T.C.

MARMARA UNIVERSITY

FACULTY of ENGINEERING

CSE4062 Introduction to Data Science and Analytics

Spring 2021

Group #1

Delivery #6: Final Presentation

Title of the Project

Machine Learning Approach to U.S. Stock Investments

Group Members

CSE 150118825 Ahmet Hakan Ekşi ahe9953@gmail.com

ENVE 150215036 Canberk Köroğlu canberkkoroglu@hotmail.com

ENVE 150215045 Erim Varış ermvrs@live.com

Lecturer

Doç. Dr. Murat Can Ganiz

Project Description

Successful predictions of stock movements have always been utterly important in the market. Billions of dollars were spent in infrastructures and R&D departments in order to predict the future outcomes of share-stocks. Can a computer master the art of finding values in stocks and beat the buy-hold strategies [1]?

It's no secret that machines are taking up a bigger and bigger share of investing, but the extent of their influence is approaching shocking proportions. It is as high as 80%, according to one major investing firm [2].

This means so much of stock trading is now in the hands of automated buyers and sellers that the market is increasingly sensitive to headlines and more prone to sharp price swings, many notable investors believe [2][3].

Aim of the project is to predict for US stocks future performances by utilizing 200+ indicators and data throughout 2014-2018. Such is that prediction will decide if one should invest or not in a specific stock.

Data Statistics

Our dataset consists of 5 csv files [1]:

- 2014_Financial_Data.csv ➔ 3808 rows, 225 columns
- 2015_Financial_Data.csv ➔ 4120 rows, 225 columns
- 2016_Financial_Data.csv ➔ 4797 rows, 225 columns
- 2017_Financial_Data.csv ➔ 4960 rows, 225 columns
- 2018_Financial_Data.csv ➔ 4392 rows, 225 columns

First column indicates specific stock name.

The attribute named "Sector" indicates sector of corresponding stock, nominal type.

The attribute named "PRICE VAR [%]" indicates the percent price variation of the corresponding stock for the next year, numeric type.

The attribute named "class" is for classification and it's generated from "PRICE VAR [%]" column, binary type. If "PRICE VAR [%]" positive for corresponding stock then class attribute is 1, else is 0.

Other attributes are financial indicators, numeric type.

	Feature name	Description	Type	Min	Max	Avg.	Std.Dev.	Entropy	# of values	missing values %
0	Stock	stock code	text						22077	0
1	Revenue	financial indicator	numeric_continuous	-627616000	1,88689E+12	5161618858	31973144008	9,6721833	20906	5,3
2	Revenue Growth	financial indicator	numeric_continuous	-12,7693	42138,6639	3,622214228	312,6481703	8,7124784	19989	9,46
3	Cost of Revenue	financial indicator	numeric_continuous	-2986887895	1,58153E+12	3258565393	25830920898	8,324618	20306	8,02
4	Gross Profit	financial indicator	numeric_continuous	-12808000000	4,6216E+11	1970452467	8735750257	9,662554	20870	5,47
5	R&D Expenses	financial indicator	numeric_continuous	-109800000	28837000000	103333292,3	767606165,7	4,607591	19939	9,68
6	SG&A Expense	financial indicator	numeric_continuous	-140159420,3	1,85683E+11	869927885,8	3804283410	9,7665056	20408	7,56
7	Operating Expenses	financial indicator	numeric_continuous	-5495511688	3,05605E+11	1368669853	5662983943	9,9325747	20375	7,71
8	Operating Income	financial indicator	numeric_continuous	-19339000000	1,56554E+11	589697890,7	2976453997	9,9155095	20976	4,99
9	Interest Expense	financial indicator	numeric_continuous	-1710953646	31523000000	97789390,14	499654291,6	7,713734	20358	7,79
10	Earnings before Tax	financial indicator	numeric_continuous	-21772000000	87205000000	492500290,6	2484345450	9,9220695	20713	6,18
11	Income Tax Expense	financial indicator	numeric_continuous	-7,38E+11	8,49E+11	130002043,9	7962080411	8,7095261	20489	7,19
12	Net Income - Non-Controlling int	financial indicator	numeric_continuous	-1587227414	6430813535	13390062,13	143753327,2	4,1062104	19818	10,23
13	Net Income - Discontinued ops	financial indicator	numeric_continuous	-15914500000	8368000000	-3240313,279	242497021,3	2,3022106	19818	10,23
14	Net Income	financial indicator	numeric_continuous	-23045000000	2,33997E+11	388672668,9	2643759192	9,8886958	20512	7,09
15	Preferred Dividends	financial indicator	numeric_continuous	-161000000	2741588000	4673906,343	53287995,57	2,2367805	19818	10,23
16	Net Income Com	financial indicator	numeric_continuous	-23045000000	2,33997E+11	387125353,8	2633920242	9,8818891	20686	6,3
17	EPS	financial indicator	numeric_continuous	-101870898,1	8028004,014	-10657,47956	896097,6629	7,508129	20776	5,89
18	EPS Diluted	financial indicator	numeric_continuous	-101870898,1	6624003,312	-10735,82333	895348,7469	7,4963166	20785	5,85
19	Weighted Average Shs Out	financial indicator	numeric_continuous	0	1,11292E+11	263176678,5	2046155695	9,8897837	20583	6,77
20	Weighted Average Shs Out (Dil)	financial indicator	numeric_continuous	0	1,11292E+11	266492638,3	2136719667	9,8966135	20140	8,77
21	Dividend per Share	financial indicator	numeric_continuous	0	10100,664	1,215391588	72,20095121	4,4100887	19818	10,23
22	Gross Margin	financial indicator	numeric_continuous	-74,3191	31	0,487843946	0,945601415	7,7400683	20878	5,43
23	EBITDA Margin	financial indicator	numeric_continuous	-24207	3090,87	-8,88059069	239,6253606	7,8290687	19630	11,08
24	EBIT Margin	financial indicator	numeric_continuous	-24242	1056,4658	-7,258787199	217,7949818	8,8118567	20403	7,58
25	Profit Margin	financial indicator	numeric_continuous	-24414	3090,87	-9,307057101	243,2054675	7,6178844	19633	11,07
26	Free Cash Flow margin	financial indicator	numeric_continuous	-23256	689,8297	-6,293523163	204,8696797	8,8038248	19786	10,38
27	EBITDA	financial indicator	numeric_continuous	-16484000000	2,33721E+11	928982019,4	4464530765	9,9391423	20323	7,94
28	EBIT	financial indicator	numeric_continuous	-18713000000	2,33997E+11	616190279,5	3324923078	9,9046551	20652	6,45
29	Consolidated Income	financial indicator	numeric_continuous	-23045000000	2,33997E+11	401690839,1	2667626334	9,894252	20510	7,1
30	Earnings Before Tax Margin	financial indicator	numeric_continuous	-24202	1056,4658	-7,214756216	215,1260637	8,7144381	20882	5,41
31	Net Profit Margin	financial indicator	numeric_continuous	-24414	1056,4658	-7,28051256	218,2950794	8,6915063	20355	7,8
32	Cash and cash equivalents	financial indicator	numeric_continuous	0	9,88E+11	1894710353	21767059845	9,8164297	20898	5,34
33	Short-term investments	financial indicator	numeric_continuous	0	8,51E+11	1388023358	19790342267	4,7598229	19365	12,28
34	Cash and short-term investments	financial indicator	numeric_continuous	0	9,88E+11	3088980159	32984076103	9,8889869	19602	11,21
35	Receivables	financial indicator	numeric_continuous	0	9,2E+11	976118398,3	9746251737	8,6234362	20972	5,01
36	Inventories	financial indicator	numeric_continuous	0	9,12E+11	504002341,9	9014646215	6,1520328	20246	8,29
37	Total current assets	financial indicator	numeric_continuous	0	1,26611E+12	5381980048	46807037160	9,9550924	19555	11,42
38	Property, Plant & Equipment Net	financial indicator	numeric_continuous	0	9,72313E+11	2714136117	17669352472	9,5890445	20418	7,51
39	Goodwill and Intangible Assets	financial indicator	numeric_continuous	0	3,8618E+11	1731173726	8752365896	7,7088979	20640	6,51
40	Long-term investments	financial indicator	numeric_continuous	-80000000	9,97E+11	2930936625	34899214793	5,5176125	19444	11,93
41	Tax assets	financial indicator	numeric_continuous	0	66501000000	152155057,5	1216324191	5,4985001	20094	8,98
42	Total non-current assets	financial indicator	numeric_continuous	0	5,46381E+16	3,40901E+12	4,31212E+14	9,9585303	16055	27,28
43	Total assets	financial indicator	numeric_continuous	0	2,0357E+13	20407238234	1,86069E+11	9,972989	20772	5,91
44	Payables	financial indicator	numeric_continuous	-20586351351	4,24269E+11	915984766,1	8495931343	9,1167398	20407	7,56
45	Short-term debt	financial indicator	numeric_continuous	-1374631268	2,75835E+11	597883917,1	5877180783	6,5883925	19371	12,26
46	Total current liabilities	financial indicator	numeric_continuous	-21077918782	2,09531E+12	7412083756	75652859628	9,933286	19549	11,45
47	Long-term debt	financial indicator	numeric_continuous	-8446486486	4,41253E+12	3294063118	35331276261	8,0306539	20169	8,64
48	Total debt	financial indicator	numeric_continuous	-9289864865	4,41253E+12	4558784756	40661822200	8,6129629	20369	7,74
49	Deferred revenue	financial indicator	numeric_continuous	-23639000	1,34879E+16	6,71468E+11	9,51574E+13	4,3365589	20091	9
50	Tax Liabilities	financial indicator	numeric_continuous	-1783580247	4,93831E+15	2,45978E+11	3,48269E+13	6,058695	20106	8,93
51	Deposit Liabilities	financial indicator	numeric_continuous	0	1,47067E+12	5079495680	59229049348	2,6629876	19807	10,28
52	Total non-current liabilities	financial indicator	numeric_continuous	-11070540541	6,17982E+15	3,88673E+11	4,87933E+13	9,718893	16041	27,34
53	Total liabilities	financial indicator	numeric_continuous	-7872214182	1,85654E+13	16089038257	1,67356E+11	9,9617698	20681	6,32
54	Other comprehensive income	financial indicator	numeric_continuous	-94785000000	1,70852E+15	83616099617	1,19618E+13	7,7625418	20401	7,59
55	Retained earnings (deficit)	financial indicator	numeric_continuous	-2,8E+11	7,18E+11	2106631255	14651964603	9,8206735	20910	5,29
56	Total shareholders equity	financial indicator	numeric_continuous	-22884000000	1,46573E+12	3622994202	19618004420	9,9493114	20705	6,21
57	Investments	financial indicator	numeric_continuous	-5283298724	2,08023E+12	9033214981	84876067375	6,854398	20159	8,69
58	Net Debt	financial indicator	numeric_continuous	-4,44365E+11	4,41253E+12	1839178489	39555328235	9,9515545	15218	31,07
59	Other Assets	financial indicator	numeric_continuous	-9,11987E+11	7,38599E+11	630986068,3	19301321722	9,7079455	15321	30,6
60	Other Liabilities	financial indicator	numeric_continuous	-99229000000	1,86621E+12	5885822448	69958134200	9,5944256	19735	10,61
61	Depreciation & Amortization	financial indicator	numeric_continuous	-83360000	7,51E+11	342114899,8	5474084531	9,4327692	20549	6,92
62	Stock-based compensation	financial indicator	numeric_continuous	-137000000	9353000000	30812480,41	213895727,8	8,4146136	20344	7,85
63	Operating Cash Flow	financial indicator	numeric_continuous	-3,4E+11	9,6E+11	1073238607	16150911846	9,8904069	20909	5,29
64	Capital Expenditure	financial indicator	numeric_continuous	-1,45585E+11	5822595000	-391562278,8	2559380997	9,3913205	20554	6,9
65	Acquisitions and disposals	financial indicator	numeric_continuous	-51000000000	69871000000	-109483677,8	1445095708	5,6651598	20044	9,21
66	Investment purchases and sales	financial indicator	numeric_continuous	-1,93007E+11	1,49872E+11	-208487721,3	4211286025	6,2973138	20563	6,86
67	Investing Cash flow	financial indicator	numeric_continuous	-1,97993E+11	1,44591E+11	-701654255,4	4969187220	9,7648729	20672	6,36
68	Issuance (repayment) of debt	financial indicator	numeric_continuous	-1,96353E+11	62675000000	41507146,9	3050560787	8,4749087	20409	7,56
69	Issuance (buybacks) of shares	financial indicator	numeric_continuous	-72069000000	1,44401E+11	-118415415,6	1544792840	8,4167144	20135	8,8
70	Dividend payments	financial indicator	numeric_continuous	-16027206771	0	-182172267,1	802303602,5	5,5747875	19921	9,77
71	Financing Cash Flow	financial indicator	numeric_continuous	-8,88E+11	9,98E+11	-189527124,9	15440104591	9,8088457	21105	4,4
72	Effect of forex changes on cash	financial indicator	numeric_continuous	-1E+12	9,26E+11	-81866622,34	11125329192	5,3047296	20283	8,13
73	Net cash flow / Change in cash	financial indicator	numeric_continuous	-9,88E+11	9,88E+11	115132222,4	29486136156	9,7145649	21059	4,61
74	Free Cash Flow	financial indicator	numeric_continuous	-1,04239E+11	98870000000	453355548,5	3164032680	9,9084494	20438	7,42
75	Net Cash/Marketcap	financial indicator	numeric_continuous	-4943,5529	501,7801	-0,957732926	43,20719745	9,2474578	18721	15,2
76	priceBookValueRatio	financial indicator	numeric_continuous	0	108458749,3	20347,0531	1294364,862	9,5263397	17084	22,62
77	priceToBookRatio	financial indicator	numeric_continuous	0	108458749,3	20344,79097	1294364,752	8,5494813	17084	22,62
78	priceToSalesRatio	financial indicator	numeric_continuous	0	377315,7961	88,04570699	3499,317796	9,5413237	19794	10,34
79	priceEarningsRatio	financial indicator	numeric_continuous	0	105828,7129	37,84104833	853,4734956	7,2217996	19801	10,31
80	priceToFreeCashFlowsRatio	financial indicator	numeric_continuous	0	60328,7741	36,45509543	645,7326721	7,1538587	19800	10,31
81	priceToOperatingCashFlowsRatio	financial indicator	numeric_continuous	-108768,6767	554659,9637	53,97084693	4242,64938	9,8523834	18857	14,59
82	priceCashFlowRatio	financial indicator	numeric_continuous	0	554582,1706	63,9417902	4537,222149	9,8490612	15013	32
83	priceEarningsToGrowthRatio	financial indicator	numeric_continuous	0	100419,288	58,40742668	954,0306986	9,8738942	12998	41,12
84	priceSalesRatio	financial indicator	numeric_continuous	0	377315,7961	78,02207275	3429,496553	9,7666659	18720	15,21
85	dividendYield	financial indicator	numeric_continuous	0	527,926557	0,286129491	8,332549131	6,0343454	18779	14,94
86	enterpriseValueMultiple	financial indicator	numeric_continuous	0	47022,618	36,45482558	571,1146465	10,001333	15133	31,45
87	priceFairValue	financial indicator	numeric_continuous	0	94309693,22	11170,94842	867520,6729	9,8142384	18503	16,19
88	ebitperRevenue	financial indicator	numeric_continuous	-24242	1056,465753	-7,751022894	224,4290589	9,9067423	19215	12,96
89	ebtperEBIT	financial indicator	numeric_continuous	-983,2076923	2263,870095	0,634464077	21,47880022	8,7069032	14070	36,27
90	niperEBT	financial indicator	numeric_continuous	-206,0833333	845,5555556	0,864169005	9,535127399	9,7375853	13482	38,93
91	grossProfitMargin	financial indicator	numeric_continuous	-223,4827586	34,81074705	0,492200918	2,023480525	8,759676	19215	12,96
92	operatingProfitMargin	financial indicator	numeric_continuous	1	1	1	0	1,4175145	19215	12,96
93	pretaxProfitMargin	financial indicator	numeric_continuous	-24577	198,6842105	-8,055263164	225,9547824	10,000674	19215	12,96
94	netProfitMargin	financial indicator	numeric_continuous	-24414	1056,465753	-7,781141393	224,6994764	9,8934873	19215	12,96
95	effectiveTaxRate	financial indicator	numeric_continuous	-28384,61538	207,0833333	-2,003388331	244,6404842	9,7303662	13482	38,93
96	returnOnAssets	financial indicator	numeric_continuous	-344,4375	345,631	-0,065041893	8,362792694	9,45363	15542	29,6
97	returnOnEquity	financial indicator	numeric_continuous	-34772,4596	11141141,67	1645,940434	114822,1261	8,9870833	19739	10,59
98	returnOnCapitalEmployed	financial indicator	numeric_continuous	-616,6446	3863,9037	-0,115455492	35,25907382	8,7565089	15568	29,48
99	nIperEBT	financial indicator	numeric_continuous	-206,0833333	845,5555556	0,864169005	9,535127399	9,7375853	13482	38,93
100	eBTperEBIT	financial indicator	numeric_continuous	-983,2076923	2263,870095	0,634464077	21,47880022	8,7069032	14070	36,27
101	eBITperRevenue	financial indicator	numeric_continuous	-24242	1056,465753	-7,751022894	224,4290589	9,9067423	19215	12,96
102	payablesTurnover	financial indicator	numeric_continuous	-41,0958	8650,3158	6,213959035	79,57121081	9,0141761	17166	22,24
103	inventoryTurnover	financial indicator	numeric_continuous	0	95827,7103	31,62418604	727,3354978	6,3488777	19565	11,38
104	fixedAssetTurnover	financial indicator	numeric_continuous	-383,600635	446511,8571	90,56735173	4343,27676	9,7395528	18943	14,2
105	assetTurnover	financial indicator	numeric_continuous	-0,125107358	134,3659007	0,691155034	1,252066062	9,6869601	19933	9,71
106	currentRatio	financial indicator	numeric_continuous	-1,4393	20590,47	7,117724292	183,9802989	9,2348108	19012	13,88
107	quickRatio	financial indicator	numeric_continuous	0	486211,2845	71,87380198	5133,271127	10,000761	19019	13,85
108	cashRatio	financial indicator	numeric_continuous	0	164465,7863	19,00038241	1686,668794	9,9883058	19019	13,85
109	daysOfSalesOutstanding	financial indicator	numeric_continuous	-5182867,019	975,8069	-402,0754599	37239,23272	6,1547221	19835	10,16
110	daysOfInventoryOutstanding	financial indicator	numeric_continuous	-473104,3422	3900070,847	602,3590147	39376,95708	8,6774324	20168	8,65
111	operatingCycle	financial indicator	numeric_continuous	68,4745	1077,7422	176,3765158	220,1073186	10,002292	19	99,91
112	daysOfPayablesOutstanding	financial indicator	numeric_continuous	-207232,4715	1043413,33	340,3941996	9865,965792	8,9073255	19840	10,13
113	cashConversionCycle	financial indicator	numeric_continuous	-262,6048	375,998	43,85739474	131,6433214	10,002292	19	99,91
114	debtRatio	financial indicator	numeric_continuous	-0,4536	133,6571	0,272899231	1,135444257	7,9146035	19763	10,48
115	debtEquityRatio	financial indicator	numeric_continuous	-2586,4359	2131,5128	0,562818691	26,65730955	8,3934875	19763	10,48
116	longtermDebtToCapitalization	financial indicator	numeric_continuous	-1,930983348	91,03812317	0,359341565	1,229186324	8,036944	19688	10,82
117	totalDebtToCapitalization	financial indicator	numeric_continuous	-2,736131494	1040,047427	0,453077447	7,519624712	8,3612482	19785	10,38

116	longtermDebtToCapitalization	financial indicator	numeric_continuous	-1,930983348	91,03812317	0,359341565	1,229186324	8,036944	19688	10,82
117	totalDebtToCapitalization	financial indicator	numeric_continuous	-2,736131494	1040,047427	0,453077447	7,519624712	8,3612482	19785	10,38
118	interestCoverage	financial indicator	numeric_continuous	-223603	361207,3684	-32,19756542	3781,263307	8,0033936	19890	9,91
119	cashFlowToDebtRatio	financial indicator	numeric_continuous	-40783,94035	863187,7715	67,17364654	6929,147805	10,000608	15801	28,43
120	companyEquityMultiplier	financial indicator	numeric_continuous	0,274651772	8740761,333	1232,574918	82998,64085	10,000734	18742	15,11
121	operatingCashFlowPerShare	financial indicator	numeric_continuous	-24942284,42	45351079,16	22904,80882	763234,3879	9,8919232	19652	10,98
122	freeCashFlowPerShare	financial indicator	numeric_continuous	-35588921,28	31932634,45	1507,182944	405433,7599	9,394083	19653	10,98
123	cashPerShare	financial indicator	numeric_continuous	0	212722714,2	83671,83018	3097123,515	9,8022353	19648	11
124	payoutRatio	financial indicator	numeric_continuous	-137,154	479,57	0,323305308	5,961775615	5,0139712	19687	10,83
125	operatingCashFlowSalesRatio	financial indicator	numeric_continuous	-1432432,432	377383,592	-11,8965239	11384,22305	10,000483	19215	12,96
126	freeCashFlowOperatingCashFlowRatio	financial indicator	numeric_continuous	-11849,1722	47,65294118	-0,698318457	95,61367088	9,7651818	15489	29,84
127	cashFlowCoverageRatios	financial indicator	numeric_continuous	-40783,94035	863187,7715	67,17364654	6929,147805	10,000608	15801	28,43
128	shortTermCoverageRatios	financial indicator	numeric_continuous	-16336,66671	863187,7715	254,5972607	10450,29591	10,000651	11333	48,67
129	capitalExpenditureCoverageRatios	financial indicator	numeric_continuous	-46880	313315,6667	-13,60738533	2438,368584	9,9996934	18428	16,53
130	dividendpaidAndCapexCoverageRatios	financial indicator	numeric_continuous	-46880	64384,91318	-33,93732918	764,973657	10,000282	19157	13,23
131	dividendPayoutRatio	financial indicator	numeric_continuous	0	746,7014925	0,821236148	10,69369456	8,2620207	12998	41,12
132	Revenue per Share	financial indicator	numeric_continuous	-3,1677	226232494,9	151639,5713	4206785,552	9,6799405	19634	11,07
133	Net Income per Share	financial indicator	numeric_continuous	-12724877,5	29790471,67	12909,93628	485543,6331	9,8550752	19632	11,07
134	Operating Cash Flow per Share	financial indicator	numeric_continuous	-24942284,42	45351079,16	22890,83379	763001,6673	9,8918604	19664	10,93
135	Free Cash Flow per Share	financial indicator	numeric_continuous	-35588921,28	31932634,45	1506,262306	405310,0343	9,394083	19665	10,93
136	Cash per Share	financial indicator	numeric_continuous	0	212722714,2	83676,08894	3097202,28	9,8023372	19647	11,01
137	Book Value per Share	financial indicator	numeric_continuous	-642647,391	616362318,8	57141,16336	4686263,306	9,7298416	19646	11,01
138	Tangible Book Value per Share	financial indicator	numeric_continuous	0	1095289855	108896,1331	8442215,673	9,8452834	19646	11,01
139	Shareholders Equity per Share	financial indicator	numeric_continuous	-660561,1875	258351241,9	202814,8394	4956871,317	9,9639907	19647	11,01
140	Interest Debt per Share	financial indicator	numeric_continuous	-20,0091	419924840,9	319578,2413	8965197,118	8,6638437	19646	11,01
141	Market Cap	financial indicator	numeric_continuous	0	9,6192E+13	30560346775	1,43251E+12	9,9018012	19006	13,91
142	Enterprise Value	financial indicator	numeric_continuous	-1,48755E+11	1,14583E+12	8478816749	32731994168	9,8081561	17037	22,83
143	PE ratio	financial indicator	numeric_continuous	0	105828,7129	37,90369287	852,9069156	7,2160316	19829	10,18
144	Price to Sales Ratio	financial indicator	numeric_continuous	0	377315,7961	87,95396542	3496,848241	9,5389403	19822	10,21
145	POCF ratio	financial indicator	numeric_continuous	-108768,6767	554659,9637	53,91658057	4240,064551	9,8503022	18880	14,48
146	PFCF ratio	financial indicator	numeric_continuous	0	60328,7741	36,42683441	645,2775071	7,148715	19828	10,19
147	PB ratio	financial indicator	numeric_continuous	0	108458749,3	20328,1694	1293759,281	9,524482	17100	22,54
148	PTB ratio	financial indicator	numeric_continuous	0	108458749,3	20325,9089	1293759,171	8,5457427	17100	22,54
149	EV to Sales	financial indicator	numeric_continuous	-4324,8114	185096,2978	53,71533787	1675,346125	9,4997363	17030	22,86
150	Enterprise Value over EBITDA	financial indicator	numeric_continuous	0	35296,196	21,73107447	308,2014987	8,3296498	17034	22,84
151	EV to Operating cash flow	financial indicator	numeric_continuous	-94684,1767	512579,0405	40,0733114	4017,128639	9,791063	17034	22,84
152	EV to Free cash flow	financial indicator	numeric_continuous	-467933,3193	84994,8903	-19,47912927	3988,397694	9,8003732	17034	22,84
153	Earnings Yield	financial indicator	numeric_continuous	-21375	12673,4414	-0,215082702	219,1828341	8,3875114	19829	10,18
154	Free Cash Flow Yield	financial indicator	numeric_continuous	-25896,7081	293,8176	-1,735278236	191,1058846	8,5980376	18880	14,48
155	Debt to Equity	financial indicator	numeric_continuous	-2586,4359	2131,5128	0,562818691	26,65730955	8,3934875	19763	10,48
156	Debt to Assets	financial indicator	numeric_continuous	-0,4536	133,6571	0,272899231	1,135444257	7,9146035	19763	10,48
157	Net Debt to EBITDA	financial indicator	numeric_continuous	-9616	22776,5294	2,882300912	236,6003344	9,9460444	14362	34,95
158	Current ratio	financial indicator	numeric_continuous	-1,4393	20590,47	6,578997558	190,4739662	9,2353996	15561	29,51
159	Interest Coverage	financial indicator	numeric_continuous	-223603	361207,3684	-32,077465	3778,42755	7,999442	19920	9,77
160	Income Quality	financial indicator	numeric_continuous	-2719,5	17675,5	2,821206668	136,3167276	9,7831828	19857	10,06
161	Dividend Yield	financial indicator	numeric_continuous	0	14,55	0,020425596	0,143164399	4,5387377	19831	10,17
162	Payout Ratio	financial indicator	numeric_continuous	-137,154	479,57	0,323061214	5,958608477	5,0055525	19708	10,73
163	SG&A to Revenue	financial indicator	numeric_continuous	-58,7976	15246	3,50988637	122,9899733	8,6307245	19853	10,07
164	R&D to Revenue	financial indicator	numeric_continuous	-0,7317	12991	4,445070404	128,8361591	4,0602287	19800	10,31
165	Intangibles to Total Assets	financial indicator	numeric_continuous	0	0,9953	0,154727445	0,209260006	7,0200682	19880	9,95
166	Capex to Operating Cash Flow	financial indicator	numeric_continuous	0	11850,1722	1,386501778	84,31568267	7,4266056	19967	9,56
167	Capex to Revenue	financial indicator	numeric_continuous	-4310	773,6667	0,12848176	31,98942754	7,6309164	19868	10,01
168	Capex to Depreciation	financial indicator	numeric_continuous	-28088,9763	35298,3333	-5,526198988	372,6075808	9,2122073	19852	10,08
169	Stock-based compensation to Revenue	financial indicator	numeric_continuous	-5,7684	3700	1,079116527	33,51364236	6,3588676	19774	10,43
170	Graham Number	financial indicator	numeric_continuous	0	8851965,516	737,5825845	69208,74595	7,8487315	17094	22,57
171	ROIC	financial indicator	numeric_continuous	-616,6446	3863,9037	-0,115455492	35,25907382	8,7565089	15568	29,48
172	Return on Tangible Assets	financial indicator	numeric_continuous	-344,4375	345,631	-0,065041893	8,362792694	9,45363	15542	29,6
173	Graham Net-Net	financial indicator	numeric_continuous	-5474,4406	283,0389	-1,659189271	45,88340617	9,3533211	18911	14,34
174	Working Capital	financial indicator	numeric_continuous	-4,61001E+15	1,47214E+13	-2,88121E+11	3,69499E+13	9,9755616	15567	29,49
175	Tangible Asset Value	financial indicator	numeric_continuous	-24215000000	2,56818E+12	16164888181	1,10548E+11	9,993728	19880	9,95
176	Net Current Asset Value	financial indicator	numeric_continuous	-1,37794E+12	80512000000	-7585341634	56846491891	9,9887559	19046	13,73
177	Invested Capital	financial indicator	numeric_continuous	-8,26288E+15	3,0528E+12	-4,00379E+11	5,87784E+13	9,9939914	19762	10,49
178	Average Receivables	financial indicator	numeric_continuous	0	4,60003E+11	914402488,4	6902478102	8,8585767	19948	9,64
179	Average Payables	financial indicator	numeric_continuous	-20369190904	7,1236E+11	950983534,7	10140227058	9,2630179	19651	10,99
180	Average Inventory	financial indicator	numeric_continuous	0	4,56E+11	392181094,2	3875869753	6,3487377	19646	11,01
181	Days Sales Outstanding	financial indicator	numeric_continuous	-473104,3422	3900070,847	747,5486148	42585,473	8,6779108	20238	8,33
182	Days Payables Outstanding	financial indicator	numeric_continuous	-207232,4715	1043413,33	340,4278665	9866,462561	8,9077961	19838	10,14
183	Days of Inventory on Hand	financial indicator	numeric_continuous	-5182867,019	975,8069	-402,1160061	37241,11019	6,1556404	19833	10,16
184	Receivables Turnover	financial indicator	numeric_continuous	-8698,6061	164428,5	44,77226775	1687,815259	8,7340214	19864	10,02
185	Payables Turnover	financial indicator	numeric_continuous	-41,0958	8650,3158	6,213959035	79,57121081	9,0141761	17166	22,24
186	Inventory Turnover	financial indicator	numeric_continuous	0	95827,7103	31,62418604	727,3354978	6,3488777	19565	11,38
187	ROE	financial indicator	numeric_continuous	-34772,4596	11141141,67	1646,107283	114827,9427	8,9871461	19737	10,6
188	Capex per Share	financial indicator	numeric_continuous	-73354000	1255872,972	-18800,31495	662354,255	9,1777569	19664	10,93
189	Gross Profit Growth	financial indicator	numeric_continuous	-5536,4833	336767,8	18,95154469	2380,55765	8,8223452	20069	9,1
190	EBIT Growth	financial indicator	numeric_continuous	-16888	20598,4286	0,807739501	233,2634268	9,5673168	19736	10,6
191	Operating Income Growth	financial indicator	numeric_continuous	-18881	13545,3387	0,462037376	188,4220477	9,5187312	20109	8,91
192	Net Income Growth	financial indicator	numeric_continuous	-33777	3159	-2,230353811	248,959622	9,6459386	19695	10,79
193	EPS Growth	financial indicator	numeric_continuous	-2125	4378,0492	0,28336706	37,74877337	9,2935033	19894	9,89
194	EPS Diluted Growth	financial indicator	numeric_continuous	-2125	39679767,04	1995,812966	281395,5581	9,2892832	19884	9,93
195	Weighted Average Shares Growth	financial indicator	numeric_continuous	-0,9999	223089,5	13,20897567	1611,269005	8,1611523	19269	12,72
196	Weighted Average Shares Diluted Growth	financial indicator	numeric_continuous	-1	3378329,472	189,6109702	25075,11989	8,3011882	18155	17,77
197	Dividends per Share Growth	financial indicator	numeric_continuous	-1	372,5714	0,15822221	3,502421548	4,2221884	19608	11,18
198	Operating Cash Flow growth	financial indicator	numeric_continuous	-18001	455999	64,1095843	5239,628223	9,5373236	20014	9,34
199	Free Cash Flow growth	financial indicator	numeric_continuous	-83431,4286	8030,3333	-3,544641565	603,6085433	9,6935924	19603	11,21
200	10Y Revenue Growth (per Share)	financial indicator	numeric_continuous	-1	1,747	0,021050618	0,156278714	9,1178861	12545	43,18
201	5Y Revenue Growth (per Share)	financial indicator	numeric_continuous	-1	8,9762	0,028228113	0,23575332	8,8964722	15676	28,99
202	3Y Revenue Growth (per Share)	financial indicator	numeric_continuous	-1	43,9195	0,04070423	0,508011672	8,756619	17611	20,23
203	10Y Operating CF Growth (per Share)	financial indicator	numeric_continuous	-0,7618	1,7327	0,037765569	0,115318076	8,3930886	12599	42,93
204	5Y Operating CF Growth (per Share)	financial indicator	numeric_continuous	-1	4,1862	0,04725043	0,197747276	7,9154553	15713	28,83
205	3Y Operating CF Growth (per Share)	financial indicator	numeric_continuous	-1	9,6961	0,064219315	0,316339779	7,5719552	17644	20,08
206	10Y Net Income Growth (per Share)	financial indicator	numeric_continuous	-1	1,3372	0,030847493	0,11101356	7,9262918	12545	43,18
207	5Y Net Income Growth (per Share)	financial indicator	numeric_continuous	-1	4,4761	0,0569213	0,204218394	7,1267944	15676	28,99
208	3Y Net Income Growth (per Share)	financial indicator	numeric_continuous	-1	17,0282	0,072422957	0,375422568	6,7738537	17611	20,23
209	10Y Shareholders Equity Growth (per Share)	financial indicator	numeric_continuous	-0,7137	2,451	0,033468202	0,126555696	9,0231715	12397	43,85
210	5Y Shareholders Equity Growth (per Share)	financial indicator	numeric_continuous	-0,8874	6,0316	0,039181994	0,203913496	8,6958194	15467	29,94
211	3Y Shareholders Equity Growth (per Share)	financial indicator	numeric_continuous	-0,981	20,0367	0,047642119	0,386584146	8,5730066	17434	21,03
212	10Y Dividend per Share Growth (per Share)	financial indicator	numeric_continuous	-1	0,9128	-0,026189597	0,229555076	6,8082258	12900	41,57
213	5Y Dividend per Share Growth (per Share)	financial indicator	numeric_continuous	-1	2,9811	-0,000595742	0,229905763	5,543252	16299	26,17
214	3Y Dividend per Share Growth (per Share)	financial indicator	numeric_continuous	-1	5,4005	0,01229793	0,263079291	4,885649	18264	17,27
215	Receivables growth	financial indicator	numeric_continuous	-1	153332,3333	10,97823889	1101,098165	8,1732668	19952	9,63
216	Inventory Growth	financial indicator	numeric_continuous	-1	293,473	0,158858254	3,268153782	5,9121922	19650	10,99
217	Asset Growth	financial indicator	numeric_continuous	-1	5468,4264	1,134220838	45,38107936	8,9476856	19690	10,81
218	Book Value per Share Growth	financial indicator	numeric_continuous	-3963,1312	7039,3259	0,783208338	66,52400566	9,1425259	19213	12,97
219	Debt Growth	financial indicator	numeric_continuous	-1,051	17646,8235	3,733922197	174,8647006	7,9443619	19462	11,84
220	R&D Expense Growth	financial indicator	numeric_continuous	-2,6622	1542,611	0,295609613	12,97412737	4,5681472	19598	11,23
221	SG&A Expenses Growth	financial indicator	numeric_continuous	-2,3699	2042	0,365149021	15,03005424	8,8285163	19655	10,97
222	Sector	specified sector of selected company stock	nominal	Communication Services	Financial Services			2,1877003	22077	0
223	PRICE VAR NEXT YEAR	annual stock value variation for following year in percent	numeric_continuous	-100,3972195	2418600,915	269,8892659	19346,1736	9,9909845	22077	0
224	Class	indicator for decision making with respect to price var value being positive or negative	boolean						22077	0
225	Year	selected year of stock performance	nominal	2014	2017			1,6048029	22077	0

1- Feature Selection & Charts

We have 225 features. One of them is for classification, one of them for generating class label, one of them is year, one of them is the Stock ID. There are 221 features left. One is nominal/categorical feature that specifies companies' sector. Other 220 features are financial indicator and all of them is numerical attribute.

Our data has considerable number of missing values. Therefore, in feature selection we first drop the attributes that has more than 33% missing values, 16 attributes. There are 206 attributes left. And then we replace missing values with mean value of attributes, considering each sector individually.

In final we choose our best 20 attributes by using ANOVA F-value. We also try with Mutual Information but with MI, our 20 attributes' missing values percentages before replacing as very close to our threshold 33%. We want to use more raw data; therefore, we choose ANOVA F-value for feature selection.

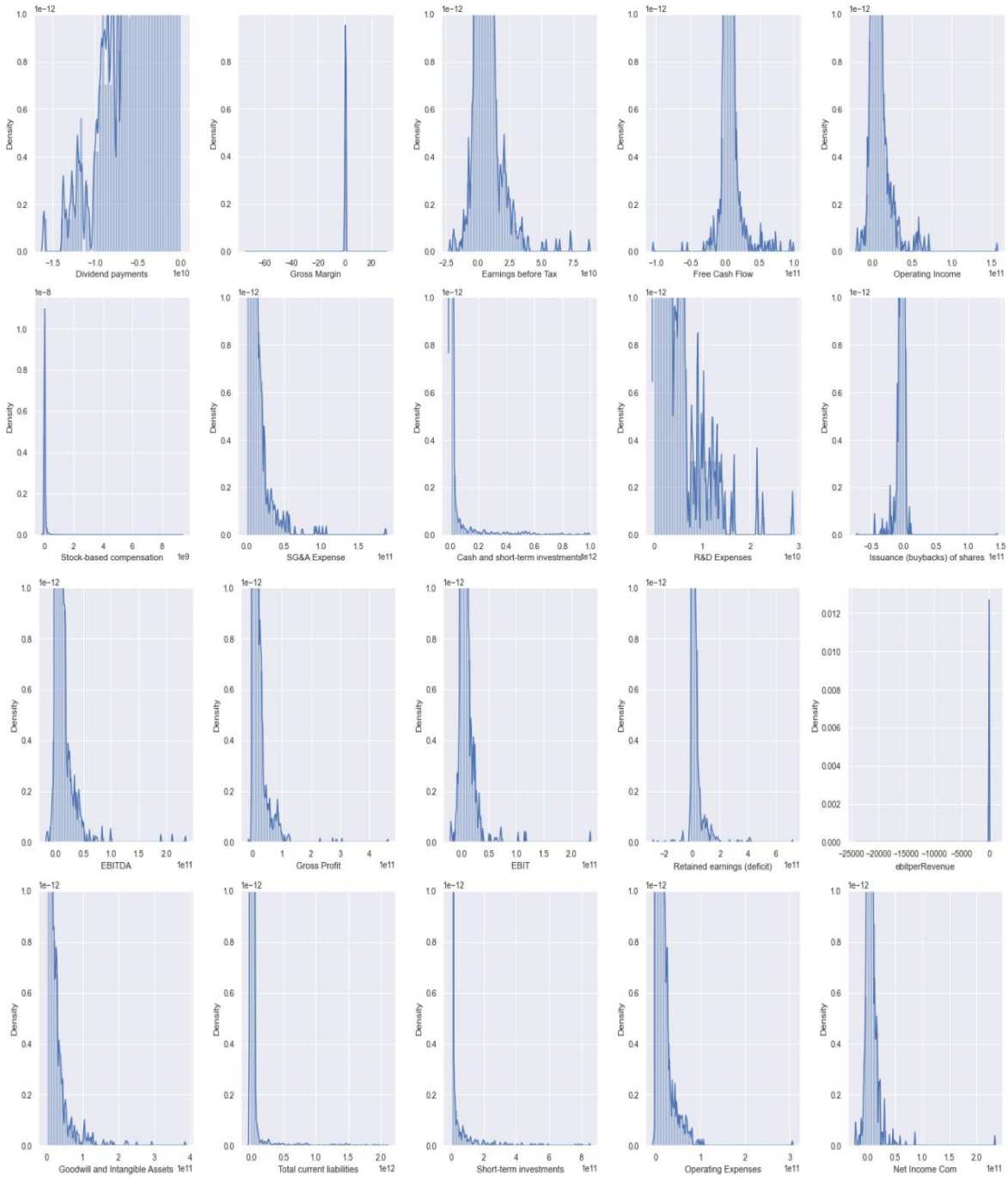

Fig. 1. Distribution charts

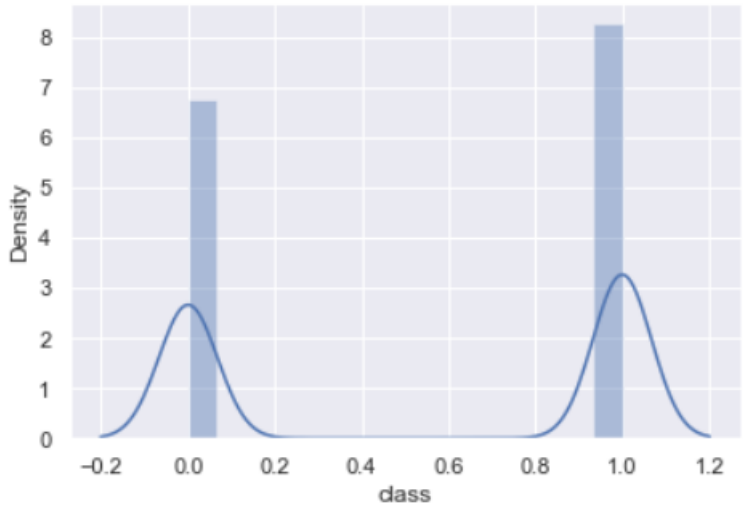

Fig. 2. Class Distribution chart

This is our class labels density chart. As you can see from the chart our stocks prices most likely to increase next year.

2- Scatter Plot Matrix

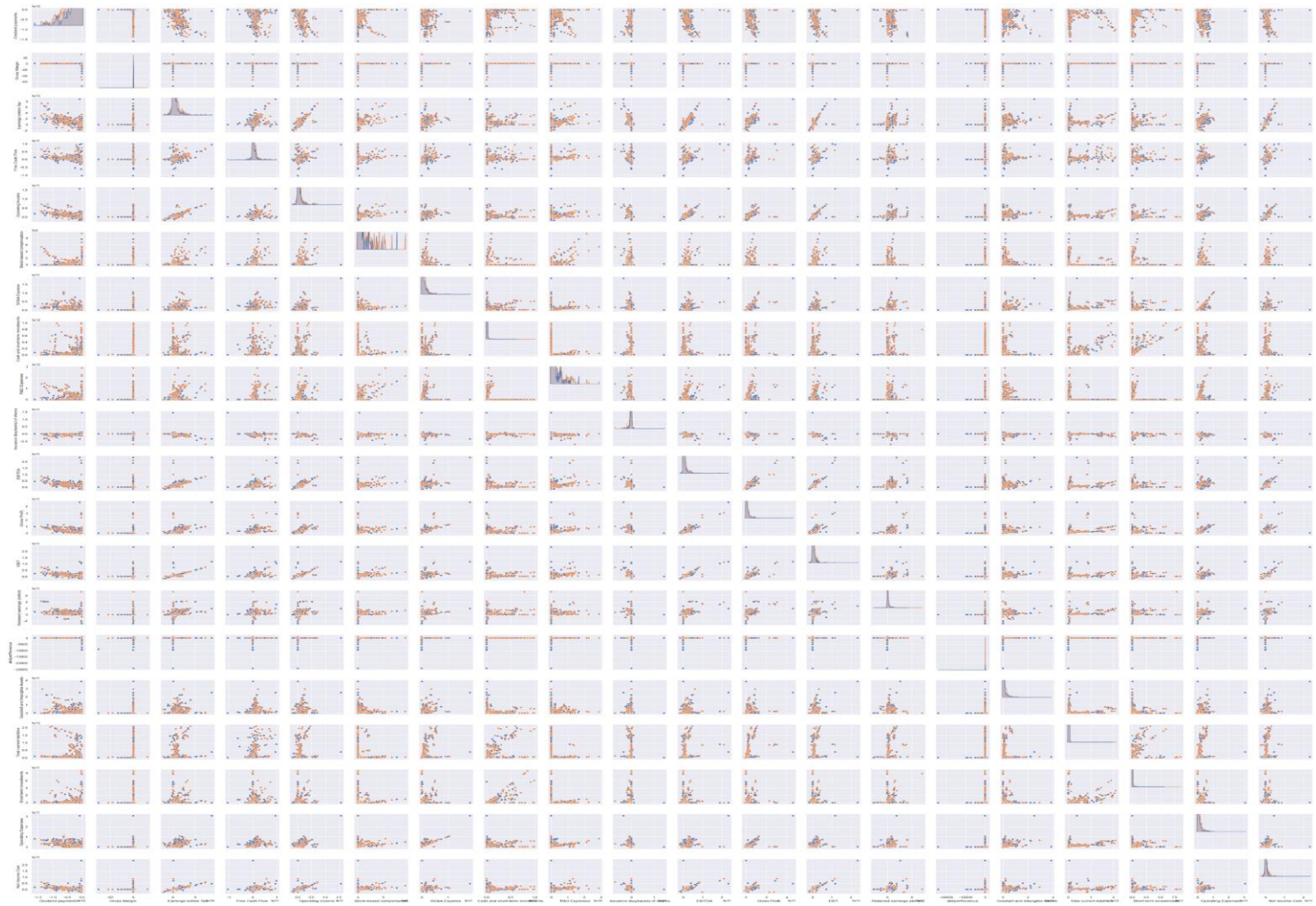

Fig. 3. Scatter Plot Matrix

In Figure 3 relationships between financial indicators are observed. X-axis shows a specific indicator, and y-axis shows another indicator, and the plot shows their relationships. Interactions between these indicators are mostly positively correlated for e.g. (3,6) or negatively correlated for e.g. (17,17) due to slopes being either negative or positive.

1- Feature Selection Methods

Due to data having large numbers of missing values we drop the attributes that has more than 33% missing values and we replace missing values with mean value of attributes.

Table 1. Mutual Information

Results for Mutual Information		
Attribute	Score	Missing Values %
Total non-current assets	0.022112280957419017	27.27725687366943
EBIT Margin	0.022054091974503587	7.582551977170811
eBITperRevenue	0.01920992042646863	12.963717896453323
Gross Profit	0.019154132780264455	5.467228337183494
payoutRatio	0.018948078374790178	10.82574625175522
Total non-current liabilities	0.018644065311503644	27.340671286859628
Earnings before Tax	0.01823924284012657	6.178375685102142
ebitperRevenue	0.018153038022910906	12.963717896453323
EPS	0.017993695863804504	5.893010825746252
Earnings Before Tax Margin	0.017921437192584877	5.41287312587761
Free Cash Flow margin	0.01782413204082678	10.377315758481679
Profit Margin	0.016968018418720643	11.070344702631699
netProfitMargin	0.01651437783387477	12.963717896453323
EBITDA Margin	0.016334752774368688	11.08393350545817
Total shareholders equity	0.016233635705729332	6.214612492639398
dividendYield	0.016230110792283314	14.938623907233772
Consolidated Income	0.016169867868345955	7.097884676360013
Other Assets	0.016166675280912646	30.601983965212664
Net Profit Margin	0.016102373873498665	7.799972822394347
Net Income	0.01600560547987806	7.088825474475699

Table 2. ANOVA F-Value

Results for ANOVA F-Value		
Attribute	Score	Missing Values %
Dividend payments	40.99933856904691	9.765819631290483
Gross Margin	32.31890334987439	5.4309915296462385
Earnings before Tax	31.54958855350051	6.178375685102142
Free Cash Flow	31.143068580384302	7.424015944195316
Operating Income	27.759629645131373	4.987090637314853
Stock-based compensation	27.303706581127024	7.849798432758074
SG&A Expense	23.464233709464327	7.5599039724600265
Cash and short-term investments	22.72453457323068	11.210762331838565
R&D Expenses	22.336960734419254	9.684286814331657
Issuance (buybacks) of shares	22.1082255055086	8.796485029668887
EBITDA	20.253538575804583	7.944920052543371
Gross Profit	19.585008492335405	5.467228337183494
EBIT	16.072282497737262	6.4546813425737195
Retained earnings (deficit)	15.51009740357719	5.286044299497214
ebitperRevenue	15.122352264938574	12.963717896453323
Goodwill and Intangible Assets	14.643294274095304	6.509036553879604
Total current liabilities	14.41323496095706	11.450831181772886
Short-term investments	14.142473710094146	12.284277755129773
Operating Expenses	13.792811693653622	7.709380803551207
Net Income Com	13.71337691993083	6.300674910540382

To conclude we choose our best 20 attributes by using ANOVA F-value. We also try with Mutual Information but with MI, our 20 attributes with missing value percentages before replacing gets very close to our threshold of 33%. We want to use more raw data; therefore, we choose ANOVA F-value for feature selection.

2- Classification Experiments

We did not use Cross Validation or any other similar methods alike. We split our train and test set according to years. Last year's csv, 2018, is our test set.

The reasoning behind this is we thought ourselves at the end of 2018 and wanted to profit next year by using the algorithm. Thus, we use features selected in ANOVA F-Value.

Methods used for classification in this experiment;

- Decision Tree with Gini Index which is calculated by subtracting the sum of squared probabilities of each class from one.
- Decision Tree with Gain Ratio which determines the information gain of all the attributes, and then computes the average information gain.
- Naïve Bayes which are based on applying Bayes' theorem with strong independence assumptions between the features.
- Artificial Neural Networks which are designed to simulate the way the human brain analyzes and processes information.
- K Nearest Neighbor assumes that similar things exist in close proximity.

Table 3. Table for Evaluation for Classification Experiments

	Experiment	Accuracy	F1-macro	F1-micro	AUC
0	Decision Tree with Gini Index	0.528916	0.510808	0.528916	0.534653
1	Decision Tree with Gain Ratio	0.521403	0.505316	0.521403	0.531224
2	Naive Bayes	0.670993	0.428472	0.670993	0.603608
3	ANN with 1 hidden layer	0.680328	0.629554	0.680328	0.682963
4	ANN with 2 hidden layer	0.561475	0.555338	0.561475	0.631891
5	KNN 3	0.570355	0.550413	0.570355	0.592899
6	KNN 9	0.602687	0.579037	0.602687	0.624357
7	KNN 149	0.665528	0.627639	0.665528	0.670177

The performance evaluation table shows the most accurate method is ANN with 1 hidden layer. Considering Area Under Curve (AUC) which measures performance across all possible classification thresholds it suggests ANN with 1 hidden layer overperforms when compared with Naïve-Bayes which is also less accurate. Also, F1-micro (micro-averages) suggests ANN with 1 hidden layer performs better. Furthermore, F1-macro (macro-averages) indicate ANN with 1 hidden layer performs better.

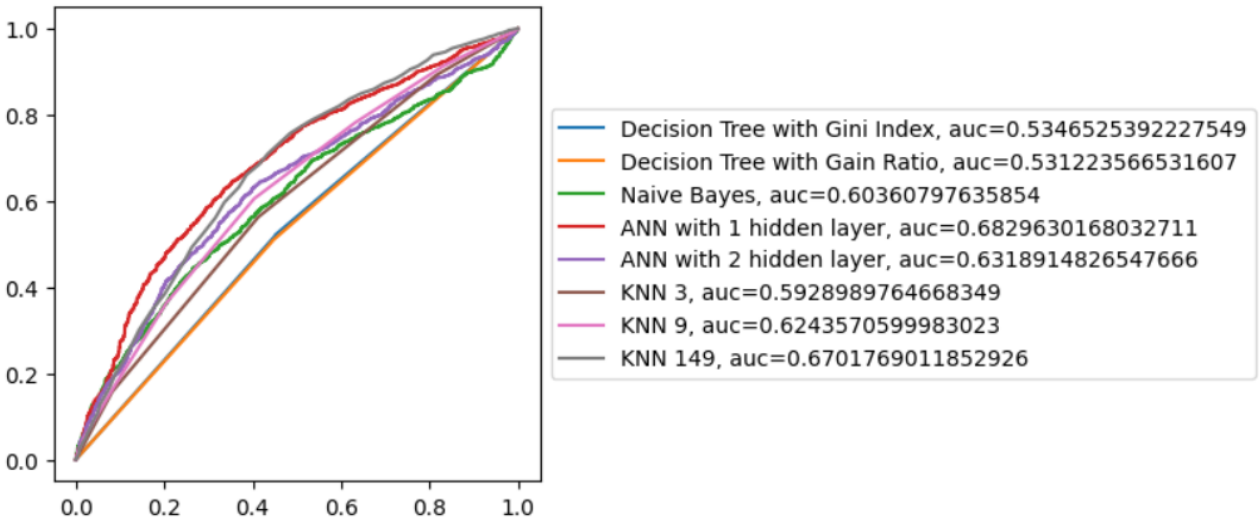

Fig.1 ROC Curve

The ROC curve shows the trade-off between sensitivity and specificity. Classifiers that give curves closer to the top-left corner indicate a better performance. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. This also shows ANN with 1 hidden layer is the best performing method followed by Naïve-Bayes.

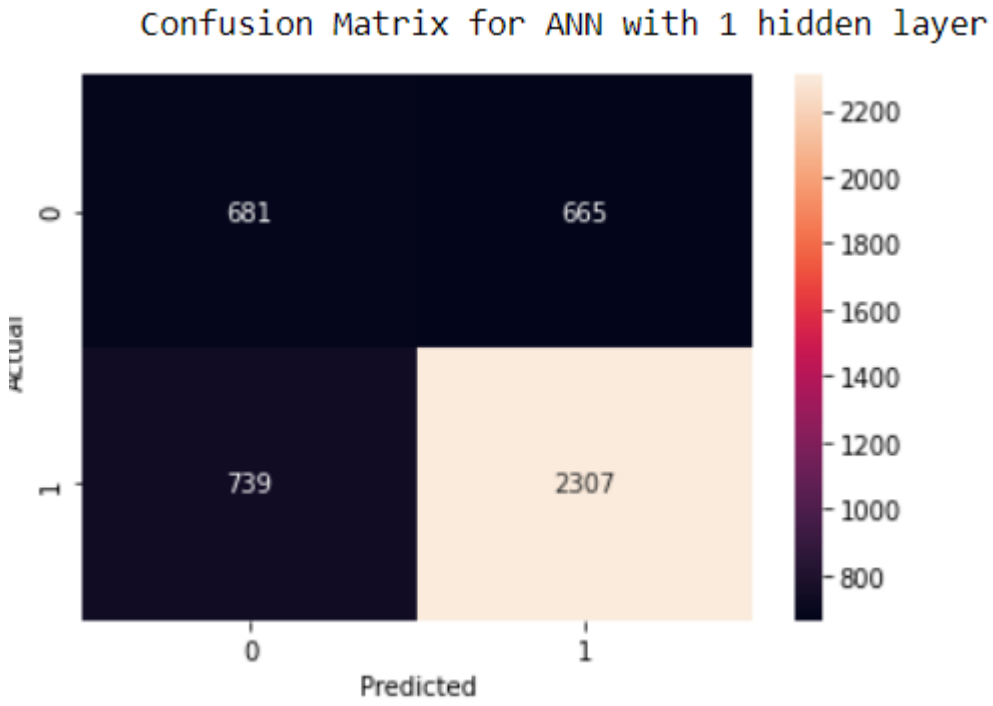

Fig.2 Confusion Matrix for ANN with 1 hidden layer

In Figure.2 it is observed that ANN with 1 hidden layer method predicted true positive cases 2307 times out of 4392 cases and true negative cases 681 times.

This suggests 68% of the algorithm is classifying correctly. Even though theoretically any algorithm can reach 100% accuracy, our achievement of 68% is good enough. Case being that if this algorithm reached maximum accuracy Group-1 would probably be very rich next year because of this achievement.

5- Statistical significance analysis between your best performing model and its closest competitor

Best Model: ANN with 1 hidden layer, Closest Competitor: Naïve-Bayes

Accuracy:

The P-value is = 0.002
The t-statistics is = 5.914
Since $p<0.05$, We can reject the null-hypothesis in terms of accuracy that both models perform equally well on this dataset. We may conclude that the two algorithms are significantly different.

F1-Macro:

The P-value is = 0.005
The t-statistics is = 4.883
Since $p<0.05$, We can reject the null-hypothesis in terms of f1_macro that both models perform equally well on this dataset. We may conclude that the two algorithms are significantly different.

F1-Micro:

The P-value is = 0.059
The t-statistics is = 2.428
Since $p>0.05$, we cannot reject the null hypothesis in terms of f1_micro may conclude that the performance of the two algorithms is not significantly different.

AUC:

The P-value is = 0.398
The t-statistics is = -0.923
Since $p>0.05$, we cannot reject the null hypothesis in terms of AUC and may conclude that the performance of the two algorithms is not significantly different.

1- Choosing K-Means Cluster Number with Elbow Method

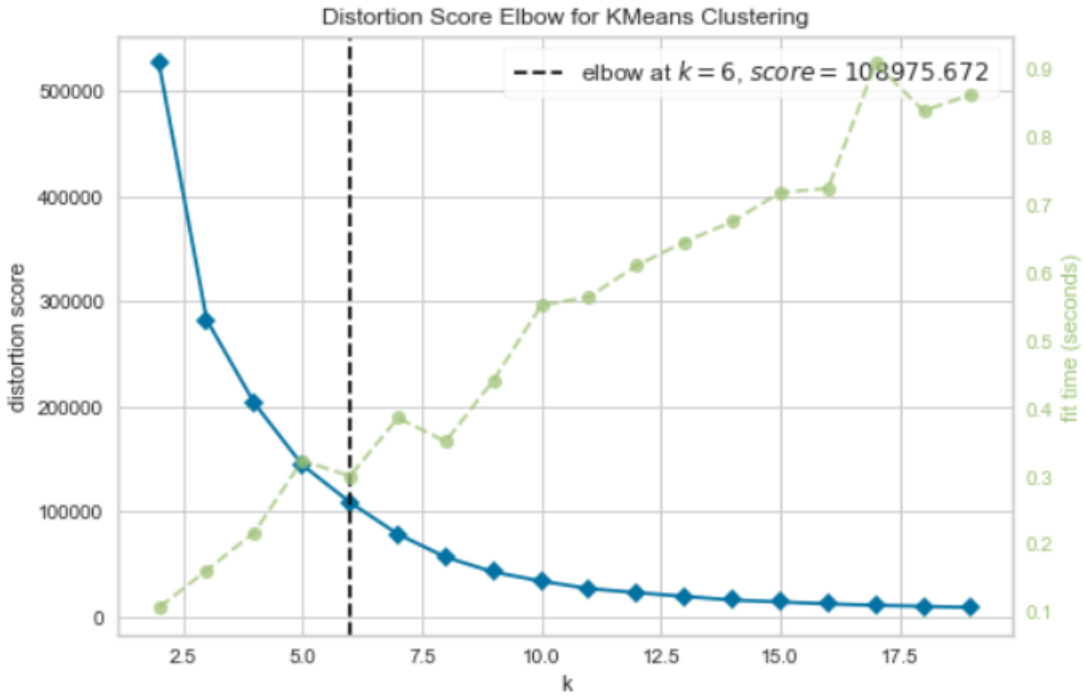

Fig.1 Distortion Score Elbow

Elbow method is used to find out exact number of clusters to both not divide information unnecessarily and add much information possible. By looking at the blue-line we can see the elbow(bend) is exactly at point k=6. Also point of infliction suggest that point k=6 is the best fitting value for K-Means clustering.

2- Clustering Experiments

There are 3 different clustering methods used in this experiment. AGNES (Agglomerative Nesting), K-Means and DBSCAN (Density-based spatial clustering of applications with noise). We used DBSCAN because our data has noise at some instances which are attributes named “Dividend Payments” and “R&D Expenses” and due to DBSCAN resulting in good performance in such instances.

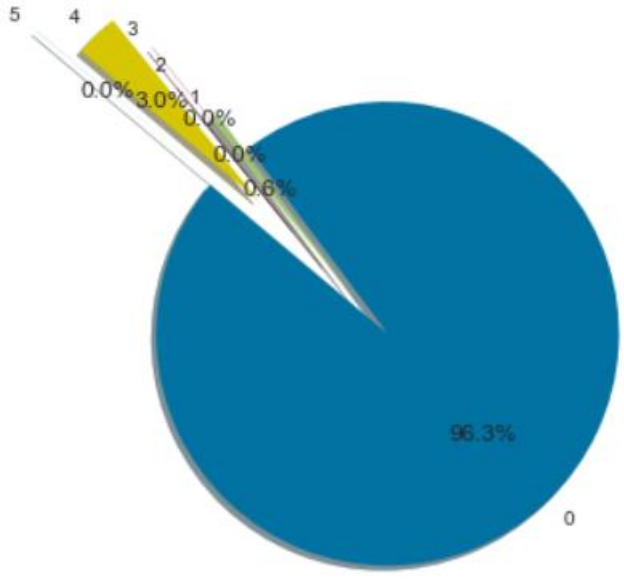

Fig.2 K-Means with 6 Clusters

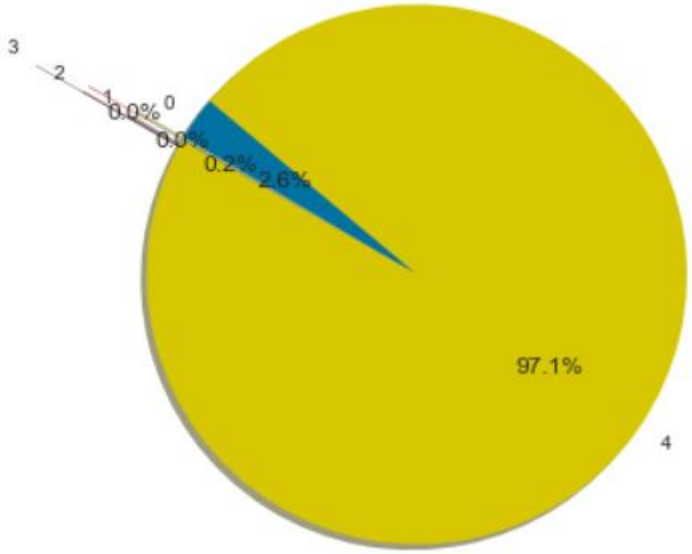

Fig.3 AGNES with 5 Clusters

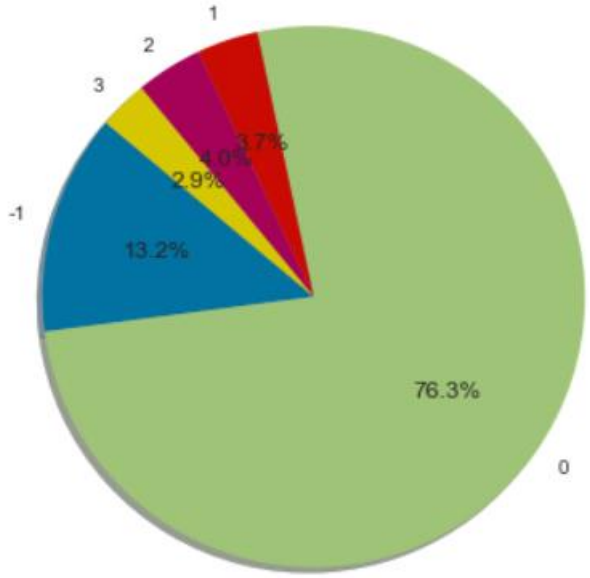

Fig.4 DBSCAN with eps=0.04 and Minimum of 72 Samples

3- Evaluation Table

Table.1 Evaluation Table

	Experiment	# of Clusters	Average Number of Instances in Clusters	Std. Dev.	SSE	NMI	Silhoutte Value	Adjusted RI
0	K-Means	6	{0: 21265, 1: 128, 2: 1, 3: 7, 4: 668, 5: 8, 'avg': 3679.5}	7867.979680	108475	0.002279	0.882208	-0.002596
1	AGNES	5	{0: 583, 1: 35, 2: 11, 3: 1, 4: 21447, 'avg': 4415.4}	8518.641432	---	0.001055	0.910104	-0.001390
2	DBSCAN	5	{-1: 2909, 0: 16839, 1: 811, 2: 874, 3: 644, 'avg': 4415.4}	6266.925071	---	0.004560	0.249948	0.002690

It is observed from the point of Silhouette Value which is the similarity of the value to its cluster, AGNES is the best followed by K-Means. DBSCAN on the other hand performed poorly.

From the point of ARI all of the algorithms performed poorly. This can be related to most financial indicators being relevant only to themselves and ARI does not see local relevancy between these attributes. Therefore, considers the values in the clusters randomly placed.

Normalized Mutual Information values suggest high number of correct clusters for the instances. It is observed that K-Means algorithm is more reliable. Although the NMI values are too low to discriminate.

Sum of Squared Error (SSE) for K-Means can be considered normal if we assume the number of instances we have.