116	0.063	0.049	0.060	1.0	942	12726.42	13533.089	0.135
	EVAX	0.100	-0.094	-0.055	0.019	-1.0	-1740	10213.80	9164.287	87.059
	PLBY	0.045	-0.083	-0.052	0.000	-1.0	-283	8178.70	7488.152	14.439
2021-04-07	**UTME**	0.157	0.089	0.101	-0.000	1.0	453	24457.47	26641.782	0.540
	ASTC	0.133	-0.081	-0.044	0.042	-1.0	-6461	18413.85	16597.094	323.078
	ISIG	0.165	0.079	0.066	0.007	1.0	1968	13854.72	14945.109	0.070
2021-04-08	**DOGZ**	-0.053	0.127	0.095	0.065	1.0	11442	24829.14	27981.921	0.022
	IPA	0.086	-0.091	-0.058	0.000	-1.0	-1401	18675.33	16908.405	70.183

The sharpe ratio (calculated here as the ratio of annual return to annual standard deviation of returns) of each strategy exiting at 9:35 with different absolute minimum predicted threshold before initiating a trade is shown below. In the plot, some sharpe ratio values are obviously excessively positive (i.e., values above 2). These values are not statistically sound since as the minimum absolute predicted threshold is increased, the sample size of trades diminishes to a

very low number. Furthermore, almost all the trading strategies with high sharpe ratios only initiate trades during the in-sample period, with no trades occurring during the out-of-sample period. This is an artifact of the difference between the distribution of the predicted price percent changes of the in-sample period, where the range of predicted variables is widely dispersed around zero, and the OOS period, where the predicted values are narrowly dispersed around zero. Hence, almost no trades are initiated in the OOS period.



4. Conclusion

In this project, exploratory data analysis, multiple predictor variables, and machine learning techniques were used to forecast price changes and develop a trading strategy.

Custom predictor variables and different target variables of the post market open price changes were built. A rigorous analysis (i.e., exploratory data analysis) of these target variables was performed. Next, motivation for utilizing a RF CART model was provided, and the model was trained and tuned on in-sample data. Analysis of the model performance was presented, including permutation feature importance using the trained model for the predictor variables. Besides the benefit of improving predictive model scores, the results provide discretionary traders with insight into which variables are most predictive of percent price change in the post U.S. regular trading session.

Next, a basic trading strategy which employs the predictive model was designed. The simulated trading results were not encouraging. This may be a result of the high price volatility which occurs immediately after the U.S. regular trading session opens. As such, further refinements to the trading strategy should be tested, such as the timing of entry and exit of positions depending on the current respective stock price dynamics.

The motivation for this project was to discover if it is possible to trade these volatile stocks profitably, or if the endeavor is better yet avoided. The tests of the simulated strategies demonstrated that the most consistent trading strategies across both in-sample and out-of-sample periods were losing strategies. This is due to higher frequency trading incurring higher commission and slippage costs. In the section on model fitting, it was shown that the combination of the predictor variables, the chosen predictive model, and the target predicted

variable results in poor performing predictions. This is most likely due to the high level of noise in financial data, especially in volatile stocks immediately after the U.S. regular trading session open. We conclude that the prediction of price percent change between 9:30 and 10:00 of overnight volatile stocks is a non-trivial task. However, we abstain from commenting on if a trading strategy on this universe of stocks using more sensitive entry and exit criteria can exhibit desirable statistical properties of a trading strategy. As such, a recommended avenue for further research is development and testing of robust trading strategies which trade using dynamic entry and exit criteria.

5. Appendix

Predictor variables categorized by type:

FUNDAMENTAL VARIABLES
stock_price_925
market_cap
short_float_pct
float
shares_outstanding
cash_$
cash_to_debt_ratio
revenue_$_per_year
dividend_$
interest_income_$
current_assets_$
enterprise_value_$
NOVEL VARIABLES
num_fitted_reg_models
num_overnight_trades
overnight_dollar_volume
max_min_price_pct_change
overnight_momentum_adj_by_r2
max_after_hours_price_velocity
last_premarket_price_ratio_to_premarket_high
last_premarket_price_ratio_to_10_day_high
last_premarket_price_ratio_to_10_day_low
TRADITIONAL VARIABLES
10_day_avg_volume
10_day_price_std_dev
overnight_volume_to_10_day_avg_volume
year_daily_closes_reg_slope
year_daily_closes_reg_fit

The source code for this project is available at http://www.github.com/nrenzoni/forecasting-and-trading-volatile-stocks-wqu.

6. References

[1] Ballings, Michel, et al. "Evaluating multiple classifiers for stock price direction prediction." Expert systems with Applications 42.20 (2015): 7046-7056.

[2] Chen, Chun-Hung, Wei-Choun Yu, and Eric Zivot. "Predicting stock volatility using after-hours information: evidence from the Nasdaq actively traded stocks." International Journal of Forecasting 28.2 (2012): 366-383.

[3] Clenow, Andreas F. Stocks on the Move: Beating the Market with Hedge Fund Momentum Strategies. Equilateral Publishing, 2015.

[4] Gerety, Mason S., and J. Harold Mulherin. "Trading halts and market activity: An analysis of volume at the open and the close." The Journal of Finance 47.5 (1992): 1765-1784.

[5] 2021 U.S. self-directed investor Satisfaction Study. J.D. Power. (2021, April 27). Retrieved October 12, 2021, from https://www.jdpower.com/business/press-releases/2021-us-self-directed-investor-satisfaction-study.

[6] Cerlymarco. (n.d.). Cerlymarco/linear-tree: A python library to build model trees with linear models at the leaves. GitHub. Retrieved October 12, 2021, from https://github.com/cerlymarco/linear-tree.