

Ask EDGAR:

**The Informational Content of Mutual Fund
Prospectuses**

Simona Abis and Anton Lines

Columbia Business School

Motivation

- How do investors allocate their capital across different asset managers?
 - ▶ Prior research mostly explained their decisions looking at past returns and other “**hard information**”, delivering low explanatory power
 - ▶ The SEC has been urging investors to be weary of past performance and to read prospectuses carefully when making investment decisions



U.S. SECURITIES AND
EXCHANGE COMMISSION

Mutual Fund Investing: Look at More Than a Fund's Past Performance

May 8, 2007

You can't open a newspaper or read a magazine without seeing ads promoting the stellar performance of "hot" mutual funds. But **past performance is not as important as you may think**, especially the short-term performance of relatively new or small funds. As with **any investment, a fund's past performance is no guarantee of its future success**. Over the long-term, the success (or failure) of your investment in a fund also will depend on factors such as:

So, look at **more than the fund's past performance when making your investment decisions**. Read the **fund's prospectus and shareholder reports**, and consider these tips:

<https://www.sec.gov/reportspubs/investor-publications/investorpubsmfperformhtm.html>

Motivation continued...

- **Are prospectuses informative?**

- ▶ Investors also have access to “soft information”
 - ★ e.g. marketing materials, regulatory disclosures, in-person meetings
- ▶ Prospectuses are the main document containing funds information
 - ★ likely proxying for other forms of communication
- ▶ The SEC has been investing in
 - ★ Educating investors about prospectuses
 - ★ Monitoring their fair disclosure

BUT

- ▶ Disclosure requirements have been shown to be very costly for funds
- ▶ The SEC has not systematically shown their usefulness
- ▶ Investors might not be paying attention to them

This Paper

- **Question: are prospectuses informative above and beyond what can be learned from “hard information”?**
 - ▶ **DATA:**

Collect and categorize all textual information disclosed by mutual funds through the EDGAR system

 - ★ Developed a comprehensive parsing algorithm which allows for automatic collection, mapping and parsing of historical filings
 - ★ Focus on active equity mutual funds in the US
 - ▶ **DESCRIPTIVE:**

Characterize prospectuses' content by conditioning on the same regulatory questions across funds and over time

 - ★ Length, complexity, sentiment
 - ★ Clustering (unsupervised learning)
 - ▶ **ANALYSIS:**
 - ★ Supervised learning to predict funds' likelihood to incur in agency-like behaviours (an example: risk-shifting)
 - ★ Work in progress:
 - Prediction of likelihood of legal actions
 - Unsupervised learning to predict future return distribution

- **Determinants of mutual fund flows:** Sirri & Tufano (1998), Jain & Wu (2000), Del Guercio & Tkac (2002), Barber, Odean & Zheng (2003), Berk & Green (2004), Del Guercio & Tkac (2008), Ivković & Weisbenner (2009), Gennaioli Shleifer & Vishny (2015)
- ▶ We'll provide new insights using variables extracted from soft information
- **Risk shifting:** Huang, Sialm and Zhang (2011); Chevalier and Ellison (1997); Brown, Harlow and Starks (1996); Ha and Ko (2017)
- ▶ We'll use supervised learning applied to mutual fund prospectuses to predict funds risk-shifting behavior

- **Textual analysis and machine learning:** Subramanian, Inslay & Blackwell (1993), Philipot & Johnson (2007), Tetlock (2007), Tetlock, Saar-Tsechansky, Macskassy (2008), Manela and Moreira (2017), Abis (2018), Ryans (2018), Kelley, Manela and Moreira (2018)
- ▶ We'll apply supervised and unsupervised learning to extract signals from mutual fund prospectuses

- **SEC regulatory changes and EDGAR usage:** Johnson (2004), Agarwal, Mullally, Tang & Yang (2015), Agarwal, Vashishtha & Venkatachalam (2017), Gao and Huang (2018)
- ▶ We'll exploit regulatory changes to study the reaction of flows to cross-sectional differences in mandatory disclosures

Roadmap

1 Introduction

2 Data

3 Descriptive Analysis

- Length, Complexity, Sentiment
- Clustering

4 Empirical Analysis

- Risk Shifting
- Work in progress...

5 Future Research

6 Conclusion

Roadmap

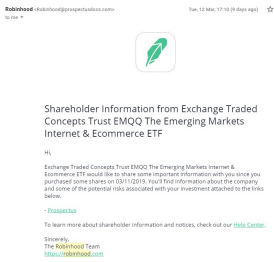
- 1 Introduction
- 2 Data
- 3 Descriptive Analysis
 - Length, Complexity, Sentiment
 - Clustering
- 4 Empirical Analysis
 - Risk Shifting
 - Work in progress...
- 5 Future Research
- 6 Conclusion

Prospectuses Availability

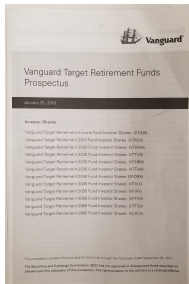
- Funds are required to publish prospectuses regularly
- There are clear guidelines regarding the information these should contain
 - ▶ Funds can be sued by the SEC for misrepresenting their behavior
- Prospectuses are publicly available through the EDGAR system since 1995
 - ▶ Sophisticated investors can automate access using the FTP of the SEC
 - ▶ Retail investors might also access prospectuses of selected funds through their online brokerage accounts or by post



The screenshot shows the SEC's EDGAR website. At the top left is the SEC logo. Below it is a navigation bar with links: ABOUT | DIVISIONS | ENFORCEMENT | REGULATION | EDGAR. On the left side, there is a sidebar with 'EDGAR Search Tools' and 'Mutual Funds' highlighted. The main content area is titled 'EDGAR | Mutual Funds' and includes the text: 'Free access to more than 21 million filings'. Below this is a search box labeled 'Mutual Fund Name' with a 'SEARCH' button.



The screenshot shows a prospectus page for Robohood Concepts Trust EMQQ. At the top, it says 'Robohood - robohood@prospectusdocs.com' and 'to view *'. There is a green leaf icon. The main heading is 'Shareholder Information from Exchange Traded Concepts Trust EMQQ The Emerging Markets Internet & Ecommerce ETF'. Below this, it says 'HL' and 'Exchange Traded Concepts Trust EMQQ The Emerging Markets Internet & Ecommerce ETF would like to share some important information with you since you purchased some shares on 03/11/2019. You'll find information about the company and some of the potential risks associated with your investment attached to the links below.' There is a link for '- Prospectus' and another for 'To learn more about shareholder information and notices, check out our [Info Center](#).' The page ends with 'Sincerely, The Robohood Team' and the website 'http://robohood.com'.



The screenshot shows the cover of a Vanguard Target Retirement Funds Prospectus. At the top right is the Vanguard logo. The title is 'Vanguard Target Retirement Funds Prospectus'. Below the title is a dark bar with the date 'March 20, 2019'. Underneath, it says 'Investor Shares' and lists several fund names with their share classes and dates, such as 'Vanguard Target Retirement Income Fund Investor Shares - 07/2006' and 'Vanguard Target Retirement 2025 Fund Investor Shares - 07/2006'. At the bottom, there is a small disclaimer: 'This prospectus contains forward-looking statements through September 30, 2019. The Securities and Exchange Commission (SEC) has not approved this prospectus and does not guarantee the accuracy of the information contained herein.'

Prospectuses Description

- They are divided in **sections** addressing different regulatory questions e.g.:
 - ▶ Principal Investment Strategies (PIS)
 - ▶ Principal Risks (PIR)
- The content, writing style and length of different sections vary substantially
 - ▶ Crucial to condition on sections when comparing text cross-sectionally
- Regulatory requirements can be satisfied in just a few sentences
 - ▶ Some funds choose to write substantially more

E.g.: Vanguard - JAG Large Cap Growth Fund

Principal Investment Strategies

The Fund invests primarily in common stocks of U.S. companies that the Fund's advisor believes have strong earnings and revenue growth potential. Under normal conditions, the Fund will invest at least 80% of the Fund's net assets plus any borrowings for investment purposes in large cap stocks defined as stocks of companies with market capitalizations of at least \$8 billion.

The advisor's employs a bottom-up, quantitatively-derived buy discipline to identify stocks the advisor believes have superior earnings and revenue growth characteristics. The cornerstone of the advisor's investment process is a proprietary multi-factor model that scores several thousand equity securities according to a variety of weighted factors measuring earnings and revenue growth, valuation, size and relative strength. The sell discipline is designed to eliminate portfolio holdings with inferior price performance and deteriorating earnings and revenue growth factors.

The Fund actively trades its portfolio investments, which may lead to higher transaction costs that may affect the Fund's performance.

Principal Risks of Investing in the Fund

As with any mutual fund, there is no guarantee that the Fund will achieve its objective. Investment markets are unpredictable and there will be certain market conditions where the Fund will not meet its investment objective and will lose money. The Fund's net asset value and returns will vary and you could lose money on your investment in the Fund and those losses could be significant.

The following summarizes the principal risks of investing in the Fund. These risks could adversely affect the net asset value, total return and the value of the Fund and your investment.

- **Equity Securities Risks.** Common stocks are subject to market risks that affect the value of the Fund. Factors such as interest rate levels, market conditions, and political events may adversely affect equity prices.
- **Management Risk.** The Portfolio Manager's judgments about the attractiveness, value and potential appreciation of particular stocks, options or other securities in which the Fund invests or sells short may prove to be incorrect and there is no guarantee that the Portfolio Manager's

- The EDGAR Mutual Fund database includes over 1 million filings
 - ▶ The historical data is highly unstructured
 - ▶ Isolating single sections for all funds over time is complex
- Parser
 - ▶ The parsing job has so far been applied to US active equity mutual funds
 - ★ We parsed both prospectuses and N-SAR filings
 - ★ It produces reliable information as of 2006
 - ★ (1994-2006 work in progress...)
 - ★ We have a total of 40,000 prospectuses correctly parsed
- Types of information
 - ▶ "Soft information" (EDGAR)
 - ★ A separate variable for each section of the prospectus (PIS and PIR)
 - ▶ "Hard information" (CRSP/Thomson Mutual Fund Databases)
 - ★ Traditional fund-level data (returns, AUM, fees, etc)

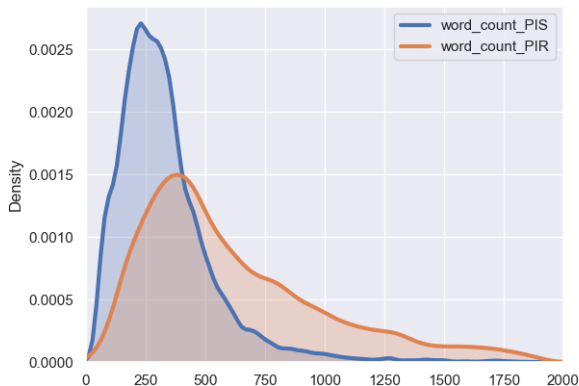
Roadmap

- 1 Introduction
- 2 Data
- 3 Descriptive Analysis**
 - Length, Complexity, Sentiment
 - Clustering
- 4 Empirical Analysis
 - Risk Shifting
 - Work in progress...
- 5 Future Research
- 6 Conclusion

PIS vs. PIR

- Strategy descriptions (**PIS**) are substantially **shorter** than Risk ones (**PIR**)
- But they are **harder** to understand - Dale Chall Score:

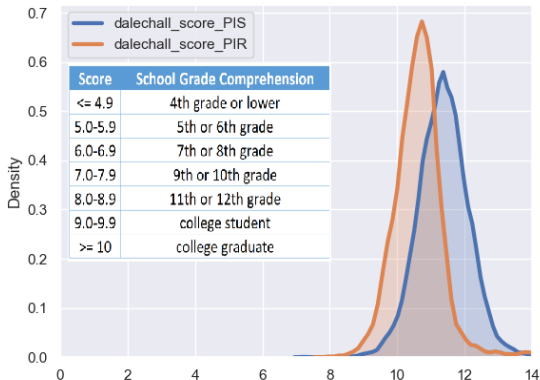
$$\begin{cases} 3.6365 \frac{\text{nonDaleChalCount}}{\text{wordCount}} > 0.5 \\ 0 & \text{otherwise} \end{cases} + 15.79 * \frac{\text{nonDaleChalCount}}{\text{wordCount}} + 0.0496 \frac{\text{wordCount}}{\text{sentCount}}$$



PIS vs. PIR

- Strategy descriptions (PIS) are substantially shorter than Risk ones (PIR)
- But they are harder to understand - Dale Chall Score:

$$\begin{cases} 3.6365 \frac{\text{nonDaleChalCount}}{\text{wordCount}} > 0.5 \\ 0 \text{ otherwise} \end{cases} + 15.79 * \frac{\text{nonDaleChalCount}}{\text{wordCount}} + 0.0496 \frac{\text{wordCount}}{\text{sentCount}}$$



PIS vs. PIR (continued)

Using Loughran and McDonald sentiment word lists we find that PIR contain:

- A higher frequency of Negative words
- A higher frequency of Uncertainty and Litigious and Constraining words
- A lower frequency of Positive words

$$Sentiment_{i,t} = \delta T_{i,t} + v_i + u_t + \epsilon_{i,t}$$

where : $i = fund$, $t = time$

$$T_{i,t} = \begin{cases} 0 & \text{if } SectionType_{i,t} = PIS \\ 1 & \text{if } SectionType_{i,t} = PIR \end{cases}$$

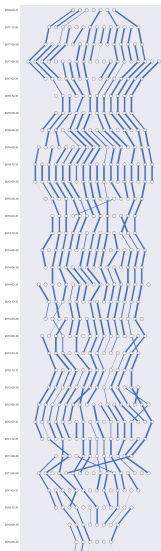
	(1)	(2)	(3)	(4)	(5)
negative	3.495***	-0.505***	0.495***	0.953***	0.252***
positive					
litigious					
uncertainty					
constraining					
T	3.495***	-0.505***	0.495***	0.953***	0.252***
T	(58.34)	(-12.29)	(20.83)	(19.79)	(16.86)
N	24716	24716	24716	24716	24716

t statistics in parentheses
* p>0.10, ** p>0.05, *** p<0.01

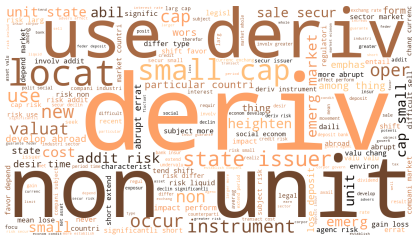
Clustering

- Unsupervised learning methods (“clustering”) can discover groupings of similar prospectuses
 - ▶ Use 4 different clustering algorithms:
 - ★ DBSCAN
 - ★ Mean-Shift
 - ★ K-Means
 - ★ Gaussian Mixture

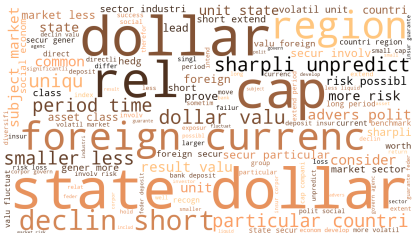
Stability and Interpretability of Clusters



June 2013 PIR Cluster 9: "Derivative Risk"



June 2013 PIR Cluster 15: "Currency Risk"



Roadmap

1 Introduction

2 Data

3 Descriptive Analysis

- Length, Complexity, Sentiment
- Clustering

4 Empirical Analysis

- Risk Shifting
- Work in progress...

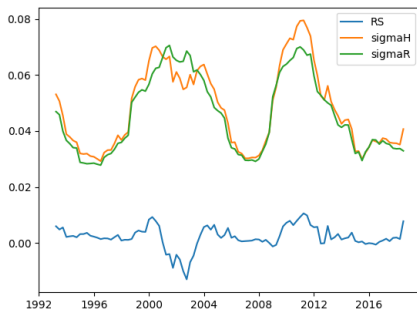
5 Future Research

6 Conclusion

Risk-Shifting

- A vast literature has shown that funds may suddenly increase their risk in order to obtain a short-term boost in returns
 - ▶ Related to career concerns and agency problems
 - ▶ Found to have a negative impact on performance
- Holdings-based measure from Huang, Sialm and Zhang:

$$RS_{i,t} = \sigma_{f,t}^H - \sigma_{f,t}^R$$
$$RH_{f,t} = \omega'_{f,t} R_t = \sum_{i=1}^n \omega_{f,t}^i R_t^i$$
$$Var(RH_{f,t}) = Var(\omega'_{f,t} R_t) = \omega'_{f,t} \Sigma \omega$$



Risk-Shifting Prediction

We predict *future risk-shifting* using the information disclosed *ex-ante* by funds in their prospectuses. We use "hard information" as a benchmark prediction.

- For each fund-quarter observation:

▶ Outcome variables:

- ★ Deciles based on the risk-shifting distribution

▶ Predictors:

- ★ Hard information: AUM, fees, returns, turnover-ratio
- ★ Soft information:

→ tf-idf matrix of unique PIS and PIR

→ text statistics (word count, complexity, dictionary frequencies)

- ★ All predictors matched at prior windows: (-48m, -36m, -24m, -12m)

- Split the sample into two randomly selected sub-samples of unique documents

- ▶ One used for model estimation (**Training Sample**; N=[4102-7905])

- ▶ The other used to assess performance (**Testing Sample**; N=[1758-3388])

- Assess model goodness by looking at:

- ▶ **Recall** = $TP / (TP + FN)$ & **Precision** = $TP / (TP + FP)$

Risk-Shifting Model Selection

- We build supervised learning models to predict future risk-shifting

Risk-Shifting Model Selection

- We build supervised learning models to predict future risk-shifting
 - ▶ Train 6 machine learning algorithms using **training sample** & Verify model accuracy using **testing sample**.

Risk-Shifting Model Selection

- We build supervised learning models to predict future risk-shifting
 - ▶ Train 6 machine learning algorithms using **training sample** & Verify model accuracy using **testing sample**.
 - ▶ Deciles prediction:

Testing-sample predictability of deciles using information at 24 months prior window

Model: Only hard information

	1	2	3	4	5	6	7	8	9	10	precision	recall	support
1	51	30	11	9	11	8	14	10	13	14	0.18	0.3	171
2	26	46	23	32	23	22	16	26	29	23	0.16	0.17	266
3	35	47	48	47	18	29	32	25	31	31	0.19	0.14	343
4	28	33	43	58	41	37	39	28	25	25	0.18	0.16	357
5	25	34	37	52	44	39	31	42	26	29	0.17	0.12	359
6	26	14	28	41	39	50	40	26	29	18	0.19	0.16	311
7	18	16	16	22	37	31	37	35	36	40	0.14	0.13	288
8	29	31	17	24	21	19	25	34	43	27	0.12	0.13	270
9	26	22	11	21	9	17	20	33	33	26	0.11	0.15	218
10	13	17	18	16	11	12	20	13	29	47	0.17	0.24	196

Risk-Shifting Model Selection

- We build supervised learning models to predict future risk-shifting
 - ▶ Train 6 machine learning algorithms using **training sample** & Verify model accuracy using **testing sample**.
 - ▶ Deciles prediction:

Testing-sample predictability of deciles using information at 24 months prior window
Model: Only prospectuses

	1	2	3	4	5	6	7	8	9	10	precision	recall	support
1	94	32	5	3	5	6	3	6	5	12	0.41	0.55	171
2	37	72	38	25	17	22	18	11	12	14	0.25	0.27	266
3	14	51	69	75	41	24	19	18	18	14	0.24	0.2	343
4	12	29	59	64	61	41	36	31	14	10	0.19	0.18	357
5	13	35	39	58	62	41	42	31	17	21	0.21	0.17	359
6	18	21	30	44	49	63	30	25	19	12	0.21	0.2	311
7	8	17	17	21	28	43	58	40	32	24	0.21	0.2	288
8	8	10	13	23	20	27	35	58	49	27	0.23	0.21	270
9	13	11	11	12	6	18	28	22	55	42	0.22	0.25	218
10	14	9	9	6	8	10	8	13	31	88	0.33	0.45	196

Risk-Shifting Model Selection

- We build supervised learning models to predict future risk-shifting
 - ▶ Train 6 machine learning algorithms using **training sample** & Verify model accuracy using **testing sample**.
 - ▶ Deciles prediction:

Testing-sample predictability of deciles using information at 24 months prior window
Model: All textual and hard information

	1	2	3	4	5	6	7	8	9	10	precision	recall	support
1	94	38	8	2	4	7	3	4	6	5	0.41	0.56	171
2	35	80	38	23	22	17	15	10	14	12	0.28	0.29	266
3	13	44	67	78	38	30	25	19	12	17	0.27	0.22	343
4	14	32	46	77	56	41	37	24	21	9	0.22	0.19	357
5	17	33	37	47	69	50	42	29	19	16	0.21	0.19	359
6	21	17	24	37	51	56	42	25	21	17	0.18	0.17	311
7	12	18	18	26	30	38	59	38	35	14	0.2	0.2	288
8	9	15	13	22	20	30	36	55	44	26	0.24	0.21	270
9	17	8	12	12	8	9	28	22	56	46	0.21	0.24	218
10	13	7	11	2	9	10	8	15	21	100	0.36	0.47	196

Risk-Shifting Model Selection

- We build supervised learning models to predict future risk-shifting
 - ▶ Train 6 machine learning algorithms using **training sample** & Verify model accuracy using **testing sample**.
 - ▶ Deciles prediction:

Testing-sample predictability of deciles using information at 24 months prior window

Model: All textual and hard information

	1	2	3	4	5	6	7	8	9	10	precision	recall	support
1	94	38	8	2	4	7	3	4	6	5	0.41	0.56	171
2	35	80	38	23	22	17	15	10	14	12	0.28	0.29	266
3	13	44	67	78	38	30	25	19	12	17	0.27	0.22	343
4	14	32	46	77	56	41	37	24	21	9	0.22	0.19	357
5	17	33	37	47	69	50	42	29	19	16	0.21	0.19	359
6	21	17	24	37	51	56	42	25	21	17	0.18	0.17	311
7	12	18	18	26	30	38	59	38	35	14	0.2	0.2	288
8	9	15	13	22	20	30	36	55	44	26	0.24	0.21	270
9	17	8	12	12	8	9	28	22	56	46	0.21	0.24	218
10	13	7	11	2	9	10	8	15	21	100	0.36	0.47	196

- ▶ Problem: low power in explaining middle deciles
- ▶ Solution: binary outcome var (1 for top & bottom 10%, 0 otherwise)
 - Use stratified sampling to split between training and testing
 - Down-sample majority/Up-sample minority class while training
 - Use the most accurate algorithm: **Random Forest with down-sampling**

Risk-Shifting Prediction Accuracy

- Testing prediction of binary risk-shifting variable
 - ▶ **Recall** = $TP / (TP + FN)$ & **Precision** = $TP / (TP + FP)$

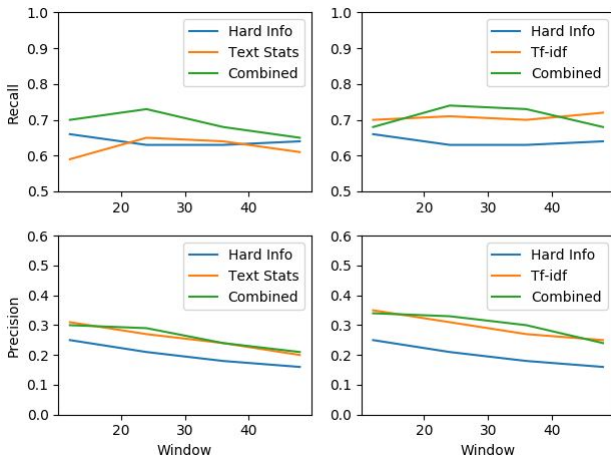
Testing-sample predictability of deciles using information at 24 months prior window

	Only Hard information				Only Prospectuses				All textual and hard information				support
	0	1	precision	recall	0	1	precision	recall	0	1	precision	recall	
0	1530	882	0.92	0.63	1848	564	0.94	0.77	1857	555	0.95	0.77	2412
1	136	231	0.21	0.63	108	259	0.31	0.71	96	271	0.33	0.74	367

Risk-Shifting Prediction Accuracy

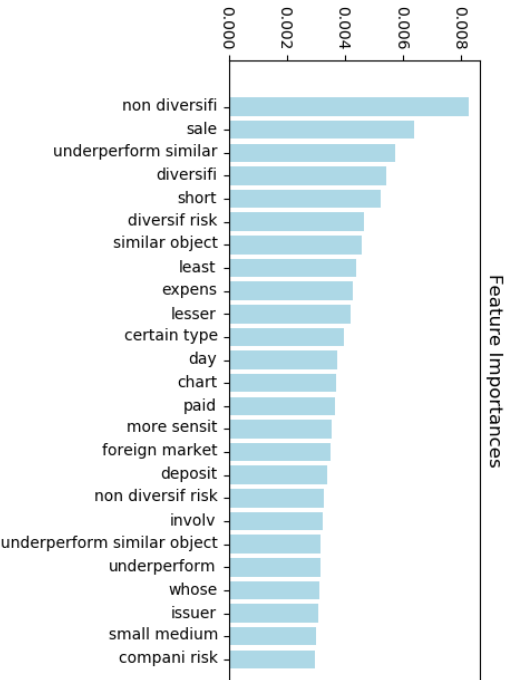
- Testing prediction of binary risk-shifting variable

▶ **Recall** = $TP / (TP + FN)$ & **Precision** = $TP / (TP + FP)$



Risk-Shifting Insights

- Features importance and holdings concentration



Work in progress...

- Beyond risk shifting, can we utilize supervised learning to predict likelihood of future lawsuits?
 - ▶ Collecting legal action data against mutual funds from N-SAR and ATV filings
- Is textual similarity associated with similar fund returns?
 - ▶ Use cluster assignment to predict future return distribution

Roadmap

1 Introduction

2 Data

3 Descriptive Analysis

- Length, Complexity, Sentiment
- Clustering

4 Empirical Analysis

- Risk Shifting
- Work in progress...

5 Future Research

6 Conclusion

Future Research

● Investors attention

- ▶ Solve a model of rational inattention with investors with different levels of financial literacy who learn from prospectuses
- ★ Utilize the model to derive testable predictions regarding the distribution of funds size

- ▶ Relate signals exacted from the prospectuses to funds size/growth
- ★ Distinguish between institutional and retail investors' assets

● Identification

- ▶ Relate textual measures to fund growth one year after inception
- ▶ Regulatory changes:

- ★ 1995: introduction of online EDGAR distribution system
- ★ 1998 (Rule 421): Readability act
- ★ 1999: Increase disclosure requirements in PIS
- ★ 2004: More frequent disclosure

Roadmap

1 Introduction

2 Data

3 Descriptive Analysis

- Length, Complexity, Sentiment
- Clustering

4 Empirical Analysis

- Risk Shifting
- Work in progress...

5 Future Research

6 Conclusion

Conclusion

- Results suggests that prospectuses contain meaningful information about:
 - ▶ Funds' trustworthiness
 - ▶ The shape of funds' future return distribution
- It remains an open question (that we plan to answer) whether investors are paying attention to this information!