Predicting and analyzing stock market behavior using magazine covers

By

Egor Isakson, Bachelor of Science Degree in Computer Science

A thesis submitted to the Graduate Committee of

Ramapo College of New Jersey in partial fulfillment

of the requirements for the degree of

Master of Science in Data Science

Spring, 2023

Committee Members:

- Scott Frees Ph.D., Advisor
- Victor Miller Ph.D., Reader
- Amanda Beecher Ph.D., Reader

Acknowledgments

This paper marks the end of my education at Ramapo College. When I joined Ramapo in 2018, I knew nothing about coding and data, and I was not even been aware of the term “Data Science”. This journey has not been without struggles and some discomfort, but it’s been most rewarding as a result. I am very grateful for all the people that I’ve met and who’s been part of it. I want to especially thank Professor Miller for introducing me to coding in finance and believing in me. I also want to thank Professor Frees for guidance and extreme level of patience throughout the progress of this Master Thesis.

Table of contents

List of tables

List of figures

Abstract

Financial magazines have been part of the financial industry right from the start. There has long been a debate whether a stock being featured in a magazine is a contrarian signal. The reasoning behind this is simple; any informational edge reaches the wide masses last, which means by the time that happens, the bulk of the directional move of the financial instrument has long been completed. This paper puts this idea to the test by examining the behavior of the stock market and the stocks that are featured on magazine covers of various financial magazines and newspapers. By going through several stages of data extraction and processing utilizing a series of most up-to-date data science techniques, ticker symbols are derived from raw colorful images of covers. The derivation results in a many-to-many relationship, where a single ticker shows up at different points in time, at the same time, with a possibility of a single cover having many tickers at once. From then, several historic price and media-related features are created in preparation for the machine learning models. Several models are utilized to look at the behavior of the stock and the index at different points in time in the upcoming future. Results demonstrate more than random results but insufficient as the sole determinant of direction of the asset.

Chapter 1: Introduction

There are several infamous examples in financial history, where a popular magazine releases a cover, following which the price of the asset listed completely changes the direction of where it was originally headed. This poses a series of questions: Do magazine covers have an impact on future asset performance in any capacity? Do they act as contrarian signals? Are historic examples of this occurrence just cherry-picked or are there predictable patterns that lead to this behavior? This question encompasses several data science concepts and is open to a lot of experimentation with data. It's a mixture of sentiment analysis and machine learning, more specifically anomaly detection.

Solving this problem will lay to rest the myth and speculations in the financial industry whether people should pay attention to magazine covers. While this may uncover a hidden edge, it will first and foremost give us insight into the problem from the analytical perspective. This thesis is also about logging useful observations, that the public would also benefit from knowing.

There are two overarching challenges that make this problem hard to solve. The first challenge is getting the raw data and building a pipeline to prepare the data for analysis. Obtaining raw image data of magazine covers requires writing code to scrape data from different sources and then storing and processing it in a structured and scalable manner. The next step of the challenge is figuring out how to derive which ticker symbols are associated with text on the covers.

The second challenge is coming up with various features and models to solve this problem. To be able to have different models requires insight, intuition, creativity to develop new features as well as labels using existing data. There will be at least 2 types of features: price-related and magazine-related. This is an area of experimentation, where through trial and error, a useful model can hopefully be built.

The difficulty also arises in the fact that there are no assurances in the beginning of this project that I will be able to solve this problem, and to be able to start producing any insights, there is a certain amount of work to be done. In case I won't be able to develop a useful model, I will record all the methods that were attempted, so that the audience can eliminate those as potential predictors.

There is no clearly identifiable target audience who might be particularly harmed by the outcome of this project; on the contrary, the public would benefit from the results of the work conducted. This paper adds awareness and more public knowledge to a rather secretive field.

While performing research and doing the analysis, the following tech stack is being used: Python language and several key third-party modules for programming, MySQL, as well as Polygon.IO and ChatGPT REST APIs.

Based on our objective, the upcoming layout of this thesis is as follows: Chapter 2 adds more background and reviews related literature. Chapter 3 discusses the techniques and thought processes involved in resolving the two primary challenges described earlier in this chapter. Chapter 4 presents the results and examines them.

Conclusions are contained in Chapter 5. Chapter 6 is used for any additional references used throughout the paper.

Chapter 2: Background

Magazines

There are only a handful of impactful financial magazines. For this research, we only focused on those, as they are most likely to provide the most meaningful signals due to their reach and size of the audience. The following were used: Barron's, Forbes, Bloomberg Businessweek, Wall Street Journal, and Financial Times. The frequency of publication of this magazine varies, however. WSJ and FT are daily issues, Bloomberg and Barron's come out weekly, and Forbes publishes 8 times a year, or less than once a month. The implication of that is that there will be some count difference between the publications in the dataset, WSJ and FT making up a significant chunk of the dataset.

Similar work

While there have been newer methods of analysis due to an increase of newer methods to track data (something as simple as publicly available Google Trends), studying sentiment analysis or media influence on the stock market is not new.

One of the more famous studies has been done by Malcolm Baker and Jeffrey Wurgler, where they explored the relationship between investor sentiment and stock returns. (Baker) They collected data using various sources, including surveys and price-based indicators, and they found that high levels of sentiment are associated with lower subsequent stock returns.

Another academic paper by Joseph E. Engelberg and Christopher A. Parsons examined how changes in media coverage of individual stocks affect trading volume, stock returns, and market volatility. They observed that coverage impacts market volatility, with higher coverage associated with increased volatility. Thus, their work implies the direct influence of media on trading activity and market dynamics.

A more recent study from 2022 explored social media rumors on the Chinese stock market. The results reveal that rumors have a significant information transmission effect on volatility, and the developed Internet Financial Forum Rumor Index can help sense potential impacts, providing guidance for information environment optimization and promoting healthy stock market development. (Zhang)

Besides sentiment analysis, an important element of this thesis is OCR. OCR is technology that enables computers to recognize printed or handwritten text and convert it into machine-readable digital text. (Srihari) This is the initial process that is going to be used in the pipeline to extract tickers out of images. As the plan is to use a pre-trained OCR model, there needs to be a realization that it is model after all, and the better its input is, the better the output is going to be. In the case of OCR, better input means input that the OCR model ingests is easier to read. One of the preprocessing techniques for OCR is binarization. (Kaushik) Binarization of the image is converting a grayscale image into a black and white image. This can be achieved with the help of a process called thresholding. (Kaushik)

A couple examples of contrarian signals

There are a few stark examples in recent memory of magazine covers acting contrarian indicators. Let us look at them.

Figure 1 shows the front cover of Barron's on the recent dollar strength.

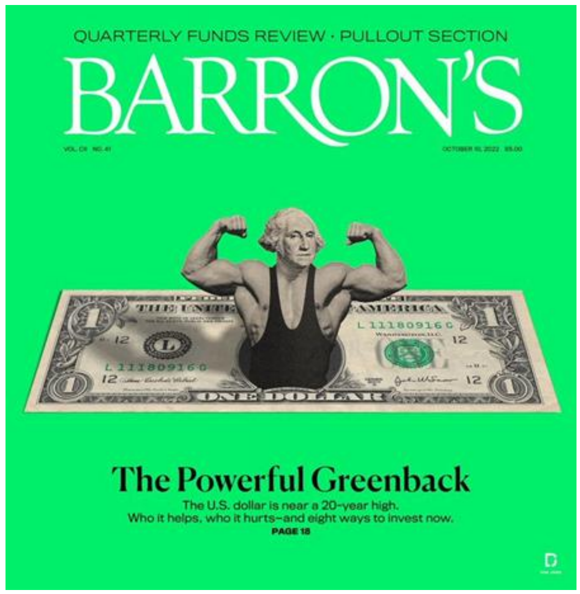

Figure 1: Barron's Dollar Cover

Figure 2 shows the chart of $DXY, the index that represents where dollar is currently valued among its few competitors (the arrow indicates the date of the publication of the above Barron's magazine).

Figure 2: $DXY Before and After Cover Release

The magazine with an extremely positive sentiment towards the dollar was within 2 weeks of the high, and the dollar has been in decline ever since.

Figure 3 shows a bullish contrarian example from the Economist.

Figure 3: The Economist Bearish Crypto Cover

Figure 4 shows the Bitcoin chart with the arrow marked when the Economist released that cover.

Figure 4: Bitcoin Chart Before and After the Economist Cover

The low of Bitcoin was less than a week ago when the magazine was released. Thus, looking at just these two examples, the magazine covers almost seem prophetic, and this idea is going to be put to the test.

Machine Learning Algorithms and Evaluation Metrics

Once the dataset is created, we trained machine learning models. Both classifier and regressor types of models were used. Let us provide a description of each.

A classifier is a type of machine learning model that is used for classification tasks. Classification is the process of assigning a predefined category or label to a given input based on its features or attributes. The goal of a classifier is to learn a mapping between the input features and the corresponding class labels. For example, identifying handwritten digits can be considered as a classification task.

A regressor is a type of machine learning model used for regression tasks. Regression involves predicting a continuous and numerical output value based on input features. The goal of a regressor is to learn a mapping between the input features and the corresponding continuous output value. For example, predicting house prices based on features like size, location, and number of bedrooms.

Several machine learning algorithms are going to be used for the dataset. Though the quality and quantity of input data takes precedence over any type of algorithm used, sometimes the algorithms may make a difference. The following will be used in this thesis: Random Forest, Gradient Boosting, and Dense Neural Network.

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. Each decision tree in the forest is trained on a random subset of the data and a random subset of features. During prediction, each tree casts a vote, and the final prediction is determined by majority voting or averaging the predictions. Random Forest is known for its ability to handle high-dimensional data and deal with noisy and correlated features.

Gradient Boosting is another ensemble learning method that builds an ensemble of weak learners, typically decision trees, in a sequential manner. Each subsequent weak learner is trained to correct the mistakes made by the previous learners. The predictions of all the weak learners are combined using weighted averaging, where the weights are determined based on the errors made in the previous iterations.

A dense neural network is a type of artificial neural network in which the information flows in one direction, from the input layer through one or more hidden layers to the output layer. Each neuron in a layer is connected to every neuron in the subsequent layer. The neural network learns by adjusting the weights and biases associated with these connections to minimize a specified loss function. Dense neural networks are good at learning complex nonlinear relationships in data.

There are also common machine learning evaluation metrics used for both regressor and logistic models. Below is the list of them.

Accuracy (classifier metric) measures the proportion of correctly predicted instances in classification tasks. Higher accuracy indicates better overall performance.

Recall (classifier metric) measures the ability of a binary classifier to correctly identify positive instances. It calculates the proportion of true positives out of all actual positives. Same as with accuracy, a higher value indicates better overall performance.

Mean Squared Error or MSE (regressor metric) is an evaluation metric for regression tasks. It calculates the average squared difference between predicted and actual values. Smaller MSE values indicate better model performance.

R-squared or R^2 (regressor metric) measures the proportion of predictable variance in the dependent variable. Higher R-squared suggests better model fit, but it does not indicate absence of errors or generalization ability.

Chapter 3: Methodology

Chapter 3 provides an overview of the data included in the dataset, specifically how the data points were created, what the features are, and which labels were used. Finally, we will discuss the data cleaning that was involved in the final stage prior to training the machine learning models.

The first step of the project was retrieving images of various financial publication covers and storing them in a single location. There were 5 publications that were used: Barron's, Forbes, Bloomberg, Wall Street Journal, and Financial Times. Images for these publications were pulled from 3 resources: Barron's images were scraped from Barrons.com, Forbes and Bloomberg were scraped from Magzter.com, Wall Street Journal and Financial Times were pulled from Kiosko.net. Each website required its own ways of extracting the image from the source, but 2 of the websites held 2 magazines each, so pulling the second magazine required merely a constant change.

With the images and their publishing dates all pulled into a central location, the image clean-up stage began. Before utilizing OCR to scrape words from images, images had to be properly prepared. It is considered good practice to convert the image color to black and white for a more successful OCR reading. Figures 5 and 6 show an example of the conversion.

Figure 5: Barron's Cover Before OCR preparation

All colorful images were converted to grey for proper OCR.

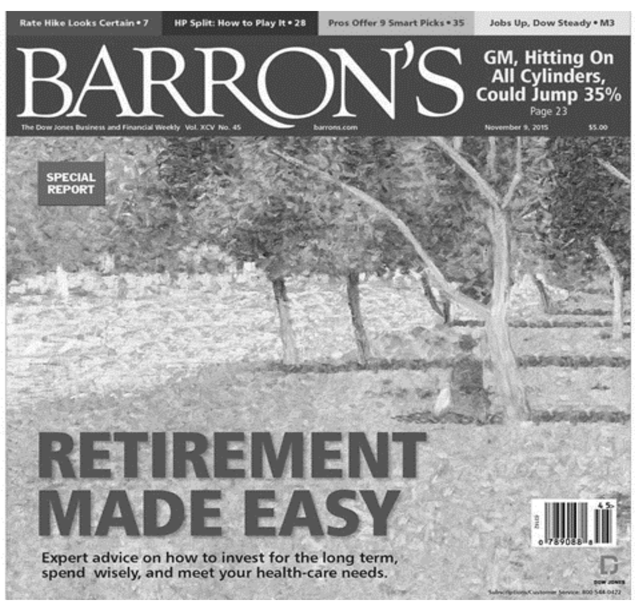

Figure 6: Barron's Cover After OCR preparation

As part of image exploration, it was also noticed that Wall Street Journal and Financial Times include too much information on a single cover page, as shown in Figure 7.


FINANCIAL TIMES

WEDNESDAY 18 MAY 2022 — INTERNATIONAL NEWSPAPER OF THE YEAR — MIDDLE EAST

Driving Deutsche
Bank's incoming head must steer a fresh course — BIG READ, PAGE 15

Owning Brexit
Johnson should embrace the deal that he signed — MARTIN WOLF, PAGE 17

Dollar danger
Rapid rise poses grave risks to poor countries — MOHAMED EL-ERIAN, PAGE 9

Ukraine relief
Injured taken from Mariupol

Briefing

Musk discussed Twitter deal with board before disclosing his stake

• Timeline in regulatory filings • New doubt over $44bn deal • Chance of litigation rises

Datawatch

Boeing's flight to Virginia deepens customer unease

Dimon's $50mn bonus under fire as JPMorgan investors throw out pay plan

World Markets

RONALD PHILLIPS
GREAT ENGLISH FURNITURE

A LARGE IRISH GEORGE III MIRROR CHANDELIER
IRISH, CIRCA 1780

Subscribe In print and online

A Nikkei Company

Figure 7: Financial Times Cover Before OCR preparation

This should not be surprising as both publications are actually newspapers, and the images show the fully flattened version of the front page. Thus, to avoid making processing less noisy, there was a decision made to crop out the upper bit and the lower half of the image. The upper bit was removed because it added no value from the analytical standpoint: it had a repetitive cover, and the same list of generic indices, which were not the center of the attention. Figure 8 is the processed version of the original image in Figure 7.

Ukraine relief
Injured taken from Mariupol

Wounded Ukrainian soldiers lie in a bus as they are evacuated by Russian troops from the Azovstal steel plant in Mariupol, Ukraine, after three months of intense battles in the frontline city.

Kyiv said more than 260 Ukrainian fighters had been evacuated, with 53 injured soldiers transferred to a Russian-occupied port town for medical treatment. An undisclosed number of Ukrainian fighters remained in the besieged plant yesterday.

Some Russian parliamentarians yesterday called for a ban on swapping Russian troops for Ukrainian soldiers from the Azov battalion, the military unit that had been leading the resistance at the steelworks.

War in Ukraine page 2

Musk discussed Twitter deal with board before disclosing his stake

● Timeline in regulatory filings ● New doubt over $44bn deal ● Chance of litigation rises

JAMES FONTANELLA-KHAN AND ANTOINE GARA — NEW YORK
TIM BRADSHAW — LONDON

Elon Musk first discussed taking Twit- investor, declaring in a filing that he had "no present plans or intentions" to make a bid. He amended that filing a day later to reflect that he was an active ability of discussing "potential business combinations and strategic alternatives" with Twitter's board and management team "from time to time". His hos- completing the deal at $54.20 a share when he said lowering the offer price was "not out of the question".

Twitter insisted it would force Musk

Briefing

▸ Senior ECB official floats half-point rise
Dutch central bank chief Klaas Knot has raised the prospect of a more aggressive shift in interest rates, sending ripples through markets and forcing the euro up against the dollar.— PAGE 3; MARKETS, PAGE 9

▸ Allianz arm admits securities fraud
The German insurer has agreed a $6bn settlement with US authorities and its investment division has pleaded guilty to fraud after a scandal at its funds business left investors with billions in losses.— PAGE 1

▸ Washington's Taiwan arms plan attacked
Business groups have told the US it is undermining the island's defences against Chinese aggression by limiting weapons shipments to those required for a D-Day scenario but neglecting other threats.— PAGE 4

▸ Biden spurns 'poison' ideology in Buffalo
The president has denounced white supremacy as a "poison" in US politics after blaming a "hateful ideology" for the mass shooting at a supermarket in Buffalo, New York.— PAGE 3

▸ Chinese poverty relief cash goes on tests
Local governments have been forced to divert funds from hunger relief and infrastructure to finance mass coronavirus testing, highlighting the financial challenges of Xi Jinping's zero-Covid policy.— PAGE 4

▸ Indian life insurance group in debut dip
Shares in Life Insurance Corporation have dropped more than 8 per cent on their stock exchange debut, as the country's biggest listing landed at a turbulent time for global asset markets.— PAGE 5; LEX, PAGE 18

▸ Mastercard rolls out 'smile to pay'
The card issuer has launched pilots for a biometric payments technology that allows shoppers to pay with a smile, as it tries to keep pace with nimble fintechs and bigger rivals such as Amazon.— PAGE 7

Datawatch

Figure 8: Financial Times Cover After OCR preparation

It is important to mention that the amount to crop out the upper and lower parts of the newspaper cover was constant for absolutely every newspaper cover. This was done mostly due to time an ever-so-slightly shifting sizing which was difficult to approximate accurately in a time-efficient manner. The issue with this hard-coded approach is that some newspaper covers did not have everything desired cropped out. This does not have a big impact. Figure 9 is an example, where the line of generic indices is visible as well as the lower half of WSJ.

THE WALL STREET JOURNAL.

DOW JONES

THURSDAY, NOVEMBER 13, 2014 - VOL. CCLXIV NO. 115 WSJ.com ★★★★ $2.00

DJIA 17612.20 ▼ 2.70 0.02% NASDAQ 4675.13 ▲ 0.3% NIKKEI 17197.05 ▲ 0.4% STOXX 600 335.09 ▼ 1.1% 10-YR. TREAS. unch., yield 2.359% OIL $77.18 ▼ $0.76 GOLD $1,158.90 ▼ $3.90 EURO $1.2488 YEN 115.47

What's News

Business & Finance

Citigroup and J.P. Morgan agreed to pay over $1 billion each to resolve currency-rigging probes, the biggest fines from a group of six banks. C1

■ Hasbro is in early talks to buy DreamWorks Animation, a deal that would bring an animation studio into the fold of a big toy maker. B1

■ A range of officials are raising concerns about tests North Dakota regulators will draw on in deciding rail-shipping rules for Bakken crude. B1

■ State Street's ties to an Ohio lobbyist are being examined as part of a broader look at the company's dealings with public pension funds. C1

■ Virgin Galactic's rocket ship had been plagued by technical problems long before its fatal crash last month. B1

■ Goldman promoted 78 people to the coveted role of partner, the second-smallest class since the bank's 1999 IPO. C1

■ Valeant is wooing doctors with consulting fees and luxury-hotel stays as it fights to buy Botox maker Allergan. B7

■ The Dow and S&P 500 slipped, breaking a streak of records. The blue chips eased 2.70 points to 17612.20. C6

2014: A Space Odyssey

SCIENCE NONFICTION

A robotic probe made the first landing on a comet and beamed pictures back to Earth, to the delight of French officials including President François Hollande, second from left, wearing 3D glasses to view results of the mission.

Rendezvous With a Comet

European Probe Makes Historic Landing on Celestial Wayfarer; a Hitch Then Success

By Gautam Naik

DARMSTADT, Germany—A robotic probe has landed on a million miles away from earth on the forbidding landscape of a small comet known as 67P/Churyumov-Gerasimenko.

comet, Rosetta bounced around the inner solar system like a cosmic billiard ball, circling the sun almost four times.

surface before settling back.

"Maybe today we didn't just land once but twice," joked Stephan Ulamec, lander man-

Russians Dial Up Military Threats

Russia is sending fresh convoys of troops and tanks into eastern Ukraine, NATO said, and threatening to dispatch warplanes on maneuvers over the Gulf of Mexico, flexing its military muscles in a Cold War-style escalation with the West.

By Naftali Bendavid in Brussels and Gregory L. White in Moscow

U.S. officials said they couldn't confirm the new incursion into Ukraine announced by the North Atlantic Treaty Organization and an international monitoring group in the region, but accused Russia of repeatedly violating the terms of a peace pact signed two months ago.

The international monitors also said that the Russian-backed rebels in Ukraine have been gaining some ground, and that the mission's surveillance drones have been shot at and jammed.

Gen. Philip Breedlove, NATO's top military commander, said Russia's goal may be to consolidate the separatist enclaves, making the area easier to control. "It is our first guess that these forces will go in to make this a more contiguous, more whole and capable pocket of

Figure 9: Wall Street Journal Cover After OCR preparation

When the image preparation for the OCR was complete, the OCR process began. Originally the approach was to use Pytesseract, a Python wrapper for C++ executable named Tesseract. Tesseract is highly configurable, it has several –psm modes, which choose a different approach of slicing and parsing text out of images. It was found that using "–psm 6" returned the most amount of useful data. However, upon closer inspection, the output text was too noisy. A decision was made to look for an alternative solution instead. Keras_ocr turned out to be a successful replacement. Table 1 is a comparison of Figure 6, where keras_ocr has much less data, so this module was used to parse text out of images.

Table 1: OCR comparison of Figure 6

pytesseract	keras_ocr
ee eae ee ESP TME Pros Offer 9 SmartPicks*35 Jobs Up, Dow Steady * M3 9) GM, Hitting On All Cylinders, Could Jump 35% Page 23 The Dow Jones Busnes and Financial Weekly Vol XCY No, 45 baronscom November 9, 2015 5.00 mage Tee eS NE i gee aN See We 2) ee re . je SPECIAL 16 gpa en a ks ee Be, OS oye es 5 REPORT | Ae go ie ees ol Eee 5 BN ee 7 Se oe. enor frag A ae ig he ES in Roe eR sy as Seer ata yy Via, Cmaeeeaath P y sae ag Pe ee an ae a 9 oe ae hy be Sa ee Se, ee BGO Es | ae Se a Gk, Hater og RH 8 co a 2 ae <= a4 os ees Ss	rate hike looks certain hp split how play 28 pros offer smart pickso 35 steady ma 7 to ite jobs up dow s barrons hitting gm on cylinders all could jump 3500 23 page financial the dow jones business and weckly vol xcv no 45 barronscom november 9 201s ss00 special report retirement asy made s ol789088 b advice expert how invest for the long to termy on spend and healthcare needs wisely meet your dos jonte sad subrcip lionecubtonss sorcos hoo da2

Ms te So ae ee ee Neth, - y . Mies I ii 4 5 qe �ae tasAS Be oe sj a cl % a �| oe Sages : eae te die eee : tits 6 �hyeneganenny MADE EASY i ! ME EAST hit Expert advice on how to invest for the long term, E., B | spend wisely, and meet your health-care needs. �a", ~ scStgdasucacee orictaisho7	

With the text available for all covers, the next step was to retrieve all the company names out of the text. This is a rather complex task for several reasons:

- Some covers had both tickers and company names of financial assets.
- Some company names had typos in them because of inaccurate OCR, so direct mapping was not always possible.
- There is some level of association between words and text required to correctly identify which ticker is exactly being talked about.

- Some tickers were no longer tradeable and/or under a different ticker than before.

There was no immediate, accurate, and time-efficient solution available. All available REST APIs and scraping tools relied on accuracy and absolutely 0% ambiguity.

Thus, a decision was made to experiment with the newest AI tool available to the public: ChatGPT. By carefully structuring the prompt and requesting ChatGPT to review every image in the dataset, the company names were extracted. There were several version iterations of the prompt. This is the final one that was used, where “textUsed” held the text from the image parsed using OCR:

```
"Identify all public companies in this text: \"{textUsed}\"\nKeep it short and provide a list separated with a pipe:"
```

Surprisingly, the model was able to parse out the company names with a rather high accuracy, tackling all the issues addressed in the bullet points above. Additionally, the model was also capable of being honest when there were no public companies able to be identified on the cover, and it was direct about it. However, there were several examples when ChatGPT got into a negative loop and refused to process the request, providing the following output: “Sorry, as an AI language model, I do not have the capability to identify all public companies in a text without any context or criteria.” ChatGPT returned the responses of company names indeed in a pipe delimited list which was easily parseable and ready to be used for the next stage.

Once all the company names have been collected, the next challenge was converting company names into tickers. When testing GPT prompts at the previous stage, there was an attempt to retrieve tickers explicitly instead of company names, but several prompts were tried unsuccessfully. Thus, converting company names into tickers created the same challenges as described earlier (except for typos, there were no typos from manually checked sample data). A decision was made to make another pass using ChatGPT but this time going through the list of company names. The following prompt was used for every list of public company names per table, where "textUsed" was a pipe-separated list of companies:

> "Assume you do not have access to real-time data. What are company symbols for the following companies: {textUsed}\nReturn a Python list of those symbols:"

Returning a Python list was an experimental approach to extract the data in a different but another parseable format from the ChatGPT's response, which also worked successfully. At this point, all the tickers at all dates listed on the covers were available: these are the data points for the final dataset. Once the data points were available, it was possible to calculate their labels and features.

Since it was not possible to predict which label exactly would be the best choice, several were created to be tested. All labels calculate the performance of a specific asset either a week, a quarter, or a year from the cover release date. There are several groups of labels used. Firstly, there are ticker and $SPY labels. $SPY is a ticker that represents S&P 500 index price. This can be a benchmark label, but also a form of baseline that can be compared to the model performance of equal period for a ticker label. Secondly, the labels are split into quantitative and categorical types. Quantitative labels all reflect percentage change from the cover date. The “bool” within categorical label columns, in retrospect, may be a confusing since all those columns have 3 categories, not 2: green, neutral, and red. Neutral is anything between 1% and -1%; it was not prudent to consider anything within this range as either red or green. Below is the full list of labels used.

Table 2: List of Labels

Labels
ticker_week_perf
ticker_quarter_perf
ticker_year_perf
spy_week_perf
spy_quarter_perf
spy_year_perf
ticker_week_bool_perf
ticker_quarter_bool_perf
ticker_year_bool_perf
spy_week_bool_perf
spy_quarter_bool_perf

spy_year_bool_perf

Next come the features. There were 2 types of features: price-based, volume-based, and magazine-based. Price-based features focused on calculating percentages from specific starting points. Volume-based features focused on calculating the ratios between recent average volume and yearly average volume. Magazine-based features focused on counts associated with the ticker and its relation to the cover. Below is the table of features used as well as their description and reason for inclusion.

Table 3: List of Features

Features used	Description	Inclusion reason
hist_ticker_week_perf	Ticker performance in the past week leading up to being listed on the cover	Check if past recent behavior can reverse after magazine release
hist_ticker_quarter_perf	Ticker performance in the past quarter leading up to being listed on the cover	Check if past recent behavior can reverse after magazine release
hist_ticker_year_perf	Ticker performance in the past year leading up to being listed on the cover	Check if past recent behavior can reverse after magazine release
hist_spy_week_perf	SPY index performance in the past week leading up to the ticker being listed on the cover	Check if past recent behavior can reverse after magazine release

hist_spy_quarter_perf	SPY index performance in the past quarter leading up to the ticker being listed on the cover	Check if past recent behavior can reverse after magazine release
hist_spy_year_perf	SPY index performance in the past year leading up to the ticker being listed on the cover	Check if past recent behavior can reverse after magazine release
rvol_buildup_week	Relative volume: average ticker volume over last week / average ticker volume over last year	Check if increase in relative volume may indicate higher significance of the event
rvol_buildup_month	Relative volume: average ticker volume over last month/ average ticker volume over last year	Check if increase in relative volume may indicate higher significance of the event
retracement_pct_from_biweekly_highs	Calculating the close price at the day of the cover release relative to the high of last 2 weeks	Check how eager the instrument is to move relative to recent highs
retracement_pct_from_biweekly_lows	Calculating the close price at the day of the cover release relative to the low of last 2 weeks	Check how eager the instrument is to move relative to recent lows
num_valid_tickers_per_cover	Number of companies listed on a specific magazine issue	Check if less tickers per cover indicate higher significance

num_ticker_mentions_past_month	Number of times this company has already been mentioned in the last 4 weeks before the event	Check if over usage of mentions may results in lack of additional movement.

At this point we have a dataset with all features and labels for every datapoint; it is time to clean the data in preparation for the machine learning training. Several entries in the dataset had null values. There were several reasons. The biggest reason is that the price data source that was used in the project only provided data for exactly the last 10 years since the day of the REST API call to pull data. March 2013 was the earliest month that the data was available for, but there were magazine covers before that, some dating 2010. Since some calculating features and labels were not possible, they had null values and thus removed from the dataset. Some data points were also removed because the label date has not actually been materialized (i.e., it is not possible to calculate a quarterly or yearly $SPY price change for yesterday's cover).

The dataset is ready with all the features and labels, so we start training the data. We are going to be using a Random Forest model, Gradient Boosting, as well as Neural Networks, both the classifier and the regressor versions for both. We are also going to train on a bigger and smaller dataset. The bigger dataset is the original size; the smaller dataset is the one without the Wall Street Journal and Financial Times. The reasoning there is because WSJ and FT are both daily issues, and it is possible they are introducing noise to the dataset because of this high frequency of releases.

Chapter 4: Analysis and Discussion

In this chapter, we are going to go over the results of the models, interpret them, and highlight several insights that were made using the clean dataset.

Before analyzing the machine learning results, it is important to mention the overall dataset sizes at various project stages. About 1700 stocks were involved in the project. Also, here is the list of the magazines and their respective cover count in the project: 3349 Wall Street Journal covers, 2129 Financial Times covers, 424 Bloomberg Businessweek covers, 375 Barron's covers, 87 Forbes covers. About 3160 covers did not have any tickers identified with them. 17579 entries were in the entire final dataset before cleaning. 11904 entries were available for model training after cleaning the data by replacing rows with null values. The model that was trained without WSJ and FT had 782 entries in total.

As planned, due to the uncertainty of what is going to work, there were quite a few models trained. There were 3 main machine learning algorithms (Random Forest, Gradient boosting, Neural Network) each trained on 12 labels. In addition, there was an additional model training with a reduced dataset, excluding WSJ and FT articles, making another 12 models. Thus, there were 48 baseline models trained in the dataset.

Table 4 shows the results of the Random Forest training (both regressor and classifier) on a full dataset.

Table 4: Labels and their evaluation metrics for Random Forest model

Label	1st Metric	2nd Metric
ticker_week_perf	MSE: 78.55	R^2 score: -1.156
ticker_quarter_perf	MSE: 489.073	R^2 score: -0.581
ticker_year_perf	MSE: 2719.776	R^2 score: -3.551
spy_week_perf	MSE: -3.551	R^2 score: -0.407
spy_quarter_perf	MSE: 160.469	R^2 score: -3.436
spy_year_perf	MSE: 382.644	R^2 score: -4.252
ticker_week_bool_perf	Accuracy: 0.407	Precision score [0.394, 0.156, 0.42]
ticker_quarter_bool_perf	Accuracy: 0.437	Precision score [0.719, 0.0, 0.414]
ticker_year_bool_perf	Accuracy: 0.534	Precision score [0.718, 0.0, 0.511]
spy_week_bool_perf	Accuracy: 0.381	Precision score [0.46, 0.216, 0.418]
spy_quarter_bool_perf	Accuracy: 0.301	Precision score [1.0, 0.0, 0.286]
spy_year_bool_perf	Accuracy: 0.532	Precision score [1.0, 0.0, 0.514]

Table 5 shows the results of the Gradient Boosting training (both regressor and classifier) on a full dataset.

Table 5: Labels and their evaluation metrics for Gradient Boosting model

Label	1st Metric	2nd Metric
ticker_week_perf	MSE: 88.308	R^2 score: -1.424
ticker_quarter_perf	MSE: 496.37	R^2 score: -0.604
ticker_year_perf	MSE: 2029.675	R^2 score: -2.396
spy_week_perf	MSE: 15.466	R^2 score: -0.639
spy_quarter_perf	MSE: 133.671	R^2 score: -2.695
spy_year_perf	MSE: 357.36	R^2 score: -3.905
ticker_week_bool_perf	Accuracy: 0.437	Precision score [0.619, 0.182, 0.426]
ticker_quarter_bool_perf	Accuracy: 0.425	Precision score [0.667, 0.000 , 0.409]
ticker_year_bool_perf	Accuracy: 0.527	Precision score [0.705, 0.000, 0.509]
spy_week_bool_perf	Accuracy: 0.403	Precision score [0.812, 0.228, 0.426]
spy_quarter_bool_perf	Accuracy: 0.317	Precision score [1.000, 0.000, 0.290]

spy_year_bool_perf	Accuracy: 0.527	Precision score [1.000, 0.000, 0.512]

Table 6 shows the results of the Neural Network training (both regressor and classifier) on a full dataset.

Table 6: Labels and their evaluation metrics for Neural Network model

Label	1st Metric	2nd Metric
ticker_week_perf	MSE: 106.151	R^2 score: -1.914
ticker_quarter_perf	MSE: 598.715	R^2 score: -0.935
ticker_year_perf	MSE: 2362.087	R^2 score: -2.952
spy_week_perf	MSE: 20.484	R^2 score: -1.171
spy_quarter_perf	MSE: 157.19	R^2 score: -3.345
spy_year_perf	MSE: 521.494	R^2 score: -6.158
ticker_week_bool_perf	Accuracy: 0.381	Precision score [0.481, 0.137, 0.42]
ticker_quarter_bool_perf	Accuracy: 0.458	Precision score [0.597, 0.065, 0.42]
ticker_year_bool_perf	Accuracy: 0.54	Precision score [0.617, 0.0, 0.522]

spy_week_bool_perf	Accuracy: 0.427	Precision score [0.424, 0.364, 0.444]
spy_quarter_bool_perf	Accuracy: 0.343	Precision score [1.0, 0.038, 0.302]
spy_year_bool_perf	Accuracy: 0.549	Precision score [0.97, 0.0, 0.527]

Table 7 shows the results of the Random Forest training (both regressor and classifier) on the dataset excluding WSJ and FT.

Table 7: Labels and evaluation metrics for Random Forest model of smaller dataset

Label	1st Metric	2nd Metric
ticker_week_perf	MSE: 40.451	R^2 score: -0.48
ticker_quarter_perf	MSE: 371.297	R^2 score: -0.124
ticker_year_perf	MSE: 1892.647	R^2 score: -1.196
spy_week_perf	MSE: 5.807	R^2 score: -0.231
spy_quarter_perf	MSE: 40.807	R^2 score: -0.018
spy_year_perf	MSE: 51.695	R^2 score: -0.313
ticker_week_bool_perf	Accuracy: 0.346	Precision score [0.571, 0.0, 0.377]

ticker_quarter_bool_perf	Accuracy: 0.564	Precision score [0.667, 0.0, 0.556]
ticker_year_bool_perf	Accuracy: 0.782	Precision score [0.333, 0.0, 0.8]
spy_week_bool_perf	Accuracy: 0.333	Precision score [0.0, 0.262, 0.455]
spy_quarter_bool_perf	Accuracy: 0.615	Precision score [0.0, 1.0, 0.618]
spy_year_bool_perf	Accuracy: 1.0	Precision score [1.0]

In short, all regressor labels have not performed adequately, with all of them having high MSE and negative R^2. The negative value suggests that the models are not able to explain the variance in the target variables and fail to capture the underlying patterns in the data. This is fair in a way, as asking the model to predict the exact percentage within a week, a quarter, or a year is a very demanding ask, and with such limited number of features, it is likely not feasible.

The classifier models performed relatively better, as the task was seeking fewer specific results, only the direction of the asset after a certain period. The models containing the entire dataset ranged between 30% and 55% accuracy. Between the ticker and SPY labels for the same period, there were some models that performed better predicting ticker direction, while there were also other models that had higher accuracy for the SPY label instead.

There is an interesting behavior happening with the smaller model. Regressor models perform similarly inadequately, but the performance range on classifier models is much wider than the regular datasets. The increasing accuracy as the time interval increases suggests 2 things: (1) It is possible that being listed on the cover results in a more reliable longer-term stock performance rather than shorter term. (2) The latter can be subject to the survivorship bias of only including those tickers that have survived and went up. At the same time, the trajectory of the general market is up more often than not.

While performing some exploratory analysis, there were several interesting observations found that are shared below. For instance, 60% of stocks that went up 100% leading up to the cover date closed green a week after the cover release as shown in Figure 10.

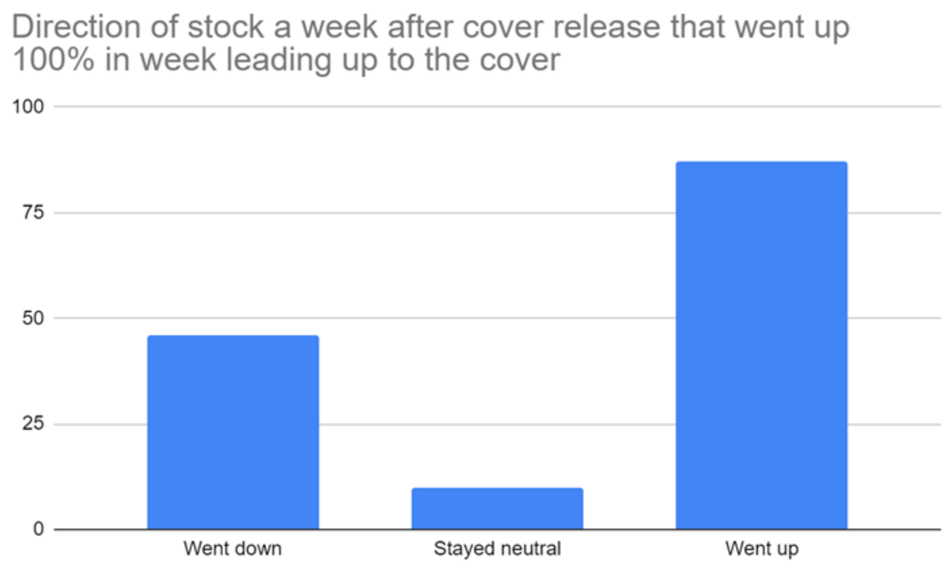

Figure 10: Additional Label Insights

Furthermore, 76% of stocks that went up 100% leading up to the cover date closed green a quarter after the cover release as shown in Figure 11.

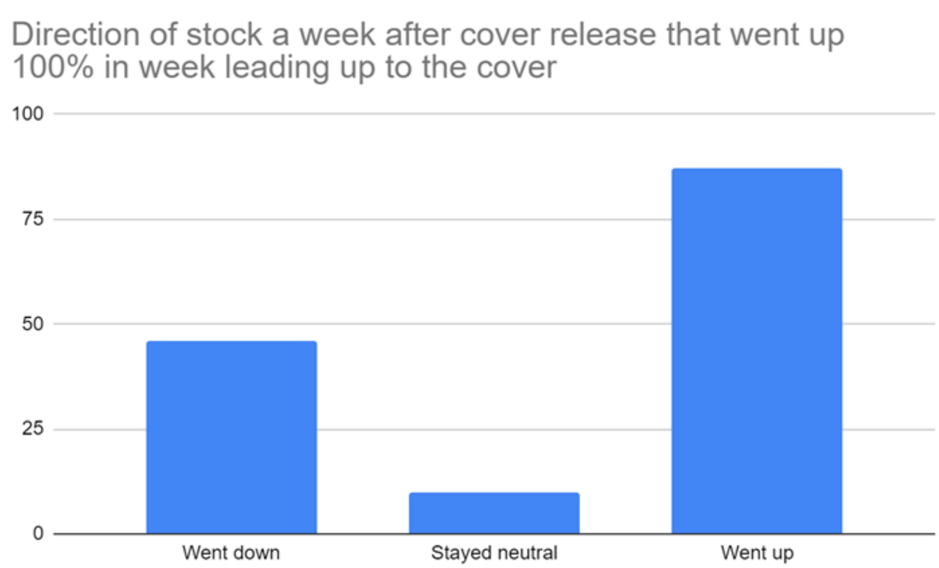

Figure 11: Additional Label Insights

This insight almost contradicts the initial hypothesis of this thesis. It is likely that those are anomalies as less than 100 instances out of an entire dataset were recorded to have this phenomenon.

Another interesting insight is that the stocks featured on the covers were more likely to have a smaller directional movement a week after the cover than a week leading up to it as shown in Figure 11. This likely is tied to volatility, and how it was not allowing the directionality to happen.

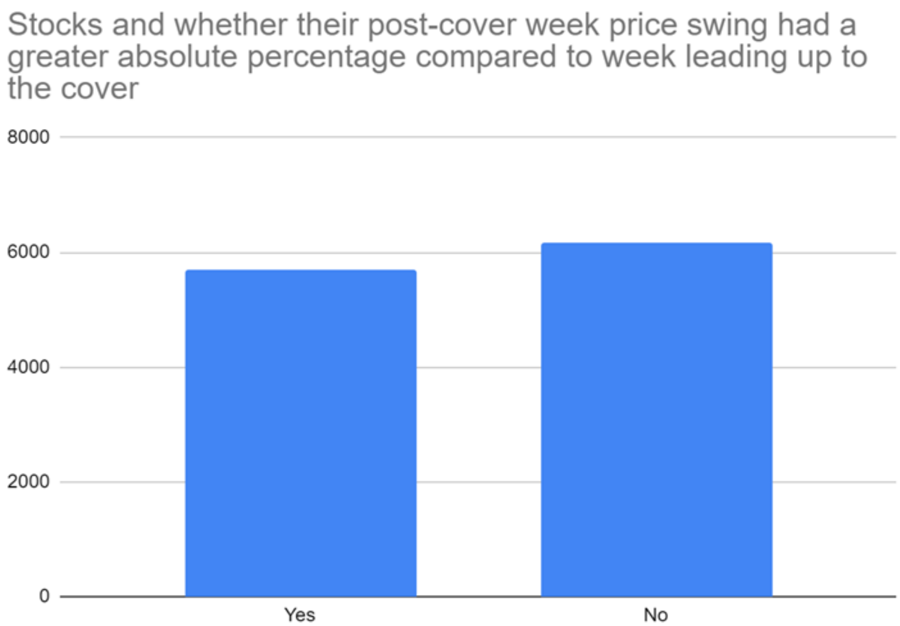

Figure 12: Additional Label Insights

Overall, regressor models were not successful in their predictions; the classifiers had more success, with the model without WSJ and FT entry points showing most success. In addition to that, there were interesting insights made while analyzing the labels and the features.

Chapter 5: Conclusions

The goal of this thesis was to analyze and predict the impact of financial magazine covers on the stocks features on them. We demonstrated the end-to-end process of how to create a dataset of features and labels per stock per cover, and we also analyzed the results of machine learning models and the additional insights made along the way.

When extracting ticker data points from magazine covers for the dataset, we had a multi-step process utilizing 2 highly trained AI models, an OCR model and a ChatGPT model. In-between of usages of both models, there were several data cleaning and data processing techniques involved.

When datapoints were available, the generation of labels and features began. There were regressor and classifier type labels, across several periods of time, both for the ticker on the cover and $SPY, an S&P 500 index. Features were price-based, volume-based, and magazine-based. Upon the completion of the dataset, the data was cleaned and trained using 3 different Machine Learning algorithms: Random Forest, Gradient boosting, and Neural Network.

The results of the models did not demonstrate a strong ability to predict the direction of the asset based on existing features alone. However, we were able to extract useful insights based on the dataset alone, without the models.

Moving forward, there are areas of improvements: some things to adjust in the current state and some add things as future work. Let us walk through the improvement at each stage of the project.

The most important item is the inclusion of sectors as part of model labels. Initially the plan was to test input data on tickers, SPY, as well as sectors of the corresponding tickers, but due to time constraints this was not implemented. Sectors could have been the sweet spot between the performance of the individual ticker and the performance of SPY. There are several classification systems of tickers into sectors, but the most ideal one would be the one that has most likely under 10 sectors in its set.

Also, there are several elements to consider in the process of deriving tickers out of images. A very limited number of images were of a very poor quality, and thus no tickers were able to be extracted from that set of images. So, higher quality images for some covers. Also, parsing text from images is not 100% reliable. There might be typos in parsing that make interpretation not possible. In addition to that, introducing ChatGPT for parsing tickers out of raw text is also not 100% reliable. There could be false positives as well as false negatives. On top of that, the approach in this paper involved extracting tickers exclusively. That means that, any metaphorical text or text that related to the industry is omitted and does not get accounted for in the training data.

When it comes to feature engineering, there are also ways to improve. One of the hottest features to add to the model is the sentiment on the cover towards the ticker mentioned. The dataset only had the mention of the ticker, but not what the sentiment towards it is; sentiment could be a pivotal feature, but it is difficult to capture it. The number 1 issue: the difficulty of parsing words out of text and then associating them with the mood towards the asset. Additionally, it would be a great addition to the model to include sentiment from other sources on the same day of the cover release or leading up to it. It could be Google Trends or posts from social media. This way we are expanding the model to include more sentiment from more sources. The challenge then is in evaluating which source's sentiment takes precedence or has heavier weight. It is no surprise that different media sources had different levels of priority across time. For instance, newspapers in the 60s most likely had more influence than they do now, as there is a shrinking audience of newspaper readers and a growing audience of social media users. Thus, when incorporating sentiment from different sources, it would be a good idea to calibrate their weight across time if possible.

Overall, sentiment and media do seem to have an influence on the stock marker behavior, but the model should likely be used in conjunction with other non-sentiment indicators to have a more accurate evaluation of the stock's behavior in the near to midterm future.

References

"An overview of Optical Character Recognition (OCR) research" by Sargur Srihari (2000).

"Investor Sentiment and the Cross-Section of Stock Returns", Malcolm Baker and Jeffrey Wurgler (2006)

“The Causal Impact of Media in Financial Markets”, JOSEPH E. ENGELBERG and CHRISTOPHER A. PARSONS

“Effect of social media rumors on stock market volatility: A case of data mining in China”, Hua Zhang

“Identify all public companies in this text: \"{textUsed}\"\nKeep it short and provide a list separated with a pipe:” prompt. ChatGPT, 13 Feb. version, OpenAI, 20 Mar. 2023, chat.openai.com/chat.

“Assume you do not have access to real-time data. What are company symbols for the following companies: {textUsed}\nReturn a Python list of those symbols:” prompt. ChatGPT, 13 Feb. version, OpenAI, 25 Mar. 2023, chat.openai.com/chat.

"Preprocessing Techniques for Optical Character Recognition" by Saurabh Kaushik and Sanket Kulkarni (2016).

Van Rossum, G. & Drake, F.L., 2009. Python 3 Reference Manual, Scotts Valley, CA: CreateSpace.

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Chollet, F. & others, 2015. Keras. Available at: https://github.com/fchollet/keras.

Magzter.com. Magzter, n.d., https://www.magzter.com/.

Kiosko.net. Kiosko.net, n.d., https://kiosko.net/.

Barron's. Barron's, n.d., https://www.barrons.com/.