IPO Investment via Text Mining

Introduction to NLP @ Open University 2018b, Final Project
Andrew Kreimer
kreimer.andrew@gamil.com
https://github.com/algonell/IPOMiner

Abstract

Let us introduce IPOMiner: Python utilities to predict future performance of upcoming IPO (Initial Public Offering). The project is a collection of data-sets and Python code to perform Text Mining on raw SEC S-1 filings. The goal of this project is to apply Text Mining tools and techniques to spot investment opportunities in upcoming IPO. The project is intended for traders and researchers as potential source for alpha generation.

1 Purpose

The project incorporates Text Mining tools and techniques to find potential trade opportunities in upcoming IPO (by buying or selling equity). The project transforms S-1 forms to buy/sell signals via summarization, sentiment analysis, keywords analysis and classification (Jurafsky and Martin, 2014; Barnes et al., 2017; Khan et al., 2015). This project is targeting only English text corpora and US equities markets.

The purpose of this project is to apply Text Mining tools and techniques in order to spot investment opportunities in upcoming IPO (Initial Public Offering). This project is interesting as we try to transform biased texts into actionable trading signals and bridge the gap between accountants and retail traders.

Public companies with good fundamentals encourage investments. Contrary, companies with bad fundamentals or hidden details about failing products and unclear future will try to obfuscate it in their earnings reports to hold investors in. IPO (Initial Public Offerings) are different as we have no earnings report or true intrinsic data, rather hype and biased news. The single most objective document is the S-1 form created by accountants which is hard to understand.

2 Organization

The system is comprised of three main modules. The first module is responsible for IPO data retrieval. This module uses the EDGAR system provided by SEC (Securities and Exchange Commission) to download and store the raw S-1 documents (EDG; SEC). This module is also responsible for historical price data retrieval via Yahoo Finance (Yah). This module is mainly used for batch processing and should be run monthly.

The second module is responsible for Text Mining and model learning. The module loads, cleans, transforms and prepares our raw data for training. In addition this module learns and stores our predictive model: an IPO investment classifier. This module is also mainly used for batch processing and should be run monthly.

The third module is our production/real-time environment. The module is responsible for real-time classification of upcoming IPO. This module should be run daily or weekly. Our existing data-sets are also available for enhancements via Feature Engineering (Kag).

All of our components rely on each other in a sequential order. The raw data retrieval output is an input to our Text Mining module which is an input to our real-time classification module. Data retrieval is our heaviest module and usually takes 2 days to finish on a single machine. A zipped raw data collection for more than 2400 S-1 filings (5.5GB) is included and cuts the running times for assets prior to 2018 (IPO).

3 Example Input

The input of our system is a collection of raw S-1 filings documents indexed by the EDGAR sys-

DROPBOX, INC.

Our Business

Our modern economy runs on knowledge. Today, knowledge lives in the cloud as digital content, and Dropbox is a global collaboration platform where more and more of this content is created, accessed, and shared with the world. We serve more than 500 million registered users across 180 countries.

Dropbox was founded in 2007 with a simple idea: Life would be a lot better if everyone could access their most important information anytime from any device. Over the past decade, we've largely accomplished that mission—but along the way we recognized that for most of our users, sharing and collaborating on Dropbox was even more valuable than storing files.

Our market opportunity has grown as we've expanded from keeping files in sync to keeping teams in sync. Today, Dropbox is well positioned to reimagine the way work gets done. We're focused on reducing the inordinate amount of time and energy the world wastes on "work about work"—tedious tasks like searching for content, switching between applications, and managing workflows.

We want to free up our users to spend more of their time on the work that truly matters. Our mission is to unleash the world's creative energy by designing a more enlightened way of working.

We believe the need for our platform will continue to grow as teams become more fluid and global, and content is increasingly fragmented across incompatible tools and devices. Dropbox breaks down silos by centralizing the flow of information between the products and services our users prefer, even if they're not our own.

By solving these universal problems, we've become invaluable to our users. The popularity of our platform drives viral growth, which has allowed us to scale rapidly and efficiently. We've built a thriving global business with over 11 million paying users.

Our revenue was $603.8 million, $844.8 million, and $1,106.8 million in 2015, 2016, and 2017, respectively, representing an annual growth rate of 40% and 31%, respectively. We generated net losses of $325.9 million, $210.2 million, and $111.7 million in 2015, 2016, and 2017, respectively. We also generated positive free cash flow of $137.4 million and $305.0 million in 2016 and 2017, respectively, compared to negative free cash flow of $63.9 million in 2015.

Figure 1: Dropbox S-1/A Filing (SEC)

Company Name	Symbol	Market	Price	Shares	Offer Amount	Date Priced
AVALARA INC	AVLR	NYSE	$24	7,500,000	$180,000,000	6/15/2018
PUXIN LTD	NEW	NYSE	$17	7,200,000	$122,400,000	6/15/2018
VERRICA PHARMACEUTICALS INC.	VRCA	NASDAQ Global	$15	5,000,000	$75,000,000	6/15/2018
US XPRESS ENTERPRISES INC	USX	NYSE	$16	18,056,000	$288,896,000	6/14/2018
CHARAH SOLUTIONS, INC.	CHRA	NYSE	$12	7,352,941	$88,235,292	6/14/2018
FAR POINT ACQUISITION CORP	FPACU	NYSE	$10	55,000,000	$550,000,000	6/12/2018
GS ACQUISITION HOLDINGS CORP	GSAHU	NYSE	$10	60,000,000	$600,000,000	6/8/2018
MEIRAGTX HOLDINGS PLC	MGTX	NASDAQ Global Select	$15	5,000,000	$75,000,000	6/8/2018
AMBOW EDUCATION HOLDING LTD.	AMBO	NYSE MKT	$4.25	1,800,000	$7,650,000	6/1/2018

Figure 2: Upcoming IPO (NAS)

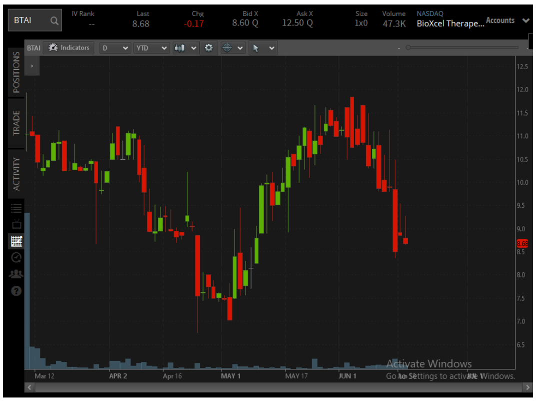
tem (EDG). S-1 filings are mandatory for private companies willing to go public. S-1 forms combine fundamental and financial data provided by accountants and should be objective and deterministic. Figure 1 shows an introduction of a raw filing.

The system uses raw IPO filings to create a unified data-set which contains multiple features and core meta-data. The system then builds a classifier for real-time prediction of upcoming IPO. Figure 2 shows a sample upcoming IPO at NASDAQ (NAS).

Finally, our system is used for real-time classification of upcoming IPO in order to spot potential investment opportunities. It is possible to buy or sell the underlying stock for various holding periods. Figure 3 shows two IPO of DBX and BTAI and their respective performance over a long-term holding period.

It is clear that DBX could be a profitable buy and hold, while BTAI could be a profitable short sell. The system provides classified directions and the underlying probabilities in order to discover upcoming IPO as potential trades similar to DBX and BTAI (tas; NAS).

(a) DBX



(b) BTAI

Figure 3: IPO Performance (tas)

Symbol	Company Name	Date Priced	Market	Offer Amount	Price	Shares	1D	1W	1M	3M
WQNI	WQN, INC.	2000-02-04	NASDAQ	35750000	13.00	2750000	-0.129856	-0.296122	0.088818	-0.667432
BBGI	BEASLEY BROADCAST GROUP INC	2000-02-11	NASDAQ	106175000	15.50	6850000	-0.058333	-0.133333	-0.175000	-0.250000
UTSI	UTSTARCOM HOLDINGS CORP.	2000-03-03	NASDAQ	180000000	18.00	10000000	0.512195	0.448171	1.042683	-0.134146
SLAB	SILICON LABORATORIES INC	2000-03-24	NASDAQ	99200000	31.00	3200000	0.033730	0.349206	0.190476	-0.142857
ALTH	ALLOS THERAPEUTICS INC	2000-03-28	NASDAQ	90000000	18.00	5000000	-0.116667	-0.054000	-0.391333	-0.300000
WBSN	WEBSENSE INC	2000-03-28	NASDAQ	72000000	18.00	4000000	0.409565	0.275362	-0.188406	-0.246377
MET	METLIFE INC	2000-04-05	New York Stock Exchange	2878500000	14.25	202000000	0.060345	0.051724	0.198276	0.500000
LPSN	LIVEPERSON INC	2000-04-07	NASDAQ	32000000	8.00	4000000	0.185714	-0.085714	-0.085714	-0.042857

Figure 4: Raw IPO Listings Data (Pedregosa et al., 2011)

Symbol	Neg Sent Signal compound	Neg Sent Signal neg	Neg Sent Signal neu	Neg Sent Signal pos	Pos Sent Signal compound	Pos Sent Signal neg	Pos Sent Signal neu	Pos Sent Signal pos	Mean Sent Len	1D	1W	1M	3M
AACC	0.750307	0.188291	-0.188526	-0.088264	0.824884	-0.047711	-2.180596	2.130875	-0.229744	0.002401	0.026411	0.073229	0.121248
AAT	0.393542	-0.595995	0.415479	0.748544	1.031044	-0.191374	0.296492	-0.236425	0.468211	-0.017185	-0.010683	-0.019508	-0.008825
ABR	2.255889	-0.648139	0.047402	1.835012	0.378669	0.187877	0.194437	-0.238990	0.029447	0.022444	-0.030923	-0.022444	-0.003990
ABTX	0.371174	-0.206152	-0.139155	0.980488	0.387190	0.206714	-2.462221	2.337360	-0.126546	0.027938	-0.008869	0.054989	0.004878
ACAD	-0.056967	0.067109	-0.039826	-0.099553	-0.135393	-0.211326	0.163203	-0.102851	-0.825017	-0.075269	-0.112903	-0.134409	-0.233871
ACFC	0.552216	0.287444	-0.384181	0.111178	-3.876276	0.948891	-3.870263	3.507450	-0.742610	0.036060	0.093228	0.130167	0.218118
ACIA	0.506249	-0.451198	0.276213	0.658924	0.380346	0.654197	0.595426	-0.752431	0.268327	0.068621	-0.011379	0.282759	1.258621
ACMR	0.093448	2.153342	-1.721236	-2.118655	-0.841036	0.580083	1.010444	-1.138132	0.273143	-0.264382	-0.308446	-0.168911	-0.351285

Figure 5: Enhanced IPO Listings (Pedregosa et al., 2011)

4 Text Mining

4.1 Data

Our Text Mining flow starts with raw NASDAQ listing data. The raw listing features include company name, pricing date, market, offer amount, initial price and number of shares. We have four target variables: 1D, 1W, 1M and 3M performances (one day, one week, one month and three months months respectively). Performance is calculated based on open prices only as retail traders do not get the best execution price or the predefined offering price (Narang, 2013). Figure 4 shows sample listings raw data.

This is our baseline data-set and from here we add more features via NLP and Feature Engineering techniques (Larose, 2006; Kag). Figure 5 shows enhanced data-set with sentiment features.

Another enhancement techniques that have been incorporated is text summarization. Document Summarization is the process of shortening documents with the effort of keeping the same variance of data and meaning (Jurafsky and Martin, 2014; Steinberger and Ježek, 2012; Allahyari et al., 2017; Lloret and Palomar, 2010). Following is a partial summary of DBX IPO S-1 form made with the gensim library (EDG; Řehůřek and Sojka, 2010):

> As such, in this registration statement we have taken advantage of certain reduced disclosure obligations that apply to emerging growth companies regarding selected financial data and executive compensation arrangements. Salesforce Ventures LLC has entered into an agreement with us pursuant to which it has agreed to purchase $100,000,000 of our Class A common stock.

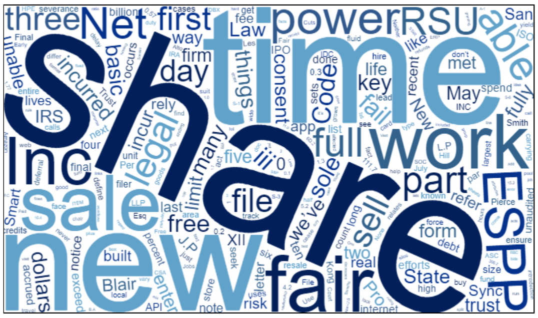


Figure 6: DBX WordCloud (wor)

Finally, keywords extraction also has been incorporated in our project due to high variance of word counts vs. meaningful keywords. Figure 6 shows word cloud visualization for the DBX raw S-1 form (EDG).

Contrary to the word cloud which is value count based, following is the list of keywords extracted in a more sophisticated manner using the gensim library (Řehůřek and Sojka, 2010). Note the difference of buzz-words and hype terms vs. actual meaningful words.

management	offering	taxing	cash
new	information	operates	addition
periodic	include	capitalization	sale
follows	stock	includable	informed
marketing	informational	including	secure
inform	news	tax	capitalizing
relating	offer	planned	manages
accountant	generated	provided	operate
accountability	includes	shared	generating
manage	relate	capitalized	manager
relation	certain	generates	dropboxers
followed	based	operating	expensive
accounting	planning	sharing	relates
company	offered	taxed	user
financial	security	generation	included
operational	financially	base	capitalizes
managing	related	providing	plan
managed	securing	service	share
capital	provides	general	additional
period	market	accounted	following
provide	generally	provider	account
dropbox	marketability	generate	business
operation	follow	operator	expense
financials	table	expensed	

For instance the word cloud contains new, RSU (Restricted Stock Units) and ESPP (Employee Stock Purchase Plan) words which target the company employees and are more of a show-off. In contrast, words like providing, capitalization and generation have more meaning. Immediate issue with the keywords extraction is domain knowledge integration. The words dropbox and dropboxers get a key-role meaning, although dropbox is a company name and dropboxers are essentially users. This issue is hard to spot systematically and must be addressed manually.

4.2 Machine Learning

Our project incorporates various Machine Learning tools and techniques. In terms of data preprocessing we use one-hot encoding to avoid ordered categorical values such as month, market and quarter. In addition, we incorporate standardization to normalize our data points as share amount and offering price vary across listings (Larose, 2006; Kag).

Once a clean and normalized data-set is available, we integrate Ensemble Classification. Our problem is a supervised binary classification task. Our ensemble is comprised of Logistic Regression and Random Forest classifiers (Larose, 2006; Kag). Ensemble Classification is a widely used methodology to improve simple classifiers generalization and boost performance (Kag).

4.3 Supervised Learning

Our problem is a supervised learning binary classification (Larose, 2006). As mentioned before, we collect four target variables (1D, 1W, 1M and 3M) performances based on open prices. We collect percent changes as open prices of respective period since the the first trading day, minus the first trading day opening price. The percentages are then converted to binary variables as positive performance and negative performance (1 and 0 respectively). Note that predicting performance as continuous values, which is known as regression problem in Machine Learning, has not been addressed in this project.

4.4 Randomness

Standard Machine Learning problems provide idealistic world problems in terms of perfect train-test splits and out of sample validation (Larose, 2006; Kag). Real-life applications on the other hand usually have lower generalization and predictive power, particularly in financial markets. The most common issue is the non-repeated data distributions and existence of market momentum (Narang, 2013; Johnson, 2010).

4.5 Hyper-Parameter Tuning

Hyper-parameter tuning is the process of fixing various parameters for our models within the train-test split in order to increase generalization and predictive power. Hyper-parameter tuning is widely used in Machine Learning tasks (Kag). Contrary, in financial markets it is always advised to keep the models and learning as simple as possible (Kag; Narang, 2013; Johnson, 2010).

5 Discussion

Previous works have shown significant integration of news related to the IPO and their influence on performance. IPO S-1 filings tend to be diversified as various companies make them. The addition of more differences and non-unified data source increases noise. Eventually, the tone and biased jargon tend to influence retail investors (Feuerriegel et al., 2014).

In addition, previous works in the field show various applications for subsets of financial documents. For instance annual reports tend to present wider and more reliable company overview (Kloptchenko et al., 2004). On the other hand, news, tone and biased jargon tend to influence retail investors (Feuerriegel et al., 2014). Another important signal boosting of buying or selling is provided by 8-K documents (Lee et al., 2014). Finally, a brief correlation is described in IPO filings and higher management opinions (Deokar and Tao, 2015).

5.1 Management Data Integration

A research has shown the ability to integrate management opinions and summaries to evalute upcoming IPO performance (Deokar and Tao, 2015). The research showed how genuine information about the companies can be extracted by combining final IPO statements and management discussion analysis. The combination provides boosted sentiment signal and better predictive modeling. In our project we have been relying only on raw IPO S-1 filings and no external related data.

5.2 Continuous Sentiment

A research of social media impact and performance of equities have shows a significant momentum effect (Makrehchi et al., 2013). The research analyzed aggregated twitter feeds grouped by equities performance (positive and negative) to reveal patterns. They showed that drastic performance of companies tend to continue the following trading days. Regarding IPO investments the moment effect is even more significant and thus can be another potential data source for our problem.

Another research showed text analysis of 8-K documents and the ability to classify performance. 8-K documents are legally required documents for major corporate changes such as bankruptcy and CEO changes. Text analysis of major financial data shows a high correlation to discrete class variable (going up, going down and staying the same) for the proceeding week (Lee et al., 2014). In our project we have not been incorporating social media signals nor 8-K documents.

5.3 Hype Analysis

Previous works in the fields have been analyzing hype and biased impact of financial news on equity performance (Johnson, 2010; Narang, 2013). A research showed how biased and manipulated language can mislead investors and intensify post IPO performance. In addition, the researched showed how language and jargon can tilt risk to reward ratios as seen by investors (Feuerriegel et al., 2014). Another research showed how opinion mining can help businesses to improve and increase sells. An important aspect of this research is messages and comments analysis regarding the topic explored. The research shows how user opinion is important for boosting our predictive modeling (Chen and Zimbra, 2010).

5.4 Language Specificity

A similar research has been conduced for Turkish companies (Bastı et al., 2015). The researched showed how listing data and various over-pricing and under-pricing indicators influence first week performance of IPO. Turkish jargon and format allowed better predictive modeling. Similarly to our research the analysis is based on a single market place. On the other hand, the research is dedicated to a small market with mainly Turkish companies, whereas the US market is a major market place for companies all over the world (Johnson, 2010).

5.5 Cluster Analysis

In contradiction to previous works in the filed, we have tried to incorporated clustering features. Clustering via KNN (K Nearest Neighbours) is widely used strategy for feature improvement (Kag; Larose, 2006). By analyzing our data-set and looking for nearest neighbours we can add new sets of features to existing features (horizontal concatenation). This enhancement adds more meta-data for existing features. Although this methodology usually improves classifiers, in our problem it did not help to improve. Figure 7 shows our experiments with and without clustering features.

		1D	1W	1M	3M
AUC	LR	0.516569	0.516569	0.516569	0.516569
	RF	0.538337	0.538337	0.538337	0.538337
f1	LR	0.735849	0.735849	0.735849	0.735849
	RF	0.726368	0.726368	0.726368	0.726368
log loss	LR	0.663721	0.663721	0.663721	0.663721
	RF	0.657215	0.657215	0.657215	0.657215

(a) With Clustering Features

		1D	1W	1M	3M
AUC	LR	0.544509	0.544509	0.544509	0.544509
	RF	0.578622	0.578622	0.578622	0.578622
f1	LR	0.732673	0.732673	0.732673	0.732673
	RF	0.735751	0.735751	0.735751	0.735751
log loss	LR	0.655687	0.655687	0.655687	0.655687
	RF	0.640521	0.640521	0.640521	0.640521

(b) Without Clustering Features

Figure 7: Clustering Features (Pedregosa et al., 2011)

5.6 Word2vec Integration

Our best improvement was achieved by incorporating keywords analysis and cross distances between companies. As we showed, simple cluster analysis is too wide to provide some kind of signal. Word2vec models essentially transform words to n-dimensional matrices (Jurafsky and Martin, 2014). The sparse representation better describes words relations. We then apply a series of aggregations on small keywords groups and eventually create better relationship features between companies and their relative performance. Figure 8 shows the combined Word2Vec build from all IPO keywords.

5.7 Multiple Target Variables

The previous works that has been covered usually incorporate single class variable and evalute the entire performance on a single outcome. In this work we classify performance and provide the underlying probabilities for multiple periods. The combination of multiple products in different periods of expiration can out-perform single direc-

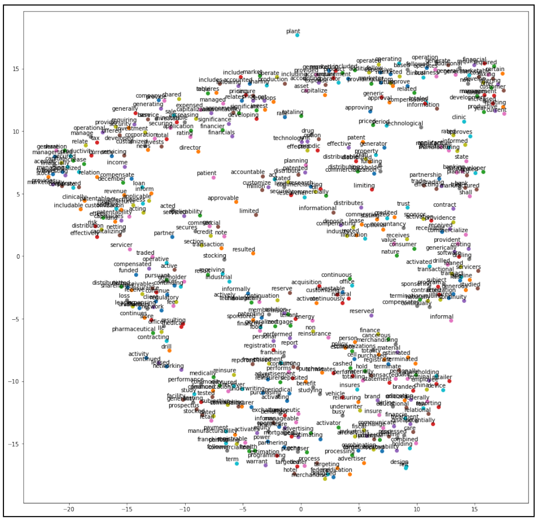


Figure 8: Keywords Word2Vect (wor)

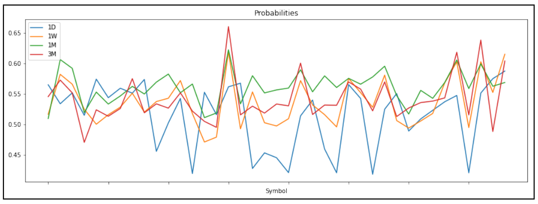


Figure 9: Predictions Probability Distribution (Pedregosa et al., 2011)

tion execution (Johnson, 2010). Figure 9 shows predictions and probabilities for upcoming IPO (June/July 2018). As expected, the mean probability is low (0.6) and the underlying predictions are ambiguous.

The signals are hard to integrate or trade which implies that most of the time the signals will be poor, and occasionally higher than mean with actionable insights. Combining all signals can increase our odds similar to the law of large numbers principle, but is hard to accomplish for a retail trader (Narang, 2013). Figure 10 shows concrete predictions for upcoming IPO (June/July 2018) (IPO).

Note how ambiguous the predictions can be for various assets. For instance ALGRU and GSAHU have probabilities over 0.55 for all periods while MGTX and NEW have no common direction.

6 Evaluation

6.1 Criteria

We incorporate mainly quantitative evaluation techniques. As we are dealing with a binary classification problem, we incorporate metrics such as AUC, f1 and logarithmic loss. AUC (Area Under the Curve) and f1 score provide balanced combinations of precision and recall. Logarithmic loss provides a more sophisticated analysis of our confidence in predictions and the underlying probabilities (Larose, 2006; Kag). The evaluation is performed for each one of the classifiers we use (Logistic Regression and Random Forests) (Larose, 2006).

We run a set of experiments from baseline to state of the art in order to constantly improve our models. We provide ROC (Receiver Operating Characteristic) curves and logarithmic loss for each instance in our test set (Larose, 2006). In addition we use feature importance visualizations to interpret most insightful features using scikit-learn and XGBoost (Pedregosa et al., 2011; Chen and Guestrin, 2016).

Symbol	1D	1W	1M	3M
VRCA	0.550516	0.506244	0.547206	0.512838
NEW	0.489007	0.493931	0.516916	0.526819
AVLR	0.508936	0.505516	0.555776	0.535742
CHRA	0.523042	0.517920	0.542729	0.538168
USX	0.537083	0.568160	0.568746	0.543398
FPACU	0.547410	0.602563	0.605661	0.618350
MGTX	0.420460	0.494677	0.558710	0.515455
GSAHU	0.551328	0.602748	0.599222	0.638340
AMBO	0.574962	0.552362	0.562572	0.488216
ALGRU	0.587435	0.615117	0.568686	0.604006

Figure 10: Predictions Probabilities (Pedregosa et al., 2011)

6.2 Simple Performance

We start our evaluation with a collection of AUC, f1 and logarithmic loss for a test set which was held out via a 0.8/0.2 train/test split respectively. The splits allow us better validation of our models, plus we have been incorporating no hyperparameter tuning to avoid over fitting. As mentioned before, our data is time-series based and due to a lack of massive amounts of train data, we do not incorporate k-fold cross validation. Figure 11 shows our first stage evaluation.

6.3 Detailed Performance

We proceed with a deeper analysis of our classifier performance by incorporating ROC (Receiver

		1D	1W	1M	3M
AUC	LR	0.479237	0.479237	0.479237	0.479237
	RF	0.512065	0.512065	0.512065	0.512065
f1	LR	0.696864	0.696864	0.696864	0.696864
	RF	0.695971	0.695971	0.695971	0.695971
log loss	LR	0.708928	0.708928	0.708928	0.708928
	RF	0.7011	0.7011	0.7011	0.7011

(a) Baseline

		1D	1W	1M	3M
AUC	LR	0.517241	0.517241	0.517241	0.517241
	RF	0.553196	0.553196	0.553196	0.553196
f1	LR	0.745455	0.745455	0.745455	0.745455
	RF	0.746411	0.746411	0.746411	0.746411
log loss	LR	0.670979	0.670979	0.670979	0.670979
	RF	0.663169	0.663169	0.663169	0.663169

(b) Sentiment Analysis

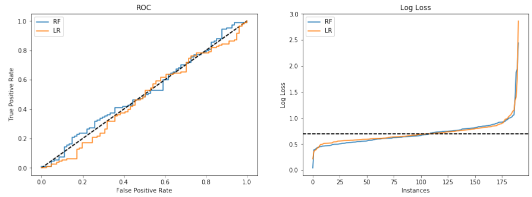
		1D	1W	1M	3M
AUC	LR	0.539319	0.539319	0.539319	0.539319
	RF	0.530908	0.530908	0.530908	0.530908
f1	LR	0.681081	0.681081	0.681081	0.681081
	RF	0.744186	0.744186	0.744186	0.744186
log loss	LR	0.676177	0.676177	0.676177	0.676177
	RF	0.661292	0.661292	0.661292	0.661292

(c) Summarization

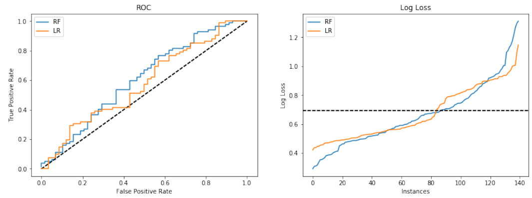
		1D	1W	1M	3M
AUC	LR	0.544509	0.544509	0.544509	0.544509
	RF	0.578622	0.578622	0.578622	0.578622
f1	LR	0.732673	0.732673	0.732673	0.732673
	RF	0.735751	0.735751	0.735751	0.735751
log loss	LR	0.655687	0.655687	0.655687	0.655687
	RF	0.640521	0.640521	0.640521	0.640521

(d) Keywords

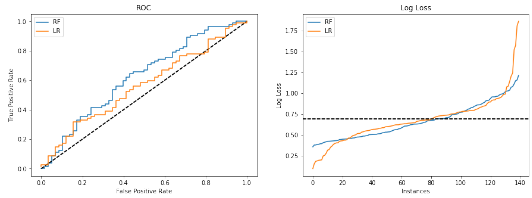
Figure 11: AUC, f1 and log loss (Pedregosa et al., 2011)



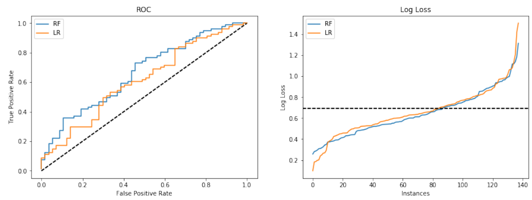
(a) Baseline



(b) Sentiment Analysis



(c) Summarization



(d) Keywords

Figure 12: ROC and logarithmic loss for instances (Pedregosa et al., 2011)

Operating Characteristic) and logarithmic loss for all instances in the test-set. ROC plots show deeper relations between precision and recall by providing various thresholds and respective precision/recall values. Classifiers use 0.5 as a default cut-off point for class labeling, which sometimes make the decision not reliable, particularly in our problem with $AUC < 0.6$. The collection of logarithmic loss values for our test-set shows the general confidence of our predictions and the respective probabilities. Figure 12 shows ROC and multiple log loss values for our test-set. Note the slight improvements over experiments both in terms of ROC and higher AUC and more instances having log loss values less than 0.693 (a coin toss).

6.4 Feature Importance

Finally, we present feature importance evaluation for each experiment. We have been incorporating

XGBoost feature importance utilities to rank the newly added features. Note that this evaluation is independent of our training and testing as different classifiers are used. Figure 13 shows the feature importance for our experiments. Note how newly added features always add more signal.

7 Lessons Learned

7.1 Raw Data

In terms of raw S-1 filings data, most of the IPO disappear within the next couple of years since the offering. From raw data scraping since 2000 there were 3500 IPO announcements on NASDAQ IPO listings but only 2400 concrete S-1 filings. Concrete data in terms of performance exists for less than 800 (out of the 2400) which means most of the companies do not survive going public. Another filtering issue is that finally a lot of companies don't get listed at all (more listings then prices and performance data). Essentially we have a 80% dropout from listing to survival.

In terms of data parsing and transformations, parsing and reshaping HTML documents is a hard task. Due to various tags, styles, left parts of scripts and images within the HTML body. This ambiguity usually creates differences in tokenized sentences and sometimes even words. The result is ambiguous parsing and Text Mining. Another issue is the data size as it is difficult to validate the results manually.

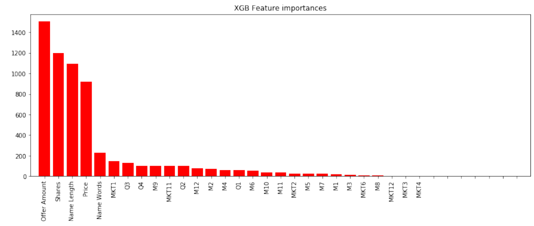
7.2 Summarization

Text summarization is a heavy memory and time task. The project showed 5% dropout due to out of memory and long calculations times (even using Google Compute machines). Another issue is the lack of state of the art summarization in the financial field. Most of the existing projects target user review summaries and tweet feeds. There are no reliable summarization projects in the professional financial world.

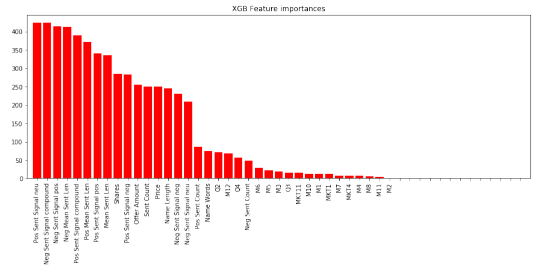
8 Difficulties

8.1 Data Retrieval

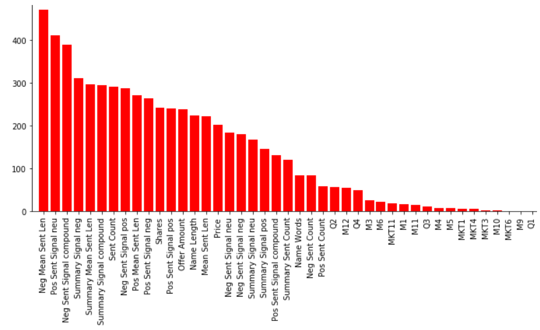
Although public S-1 forms our published by SEC, retrieving, storing and transforming them took most of the time. Instead of accessing the raw data via HTTP requests, the S-1 forms were downloaded and stored on disk. Next, the historical performance is another batch process for retriev-



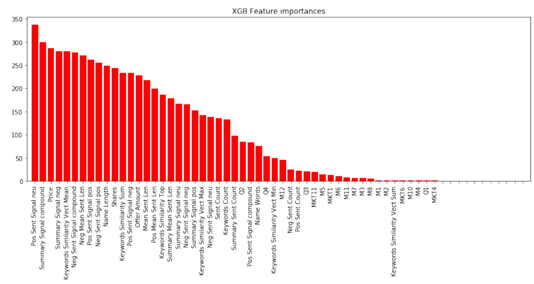
(a) Baseline



(b) Sentiment Analysis



(c) Summarization



(d) Keywords

Figure 13: Feature Importance (Chen and Guestrin, 2016)

ing historical price data and calculating target variables (1D, 1W, 1M and 3M performance). Another long running batch processes are summarization and keywords extraction which required GCE (Google Compute Engine) assistance due to out of memory issues with particularly large IPO.

8.2 Amendments

Most S-1 filings are followed by S-1/A amendments (and sometimes up to 10 more amendments). This data has not been incorporated in this project. We always chose the most up to date filing. Changes in S-1 filings could improve the long term inspection and opinion of the exchange about the company.

8.3 Repetitive Words

S-1 filings have common keywords that appear over 300 times in most of the documents. The following is a short list of the common words.

share	common
stock	marketing
company	development
month of IPO	previous year to IPO
believe	

Those words can mislead to some importance of particular months. As shown in the baseline model, month or quarter of IPO has no predictive power. Another problem that accountants add many positive words but usually with negative numbers. The right approach would be standardizing all of the S-1 filings to remove the common words, although it is computationally hard.

8.4 Numerical Data

Special care must be taken dealing with numerical data in our S-1 filings. Tables with current cash flow, debt, holdings and company effective risk to reward are hard to integrate. Tables tend to diverse with different colors, styles and shapes. This data must be incorporated manually.

8.5 Domain Knowledge

Many raw data files have implied terminology related to the relevant IPO. In the Dropbox example we have seen dropboxers extracted as a noun or meaningful keyword although essentially it should be replaced with users. It will be interesting to incorporate domain knowledge (both financial and accounting) to our models such as the NLTK Reuters corpus for financial news analysis (NLP).

References

5 heroic tools for natural language processing. https://goo.gl/UwECfQ.

Electronic data gathering, analysis, and retrieval system. https://www.sec.gov/edgar.shtml.

Free online word cloud generator and tag cloud creator. https://www.wordclouds.com/.

Ipominer. https://github.com/algonell/IPOMiner.

Kaggle. https://www.kaggle.com/.

National association of securities dealers automated quotations. https://www.nasdaq.com/.

tastytrade. https://www.tastytrade.com/.

U.s. securities and exchange commission. https://www.sec.gov/.

Yahoo finance. http://finance.yahoo.com/.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: A brief survey. *arXiv preprint arXiv:1707.02268*.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2017. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. *arXiv preprint arXiv:1709.04219*.

Eyup Bastı, Cemil Kuzey, and Dursun Delen. 2015. Analyzing initial public offerings' short-term performance using decision trees and svms. *Decision Support Systems*, 73:15–27.

Hsinchun Chen and David Zimbra. 2010. Ai and opinion mining. *IEEE Intelligent Systems*, 25(3):74–80.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Amit Deokar and Jie Tao. 2015. Text mining for studying managements confidence in ipo prospectuses and ipo valuations.

Stefan Feuerriegel, Jonas Schmitz, and Dirk Neumann. 2014. What matters most? how tone in initial public offering filings and pre-ipo news influences stock market returns.

Barry Johnson. 2010. *Algorithmic Trading & DMA: An introduction to direct access trading strategies*, volume 200. 4Myeloma Press London.

Dan Jurafsky and James H Martin. 2014. *Speech and language processing*, volume 3. Pearson London:.

Aamera ZH Khan, Mohammad Atique, and VM Thakare. 2015. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)*, page 89.

Antonina Kloptchenko, Tomas Eklund, Jonas Karlsson, Barbro Back, Hannu Vanharanta, and Ari Visa. 2004. Combining data and text mining techniques for analysing financial reports. *Intelligent systems in accounting, finance and management*, 12(1):29–41.

Daniel T Larose. 2006. *Data mining methods & models*. John Wiley & Sons.

Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. 2014. On the importance of text analysis for stock price prediction. In *LREC*, pages 1170–1175.

Elena Lloret and Manuel Palomar. 2010. Challenging issues of automatic summarization: relevance detection and quality-based evaluation. *Informatica*, 34(1).

Masoud Makrehchi, Sameena Shah, and Wenhui Liao. 2013. Stock prediction using event-based sentiment analysis. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01*, pages 337–342. IEEE Computer Society.

Rishi K Narang. 2013. *Inside the Black Box: A Simple Guide to Quantitative and High Frequency Trading*, volume 883. John Wiley & Sons.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Josef Steinberger and Karel Ježek. 2012. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275.