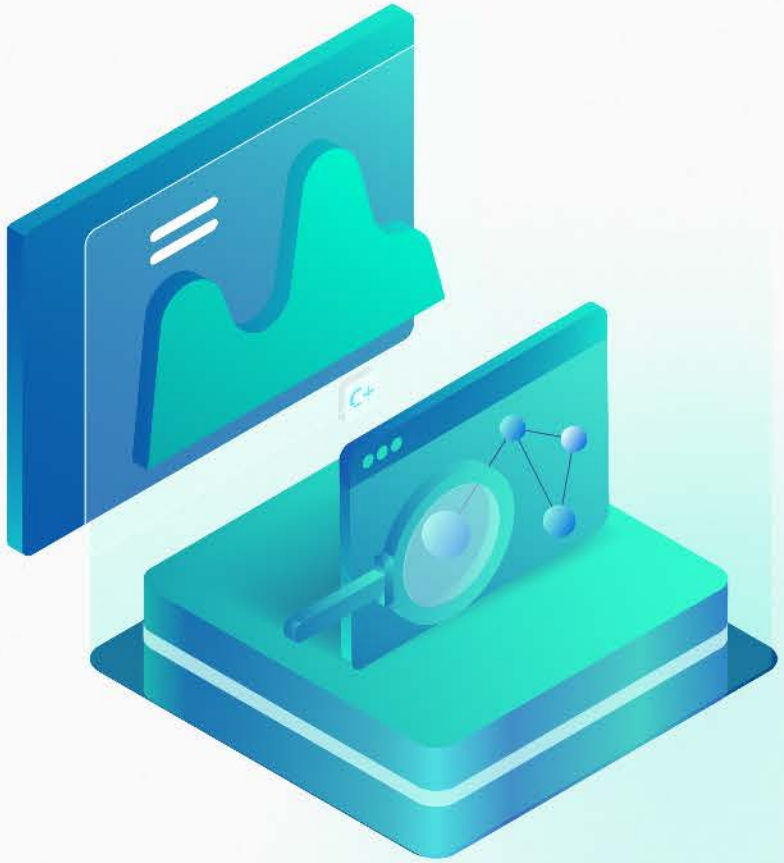
Stock Prediction

With Tweet sentiment and financial data



Group: Fight Potatoes

Wenxin Gu | Siwen Chen | Carolyn Liu

Project Objective

Output: Healthcare Sector Stock Price Prediction

Input:

Financial Data

Generate financial ratios from quarter report

- Profitability
- Liquidity
- Growth
- Credibility

Generate stock trend indicators using stock price

- RSI
- MACD

Tweets Data

Generate sentiment indicators using Covid-related Tweets

- Polarity
- Subjectivity

Phase 1: Data Collection

Goal: Increase the Stock Price Prediction Accuracy

Phase One

	Independent Variables	Dependent variables
Collect Data	• Quarterly data: Financial ratios • Daily data: Tweets sentiment indicators	• Dummy stock price: -1/1 • Stock price
Data processing	Transfer all data into daily data • Transfer the quarterly data by filling the value by the last valid observation	Fill weekend data • Stock Market only has data on weekdays • Fill by taking average
Feature selection	Split into training and testing • Use 2020.3.1 – 2021.12.31 as training • Use 2022.1.1 – 2022.4.24 as testing	Information Gain Method • Select the feature that has an importance larger than 0 using training data
Modeling	Classifier • Ran logistic regression for each of the 24 stocks	SGD Regression • Tried SGD regression to predict the stock price instead of classifying

Outcomes

24 sheets
(per stock)
*****
785 rows
(2020.3.1 – 2022.4.24)
*****
21 features

	G	H	I	J	K	L	M	N	O	P	Q
/ (D	Current Ra	Debt Ratio	Dividends	Dummy_C	Gross Prof	Interest Co	Inventory	Log Reven	Net Acquis	Net Profit	Quick Rati
737	1.533173	0.254655	0.179577	1	0.539137	16.66667	2.013255	9.135769	0.122725	0.141917	1.111538
038	1.639535	0.259131	-0.60215	1	0.532793	12.64706	1.922096	9.13258		0.145173	1.158562
211	1.630334	0.263141	0.196429	1	0.530695	6	1.650667	9.092721		0.081583	1.136247
588	2.469559	0.243348	0.210526	1	0.530531	12.77778	1.690349	9.100715		0.157811	1.741248
263	2.32788	0.24504	0.157143	1	0.531355	16.61111	2.059722	9.171141		0.149697	1.689843
474	2.064612	0.258321	0.299492	1	0.541344	17.26316	2.050331	9.189771		0.186047	1.432128
597	1.998862	0.281977	0.133787	1	0.535738	15.15789	1.927939	9.18327	0.052606	0.141639	1.396473
857	2.106729	0.272424	0.210714	1	0.537201	16	1.938875	9.200303	-9.53E-05	0.166456	1.479118
519	2.224239		0.155263	1	0.542169	19.75	2	9.220108	0	0.266265	1.608314
534	2.193182		0.35	1	0.543608	18.8	1.904437	9.223755		0.169056	1.491793
236	3.17735	0.748785	0.512532	1	0.769531	9	4.800883	9.939719	0.012736	0.321806	2.909978
514	3.142371	0.735162	0.477778	1	0.774684	8.418224	4.674078	9.935457	0.000132	0.349228	2.884099
326	0.862452	0.584712	0.601648	1	0.644029	2.614007	2.568367	10.01808	0.256256	-0.07079	0.58403
577	0.951532	0.581864	0.374131	1	0.608588	5.322581	3.713874	10.11066	0.005882	0.178887	0.676812
109	0.843411	0.571554	0.459235	1	0.662	6.558252	4.186707	10.1417	0.002657	0.002598	0.603643
376	0.831555	0.568302	0.495202	1	0.676172	6.596463	3.976161	10.11428	0.001316	0.273098	0.606084
536	0.906394	0.555092	0.492013	1	0.67598	7.60396	4.122563	10.14485	0.000993	0.054875	0.645447
022	1.01486	0.542268	0.299948	1	0.693906	8.02735	4.635423	10.15661	0.003305	0.221657	0.754565

Phase 1: Data Processing

Goal: Increase the Stock Price Prediction Accuracy

Phase One

	Independent Variables	Dependent variables
Collect Data	• Quarterly data: Financial ratios • Daily data: Tweets sentiment indicators	• Dummy stock price: -1/1 • Stock price
Data processing	Transfer all data into daily data • Transfer the quarterly data by filling the value by the last valid observation	Fill weekend data • Stock Market only has data on weekdays • Fill by taking average
Feature selection	Split into training and testing • Use 2020.3.1 – 2021.12.31 as training • Use 2022.1.1 – 2022.4.24 as testing	Information Gain Method • Select the feature that has an importance larger than 0 using training data
Modeling	Classifier • Ran logistic regression for each of the 24 stocks	SGD Regression • Tried SGD regression to predict the stock price instead of classifying

Outcomes

In total:
24 stocks (sheets) *
785 rows * 21 features

	A	B	C	D	E	F	G	H	I	J	K
1	Ticker	Report Date	Stock price	Asset Turn	CapEx / (D	Current Ra	Debt Ratio	Dividends	Dum	Gross Prof	Interest Co
2	PKI	2020/3/1	86.26	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
3	PKI	2020/3/2	86.38	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
4	PKI	2020/3/3	85.81	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
5	PKI	2020/3/4	88.71	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
6	PKI	2020/3/5	84.20	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
7	PKI	2020/3/6	83.11	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
8	PKI	2020/3/7	81.10	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
9	PKI	2020/3/8	79.08	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
10	PKI	2020/3/9	77.06	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
11	PKI	2020/3/10	78.55	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
12	PKI	2020/3/11	72.49	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
13	PKI	2020/3/12	70.12	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
14	PKI	2020/3/13	73.47	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
15	PKI	2020/3/14	71.11	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
16	PKI	2020/3/15	68.75	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
17	PKI	2020/3/16	66.39	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
18	PKI	2020/3/17	66.79	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
19	PKI	2020/3/18	66.90	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
20	PKI	2020/3/19	66.70	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438
21	PKI	2020/3/20	70.55	0.123192	0.486733	1.80168	0.317197	0.041475	1	0.49433	1.916438

XRAY | ZTS | WST | WAT | TMO | TFX | TECH | STE | PKI | MDT | JNJ | ILMN | IDXX | HC

Phase 1: Data Processing (Continue)

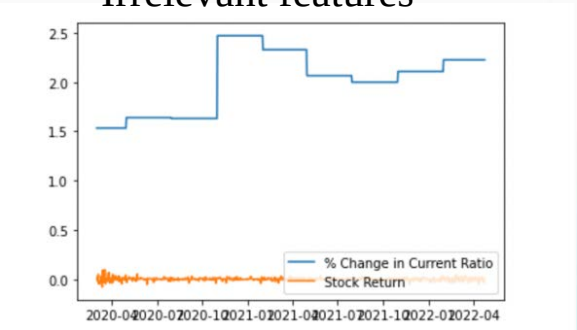
Goal: Increase the Stock Price Prediction Accuracy

Phase One

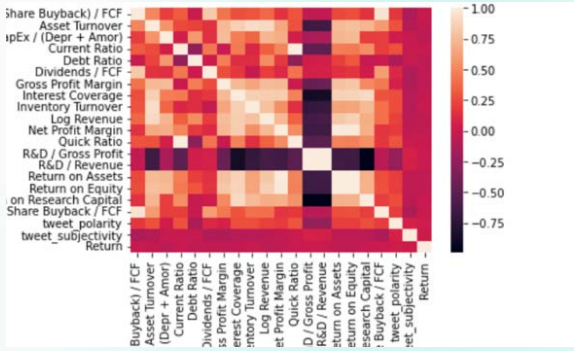
	Independent Variables	Dependent variables
Collect Data	• Quarterly data: Financial ratios • Daily data: Tweets sentiment indicators	• Dummy stock price: -1/1 • Stock price
Data processing	Transfer all data into daily data • Transfer the quarterly data by filling the value by the last valid observation	Fill weekend data • Stock Market only has data on weekdays • Fill by taking average
Feature selection	Split into training and testing • Use 2020.3.1 – 2021.12.31 as training • Use 2022.1.1 – 2022.4.24 as testing	Information Gain Method • Select the feature that has an importance larger than 0 using training data
Modeling	Classifier • Ran logistic regression for each of the 24 stocks	SGD Regression • Tried SGD regression to predict the stock price instead of classifying

Outcomes

Irrelevant features



Redundant features



Phase 1: Feature Selection

Goal: Increase the Stock Price Prediction Accuracy

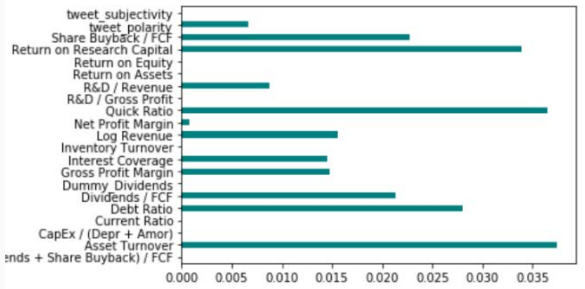
Phase One

	Independent Variables	Dependent variables
Collect Data	• Quarterly data: Financial ratios • Daily data: Tweets sentiment indicators	• Dummy stock price: -1/1 • Stock price
Data processing	Transfer all data into daily data • Transfer the quarterly data by filling the value by the last valid observation	Fill weekend data • Stock Market only has data on weekdays • Fill by taking average
Feature selection	Split into training and testing • Use 2020.3.1 – 2021.12.31 as training • Use 2022.1.1 – 2022.4.24 as testing	Information Gain Method • Select the feature that has an importance larger than 0 using training data
Modeling	Classifier • Ran logistic regression for each of the 24 stocks	SGD Regression • Tried SGD regression to predict the stock price instead of classifying

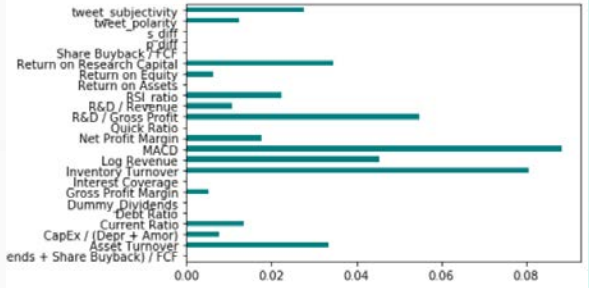
Outcomes

Information Gain Method showing the importance of different features

Example: Stock 1



Example: Stock 2



Phase 1: Modeling

Goal: Increase the Stock Price Prediction Accuracy

Phase One

	Independent Variables	Dependent variables
Collect Data	• Quarterly data: Financial ratios • Daily data: Tweets sentiment indicators	• Dummy stock price: -1/1 • Stock price
Data processing	Transfer all data into daily data • Transfer the quarterly data by filling the value by the last valid observation	Fill weekend data • Stock Market only has data on weekdays • Fill by taking average
Feature selection	Split into training and testing • Use 2020.3.1 – 2021.12.31 as training • Use 2022.1.1 – 2022.4.24 as testing	Information Gain Method • Select the feature that has an importance larger than 0 using training data
Modeling	Classifier • Ran logistic regression for each of the 24 stocks	SGD Regression • Tried SGD regression to predict the stock price instead of classifying

Outcomes

Outcome for the classifier model

```
logit_accuracy_results

{'XRAY': 0.43859649122807015,
 'ZTS': 0.6842105263157895,
 'WST': 0.42105263157894735,
 'WAT': 0.3333333333333333,
 'TMO': 0.4473684210526316,
 'TFX': 0.5701754385964912,
 'TECH': 0.38596491228070173,
 'STE': 0.43859649122807015,
 'PKI': 0.49122807017543857,
 'MDT': 0.47368421052631576,
 'JNJ': 0.5350877192982456,
 'ILMN': 0.5877192982456141,
 'IDXX': 0.35964912280701755,
 'HOLX': 0.45614035087719296,
 'DXCM': 0.5263157894736842,
 'A': 0.37719298245614036,
 'ABBV': 0.6140350877192983,
 'ABT': 0.4473684210526316,
 'AMGN': 0.4473684210526316,
 'BAX': 0.4649122807017544,
 'BSX': 0.4649122807017544,
 'BIO': 0.6666666666666666,
 'COO': 0.5175438596491229,
 'DHR': 0.3508771929824561}
```

Phase 1: Modeling (Continue)

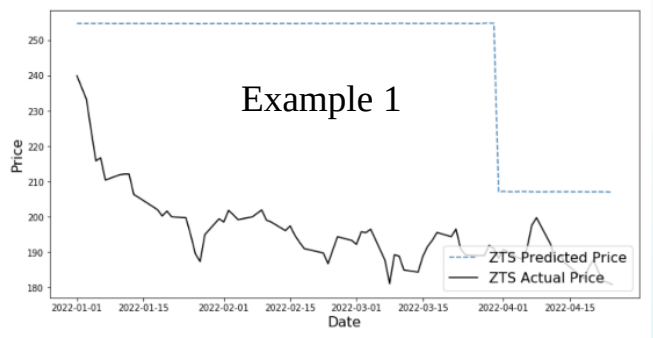
Goal: Increase the Stock Price Prediction Accuracy

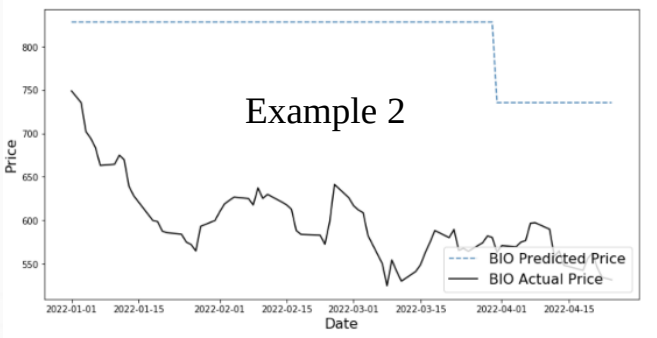
Phase One

	Independent Variables	Dependent variables
Collect Data	• Quarterly data: Financial ratios • Daily data: Tweets sentiment indicators	• Dummy stock price: -1/1 • Stock price
Data processing	Transfer all data into daily data • Transfer the quarterly data by filling the value by the last valid observation	Fill weekend data • Stock Market only has data on weekdays • Fill by taking average
Feature selection	Split into training and testing • Use 2020.3.1 – 2021.12.31 as training • Use 2022.1.1 – 2022.4.24 as testing	Information Gain Method • Select the feature that has an importance larger than 0 using training data
Modeling	Classifier • Ran logistic regression for each of the 24 stocks	SGD Regression • Tried SGD regression to predict the stock price instead of classifying

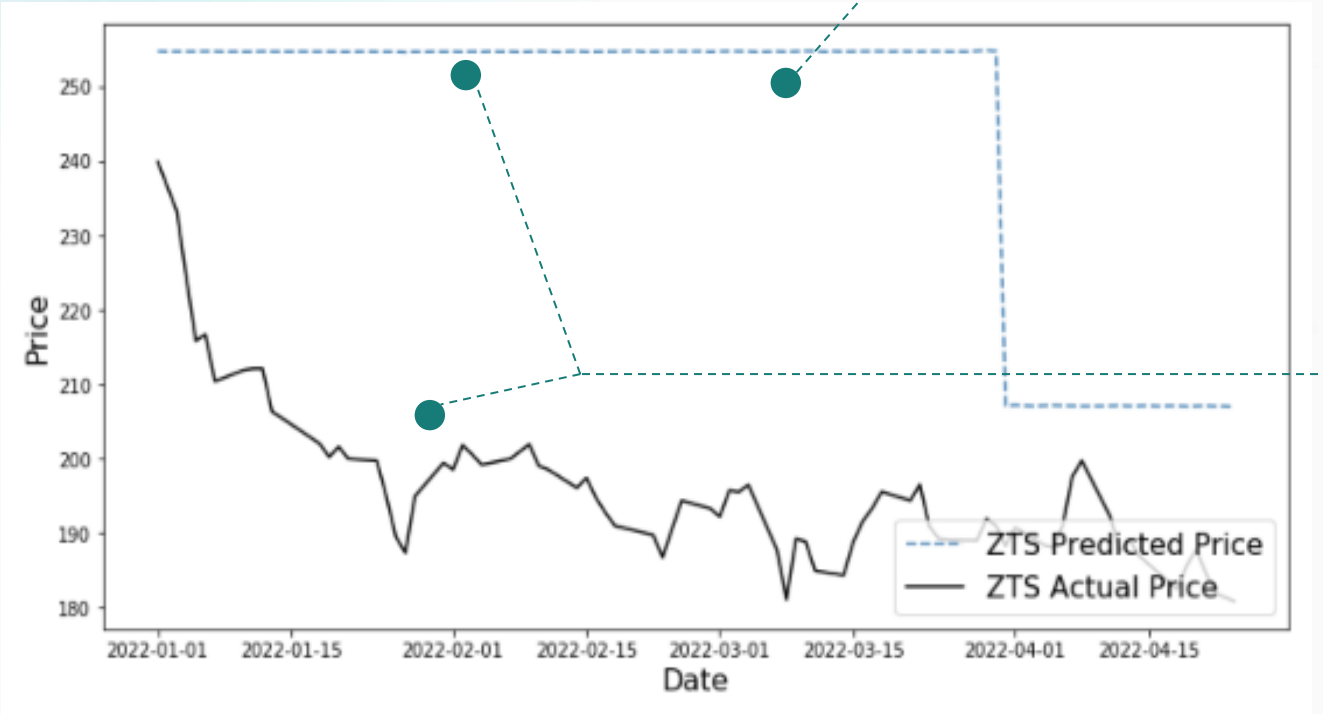
Outcomes

Outcome for the SGD regression





Phase 1: Problems



Fluctuation too small?

- Potential Reason:
  tweets data not so informative?
- Possible Solution:
  - Normalization?
  - Add more daily data?

The predicted value is too high comparing to the actual value

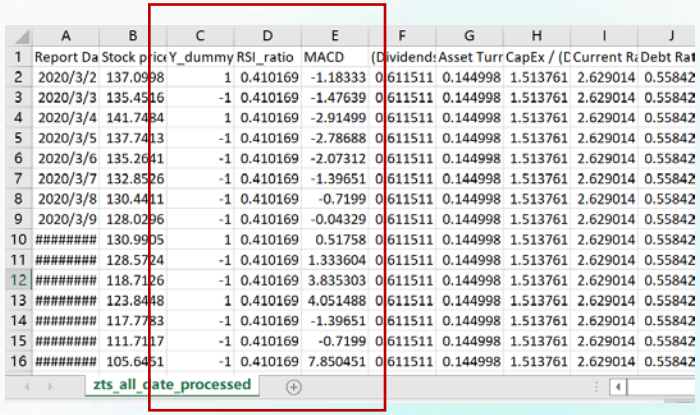
- Potential Reason:
  Using the past two years data is too old for the current situation?
- Possible Solution:
  Splitting the data into smaller windows?

Phase 2: Data collection

Goal: Increase the Stock Price Prediction Accuracy

	Phase Two	
Collect Data	Phase 1	Based on Phase 1, We add:
	• Quarterly data: Financial ratios • Daily data: Tweets sentiment indicators • Dummy stock price: 1, 0	• Stock trend indicator: RSI, MACD • Stock price: 1, 0, -1
Data processing	Phase 1	Based on Phase 1, We add:
	• Transfer all data into daily data • Fill weekend and missing data	• Normalization • Change in variable
Feature selection	Phase 1	Based on Phase 1, We add:
	• Split into training and testing • Select feature using Information Gain Method	• Add Fisher's score to select feature together with information gain method
Modeling	Phase 1	Based on Phase 1, We add:
	• Ran logistic regression for each of the 24 stocks • Tried SGD regression	• Focus on 2 stocks instead of 24 stocks • Divide into 5 smaller windows

Outcomes



	A	B	C	D	E	F	G	H	I	J
1	Report Da	Stock pric	Y_dummy	RSI_ratio	MACD	(Dividend:	Asset Turr	CapEx / (D	Current Ra	Debt Rat
2	2020/3/2	137.0998	1	0.410169	-1.18333	0.611511	0.144998	1.513761	2.629014	0.55842
3	2020/3/3	135.4516	-1	0.410169	-1.47639	0.611511	0.144998	1.513761	2.629014	0.55842
4	2020/3/4	141.7484	1	0.410169	-2.91499	0.611511	0.144998	1.513761	2.629014	0.55842
5	2020/3/5	137.7413	-1	0.410169	-2.78688	0.611511	0.144998	1.513761	2.629014	0.55842
6	2020/3/6	135.2641	-1	0.410169	-2.07312	0.611511	0.144998	1.513761	2.629014	0.55842
7	2020/3/7	132.8526	-1	0.410169	-1.39651	0.611511	0.144998	1.513761	2.629014	0.55842
8	2020/3/8	130.4411	-1	0.410169	-0.7199	0.611511	0.144998	1.513761	2.629014	0.55842
9	2020/3/9	128.0296	-1	0.410169	-0.04329	0.611511	0.144998	1.513761	2.629014	0.55842
10	########	130.9905	1	0.410169	0.51758	0.611511	0.144998	1.513761	2.629014	0.55842
11	########	128.5724	-1	0.410169	1.333604	0.611511	0.144998	1.513761	2.629014	0.55842
12	########	118.7126	-1	0.410169	3.835303	0.611511	0.144998	1.513761	2.629014	0.55842
13	########	123.8448	1	0.410169	4.051488	0.611511	0.144998	1.513761	2.629014	0.55842
14	########	117.7783	-1	0.410169	-1.39651	0.611511	0.144998	1.513761	2.629014	0.55842
15	########	111.7117	-1	0.410169	-0.7199	0.611511	0.144998	1.513761	2.629014	0.55842
16	########	105.6451	-1	0.410169	7.850451	0.611511	0.144998	1.513761	2.629014	0.55842

zts_all_date_processed

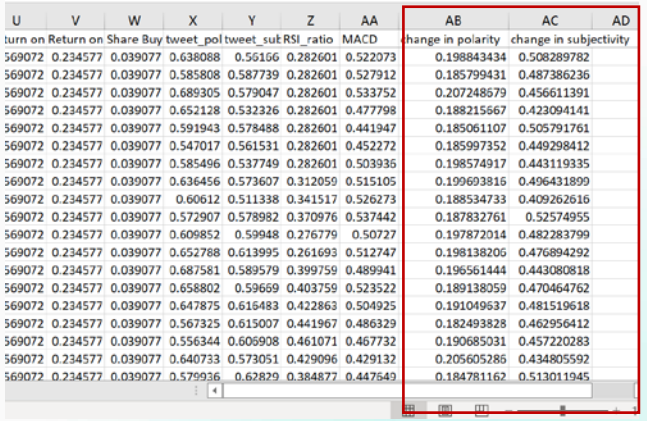
Phase 2: Data Processing

Goal: Increase the Stock Price Prediction Accuracy

Phase Two

	Phase 1	Based on Phase 1, We add:
Collect Data	• Quarterly data: Financial ratios • Daily data: Tweets sentiment indicators • Dummy stock price: 1, 0	• Stock trend indicator: RSI, MACD • Stock price: 1, 0, -1
Data processing	• Transfer all data into daily data • Fill weekend and missing data	• Normalization • Change in variable
Feature selection	• Split into training and testing • Select feature using Information Gain Method	• Add Fisher's score to select feature together with information gain method
Modeling	• Ran logistic regression for each of the 24 stocks • Tried SGD regression	• Focus on 2 stocks instead of 24 stocks • Divide into 5 smaller windows

Outcomes



U	V	W	X	Y	Z	AA	AB	AC	AD
turn on	Return on	Share Buy	tweet_pol	tweet_sub	RSI_ratio	MACD	change in polarity	change in subjectivity	
569072	0.234577	0.039077	0.638088	0.56166	0.282601	0.522073	0.198843434	0.508289782	
569072	0.234577	0.039077	0.585808	0.587739	0.282601	0.527912	0.185799431	0.487386236	
569072	0.234577	0.039077	0.689305	0.579047	0.282601	0.533752	0.207248679	0.456611391	
569072	0.234577	0.039077	0.652128	0.532326	0.282601	0.477798	0.188215667	0.423094141	
569072	0.234577	0.039077	0.591943	0.578488	0.282601	0.441947	0.185061107	0.505791761	
569072	0.234577	0.039077	0.547017	0.561531	0.282601	0.452272	0.185997352	0.449298412	
569072	0.234577	0.039077	0.585496	0.537749	0.282601	0.503936	0.198574917	0.443119335	
569072	0.234577	0.039077	0.636456	0.573607	0.312059	0.515105	0.199693816	0.496431899	
569072	0.234577	0.039077	0.60612	0.511338	0.341517	0.526273	0.188534733	0.409262616	
569072	0.234577	0.039077	0.572907	0.578982	0.370976	0.537442	0.187832761	0.52574955	
569072	0.234577	0.039077	0.609852	0.59948	0.276779	0.50727	0.197872014	0.482283799	
569072	0.234577	0.039077	0.652788	0.613995	0.261693	0.512747	0.198138206	0.476894292	
569072	0.234577	0.039077	0.687581	0.589579	0.399759	0.489941	0.196561444	0.443080818	
569072	0.234577	0.039077	0.658802	0.59659	0.403759	0.523522	0.189138059	0.470464762	
569072	0.234577	0.039077	0.647875	0.616483	0.422863	0.504925	0.191049637	0.481519618	
569072	0.234577	0.039077	0.567325	0.615007	0.441967	0.486329	0.182493828	0.462956412	
569072	0.234577	0.039077	0.556344	0.606908	0.461071	0.467732	0.190685031	0.457220283	
569072	0.234577	0.039077	0.640733	0.573051	0.429096	0.429132	0.205605286	0.434805592	
569072	0.234577	0.039077	0.579936	0.62829	0.384877	0.447649	0.184781162	0.513011945	

Phase 2: Feature Selection

Goal: Increase the Stock Price Prediction Accuracy

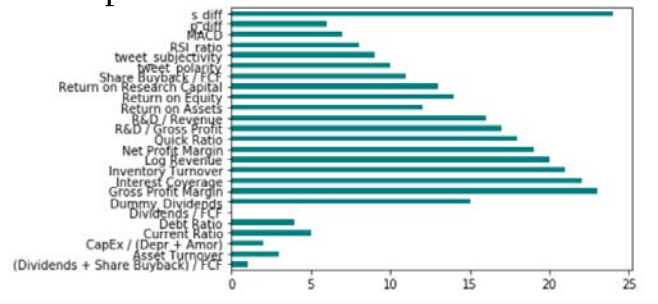
Phase Two

	Phase 1	Based on Phase 1, We add:
Collect Data	• Quarterly data: Financial ratios • Daily data: Tweets sentiment indicators • Dummy stock price: 1, 0	• Stock trend indicator: RSI, MACD • Stock price: 1, 0, -1
Data processing	• Transfer all data into daily data • Fill weekend and missing data	• Normalization • Change in variable
Feature selection	• Split into training and testing • Select feature using Information Gain Method	• Add Fisher's score to select feature together with information gain method
Modeling	• Ran logistic regression for each of the 24 stocks • Tried SGD regression	• Focus on 2 stocks instead of 24 stocks • Divide into 5 smaller windows

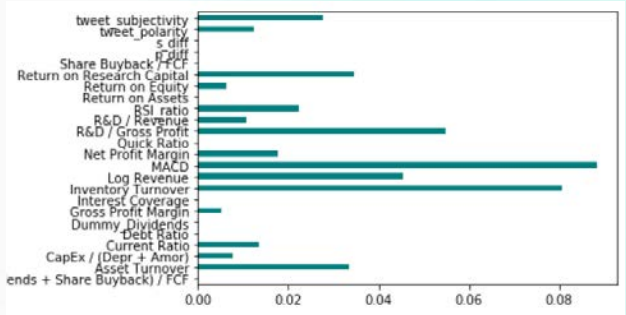
Outcomes

Use both the method to select features

Example: Fisher's score



Example: Information gain



Phase 2: Modeling

Goal: Increase the Stock Price Prediction Accuracy

Phase Two

	Phase 1	Based on Phase 1, We add:
Collect Data	• Quarterly data: Financial ratios • Daily data: Tweets sentiment indicators • Dummy stock price: 1, 0	• Stock trend indicator: RSI, MACD • Stock price: 1, 0, -1
Data processing	• Transfer all data into daily data • Fill weekend and missing data	• Normalization • Change in variable
Feature selection	• Split into training and testing • Select feature using Information Gain Method	• Add Fisher's score to select feature together with information gain method
Modeling	• Ran logistic regression for each of the 24 stocks • Tried SGD regression	• Focus on 2 stocks instead of 24 stocks • Divide into 5 smaller windows • Other models: Decision Tree, SVM, etc.

Outcomes

5 windows:

- bio_test_group0_2021-01-01_2021-06-30
- bio_test_group1_2021-04-01_2021-09-30
- bio_test_group2_2021-07-01_2021-12-31
- bio_test_group3_2021-10-01_2022-04-24
- bio_test_group4_2022-01-01_2022-04-24
- bio_train_group0_2020-03-02_2020-12-31
- bio_train_group1_2020-06-01_2021-03-31
- bio_train_group2_2020-09-01_2021-06-30
- bio_train_group3_2021-01-01_2021-09-30
- bio_train_group4_2021-03-01_2021-12-31
- zts_test_group0_2021-01-01_2021-06-30
- zts_test_group1_2021-04-01_2021-09-30
- zts_test_group2_2021-07-01_2021-12-31
- zts_test_group3_2021-10-01_2022-04-24
- zts_test_group4_2022-01-01_2022-04-24
- zts_train_group0_2020-03-02_2020-12-31
- zts_train_group1_2020-06-01_2021-03-31
- zts_train_group2_2020-09-01_2021-06-30
- zts_train_group3_2021-01-01_2021-09-30
- zts_train_group4_2021-03-01_2021-12-31

Phase 2: Modeling (Continue)

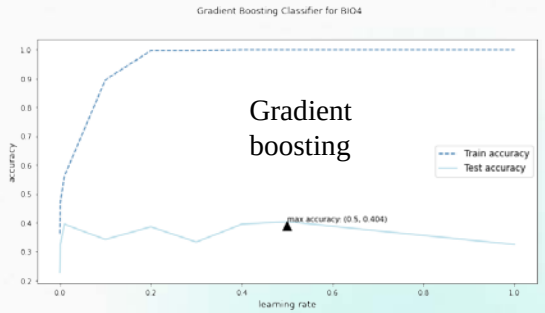
Goal: Increase the Stock Price Prediction Accuracy

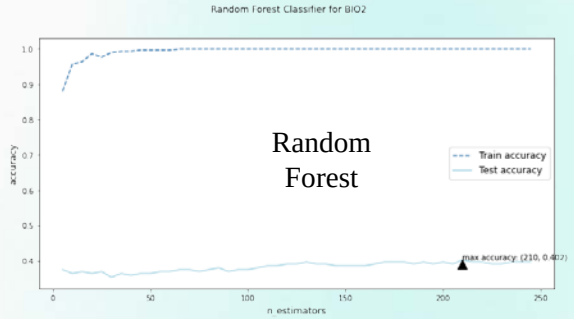
Phase Two

	Phase 1	Based on Phase 1, We add:
Collect Data	• Quarterly data: Financial ratios • Daily data: Tweets sentiment indicators • Dummy stock price: 1, 0	• Stock trend indicator: RSI, MACD • Stock price: 1, 0, -1
Data processing	• Transfer all data into daily data • Fill weekend and missing data	• Normalization • Change in variable
Feature selection	• Split into training and testing • Select feature using Information Gain Method	• Add Fisher's score to select feature together with information gain method
Modeling	• Ran logistic regression for each of the 24 stocks • Tried SGD regression	• Focus on 2 stocks instead of 24 stocks • Divide into 5 smaller windows • 6 other models: Decision Tree, SVM, etc.

Outcomes

A glimpse of different models: accuracy slightly increase





Phase 2: Modeling (Continue)

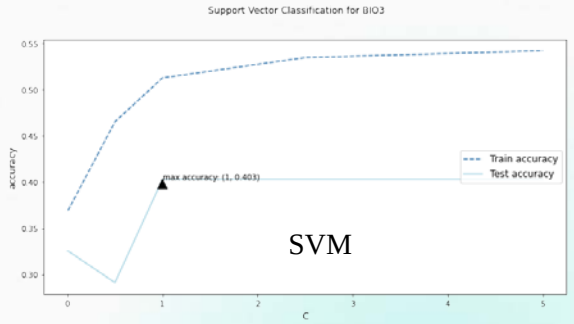
Goal: Increase the Stock Price Prediction Accuracy

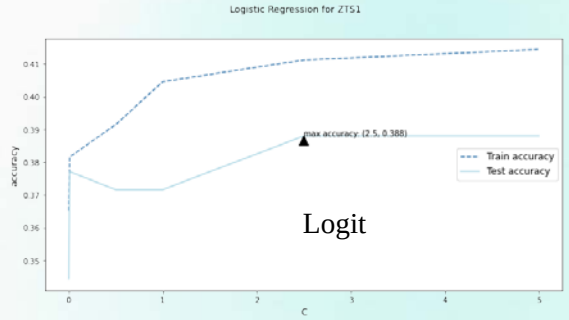
Phase Two

	Phase 1	Based on Phase 1, We add:
Collect Data	• Quarterly data: Financial ratios • Daily data: Tweets sentiment indicators • Dummy stock price: 1, 0	• Stock trend indicator: RSI, MACD • Stock price: 1, 0, -1
Data processing	• Transfer all data into daily data • Fill weekend and missing data	• Normalization • Change in variable
Feature selection	• Split into training and testing • Select feature using Information Gain Method	• Add Fisher's score to select feature together with information gain method
Modeling	• Ran logistic regression for each of the 24 stocks • Tried SGD regression	• Focus on 2 stocks instead of 24 stocks • Divide into 5 smaller windows • 6 other models: Decision Tree, SVM, etc.

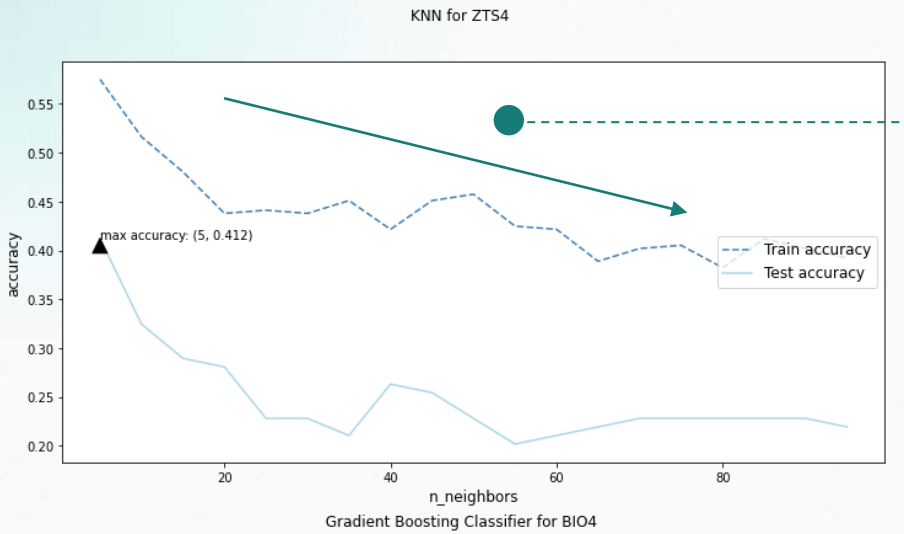
Outcomes

A glimpse of different models: accuracy slightly increase



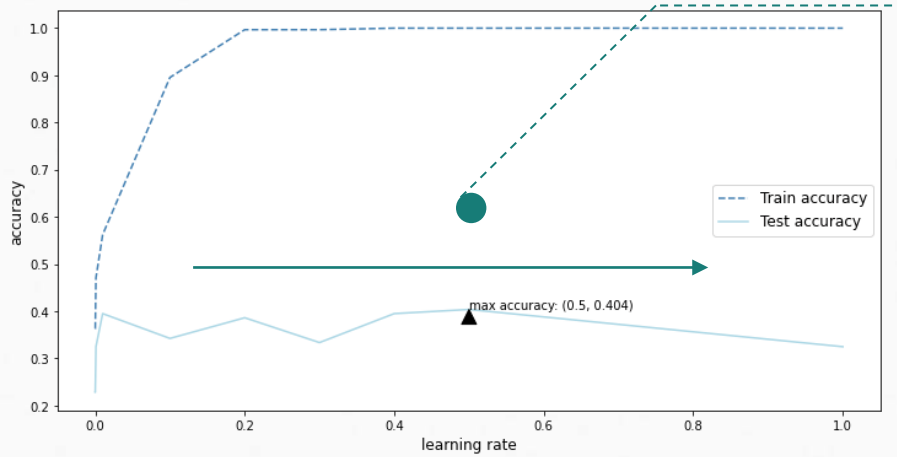


Phase 2: some interesting findings



KNN method accuracy decrease as k increasing

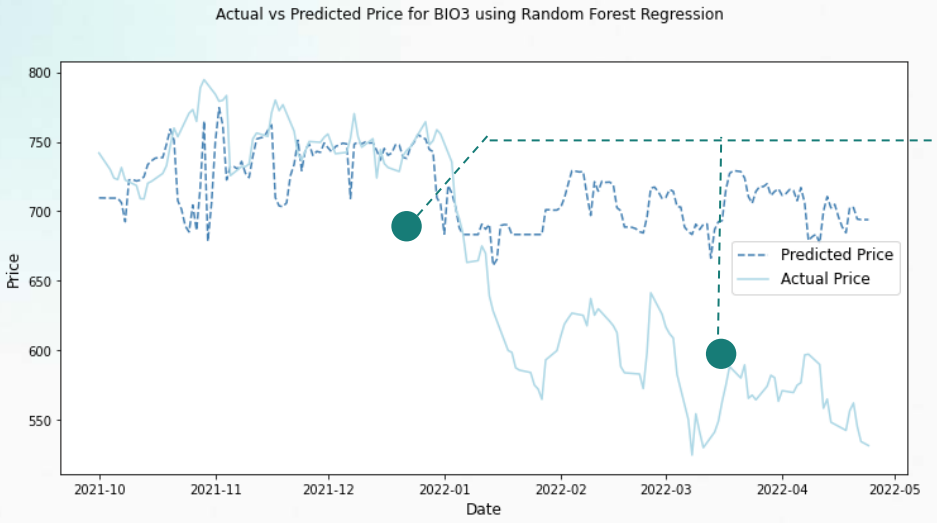
- Potential Reason:
  similar datapoint indicates the similar time period?
  similar time period datapoint has similar trend?
  As k increase, it tends to differ a lot



Gradient Boosting accuracy remain unchanged as not matter what learning rate to choose

- Potential Reason:
  the data itself might not be so informative

Phase 2: Problems



Predict well at first but then worsen

- Potential Reason:
  the window method works?
- Possible Solution:
  use smaller window?

Shall we use the current day information for the current day stock price prediction?

- Possible Solution:
  Shift Y and use previous X to predict current Y?

Only use training data to check accuracy?

- Possible Solution:
  Validation data?

Manually selecting parameters causing not choosing the optimal parameters?

- Possible Solution:
  Use Loop to select the optimal parameters

Phase 3: Data Processing

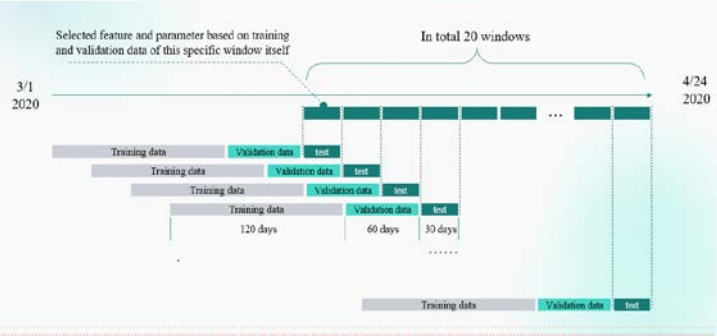
Goal: Increase the Stock Price Prediction Accuracy

Phase Three

	Phase 1 + 2	Based on Phase 1 + 2, We add:
Data Processing	• Transfer all data into daily data • Fill weekend and missing data • Normalization • Change in variable	• 20 windows are applied to use the closer date to predict the data
Feature selection	• Split into training and testing • Select feature using Information Gain Method • Add Fisher's score to select feature together with information gain method	• Use validation data to select feature instead of training data • Loop to select optimal bar for feature selection
Modeling	• Focus on 2 stocks instead of 24 stocks • Divide into 5 smaller windows • tried 7 models: Decision Tree, SVM, Random Forest, etc.	• Loop to select optimal number for model parameter • Shift Y for one day, use the Xs of day t-1 to predict Y of day t

Outcomes

Window illustration



Phase 3: Data Processing

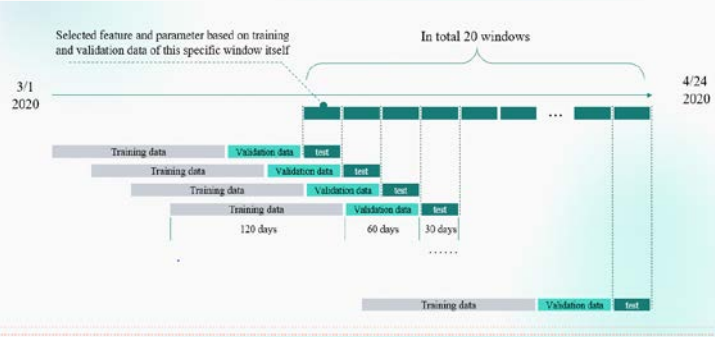
Goal: Increase the Stock Price Prediction Accuracy

Phase Three

	Phase 1 + 2	Based on Phase 1 + 2, We add:
Data Processing	• Transfer all data into daily data • Fill weekend and missing data • Normalization • Change in variable	• 20 windows are applied to use the closer date to predict the data
Feature selection	• Split into training and testing • Select feature using Information Gain Method • Add Fisher's score to select feature together with information gain method	• Use validation data to select feature instead of training data • Loop to select optimal bar for feature selection
Modeling	• Focus on 2 stocks instead of 24 stocks • Divide into 5 smaller windows • tried 7 models: Decision Tree, SVM, Random Forest, etc.	• Loop to select optimal number for model parameter • Shift Y for one day, use the Xs of day t-1 to predict Y of day t

Outcomes

Window illustration



Phase 3: Data Processing

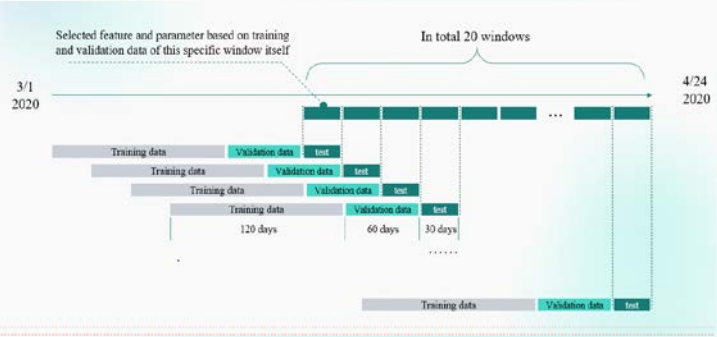
Goal: Increase the Stock Price Prediction Accuracy

Phase Three

	Phase 1 + 2	Based on Phase 1 + 2, We add:
Data Processing	• Transfer all data into daily data • Fill weekend and missing data • Normalization • Change in variable	• 20 windows are applied to use the closer date to predict the data
Feature selection	• Split into training and testing • Select feature using Information Gain Method • Add Fisher's score to select feature together with information gain method	• Use validation data to select feature instead of training data • Loop to select optimal bar for feature selection
Modeling	• Focus on 2 stocks instead of 24 stocks • Divide into 5 smaller windows • tried 7 models: Decision Tree, SVM, Random Forest, etc.	• Loop to select optimal number for model parameter • Shift Y for one day, use the Xs of day t-1 to predict Y of day t

Outcomes

Window illustration



Phase 3: Data Processing

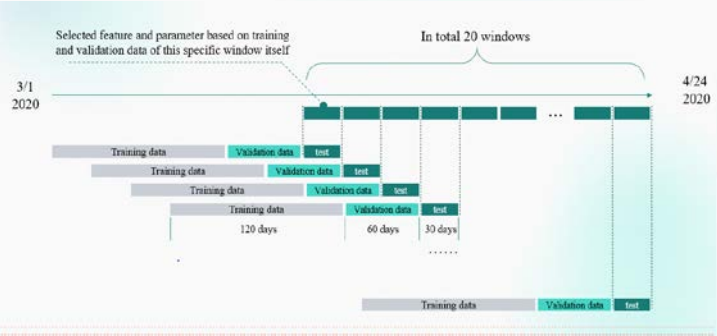
Goal: Increase the Stock Price Prediction Accuracy

Phase Three

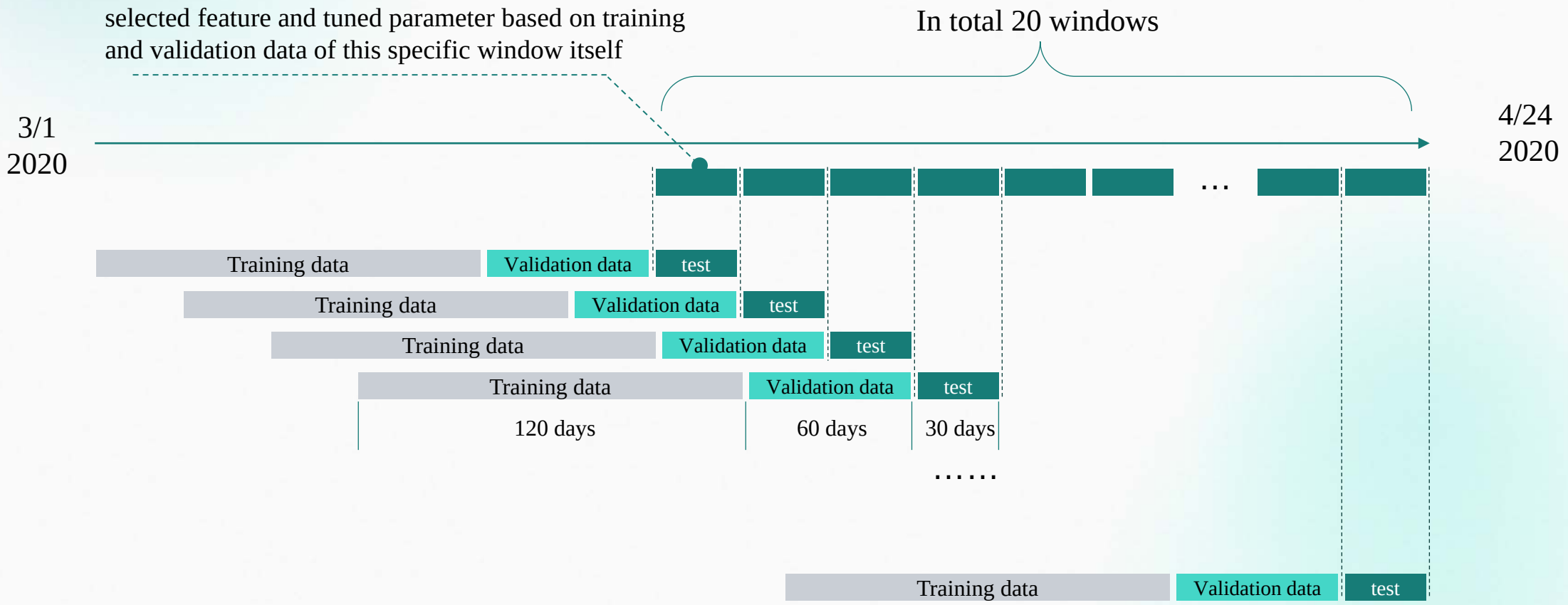
	Phase 1 + 2	Based on Phase 1 + 2, We add:
Data Processing	• Transfer all data into daily data • Fill weekend and missing data • Normalization • Change in variable	• 20 windows are applied to use the closer date to predict the data
Feature selection	• Split into training and testing • Select feature using Information Gain Method • Add Fisher's score to select feature together with information gain method	• Use validation data to select feature instead of training data • Loop to select optimal bar for feature selection
Modeling	• Focus on 2 stocks instead of 24 stocks • Divide into 5 smaller windows • tried 7 models: Decision Tree, SVM, Random Forest, etc.	• Loop to select optimal number for model parameter • Shift Y for one day, use the Xs of day t-1 to predict Y of day t

Outcomes

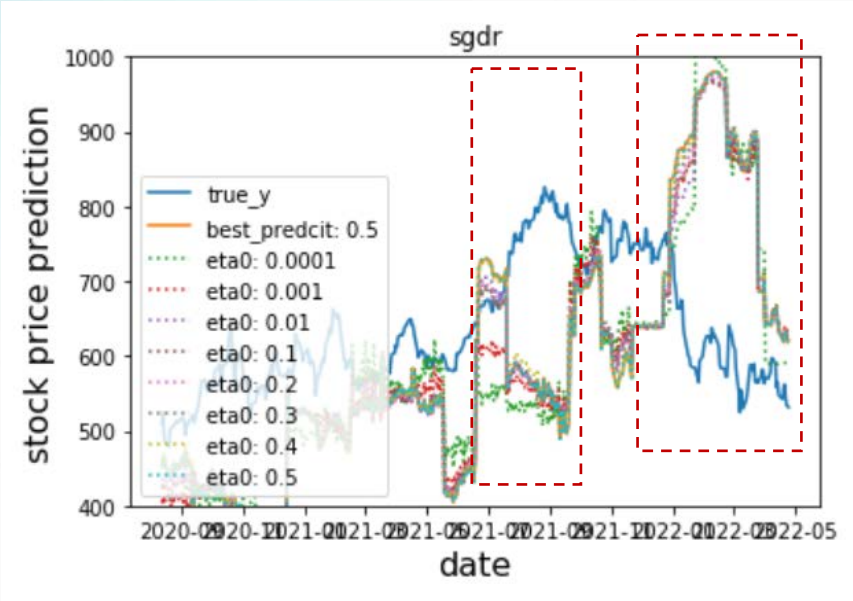
Window illustration

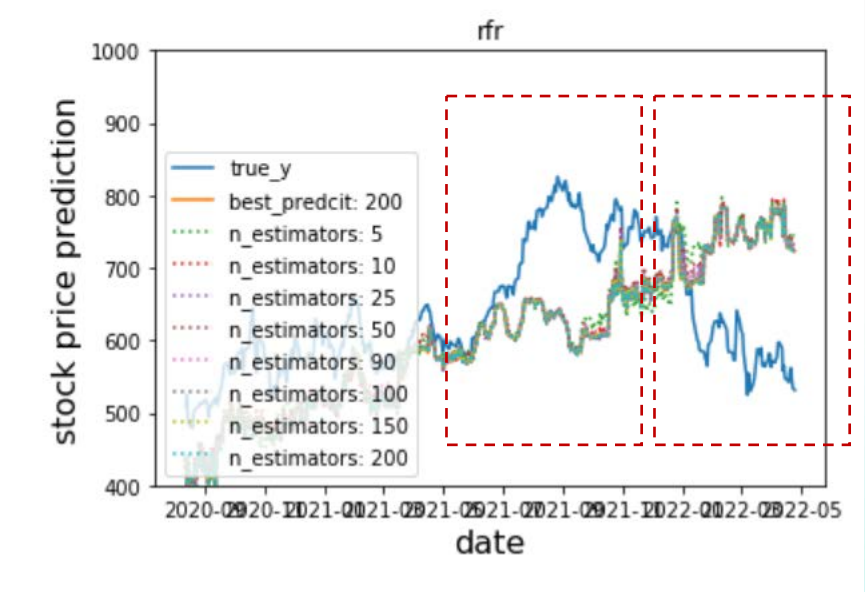


Phase 3: Modeling



Phase 3: a better result

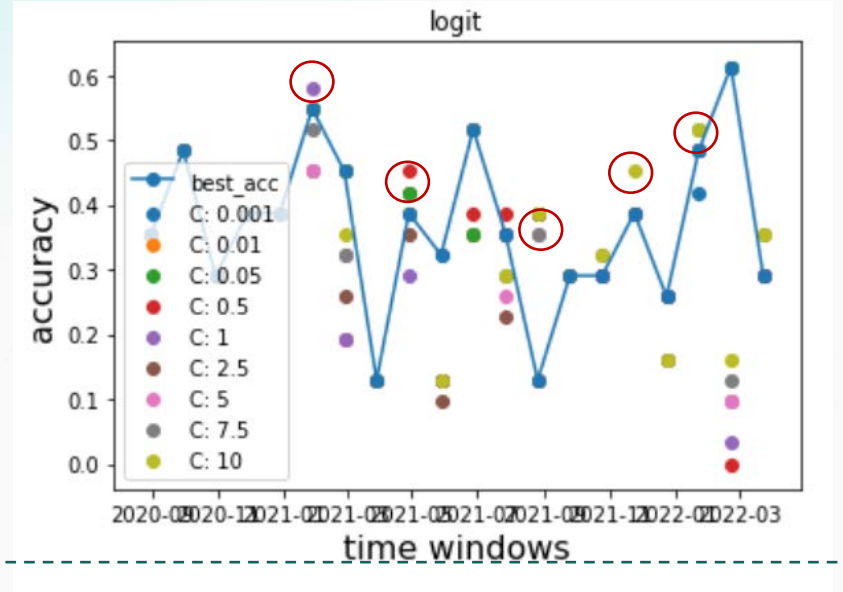




Did capture the smaller trend, but did not capture the larger trend

- Potential Reason:
  the factor capturing the general trend should have higher weight?
- Potential Solution:
  Ensemble methods?

Phase 3: some interesting findings



Validation accuracy

Test accuracy

```
other: 0.39344262295081966 test acc: 0.41935483870967744
current best:  0.39344262295081966 test acc: 0.41935483870967744
other: 0.39344262295081966 test acc: 0.41935483870967744
other: 0.39344262295081966 test acc: 0.41935483870967744
other: 0.39344262295081966 test acc: 0.41935483870967744
other: 0.4426229508196721 test acc: 0.41935483870967744
current best:  0.4426229508196721 test acc: 0.41935483870967744
other: 0.4426229508196721 test acc: 0.3548387096774194
other: 0.4262295081967213 test acc: 0.45161290322580644
other: 0.4098360655737705 test acc: 0.5161290322580645
other: 0.4098360655737705 test acc: 0.5806451612903226
```

VS

The parameter having higher validation accuracy might not always mean a better accuracy in testing

Summary: What do we learned?

- Real-life accuracy is not always as expected.
- Be careful with data.
- Ethical issues