SCS_3253 Machine Learning – Term project

Finding Signal in insider trading data

TEAM #2: DAVID CHASMAR, FERNANDO ESPINOSA, TIANYOU ZHENG, TAL NIR

Background

- **Securities Exchange Act (section 16)**
  - Regulating the behavior of corporate insiders, including directors, officers, and significant stockholders
  - Designed to promote transparency and reduce fraudulent activities
- Requires senior executives, directors, and large-block shareholders to report their company stock holdings and trading activities
- These reports submitted to SEC's Electronic Data Gathering, Analysis, and Retrieval (**EDGAR**) system
- The data sets is available in a flat file format to enable analysis
- The Insider Transactions data sets consists of XML data submitted from January 2003 through current period.

Project Objective

Original goal was

- Anomaly detection suggesting suspicious activity and leading to investigation into potentially illegal insider activity

As we worked with the data we shifted to focus on -

- **Building a model to find 'signal' in the insiders trading activity** that allows to:
  - Develop strategies based on insider trading transaction trends
  - Adjust portfolio allocation to increase/decrease exposure to stocks with heavy insider activity

- **Approach** - **Build a model to using insiders transaction data** to predict future recommended action and share price change **based on the public market performance** in the days following the insider transaction.

- As such we needed to combine standard stock market performance data with the insider trading data.

Methodology – Dataset

Our work utilized the Layline insider trading dataset (Layline Insider Trading Dataset (kaggle.com)) and stock trading data.

The dataset was built from the SEC published data by a team of researchers from Harvard**. The dataset is created from the original regulatory filings and includes all information reported by insiders since 2003.

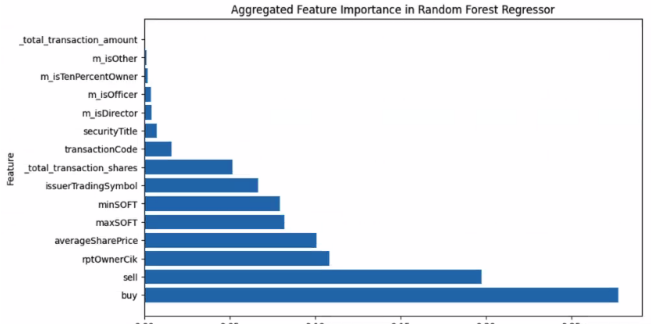
The data combines information about the Owner and Issuer as well about actual trading transactions made by insiders.
The original source data from 7 distinct datasets in combined into a panel dataset with 59 features and millions of records.

In addition historical stock market data was used to show the performance in the general market.

Source: Balogh, A. Insider trading. Sci Data **10, 237 (2023). https://doi.org/10.1038/s41597-023-02147-6

Methodology – Data Inspection

- Scope of the data narrowed to focus on:
  - Specific trading symbols – Companies of interest
  - Relevant transaction only - non derivatives, common shares
  - Specific Buy and Sell transaction codes
- Consideration about features
  - Features used in multiple formats across the dataset (date, type)
  - Aggregation required to group transactions broken by regulatory reporting model
- Feature importance review



Methodology – Data Cleansing

- Data issues
  - Share price errors – ‘fat finger’ errors, format errors
- Data Scaling and Normalization –
  - Scaling of numerical values
  - Share price – manually correcting for stock splits
- Data enrichment –
  - Market stock information with features of volume, price, etc. were added

Methodology – Data Enrichment

- Market stock information with share price was used to find Share price change:
  - Define time horizon – 30 days
  - Calculate share price change

$$\frac{(\boldsymbol{Share\ price\ on\ trx\ date\ -\ Share\ price\ at\ end\ of\ horizon})}{\boldsymbol{Shre\ price\ on\ transaction\ date}} \times \mathbf{100}$$

- Define threshold of price impact to categorize the impact:

$$\boldsymbol{Relevant\ decrease} < -5\% \qquad \boldsymbol{price\ change} \qquad +5\% > \boldsymbol{Relevant\ increase}$$

- Calculate the volume of transactions over past 90 days
- Use Price Change to generate **Signal** for the dataset:
  - **Buy/Sell instruction**
  - **Share Price Change prediction**

Machine Learning Methodology

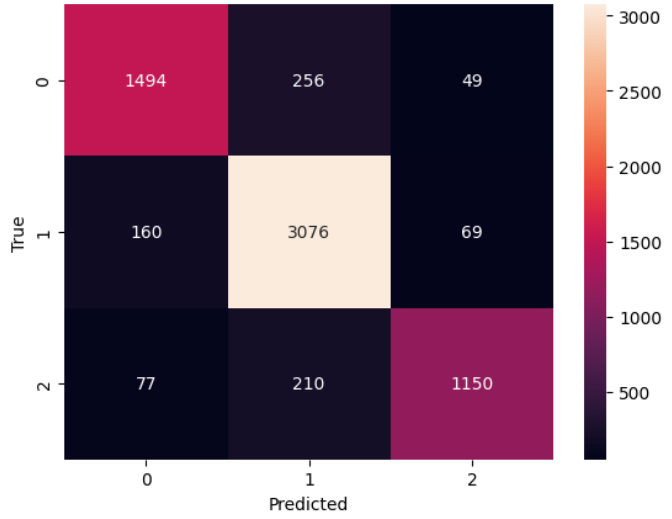
- Feature selection – lead to focus on numeric features
- Scaling – numeric values were scaled
- Test/train split –
  - spit separated 80%/20% (random)
  - Split by Stock
  - Split by Date
- Machine Learning Model build:
  - Recommendation model – Buy/Sell/Ignore:
    - Random forest
    - Ensemble Voting Classifier
  - Prediction model – Share price change:
    - Random forest

Machine Learning Models-Results

RANDOM FOREST CLASSIFIER

Accuracy: 0.8744840238495643

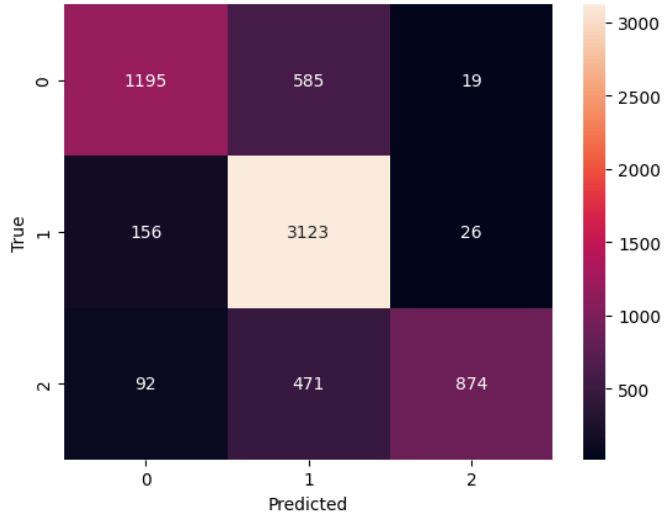
	precision	recall	f1-score	support
Buy	0.86	0.83	0.85	1799
Ignore	0.87	0.93	0.90	3305
Sell	0.91	0.80	0.85	1437



VOTING CLASSIFIER (RF, SVC)

Accuracy: 0.7937624216480661

	precision	recall	f1-score	support
Buy	0.83	0.66	0.74	1799
Ignore	0.75	0.94	0.83	3305
Sell	0.95	0.61	0.74	1437

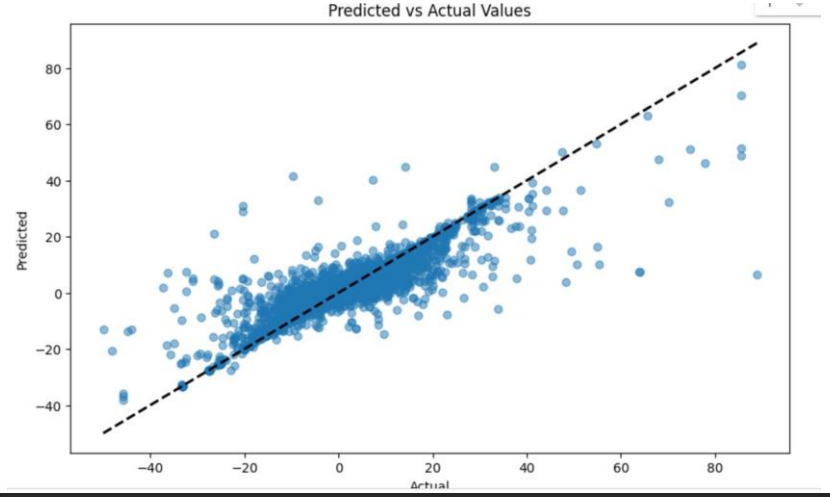


Machine Learning Models-Results

RANDOM FOREST PREDICTOR

	precision	recall	f1-score	support
0	0.91	1.00	0.95	103876
1	0.96	0.11	0.20	11961
accuracy			0.91	115837
macro avg	0.93	0.55	0.57	115837
weighted avg	0.91	0.91	0.87	115837

Precision: 0.9575091575091575



Conclusions & Open Items

- Open:
  - Share split and how they affect the dataset
  - Usability of the model on other stocks data
  - Focus made on the specific ‘known’ companies (Unicorns) – not clear how it would scale to others

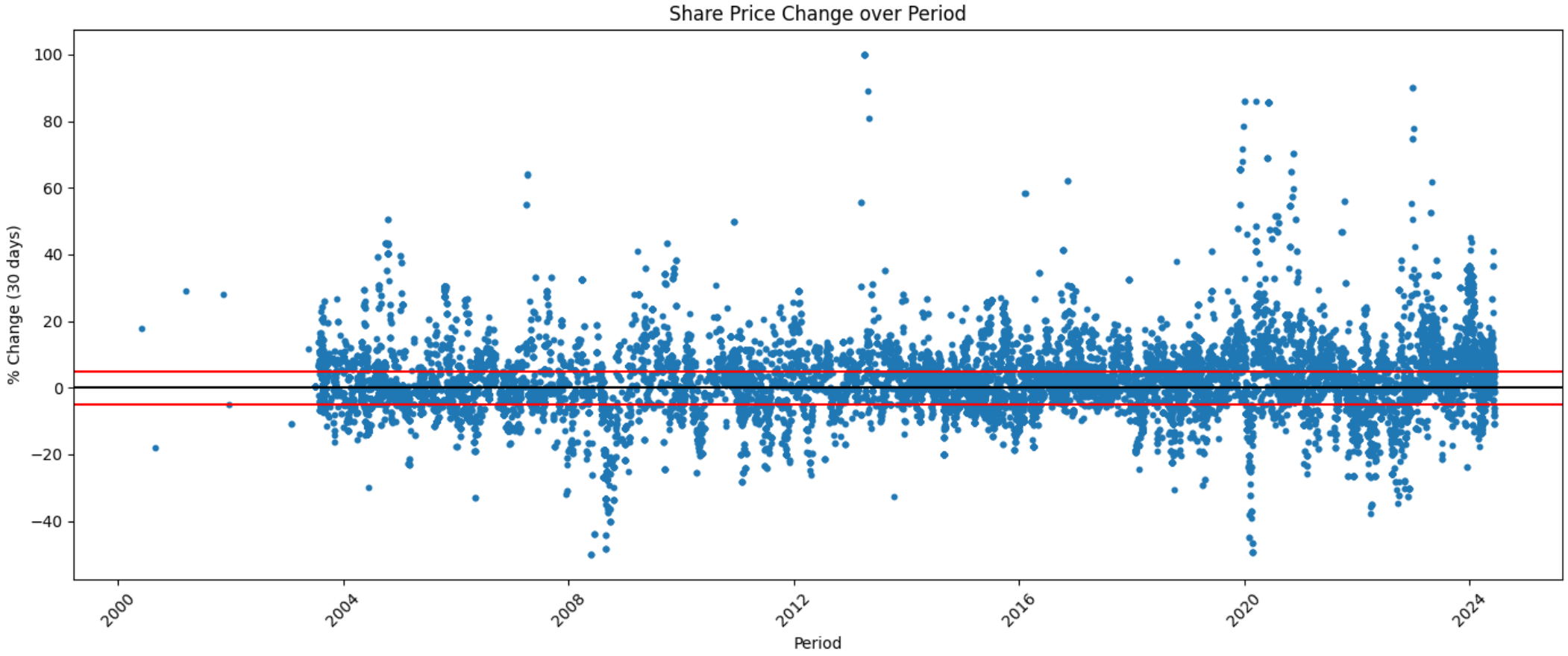
- Conclusions:
  - Not clear if true signal can be found in the insider traders data to predict direction and value of stocks
  - Potentially for anomaly detection to direct investigative resources should still be attempted

Challenges & Lessons Learned

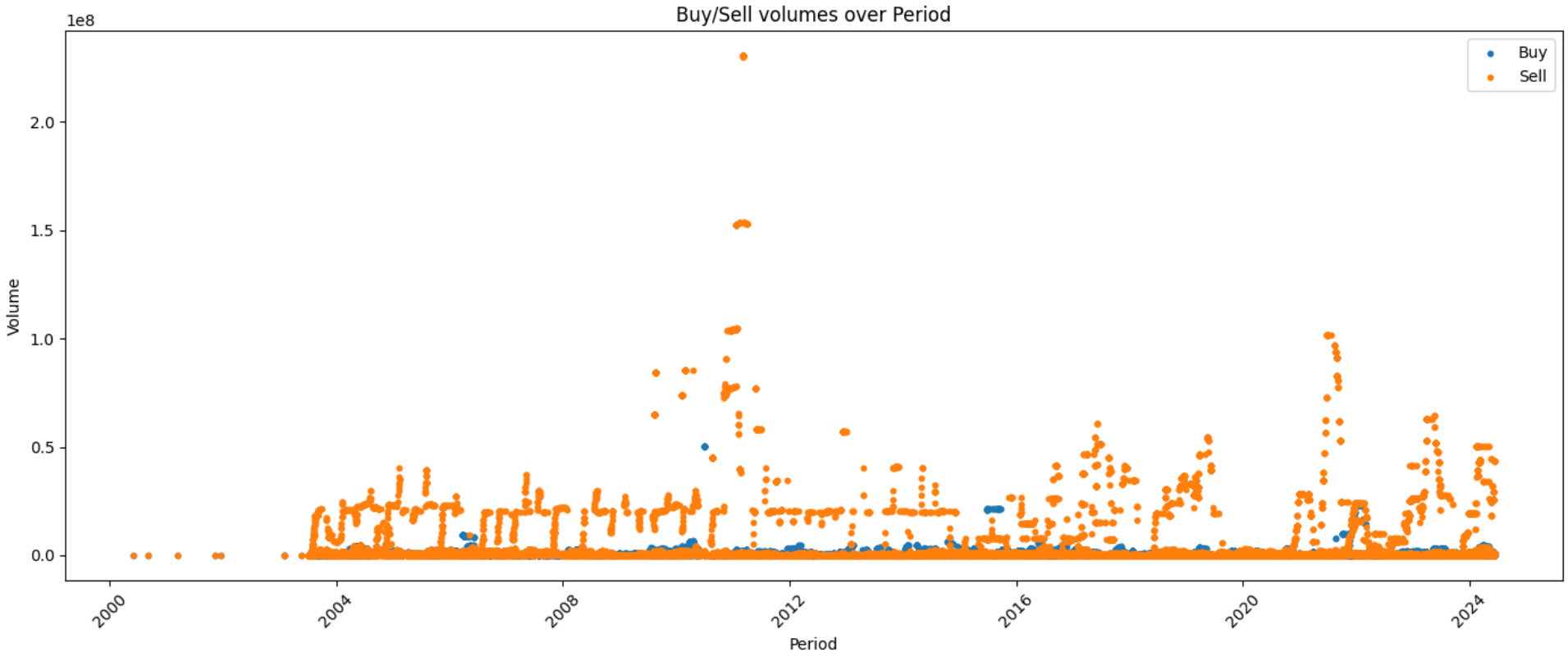
- Understanding the data – cited research assisted us but there was still a lot of cleansing to do
- Collaboration is key – brainstorming ideas and discussing doubts help create a direction and validate path to evolve model
- Trial and Error – fail quick and move to next option

Q&A

Visualizing Share Price Change %



Visualizing Buy/Sell Volume



Feature Importance