JPX Tokyo Stock Exchange Prediction

Data Science Capstone Final Project

Adam Malecek
Yisyuan Lee

Outline

1

Introduction

2

Exploratory Data Analysis

3

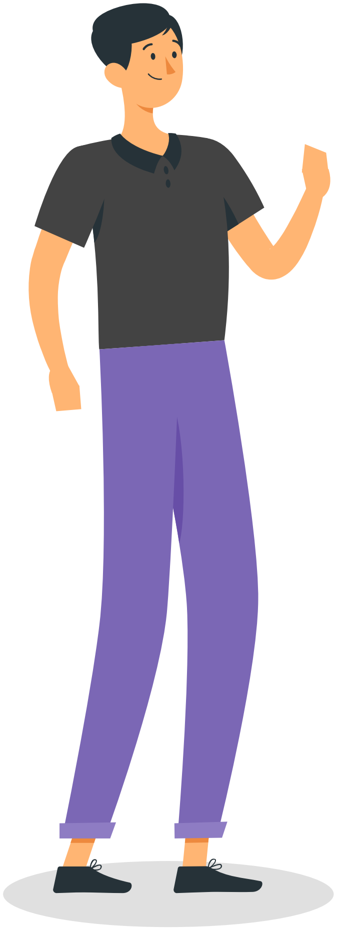
Feature Engineering

4

Model and Tuning

5

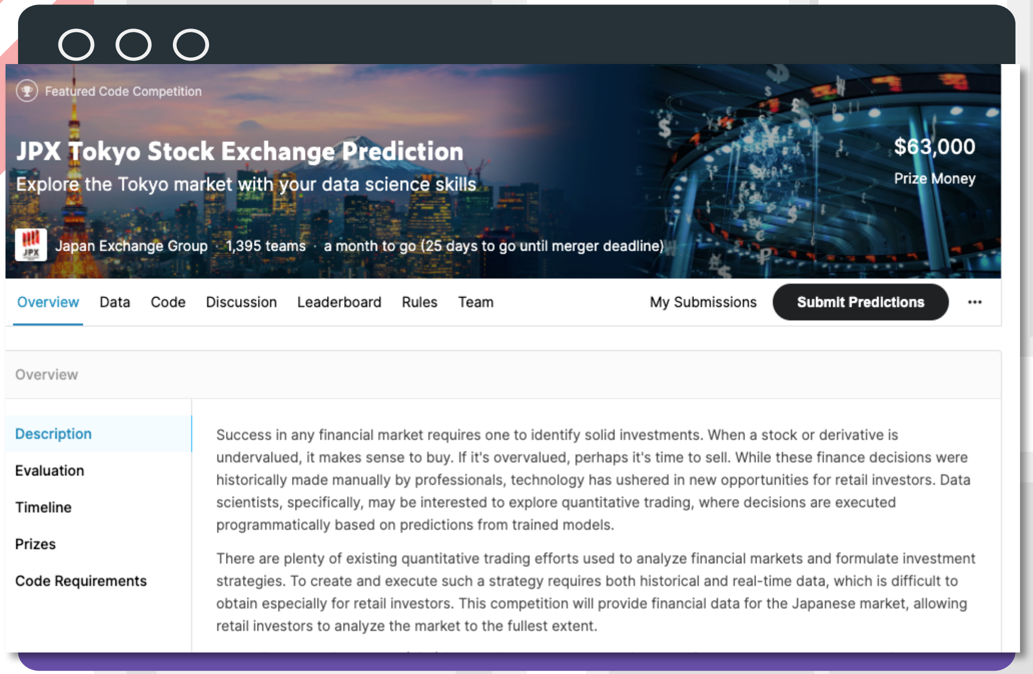
Conclusion



1

Introduction

Competition Introduction



Hosted by Japan Exchange Group, Inc. (JPX)

GOAL Predict future returns of 2000 stocks

Steps:

1. Ranks the stocks from highest to lowest expected returns
2. Evaluate on difference in returns between the top and bottom 200 stocks

Dataset

	RowId	Date	SecuritiesCode	Open	High	Low	Close	Volume	AdjustmentFactor	ExpectedDividend	SupervisionFlag	Target
0	20170104_1301	2017-01-04	1301	2734.0	2755.0	2730.0	2742.0	31400	1.0	NaN	False	0.000730
1	20170104_1332	2017-01-04	1332	568.0	576.0	563.0	571.0	2798500	1.0	NaN	False	0.012324
2	20170104_1333	2017-01-04	1333	3150.0	3210.0	3140.0	3210.0	270800	1.0	NaN	False	0.006154
3	20170104_1376	2017-01-04	1376	1510.0	1550.0	1510.0	1550.0	11300	1.0	NaN	False	0.011053
4	20170104_1377	2017-01-04	1377	3270.0	3350.0	3270.0	3330.0	150800	1.0	NaN	False	0.003026
...	...	...	...	...	...	...	...	...	...	...	...	...
2332526	20211203_9990	2021-12-03	9990	514.0	528.0	513.0	528.0	44200	1.0	NaN	False	0.034816
2332527	20211203_9991	2021-12-03	9991	782.0	794.0	782.0	794.0	35900	1.0	NaN	False	0.025478
2332528	20211203_9993	2021-12-03	9993	1690.0	1690.0	1645.0	1645.0	7200	1.0	NaN	False	-0.004302
2332529	20211203_9994	2021-12-03	9994	2388.0	2396.0	2380.0	2389.0	6500	1.0	NaN	False	0.009098
2332530	20211203_9997	2021-12-03	9997	690.0	711.0	686.0	696.0	381100	1.0	NaN	False	0.018414

2332531 rows × 12 columns

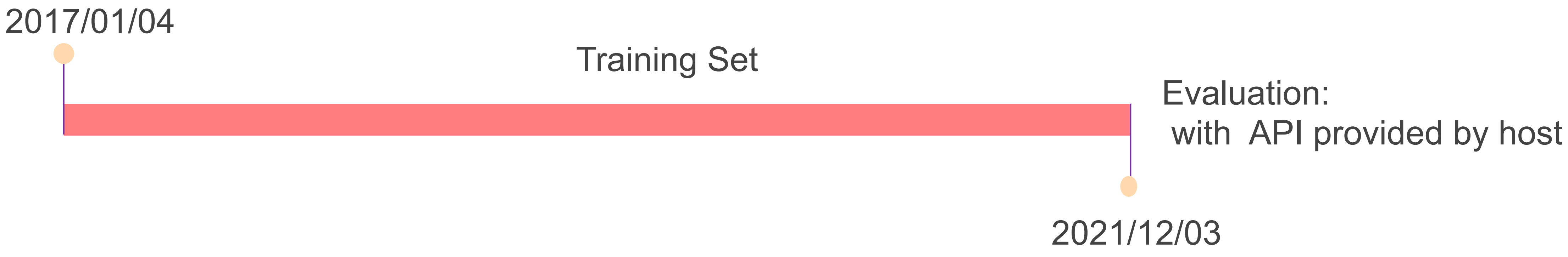
Main Dataset - stock_price.csv

2000 stocks daily stock price

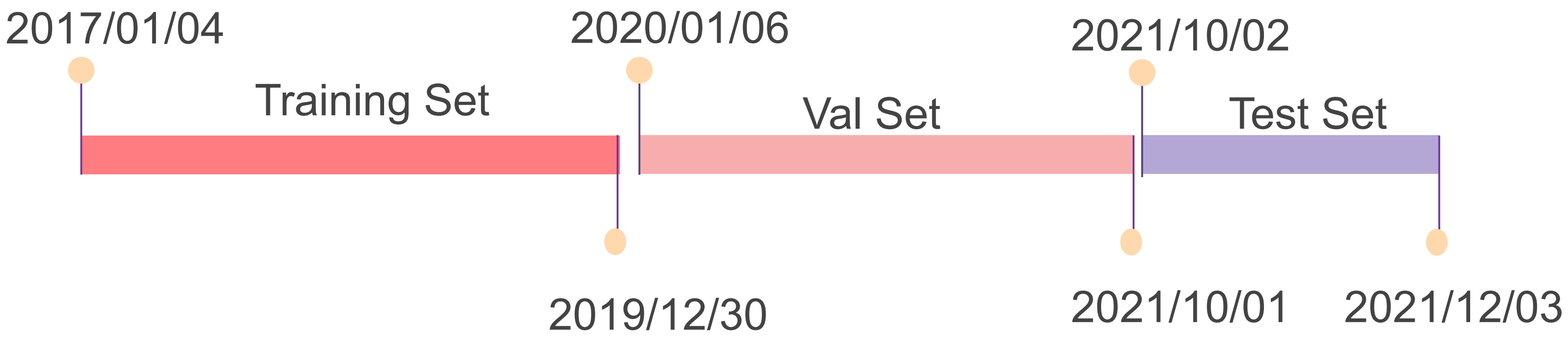
- Row Id
- Date
- Securities Code
- Open
- High
- Low
- Close
- Volume – number of traded stocks on a day
- Adjustment Factor - to calculate theoretical price/volume when split/reverse-split happens
- Expected Dividend - Expected dividend value for ex-right date.
- Supervision Flag - Flag of Securities Under Supervision & Securities to Be Delisted
- Target - Change ratio of adjusted closing price between t+2 and t+1 where t+0 is TradeDate

Train Val Test Split

Provided Dataset



Split with train, validation, test set:



JPX Competition Metric Definition

1. The model will use the price($C_{(k,t)}$) until the business day and other data every business day as input data for a stock(k), and predict rate of change($r_{(k,t)}$) of the top 200 stocks and bottom 200 stocks on the following business day ($C_{(k,t+1)}$) to next following business day ($C_{(k,t+2)}$)

$$r_{k,t} = \frac{C_{k,t+2} - C_{k,t+1}}{C_{k,t+1}}$$

2. With top 200 stock predicted $up_i (i = 1,2, \ldots, 200)$, multiply by their respective rate of change with linear weights of 2-1 for rank 1-200 and denote their sum as S_{up}

$$S_{up} = \frac{\sum_{i=1}^{200} (r_{up_i,t} * linearfuntion(2,1))}{Average(linearfunction(2,1))}$$

3. Within bottom 200 stocks predicted $\mathrm{down}_i (i = 1,2, \ldots, 200)$, multiply by their respective rate of change with linear weights of 2-1 for rank 1-200 and denote their sum as S_{down}

$$S_{down} = \frac{\sum_{i=1}^{200} (r_{down_i,t} * linearfuntion(2,1))}{Average(linearfunction(2,1))}$$

JPX Competition Metric Definition

4. The result of subtracting S_{down} from S_{up} is R_{day} and is called "**daily spread return**" .

$$R_{day} = S_{up} - S_{down}$$

5. The daily spread return is calculated every business day the public/private period and obtained as a time series for that period. The mean/standard deviation of the time series of daily spread returns is used as the score. Score calculation formula (x is the business day of public/private period)

$$Score = \frac{Average(R_{day_1 \sim day_x})}{STD(R_{day_1 \sim day_x})}$$

6. The Kaggler with the largest score for the private period wins.

2 Exploratory Data Analysis

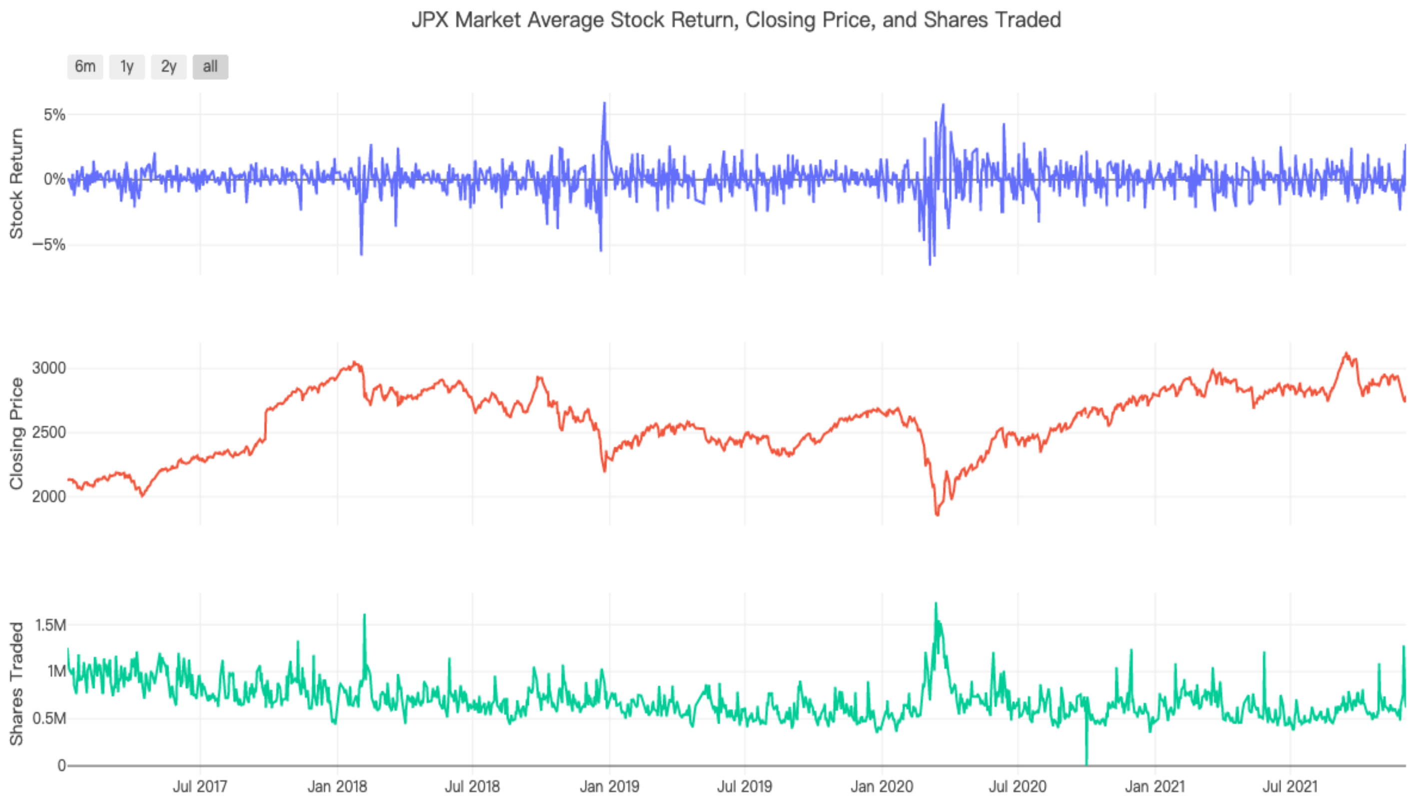
Market's Average Stock Return, Closing Price, Shared Traded

Closing price changes dramatically

Return has large fluctuation

Volume of shared traded increase

⇒ Volume might be important for the forecasting

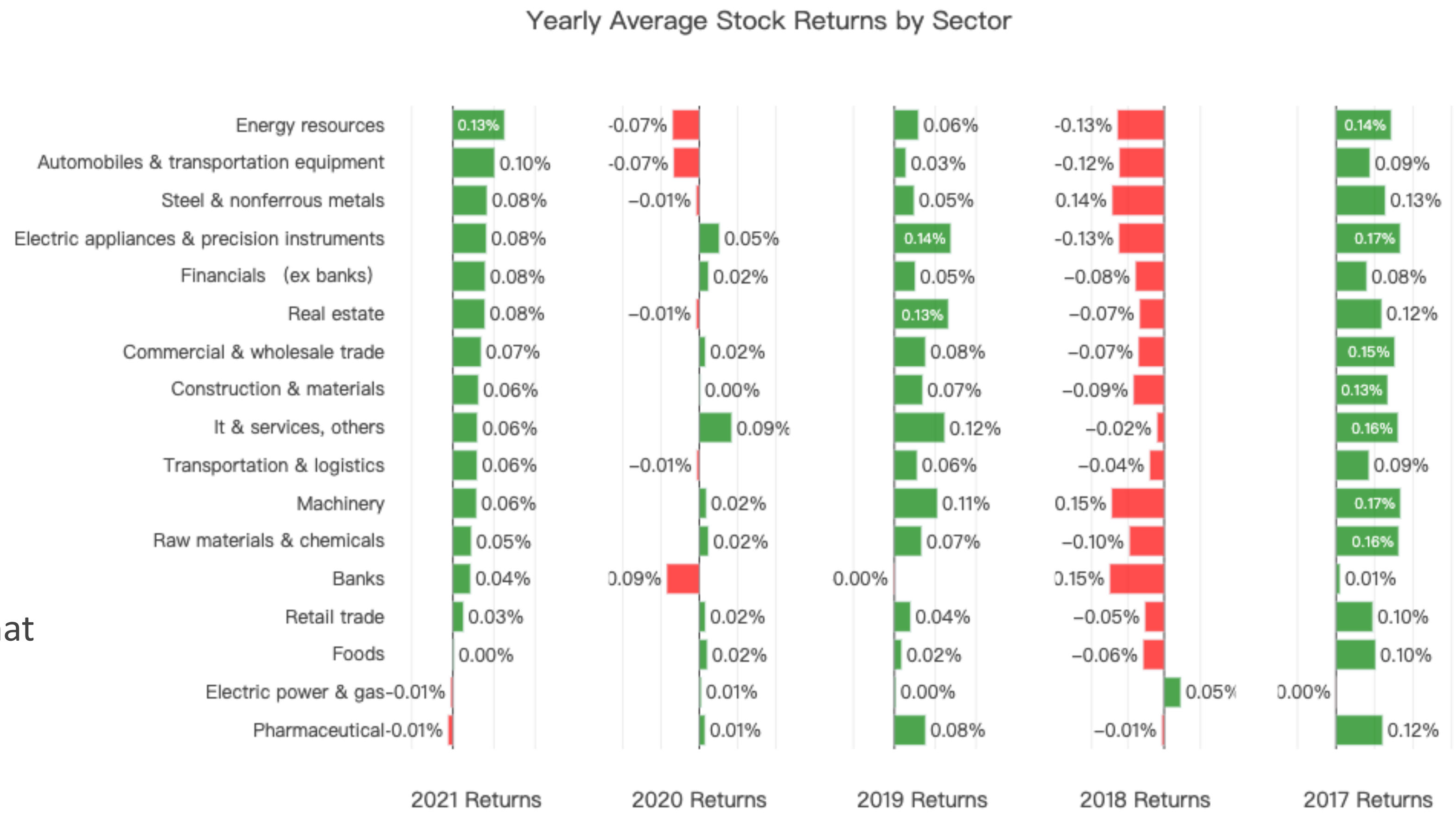


Year Average Stock Return by Sector

- Sector information is from the supplementary files

- 2021 most sectors of stocks have positive return
- 2018 most sectors of stocks have negative return

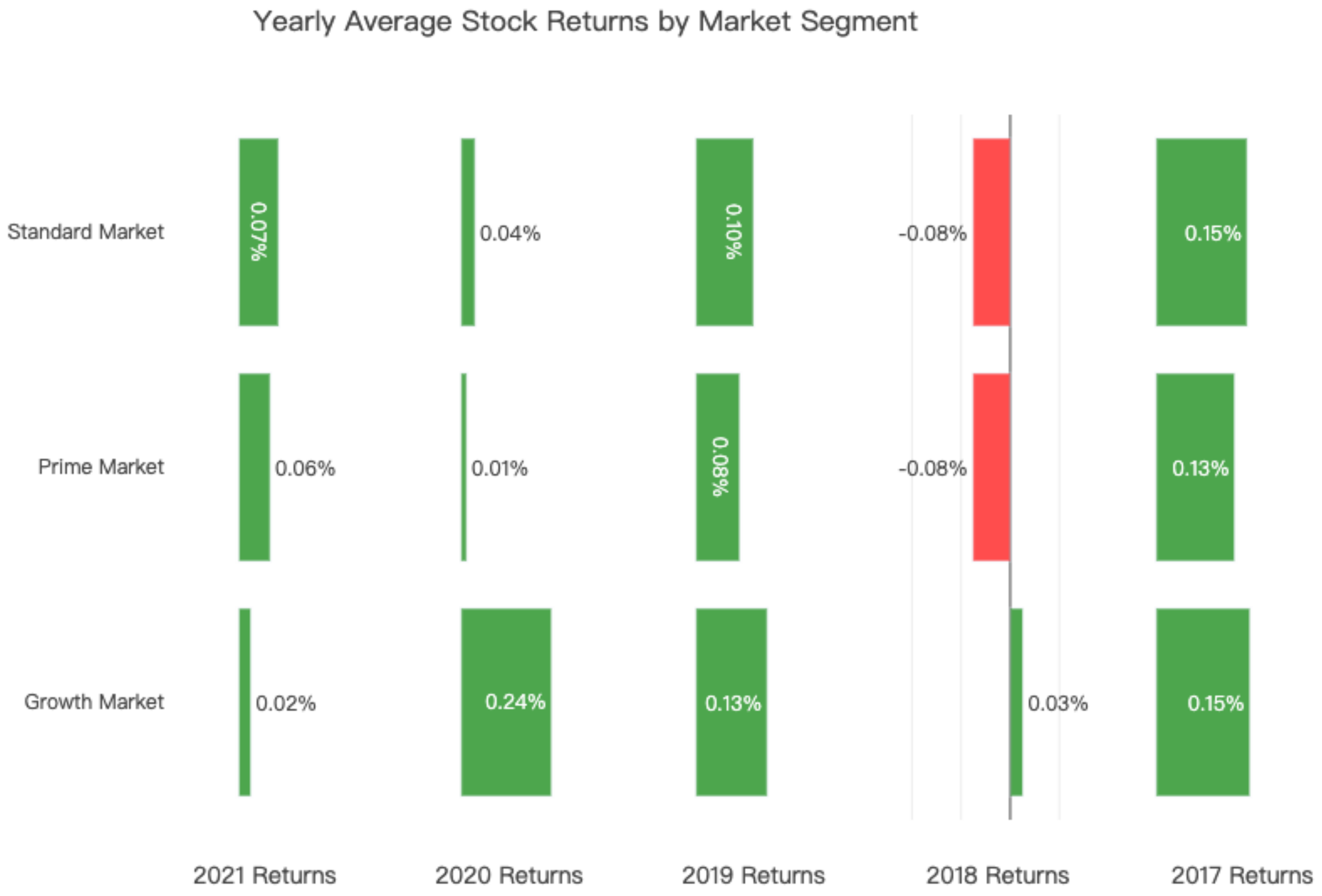
⇒ There is no any sector of stocks that could always obtain high return.



Year Average Stock Return by Segment

- Stocks are trading in different markets
- Segment information is from the supplementary files

- In growth market, average stock return is positive in each year.
- The average return rate of Standard Market is mostly higher than the Prime Market each year.

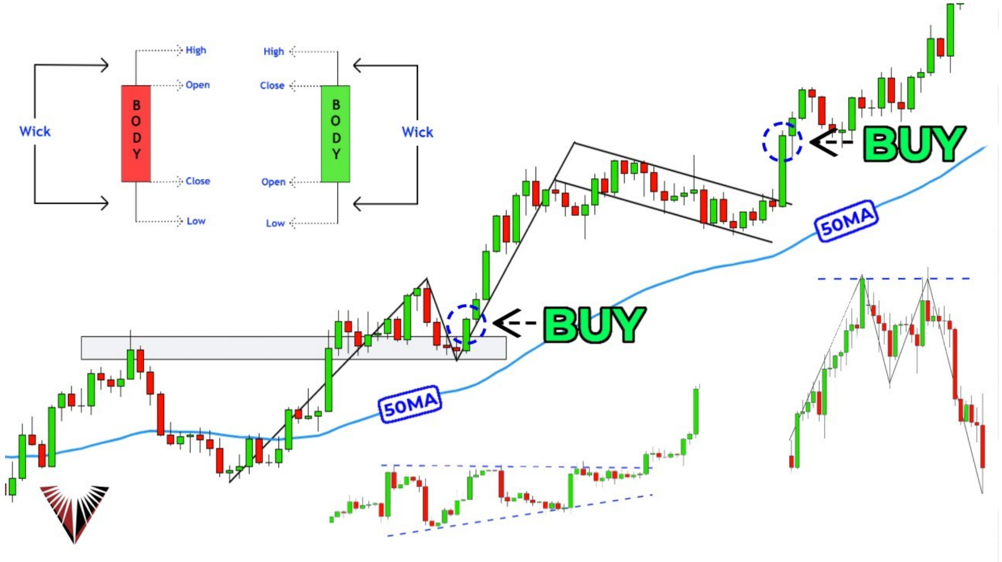


3

Feature Engineering

Feature Engineering with financial data

- Only few features:
  - Open
  - Close
  - Low
  - High
  - Volume
- Lot of expert knowledge gathered
- Technical Analysis



Technical Analysis package

- Open, Close, High, Low, Volume features used to generate new features
- 86 new features:
  - Volume features
    - Force index
  - Volatility features
    - Bollinger Bands
  - Trend features
    - Moving Average Convergence Divergence
  - Momentum features
    - Relative Strength Index

Resulting dataset

- 2000 Stocks
- 1202 Days (not every stock has data on every day) from 2017-01-04 to 2021-12-03.
- 2332531 rows
- 91 features
- Train-Val-Test split:
    - Train: 2017-01-04 – 2019-12-30
    - Val: 2020-01-06 – 2021-10-01
    - Test: 2021-10-02 – 2021-12-03

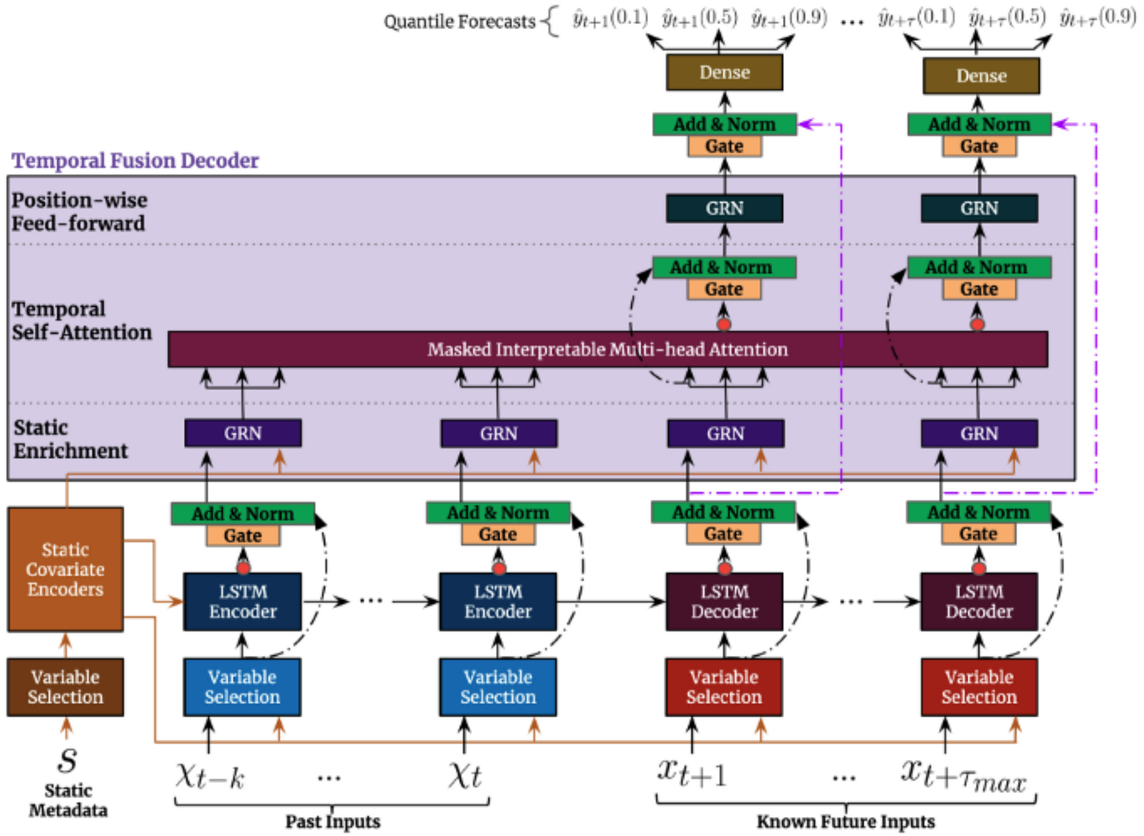
4

Model and Tuning

Method 1
Temporal Fusion Transformer

Temporal Fusion Transformer

- Recurrent layers for local processing
- Self-attention layers for learning long-term dependencies
- Selection of relevant features
- Gating layers to suppress unnecessary components
- Interpretability
- Different types of features:
    - Time varying categorical, Time varying reals
    - Static categorical, Static reals

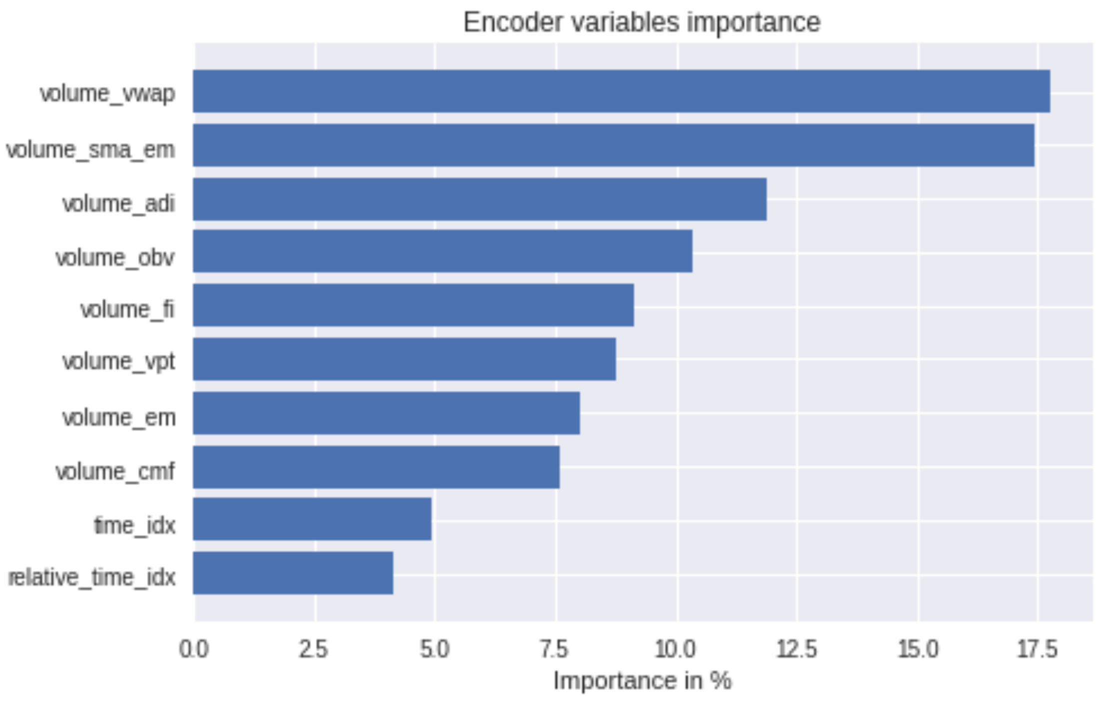


Training

- Model tuning to select the best architecture
- 1.1 M trainable parameters
- Training process:
  - 1 epoch ~1.5 h
  - Early stopping
  - Reduce on plateau

Interpretation

- Few features bare most of importance



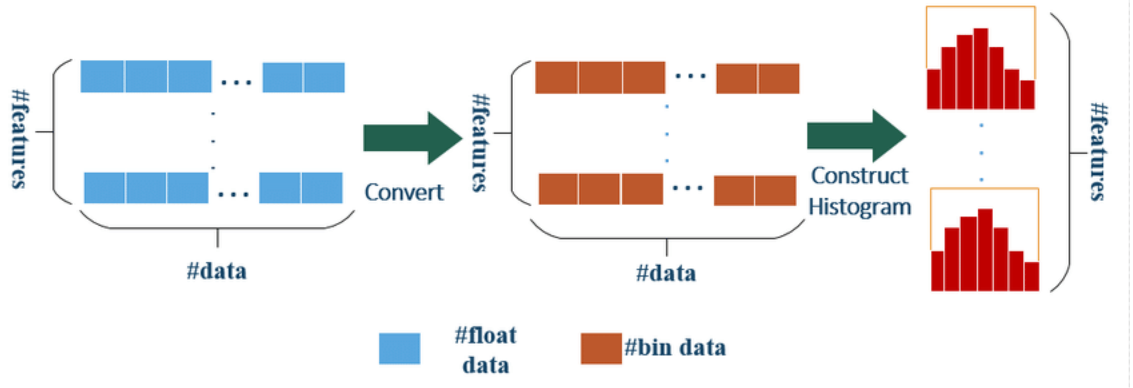
Results

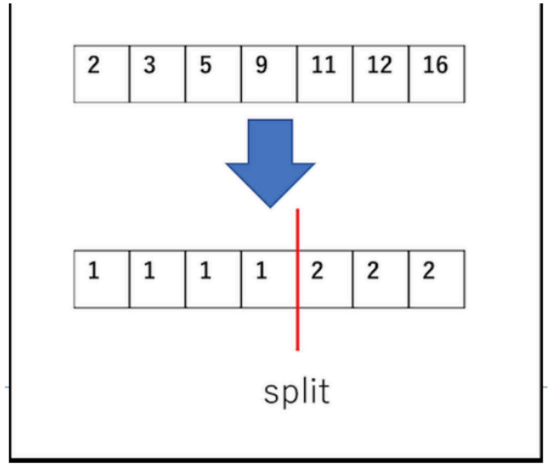
- So far only trained for few epochs
- Predictions close to zero
- Still may work for ranking

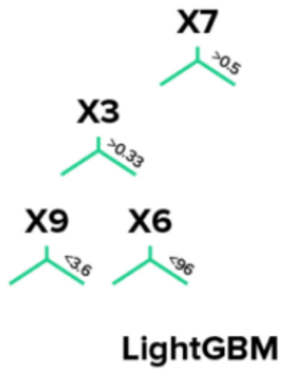
Method 2 – Light GBM

Light Gradient Boosting Machine

- Gradient Boosting Method
- Histogram-based Boosting
- Exclusive Feature Bundling
- Leaf-wise tree
- High Light: Speed







LGBM-Training

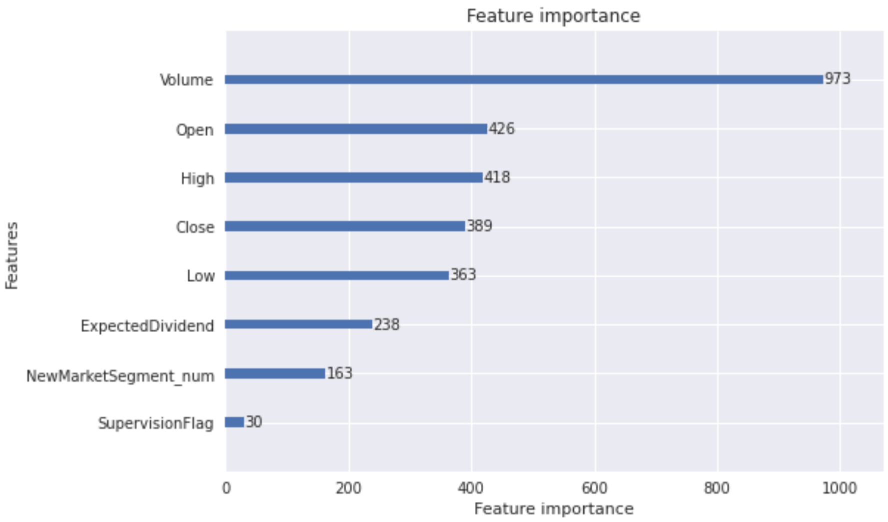
- Manually create features:
  - Log return of the day
  - Segment of the Stocks as (1,2,3) corresponding to each market segment
- Load features from TA packages:
  - Momentum RSI
  - Trend MACD
  - Trend KST

- Tune with original features:

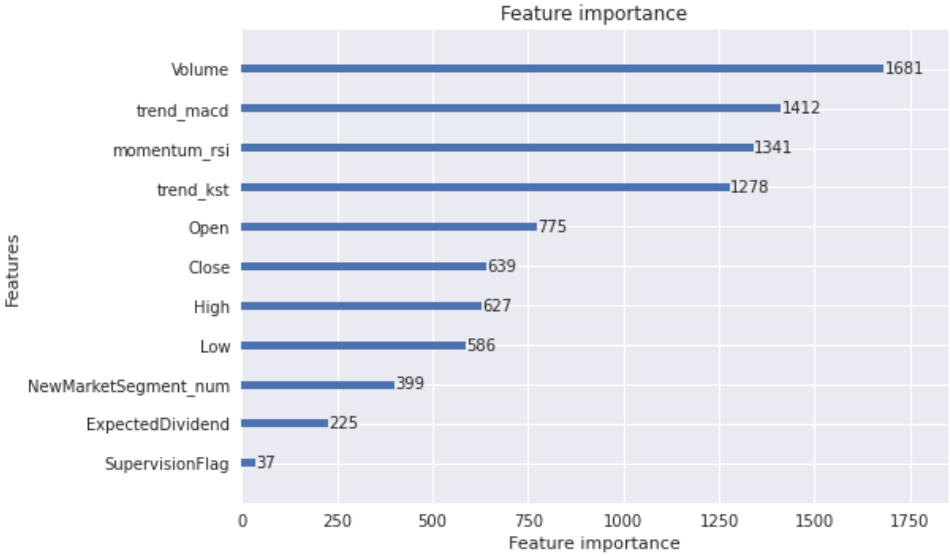
Features:	Securities Code	Open	Close	High	Low	Volume	Adjustment Factor	Expected Dividend	Supervision Flag	return	Segment	Momentum rsi	Trend macd	Trend kst	val	test	Submit
1st try	V	V	V	V	V	V	V	V	V						0.17	0.05	0.25
2nd try		V	V	V	V	V		V	V	V	V				0.14	-0.18	0.33
3rd try			V	V	V	V		V	V		V	V	V	V	0.16	0.16	-0.17
4th try			V	V	V	V		V	V		V				0.19	0.19	0.34

LGBM-Feature Importance

Best Model



Worst Model



- Volume is the most important feature overall
- Segments does influence the model

=> Match our hypothesis in EDA part

- MACD, RSI, KST are important for the models
- => These indicators might not be useful to predict the overall dataset.

5

Conclusion

Results in the Competition

- First place has the score 7.814 so far.
- Our Score is 0.342 so far

Prize Contenders

#	Team	Members	Score	Entries	Last	Code
1	Mohammad Kaif Ahmed		7.814	15	1h	
2	Oh SeungJae		5.427	5	5h	
3	Paulo Pinto		2.999	139	16h	<>

644	**YiSyuan Lee**		0.342	11	6m	<>

Your Best Entry!
Your submission scored , which is not an improvement of your previous score. Keep trying!

Difficulties

- Out of RAM issues happen
- Training session Heavy with Temporal Fusion Transformer method
- Submission with API fail

Conclusion

- The insight we get from the Exploratory Data Analysis matches our results from feature importance plot.
  - Volume and Market Segment are important for the training
- More features are not always better for the model training
  - Load all indicators from TA packages caused some issues for training process and results

Future Work…

- Add more information from the supplementary files
- Feature selection with small part of expanded features

Thank you