Quality and Textual Analysis Strategy for Predicting Stock Price Change

Rachel Ahn, Matthew Tan, Kimberly Te, Andrew Matangaidze, and Jialu Sun

Overview

- Background
- Objective
- Dataset & Signals
- Baseline Results
- Next Steps

Background

Literature Review: Fundamentals

The Excess Returns of "Quality" Stocks: A Behavioral Anomaly (Bouchaud et al 2016)

- Systematic bias in analyst expectations when accounting for company quality
- Dataset: 136967 companies (global)

	Mistake		Forecast		Realized	
	(1)	(2)	(3)	(4)	(5)	(6)
Op. Cash Flows	-.063***	-.069***	-.012**	-.005	.05***	.064***
	(-6.2)	(-6.4)	(-2.4)	(-1.1)	(6.5)	(7.1)
Rolling volatility		.14***		.13***		-.0075
		(14)		(32)		(-1)
Book to Market		-.044***		-.011**		.033***
		(-3.8)		(-2.5)		(3.6)
r2	.27	.28	.24	.29	.26	.27
N	136967	133917	136967	133917	148975	145486
Month FE	YES	YES	YES	YES	YES	YES
Cluster	Firm	Firm	Firm	Firm	Firm	Firm

$$\text{mistake}_{i,t} = \beta \text{Quality}_{i,t} + \text{controls}_{i,t} + \epsilon_{i,t}$$

Literature Review: Text Analysis

On the Importance of Text Analysis for Stock Price Prediction (Lee et al., 2014)

- Dataset: 8k reports
- Features included unigram words and event categories
- Results showed promise but not concrete evidence for trading

Feature	B1	B2	Uni	NMF	E
Earnings surprise	√	√	√	√	√
Recent movements		√	√	√	√
Volatility index		√	√	√	√
Event category		√	√	√	√
Unigrams			√		√
NMF vector				√	√

Table 5: The list of features used in each model. B1: Baseline1, B2: Baseline2, Uni: Unigram model, NMF: NMF model, E: Ensemble model

System	Accuracy
Random guess	33.3
Majority class	34.9
Baseline1	49.4
Baseline2	50.1
Unigram model	54.4
NMF 50	54.7
NMF 100	55.4
NMF 200	55.3
Ensemble	55.5

Objective

Objective

- To predict the **weekly percentage change in stock price** using **quarterly fundamentals signals** and **daily textual signals** from news articles and 8k reports

Dataset & Signals

Dataset

- **Dataset:** Compustat
  - Stock price history
  - Fundamentals
  - Textual Analysis: Key Developments Dataset
- **Universe:** S&P1500 (2000 - 2020)
  - Currently only using (2000 - 2005)

Textual Analysis: Dataset

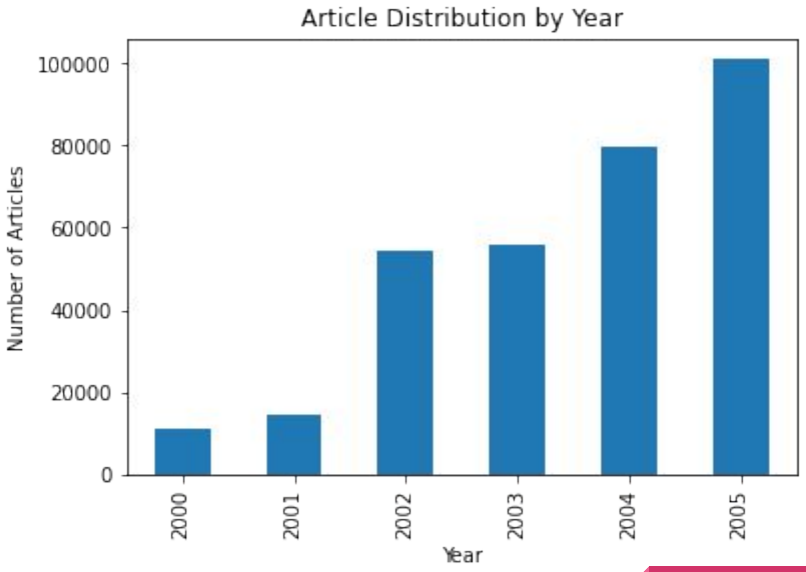
- **Dataset:** Capital IQ Key Developments
  - **Text:** Summaries of situations and events from news aggregators (e.g. financial articles), stock exchanges, regulatory websites (e.g. 8k reports), company websites (e.g. call transcripts)
  - **Events:** Categories of situation (e.g. bankruptcy, strategic alliances)
- **Features:**
  - Event type
  - Unigrams of words
  - Sentiment

Textual Data

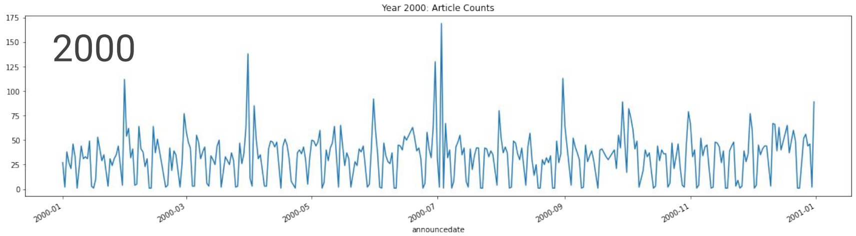
- **Pre-processing pipeline**
  - Tokening
  - Normalize text through removing stop words, numbers, names, punctuations etc
  - Lemmatizing, and vectorizing
  - Filter event categories using financial intuition
  - Aggregating and match textual data → weekly price time intervals

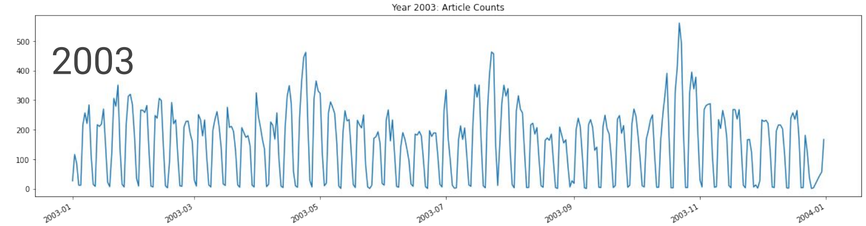
Textual Analysis: Article Frequency

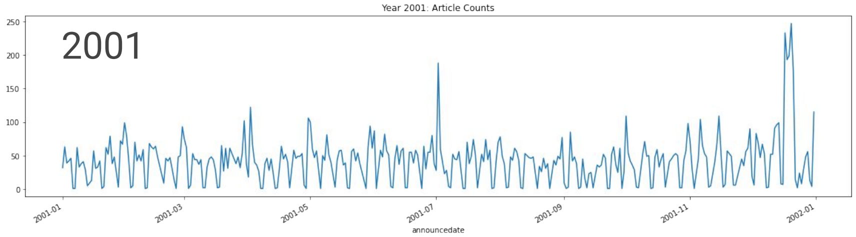
- **Time:** 2000-2005
- Increasing frequency of articles over time
- Articles scrapped from online sources
- Correlates with increased online usage

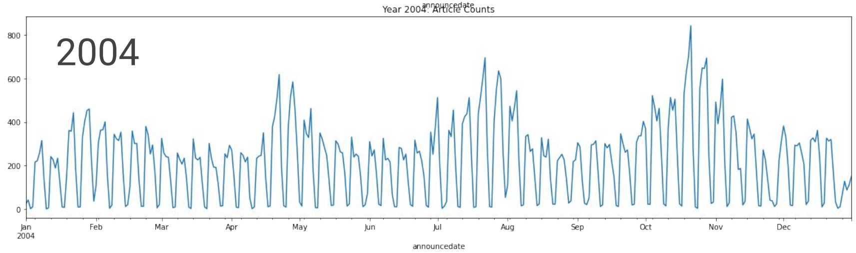


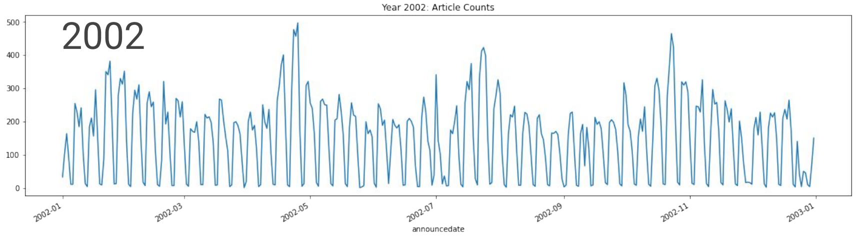
Textual Analysis: Article Counts over Time

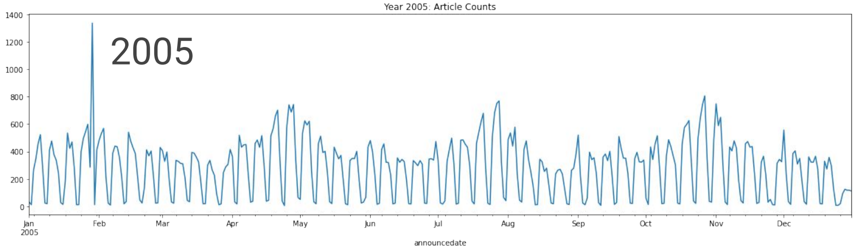








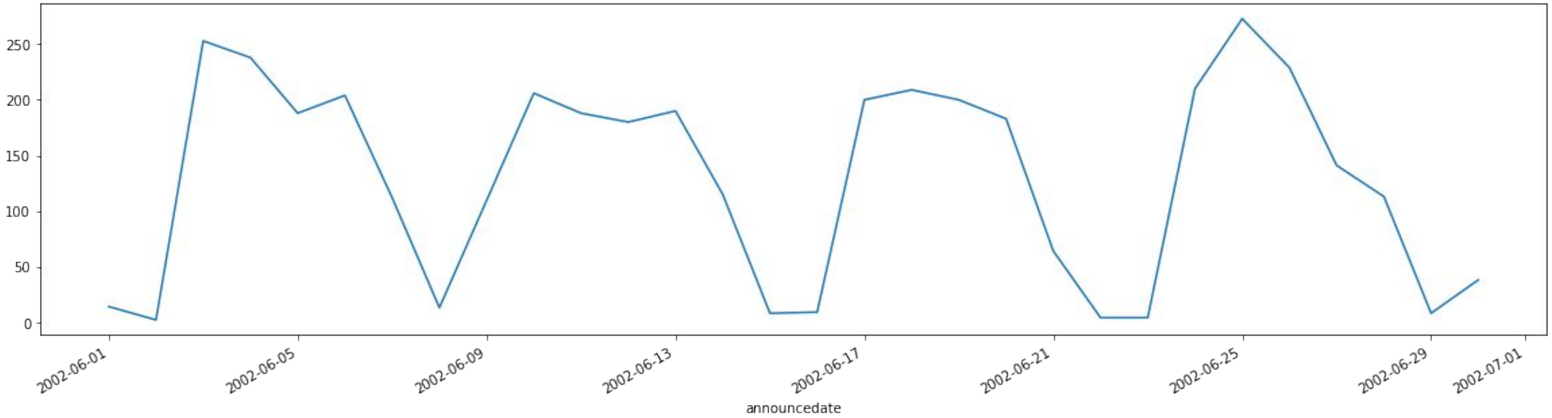




Textual Analysis: Time

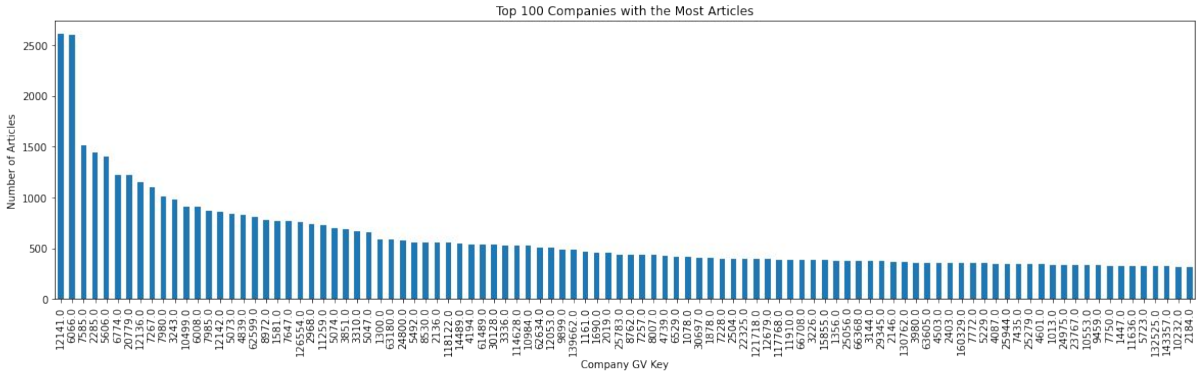
- Cyclic monthly trends
- 4 cycles per month, where drops occur on the weekends

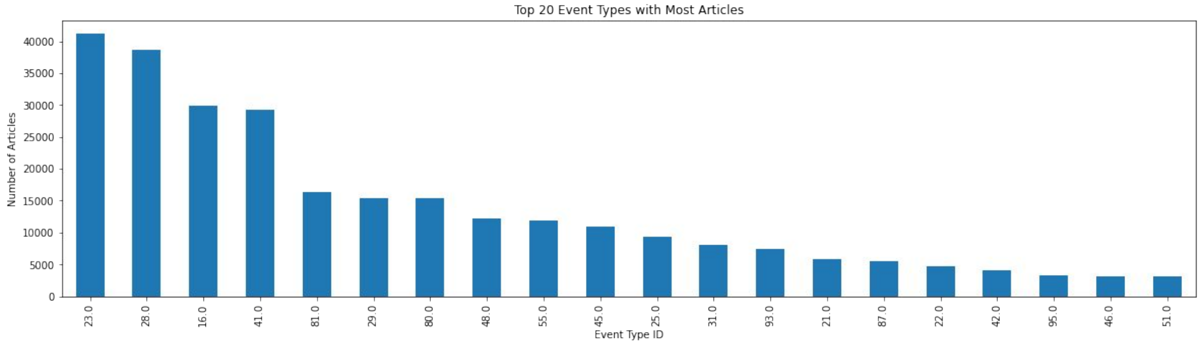
June 2002 Cycle



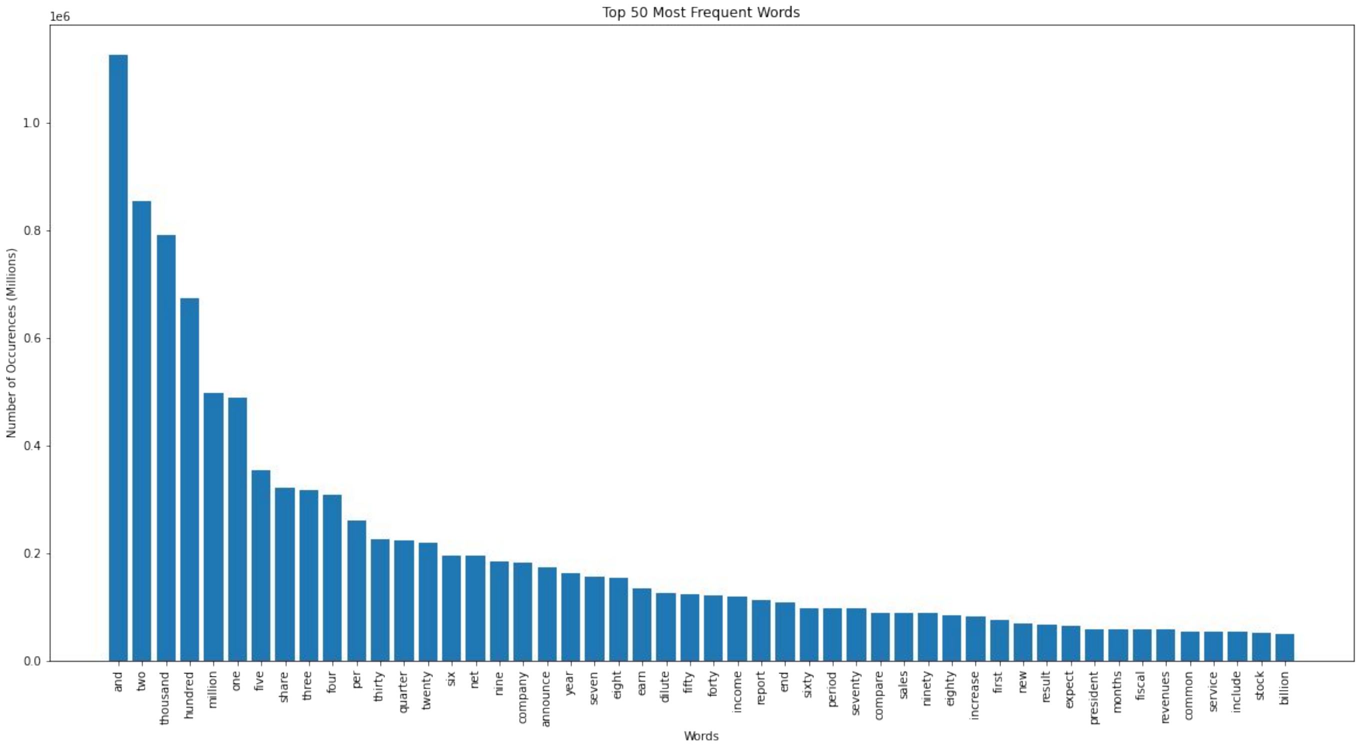
Textual Analysis: Event Types & Companies

- **Top Companies:** Microsoft, IBM
- **Top Event:** Client Announcements, Announcement of Earnings

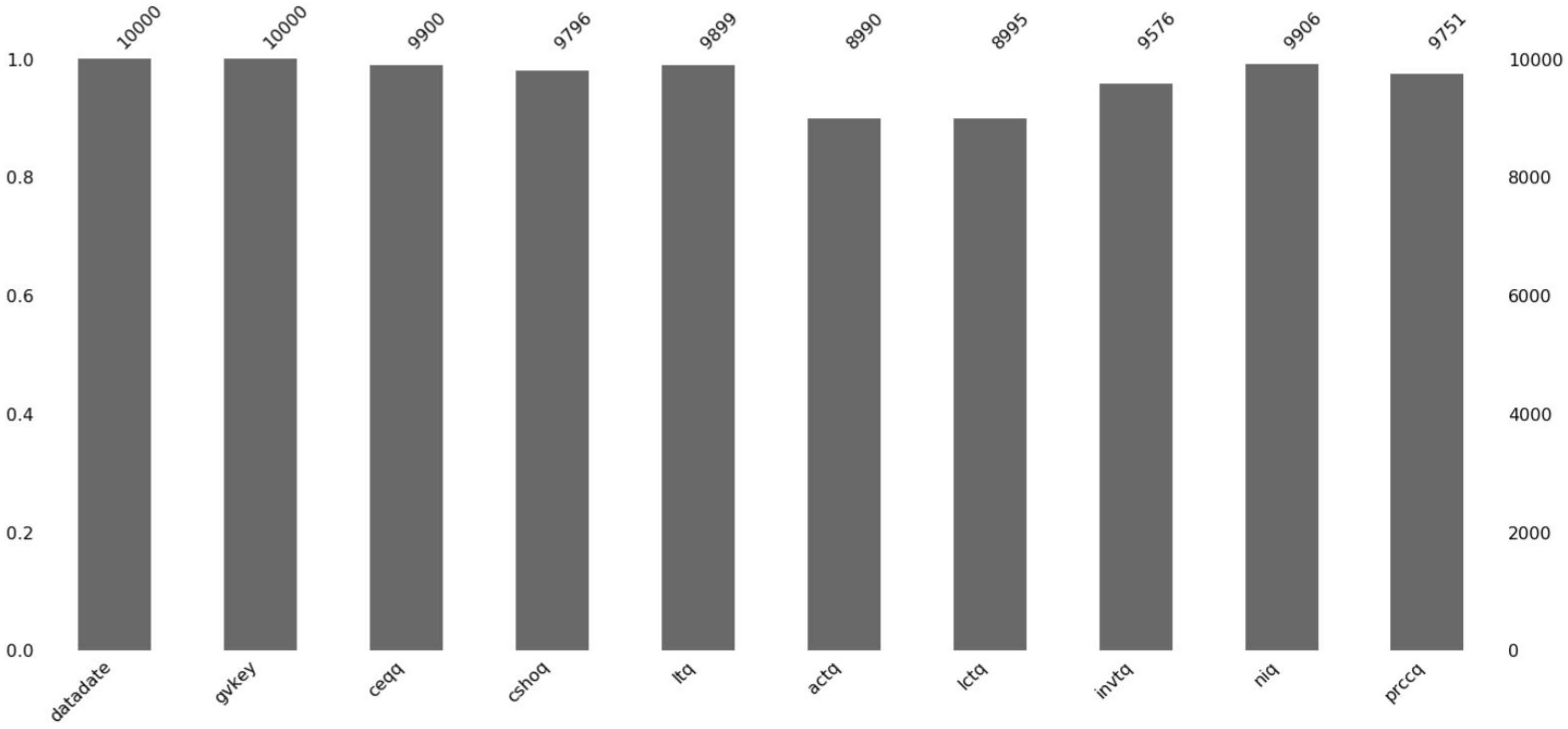




Textual Analysis: Unigram Frequency



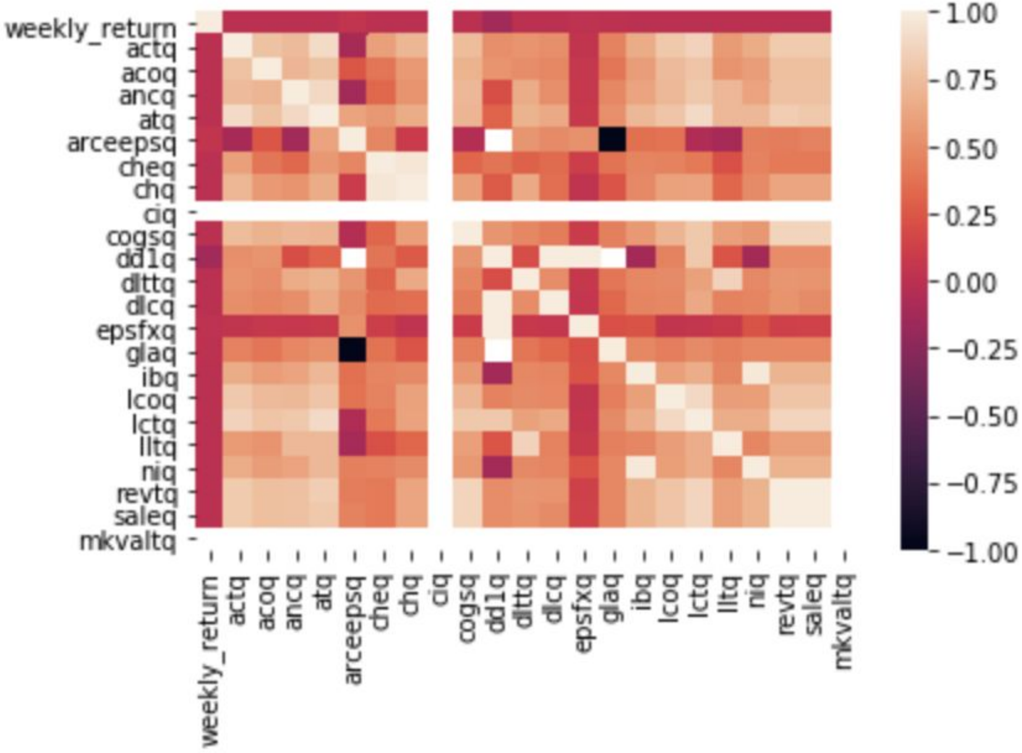
Fundamentals Data completeness



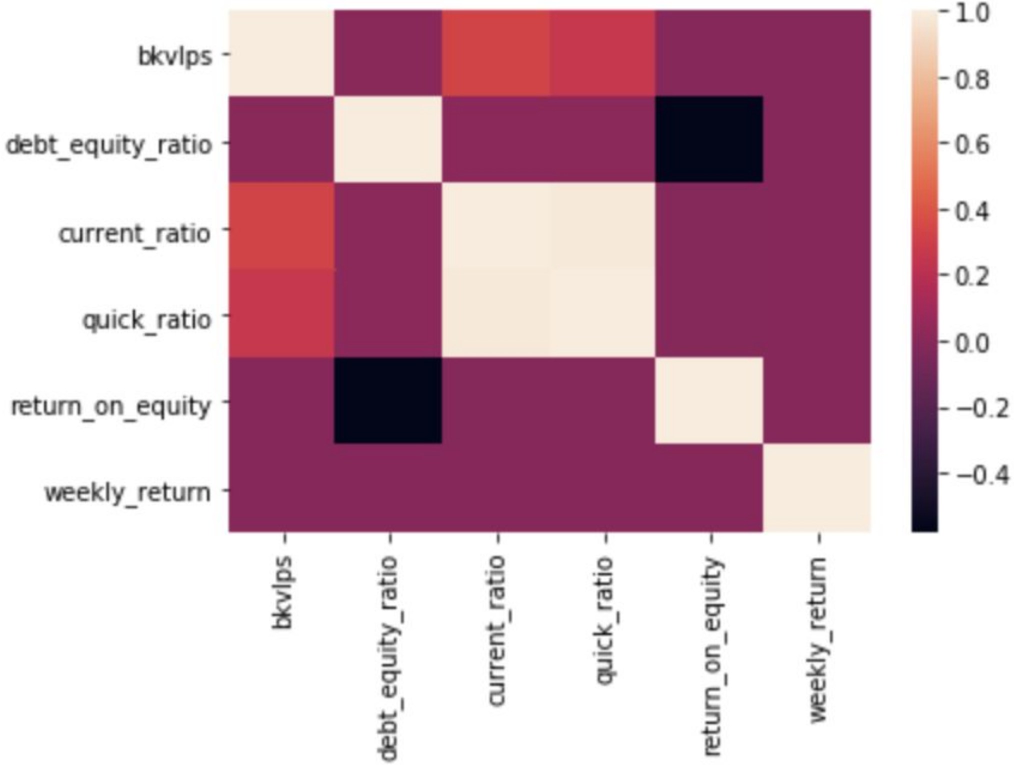
Fundamentals Data

- **Only using S&P500, 2000 - 20005 data**
- **Pre-processing pipeline**
  - Normalize stock prices, converting daily to weekly
  - Match quarterly fundamental features → weekly price time intervals
  - Filter based on column sparsity
  - Filter promising fundamental features using financial intuition

Fundamentals Data



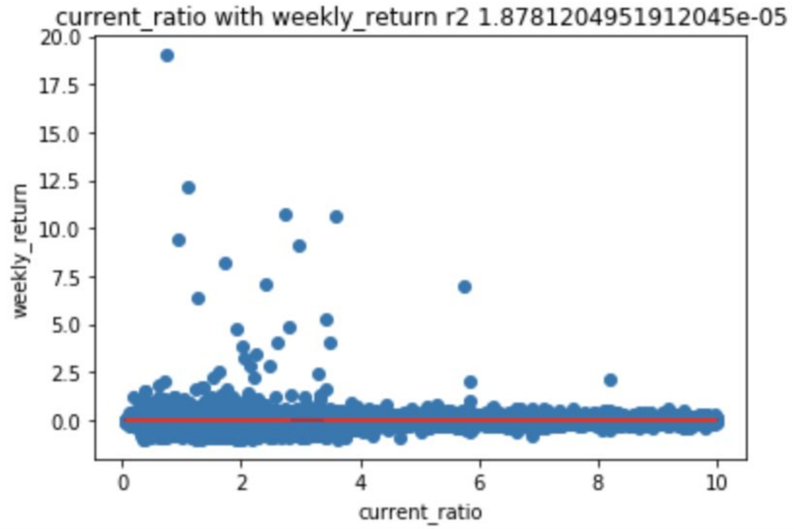
Fundamentals Data



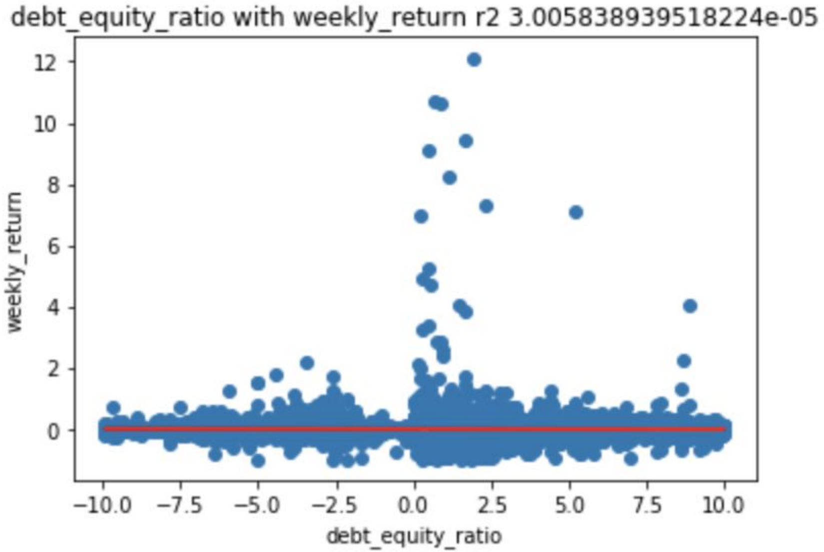
Baseline Results

Fundamentals Baseline

current_ratio



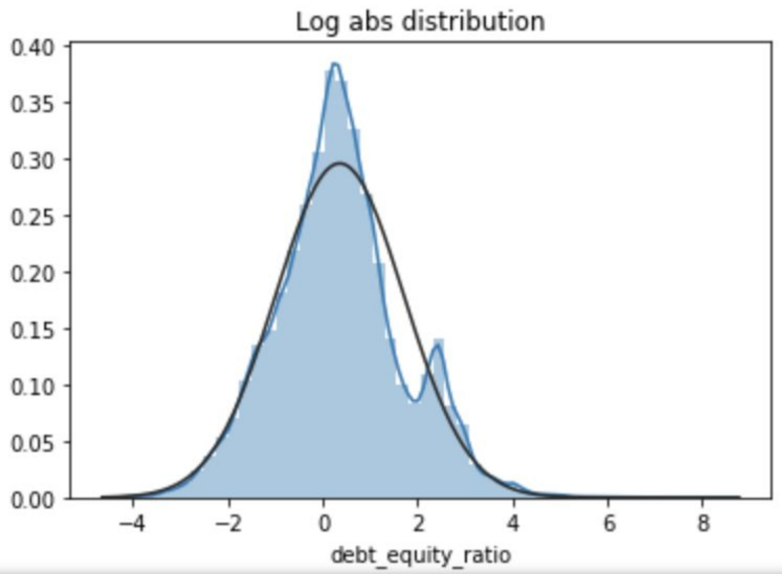
debt_equity_ratio



Fundamentals Baseline

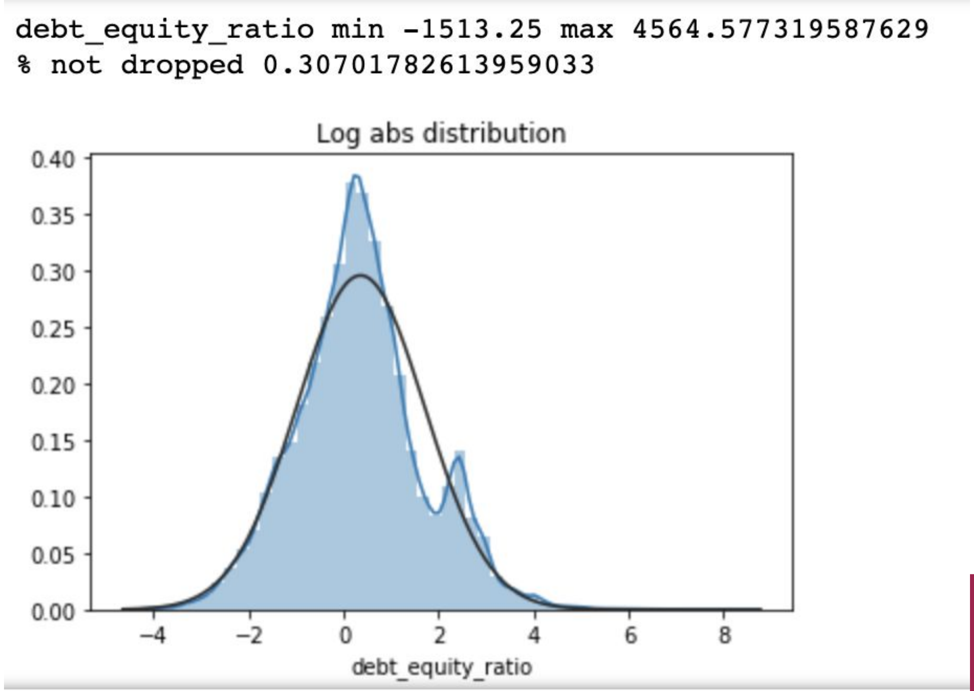
- Results mostly have low r^2 values.
- Analysis
    - Significant amount of data that was dropped during processing

```
debt_equity_ratio min -1513.25 max 4564.577319587629
% not dropped 0.30701782613959033
```



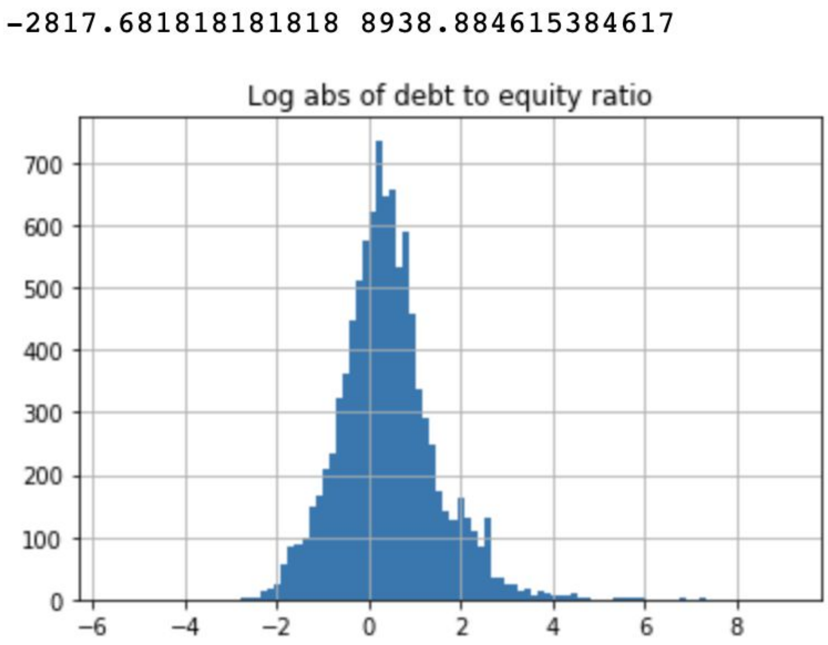
Fundamentals Baseline

- Results mostly have low r^2 values.
- Analysis
  - Significant amount of data that was dropped during processing
  - Despite normalizing with ratios, the range of values is large.
  - Traced the issue with matching with time series data.



Fundamentals Baseline

- Results mostly have low r^2 values.
- Analysis
  - Significant amount of data that was dropped during processing
  - Despite normalizing with ratios, the range of values is large.
  - Traced the issue with matching with time series data.



Next Steps

Next Steps: Fundamentals

- Build the data pipeline and find a good way to deal with missing data.
- Regress on some residual (ex. mistake) instead of % improvement
- Add macroeconomic variables since there is a “flight to quality” during high volatility regimes

Next Steps: Text

- Adding textual features from aggregated textual data across event categories to fundamentals numerical features for a complete model
- Textual features per event category plus fundamentals features model
- Examine SocialSent sentiment classifier to extract sentiment from text data
- Test and analyze experimental model results

Next Steps: Modeling

- Individually find signal between fundamental (quality) and text data.
  - We expect that simple regressions, random forest, bagging should be able to capture some signal (based on literature)
- Test the effectiveness of the regression on a rolling basis.