MS&E 448 Final Presentation
High Frequency Algorithmic Trading

Francis Choi George Preudhomme Nopphon Siranart
Roger Song Daniel Wright

Stanford University

June 6, 2017

Overview

- Review our strategy and progress from the midterm
- Changes in Data Processing
- Changes to Models
- Strategy and Simulations
- Results
- Evaluation and Next Steps

Recall from the Midterm

- **Goal:** Next-minute price movement prediction based on order book dynamics
- **Data:** Minute-by-Minute consolidated book for S&P 500 ETF (IVV)
- **Model:** Random Forest three-way classifier
- **Labels:** Mid-price changes and spread-crossing
- **Trading Strategy:** Accumulating positions and closing them out at the end of the day
- **Results:** Still not generated profit

After the Midterm

Data Processing

- Changing the data from minute by minute to second by second
- Change from three-way classification to binary classification (no longer using spread crossing label)
- Train and test on a rolling window basis - 2 weeks training period prior to each day

Data (Example)

jupyter GettingData Last Checkpoint: Yesterday at 3:35 PM (unsaved changes)

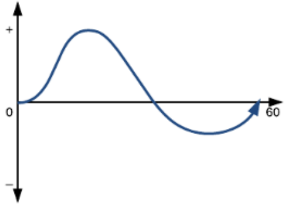
In [58]:
```
df[['asks_nord_0', 'asks_price_0','asks_size_0', 'bids_nord_0', 'bids_price_0', 'bids_size_0']]
```

Out[58]:

	asks_nord_0	asks_price_0	asks_size_0	bids_nord_0	bids_price_0	bids_size_0
0	12	208180000	3800	2	208150000	600
1	3	208180000	800	3	208160000	700
2	3	208180000	800	2	208160000	600
3	4	208180000	1200	1	208160000	200
4	2	208170000	300	8	208160000	900
5	2	208150000	300	1	208140000	200
6	1	208150000	200	1	208140000	200
7	4	208150000	800	6	208140000	800
8	3	208150000	400	5	208140000	700
9	4	208150000	800	1	208130000	200
10	4	208150000	800	12	208120000	4603
11	4	208150000	800	1	208130000	200
12	4	208150000	800	1	208130000	200
13	5	208150000	1000	1	208130000	200
14	4	208150000	800	1	208130000	200
15	5	208150000	1000	1	208130000	200
16	8	208150000	1300	1	208130000	200
17	9	208150000	1500	9	208120000	2503

New Labels

- **AREA**
  - Time-weighed PnL over the next period (area under the price movement curve)



- **VWAP**
  - Volume-weighted average price (VWAP) based on inner bid and ask.
  - Whether it goes up or down in the window.

After the Midterm

Adding new features

- **Bid-Ask Volume Imbalance** Quantity indicating the number of shares at the bid minus the number of shares at the ask in the current order book.
- **VWAP** A variation on mid-price where the average of the bid and ask prices is weighted according to their inverse volume.
- **Second Order Derivatives** Expand feature universe to encompass multiple time periods.

Model

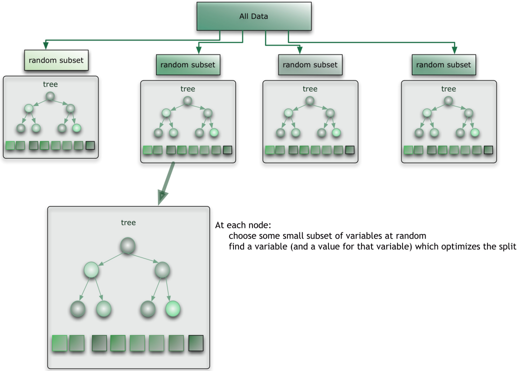
Logistic Regression

- Outputs probability (how confident we are) on each trade
- Advantages over random forest: it trains much faster, the coefficients have an interpretation

Model

Random Forest

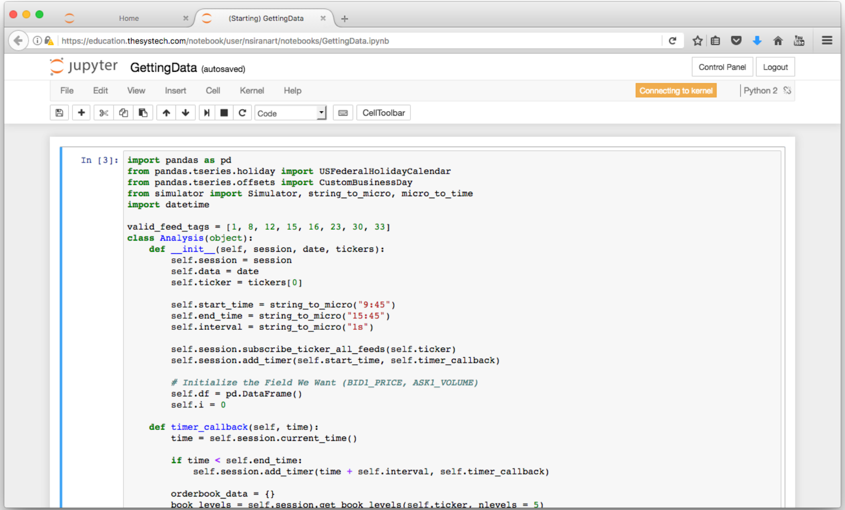
- Again, outputs probability (how confident we are) on each trade
- One key advantage over logistic regression - doesn't assume any functional form and slightly higher accuracy



Strategy

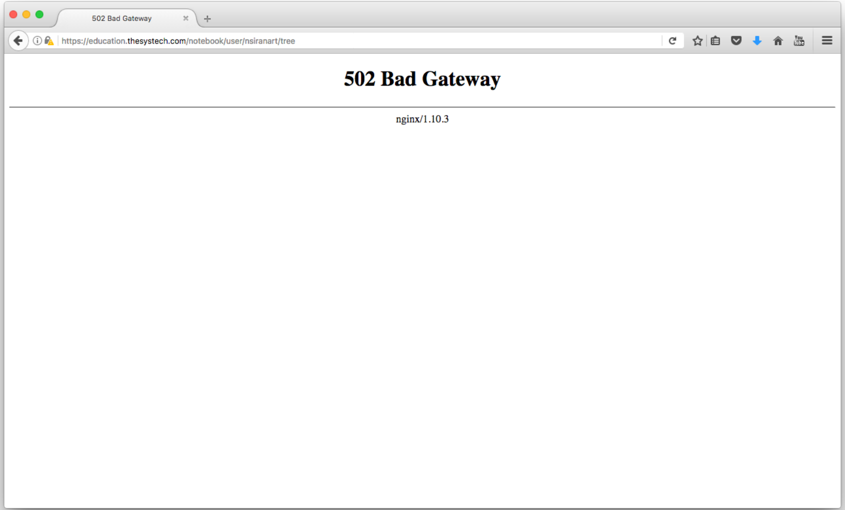
- Train the model on a rolling backwards window.
- At each second, use the model to arrive at a prediction with a probability estimate.
- If the probability estimate is above the threshold, make the predicted trade with the size weighted accordingly
- Close out the trade at the end of the trading window.

Here is what we think it looks like



Thesys Simulator

Here is what it actually looks like



Thesys Simulator

- Very frustrating and very slow
- We decided to just pull the data from Thesys and do the simulations manually.

Results

- We choose 10 stocks and ETFs to test our trading strategies, chosen based on liquidity
- These include XLF, CSCO, EEM, IVV, IWM, QQQ, UVXY, VXX, XLE, SPY
- Training Period - 2 weeks from 01/05/2015 - 01/16/2015
- Test Period - 2 weeks from 01/19/2015 - 01/30/2015
- We use PnL per trade as a performance metric

Tuning Parameters

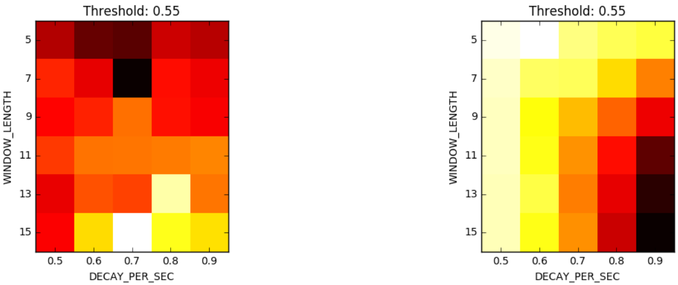


Figure: Heat map of accuracy for different decay and window length parameters (Left) XLE (Right) XLF

Accuracy of Model: Logistic Regression

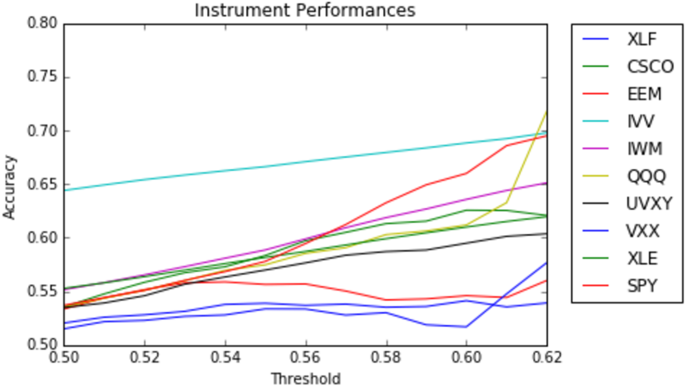


Figure: Prediction accuracy vs prediction threshold for the logistic regression model

Accuracy of Model: Random Forest

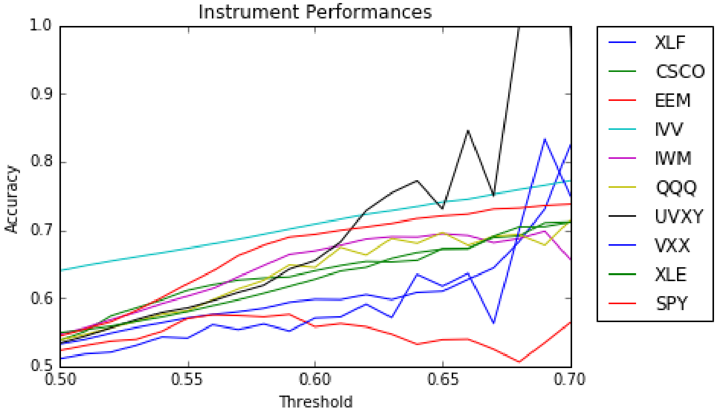


Figure: Prediction accuracy vs prediction threshold for the random forest model.

Accuracy of Model: Difference

Overall, Random Forest has slightly better accuracy across threshold values.

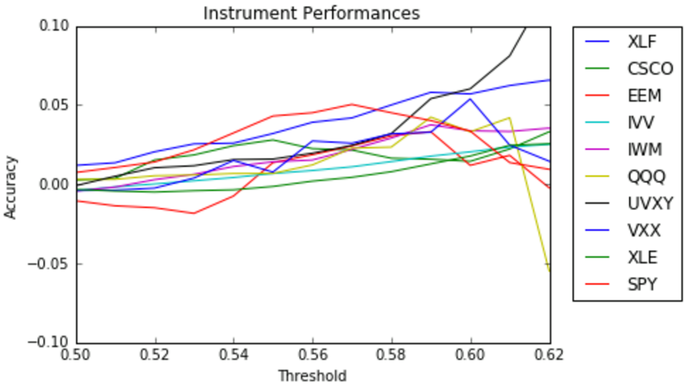


Figure: Prediction accuracy RF - LR vs prediction threshold.

Cumulative PnL (XLF)

PnL stably increasing throughout the day - High Sharpe Ratio !!

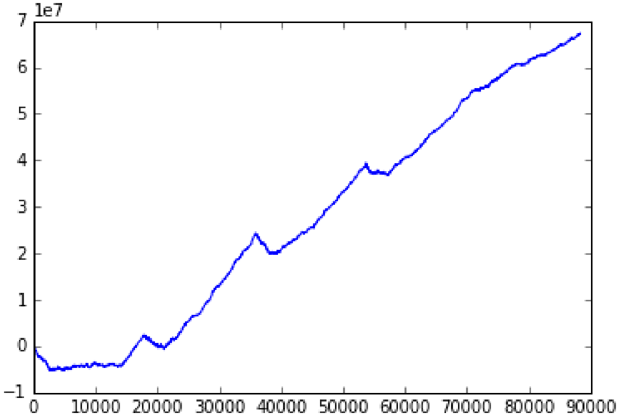


Figure: Cumulative PnL within a day

Trading PnL (XLF)

Logistic Regression with VWAP label performs best in this case

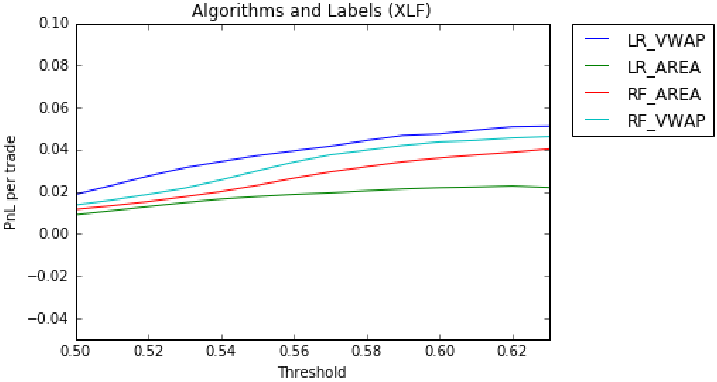


Figure: PnL per Trade vs prediction threshold for each algorithm and label

Trading PnL (XLF)

Tuning hyperparameters improves the model significantly

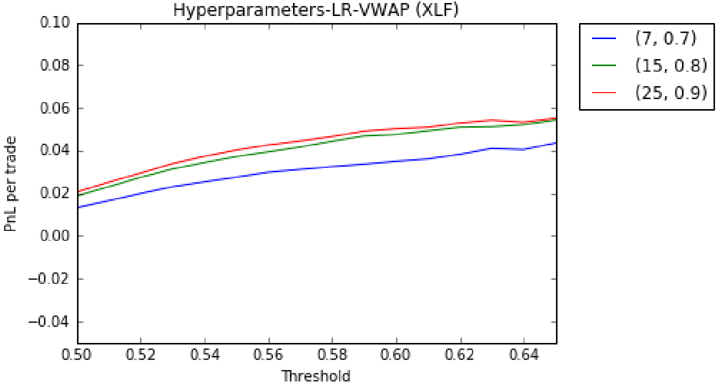


Figure: PnL per Trade vs prediction threshold for different hyperparameters

Random Forest with AREA label performs best for MSFT

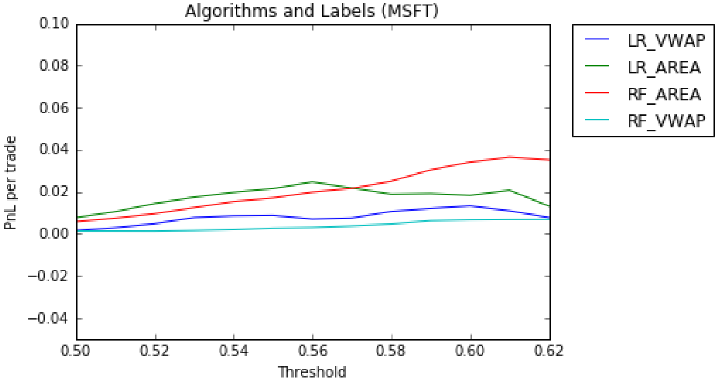


Figure: PnL per Trade vs prediction threshold for each algorithm and label

Trading PnL (MSFT)

A combination of non-optimal hyperparameters, models and labels performs poorly.

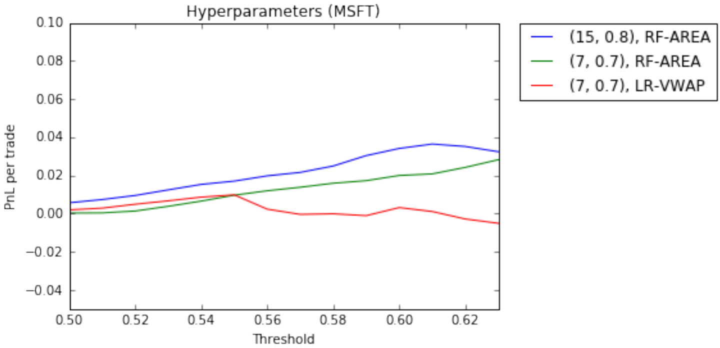


Figure: PnL per Trade vs prediction threshold for different hyperparameters

Multiple Stocks

Random Forest with AREA labels. Window = 15, decay = 0.8

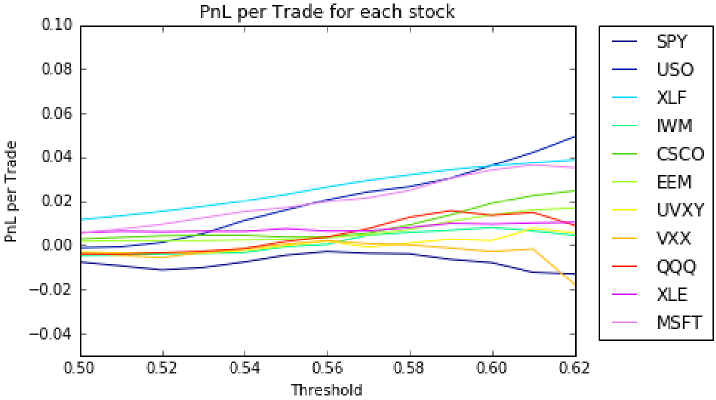


Figure: PnL per Trade vs prediction threshold for different stocks

Logistic Regression with AREA labels. Window = 15, decay = 0.8

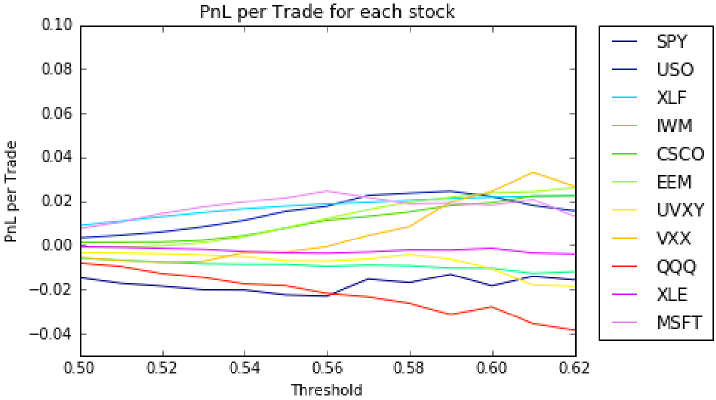


Figure: PnL per Trade vs prediction threshold for different stocks

Evaluating Our Strategy

Strengths:

- High accuracy rates: model is doing a good job
- High PnL per trade with small variance especially when training on a longer period of time
- The model can be generalized to multiple stocks/ETFs
- Perform well even in tumultuous historical periods and on hypothetical scenarios

Limitations:

- Have to tune hyperparameters for each stock
- High prediction accuracy does not always mean profit: label isn't exactly a prediction of PnL
- Interpretability of the model

Future Work and Areas for Improvement

- Within 10 weeks, we can't make the perfect trading strategy: there is still a lot we could improve.
- Some ideas for further work:
  - Training on a longer period of time
  - More sophisticated features: right now we only use the order book data, could try including external features (such as an index like the VIX, or data on correlated securities, etc.)
  - Converting to a strategy that trades at bid and ask (rather than midprice)
  - Modifying strategy to handle scaled-up trade quantities
  - Risk Management

Conclusion

- Idea: use machine learning techniques on the order book to make price movement predictions. Trade on these predictions to make $$$
- Models: Random forest, logistic regression
- Data: Second-by-second orderbook data from Thesys
- Calibrated trading frequency, prediction label, hyperparameters of models
- Performed simulations on historical data
- Promising results that can be built upon

The End

Questions?