Quarterly Earnings Predictor

Based on News Articles

By Matt Elmajian

Problem

- Wall Street Analysts provide Estimated Earnings for Publicly traded companies
- Companies announce earnings approximately 1 month after closing their books
- Can articles posted within one week of a company’s earnings announcement predict whether or not they will outperform or underperform their estimate

Data Wrangling

- AlphaVantage API
  - Provides API to pull companies’ earnings estimates and actuals dating back to 1990
  - Provides API to pull news articles dating back to 2022
- Pull data on 30 companies over their last 10 earnings reports and associated articles
  - Companies: ['AAPL', 'NVDA', 'MSFT', 'AVGO', 'META', 'ORCL', 'CRM', 'AMD', ‘UBER', 'ADBE', 'IBM', 'QCOM', 'MU', 'ANET', 'GOOG', 'TSM', ‘TSLA', 'NFLX', 'BABA', 'CSCO', 'SONY', 'SHOP', 'PLTR', 'INTU', 'ADP', 'DELL', 'ABNB', 'SPOT', 'PYPL', ‘CRWD’]
- Use BeautifulSoup to web scrape articles based on url’s provided

API Outputs

- Earnings API Output

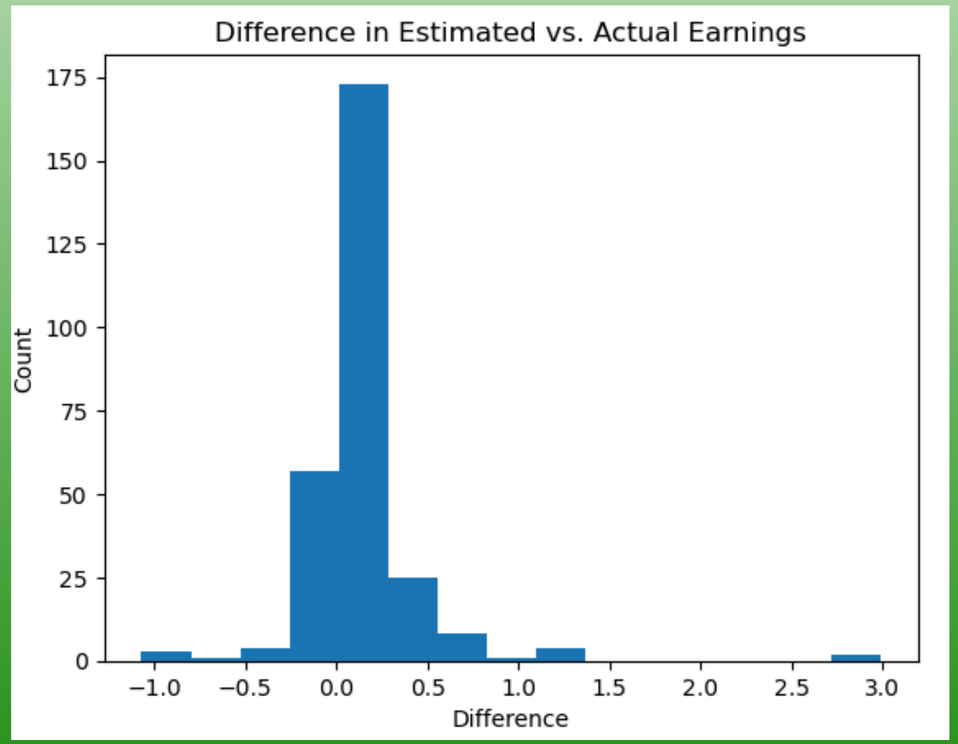
	fiscalDateEnding	reportedDate	reportedEPS	estimatedEPS	surprise	surprisePercentage	reportTime	ticker
0	2024-06-30	2024-08-01	1.4	1.35	0.05	3.7037	post-market	AAPL
1	2024-03-31	2024-05-02	1.53	1.5	0.03	2	post-market	AAPL
2	2023-12-31	2024-02-01	2.18	2.1	0.08	3.8095	post-market	AAPL
3	2023-09-30	2023-11-02	1.46	1.39	0.07	5.036	post-market	AAPL
4	2023-06-30	2023-08-03	1.26	1.19	0.07	5.8824	post-market	AAPL

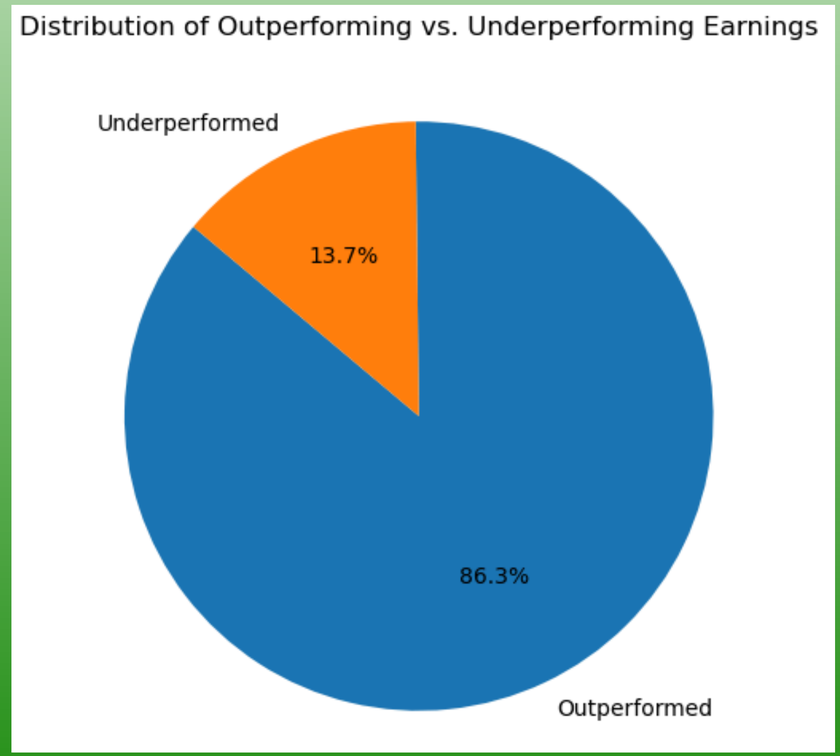
- News API Output

	title	url	time_published	authors	summary	banner_image	source
0	Advanced Energy Q2 Earnings Top Estimates, Rev...	https://www.benzinga.com/news/earnings/24/07/4...	20240731T201919	[Zacks]	Advanced Energy Industries AEIS reported non-G...	https://staticx-tuner.zacks.com/images/charts/...	Benzinga
1	AMD Q2 Earnings Beat Estimates, Shares Up on S...	https://www.zacks.com/stock/news/2313423/amd-q...	20240731T181200	[Zacks Investment Research]	AMD's second-quarter 2024 results reflect robu...	https://staticx-tuner.zacks.com/images/article...	Zacks Commentary

EDA - Target Variable

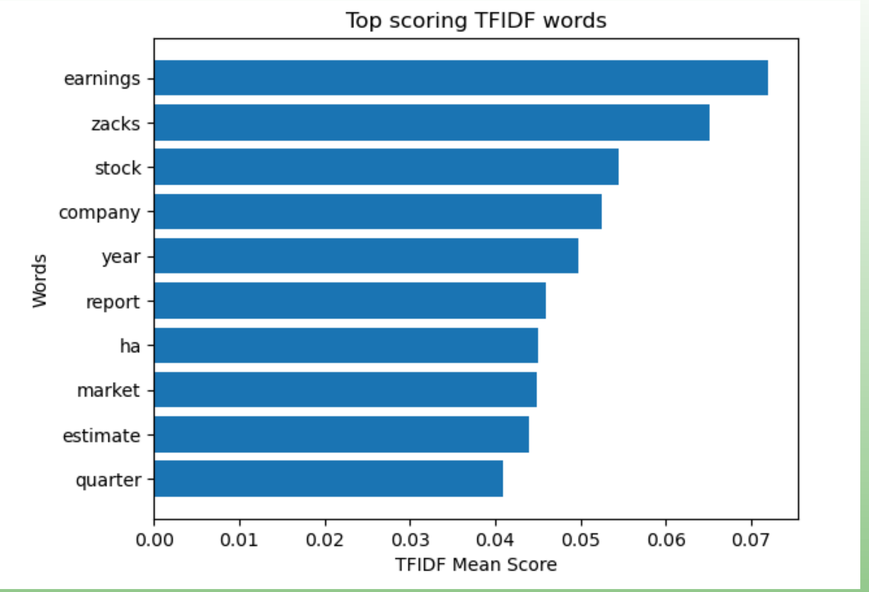
- Over the last two years, many of the companies tested outperformed their estimates.
  - 86.3% at or above estimates

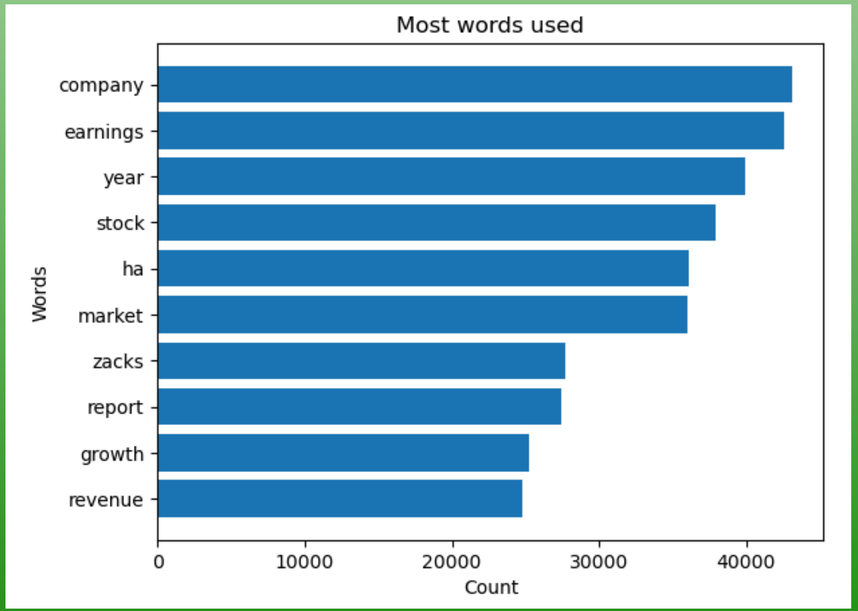




EDA

- Cleanings & Preprocessing:
  - Remove numbers
  - Remove punctuation
  - Remove Emojis
  - Lemmatize words
- Used TFIDF & Count Vectorizer





EDA/Modeling

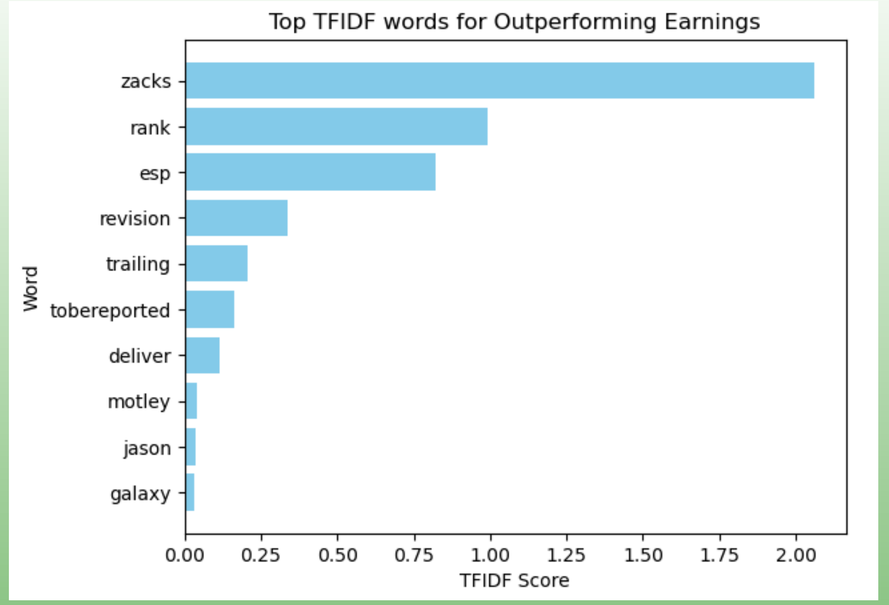
- Using Naive Bases Algorithm to determine most predictive words based on sentiment
  - Vader Sentiment algorithm on each article
- Majority of articles came back with positive sentiment
- Used these words to find relationships between words used and earnings reports

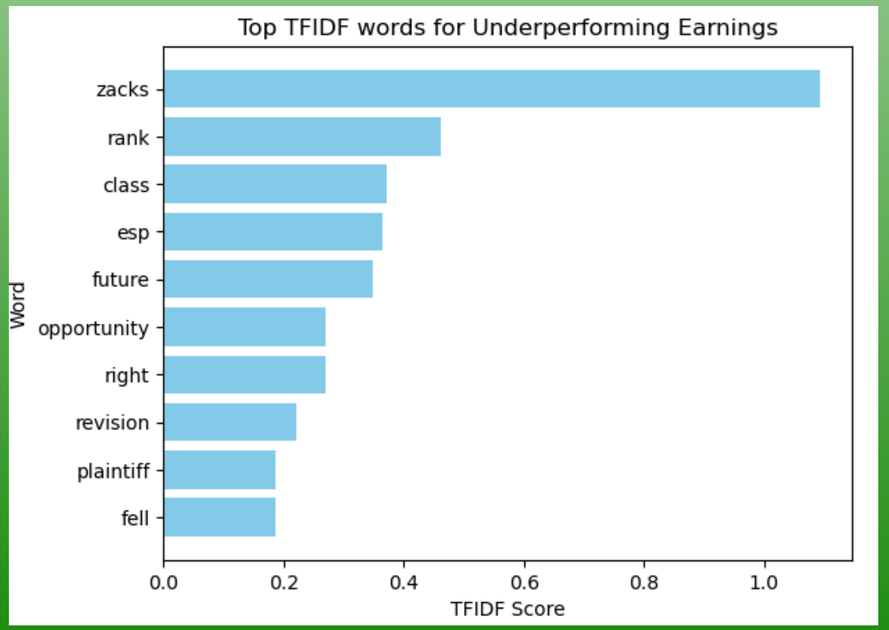
	word	probability
0	esp	0.999670
1	etfs	0.999040
2	trailing	0.998887
3	deliver	0.998563
4	rank	0.998464
5	zacks	0.998390
6	jason	0.998251
7	motley	0.998226
8	galaxy	0.998221
9	tobereported	0.998110
10	revisions	0.998017
11	dylan	0.997950
12	expense	0.997889
13	yeah	0.997718
14	score	0.997612
15	style	0.997542
16	efficiency	0.997496
17	surpassed	0.997486
18	opportunities	0.997456
19	pegged	0.997400

	word	probability
0	court	0.814771
1	class	0.814396
2	fell	0.813555
3	housing	0.802341
4	futures	0.801251
5	filed	0.800502
6	rights	0.787389
7	debt	0.787161
8	ukraine	0.768223
9	biden	0.754843
10	plaintiff	0.751073
11	cramer	0.750915
12	lawsuit	0.738893
13	falcon	0.728038
14	failed	0.726175
15	crude	0.700571
16	russia	0.665005
17	defendants	0.654167
18	misleading	0.636024
19	fear	0.567789

EDA/Modeling

- Highest scoring words for articles released prior to outperforming and underperforming earnings
  - Only inclusive of top 20 most predictive words for outperforming and top 20 for underperforming
- Only few differences:
  - Words more specific to Underperforming: opportunity, plaintiff, fell





EDA/Modeling

- Used most predictive words as features in final dataset to run through binary classification models (outperform or underperform earnings)
  - Tested models with values for key words being Count, and TFIDF scores
  - Created models both including and excluding average sentiment as a feature
- Algorithms Tested: Support Vector Machine, Random Forest Classifier
- Error metric: Balanced Accuracy

Models

- Train vs. Test sets: 70/30
- Using TFIDF as values for the word features tended to score better than simple Counts
- Most models scored at or below 50% indicating little to no predictive power
- Small Dataset could have caused predictive issues

Models without Sentiment		
Random Forest - TFIDF	No hyperparameter tuning	Balanced Accuracy: 50%
Random Forest - TFIDF	Criterion: 'entropy', min_samples_leaf: 2, n_estimators: 50, min_samples_split: 5, max_depth: 10, bootstrap: False	Balanced Accuracy: 50%
Random Forest - Count	No hyperparameter tuning	Balanced Accuracy: 48.6%
Random Forest - Count	Criterion: 'entropy', n_estimators: 50, min_samples_split: 5, max_depth: 10, min_samples_leaf: 1, bootstrap: False	Balanced Accuracy: 49.3%
SVM - TFIDF	C = 100, kernel = 'poly'	Balanced Accuracy: 52.8%
SVM - Count	C = .1, kernel = 'rbf'	Balanced Accuracy: 50%

Models with Sentiment		Balanced Accuracy: 50%
Random Forest - TFIDF	n_estimators: 200, min_samples_split: 2, min_samples_leaf: 1, max_depth: 50, bootstrap: False	Balanced Accuracy: 53.1%
Random Forest - Count	n_estimators: 200, min_samples_split: 5, min_samples_leaf: 4, max_depth: None	Balanced Accuracy: 50%
SVM - TFIDF	C = 100, kernel = 'rbf'	Balanced Accuracy: 68.7%
SVM - Count	C = .1, kernel = 'rbf'	Balanced Accuracy: 50%

Conclusion/ Final Model

- Final Model: SVM with Sentiment and TFIDF values for word features
- Ran into issues with small dataset when trying to predict outperform or underperform (277 observations)
- Only achieved 69% balanced accuracy score

	precision	recall	f1-score	support
0	0.83	0.38	0.53	13
1	0.90	0.99	0.94	71
accuracy			0.89	84
macro avg	0.87	0.69	0.73	84
weighted avg	0.89	0.89	0.88	84

Next Steps:

- Try to find API that allows access to articles prior to 2022 for further research
- Include more companies from varying sectors (more negative earnings reports)
- Test articles released 10 days prior to Fiscal Quarter End as opposed to Reporting Date
- Build out model that runs with only one company, but tracks articles/earnings over the last 10-20 years