SPRINGBOARD JANUARY 2021

E. CHAE

Finance modeling

- This is an analysis of financial data on an estimated 4000 publicly traded U.S. stocks as sourced on https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks20142018?select=2015_Financial_Data.csv
- Independent variables analyzed to uncover associations with annual change in stock price cover data reported on 10K filings with SEC
- Data science methods applied include PCA, LLE, Random Forest, stratification, XGBoost, and affinity propagation clustering
- Operating cash flow is significantly associated with annual changes in stock price
  - This is a consistent finding across all analyses

Finance modeling pre-processing

Pre-filter	2014	2015	2016	2017	2018
Number of companies	3808	4120	4797	4960	4392

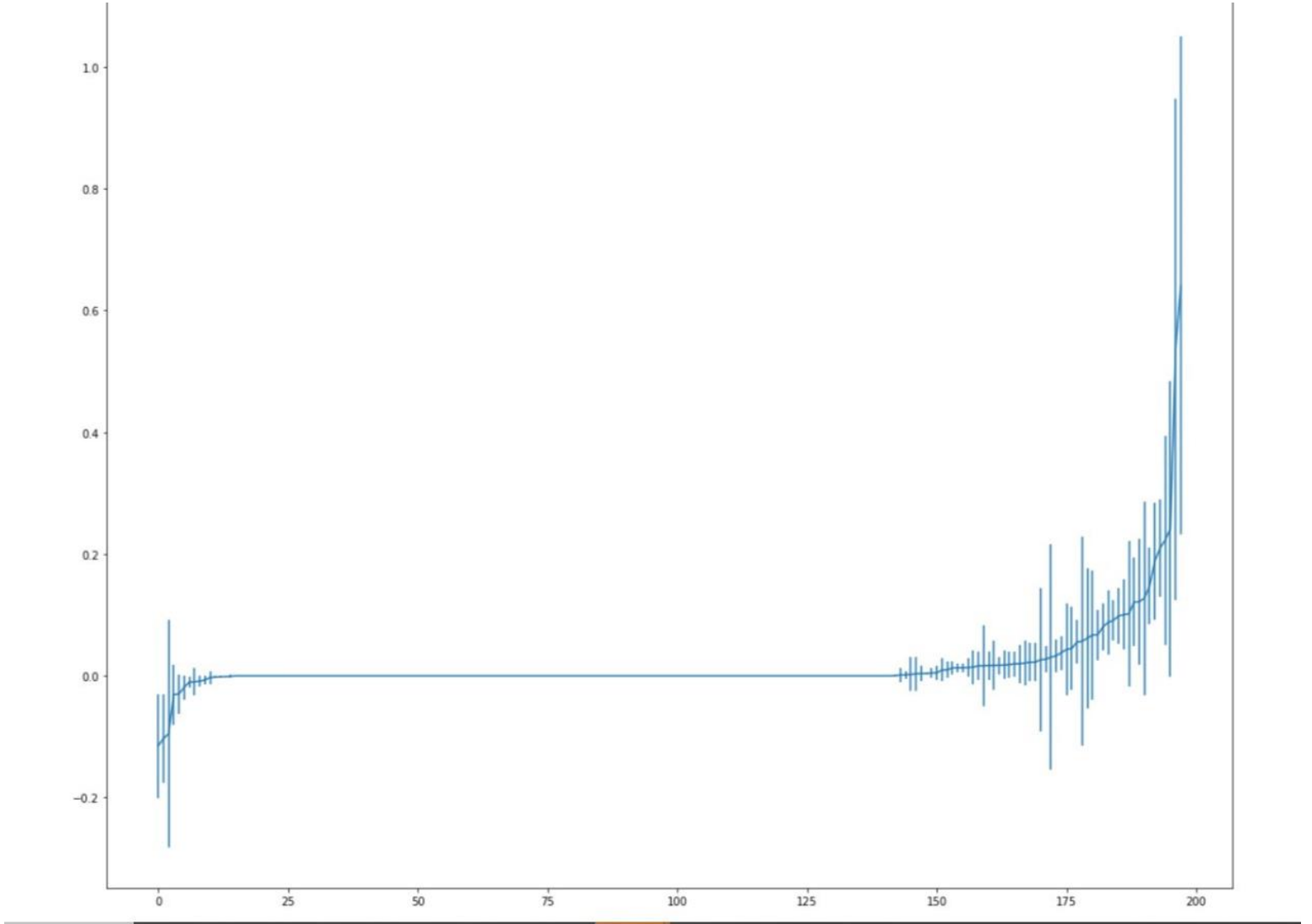
Post-filter	2014	2015	2016	2017	2018
Number of companies	513	597	740	758	793

Post-filter	2014	2015	2016	2017	2018
Average of probabilities on all variables of selecting a subset with these distributions of features at random	0.688	0.673	0.689	0.668	0.677

EDA
Identifying Top Relevant Variables I

- Reviewed differences in means in each of 198 variables to identify significant ones

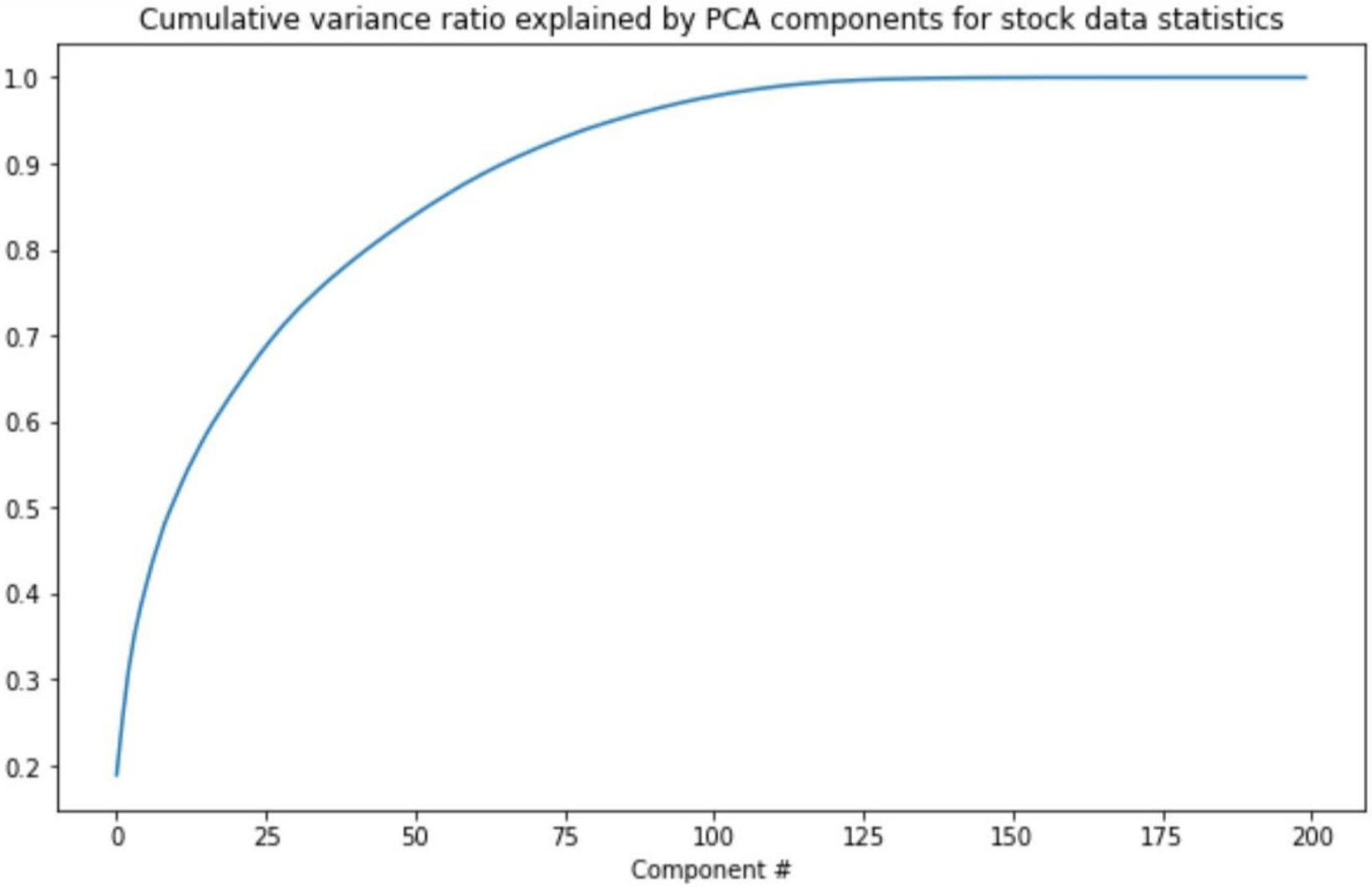
- As visible on graph the majority of variables are not significantly different on mean value when grouped by change in annual stock price.



EDA
Identifying Top Relevant Variables 2

- Performed a PCA to identify top features contributing to linear variation

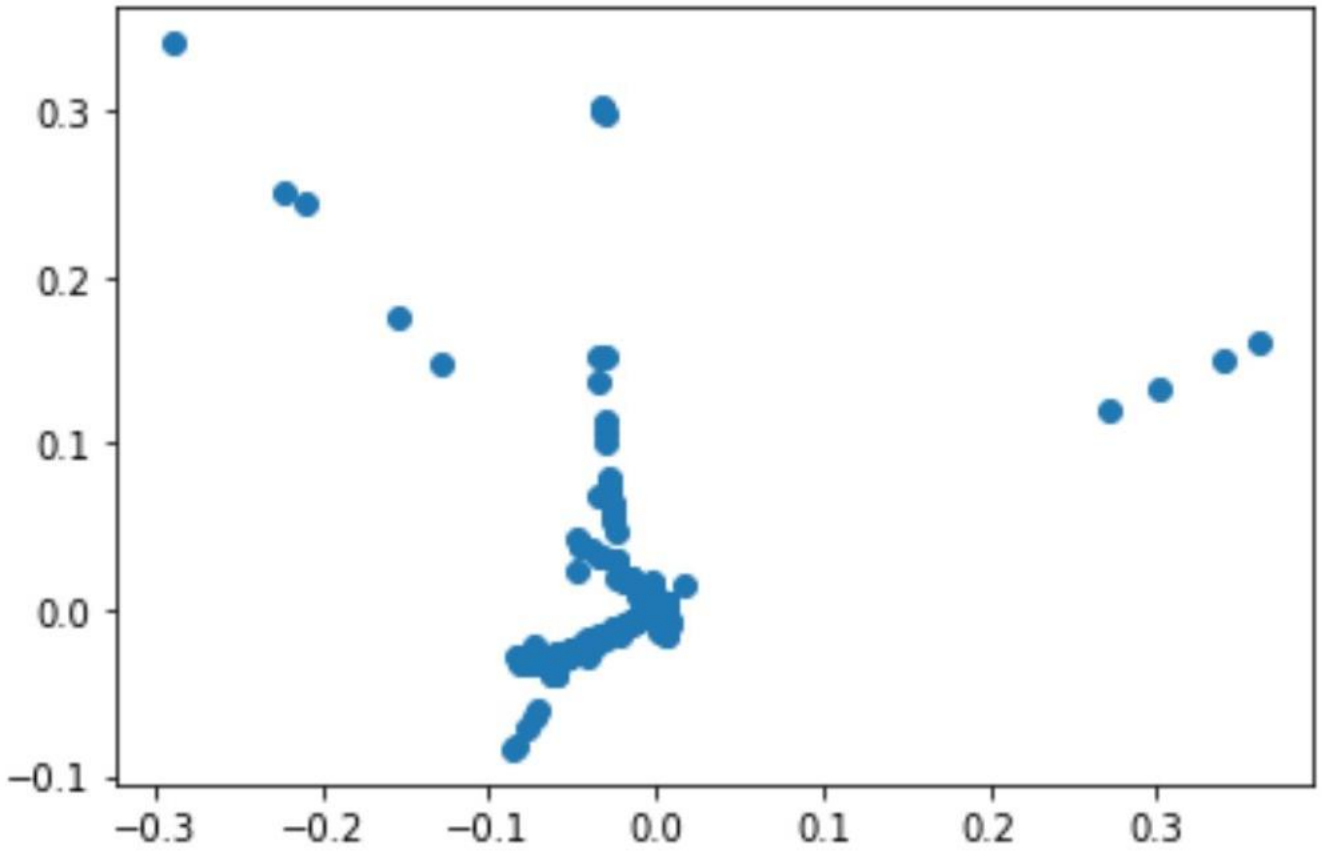
- As visible on graph a small group of features are responsible for majority of linear variation



EDA
Identifying Top Relevant Variables 3

- Performed LLE analysis to identify non-linear variation

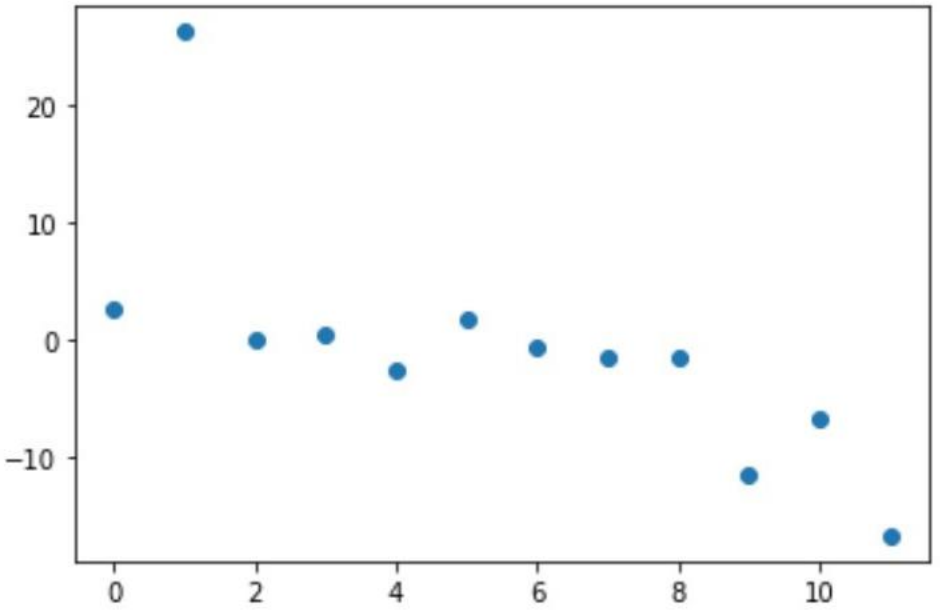
- As evident based on non-random pattern of graph there is substantial non-linear variation



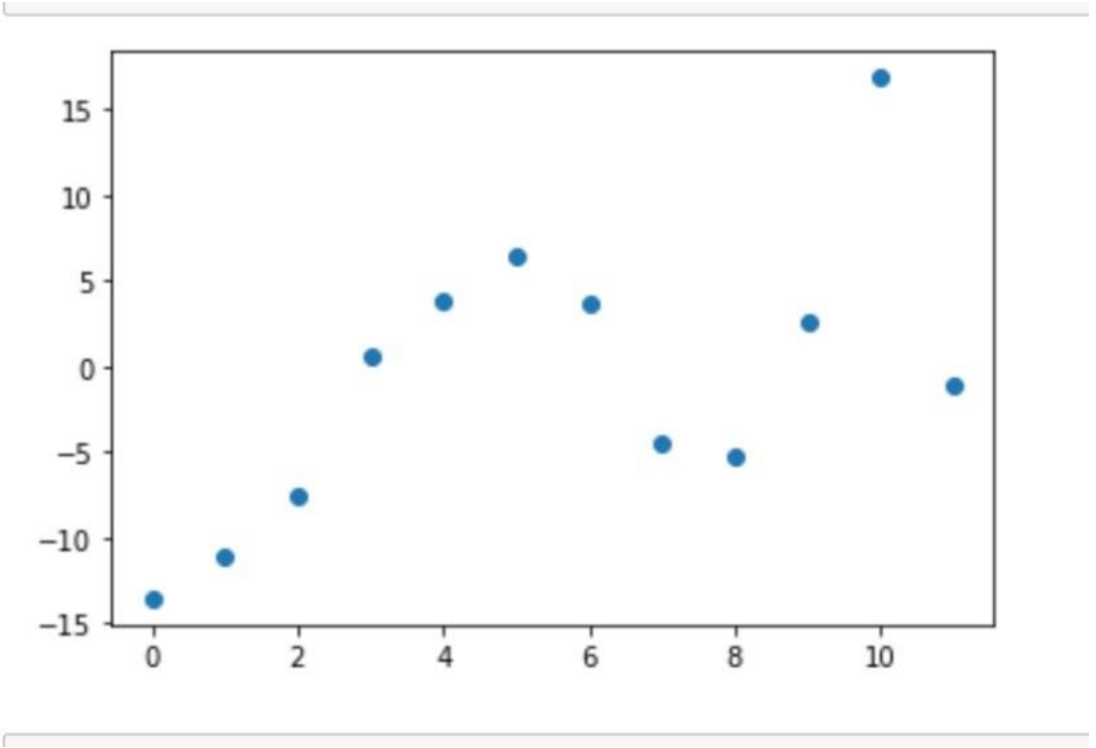
EDA
Stratification analyses

- Stratified on selected variables to plot continuous price by intervals
  - This identifies linear and non-linear variation in data not adjusted by other variables

- Capex to Operating cash flow



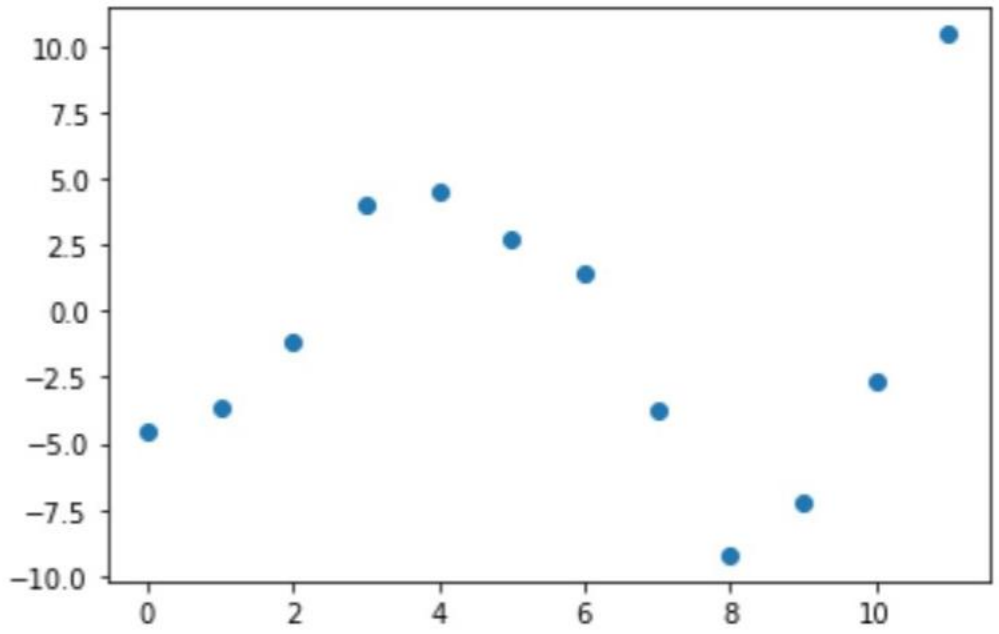
- Revenue Growth



EDA
Stratification analyses

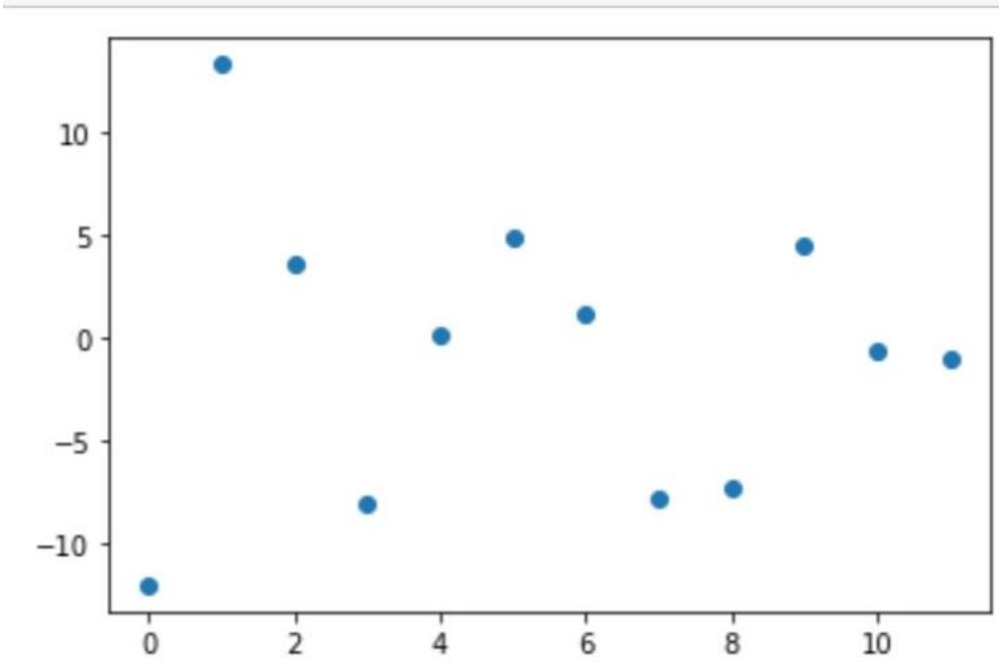
- Additional example of non-random plot

- Earnings Yield



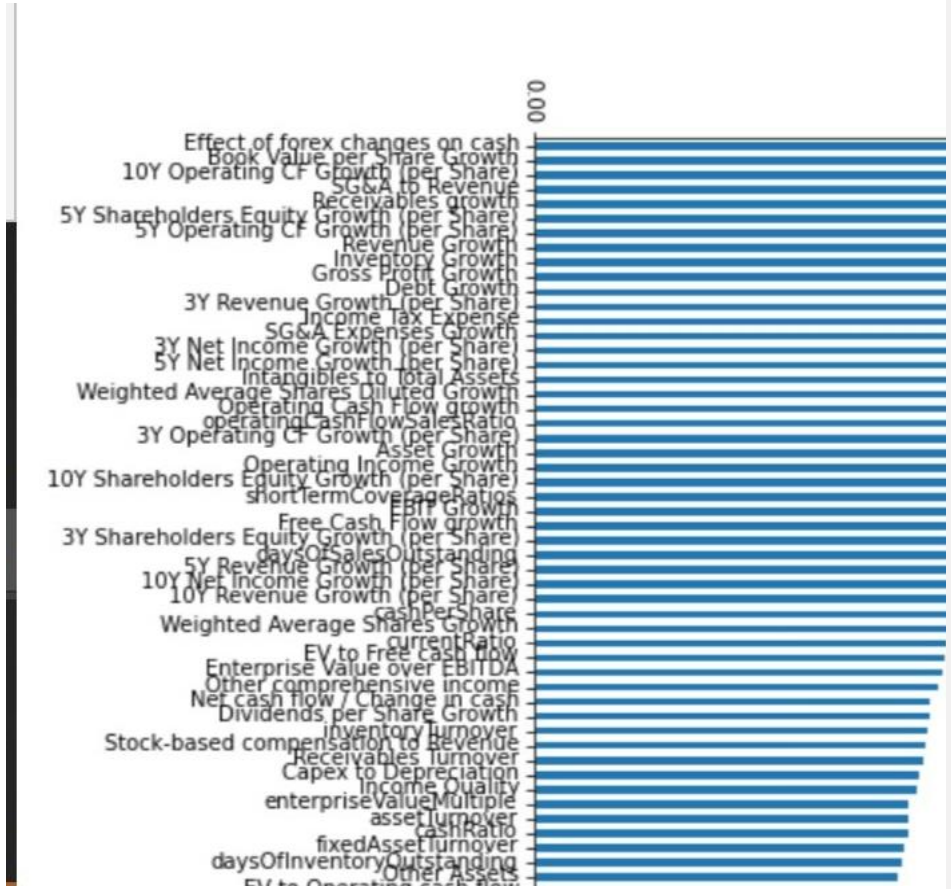
- Example of random plot to compare with previous graphs

- Free Cash Flow



Modeling with Random Forest & cross validation

- As visible on graph a group of features, including variables relevant to operating cash flow, are responsible for majority of quantifiable variation
- This finding is consistent model-to-model and by subset of each year

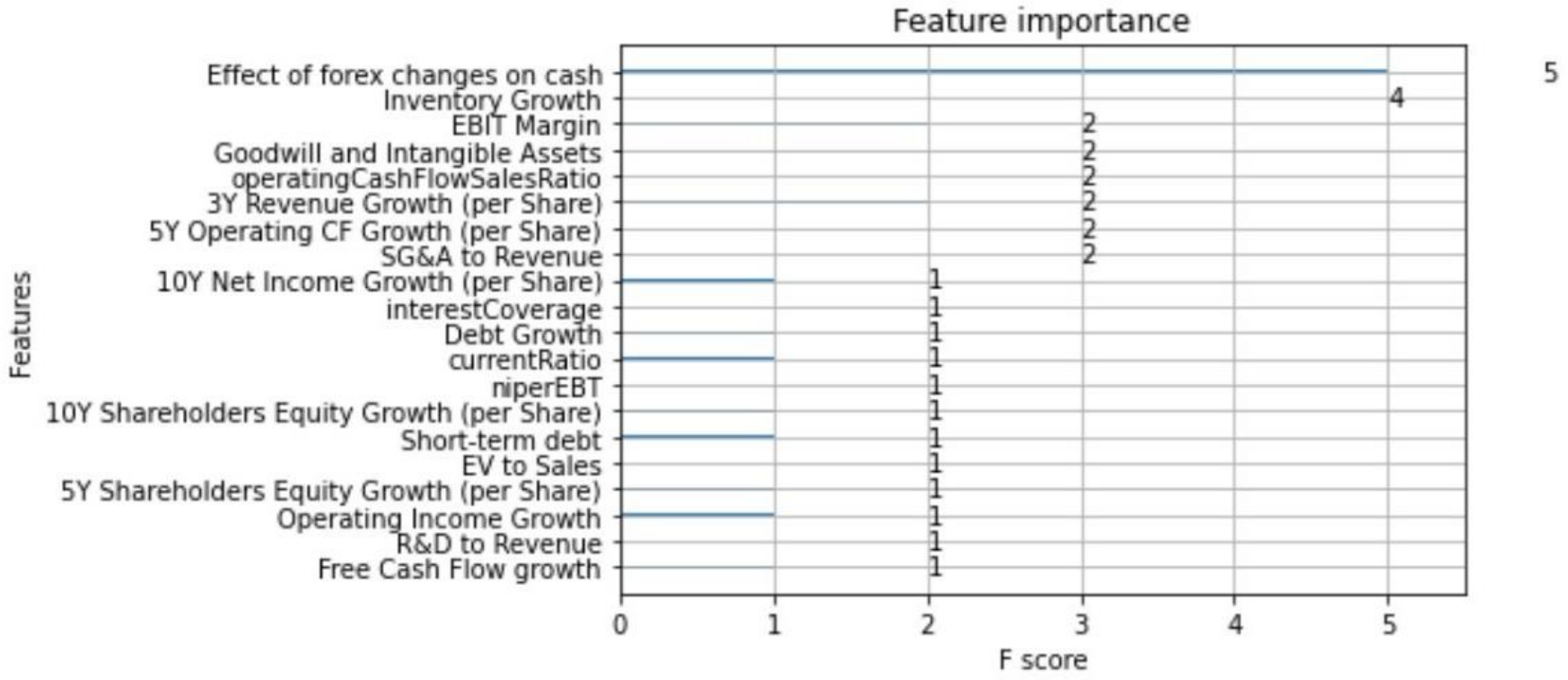


- Effect of Forex Changes on Cash
- Book Value per share Growth
- 10Yr Operating CF Growth (per share)
- SG&A to Revenue
- Receivables Growth
- 5Yr Shareholder's Equity Growth (per share)
- 5Yr Operating CF Growth (per share)
- Revenue Growth
- Inventory Growth
- Gross Profit Growth
- Debt Growth
- 3 Yr Revenue Growth (per share)
- Income Tax Expense
- SG&A Expenses Growth
- Inventory Growth
- 3Yr Net Income Growth (per share)

Modeling with XGBoost

- Performed XGBClassifier to identify top features contributing to variation

- As visible on graph top 20 features of 199 variables on all years of data comprise majority of variation in a model with ~0.48 RMSE

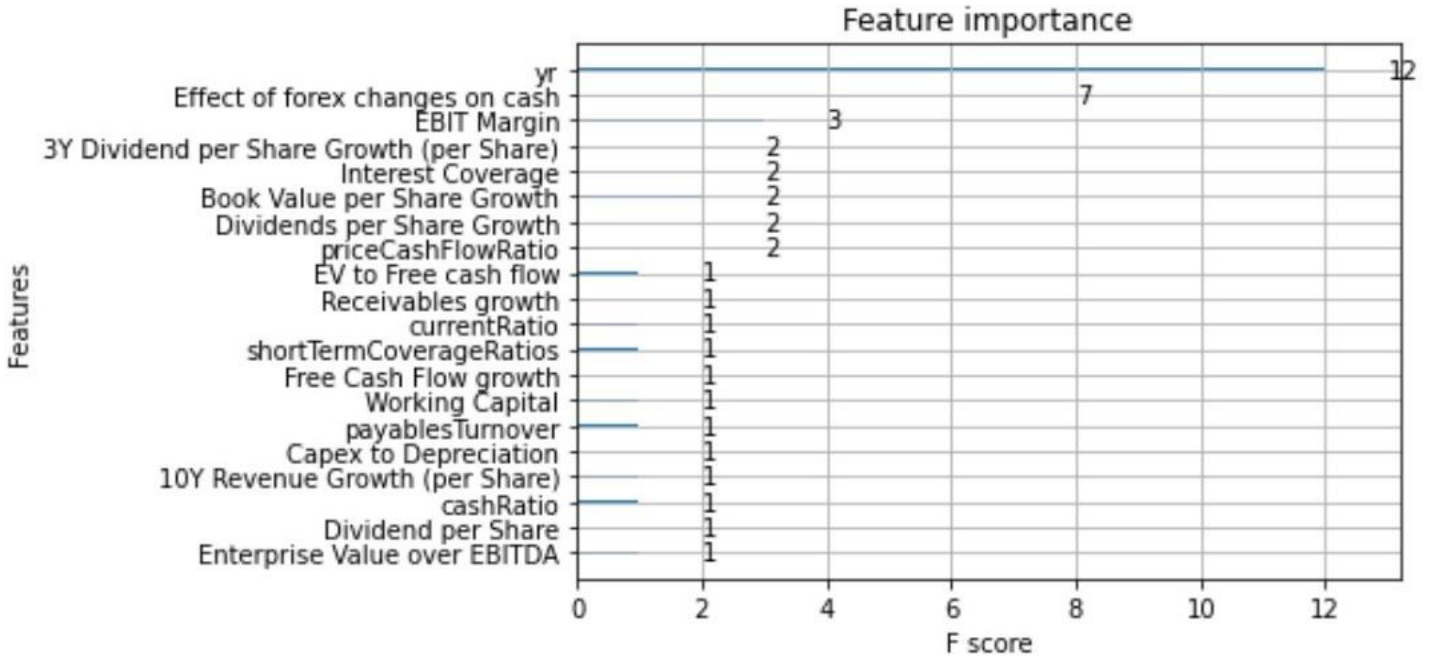


Review – whether to add information on XGBClassifier with 2016 data including variables based on financial theory. Engineered feature is top 4.

Modeling with XGBoost

- Performed XGBClassifier to identify top features contributing to variation with hypertuned parameters

- As visible on graph top 20 features of 199 variables, including macroeconomic variable covering interest rate, on all years of data comprise majority of variation



Supplementary application of machine learning to finance

- Data science code could give financial clients investment options
  - Some institutional or private wealth clients might prefer to invest in a particular company over another
    - For example – a client might like a certain company because of peer-to-peer networking preferences, social responsibility of executive management, etc.
- This is a sample of cluster analysis, affinity propagation, that identifies companies with similar returns in the four-year period covered by these data

Label	Company	Stock Price ChangeYr1	Stock Price ChangeYr2	Stock Price ChangeYr3	Stock Price ChangeYr4	SparkLines	Sector
0	GRA	-29.96228278	3.108679505	-7.140644767	7.912022399		Basic Materials
0	DIS	2.724909837	2.923334183	-0.367842325	34.34665181		Consumer Cyclical
0	IPG	5.296554033	-11.40189696	5.598865453	19.20942576		Consumer Cyclical
0	THRM	-25.68606138	-6.480119922	24.74259124	9.227362343		Consumer Cyclical
0	MNRO	-10.49863092	0.878090854	19.06960142	15.4671812		Consumer Cyclical
0	KR	-15.03650314	-17.0682515	-0.990448919	8.690633021		Consumer Defensive
0	FLO	-3.46230212	0.424385727	-0.979235372	22.52823288		Consumer Defensive
0	EPD	8.741576179	4.741643526	-2.674145956	19.92462041		Energy
0	BMY	-11.21962864	5.878996919	-12.93831545	26.64174768		Healthcare
0	MDT	-4.001501267	16.91339538	13.13922826	31.74119997		Healthcare
0	HSIC	-2.28648982	-10.37004154	9.864278065	11.36012526		Healthcare
0	UHS	-9.489575151	6.597315605	1.154187513	22.69436542		Healthcare
0	SNA	3.276889076	3.715060695	-16.11116184	17.7768145		Industrials
0	MATX	-11.79951816	-15.15938205	6.188329325	29.39309049		Industrials
0	PLAB	-6.842538466	-24.84581986	12.55813809	61.14520101		Technology
0	PPL	5.508848557	-5.139197474	-1.902321534	34.46355476		Utilities
1	SGC	22.18109703	45.58138984	-31.36767518	-23.33825991		Consumer Cyclical
1	NUS	30.8456918	43.57168645	-8.762576324	-30.42018776		Consumer Defensive
1	MMSI	46.73311619	64.88549429	28.15155534	-41.31579161		Healthcare
1	CVTI	5.683065349	39.39834397	-31.67259608	-34.36548323		Industrials
1	ARCB	34.5204251	25.28338193	-6.576158945	-19.30555489		Industrials
1	KEQU	43.48141058	20.7962702	21.18893149	-56.69874477		Industrials
2	BLL	5.592738493	1.051436634	20.17158821	45.99209743		Consumer Cyclical
2	HELE	-6.114508267	12.36151426	36.71703684	37.99217327		Consumer Defensive
2	LHCG	6.353268477	34.14367106	52.35313028	47.96992134		Healthcare
2	ENSG	3.138401088	0.025375906	67.30557798	27.46090884		Healthcare
2	ACIW	-12.19158417	22.14440048	21.94799218	39.60943346		Industrials
3	CENT	148.4234303	16.10977815	-11.23421916	-11.07612789		Consumer Defensive
3	OKE	142.8304216	-4.347816211	3.936205794	46.51230673		Energy
3	PMD	163.5955247	-11.71017487	-20.74483184	-39.26527368		Healthcare