Detecting Market Manipulation in Small – Cap Equities

- Zhihao Ai
- Tolulope Bukola
- Yang Liu
- Zhi Qi
- Bowen Zhou

Fidelity INVESTMENTS

What is Market Manipulation?

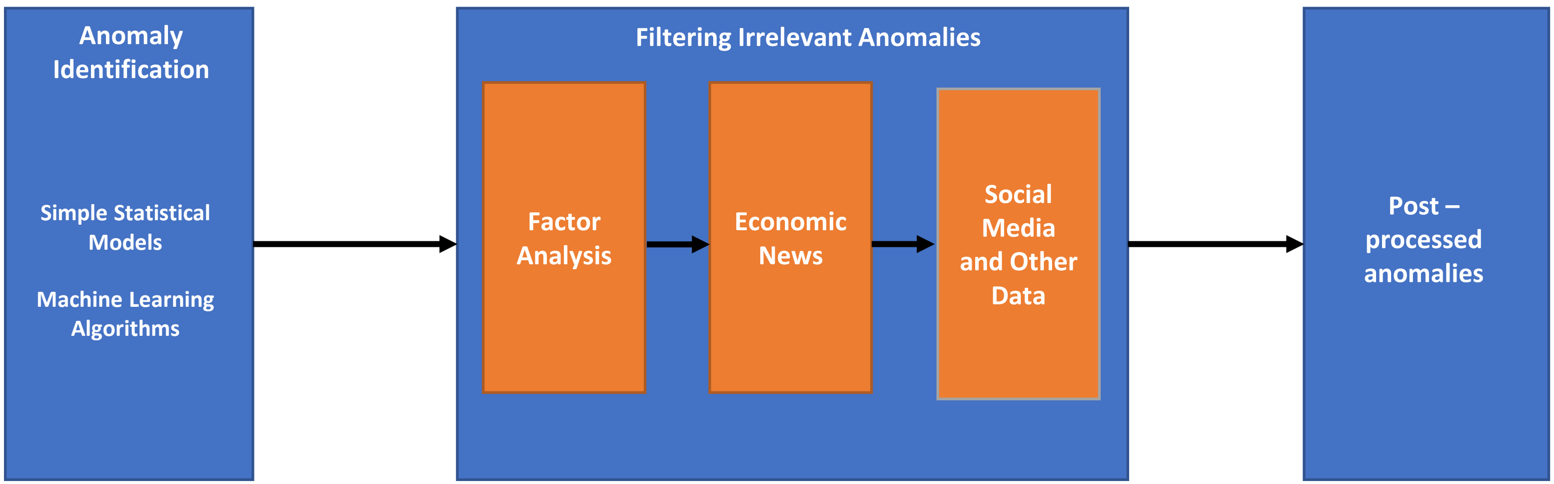
- U.S. Securities and Exchange Commission: "Transactions which create or maintain an artificial price for a trade-able security." An "artificial price" is any price different from one which would prevail in a free market.

- **The project focuses on a particular type of market manipulation: the "pump-and-dump".** A pump-and-dump is a form of market manipulation that typically has three steps:
  1. Acquire shares in a company
  2. Inflate the prices of a company through dissemination of false or misleading statements
  3. Sell shares in the company at the inflated price for a profit

- 50% of manipulated stocks are penny stocks trading in over-the-counter (OTC) markets (Renault, 2016).

Process

1. Simple statistical and machine learning models can detect anomalies based on price and volume data
2. Anomalies must be narrowed down to likely cases by eliminating price/volume movements explainable by factor returns, economic news, or legitimate (non-manipulative) company related news



Operating plan

Tasks

Phase 1 : Literature review and data gathering	▪ Review literature on detection of market manipulation and pump and dumps ▪ Create labeled dataset of market manipulation cases
Phase 2 : Model Evaluation	▪ Evaluate supervised and unsupervised models for anomaly detection ▪ Select simple models as basis for market manipulation detection algorithm
Phase 3: Feature Engineering	▪ Consider permutations of price and volume, including measures of volatility, momentum, as model features. ▪ Add general market performance, via factor returns, as model feature
Phase 4: News and Social Media	▪ Add economic news/ sentiment ▪ Add company specific news and social media sentiment

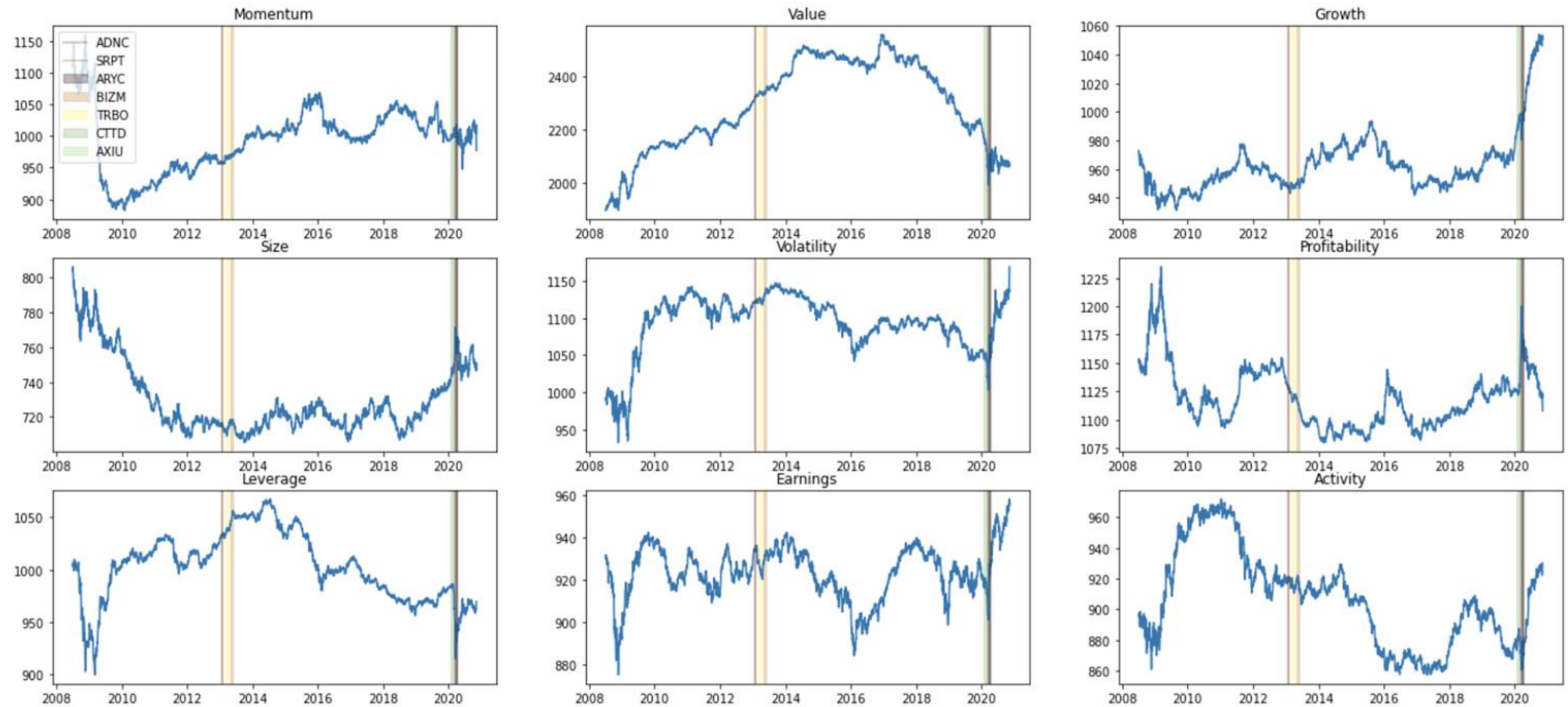
Data Gathering – Market Manipulation Cases

- The team examined over ~5,000 civil actions conducted by the SEC between 1996 and 2020. These included ~500 related to market manipulation, and 50 for which the defendants were ultimately convicted of pump-and-dumps.
- ~8 pump-and-dumps were ultimately selected as good test cases, with clearly identified dates.
- The algorithms were also tested on a machine generated Time Series Anomaly benchmark.

Ticker	Name	Date(s)
TRBO	Turbo Global Partners, Inc.	March 14 – April 8, 2020
ARYC	Arrayit Corporation	March 02 – April 13, 2020
BIZM	Biozoom	May – June, 2013
ADNC	Audience Inc	Jan 29, 2013
SRPT	Sarepta Therapeutics	Jan 20, 2013
AXIU	Axius	Feb 16 - 17, 2012
CTTD	CO2 Tech	Jan. 20 – February, 2007
DPRK	Deep Rock Oil and Gas	Aug. 16 – Sep., 2005
N/A	Numenta Time Series Anomaly Benchmark	

Data Gathering- Factor Data

- We managed to collect 4145 pieces of factor return data, starting from '2005-01'
- We took 3191 of them, since daily factor returns data began in '2008-07'
- Features: Momentum, Value, Growth, Size, Volatility, Profitability, DivYield, Leverage, Earnings and Activity
- For factor analysis we dropped ticker DPRK as the market manipulation happened before 2008



Evaluation of Unsupervised Models

Simple statistical models demonstrated high recall but low precision in identifying manipulation

	3 months			1 year		
	Recall	Precision	F1	Recall	Precision	F1
PersistAD	0.9231	0.4615	0.6154	0.9231	0.1062	0.1905
MinClusterAD	0.5385	0.6364	0.5833	0	0	0
PcaAD	0.9231	0.3750	0.5333	0.8462	0.0859	0.1560
RegressionAD	0.7692	0.2326	0.3571	0.6154	0.1404	0.2286
AutoregressionAD	0.7692	0.2703	0.4000	0.8462	0.0759	0.1392

Evaluation of Supervised Approaches

Supervised models showed good results that improved when factor returns were added

	Recall	Precision	F1
Decision Tree	0.7901	0.8348	0.8118
Random Forest	0.7984	0.9023	0.8472
SVM	0.0453	0.6111	0.0843
Logistic Regression	0.0823	0.5556	0.1434

	Recall	Precision	F1
Decision Tree	0.8148	0.8498	0.8319
Random Forest	0.8107	0.9163	0.8603
SVM	0.1523	0.6981	0.2500
Logistic Regression	0.2469	0.7059	0.3659

Social Media (Twitter) Analysis

- According to the SEC, a decent portion of market manipulation cases are done via spreading false information on Twitter.
- Manipulators looks to create either hype or fear around a particular stock to drive to stock price in either upwards or downwards direction.
- In those cases, it would be helpful to analyze both the volume and sentiment distribution of the recent tweets.

Gathering Twitter Data

- Necessary libraries: tweepy, textblob, pandas, re.
- Tweepy is a wrapper package for Twitter API.
- To set up Twitter API using tweepy, a twitter developer account is needed to get API keys.

```
# to get api keys, go to developer.twitter.com

consumer_key= 'hiyPAwor0laDNTWOAprAtehfr'
consumer_secret= '6LZ8wUfblvRHU5iQdU1XXBO7aTjqas63AjUV2IRPF0xUbJ6qIh'
access_token= '1108406328295387136-pcUtMtutcP8dipriyRAdmxaKSsMCa2'
access_token_secret= 'FPhlyI2trM7IELPqXwGGUvgROU9wBlD9gY9aVqfhIqoXw'


auth = tw.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tw.API(auth, wait_on_rate_limit=True)
```

- Then download tweets using that contains predefined search term, in this case, we use the $FTCH as example, a small cap e-commerce stock.
- Get date & time of each tweet.

```
# retrieving tweets of the following stock


search_term = "$FTCH -filter:retweets"

tweets = tw.Cursor(api.search,
                   q=search_term,
                   lang="en",
                   since='2020-12-01').items()

tweets_date = [tweet.created_at.date() for tweet in tweets]
print(len(tweets_date))
```

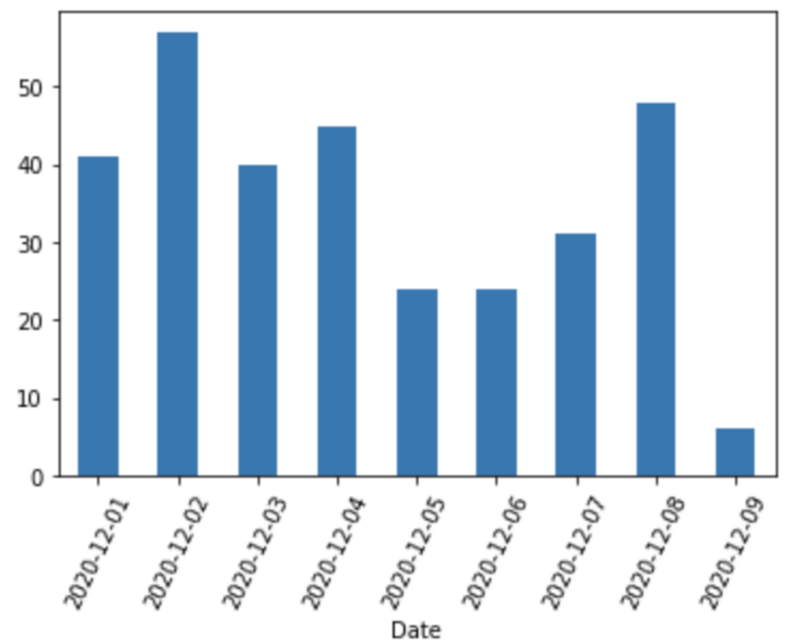
```
316
```

Twitter ‘Volume’ Analysis

- Assumption: manipulation usually results in dramatically increased attention around a particular stock.
- In the actual case of manipulation, we shall see an increase in tweet volume.
- Code to analyze volume:

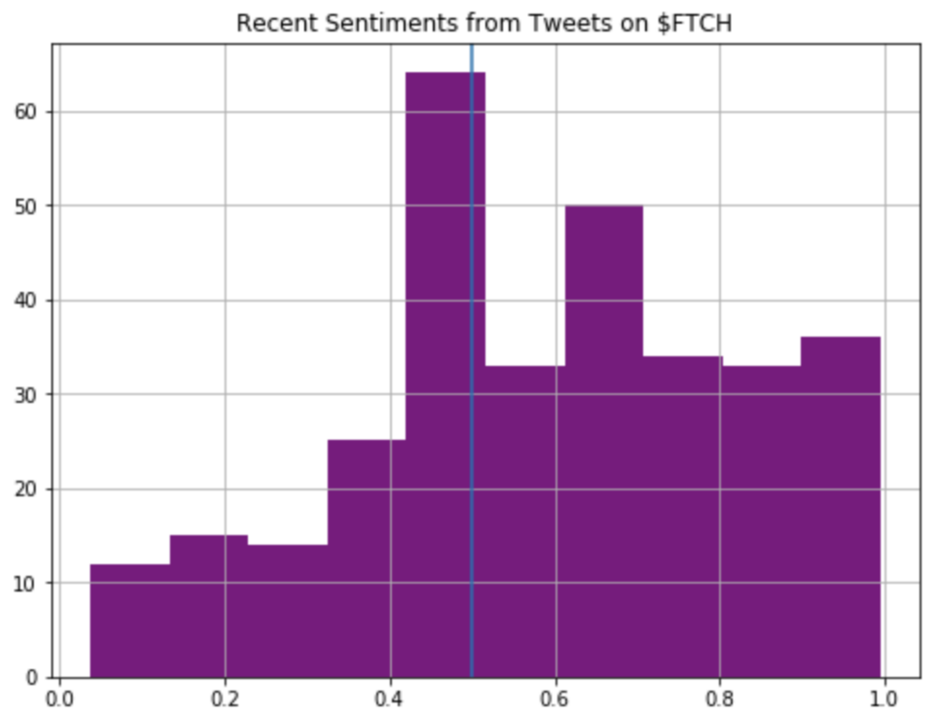
```
date_df = pd.DataFrame(tweets_date, columns = ["Date"])
date_df["Date"] = date_df["Date"].astype("datetime64")
date_df.groupby([date_df["Date"].dt.date]).count().plot(kind="bar", legend = False)
_ = plt.xticks(rotation=65)
```

- Graph of volume for ‘$FTCH’ for the week from 12/01 to 12/09:



Sentiment Analysis

- Besides volume, it is also helpful to look at the sentiments within those tweets.
- To analyze sentiments, we will use a naive bayes model. For each tweet, it offers a probability from 0 to 1, where 1 is absolutely positive, and 0 is absolutely negative.
- Achieved in Python using a package called 'Textblob'.
- Graph of sentiment distribution of '$FTCH' for the week from 12/01 to 12/09:



Our Limitations

- Although we have developed models and techniques to capture anomalies on social media, we can’t test them on actual cases of manipulation.
- Twitter API only allows users to retrieve tweets that are 7-8 days old, while sources that contains historical tweets requires an expensive membership.
- Further, fraudulent tweets are often deleted in a short period of time, therefore it is hard retrieve them.

Future Work

- Conduct analysis using more granular tick data. Prior research (Li, 2017) suggests this is not as effective, but it would nonetheless be a useful avenue for exploration if data could be obtained in a cost-effective manner.
- Expand the usage of news and social – media sources for volume and sentiment analysis. Historical social media data was too expensive to obtain, but twitter and stocktwits data are likely to be very useful.
- Expand the use of textual data beyond sentiment analysis using NLP.
- Utilize stock market returns (SPX, RTY) and volatility (VIX) returns to filter out false positives in cases in manipulation. Abnormal price or volume detection could just be a result of volatile market days.
- Incorporate Bloomberg economic announcement data as a signal or filter for manipulation. We never fully made use of the over 20k+ events data we have.
- Locate more recent examples of stock manipulation such as in the past year or two. The corresponding market data for these events would be much easier to acquire and so would the data from the social/news aspect.
- Identify other forms of stock manipulation. In our project, we mainly focused on pump-and-dump. However, there are other forms of abnormal behaviour out there that have not been fully investigated.
- Download company filings data from the SEC EDGAR website, which contain registration statements, quarterly reports, shareholder information, etc. Many stock movements might be explained by such filings which are available to the public.

References

1. Tsatsral Amarbayasgalan, Van Huy Pham, Nipon Theera-Umpon, Keun Ho Ryu. Unsupervised Anomaly Detection Approach for Time-Series in Multi-Domains Using Deep Reconstruction Error. *Symmetry, 2020*
2. David Diaz, Babis Theodoulidis, Pedro Sampaio. Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices. *Expert Systems with Applications, 2011.*
3. David Diaz, Koosha Golmohammadi, Osmar R. Zaiane,. Detecting stock market manipulation using supervised learning algorithms. *International Conference on Data Science and Advanced Analytics (DSAA), 2014.*
4. Seyed Koosha Golmohammadi. Time Series Contextual Anomaly Detection for Detecting Stock Market Manipulation. *University of Alberta, 2016.*
5. Seyed Koosha Golmohammadi, Osmar Zaiane. Sentiment Analysis on Twitter to Improve Time Series Contextual Anomaly Detection for Detecting Stock Market Manipulation. *University of Alberta, 2017.*
6. Aihua Li, Jiede Wu, Zhidong Liu. Market Manipulation Detection Based on Classification Methods. *Procedia Computer Science 122, 2017*.
7. Pankaj Malhotra, Anusha Ramakrishnan, Guarangi Anand, Lovekesh Vig, Puneet Agarwal, Guatam Shroff. LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection. *ICML 2016 Anomaly Detection Workshop, 2016.*
8. Pankaj Malhotra, Lovekesh Vig, Guatam Shroff, Puneet Agarwal. Long Short Term Memory Networks for Anomaly Detection in Time Series. *TCS Research and Jawaharlal Nehru University, 2015.*
9. Thomas Renault. Pump-and-dump or news? Stock market manipulation on social media. *European Financial Management Association, 2016.*
10. Haya Al-Thani. Detecting Market Manipulation in Stock Market Data. *Qatar University College of Engineering, 2017.*

Thanks To…

- Professor Sining Chen *(Faculty Advisor, Columbia)*
- Professor Ali Hirsa *(Faculty Advisor, Columbia)*
- Yash Madane *(Industry Advisor, Fidelity)*
- Michael Threlfall *(Industry Advisor, Fidelity)*
- Max Voelker *(Industry Advisor, Fidelity)*

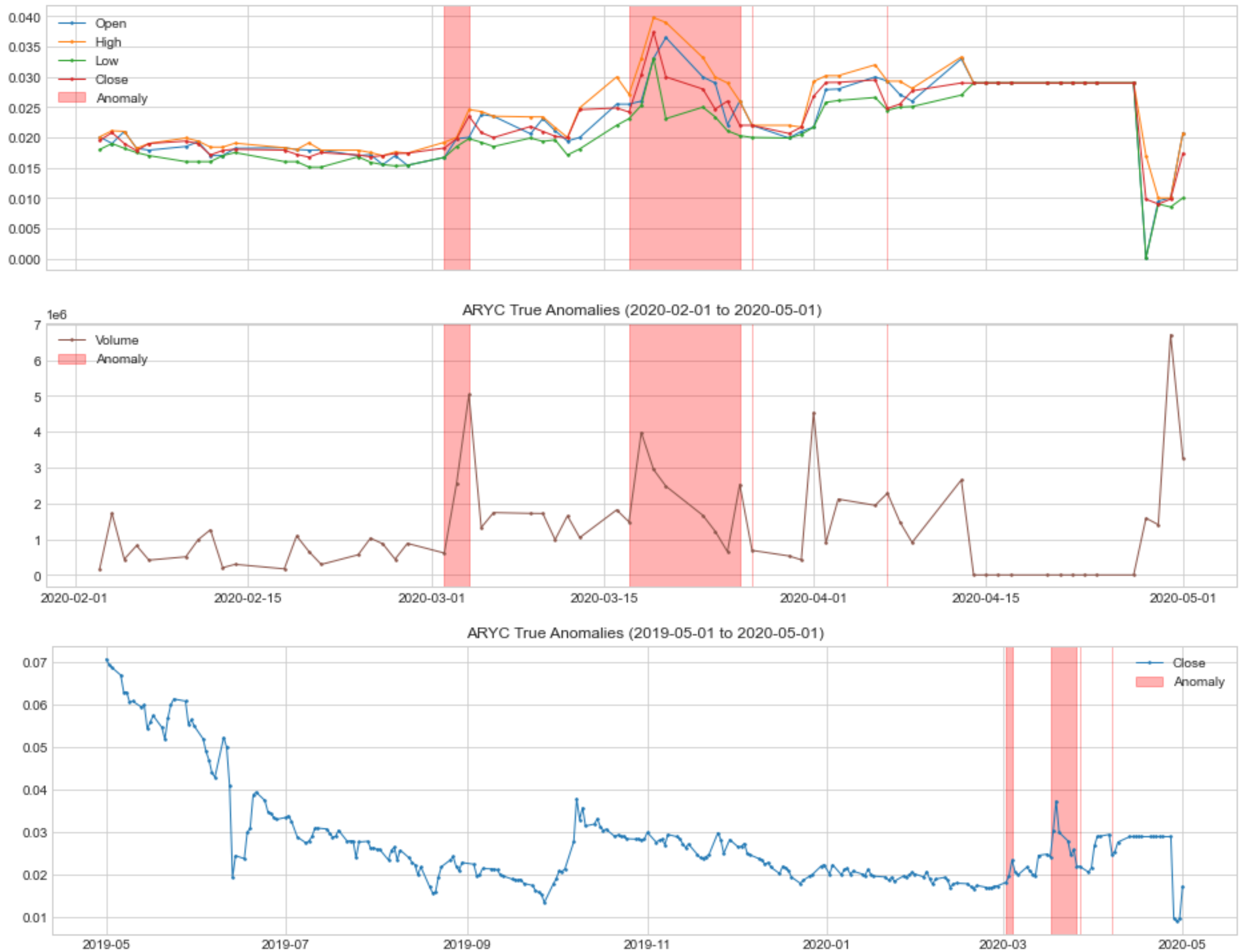
APPENDIX

Review: time series vs. subsequence anomaly detection

	Time series anomaly detection	Subsequence anomaly detection
Description	▪ Compare time series to model time series	▪ Detect anomalous patterns within segment of a time series
Models	▪ K Nearest Neighbors (KNN) ▪ Support Vector Machines (SVM) ▪ Random Forest ▪ Naïve Bayes ▪ Contextual Anomaly Detection (CAD)	▪ Long Short Term Memory (LSTM) ▪ Convolutional Neural Nets (CNN) ▪ Recurrent Neural Networks (RNN)
Cost/ Benefit	▪ Requires significant amount of (preferably) labeled data ▪ Recall is mediocre (~40% recall) even in in lab tests ▪ Largely stale problem with improvements likely to be based on better data gathering	▪ **Less data intensive – can work on a single time series** ▪ Recall can be high, but precision is low (will cure by addition of social media/ other data) ▪ State of the art problem with new solutions currently in development based on deep learning

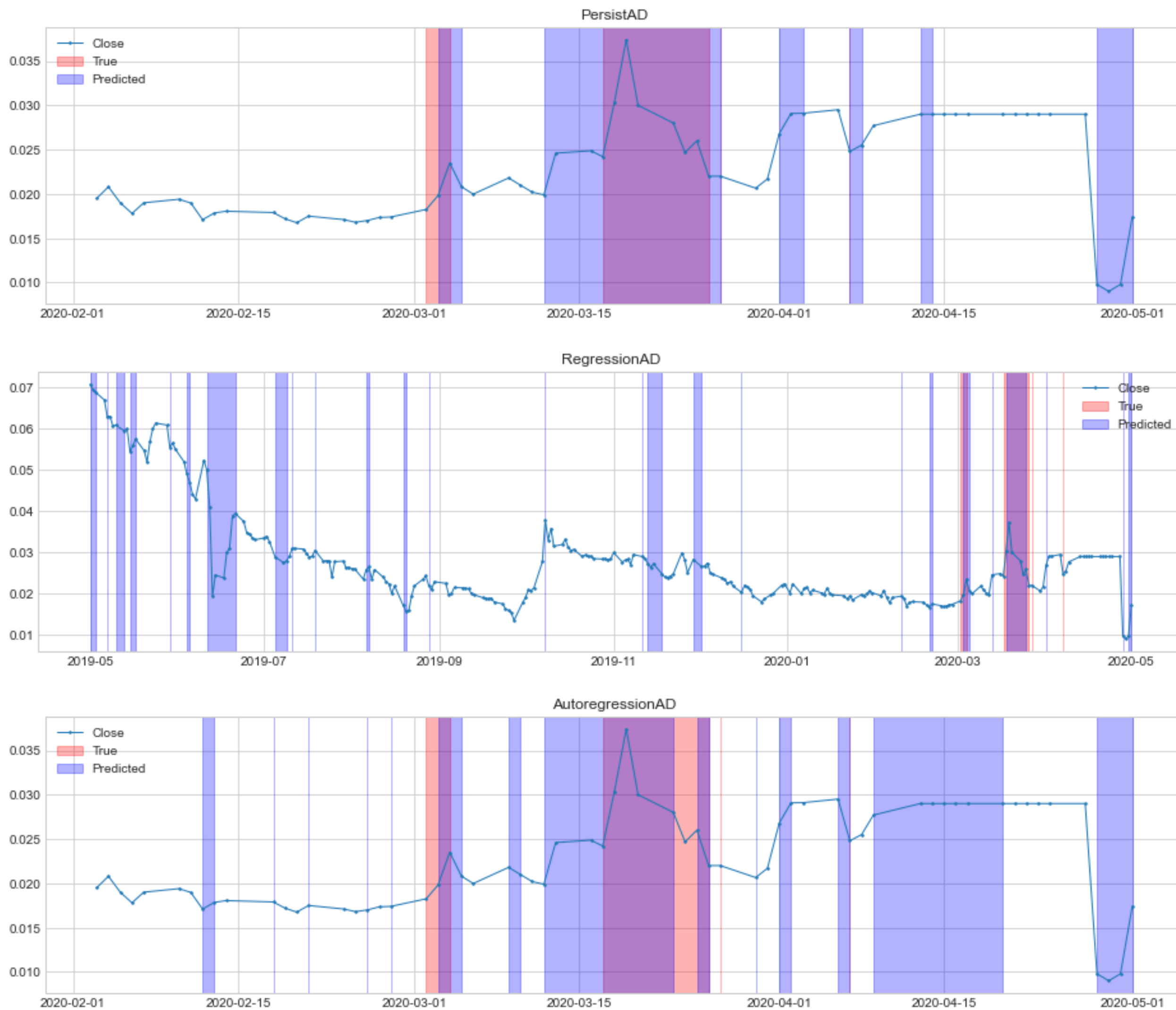
Example Anomaly Detection: ARYC

According to the complaint on the SEC website, ARYC was manipulated during March 2, 2020 to April 13 by means of pump-and-dump and spoofing. Below are visualizations of the EOD data from 02/01/2020 to 05/01/2020, with the red regions/lines representing known cases of market manipulation.



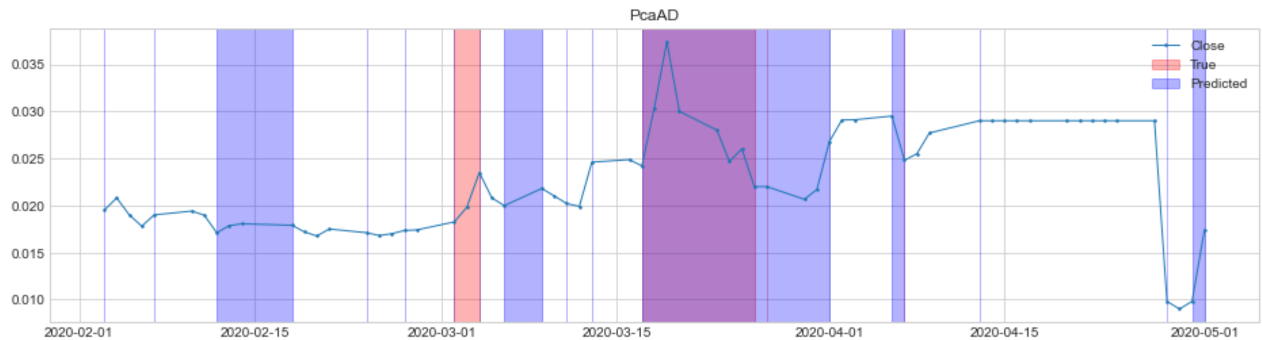
Example Anomaly Detection: ARYC

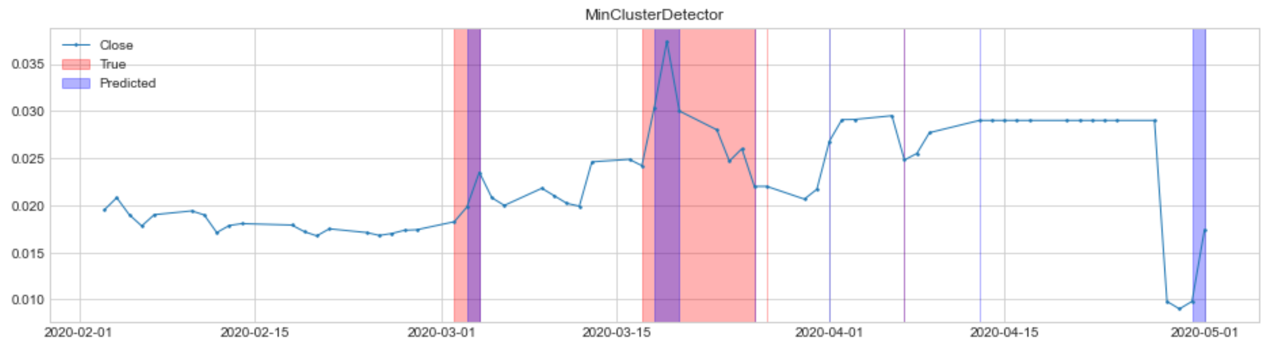
Below are comparisons of known cases of manipulation (red) to anomalies identified by simple statistical models (blue).



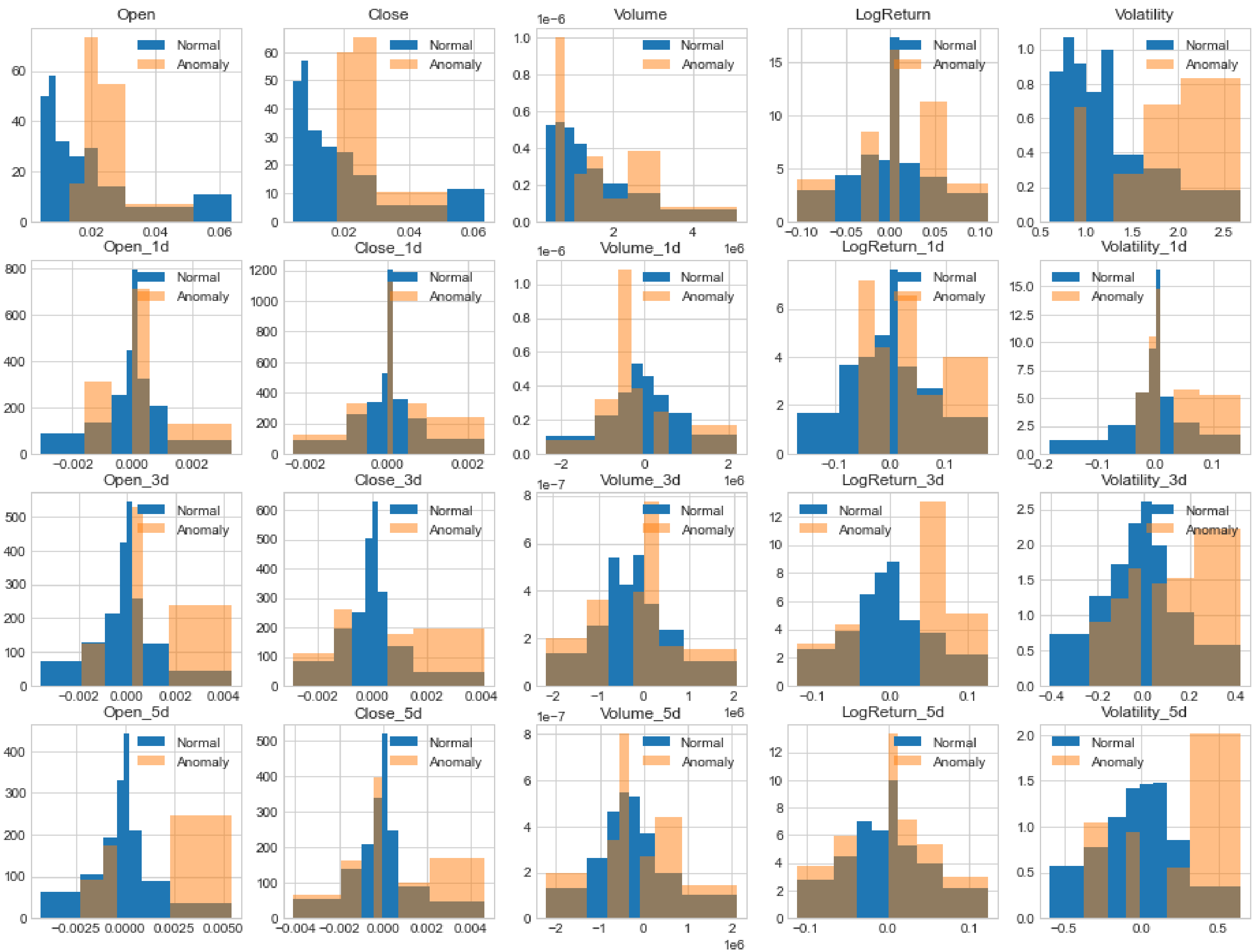
Example Anomaly Detection: ARYC

Below are comparisons of known cases of manipulation (red) to anomalies identified by simple statistical models (blue).





Distribution of features - ARYC



Possible Extension- Predicting days likely to be manipulated

- We labeled each day with 1 and 0, 1 if market manipulation happened
- 63.7% of the days happened market manipulation
- Then, we can turn the question into a classification problem: identifying/predicting the probabilities that a day likely to happen market manipulation
- Optionally, we can also count the number of 1s as an additional feature

```
factor.label.value_counts()
```

```
1    2033
0    1158
Name: label, dtype: int64
```

Possible Extension- Predicting days likely to be manipulated

Logistic Regression

	precision	recall	f1-score	support
0	0.53	0.28	0.37	290
1	0.68	0.85	0.75	508
accuracy			0.65	798
macro avg	0.60	0.57	0.56	798
weighted avg	0.62	0.65	0.61	798

Naive Bayes

	precision	recall	f1-score	support
0	0.72	0.48	0.58	290
1	0.75	0.89	0.82	508
accuracy			0.74	798
macro avg	0.74	0.69	0.70	798
weighted avg	0.74	0.74	0.73	798

Random Forest

	precision	recall	f1-score	support
0	0.98	0.99	0.98	290
1	0.99	0.99	0.99	508
accuracy			0.99	798
macro avg	0.99	0.99	0.99	798
weighted avg	0.99	0.99	0.99	798

Literature summary

Detecting Stock Market Manipulation Using Supervised Learning Algorithms - Zaiane

- Dataset from Diaz et. al., based on cases through SEC during 2003
  - 8 manipulated stocks
  - 25 similar stocks and 31 dissimilar stocks
  - 175,738 hourly transactional data
- Experiment with a few supervised learning algorithms
  - F2 score as metric, penalize false negatives more
  - Naive Bayes performs the best but precision is quite low

Algorithm	Sensitivity	Specificity	Accuracy	F_2 measure
Naïve Bayes	**0.89**	**0.83**	**0.83**	**0.53**
CART	0.54	0.97	0.94	0.51
Neural Networks	0.68	0.81	0.80	0.40
CTree	0.43	0.95	0.93	0.40
C5.0	0.43	0.92	0.89	0.35
Random Forest	0.32	0.96	0.92	0.30
kNN	0.28	0.96	0.93	0.26