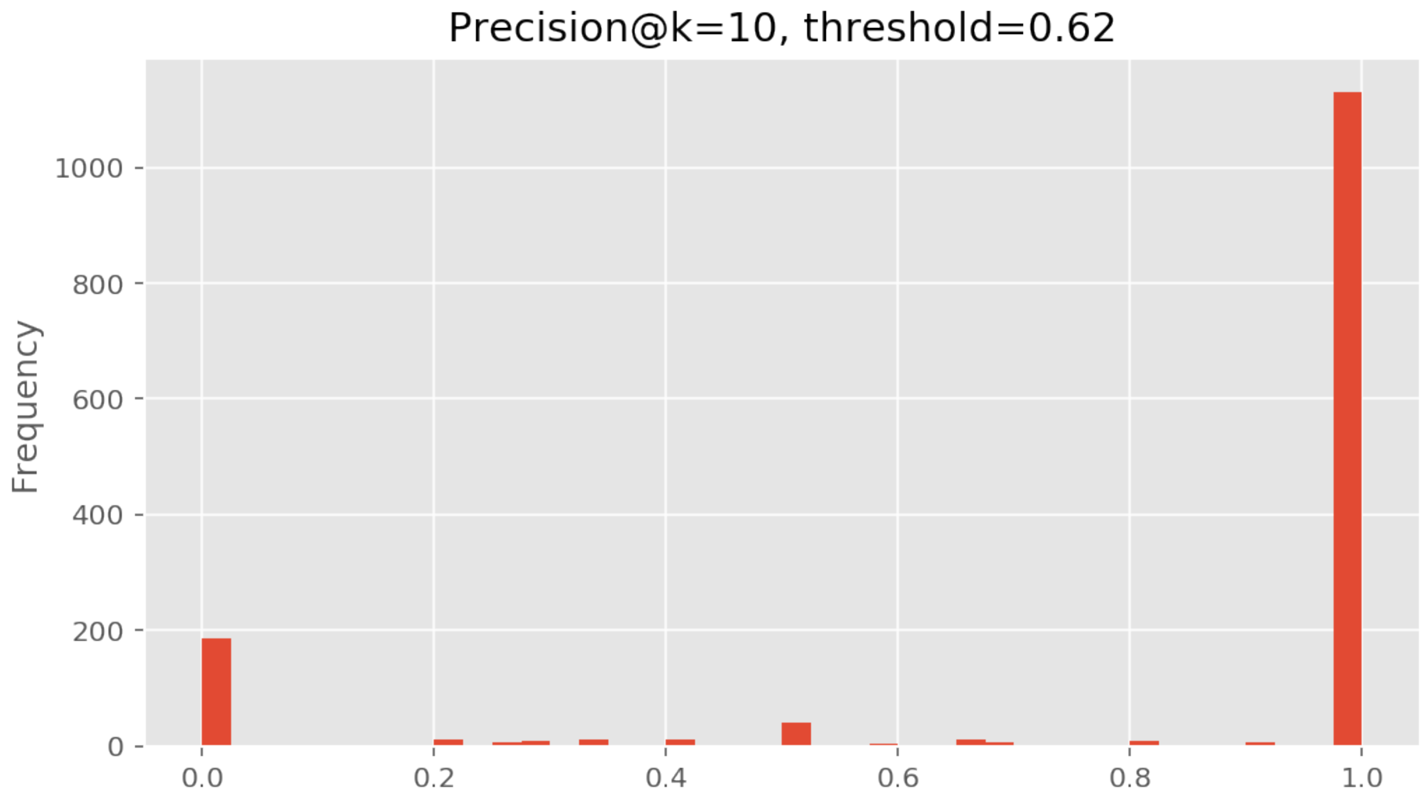
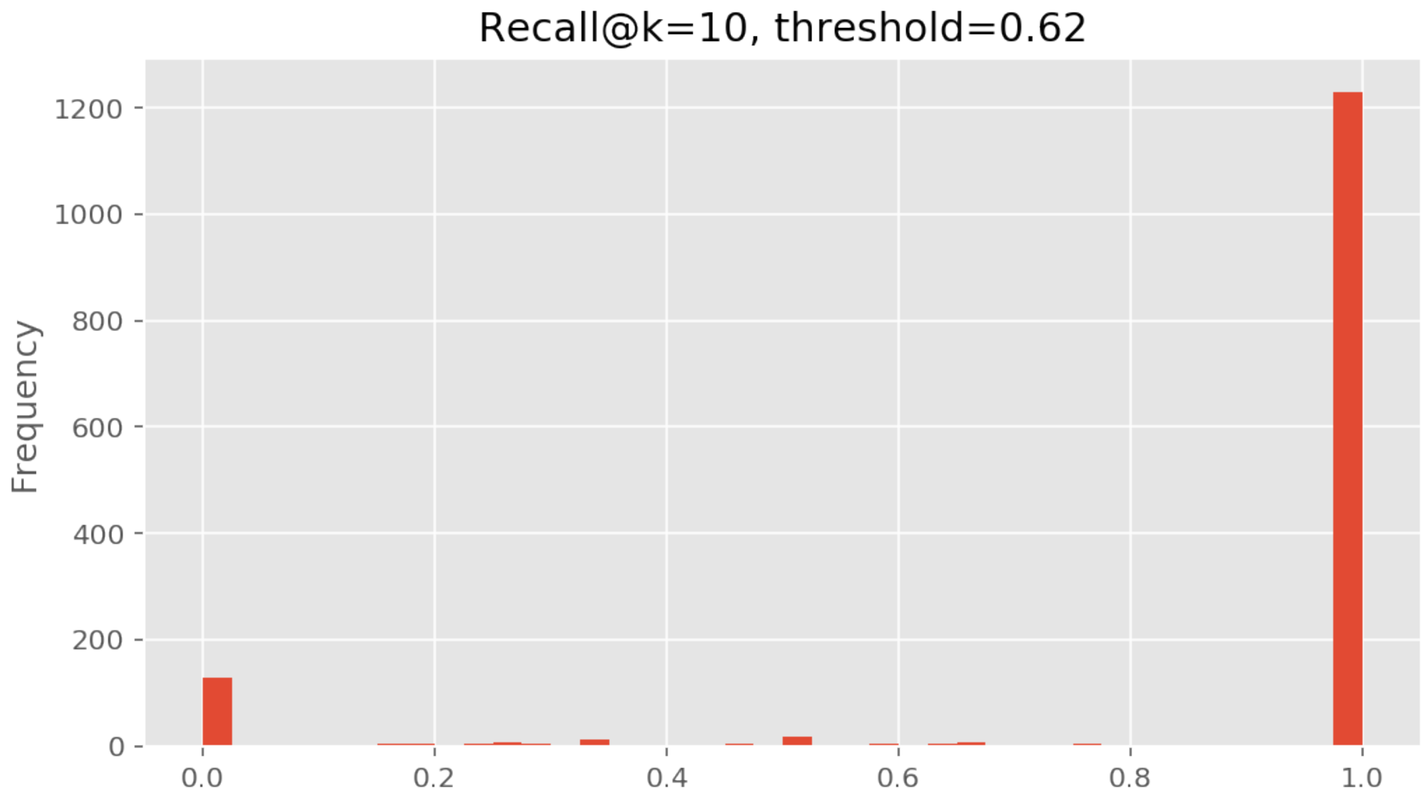
DSI-9 Capstone Project

Recommendation engine for equities investors

tl;dr

- Recommendations for 1561 investors based on 73769 individual shareholdings across 449 companies
- 0.84 mean precision@k=10 with threshold set at the baseline





Context

- Lots of buy-side firms invest on behalf of large numbers of other people e.g. pension funds
- They generally source ideas from whereever and then whittle them down
- This includes the sell-side who offer up ideas and try to earn trading commission
- Should be able to come up with recommendation model(s) to do this

Three methodologies

Will use three separate methodologies and combine the outputs:

1. Content-based filtering
2. Collaborative filtering
3. Classification models

Content-based filtering

Financial data

- There is a lot of data in financial markets because people are trying to forecast the future, particularly financial results in future periods
- Equity research is written by sell-side firms to provide (gu)estimates
- 3rd party data providers provide mean of estimates from these firms, which are taken to be the market estimate of future financial performance

Numis **Banks**
For FCA purposes this is a Marketing Communication

4 March 2019 06:30 GMT

Banks

Estimate Changes

All about distributions

The 1,500 pages of disclosure we get from domestic banks over full year provides a unique opportunity to check our assumptions and see if we've missed anything - particularly as our cash flow analysis relies on several items only detailed in the accounts. This time last year we were alerted to the drags from capitalised software spend. This year it is the benefits arising from equity-settled award schemes versus treasury share debits. We are also a step closer to understanding the implications of IFRS 9 in a more normalised credit environment. Distributions remain front and centre to our call on banks. They all screen reasonably but only RBS stands out.

- **Capital generation from award schemes:** In the last few years this hasn't been a big consideration of ours with add backs to profits for equity-settled awards being largely offset by Treasury share movements. In 2018 though, LBG benefited from net ET1 build of 22bps (0.6p) related to these items with little neutralisation of related share issuance taking place. Distributions of 5.7p were declared for the year but as we saw in 2018, if such awards are not neutralised through open market purchases, some of this flows back to employees through an increase in share count - at LBG equivalent to half of last year's buyback. There is nothing unusual about such practice but it does have implications in an environment where buybacks comprise a large portion of overall "distributions". LBG's "underlying" FCF is strong and we have made upgrades elsewhere (see below) but it is an important consideration. Elsewhere, Barclays largely neutralises share awards and RBS' mainly pays in cash / bonds;
- **Stage 1 / 2 TTC provisions:** new disclosure (from RBS) appears to confirm our view that as credit conditions normalise (whenever that may be) both traditional impairment charges and provisions against performing assets will increase. We now expect that the latter will add an average 10bps to impairment charges over the period of transition, slightly lower than our previous estimate but still material in the context of distributions;
- **Small changes to PPP estimates.** However, EPS upgrades average 4% over 2019-2022E driven by lower impairment and "exceptional" items, and higher buybacks.
- **FCF estimates increased more markedly.** EPS changes are amplified at the FCF level by deductions between the two. We have also modestly reduced the size of these deductions and at LBG assume a further excess insurance dividend in 2019 and lower mortgage RWA add-ons in 2020. Here FCF increases to 5.1p per annum, the low end of that implied by company guidance (but 4.4p excluding that attributable to awards, see above). As a consequence, target prices for all three domestic banks are increased.
- **RBS remains the standout.** While FCF yield across the sector suggests limited downside for all, our distribution potential estimates at RBS - increased to an average 26p per annum in 2019-2022E stepping up to 31p post 2022E - make it a clear standout for us. This is particularly the case when considered in the context of risk and its relatively prudent approach to most areas with which we have concern.

Analysts

Jonathan Pierce
Tel: +44 (0)20 7260 1609
j.pierce@numis.com

Tom Rayner
Tel: +44 (0)20 7260 1271
t.rayner@numis.com

This research was prepared and approved by

Numis Securities Limited
The London Stock Exchange Building
10 Paternoster Square
London EC4M 7LT

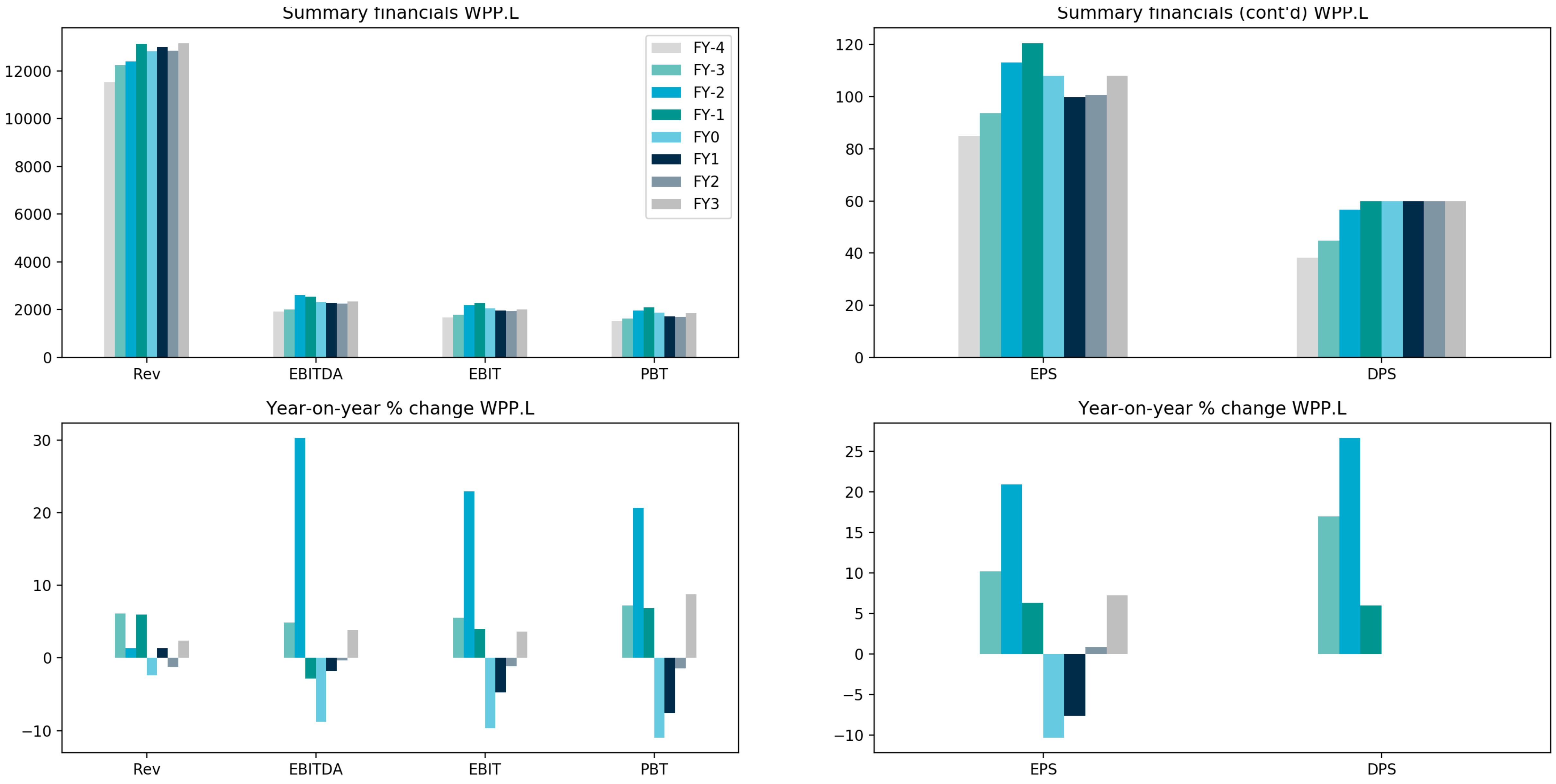
+44 (0)20 7260 1000 | mail@numis.com
www.numis.com

Historic and estimated financial information

Below is for WPP, a large advertising conglomerate

(£m, Dec YE)	FY-4	FY-3	FY-2	FY-1	FY0	FY1	FY2	FY3
Rev	11528.9	12235.2	12397.8	13139.6	12826.6	13004.5	12842.6	13149.3
EBITDA	1909.5	2002.4	2608.1	2534.1	2311.1	2269.06	2260.8	2347.71
EBIT	1680.6	1774	2180.3	2267.1	2047.3	1950.61	1928.28	1998.47
PBT	1512.6	1622.3	1957.9	2092.5	1862.8	1721	1696.2	1844.68
EPS (p)	0.849	0.936	1.132	1.204	1.08	0.9979	1.0066	1.0796
DPS (p)	0.382	0.4469	0.566	0.6	0.6	0.6	0.6	0.6

Historic and estimated financial information, charted



Feature engineering from financial data

Obvious simple metrics familiar to anyone who looks at accounts growth, CAGR, margins

Other features that have been added to attempt to capture trends:

- latest period relative to peak period i.e. $FY_3/\max(FY_1, FY_2, FY_3)$
- lowest period relative to the latest period i.e. $\min(FY_1, FY_2, FY_3)/FY_3$
- growth periods out of the number of periods less 1 e.g. if growth in FY2 and FY3 it will be 1
- positive periods proportion i.e. the proportion of the periods for which the metric is > 0
- Mean and sd for each line item are used to calculate $\log_{10}(\bar{x}/s_x)$ and then discarded, with $\log_{10}(\bar{x}/s_x)$ being retained as a feature

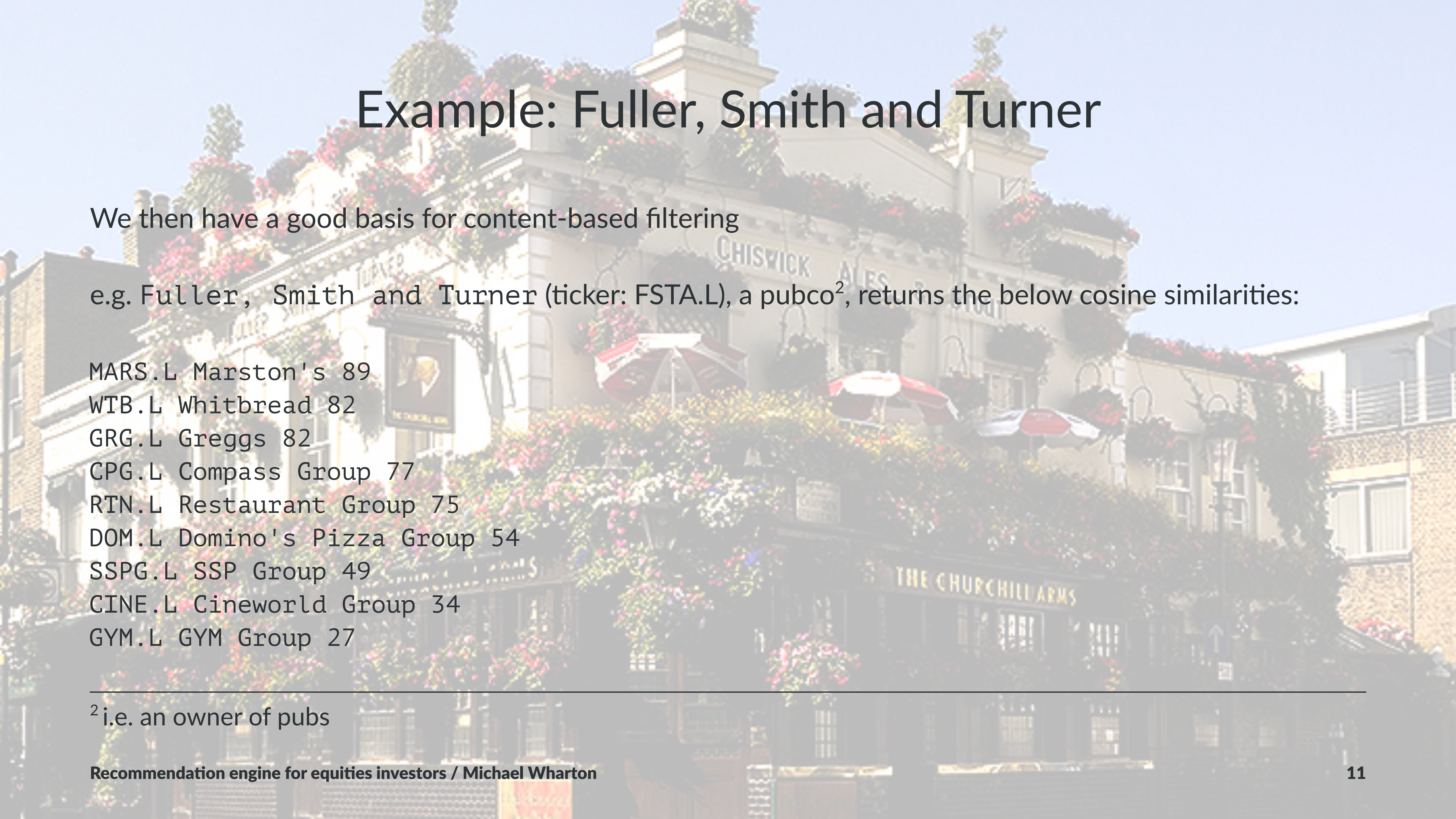
Non-financial features

- Business type e.g. for `Antofagasta`

Economic Sector Name	Basic Materials
Business Sector Name	Mineral Resources
Industry Group Name	Metals & Mining
Industry Name	Specialty Mining & Metals
Activity Name	Copper Ore Mining

- $\log$ of total employees
- NTM P/E for valuation[1]

[1] share price / next twelve months EPS



Example: Fuller, Smith and Turner

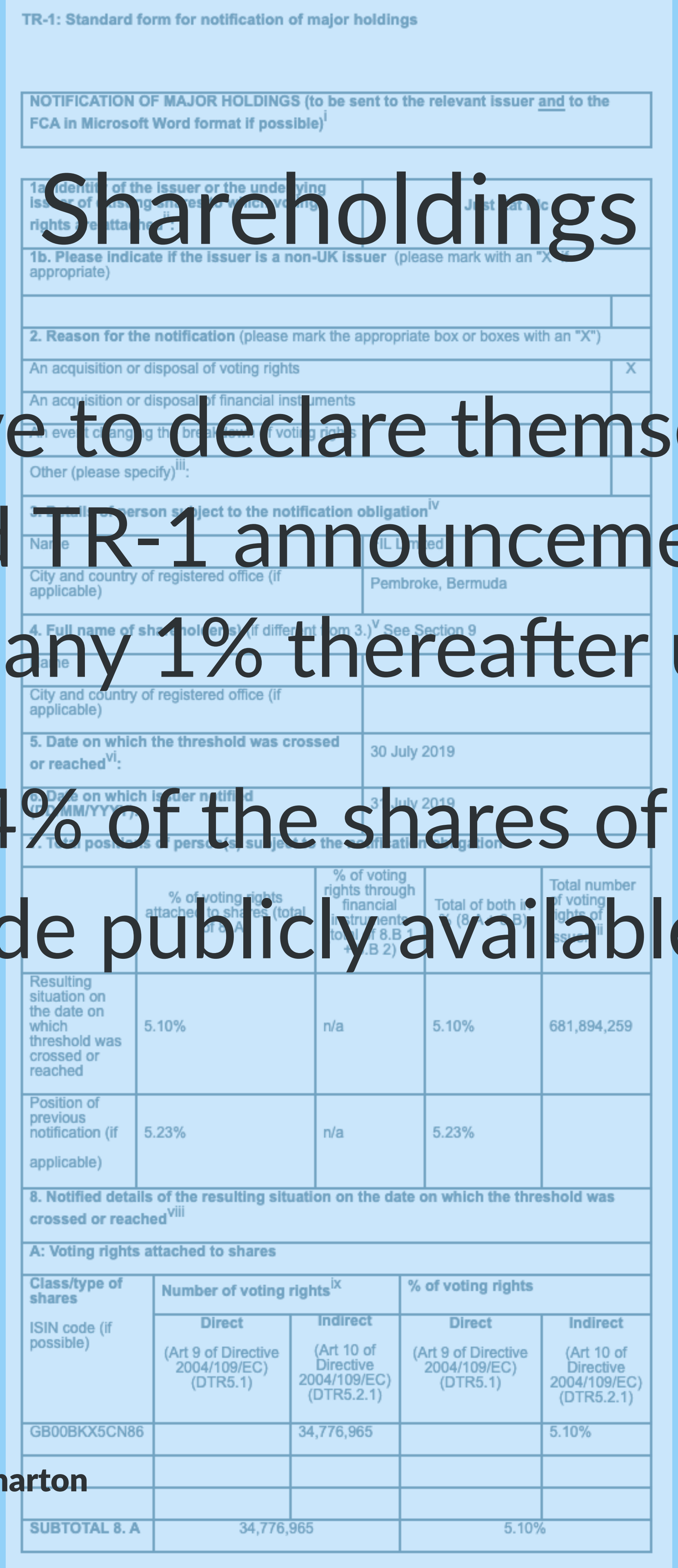
We then have a good basis for content-based filtering

e.g. `Fuller, Smith and Turner` (ticker: FSTA.L), a pubco[2], returns the below cosine similarities:

```
MARS.L Marston's 89
WTB.L Whitbread 82
GRG.L Greggs 82
CPG.L Compass Group 77
RTN.L Restaurant Group 75
DOM.L Domino's Pizza Group 54
SSPG.L SSP Group 49
CINE.L Cineworld Group 34
GYM.L GYM Group 27
```

[2] i.e. an owner of pubs

Collaborative filtering


TR-1: Standard form for notification of major holdings

NOTIFICATION OF MAJOR HOLDINGS (to be sent to the relevant issuer and to the FCA in Microsoft Word format if possible)[i]

1b. Please indicate if the issuer is a non-UK issuer (please mark with an "X" if appropriate)

2. Reason for the notification (please mark the appropriate box or boxes with an "X")

An acquisition or disposal of voting rights	X
An acquisition or disposal of financial instruments	
Other (please specify)[iii]:	

City and country of registered office (if applicable)	Pembroke, Bermuda
City and country of registered office (if applicable)	
5. Date on which the threshold was crossed or reached[vi]:	30 July 2019

	% of voting rights attached to shares	% of voting rights through financial instruments	Total of both	Total number of voting rights of issuer
Resulting situation on the date on which threshold was crossed or reached	5.10%	n/a	5.10%	681,894,259
Position of previous notification (if applicable)	5.23%	n/a	5.23%	

8. Notified details of the resulting situation on the date on which the threshold was crossed or reached[viii]

A: Voting rights attached to shares

Class/type of shares ISIN code (if possible)	Number of voting rights[ix] Direct (Art 9 of Directive 2004/109/EC) (DTR5.1)	Indirect (Art 10 of Directive 2004/109/EC) (DTR5.2.1)	% of voting rights Direct (Art 9 of Directive 2004/109/EC) (DTR5.1)	Indirect (Art 10 of Directive 2004/109/EC) (DTR5.2.1)
GB00BKX5CN86		34,776,965		5.10%
SUBTOTAL 8. A	34,776,965		5.10%	

Shareholdings

- Material investors have to declare themselves through regulatory announcements called TR-1 announcements when they go through 3% of ISC (or any 1% thereafter up or down)
- i.e. if you go and buy 4% of the shares of a UK public company, this information is made publicly available

Percentage shareholdings matrix

	888.L	AAAA.L	AAL.L	ABF.L	ACAA.L
AXA Investment Managers UK Ltd.	nan	nan	nan	nan	nan
Aberdeen Asset Investments Limited	0.2625	0.53	0.0558	0.2931	0.264
Aberdeen Standard Investments (Edinburgh)	10.1594	nan	0.6222	0.4983	0.423
Artemis Investment Management LLP	2.6826	nan	0.6879	nan	nan
Aviva Investors Global Services Limited	2.8267	0.3354	0.2566	0.1353	0.1035
BlackRock Advisors (UK) Limited	0.0524	0.0622	0.1294	0.2875	0.3374
BlackRock Institutional Trust Company, N.A.	0.8694	0.9654	5.8374	0.9182	0.4552
BlackRock Investment Management (UK) Ltd.	0.9544	0.2194	0.2496	0.3542	5.6017

Shareholdings ~ ratings

idx	Investor	Co	Holding
0	0	2	0.062
1	0	35	0.0103
2	0	43	0.0431
3	0	66	0.0397
4	0	160	0.0032

Shareholdings ~ ratings

idx	User	Item	Rating
0	0	2	0.062
1	0	35	0.0103
2	0	43	0.0431
3	0	66	0.0397
4	0	160	0.0032

Collaborative filtering

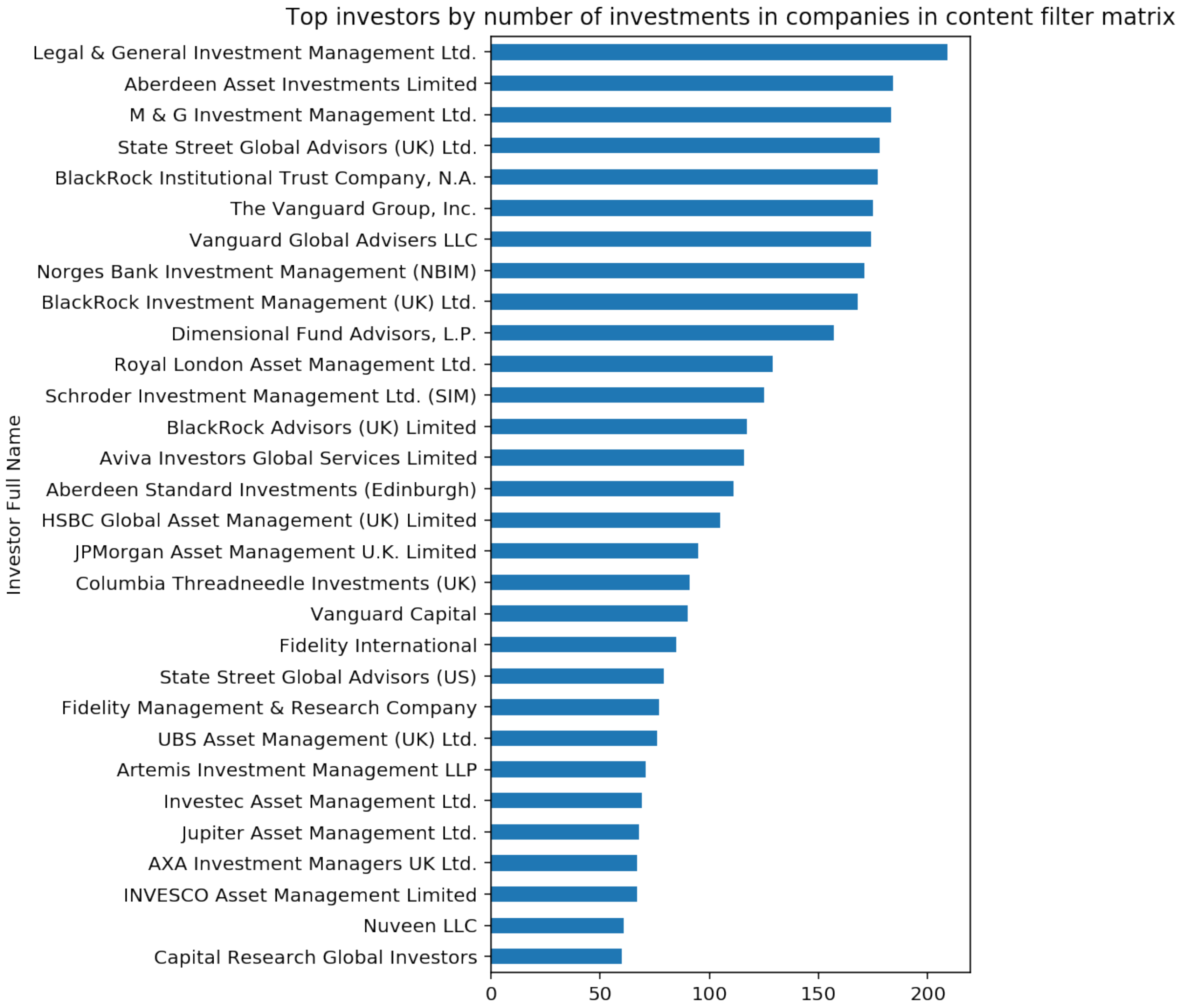
- Cross validation scores for different collaborative filtering algorithms
- Used SVD, which scored RMSE of 1.09 on test data

	RMSE
SVD	1.0393
SlopeOne	1.0485
KNNWithMeans	1.0524
BaselineOnly	1.0536

Classification models

Invested/not invested is a binary classification problem

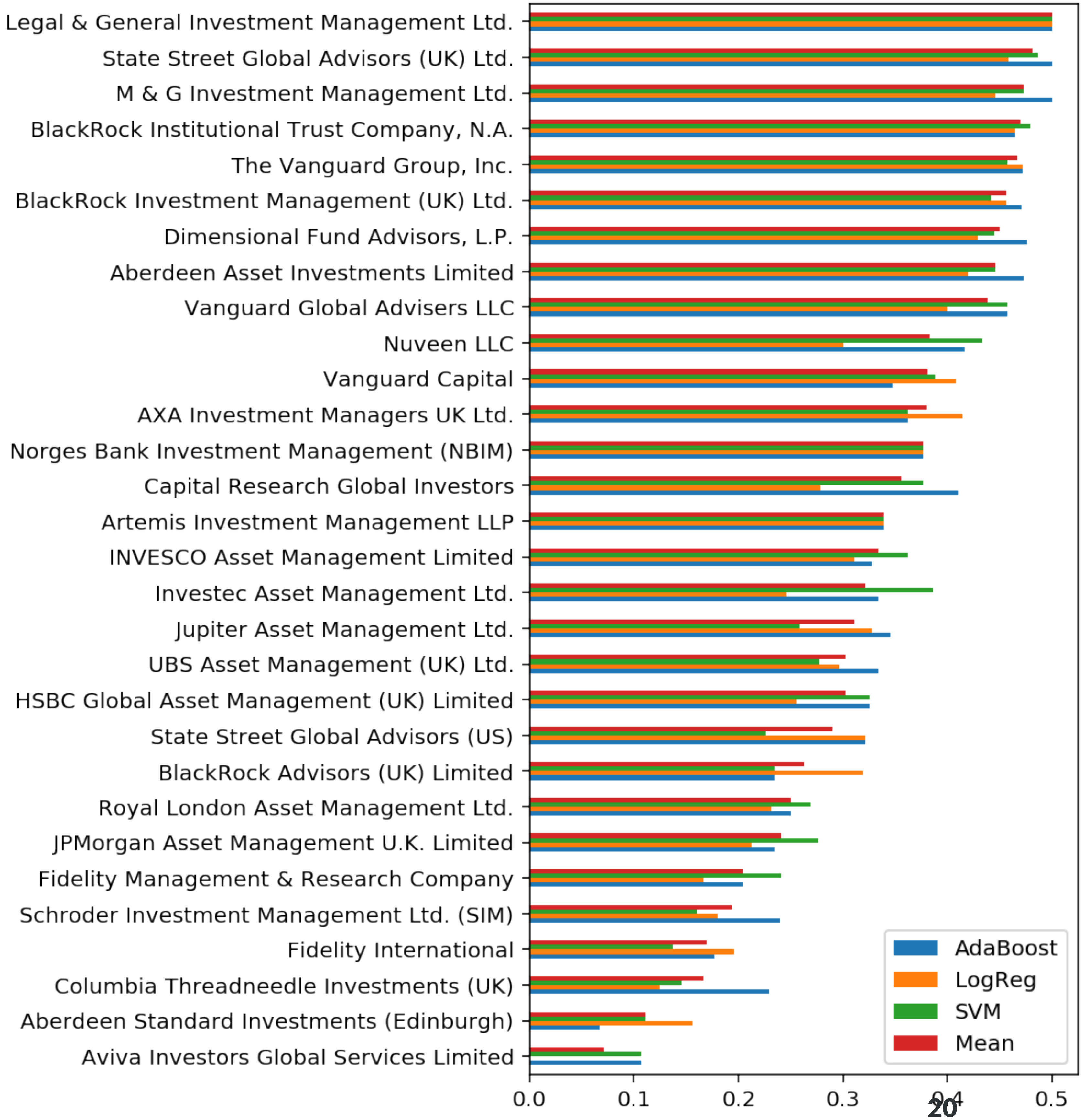
- Suggests classification models could be suitable
- Top 30 picked as class imbalance becomes quite silly after that (baseline for Capital Research ~ 0.75)
- Can use classification model for each of these investors to predict which companies they will be invested in



So with a bit of random oversampling to address class imbalance, we can fit and test classification models

Improvement in accuracy over baseline shown over

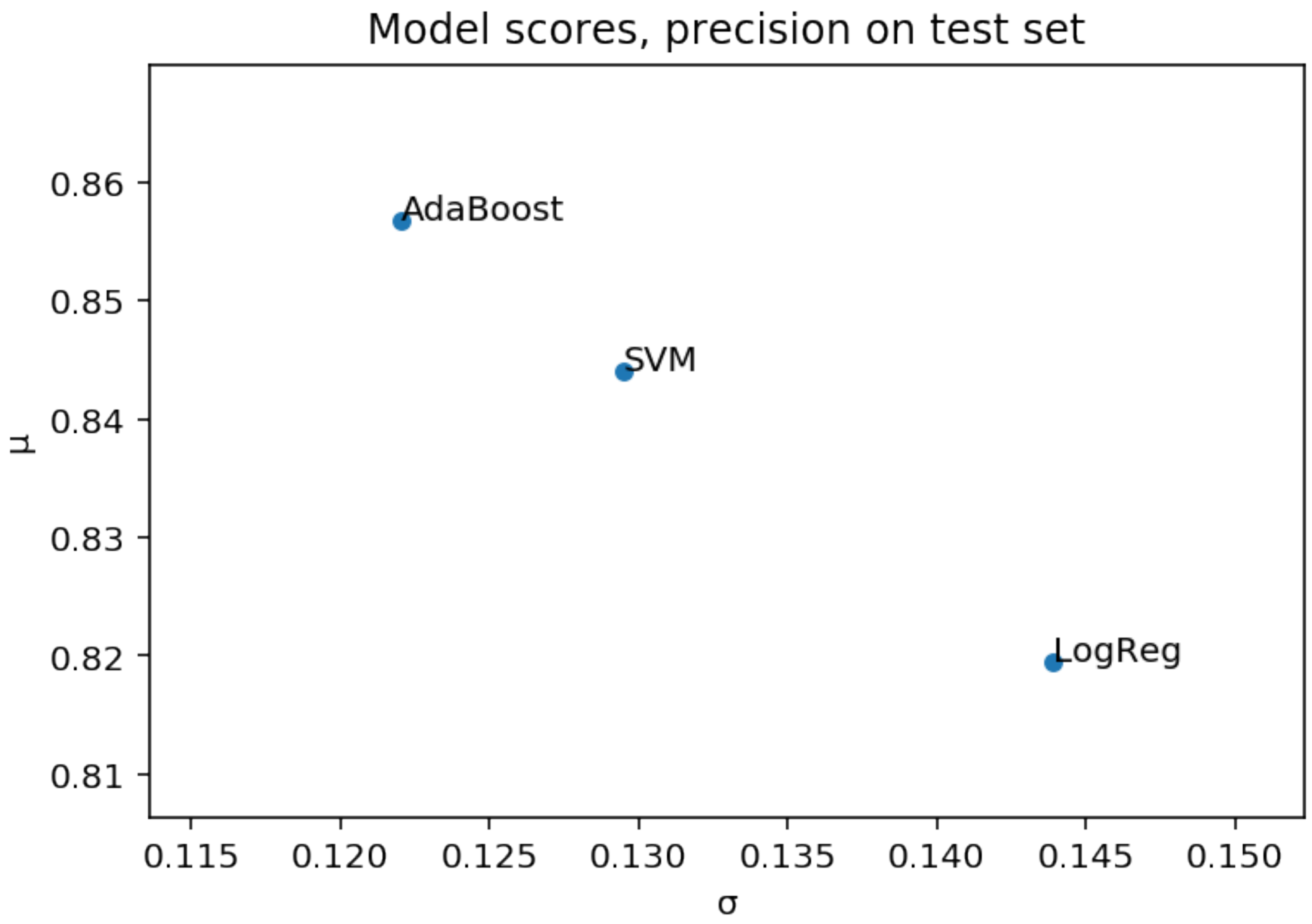
- L&G is clearly over-fit, no doubt as a result of oversampling
- Improvement over baseline is generally good – mean mean improvement over baseline is 0.331
- Best model for precision is then used to predict for each investor



Model evaluation, comparing precision scores $(n = 30)$

	AdaBst	**LogReg**	**SVM**
μ	0.85679	0.81954	0.84407
σ	0.12204	0.14388	0.12949

AdaBoost is the most precise overall



Example: BlackRock

- Can use the model with the highest precision for each investor e.g. below are shown results for `BlackRock Advisors (UK) Limited`:

	AdaBoost	LogReg	SVM	Best
Accuracy	0.744681	0.829787	0.744681	LogReg
Baseline	0.510638	0.510638	0.510638	-
Improvement	0.234043	0.319149	0.234043	LogReg
PR AUC	0.822325	0.873067	0.839429	LogReg
Precision	0.866667	0.857143	**0.923077**	**SVM**
ROC AUC	0.740942	0.828804	0.740036	LogReg
Recall	0.565217	0.782609	0.521739	LogReg
Train_CV_mean	0.801938	0.759767	0.759599	AdaBoost
Train_CV_sd	0.0499497	0.0515466	0.053361	SVM

How does this generate recommendations?

- Probabilities of positive class taken, those shares already invested in removed, top k once ordered by probability are recommendations
- Example for `Blackrock`, as before, and using the `SVM` outputs

	p	pred	hold
WEIR.L	0.892764	1	0
SDR.L	0.783258	1	0
AGGK.L	0.422853	0	0
FSTA.L	0.40612	0	0
MARS.L	0.404967	0	0

Creating a single list

Combining the ≤ 3 model results

- Lists need to be scaled before they are combined
- MinMax scaler used to transform the recommendation strength of each list into $[0, 1]$ range
- Lists are then combined, ranked, and the top k recommended

Summary

- Possible to create recommendation system for equities and good quality data is available for this task
- Someone who professionally sources investment ideas may find it useful to have suggestions that they can then look into further
- Could be extended to cover e.g. US, European equities
- Content filtering in this context is useful for other task (e.g. M&A screening)

Limitations

- Universe of 211 companies is not gigantic and limiting
- Banks and insurers not included because they use different accounting and so e.g. do not have EBITDA - not that much of an issue in practice as investors in financials tend to be specialists

Model suitability

- Content filtering works well in terms of feel i.e. it returns what could be expected
- Classification model probabilities are intuitive and this works well in the middle of the top 30 investors albeit not well for very material class imbalance so cannot be use for many investors
- Collaborative filtering is a bit black-box like so makes the most sense only in combination with one of the other two approaches I think