ECMM445 Learning from Data

I. INTRODUCTION

In today's world of data, market and corporate analysis has taken a new turn with researchers studying the sample data of large firms. This analysis depends upon various factors such as firm's monetary performance, nature of business, market risks etc. Following on the same note, I will be working with Fortune500 dataset from the year 2017 which ranks the greatest corporations by revenue for their corresponding financial years [1]. In this report the target is to first perform feature selection and then compare different styles of regression models – Multivariable Linear Regression, Polynomial, Ridge and LASSO. I will also predict the profit values and draw a contrast between the model performances and accuracy in reference to the original profit values. This dataset can be used for exploratory analysis on the companies and give insights like: What role does ranking play in profitability? How does employee strength affect the revenue made by the company? Is profit a good and only factor for estimating the stock price of the company?

II. METHODOLOGY

The dataset collected from the Fortune website [1] has relevant information about top 500 companies with 23 features in total. The data is first studied to understand the types of variables, ways to handle null values, remove duplicates and even normalised for better results. The dataset contains some categorical as well as continuous variable. The dataset summary for numerical columns can be seen in Table 1.

	Rank	Employees	Hqzip	Revenues	Revchange	Profits	Prftchange	Assets	Totshequity
count	500.000	500.000	500.000	500.000	500.000	500.000	500.000	500.000	500.000
mean	250.492	56350.132	46791.714	24111.748	3.758	1779.480	26.579	80389.340	13640.147
std	144.477	123452.026	30160.385	38337.353	19.967	3937.559	649.036	270425.701	30523.154
min	1.000	83.000	1104.000	5145.000	-57.500	-6177.000	-1499.600	437.000	-12688.000
25%	125.750	11900.000	19099.000	7245.000	-3.825	235.725	-20.300	8436.500	1997.500
50%	250.500	25000.000	46244.500	11384.000	1.900	683.600	2.250	19324.500	4981.000
75%	375.250	56825.250	75045.000	22605.250	7.325	1770.775	20.450	48126.000	12467.750
max	500.000	2300000.000	98188.000	485873.000	197.300	45687.000	12450.000	3287968.000	283001.000

Table 1: Data describe

The data contains 5 integer type columns, 4 float type columns and other object type columns. Some of the interesting variables are the Employee (count of the employees) which has a huge spread between 83 and 2300000, Sector (Business area) containing values like “Energy”, “Business Services”, “Household Products” etc, Revenues (Revenue generated in one fiscal year) with a standard deviation of 38337.353 and Profits (Profit or loss incurred for the year) where losses can be seen as much as 6177 and profits as large as 485873 which will play an important role in the analysis further.

Data mining [3] by many ways is done to transform raw data into a meaningful and understandable format since real world data is inconsistent, contain errors and may be in shortage of behaviours and trends. After cleaning and pre-processing of the data, the steps for feature selection are taken to determine the explanatory variables for the predictor variable (Profits) by the process of backward elimination as shown in Fig. 1.

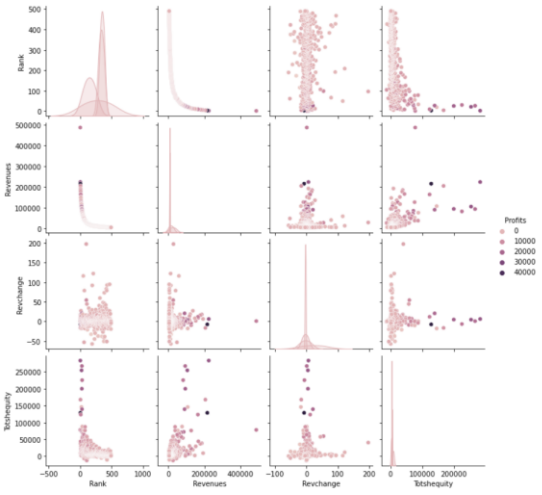
Dep. Variable:	y	**R-squared:**	0.475
Model:	OLS	**Adj. R-squared:**	0.472
Method:	Least Squares	**F-statistic:**	146.8
Date:	Tue, 06 Dec 2022	**Prob (F-statistic):**	9.95e-68
Time:	22:27:58	**Log-Likelihood:**	-4597.2
No. Observations:	490	**AIC:**	9202.
Df Residuals:	486	**BIC:**	9219.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	180.3547	154.905	1.164	0.245	-124.011	484.721
x1	-0.0035	0.002	-2.260	0.024	-0.006	-0.000
x2	0.0581	0.005	11.200	0.000	0.048	0.068
x3	0.0049	0.001	9.632	0.000	0.004	0.006

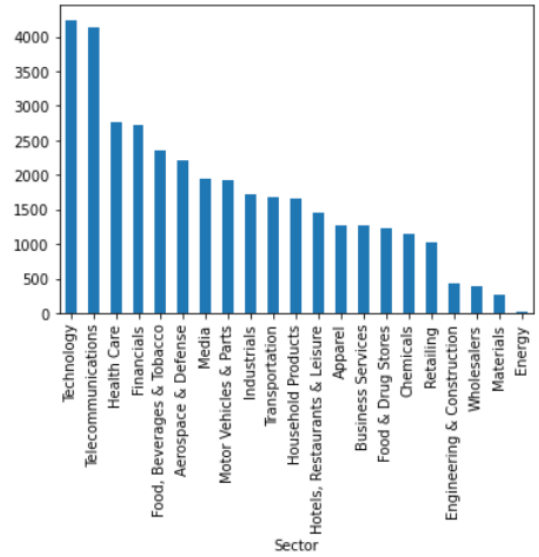
Omnibus:	427.927	**Durbin-Watson:**	1.801
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	23965.752
Skew:	3.404	**Prob(JB):**	0.00
Kurtosis:	36.578	**Cond. No.**	3.42e+05

Fig 1: Final OLS Regression Result

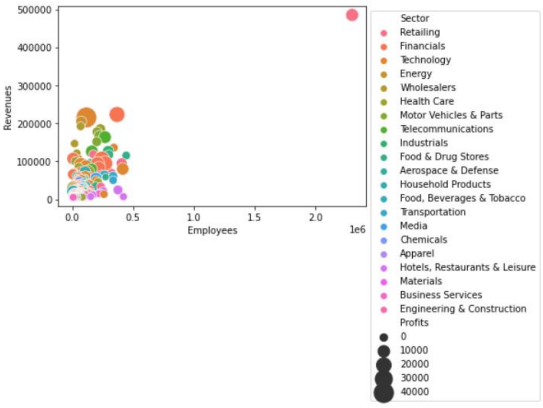
A very crucial step is to understand the relationship between the variables which can be done in many ways: check for multicollinearity [2], perform one hot encoding (if required) to include categorical data, scatter plots, bar graphs, heatmaps etc. Fig. 2 and Fig. 3, visualizes some of these methods to understand the variables better.



(i) Pair-plot between variables

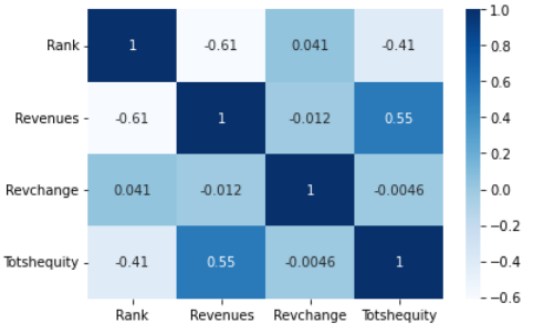


(ii) Count by Sectors' Total Profit



(iii) Scattering by Sectors, Revenue and Employees

Fig. 2: Methods to understand relationship between variables by visuals methods



(i) Heatmap between variables

	Rank	Revenues	Revchange	Totshequity	Profits
Rank	1.000000	-0.608803	0.041107	-0.413576	-0.450021
Revenues	-0.608803	1.000000	-0.011591	0.546510	0.598746
Revchange	0.041107	-0.011591	1.000000	-0.004636	0.018119
Totshequity	-0.413576	0.546510	-0.004636	1.000000	0.716337
Profits	-0.450021	0.598746	0.018119	0.716337	1.000000

(ii) Correlation table for all the dependant variables

Fig. 3: Methods to understand relationship between variables by numerical methods

Fig. 2 (i) helps to identify the relationship and is very easy to find the linearity between the variables. None of the variables show a very strong linear relationship to each other which gives us a green signal to include these in the final model.

Fig. 2 (ii) provides a total profit per sector which helps to deep dive into the dataset further. Technology sector contributes to the maximum profit of the organisation whereas Energy sector is not sufficiently contributing. This analysis can be used by companies to focus on the profit generating sectors.

Fig. 2 (iii) shows the scattering of sectors, revenues and employee counts. We can identify an outlier in the data during this process.

After establishing the relationship between the variables, the next step is to check again if any more variables need to be removed. In this case, the final explanatory variables that explain 'Profits' of the organisation are Rank (rank established according to Fortune500), Revenues (Revenue in millions), Revchange (change in revenue from last year) and Totshequity (total equity share). These variables can now be taken further to fit into the model. Fig.4 shows the relationship of each variable with the predictor variable.

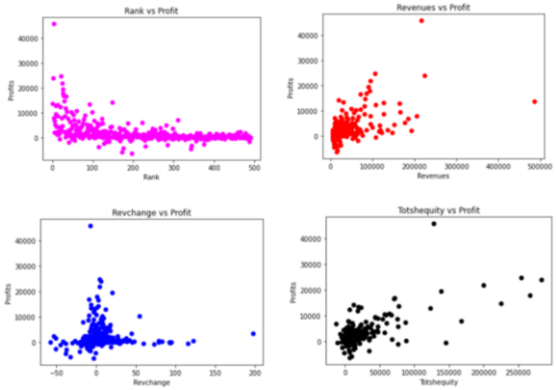
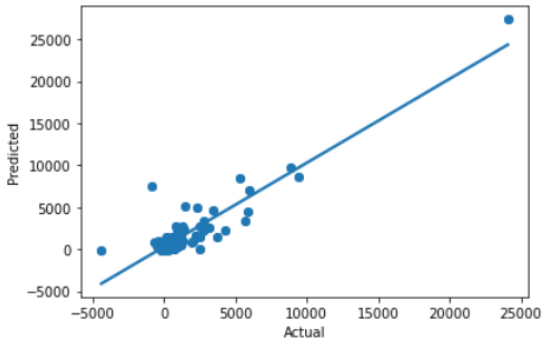


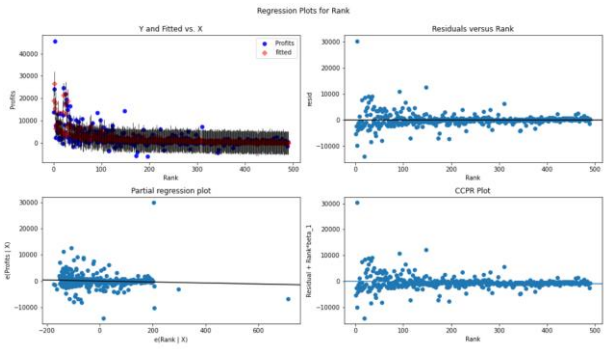
Fig. 4: Relationship between explanatory and predictor variables

III. RESULTS

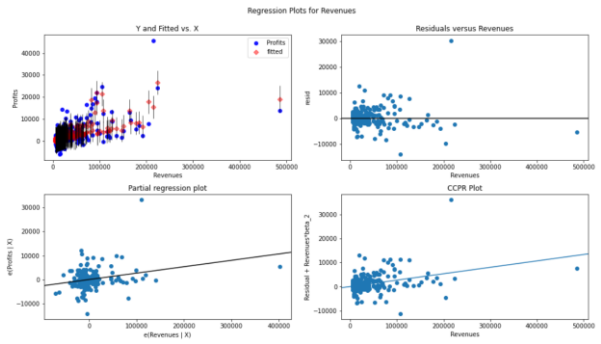
After being tested on many models of regression, various results were observed. We will now visualize the results of multivariable linear regression and then compare the results of all the other models. Below figures (Fig.5) are the results of multivariate linear regression:



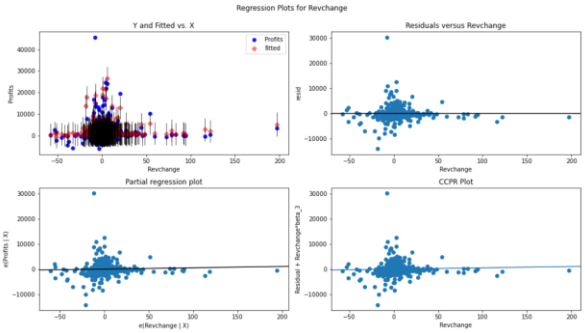
(i) Simple scatter plot analysis



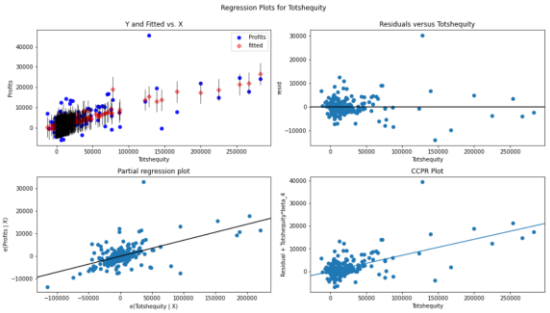
(ii) Residual plot wrt rank



(iii) Residual plot wrt Revenues



(iv) Residual plot wrt Revchange



(v) Residual plot wrt Totshequity

Fig. 5: Output results for multivariate linear regression

As seen in the graphs, this model is not doing a great job. It defines a relationship between quantitative dependant variable and more than two independent variables. The R-square of this model turns out to be 0.79353. The predicted values are not very close to the actual values. Advantages of using this model is the capacity to determine correlation of predictors to the criterion values. The second advantage is that we are able to locate outliers or any other kind of anomality. It also has a few drawbacks; it is not a good choice specially if the input data is incomplete.

The next model is the polynomial regression model. It is a modified form of multiple linear regression. The relationship between X and Y is modelled as the nth degree polynomial [4] [5]. This increases the complexity of the data and fits better than the multiple linear regression. We tried to run the model on various polynomial degrees. Fig. 6 shows the value of R-square for different values of degree.

Model	R-square
Polynomial (degree=2)	0.632967
Polynomial (degree=3)	0.735197
Polynomial (degree=4)	0.803459

Fig. 6: Calculated R-square for degrees 2,3 and 4.

The best results are provided by degree=4. The next one for comparison is the Ridge with varying alphas. The premium advantage of Ridge is that it avoids overfitting the model and does not even require any unbiased estimators. It is perfect for improving the least-square estimate in case of multicollinearity. The disadvantage for using this model is the trade variance for bias.

In Fig.7 we can see the various R-square values for all the Ridge's.

Model	R-square
Ridge (alpha=0.5)	0.79344
Ridge (alpha=1.0)	0.79336
Ridge (alpha=2.0)	0.79319

Fig. 7: Calculated R-square for alpha 0.5,1 and 2

Finally, the last one, LASSO was also run and output was recorded in the below Fig.8. The advantages to LASSO are that it automatically selects the features along with reducing overfitting. The disadvantage of LASSO is that they may provide a little unstable estimate. So, this model is also not the correct one.

Model	R-square
LASSO (alpha=0.5)	0.79353
LASSO (alpha=1.0)	0.79357
LASSO (alpha=2.0)	0.79361

Fig. 8: Calculated R-square for alpha 0.5,1 and 2

To finalise, it is very clear that polynomial with degree 4 is the most suitable model for this kind of data.

These models slightly differ when it comes to predictions and error calculations.

The maximum RMSE (Root Mean Square Error) occurred in Polynomial with degree 2 followed by Polynomial with degree 3. The lowest RMSE can be observed in Polynomial regression with degree 4 followed by LASSO with alpha 2.

The models have been arranged in the order of their RMSE values:

Polynomial (degree=4) < LASSO (alpha=2.0) < LASSO (alpha=1.0) < LASSO (alpha=0.5) < Multivariable Linear Regression < Ridge (alpha=0.5) < Ridge (alpha=1.0) < Ridge (alpha=2.0) < Polynomial (degree=3) < Polynomial (degree=2)

Arranging them according to their R-square values:

Polynomial (degree=4) > LASSO (alpha=2.0) > LASSO (alpha=1.0) > Multivariable Linear Regression > LASSO (alpha=0.5) > Ridge (alpha=0.5) > Ridge (alpha=1.0) > Ridge (alpha=2.0) > Polynomial (degree=3) > Polynomial (degree=2)

I have predicted the Profit values using Polynomial (degree=4) and plotted against the actual values from the dataset in Fig. 9.

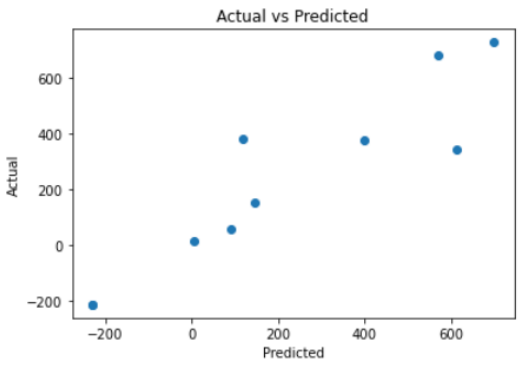


Fig. 9: Actual vs predicted for Polynomial regression

IV. DISCUSSION

After comparing numerous models, the polynomial regression explains the data more closely than any of the others. The beauty of a polynomial regression model is that it doesn't really matter if the variables have a linear relationship. So, when linear regression fails, we can always rely on polynomial method. With the highest R-square and lowest RMSE, the model has proved its worth statistically as well. Although, we have not touched the classification models there might be a chance that this data works better with the classification models. Label encoding was performed for the categorical columns but none of them actually made any difference. The data was a little less for regression models and more detailed data like maybe analysing 1000 companies would have led to better results. Linear regression is also very sensitive to outliers and after outlier treatment was done, we still did not get a very convincing improvement. Going back to the original target, we did compare and contrast various models in details which gave me a chance to not be restricted on a certain type of model rather gave me an opportunity to work on different kinds of model closely. The analysis could have gone further if unsupervised learning was involved which could have opened doors to many possibilities.

References*

[1] ‘Fortune 500’, *Fortune*, 22-May-2022. [Online]. Available: https://fortune.com/fortune500/. [Accessed: 07-Dec-2022].

[2] *Psu.edu*. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=53fef985237ae14efddeaf202d44c35ce714d8e2. [Accessed: 07-Dec-2022].

[3] A. K. Pujari, *Data Mining Techniques*. London, England: Sangam Books, 2001.

[4] M. Cenite, ‘Google Books’, in *The SAGE Guide to Key Issues in Mass Media Ethics and Law*, 2455 Teller Road, Thousand Oaks California 91320: SAGE Publications, Inc., 2015, pp. 847–858.

[5] E. Ostertagová, ‘Modelling using polynomial regression’, *Procedia Eng.*, vol. 48, pp. 500–506, 2012.

*