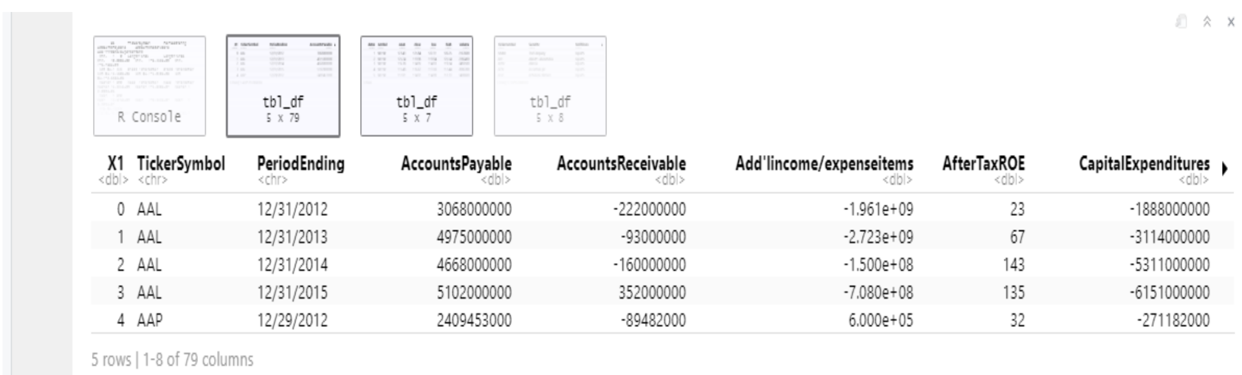
Analysis of New York Stock Data and Profitable Stocks

Ananya Gangavarapu
Data Science Class

Introduction

In this project, we desire to analyze the New York Stock Data and see what variables contribute to gains and losses within the stock market and which of 4 companies (Apple, Microsoft, Facebook, and Amazon) is most profitable. We ideally want to show how we can determine which stocks are the most profitable using regression analysis. We will utilize the New York Stock Exchange Dataset on Kaggle for our data analysis, which contains 4 csv files which contain the necessary data for our analysis. We will do regression analysis to determine the characteristics that companies have for gaining or losing money and we will also do time series analysis to analyze the opening and closing values for each stock on a daily basis as well as the high values and the low values. We ultimately will be able to find profitable stocks and what the considerations are for profitable stocks.



R Console | tbl_df 5 x 79 | tbl_df 5 x 7 | tbl_df 5 x 8

X1 <dbl>	TickerSymbol <chr>	PeriodEnding <chr>	AccountsPayable <dbl>	AccountsReceivable <dbl>	Add'lincome/expenseitems <dbl>	AfterTaxROE <dbl>	CapitalExpenditures <dbl>
0	AAL	12/31/2012	3068000000	-222000000	-1.961e+09	23	-1888000000
1	AAL	12/31/2013	4975000000	-93000000	-2.723e+09	67	-3114000000
2	AAL	12/31/2014	4668000000	-160000000	-1.500e+08	143	-5311000000
3	AAL	12/31/2015	5102000000	352000000	-7.080e+08	135	-6151000000
4	AAP	12/29/2012	2409453000	-89482000	6.000e+05	32	-271182000

5 rows | 1-8 of 79 columns

Fig 1 - Fundamental.csv

R Console | tbl_df 5 x 79 | tbl_df 5 x 7 | tbl_df 5 x 8

	dates <dbl>	symbol <chr>	open <dbl>	close <dbl>	low <dbl>	high <dbl>	volume <dbl>
	1	WLTW	123.43	125.84	122.31	126.25	2163600
	2	WLTW	125.24	119.98	119.94	125.54	2386400
	3	WLTW	116.38	114.95	114.93	119.74	2489500
	4	WLTW	115.48	116.62	113.50	117.44	2006300
	5	WLTW	117.01	114.97	114.09	117.33	1408600

5 rows

Fig 2 - Prices.csv

R Console | tbl_df 5 x 79 | tbl_df 5 x 7 | tbl_df 5 x 8

Tickersymbol <chr>	Security <chr>	SECfilings <chr>	GICSSector <chr>	GICSSubIndustry <chr>	AddressofHeadquarters <chr>
MMM	3M Company	reports	Industrials	Industrial Conglomerates	St. Paul, Minnesota
ABT	Abbott Laboratories	reports	Health Care	Health Care Equipment	North Chicago, Illinois
ABBV	AbbVie	reports	Health Care	Pharmaceuticals	North Chicago, Illinois
ACN	Accenture plc	reports	Information Technology	IT Consulting & Other Services	Dublin, Ireland
ATVI	Activision Blizzard	reports	Information Technology	Home Entertainment Software	Santa Monica, California

5 rows | 1-6 of 8 columns

Fig 3 - Security.csv

Model Analysis

For determining the gain and loss functions of each company, I utilized multiple regression analysis. For the sake of simplicity, I had utilized the capital expenditures, which is the amount each company spends on assets and equity, to be the loss function and the gain function to be the total revenue of the company (*1*). I had decided to utilize the forward selection process to help with variable selection, specifically R's olsrr forward selection function.

```
fit <- lm(TotalRevenue ~ EarningsPerShare + GrossProfit + EarningsBeforeTax + ProfitMargin + NetIncome + NetBorrowings + Goodwill, data =
fundamental)
fit2 <- lm(CapitalExpenditures ~ TotalAssets + TotalCurrentAssets + OtherEquity + Investments + LongTermDebt + OtherLiabilities +
OtherCurrentLiabilities + IncomeTax + TotalCurrentLiabilities, data = fundamental)
ols_step_forward_p(fit, details = TRUE)
ols_step_forward_p(fit2, details = TRUE)
```
```
```

Fig 4 - forward step variable selection function.

Using this approach, I was able to find the variables needed for the gain function and the loss functions, which can be seen in Fig 4. I was also able to gauge the usefulness of the models using the ols_step_forward_p function.

(Intercept)	5658504901.404	773194385.249		7.318	0.000	4.142037e+09	7.174973e+09
GrossProfit	2.120	0.069	0.704	30.933	0.000	1.986000e+00	2.255000e+00
EarningsBeforeTax	3.599	0.504	0.503	7.139	0.000	2.611000e+00	4.588000e+00
NetIncome	-3.749	0.685	-0.366	-5.469	0.000	-5.094000e+00	-2.405000e+00
ProfitMargin	-175388983.212	30936224.008	-0.075	-5.669	0.000	-2.360643e+08	-1.147137e+08
NetBorrowings	0.306	0.103	0.040	2.960	0.003	1.030000e-01	5.090000e-01
Goodwill	-0.121	0.080	-0.023	-1.508	0.132	-2.780000e-01	3.600000e-02

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	GrossProfit	0.6679	0.6677	45.1252	90136.4952	23610626684.3418
2	EarningsBeforeTax	0.6770	0.6766	-2.9518	90088.8633	23290471047.6940
3	NetIncome	0.6840	0.6834	-39.1752	90052.0264	23044392985.8326
4	ProfitMargin	0.6894	0.6887	-67.0051	90023.1182	22851728405.6446
5	NetBorrowings	0.6912	0.6903	-74.7632	90014.8899	22792620985.3809
6	Goodwill	0.6916	0.6905	-74.9314	90014.6092	22784450969.3684

Fig 5 - Output of the profit function

I was able to minimize the error for some of the parameters. However, for others such as the Profit Margin, there was a large standard error within prediction. I was able to get a semi-good Adj-R-Squared value of 69.05% (Adj-R-Squared was used due to more resistance against additional variables as opposed to regular R-Squared). RMSE and AIC are mainly large thanks to the large values of the individual data points in the data set, which would contribute to semi-large resultant error.

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-171252545.109	50401420.266		-3.398	0.001	-2.70105e+08	-72400056.211
IncomeTax	-0.549	0.034	-0.337	-15.963	0.000	-6.17000e-01	-0.482
TotalCurrentLiabilities	-0.074	0.011	-0.237	-6.902	0.000	-9.50000e-02	-0.053
Investments	0.053	0.009	0.142	5.883	0.000	3.50000e-02	0.071
TotalCurrentAssets	0.048	0.007	0.215	7.030	0.000	3.40000e-02	0.061
OtherCurrentLiabilities	0.078	0.004	2.851	22.150	0.000	7.10000e-02	0.085
LongTermDebt	0.018	0.003	0.168	5.241	0.000	1.10000e-02	0.025
TotalAssets	-0.044	0.002	-2.972	-20.502	0.000	-4.80000e-02	-0.039
OtherLiabilities	0.051	0.003	0.918	20.251	0.000	4.60000e-02	0.056

Fig 6 - Output of the loss function (Parameters)

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	IncomeTax	0.3867	0.3864	1133.8143	81894.1465	2334342817.8012
2	TotalCurrentLiabilities	0.4843	0.4837	672.6084	81587.4610	2141166541.5055
3	Investments	0.5108	0.5100	548.9423	81495.5892	2086062730.0069
4	TotalCurrentAssets	0.5201	0.5190	506.9443	81463.5176	2066785632.9136
5	OtherCurrentLiabilities	0.5246	0.5233	487.3598	81448.5613	2057549801.9153
6	LongTermDebt	0.5366	0.5350	432.6515	81405.2280	2032101942.4893
7	TotalAssets	0.5403	0.5385	416.6642	81392.6043	2024346966.5578
8	OtherLiabilities	0.6267	0.6250	8.6331	81023.8247	1824737497.1843

Fig 7 - Output of the loss function (Selection summary)

As for the loss function, we were able to get an Adjusted-R-Squared value of 62.5%, which isn't too bad but could be optimized. However, we were able to get very little standard error within the parameters of the model except for the intercept. Overall, both models aren't the best fits for the data and we cannot truly extrapolate whether these variables have a relationship, but we can make estimates and see some relationship between the variables since we have somewhat low error standard error within the parameters and somewhat good fit.

Fig 8 - Time Series plots for companies Apple, Microsoft, Facebook, and Amazon respectively from top to bottom and left to right

I had also performed more in-depth analysis into 4 specific stocks that were well known to have very high revenues: Apple, Microsoft, Facebook, and Amazon. I had developed time series plots with the high and low prices and well as the opening and closing prices for each day. We see that Apple has a drastic decrease in the stock prices while companies like Facebook seem to almost increase monotonically. We see from the 4 plots that Amazon has the highest opening prices in comparison to the other plots, thus we can conclude that Amazon is one of the more profitable companies.

Discussion

With the analysis done, we were able to gain the gain and loss functions of the stocks within the New York Stock Exchange. To a certain degree, we were able to find what influences the various actions that companies take and what composes those actions through various models generated. While some of these models have prediction inaccuracies, they do provide somewhat of a picture for seeing what goes into the various expenses that companies take.

Various steps could have been improved in the implementation of the models. For example, better selection procedures could have been used in order to complete this task. Especially with this dataset, forward selection is very time consuming and can lead to flawed results. I would try to utilize a more effective variable selection method in the future, such as best subsets regression. I would also try to use a simpler dataset in future trials, as there were simply too many variables and especially with forward selection, it was very difficult to find satisfactory variables to use. I would also try to do analysis with more stocks in future analyses and to try to see a wide demographic of companies in order to see the composition of these stocks.

Conclusion

In this project, we attempted to see how companies are able to maintain themselves within the stock market using regression models that can allow us to see what variables contribute most to a company's success as well as comparing various companies using time series analysis. From our analysis, we are able to find some sort of relationship between various variables listed in the New York Stock Exchange Data and also do in-depth analysis with 4 companies using time series analysis to see what companies are profitable.

From this process, we are able to say that we have somewhat found the means to find good companies to invest stocks in, and we can use this method to find good and efficient stocks to buy shares from. Finding a criterion for finding good companies to invest in is really important, especially in COVID-19, when people are struggling financially.

Additional Work

I had tried various other variables for the various regression models, in the hopes of potentially getting a better R-squared value. Unfortunately, I wasn't able to reduce the error that was added due to a large number of variables so I limited it to the variables that I had in the model.