CS295P: Stock Prediction Project

Group S10
Hanyan Wang, Zhihui Xia, Shuqing Ye

Introduction

Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange. The successful prediction of a stock's future price could yield significant profit. In this project we do **fundamental analysis** on the company profiles and apply machine learning models on the data to find out which features are more important. Then the machine learning will generate a score on each company depending on their data. Our portfolio will pick the top 20 companies with the highest score.

Data

The training data is collected from Yahoo Finance during the last quarter of 2001, 2006, 2011.

Approach

Our project is divided into three parts, Parsing, Prediction, and Trade/Adversary. The parsing stage will be responsible for reading the HTML, extracting the data, and generating the database for machine learning. The prediction section will run machine learning models on the database and give a portfolio with 20 best companies. Using the generated trading file, the Adversary will display the result of our portfolio.

Parsing

1.HTML to CSV

The raw data is a pile of HTML files. We use Jsoup, a Java library, and CSSselector to extract and manipulate data.
Take Apple’s profile in 2001-11 for example. It shows much useful information which can be used to do fundamental analysis. We extracted these elements from HTML files, removed unnecessary symbols, converted them into string objects and stored them in csv files.

Statistics at a Glance -- NasdaqNM:AAPL As of 31-Oct-2001

Price and Volume	
52-Week Low on 20-Dec-2000	$13.625
Recent Price	$17.56
52-Week High on 30-Apr-2001	$27.12
Beta	1.31
Daily Volume (3-month avg)	4.99M
Daily Volume (10-day avg)	6.34M
Stock Performance	
[one-year price graph]	
big chart [1d | 5d | 3m | 6m | 1y | 2y | 5y | max]	
52-Week Change	-14.3%
52-Week Change relative to S&P500	+14.9%
Share-Related Items	
Market Capitalization	$6.16B
Shares Outstanding	350.9M
Float	312.3M
Dividends & Splits	
Annual Dividend	none
Last Split: factor 2 on 21-June-2000	

Per-Share Data	
Book Value (mrq*)	$11.17
Earnings (ttm)	-$0.14
Earnings (mrq)	$0.18
Sales (ttm)	$15.26
Cash (mrq*)	$12.36
Valuation Ratios	
Price/Book (mrq*)	1.57
Price/Earnings	N/A
Price/Sales (ttm)	1.15
Income Statements	
Sales (ttm)	$5.36B
EBITDA (ttm*)	-$344.0M
Income available to common (ttm)	-$37.0M
Profitability	
Profit Margin (ttm)	-0.7%
Operating Margin (ttm)	-6.4%
Fiscal Year	
Fiscal Year Ends	Sep 29
Most recent quarter (fully updated)	29-Sep-2001
Most recent quarter (flash earnings)	30-Sep-2001

Management Effectiveness	
Return on Assets (ttm)	-0.60%
Return on Equity (ttm)	-0.96%
Financial Strength	
Current Ratio (mrq*)	3.39
Debt/Equity (mrq*)	0.08
Total Cash (mrq)	$4.34B
Short Interest As of 10-Sep-2001	
Shares Short	6.64M
Percent of Float	2.1%
Shares Short (Prior Month)	6.30M
Short Ratio	1.44
Daily Volume	4.61M

See Profile Help for a description of each item above; **M** = millions; **B** = billions; **mrq** = most-recent quarter; **ttm** = trailing twelve months; (as of 30-Sep-2001, except **mrq*/ttm*** items as of 29-Sep-2001)

In our implementation, the entry for each company consists of 41 attributes, including 52-Week Low (lowest price within 52 weeks), recent price, beta, market capitalization, DCV (daily cash volume), etc.

Considering the fact that some companies with dots in their names are frequently lacking fundamental data, since they are either foreign stocks or don't trade or work the same way as other companies on the major exchanges, we decided to skip these companies and not to put their data into our database.

2. Meric of a Company's performance

We use the formula below to measure a company (stock)'s performance, over an identical period of time:

$$Company\ Return\ -\ Market\ Return$$

`Company Return` is a company's monthly rate of return; `Market Return` indicates by how many percent the S&P 500 index changes during that month. By comparing the company with the market, we can evaluate its performance to see if it outperforms its peers, regardless of what its raw return is.

3. Output

The figure below shows the fundamental data of every company in 2001-10. Some cleaning and processing will be handled in the next phase.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Company	52-Week	Recent Pr	52-Week	Beta	Daily Vol	Daily Vol	52-Week	52-Week	Market C	Shares O	Float	Annual D	Dividend	Last Split
2	AA	25.875	33.17	45.71	1.17	3.42M	2.88M	15.60%	55.90%	28.2B	849.3M	738.9M	0.6	1.81%	none
3	AACE	7.25	9.3	14	0.63	17.5K	13.0K	-24.80%	1.40%	93.5M	10.1M	4.50M	none	none	none
4	AAE	none	none	none	none	none	none	none	none	none	none	none	none	none	none
5	AAII	6.5	17.21	23.22	1.02	23.5K	9,000	105.50%	177.20%	308.5M	17.9M	4.30M	none	none	none
6	AAON	none	none	none	none	none	none	none	none	none	none	none	none	none	none
7	AAPL	13.625	16.2	27.12	1.31	5.65M	6.49M	-25.50%	-1.70%	5.68B	350.8M	312.2M	none	none	none
8	AAR	none	none	none	none	none	none	none	none	none	none	none	none	none	none
9	AAT	none	none	none	none	none	none	none	none	none	none	none	none	none	none
10	AATK	0.5	2.92	5.109	-1.39	20.1K	109.0K	-20.80%	4.50%	13.2M	4.53M	3.40M	none	none	none
11	AAUK	9.46	12.55	18.25	0.98	183.2K	197.0K	-6.60%	23.20%	29.4B	2.34B	1.62B	none	none	none
12	ABCB	8	13	13.5	0.3	8,364	4,000	36.80%	80.60%	113.8M	8.76M	7.80M	0.48	3.69%	none
13	ABCW	12.875	16.25	18.51	0.53	74.7K	26.0K	7.40%	41.80%	367.5M	22.6M	17.9M	0.33	2.03%	none
14	ABF	7	9.16	14.2	1.54	391.4K	372.0K	-1.00%	30.70%	440.6M	48.1M	37.0M	0.16	1.75%	none
15	ABFI	4.813	17.05	21.7	0.57	29.0K	57.0K	210.00%	309.10%	43.9M	2.58M	1.10M	0.32	1.88%	none
16	ABFS	12.125	20.85	28.25	0.92	286.8K	132.0K	48.90%	96.50%	430.4M	20.6M	15.1M	none	none	none
17	ABGX	15.313	25.15	94.875	2.43	1.51M	1.66M	-65.70%	-54.70%	2.17B	86.2M	64.6M	none	none	none
18	ABI	18.5	24.99	133.313	1.59	1.19M	1.01M	-76.30%	-68.70%	5.29B	211.6M	211.3M	0.17	0.68%	none
19	ABIZ	0.76	0.95	10.125	3.2	285.0K	466.0K	-89.50%	-86.20%	127.8M	134.5M	27.6M	none	none	none
20	ABJ	none	none	none	none	none	none	none	none	none	none	none	none	none	none
21	ABK	42.2	53.77	64	0.67	552.2K	679.0K	10.40%	45.70%	5.69B	105.9M	75.2M	0.36	0.67%	none
22	ABLE	1.625	4.05	6.75	0.83	3,045	3,000	8.00%	42.50%	8.10M	2.00M	1.00M	none	none	none
23	ABM	24.96	26.07	38.2	0.41	67.6K	148.0K	-1.60%	29.80%	633.5M	24.3M	16.0M	0.66	2.53%	none
24	ABMC	0.25	0.86	1.5	0.18	28.0K	37.0K	-25.60%	-1.80%	17.6M	20.5M	12.7M	none	none	none
25	ABMD	10.5	19.21	37.75	0.54	144.0K	189.0K	-46.20%	-29.00%	400.0M	20.8M	12.5M	none	none	none
26	ABN	14.3	16.15	26.063	1.15	154.2K	218.0K	-29.40%	-6.80%	24.3B	1.50B	1.34B	none	none	none
27	ABS	21	31.99	36.99	0.17	1.53M	1.79M	33.60%	76.40%	13.0B	406.1M	300.5M	0.76	2.38%	none
28	ABT	42	51.77	56.25	0.45	3.19M	4.02M	8.60%	43.30%	80.3B	1.55B	1.44B	0.84	1.62%	none
29	ABTL	0.7	0.82	6.813	0.43	89.4K	77.0K	-87.10%	-83.00%	25.3M	30.9M	16.7M	none	none	none
30	ABV	13.2	14.45	28.7	1.24	295.5K	595.0K	-36.50%	-16.20%	5.58B	386.5M	255.3M	0.51	3.50%	none
31	ABX	12.313	16.86	19.38	0.6	1.94M	2.65M	16.80%	54.10%	6.68B	396.0M	316.8M	0.22	1.30%	none
32	ABY	6.08	6.08	9.313	1.31	228.3K	174.0K	-29.50%	-7.00%	2.68B	440.0M	435.6M	0.26	4.28%	none
33	AC	37.4	46	59.35	1.05	153.8K	226.0K	-2.60%	28.50%	11.4B	247.8M	104.1M	2.84	6.17%	none
34	ACA	none	none	none	none	none	none	none	none	none	none	none	none	none	none
35	ACAI	8.04	15.42	30.06	1.55	610.3K	819.0K	2.60%	35.40%	674.3M	43.7M	28.9M	none	none	none
36	ACAP	13.5	23.1	23.57	none	58.7K	60.0K	none	none	none	none	none	none	none	none
37	ACAS	20.25	26.1	29.5	0.96	217.7K	244.0K	18.00%	55.70%	879.6M	33.7M	25.6M	2.24	8.58%	none
38	ACAT	11.375	12.48	16.46	0.33	77.0K	146.0K	4.00%	37.20%	298.5M	23.9M	9.30M	0.24	1.92%	none
39	ACDO	23.25	31.34	62	0.8	435.5K	711.0K	-7.50%	22.10%	810.1M	25.8M	22.7M	none	none	none
40	ACEC	0.688	1.49	7.063	2.4	6,273	5,000	-76.60%	-69.20%	13.8M	9.28M	4.80M	none	none	none
41	ACET	7.625	10.29	10.587	-0.03	5,500	14.0K	8.30%	42.90%	66.9M	6.50M	2.90M	0.32	3.11%	none
42	ACF	20.375	33.02	64.9	1.45	1.64M	1.53M	19.00%	57.00%	2.78B	84.3M	64.9M	none	none	none
43	ACG	7.125	8.67	8.94	0.26	277.4K	252.0K	16.60%	53.80%	1.43B	164.7M	163.0M	0.84	9.69%	none
44	ACI	9.375	17.5	38.4	-0.28	601.0K	358.0K	59.10%	109.90%	922.3M	52.7M	33.7M	0.23	1.31%	none

	A	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	Company	Book Valu	Earnings (	Earnings (	Sales (pe	Cash	Price/Boo	Price/Earn	Price/Sale	Sales (ttm	EBITDA	Income	Profit Ma	Operating	Fiscal Yea
2	AA	12.67	1.65	0.39	28.05	0.57	2.62	20.07	1.18	24.4B	4.25B	1.44B	7.10%	10.40%	31-Dec
3	AACE	5.56	0.11	0.2	20.52	13.08	1.67	81.58	0.45	208.4M	26.5M	1.21M	0.60%	1.00%	30-Jun
4	AAE	none	none	none	none	none	none	none	none	none	none	none	none	none	none
5	AAII	3.8	0	0.06	6.26	0.11	4.53	none	2.75	111.3M	9.77M	-36.0K	0%	2.30%	31-Dec
6	AAON	none	none	none	none	none	none	none	none	none	none	none	none	none	none
7	AAPL	11.01	0.15	0.17	16.43	12.04	1.47	106.58	0.99	5.78B	-223.0M	67.0M	1.20%	-5.40%	30-Sep
8	AAR	none	none	none	none	none	none	none	none	none	none	none	none	none	none
9	AAT	none	none	none	none	none	none	none	none	none	none	none	none	none	none
10	AATK	1.02	-0.51	-0.14	0.98	0.05	2.87	none	2.98	4.55M	-1.35M	-2.33M	-50.80%	-34.40%	31-Dec
11	AAUK	9.8	2.21	1.74	10.08	0.59	1.28	5.67	1.25	15.8B	4.32B	3.43B	27.00%	20.90%	31-Dec
12	ABCB	10.16	1.15	0.25	8.49	5.82	1.28	11.29	1.53	72.1M	none	9.74M	13.50%	none	31-Dec
13	ABCW	9.92	1.22	0.35	10.08	2.76	1.64	13.34	1.61	230.9M	none	27.9M	12.10%	34.70%	31-Mar
14	ABF	17.39	-0.85	-0.13	68.36	0.55	0.53	none	0.13	3.29B	172.3M	-40.7M	-1.20%	-1.20%	31-Dec
15	ABFI	24.78	2.47	0.45	57.75	33.75	0.69	6.91	0.29	183.3M	100.2M	7.86M	4.30%	7.20%	30-Jun
16	ABFS	15.26	2.65	0.4	71.04	0.18	1.37	7.88	0.29	1.73B	118.7M	60.2M	3.70%	6.90%	31-Dec
17	ABGX	9.79	-0.29	-0.17	0.81	6.68	2.57	none	30.85	71.4M	-18.9M	-25.4M	-35.60%	-35.60%	31-Dec
18	ABI	4.92	0.96	0.22	7.34	1.86	5.08	25.98	3.4	1.62B	346.7M	212.4M	13.10%	17.30%	30-Jun
19	ABIZ	1.92	-4.95	-0.61	4.86	0.02	0.49	none	0.2	429.7M	-96.0M	-398.3M	-84.00%	-55.00%	31-Dec
20	ABJ	none	none	none	none	none	none	none	none	none	none	none	none	none	none
21	ABK	26.35	3.63	0.99	6.13	0.33	2.04	14.83	8.77	665.1M	none	393.5M	59.20%	84.50%	31-Dec
22	ABLE	2.38	-0.67	-0.45	16.49	0.5	1.7	none	0.25	34.0M	-789.0K	-1.37M	-4.00%	-5.30%	30-Jun
23	ABM	14.84	1.93	0.52	78.12	0.1	1.76	13.52	0.33	1.93B	93.8M	47.7M	2.50%	4.00%	31-Oct
24	ABMC	0.06	-0.1	-0.01	0.38	0.01	13.93	none	2.25	6.88M	-2.35M	-1.82M	-26.40%	-33.40%	30-Apr
25	ABMD	4.93	-0.76	-0.3	1.21	4.23	3.9	none	15.86	25.0M	-17.8M	-15.7M	-62.80%	-85.10%	31-Mar
26	ABN	none	1.32	0.41	none	none	none	12.26	none	none	none	none	none	none	none
27	ABS	13.8	1.04	-0.37	91.38	0.34	2.32	30.88	0.35	37.4B	2.20B	427.0M	1.10%	3.10%	1-Feb
28	ABT	5.52	1.09	0.34	9.39	0.57	9.38	47.49	5.51	14.7B	3.08B	1.71B	11.70%	14.60%	31-Dec
29	ABTL	2.4	-2.55	-1.8	3.28	3.31	0.34	none	0.25	66.7M	-55.9M	-51.9M	-81.30%	-88.60%	31-Dec
30	ABV	5.22	0.4	0.13	3.6	2.88	2.77	36.22	4.01	1.25B	203.2M	138.6M	10.00%	13.30%	31-Dec
31	ABX	7.89	-1.95	0.17	3.41	1.72	2.14	none	4.94	1.35B	-654.0M	-772.0M	-57.10%	-70.70%	31-Dec
32	ABY	4.7	0.72	0.16	9.86	0.05	1.29	8.49	0.62	4.36B	1.26B	316.8M	7.60%	18.80%	31-Dec
33	AC	16.36	2.71	0.63	11.97	0.85	2.81	16.95	3.84	2.91B	1.10B	650.2M	22.60%	24.00%	31-Dec
34	ACA	none	none	none	none	none	none	none	none	none	none	none	none	none	none
35	ACAI	4.53	0.52	0.29	11.79	3.11	3.41	29.43	1.31	523.4M	48.4M	23.4M	4.50%	6.70%	31-Dec
36	ACAP	none	none	none	none	none	none	none	none	none	none	none	none	none	none
37	ACAS	16.51	-1.35	0.4	2.92	0.07	1.58	none	8.95	78.5M	68.3M	-32.9M	-41.90%	78.20%	31-Dec
38	ACAT	7.12	1.1	0.08	21.57	1.38	1.75	11.34	0.58	521.9M	46.8M	26.9M	5.20%	6.90%	31-Mar
39	ACDO	7.31	0.66	0.18	17.69	2.18	4.29	47.7	1.77	462.1M	30.8M	17.3M	3.70%	5.70%	30-Jun
40	ACEC	1.11	-0.75	-0.27	2.57	0.62	1.35	none	0.58	24.2M	-5.20M	-6.93M	-28.60%	-28.80%	30-Jun
41	ACET	10.64	0.7	0.03	28.98	1.28	0.97	14.62	0.35	178.2M	6.76M	4.25M	2.40%	3.30%	30-Jun
42	ACF	12.71	2.58	0.81	9.53	0.92	2.6	12.78	3.46	818.2M	498.1M	222.9M	27.20%	44.30%	30-Jun
43	ACG	8.22	0.31	0.14	0.51	0.04	1.05	27.88	16.9	111.6M	100.0M	69.2M	62.00%	77.80%	31-Dec
44	ACI	11.19	0.28	0.02	35.03	0.04	1.56	61.62	0.5	1.46B	124.8M	11.4M	0.80%	6.10%	31-Dec

Prediction

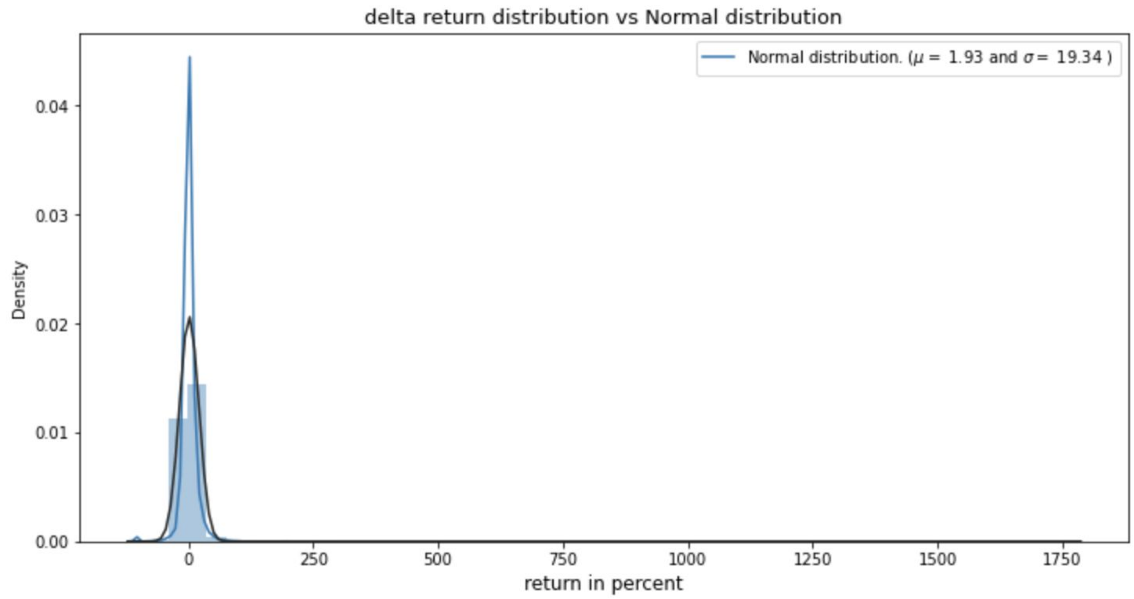
1. Data Processing

We read 9 months data from 2001, 2006 and 2011 and concatenate them. We only select companies with daily cash volume larger than one million.

In the raw data, there are empty cells and special symbols like ‘M’, ‘K’, ‘%’, we need to replace empty cells with NaN and convert special symbols to floating-point numbers. After converting, we get something like below:

	52-Week Low	Recent Price	52-Week High	Beta	Daily Volume (3-month avg)	Daily Volume (10-day avg)	52-Week Change	52-Week Change relative to S&P500	C
0	25.875	33.17	45.71	1.17	3420000.0	2880000.0	15.6	55.9	
5	13.625	16.20	27.12	1.31	5650000.0	6490000.0	–25.5	–1.7	
6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9	9.460	12.55	18.25	0.98	183200.0	197000.0	–6.6	23.2	
12	7.000	9.16	14.20	1.54	391400.0	372000.0	–1.0	30.7	

Our label is called `delta return`, which is computed as stock return - market return in the corresponding month. We plot the distribution of `delta return` as follows, blue curve represents our return and black curve represents normal distribution. We want to convert our distribution to the normal one since it helps the training process.



We drop rows and columns with too many missing values. Finally, we fill these missing values with simple imputer to avoid errors in training.

```
#fill missing values

from sklearn.impute import SimpleImputer
my_imputer = SimpleImputer()
train_filled = pd.DataFrame(my_imputer.fit_transform(train))
train_filled.columns=train.columns
train_filled.index=train.index
```

After preprocessing, our data has 23743 entries and 33 columns.

2. Model

We use three models to train on our data, namely CatBoost, Decision Tree and Linear Regression. 5 fold cross validation is applied on the data to determine the best model and root mean squared error (RMSE) is set as the loss function.

The following table shows the result for training, the second column is the mean of RMSE and the third column is the stand deviation of RMSE. CatBoost works the best among three and we will use it to train the whole dataset.

CatBoost	0.196775	0.019617
DecisionTree	0.287104	0.020307
LinearRG	0.213732	0.033751

After training on nine months data in total, we plot the top 20 important features that contribute to the model. All features contribute more or less to the model.

	Feature Id	Importances
0	52–Week High	7.383566
1	52–Week Change	6.866093
2	Short Ratio	5.515618
3	Daily Volume (3–month avg)	5.287328
4	52–Week Change relative to S&P500	5.074821
5	Beta	4.859244
6	DCV	4.667107
7	Book Value	4.391858
8	Return Assets	4.025941
9	Shares Outstanding	3.885802
10	Daily Volume (10–day avg)	3.624967
11	Percent of Float	3.428175
12	Price/Sales	3.411327
13	Float	3.080421
14	Shares Short	2.894444
15	Shares Short (Prior Month)	2.686684
16	Current Ratio	2.552755
17	Recent Price	2.457203
18	52–Week Low	2.411475
19	Cash	2.369788

3. Output

We save the trained model under the current directory. When there is a need to predict, we load the model and preprocess the test csv produced by the parser. It is worth noting that this test csv doesn’t have label information. We predict the top 20 stocks and export them as "year-month-portfolio.csv" based on the given year and month.

Trades and Adversary

1. Trades

We start from $100,000 and distribute the money evenly over the 20 companies. As a long-term trader, we only trade on a monthly basis -- buy at the beginning of a month, and sell at the end of that month. To make things easier, each transaction is as close to the market close time, i.e. 16:00, as possible. Here is an example of our trading:

```
2001-10-02 15:59 buy 2647 shares of ARBA
2001-10-02 15:59 buy 356 shares of AVGN
......
2001-10-31 15:59 sell 2647 shares of ARBA
2001-10-31 15:59 sell 356 shares of AVGN
```

2. Adversary

We also build an Adversary to mimic our "adversary" in the real market, such as brokers, the exchange and other actors, and evaluate the proposed trades. Adversary is implemented based on the following rules:

A. All orders are market orders that are filled immediately at the current ask for buys, and at the current bid for sells.
B. You will only be able to buy at most 1% of the DCV at the current ask price. After that, the price goes up by at least X for each 1% of the DCV, where x = bid-ask spread / 2.
C. Similarly, if the trader tries to sell more than 1% of the DCV, then the price the trader gets should decrease by X for each 1% of DCV.
D. If there is no transaction at the time requested for the trade, choose the line with the closest line after. If there is no line after, go for the closest line before.
E. If the bid and ask is N/A, approximate the price as follows. If the whole day daily cash volume is DCV, β is the bid-ask spread, log is the natural logarithm, then the following is a half decent approximation to their relationship:

$$\frac{-5log(\beta)}{log(DCV)} \approx 1$$

Thus the current asking price = current price + β, current bidding price = current price - β.

Testing Result

For the testing process, we use the data of two years to train the model, and then apply the prediction on the data of the third year. For example, if we wanted to make predictions on 2001-10, we will use data from 2006 to 2011 as training data, then output the portfolio of 2001-10.
Here is our result in each month of Quarter 4 over 3 years, compared with monthly market return.

Month	Our adversary	Prof’s adversary	Portfolio Return	Market Return
2001-10	$121171.44	$121583.91	21.6%	0.8%
2001-11	$100179.65	$100125.99	0.13%	5.11%
2001-12	$116653.12	$117097.12	17.1%	1.61%
2006-10	$93481.44	$93807.11	-6.2%	3.50%
2006-11	$96321.79	$96620.38	-3.4%	2.40%
2006-12	$97162.27	$97676.89	-2.3%	1.55%
2011-10	$126235.41	$128915.40	28.9%	14.02%
2011-11	$105740.69	$103531.11	3.53%	2.35%
2011-12	$102122.41	$102373.86	2.37%	1.05%

Generally speaking, our portfolio’s performance is great. Among 9 months from 2001 to 2011, we obtained positive returns in 6 months. It is worth noting that in October and December 2001, and October 2011 our return is significantly higher than the market, especially in 2001-10, we achieved a 21.6% return, which is 27 times of the market return in that month.

But our portfolio cannot always outperform the market. One major reason is that certain stocks we bought at the beginning of one month, due to delisting, were not trading any more at the end of that month, and tagged $-1 by Adversary, thus reducing our earnings by a large margin. This accounts for the low income in November 2001 and loss in October 2006.

In all, according to the results of Adversary, our portfolio can beat the market and get excess returns.

Please NOTE that for the final prediction application, we will train the data from three years, and generate the final model. Therefore, the result from our final application will be different from the above table.

Appendix

Language and Environment

Python 3.6
Java 1.8

Openlab
Google Colab (model training)

Dependencies

Jsoup >= 1.13.1
OpenCSV >= 5.3
shap >= 0.39.0
scikit-learn >= 0.24.1
Catboost >= 0.24.4
pandas >= 1.1.5
numpy>=1.13.3