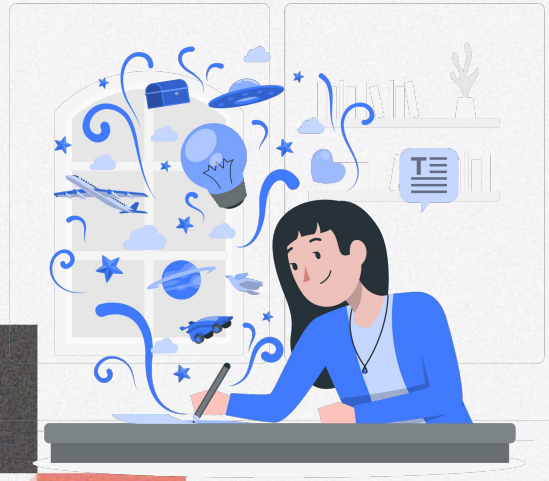
SC1015 MINI PROJECT: STOCKS

Done by:

1. Yeoh Wei Yang (U2121112A)
2. Chantharojwong Kasidis (U2020731L)
3. Harvey Zhang Tianren (U2122097A)

Agenda



1. Motivation
2. Exploratory data analysis
3. Model building
4. Applying our model

1

Motivation

Why stocks?

Wealth

A method of generating income passively

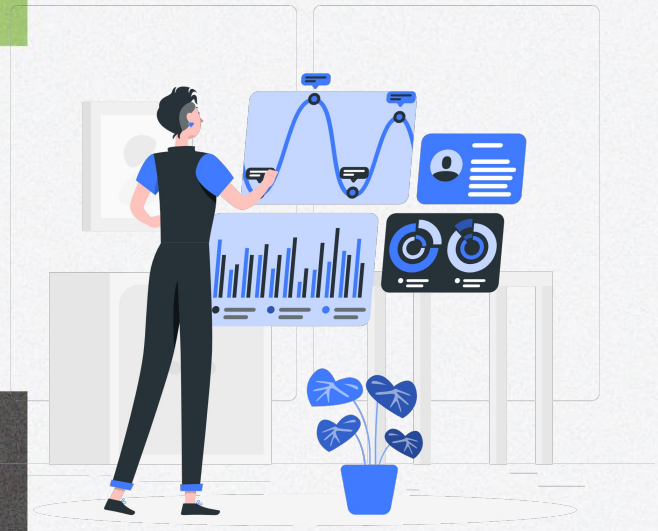
Inflation

Growth of money can beat out inflation

Potential

Potential for high returns

However!

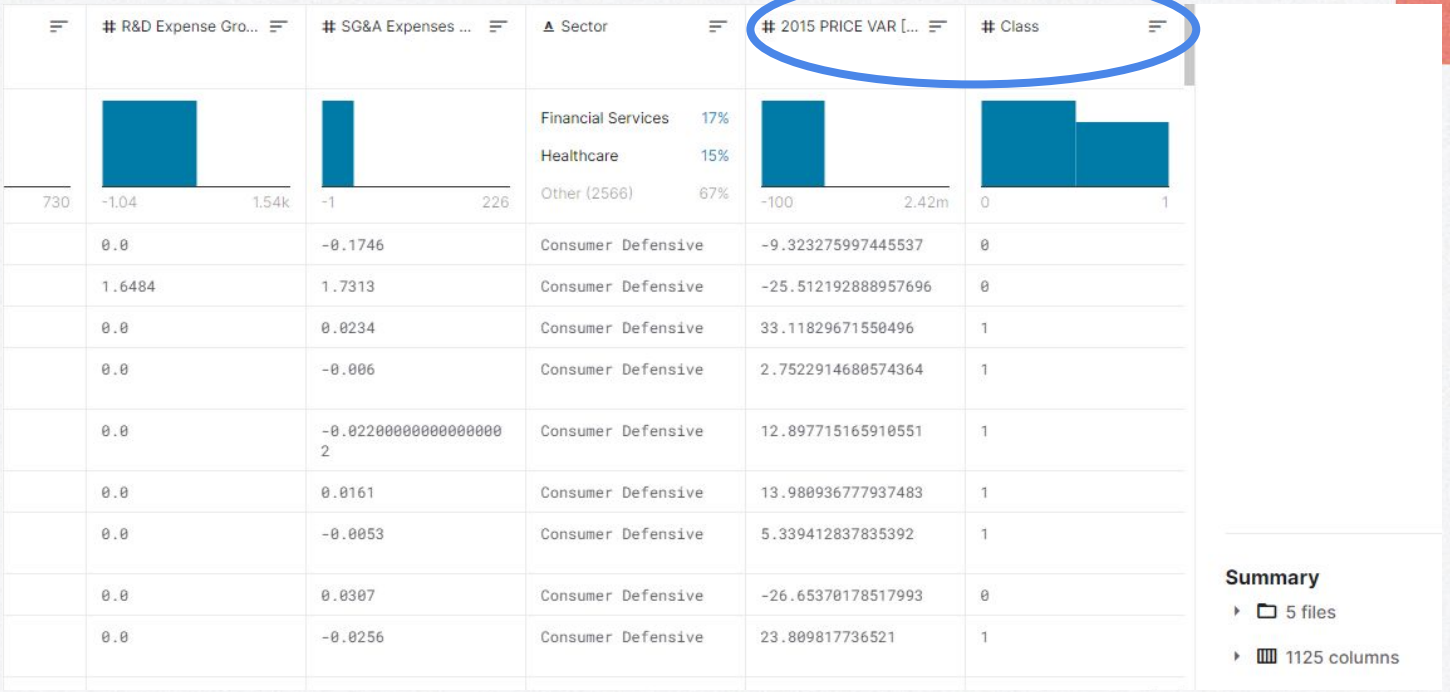


Find the top factors that affect the performance of a stock, so as to develop a safe stock investing strategy to ensure positive rates of return for investors while minimizing losses

2

Exploratory data analysis

Understanding the dataset



	# R&D Expense Gro...	# SG&A Expenses ...	Sector	# 2015 PRICE VAR [...	# Class
	730	-1.04 1.54k	-1 226	Financial Services 17% Healthcare 15% Other (2566) 67%	-100 2.42m	0 1
	0.0	-0.1746	Consumer Defensive	-9.323275997445537	0
	1.6484	1.7313	Consumer Defensive	-25.512192888957696	0
	0.0	0.0234	Consumer Defensive	33.11829671550496	1
	0.0	-0.006	Consumer Defensive	2.7522914680574364	1
	0.0	-0.022000000000000002	Consumer Defensive	12.897715165910551	1
	0.0	0.0161	Consumer Defensive	13.980936777937483	1
	0.0	-0.0053	Consumer Defensive	5.339412837835392	1
	0.0	0.0307	Consumer Defensive	-26.65370178517993	0
	0.0	-0.0256	Consumer Defensive	23.809817736521	1

Summary

- 5 files
- 1125 columns

Understanding the dataset

In the dataset:

1) 200+ financial indicators of a year

2) Price Var [%] :

- Measure of percent price variation of the next year

3) Class :

- Binary classification
- “0” for stocks which value will drop (don’t buy)
- “1” for stocks which value will increase (buy)

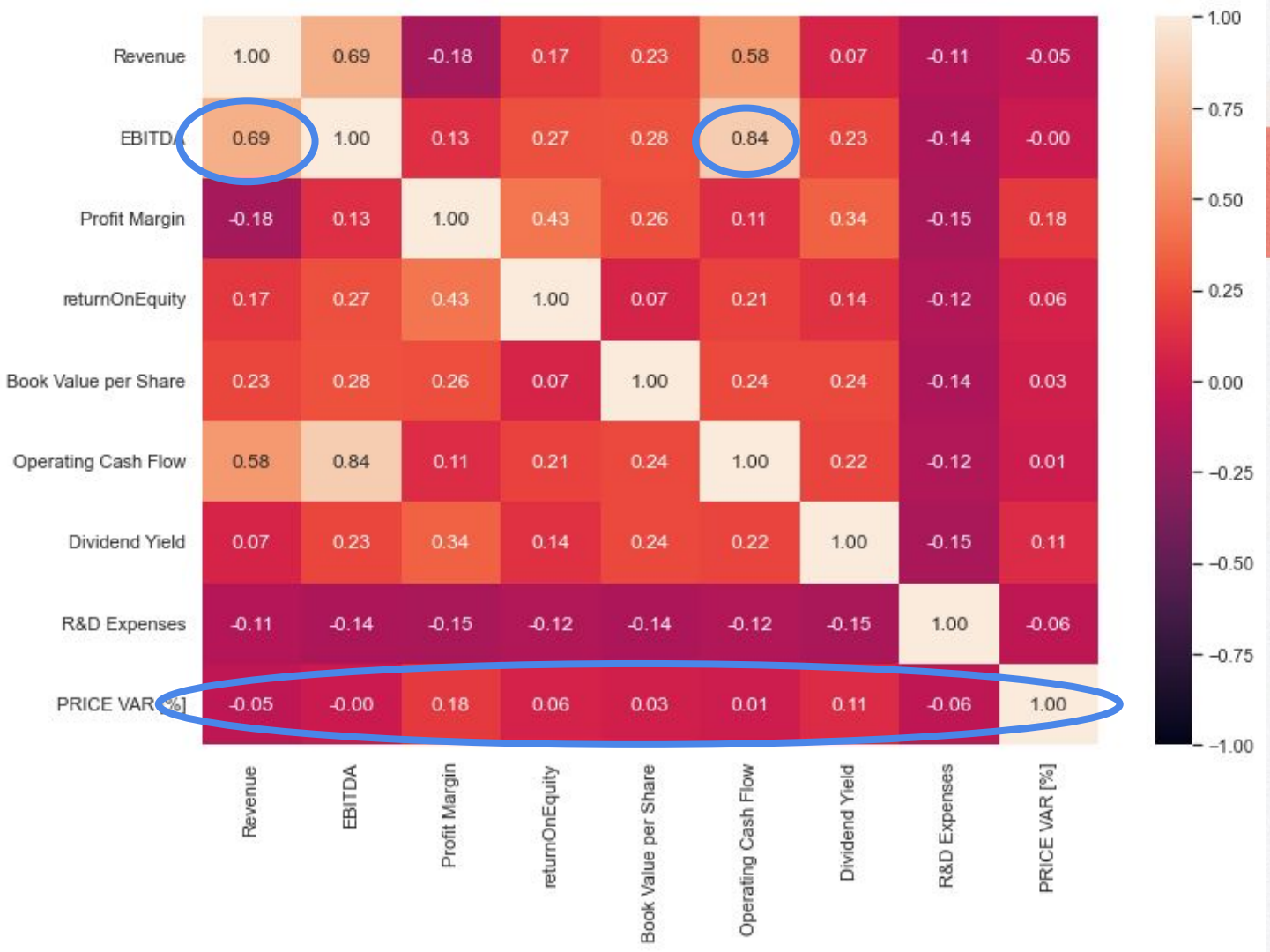
Data preparation and cleaning

- Check for missing values
- Fill missing values with median
- Remove outliers
- Select a few indicators to work with

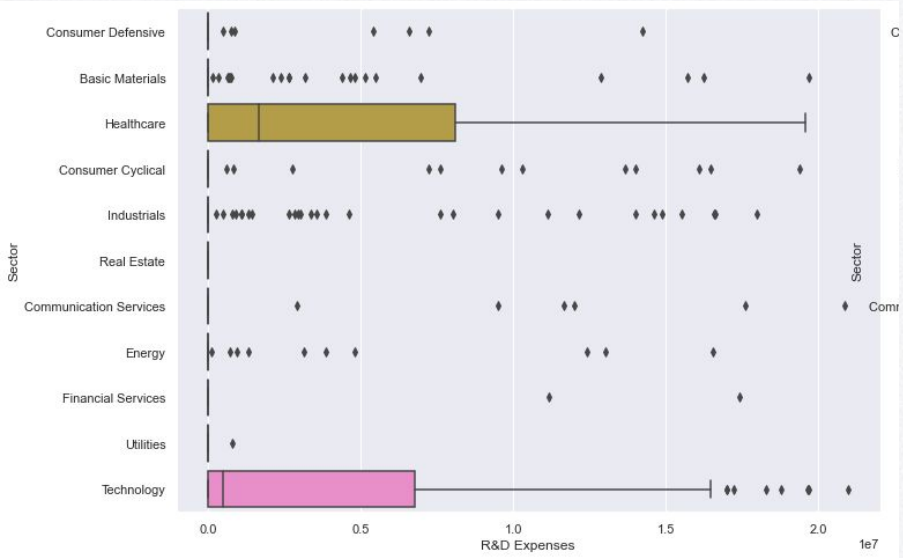
Out[13]:

	Revenue	EBITDA	Profit Margin	returnOnEquity	Book Value per Share	Operating Cash Flow	Dividend Yield	R&D Expenses	PRICE VAR [%]	Class	Sector
14	5.727000e+09	683400000.0	0.066	0.2041	6.706	634100000.0	0.0173	0.0	18.603845	1	Consumer Defensive
22	4.551600e+09	241900000.0	-0.021	-0.1154	2.211	536500000.0	0.0117	0.0	26.681241	1	Consumer Defensive
23	2.464867e+09	771439000.0	0.196	0.3189	3.020	597491000.0	0.0000	0.0	37.721889	1	Consumer Defensive
30	3.297600e+09	743500000.0	0.126	0.1969	7.779	540300000.0	0.0157	0.0	10.792636	1	Consumer Defensive
32	5.973810e+08	183876000.0	0.122	0.1289	10.909	111582000.0	0.0000	0.0	49.607672	1	Consumer Defensive
...	...	...	...	...	...	...	...	...	...	...	...
3802	1.185080e+08	9650000.0	0.034	0.1324	4.249	7612000.0	0.0000	0.0	-2.453386	0	Technology
3803	4.952987e+07	-53213.0	-0.002	-0.0097	4.505	523987.0	0.0000	0.0	29.362884	1	Technology
3804	1.532400e+08	20887000.0	0.085	0.3646	2.426	-1587000.0	0.0000	11326000.0	-31.167763	0	Technology
3806	3.407580e+08	8512000.0	0.017	0.1456	8.489	5745000.0	0.0395	0.0	7.779579	1	Technology
3807	4.033737e+07	4959141.0	0.060	0.0721	1.645	4012331.0	0.0000	3379920.0	-34.099613	0	Technology

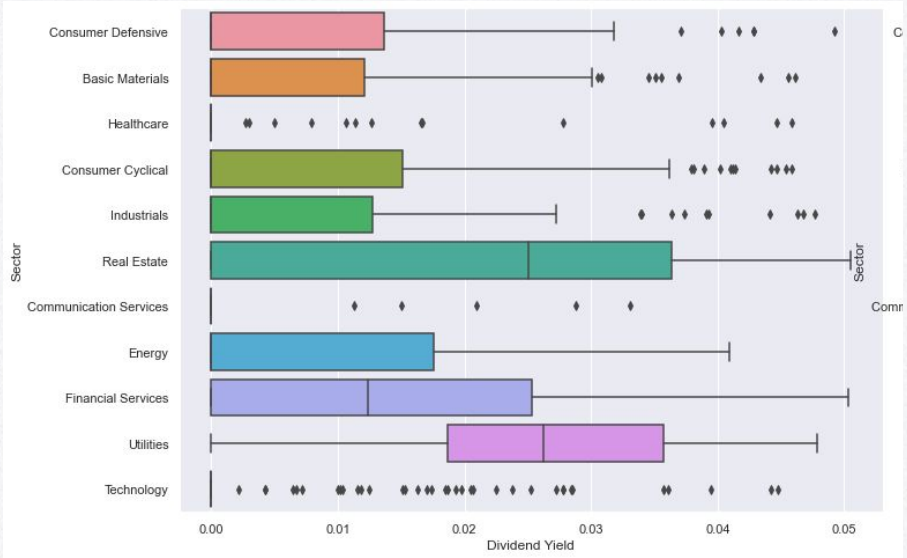
1521 rows × 11 columns

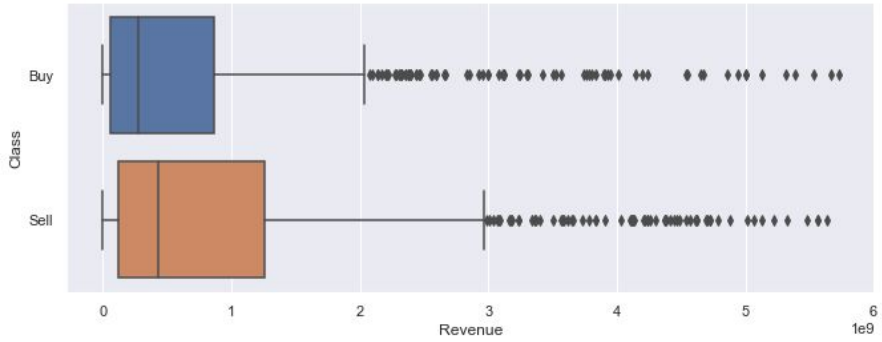

Revenue 1.00 0.69 -0.18 0.17 0.23 0.58 0.07 -0.11 -0.05
EBITDA 0.69 1.00 0.13 0.27 0.28 0.84 0.23 -0.14 -0.00
Profit Margin -0.18 0.13 1.00 0.43 0.26 0.11 0.34 -0.15 0.18
returnOnEquity 0.17 0.27 0.43 1.00 0.07 0.21 0.14 -0.12 0.06
Book Value per Share 0.23 0.28 0.26 0.07 1.00 0.24 0.24 -0.14 0.03
Operating Cash Flow 0.58 0.84 0.11 0.21 0.24 1.00 0.22 -0.12 0.01
Dividend Yield 0.07 0.23 0.34 0.14 0.24 0.22 1.00 -0.15 0.11
R&D Expenses -0.11 -0.14 -0.15 -0.12 -0.14 -0.12 -0.15 1.00 -0.06
PRICE VAR [%] -0.05 -0.00 0.18 0.06 0.03 0.01 0.11 -0.06 1.00
Revenue
EBITDA
Profit Margin
returnOnEquity
Book Value per Share
Operating Cash Flow
Dividend Yield
R&D Expenses
PRICE VAR [%]
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

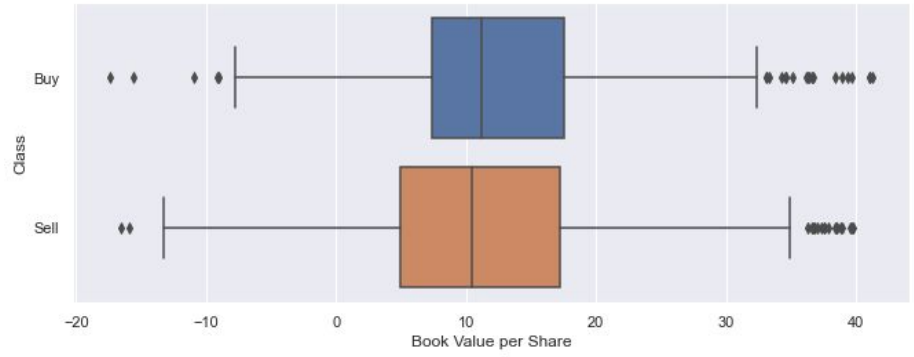
R&D expenses against sector

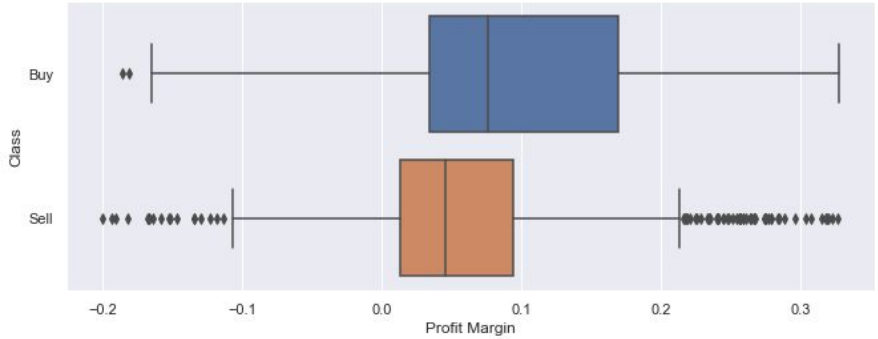


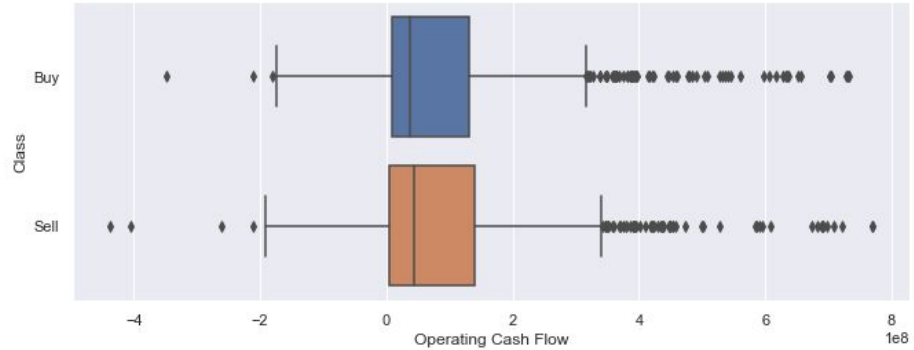
Dividend yield against sector

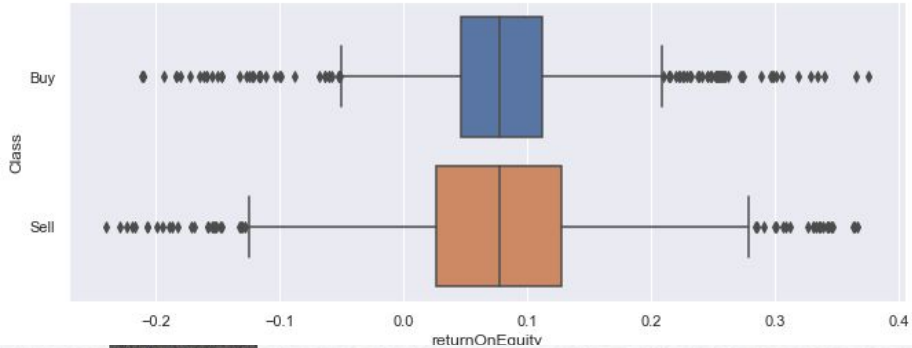



Buy
Sell
Class
0
1
2
3
4
5
6
Revenue
1e9


Buy
Sell
Class
−20
−10
0
10
20
30
40
Book Value per Share


Buy
Sell
Class
−0.2
−0.1
0.0
0.1
0.2
0.3
Profit Margin


Buy
Sell
Class
−4
−2
0
2
4
6
8
1e8
Operating Cash Flow

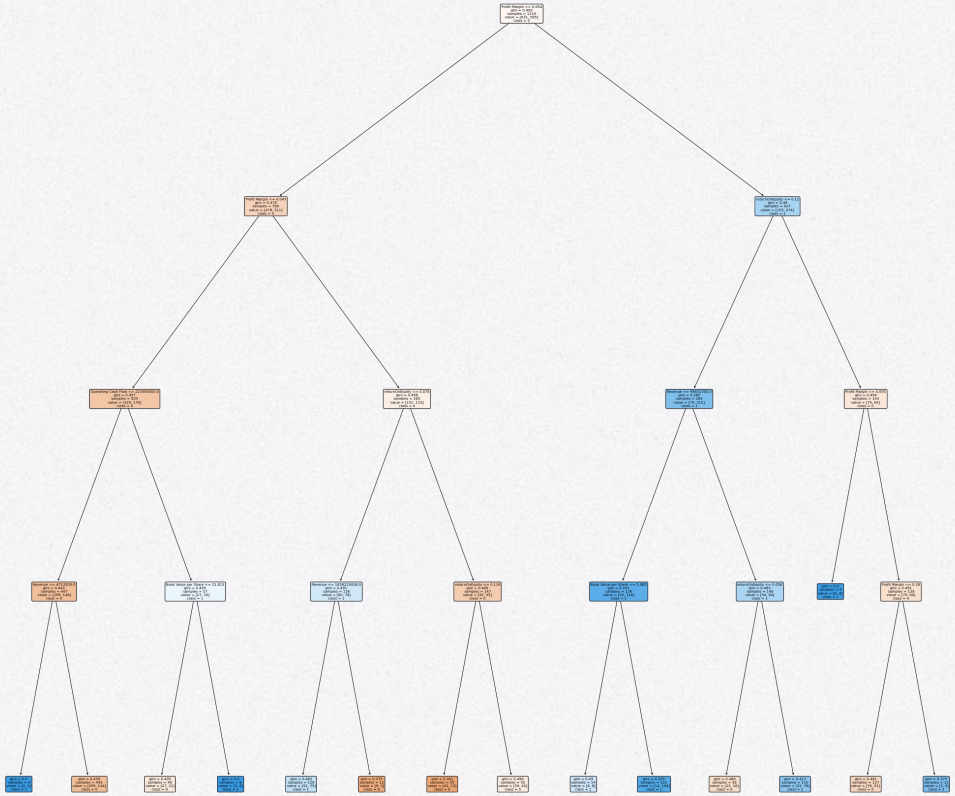

Buy
Sell
Class
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
returnOnEquity

3

Machine learning

- MultiVariate Decision Tree
- Random Forest
- GridsearchCV

Decision Tree

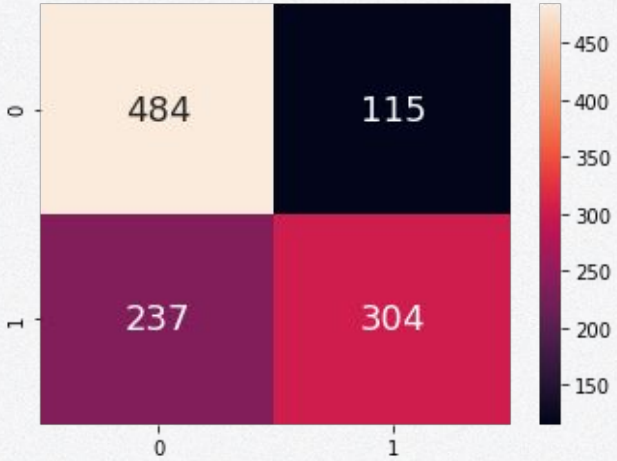


- A Decision Tree is used to classify the response variables against Class
- A classification accuracy of 0.67 is obtained on the train set and 0.57 for the test set
- (A more detailed image will be in our GitHub repository)

Random Forest

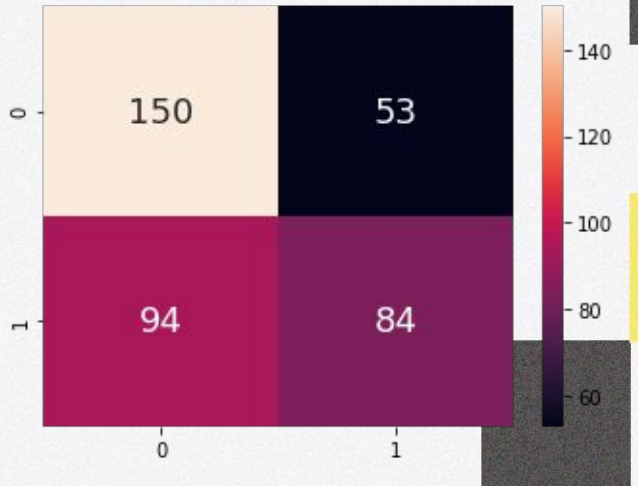
Train Dataset

Classification accuracy: 0.691
True Positive Rate: 0.561
True Negative Rate: 0.808
False Positive Rate: 0.438
False Negative Rate: 0.191



Test Dataset

Classification accuracy: 0.614
True Positive Rate: 0.472
True Negative Rate: 0.738
False Positive Rate: 0.528
False Negative Rate: 0.261



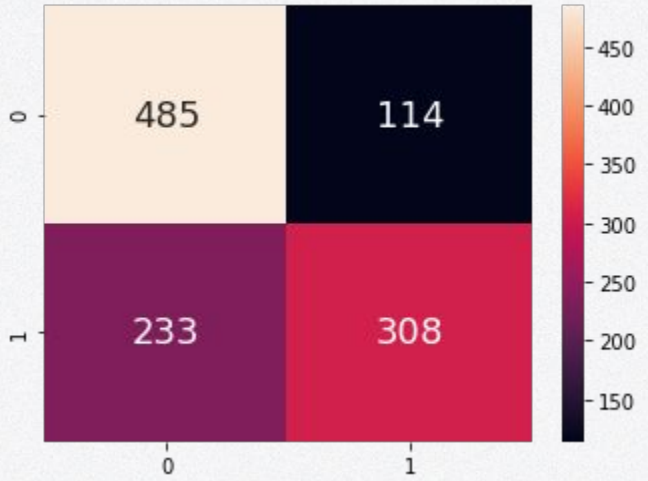
Cross Validation GridSearch

- Find the best hyperparameters for our random forest model
- GridSearch was run with n_estimators over a range of (100,400) and depth over a range of (2,4)
- K-fold Cross Validation, k is set to 5
- Best parameters: n_estimators = 128, depth = 4

After GridSearchCV

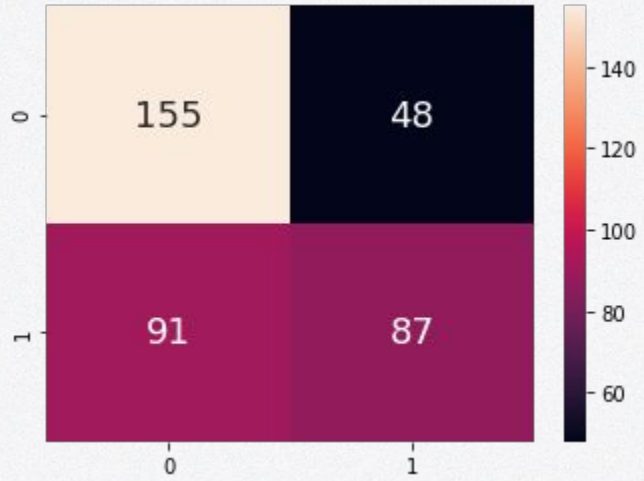
Train Dataset

Classification accuracy: 0.691 -> 0.695
True Positive Rate: 0.561 -> 0.569
True Negative Rate: 0.808 -> 0.809
False Positive Rate: 0.438 -> 0.430
False Negative Rate: 0.191 -> 0.190



Test Dataset

Classification accuracy: 0.614 → 0.634
True Positive Rate: 0.472 → 0.488
True Negative Rate: 0.738 → 0.763
False Positive Rate: 0.528 → 0.511
False Negative Rate: 0.261 → 0.236



4

Applying our model

Research topic

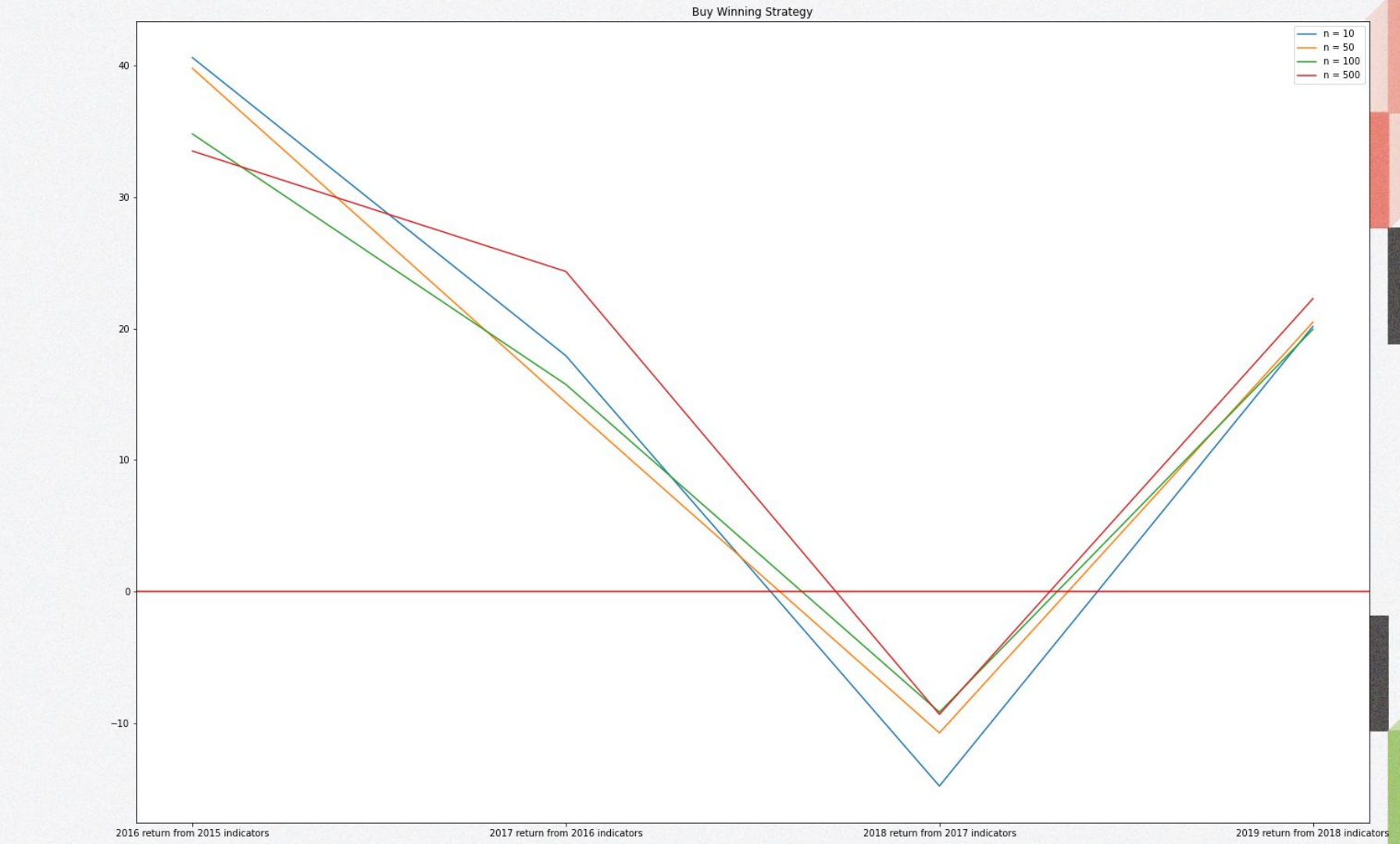
Find the top factors that affect the performance of a stock, so as to develop a safe stock investing strategy to ensure positive rates of return for investors while minimizing losses

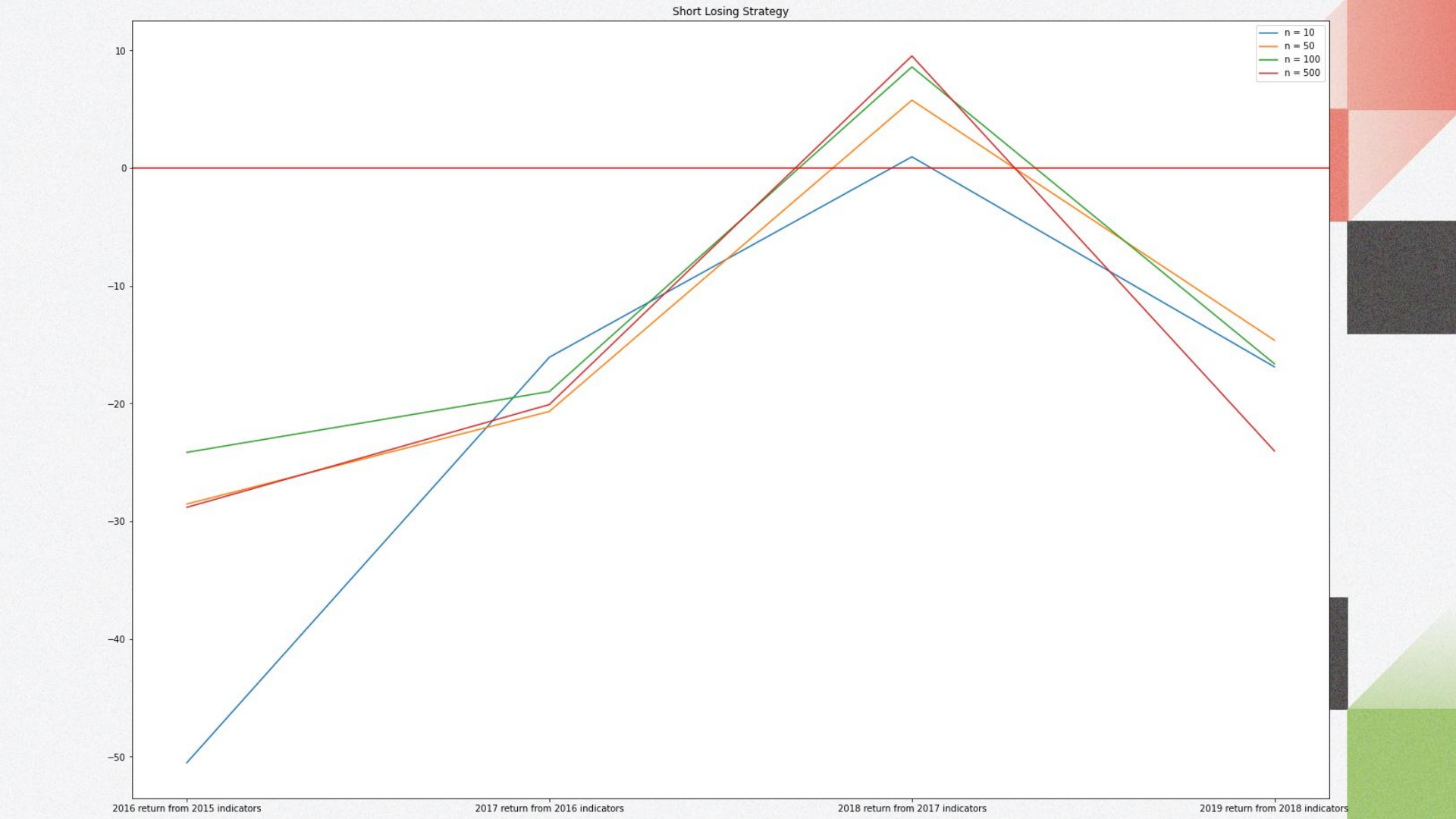
Investing strategies

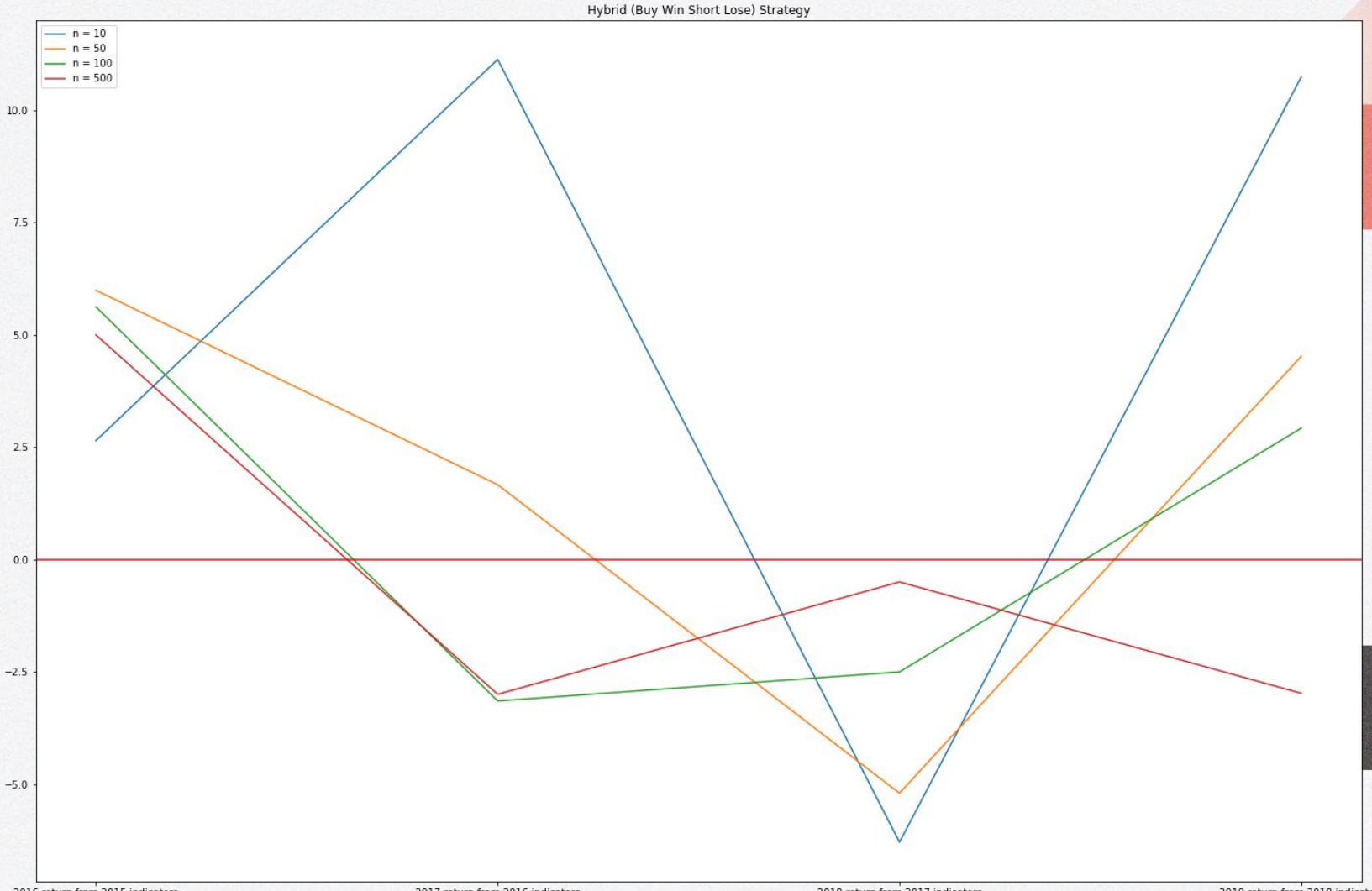
01 Buy top n stocks with highest Class “1” probability

02 Short sell top n stocks with lowest Class “1” probability

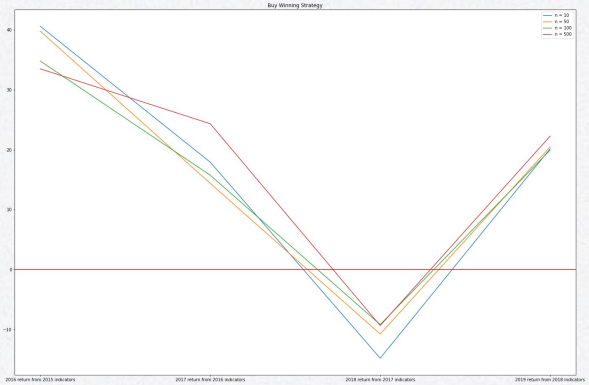
03 Buy top n/2 stocks with highest Class “1” probability and Short sell top n/2 stocks with lowest Class “1” probability


Buy Winning Strategy
n = 10
n = 50
n = 100
n = 500
40
30
20
10
0
−10
2016 return from 2015 indicators
2017 return from 2016 indicators
2018 return from 2017 indicators
2019 return from 2018 indicators

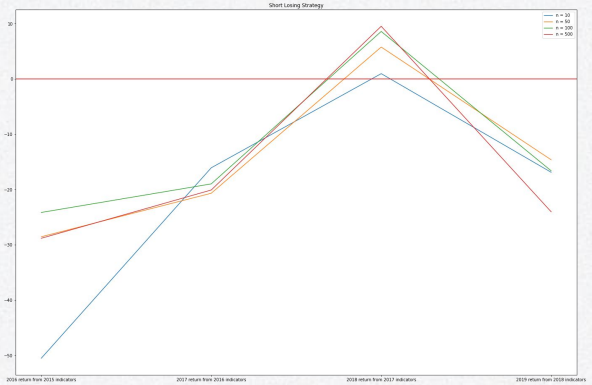

Short Losing Strategy
n = 10
n = 50
n = 100
n = 500
10
0
−10
−20
−30
−40
−50
2016 return from 2015 indicators
2017 return from 2016 indicators
2018 return from 2017 indicators
2019 return from 2018 indicators


Hybrid (Buy Win Short Lose) Strategy
n = 10
n = 50
n = 100
n = 500
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0

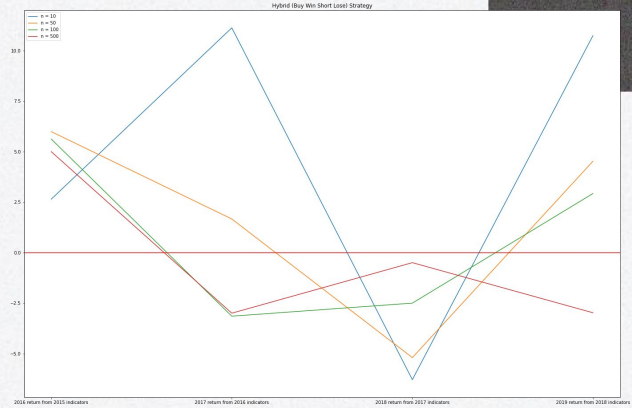
Strategy 1: Buy winning



Strategy 2: Short sell

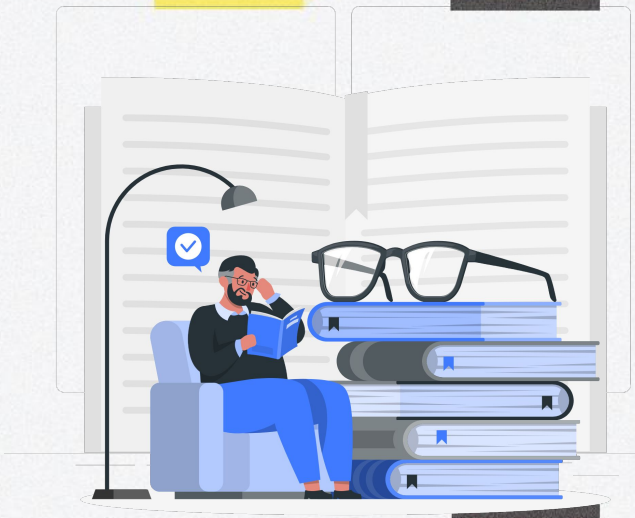


Strategy 3: Combination



Learning something new

- Cross Validation GridSearch
- Time series



Outcomes

With our machine learning model + strategy investors can:

- Make informed decisions on which stocks to invest in
- Maximize profit whilst minimizing losses

Assumptions

- Based on US stocks
- Long term investments

Limitations

- Data used is from 2014-2018 (model might change if there is a drastic change in the market)
- Dataset was too large to pick best predictors
- Focus on maximizing stock price change and neglect stock dividends

Thank you!