MS&E 448

Trading forex with a distributed quote book

Jingbo Yang, Xiaoye Yuan, Carolyn Kao, Jiachen Ge, Jon Braatz, Sunny Shah

With data provided by Integral
Under guidance of Dr. Lisa Borland and Dr. Enguerrand Horel.

Overview

- Data from Integral
    - Preprocessing
- Latency arbitrage
- Machine learning based
    - Classification (up/down)
    - Regression (rate of return)
- Final report and beyond

Data from Integral

8 Currency pairs

- USD/CAD, USD/CHF, USD/JPY, USD/SEK, AUD/USD, EUR/USD, GBP/USD, NZD/USD

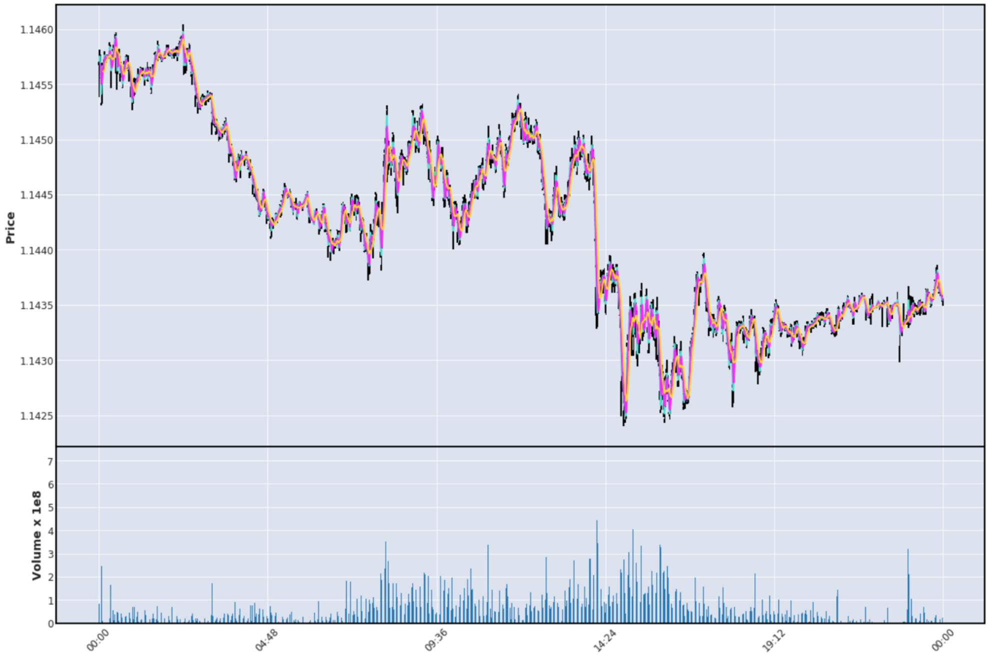
Across 1 month, 5 liquidity providers (LPs)

- February 1st, 2019 - March 1st, 2019
- Sunday: starting at 1800 (discard, too few trades)
- Monday–Thursday: 24 hours
- Friday: Ends at 2200 (discard after 1800)
- 25 days, ~400 active hours

provider	currency pair	time	bid price	bid volume	ask price	ask volume
LP-1	EURUSD	02.25.2019 00:00:00.819	1.13417	1000000	1.13424	1000000
LP-1	EURUSD	02.25.2019 00:00:00.819	1.13417	1000000	1.13423	1000000
LP-1	EURUSD	02.25.2019 00:00:00.819	1.13417	1000000	1.13423	1000000
LP-1	EURUSD	02.25.2019 00:00:00.841	1.13411	1000000	1.13423	1000000
LP-1	EURUSD	02.25.2019 00:00:00.841	1.13411	1000000	1.13423	1000000

Data Preprocessing

- We have 7 currency pairs and asynchronous quote updates
- Aggregate data over past 10, 30, 60, 300, and 600 seconds to get:
  - Open
  - Close
  - High
  - Low
  - Volume



Since Midterm Presentation

- Enabled a larger set of data
  - USD/CAD, USD/CHF, USD/JPY, EUR/USD, GBP/USD, NZD/USD
- Latency arbitrage
- More and better models
  - Linear regression using LightGBM, Kernel Ridge, Linear Regressor
  - Classification using LightGBM, KNN
  - Hyperparameter-tuning and error analysis

Latency Arbitrage

What is latency arbitrage?

Take advantage of price difference between quotes by liquidity providers (LP)

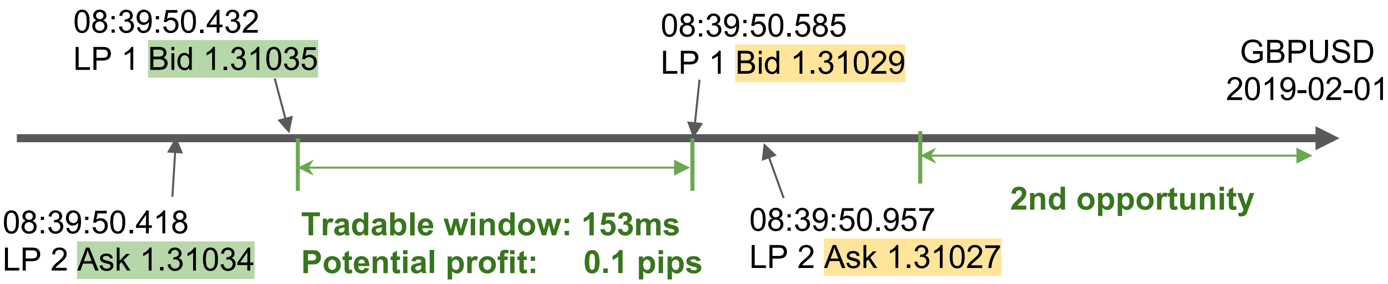
Trading window

Start: as soon as highest bid > lowest ask (negative spread)

End: LP change bid/ask so spread is no longer negative

Only interested in windows long enough for data to make round trip from NYC to Chicago

Capturing Fillable Orders



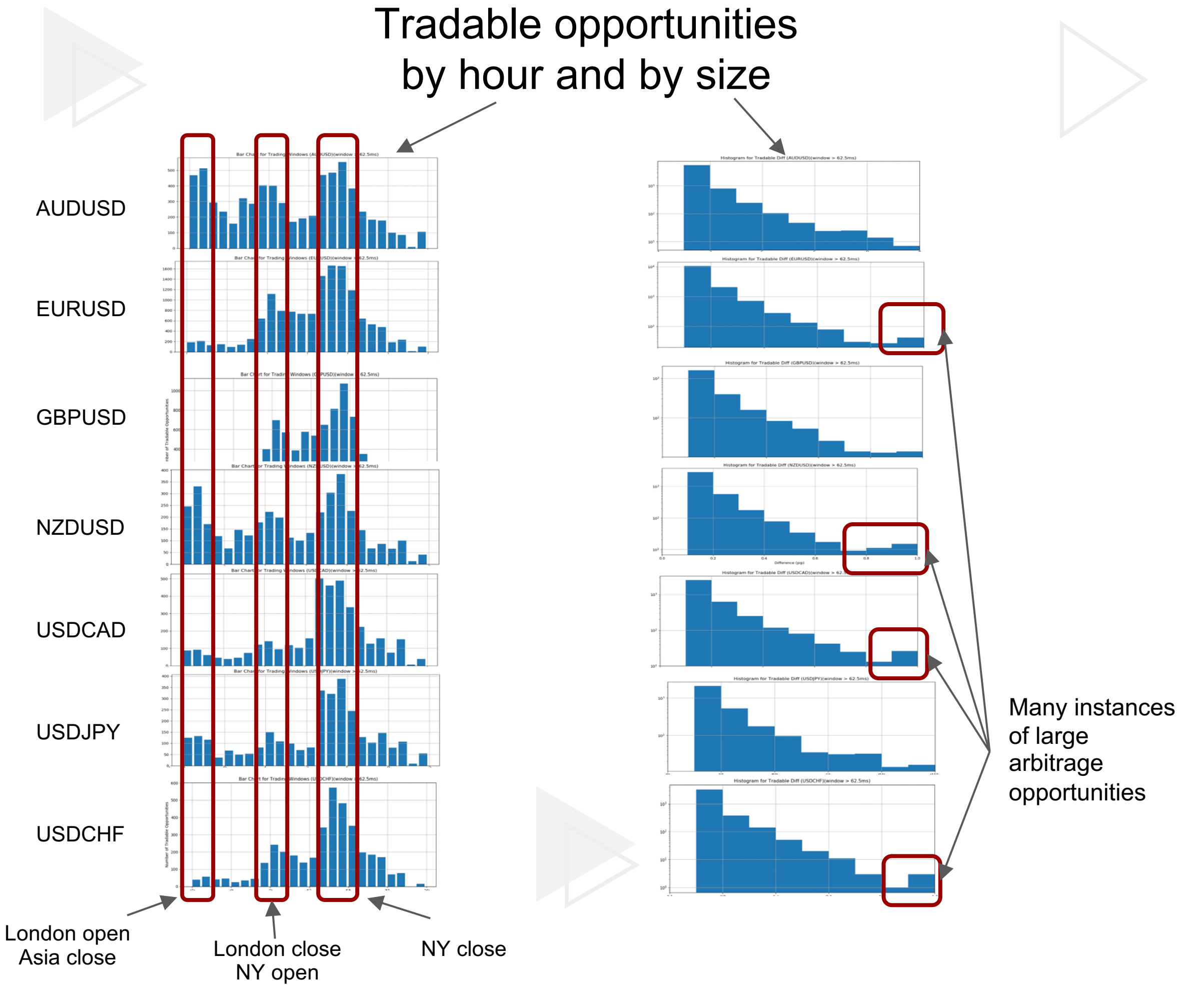
Network round trip time between Google servers

- NYC/Chicago: 46ms
- NYC/San Francisco: 140ms
- NYC/London: 143ms
- NYC/Hong Kong: 495ms

We picked **62.5ms** for minimum tradable window.

Observations

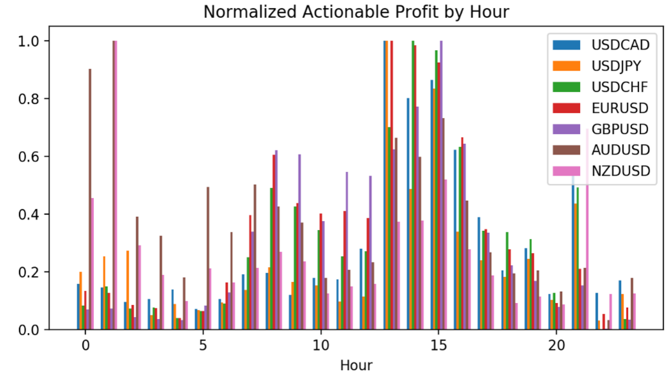
- Pairs have different activity patterns
- Lots of low spread and short window opportunities but large/long ones do exist

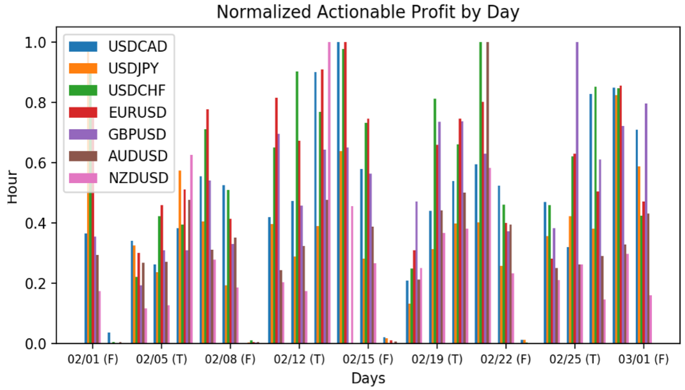


Trading Profits

Most profitable in NY afternoon and on Thursdays

EURUSD with most actionable profit





Pair	Potential Profit (0ms) (pips)	Actionable Profit (62.5ms+) (pips)
EURUSD	85000.0	**2152.5**
GBPUSD	50100.0	1431.6
AUDUSD	73700.0	999.5
NZDUSD	**92300.0**	656.6
JPYUSD	10380000	61820.0
USDCAD	44300.0	685.7
USDCHF	38000.0	513.4

Machine Learning Signal Generation

ML Modeling

Input

- Open/High/Low/Close/Volume of one currency pair
- Scale input by last close (mean of last close is 1)

Models

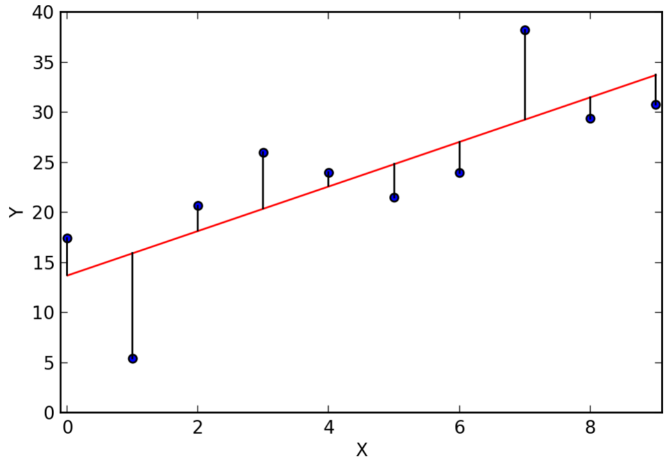
- Regression
  - Target: Return for next close compared to current close
  - LightGBM, Kernel Ridge, Linear regressor
- Binary classification
  - Target: Average of next close is higher/lower than current close
  - LightGBM and KNN classifier

Regression - Linear Regression

- Linear regression assumes the return at the end of the next time step is an affine function of the previous open, high, low, close, and volume, plus a mean-zero "error term"
  - Or, a "factor model" using those previous statistics as factors

- Parameters: None

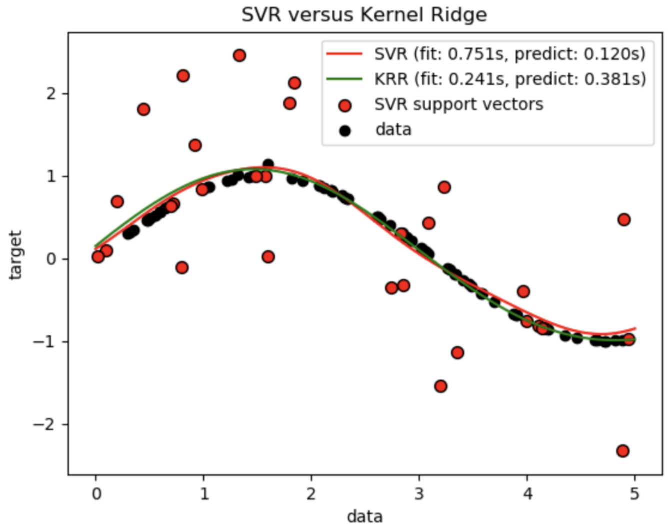
- Loss function: Mean Squared Error (MSE)



Regression - Kernel Ridge

- Kernel ridge regression (KRR) combines Ridge regression (linear least squares with L2-norm regularization) with the kernel trick.

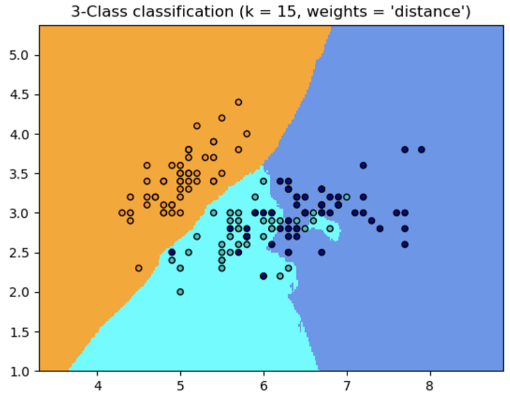
- Parameters:
    - alpha (Regularization strength)
    - kernel (Kernel mapping used internally)
    - gamma (Gamma parameter for RBF kernel)
    - degree (Degree of the polynomial kernel)

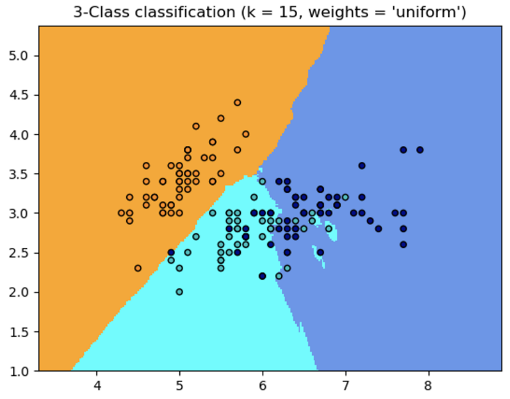


Binary Classification - KNN

- K Nearest Neighbors (KNN) classification
  - Find distances between a query and all data examples
  - Selected k nearest neighbors to the query
  - Vote for the most frequent label

- Parameters:
  - n_neighbors (number of neighbors to use)
  - Leaf_size
  - P (Measure distance using LP norm)

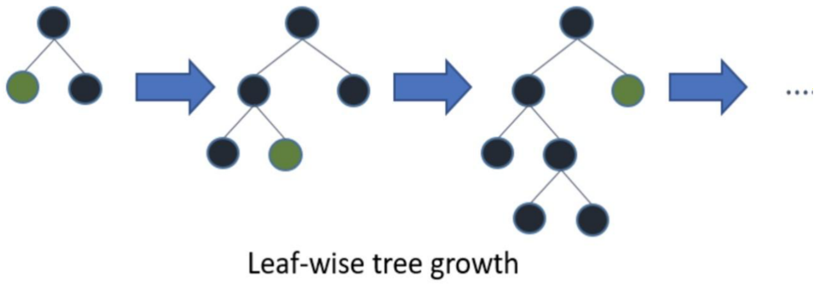




Classification/Regression - LightGBM

- Light Gradient Boosting Machine (LightGBM)
  - gradient boosting framework
  - tree based learning algorithms.
  - grows tree vertically (leaf-wise), by choosing the leaf with max delta loss.

- Parameters:
  - max_depth
  - num_leaves
  - min_split_gain
  - reg_alpha
  - reg_lambda



Training Pipeline

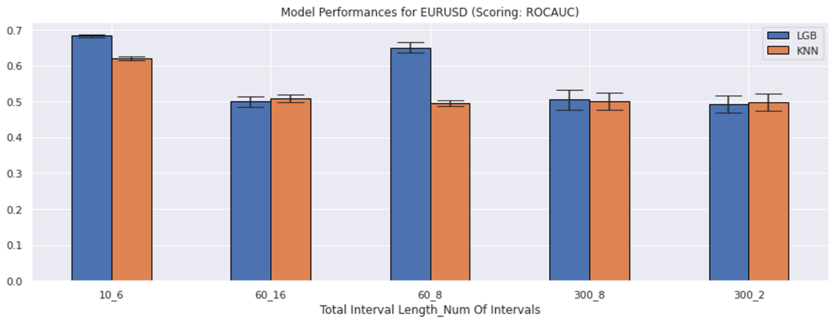
- Train (cross val) Test
  - Train on first 20 days
  - Test of last week (5 days)
- Hyperparameter search using GridSearchCV
  - LGBM
  - KNN
  - Kernel Ridge

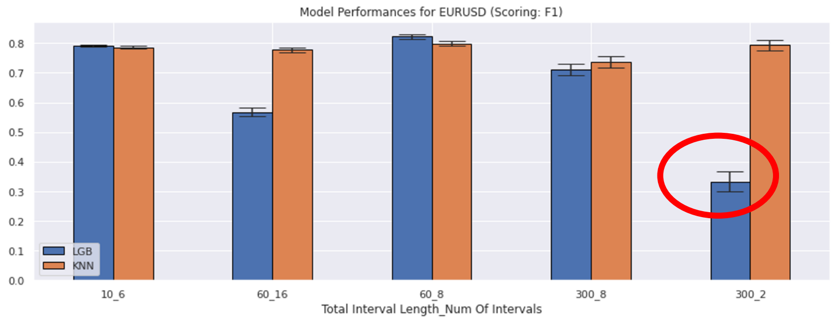
Sun.	Mon.	Tue.	Wed.	Thur.	Fri.	
					2/1	Train
2/3	2/4	2/5	2/6	2/7	2/8	
2/10	2/11	2/12	2/13	2/14	2/15	
2/17	2/18	2/19	2/20	2/21	2/22	
2/24	2/25	2/26	2/27	2/28	3/1	Test

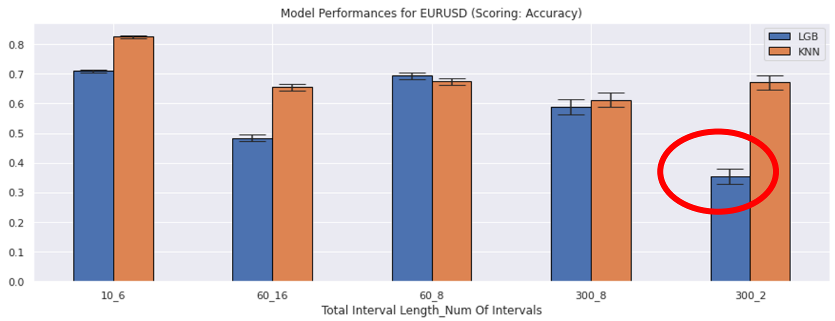
Binary Classification

Sample Statistics: EURUSD

- x_axis: interval length (sec)_numinterval
- y_axis: Scoring (ROCAUC, F1, ACC)
- Observation across **classifiers**
    - 10_6: Stably high performance
    - 60_16: KNN outperforms LGB
    - 60_8: LGB sometimes dominates
    - 300_8: Tie
    - 300_2 KNN strictly outperforms
- Observation across **intervals**



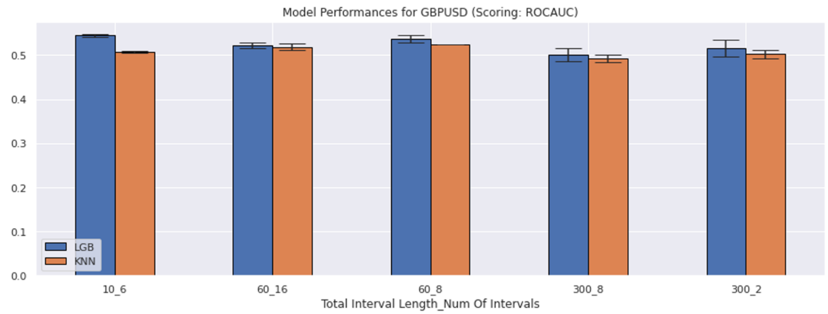


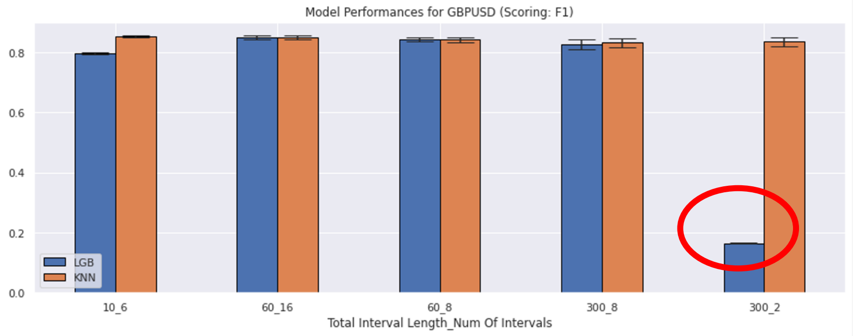


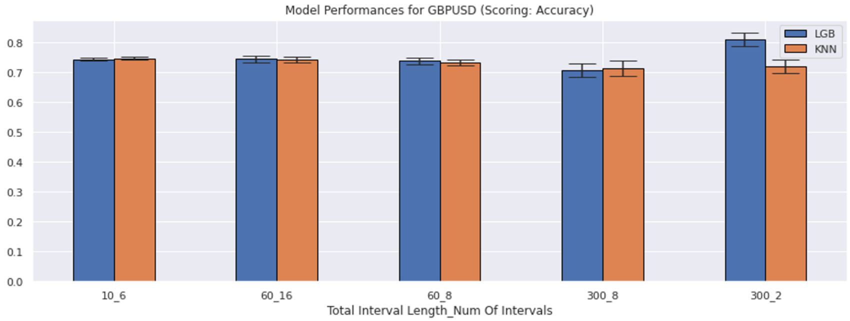
Binary Classification

Sample Statistics: GBPUSD

- x_axis: interval length (sec)_numinterval
- y_axis: Scoring (ROCAUC, F1, ACC)
- Observation across **classifiers**
  - 10_6: Still, stably high
  - 60_16: Tie; KNN no longer prevails
  - 60_8: Tie; LGB no longer dominates
  - 300_8: Tie; pattern preserved
  - 300_2: Still, seeing unstable LGB
- Generally **more** balanced and **accurate**
- F1 again scores best esp. for KNN





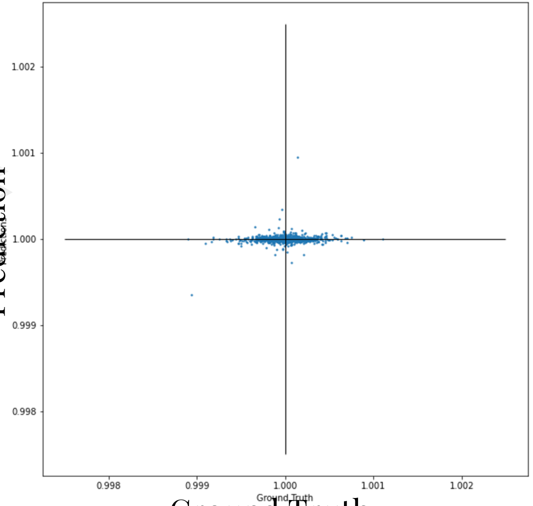


Regression

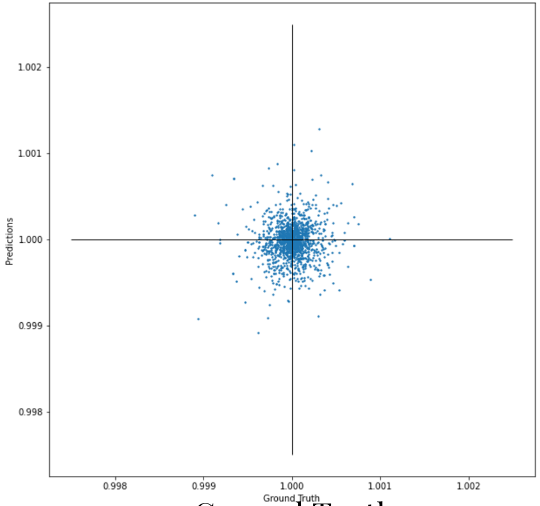
- Using kendaltau as regression matrix
- Sample statistics (fixing interval length= 300 and #interval = 8 for EURUSD)

Regressors: Linear Regression | Kernel Ridge | Light GBM etc.

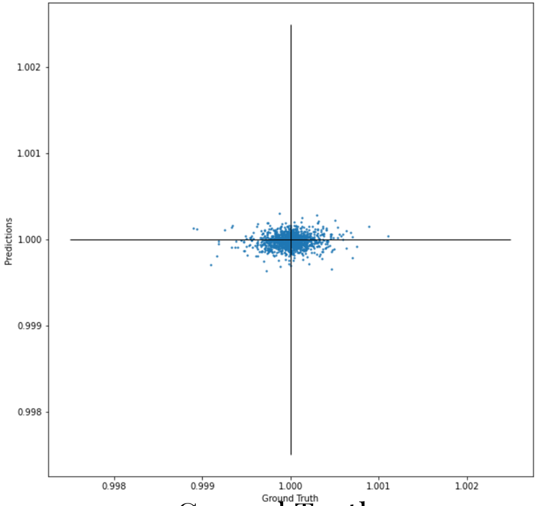
Prediction



Ground Truth



Ground Truth

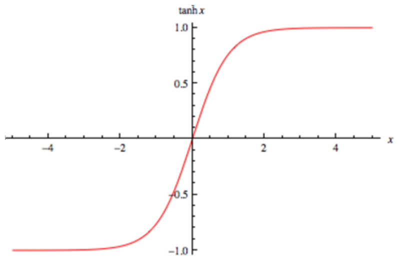


Ground Truth

Trading Profit (with regressor)

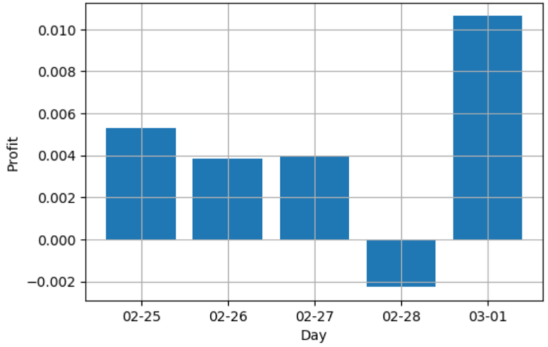
Make prediction based on kernel ridge regression for return value

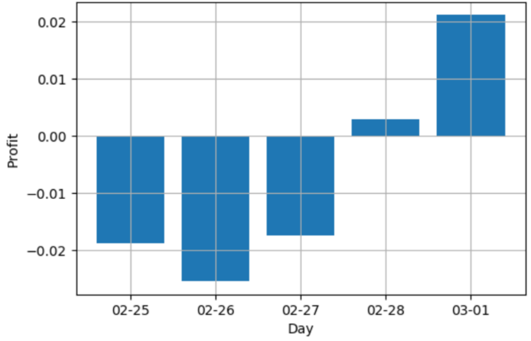
- 5min, look back 2 intervals (10 minutes)
- Take action after every time interval
- Exit immediately at the next close
- Size each trade with minimum “confidence” then with tanh activation

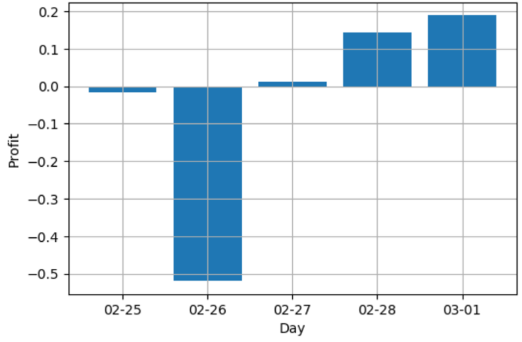


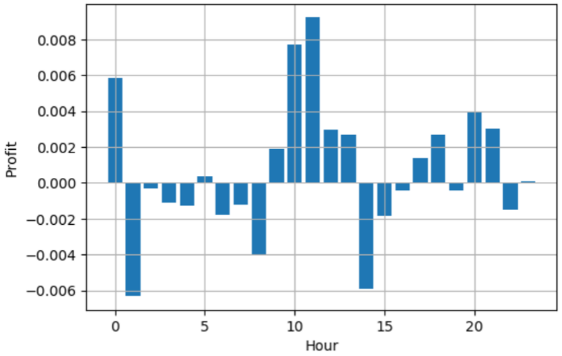
Trading profit on 5 min interval

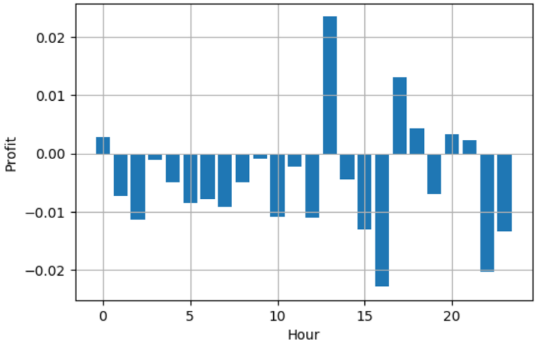
EURUSD	GBPUSD	USDCAD

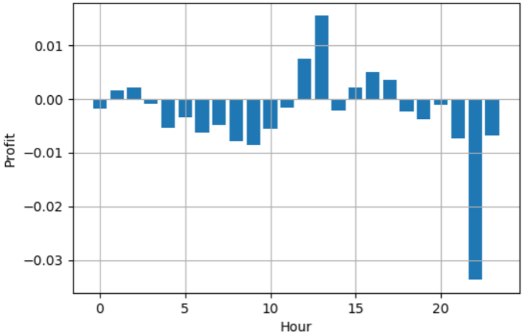












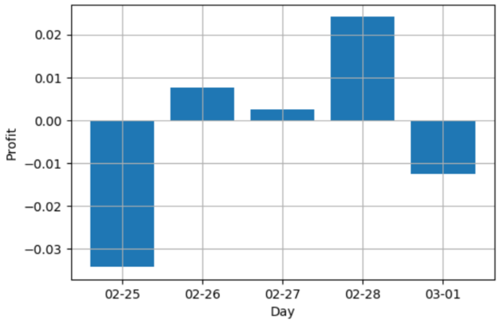
Trading Profit (with classifier)

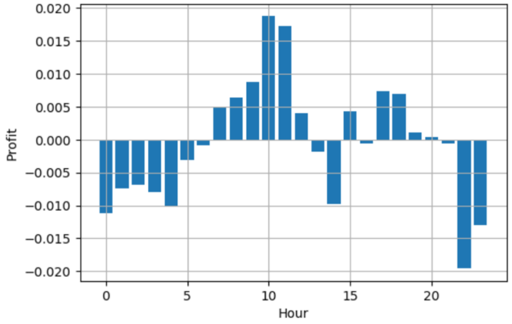
Make prediction based on KNN classifier

- 5min, look back 4 intervals (20 minutes)
- Take action after every time interval
  - Long if score > 0.85
  - Short if score < 0.15
- Exit immediately at the next close
- Size each trade with probability of up/down

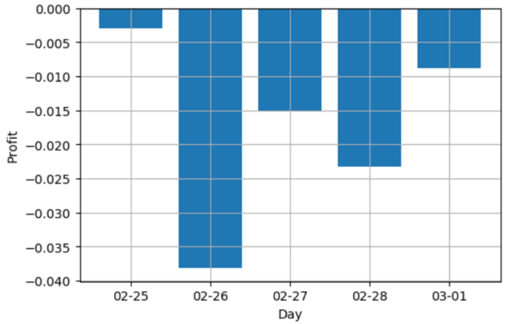
Trading profit on 5 min interval

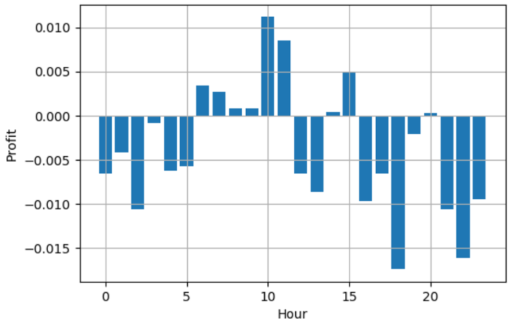
EURUSD



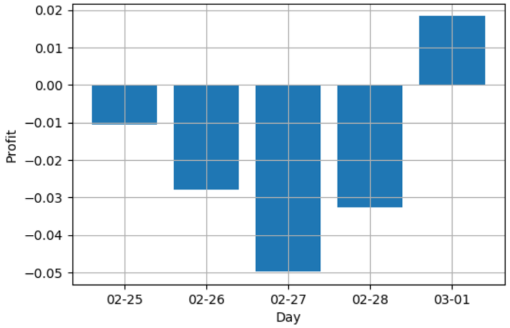


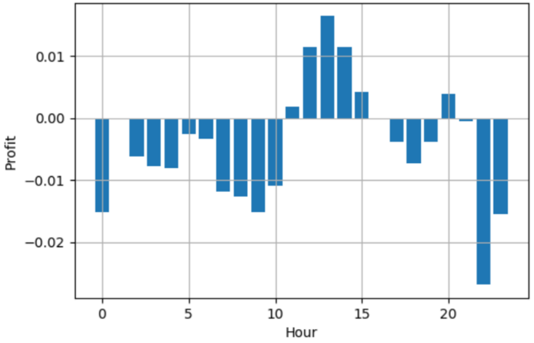
GBPUSD





USDCAD





Observations for ML Models

Can achieve high F1/accuracy but NOT r2/ranking score

- Shorter intervals lead to higher scores but more difficult to act on without incurring many mistakes
  - E.g. Classification model is far less confident, low predicted probabilities
- Regression/classification behave similarly, capture the same signals
- Regressor seems to be more reliable

Need to correlate to actual price movement of the pairs

Next Steps

Achievements so Far

Analyze properties of data from Integral

- Preprocess to synchronize across LP by fixed time intervals

Trading Strategies

- Latency arbitrage
    - Small profit after accounting for network communication time
- Machine learning modeling with lots of hyperparameter search
    - Classification
        - KNN
    - Regression
        - Kernel ridge
    - Difficult to translate to consistent and verifiable profit

Final Report and Beyond

Create model to predict when to execute latency arbitrage

Explore the use of other position sizing strategies

Trade a portfolio of currencies (cross-section) rather than a single one

- Use Modern Portfolio Theory to choose optimal portfolio weights to maximize Sharpe Ratio using estimates of mean return and covariance matrix of returns across pairs

Thanks