ECMM445

Learning From Data

Continuous Assessment 1

CONTENTS

INTRODUCTION

This report aims to extract meaningful insights from a Fortune 500 dataset, initially through exploratory data analysis, and later through statistical and machine learning techniques of clustering, regression and classification.

Research Questions and Objective of Analysis

How many 'types' of companies exist in the data?

- To cluster companies based on common features.

What are the optimal set of covariates for predicting company Revenue?

- To create models for predicting Revenue based on other features.

Can a classification be built to predict the cluster a company should belong to?

- To create model for classifying companies into clusters.

Existing Literature

Below is a brief description of the algorithms and methodologies used for exploring the dataset.

1. Principal Component Analysis, proposed by Karl Pearson around 1900 was described by its author as as finding "lines and planes of closest fit to systems of points in space." It is therefore, used to find a set of axes along which, if projected, the data retains *most* of its variance.
2. Ester M., et al (1996) introduced the DBSCAN algorithm, which is especially useful for detecting dense clusters of points. It, however, leaves outliers which do not belong to any cluster if they are sufficiently far enough of regions of high density.
3. Usually attributed to Stuart Lloyd of Bell Labs, the family of K-means algorithms was first suggested in 1957, though published much later in 1982 as a non-parametric clustering algorithm. They are hard clustering algorithms, i.e., they assign all data points to a cluster.
4. Multicollinearity may be detected by performing multiple correlation tests, however, a more readily available metric (that bypasses the family-wise error rate of multiple correlation tests) is the variance inflation factor[1].
5. Schwarz's Bayesian Information Criterion is a metric used to determine goodness-of-fit of a model[2]. It penalizes models for being too complex, thereby preventing overfitting.
6. Regularized Regression – both Ridge[3] and Lasso[4] Regression include terms based on p-norms to penalize models, thereby mediating the effect of coefficients and preventing overfitting.
7. Ensemble models for regression – XGBoost[5] and Random Forest[6] are some of the most widely-used tree-based ensemble models for prediction
8. Multi-Layer Perceptron for regression and classification[7]

Possible Obstacles

From a regression standpoint, depending on the target variable, the predictors just may not have enough signal to build a robust model. High multicollinearity in data can undermine a regression's inference capacity, yet taking them out may negatively affect the R-squared value. In such a scenario, other ways of exploring this data will be looked at such as classification.

[1] Edward R. Mansfield & Billy P. Helms (1982) Detecting Multicollinearity, The American Statistician, 36:3a, 158-160, DOI: 10.1080/00031305.1982.10482818
[2] A. Maydeu-Olivares & C Garcıa-Forero (2010) Goodness-of-fit Testing
[3] Donald W. Marquardt & Ronald D. Snee (1975) Ridge Regression in Practice, The American Statistician, 29:1, 3-20, DOI: 10.1080/00031305.1975.10479105
[4] J Ranstam, J A Cook, LASSO regression, *British Journal of Surgery*, Volume 105, Issue 10, September 2018, Page 1348, https://doi.org/10.1002/bjs.10895
[5] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting Systemhttps://doi.org/10.1145/2939672.2939785
[6] Breiman, L. Random Forests. Machine Learning **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
[7] Murtagh F. (1991), Multilayer perceptrons for classification and regression, Neurocomputing, Volume 2, Issues 5–6

METHODOLOGY & DATASET

The dataset is taken from https://data.world/aurielle/fortune-500-2017. It contains various numeric and categorical information regarding Fortune 500 companies as of 2017.

Checks and Transformations

Data does not have any null values. Rank is completely dependent on Revenues so it is removed.

Hqzip is cast into string format. For ZIPs with 4 characters, a '0' is appended to the front. Prftchange (Change in Profit) is a string – it is cast into a float.

Statistic	Employees	Revenues	Revenue change	Profits	Profit change	Assets	Total share equity
count	500	500	500	500	500	500	500
mean	5.64E+04	24111.748	3.758	1779.4798	26.5792	8.04E+04	13640.147
std	1.23E+05	38337.3533	19.9675	3937.55872	649.03636	2.70E+05	30523.154
min	8.30E+01	5145	-57.5	-6177	-1499.6	4.37E+02	-12688
25%	1.19E+04	7245	-3.825	235.725	-20.3	8.44E+03	1997.5
50%	2.50E+04	11384	1.9	683.6	2.25	1.93E+04	4981
75%	5.68E+04	22605.25	7.325	1770.775	20.45	4.81E+04	12467.75
max	2.30E+06	485873	197.3	45687	12450	3.29E+06	283001

Categorical summaries are as follows.

Statistic	Title	Sector	Industry	Hqcity	Hqstate
count	500	500	500	500	500
unique	500	21	73	239	37
top	Walmart	Financials	Utilities: Gas and Electric	New York	NY
freq	1	84	22	45	54

Basic Exploratory Data Analysis

1. Top 5 sectors in terms of Revenues are shown

Sector	Employees
Retailing	5,819,887
Financials	3,161,998
Technology	3,049,348
Health Care	2,341,423
Hotels, Restaurants & Leisure	1,838,742

It makes intuitive sense for retailing to employ the highest number of workers as many of them are likely part-time workers. Among retail, the industry with the largest workforce is General Merchandisers.

Sector	Industry	Employees
Retailing	General Merchandisers	3,377,600
Retailing	Specialty Retailers: Other	1,623,143
Financials	Commercial Banks	1,451,368
Food & Drug Stores	Food and Drug Stores	1,390,095

2. Sectors with most assets

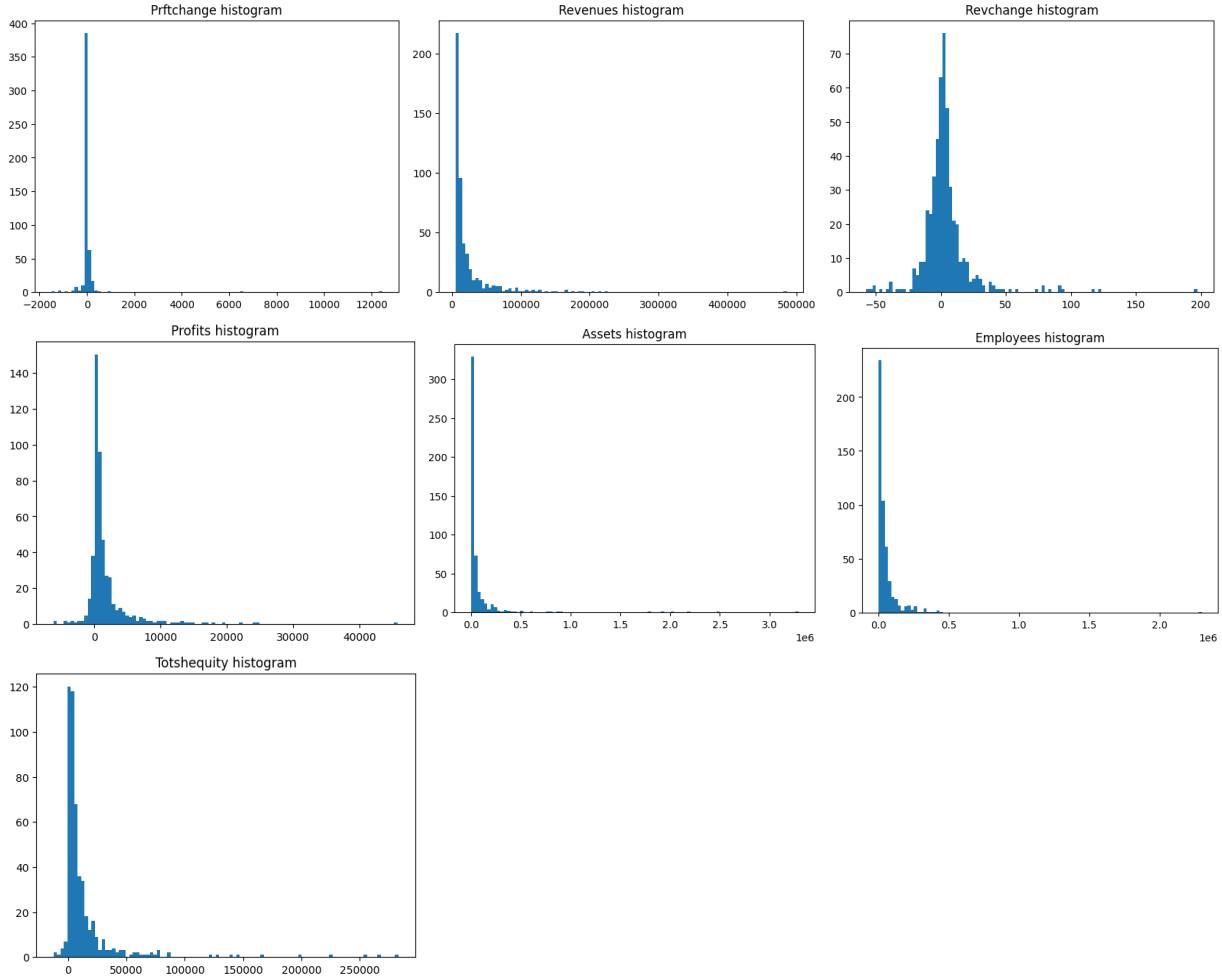
Sector	Assets
Financials	26,816,588
Energy	2,779,104
Technology	2,080,354
Health Care	1,598,998
Telecommunications	1,149,227

3. Sector with highest profit

Sector	Profits
Financials	228461.2
Technology	177901.4
Health Care	105383.9
Food, Beverages & Tobacco	56308.4
Retailing	46613.9
….	….
Energy	370.6

It is interesting to note that even though Energy ranks second on Assets, it makes the least profit among all the sectors.

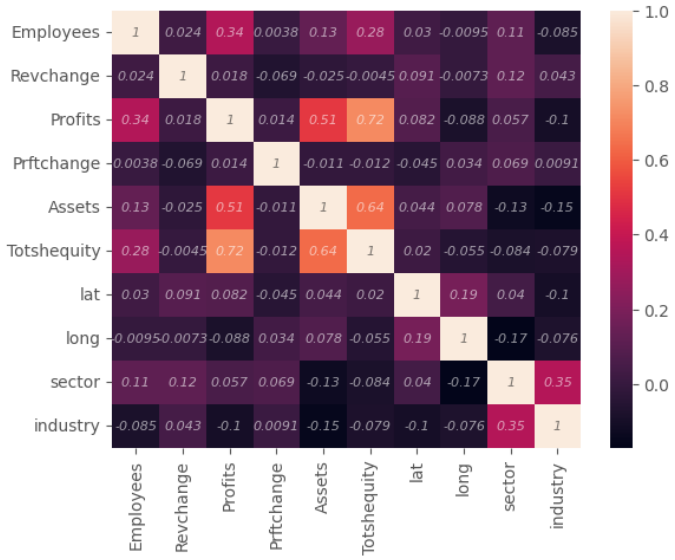
4. What do distributions of numerical values look like?



Feature Engineering

Hqzip is approximately geocoded into latitude and longitude with `pgeocode`. Categorical variables, sector and industry are encoded as numbers through `sklearn.preprocessing.LabelEncoder`.

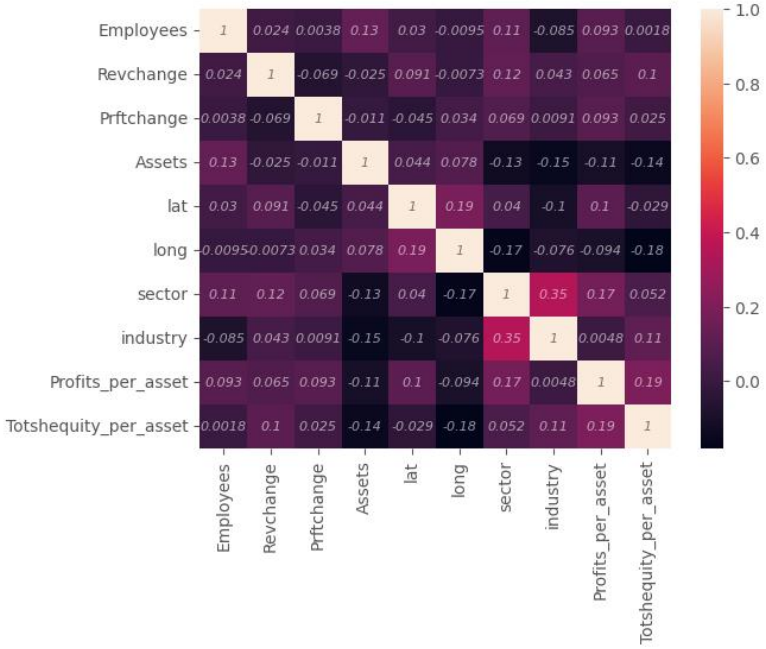
`seaborn.heatmap` plots the correlogram of the data.



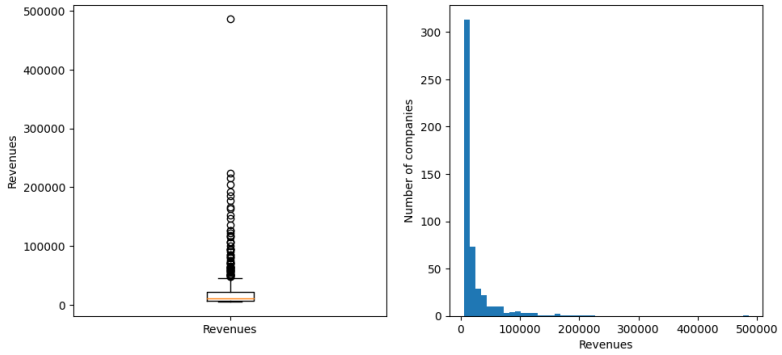
Some variables are closely related –

Variables		Correlation
assets	profits	0.51
assets	totshequity	0.64
profits	totshequity	0.72
profits	employees	0.34
totshequity	employees	0.28

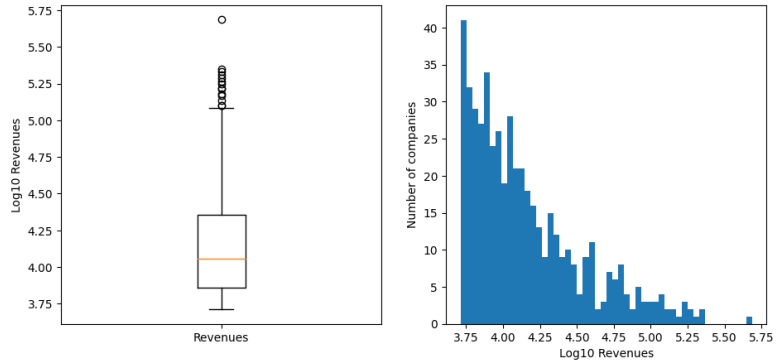
Expressing profits and totshequity as a function of assets can remove much of the multicollinearity in the data. They are swapped out for Profits_per_unit_Asset and Totshequity_per_unit_Asset.



Finally, all numerical covariates are scaled using `sklearn.preprocessing.StandardScaler()`.



Revenues follows a highly right-skewed distribution and there is **one very obvious outlier**.



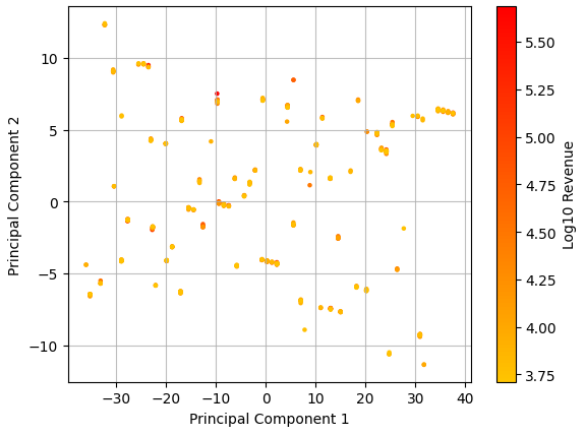
Taking a logarithm of Revenues makes the distribution slightly less skewed. It is worth looking at the correlation between each of the independent variables and the dependent variable (Revenue, in this case).

Covariate	Correlation with log(Revenues)
Employees	0.483058017
Revchange	0.028993547
Prftchange	0.025084715
Assets	0.380907409
lat	0.054539964
long	0.018492734
sector	0.049428754
industry	0.05286655
Profits_per_asset	0.035514893
Totshequity_per_asset	0.022903708

Note: Covariates barely correlated to the target **may** be left out during further predictive modelling.

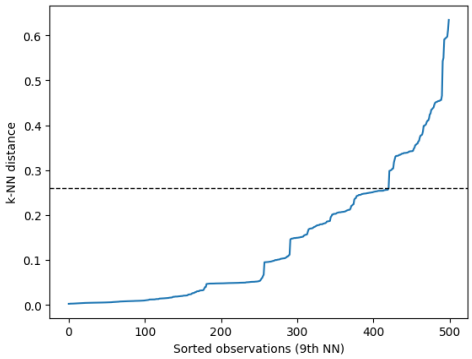
Clustering

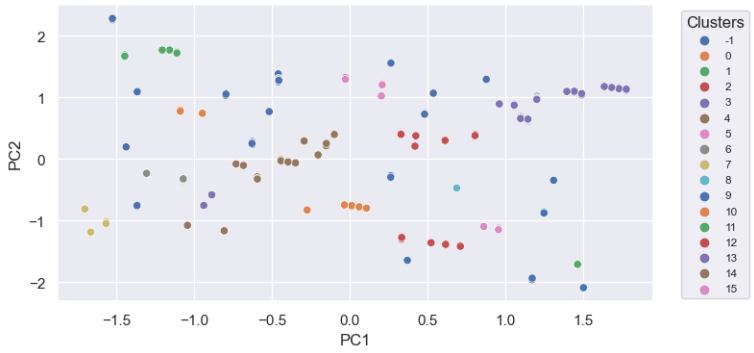
Principal Component Analysis is performed on the dataset (with dependent and independent variables). The first two principal components are `array([0.92257899, 0.06093741])`. Combined, they explain up to `98.2%` of the variance in the data. They are used as axes for clustering.



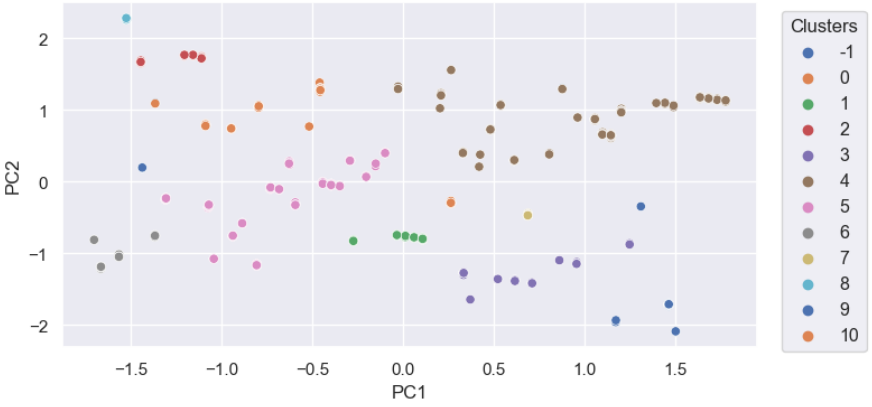
Both principal components are standardized using `sklearn.preprocessing.StandardScaler( )`.

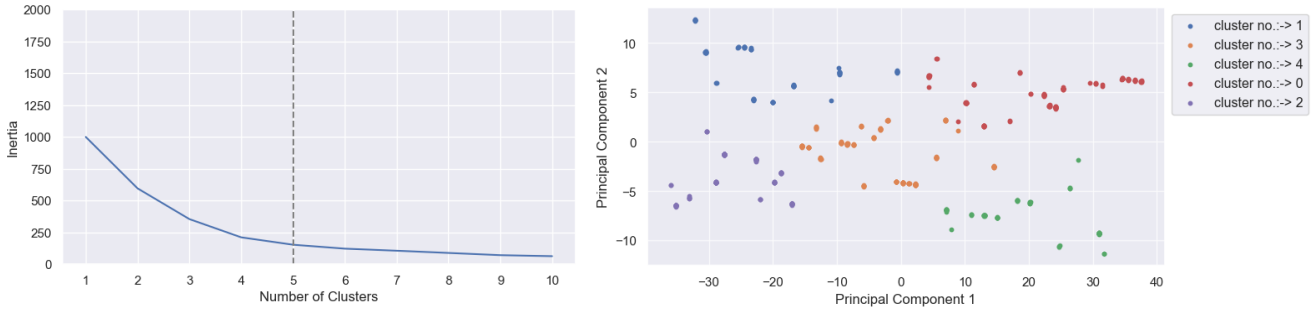
DBSCAN and k-Means are used for clustering. To apply DBSCAN, minimum samples (`min_samples`) to form a cluster and epsilon (`eps`) need to be known. The former is a judgement decision. The latter can then be approximately calculated from a sorted distance plot of k-Nearest Neighbours. As a rule of thumb, k can be min_samples – 1. Assuming `min_samples` = 10, the k-NN curve is as follows. No clear "crook" is seen in the elbow – epsilon is approximately set at 0.25.





There are many outlier points (-1 in the clusters denotes an outlier). For lowest number of outliers (6), samples and epsilon are respectively 5 and 0.4 – data is refit accordingly.





K-Means clustering algorithm is then applied. Ideal number of clusters are decided by plotting inertia against number of clusters[8]. The rate of change starts to plateau at 5 clusters.
Below is the result for k-means fitted with `n_clusters=5`.

Regression

`Walmart` is removed as it has an outlier-like Revenue. Variance Inflation Factor is checked for all independent variables. Sector and Industry are moderately related.

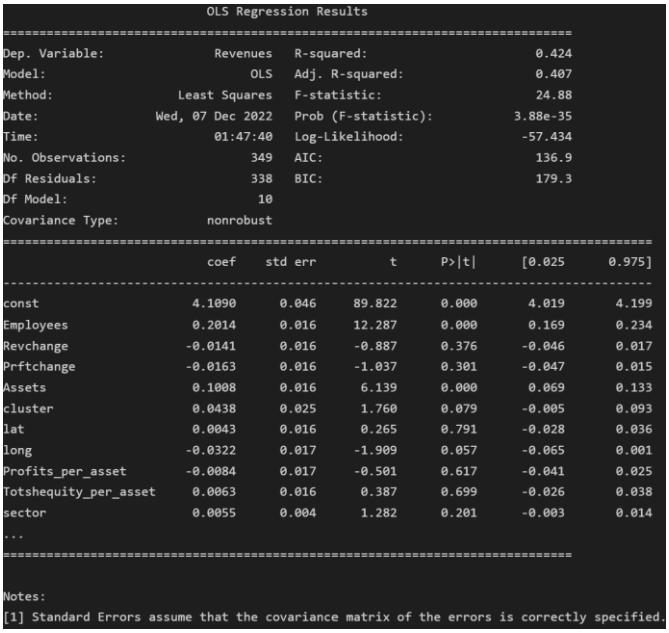
feature	VIF
Employees	1.052998
Revchange	1.040126
Prftchange	1.036324
Assets	1.054176
lat	1.105861
long	1.15727
Profits_per_asset	1.198158
Totshequity_per_asset	1.22852
sector	3.498679
industry	3.440126

Both cannot be used together in the model. After, removing industry, the corrected factors are →

feature	VIF
Employees	1.107303
Revchange	1.056591
Prftchange	1.018266
Assets	1.101016
cluster	1.279129
lat	1.091452
long	1.153139
Profits_per_asset	1.1615
Totshequity_per_asset	1.109326
sector	1.188361

Data is split into 70:30 proportion for training and testing, before being put through the same feature engineering pipeline. This is done separately so as to avoid train-test contamination. Ordinary least squares regression is utilized from the `statsmodels.api` library.

[8] Rena Nainggolan *et al* 2019 *J. Phys.: Conf. Ser.* 1361 012015 DOI 10.1088/1742-6596/1361/1/012015



OLS Regression Results

Dep. Variable:	Revenues	R-squared:	0.424
Model:	OLS	Adj. R-squared:	0.407
Method:	Least Squares	F-statistic:	24.88
Date:	Wed, 07 Dec 2022	Prob (F-statistic):	3.88e-35
Time:	01:47:40	Log-Likelihood:	-57.434
No. Observations:	349	AIC:	136.9
Df Residuals:	338	BIC:	179.3
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.1090	0.046	89.822	0.000	4.019	4.199
Employees	0.2014	0.016	12.287	0.000	0.169	0.234
Revchange	-0.0141	0.016	-0.887	0.376	-0.046	0.017
Prftchange	-0.0163	0.016	-1.037	0.301	-0.047	0.015
Assets	0.1008	0.016	6.139	0.000	0.069	0.133
cluster	0.0438	0.025	1.760	0.079	-0.005	0.093
lat	0.0043	0.016	0.265	0.791	-0.028	0.036
long	-0.0322	0.017	-1.909	0.057	-0.065	0.001
Profits_per_asset	-0.0084	0.017	-0.501	0.617	-0.041	0.025
Totshequity_per_asset	0.0063	0.016	0.387	0.699	-0.026	0.038
sector	0.0055	0.004	1.282	0.201	-0.003	0.014
...						

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Many of the independent variables are not contributing/ have a non-significant effect. BIC is noted to be ~179. Using only Employees and Assets, a more robust regression model is built. BIC for this model is lower at ~147.

Predictions on testing data yield a 57% mean absolute percentage error. Regularized versions - Lasso and Ridge regression both produce MAPE of ~60%. Polynomial regression of degree 3 with ElasticNet is applied. Upon tuning some hyperparameters, MAPE is brought down to 43%. Next, a Multi-Layer Perceptron is used for regression. MAPE hovers around 52% with several hyperparameter combinations. Ensemble models are often a powerful tool for tabular data. An XGBoost model is trained to bring down the MAPE to ~44%. With a RandomForestRegressor, the MAPE is further brought down to 39%.

Classification

With cluster as the target variable, data is split into 70:30 proportion. Using an MLP Classifier with hidden_layer_size of 35, accuracy_score of ~87% is reached. Cross-validating this with multiple hyperparameters over 10 folds results in accuracy of (93.3+/-0.5)%.

RESULTS

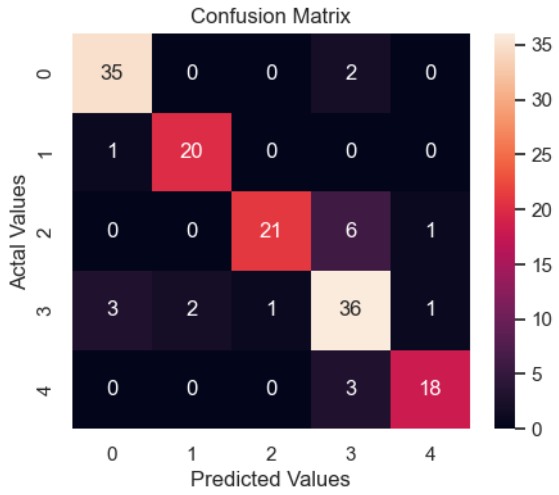
This section summarises training outcomes and modelling. Hyper-parameters with multiple entries denote all possible combinations have been tested.

Model	Hyper-Parameter(s)	Best Hyper-Parameter(s)	Metric	Metric Value
CLUSTERING				
DBSCAN	eps = [0.01,0.05,0.1,0.25,0.4] min_samples = [5,10,20,25,30,50]	eps = 0.4 min_samples = 5	Number of outliers	6
k-Means	"init": "random", "n_init": 10, "max_iter": 300, "n_clusters":[range(1,11)]	"init": "random", "n_init": 10, "max_iter": 300, "n_clusters" = 5	Inertia	~200
REGRESSION				
Linear Regression (all variables)	-	-	R-squared, BIC, MAPE	0.42, 179.3, 0.576
Linear Regression (restricted variables)	-	-	R-squared, BIC, MAPE	0.41, 147.6, 0.579
Ridge Regression	alpha=[0.0001 , 0.001 , 0.01 , 0.1 , 0.5 , 1.0 , 1.5 , 2.0 , 20]	alpha=0.0001	MAPE	0.605
Lasso Regression	alpha=[0.0001 , 0.001 , 0.01 , 0.1 , 0.5 , 1.0 , 1.5 , 2.0 , 20]	alpha=0.0001	MAPE	0.605
ElasticNet with Polynomial Regression	alpha = [0.0001 , 0.1 , 1 , 5] l1_ratio = [0.001, 0.01, 0.1 , 0.25]	alpha = 0.0001 l1_ratio = 0.001	MAPE	0.432
MLP	hidden_layers = [20,50,100,500,1000] alpha=[0.01 , 0.1 , 1.0]	Hidden_layers=1000 , alpha = 1.0	MAPE	0.523
XGBoost	n_estimators = [50,100,500,1000] max_depth = [2,5,10,20]	n_estimators = 500 max_depth = 2	MAPE	0.4402
Random Forest Regressor	n_estimators = [50,100,500,1000] max_depth = [2,5,10,20]	n_estimators = 500 max_depth = 10	MAPE	0.3964
CLASSIFICATION				
MLP Classifier	hidden_layers = [20,35,50,100,500] alpha = [0.001 , 0.1 , 1.0]	Hidden_layers = 500 Alpha = 5.0	Accuracy	0.93

Results from clustering and cluster profiling (*see: Discussion*) hint at there being 5 different 'groups' of companies in the Fortune-500. For regression, `RandomForestRegressor` has the `feature_importances_` method which shows the weighted importance of each feature:

features	importance
Employees	0.359325
Revchange	0.026788
Prftchange	0.027343
Assets	0.39633
cluster	0.007132
lat	0.037435
long	0.025507
Profits_per_asset	0.053174
Totshequity_per_asset	0.025657
sector	0.011359
industry	0.029951

Classification can be termed a success as the MLP has ~93% accuracy over 5 clusters. A confusion matrix is generated with the classifier predicting unseen data. There are not too many misclassifications, although the model tends to overpredict Cluster 3 (this can be due to the abundance of cluster 3 in the training data or because cluster 3 is the most centrally positioned cluster – sharing boundaries with others).



DISCUSSION

For clustering, K-means appears to be a better choice because of a lower number of clusters as well as absence of outliers.

Cluster Profiling

Sectors appear to fall into either the grouping of clusters 0 and 1 or clusters 2,3 and 4. With the exception of Industrials and Materials, the sectors follow this bi-partite grouping.

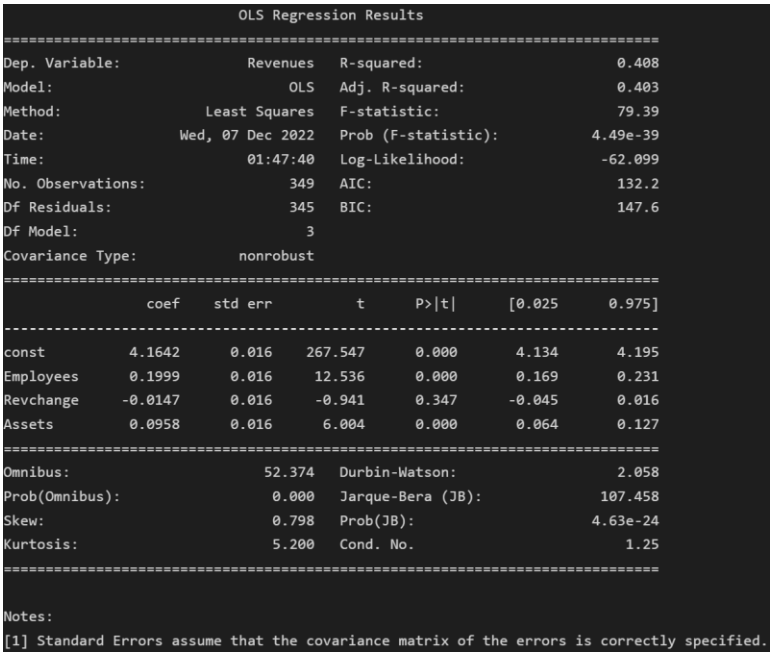
	cluster				
Sector	0	1	2	3	4
Aerospace & Defense	0	0	12	0	0
Apparel	0	0	5	0	0
Business Services	0	0	14	0	6
Chemicals	0	0	14	0	0
Energy	0	0	7	0	50
Engineering & Construction	0	0	7	6	0
Financials	0	0	31	42	11
Food & Drug Stores	0	0	0	7	0
Food, Beverages & Tobacco	0	0	3	18	3
Health Care	0	0	0	38	0
Hotels, Restaurants & Leisure	0	0	0	10	0
Household Products	0	0	0	11	1
Industrials	0	13	0	6	0
Materials	10	3	0	6	0
Media	2	9	0	0	0
Motor Vehicles & Parts	9	0	0	0	0
Retailing	29	18	0	0	0
Technology	23	20	0	0	0
Telecommunications	10	0	0	0	0
Transportation	11	6	0	0	0
Wholesalers	29	0	0	0	0

Looking at the cluster map from k-means, it is easy to see that clusters 0 and 1 occupy the top half, whereas 2,3 and 4 are located towards the bottom.

A histogram of numeric variables for each cluster is shown below.

Clustering seems to have done an acceptable job as none of the distributions are very skewed (with the exception of a very high value along Revenue, which has been discussed earlier – this will be removed during modelling).

Results from regressions point towards there not being sufficient predictive power in the covariates. Because of this, the models are likely underfitting the data. Therefore, regularization, (which helps prevent overfitting) is not improving the score, as seen above. With the most stable regression model (in terms of BIC), it can be seen that Employees and Assets are the only two statistically significant covariates for Revenues. This agrees with Random Forest model that Employees and Assets are the most important features. Positive slope for both indicates that Revenues increase with these factors. It is, however, worth noting that relationships between predictors and the target are not very linear to begin with.



OLS Regression Results

Dep. Variable:	Revenues	R-squared:	0.408
Model:	OLS	Adj. R-squared:	0.403
Method:	Least Squares	F-statistic:	79.39
Date:	Wed, 07 Dec 2022	Prob (F-statistic):	4.49e-39
Time:	01:47:40	Log-Likelihood:	-62.099
No. Observations:	349	AIC:	132.2
Df Residuals:	345	BIC:	147.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.1642	0.016	267.547	0.000	4.134	4.195
Employees	0.1999	0.016	12.536	0.000	0.169	0.231
Revchange	-0.0147	0.016	-0.941	0.347	-0.045	0.016
Assets	0.0958	0.016	6.004	0.000	0.064	0.127

Omnibus:	52.374	Durbin-Watson:	2.058
Prob(Omnibus):	0.000	Jarque-Bera (JB):	107.458
Skew:	0.798	Prob(JB):	4.63e-24
Kurtosis:	5.200	Cond. No.	1.25

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

To improve predictive power of regression, further data can be brought in to provide more predictive power – e.g., one may fetch last few years of stock market activity with the companies' tickers, as an indicator of financial health.

More intricate clustering methodologies such as hierarchical clustering may be explored especially since there exists a tiered structure within the data.

As an extension, neural networks may be applied for regression. There has been research[9] into utilization of convolutional neural networks on tabular data, in essence, bypassing the fact that there is no spatial correlation between neighbouring cells in tabular data (which is what vanilla CNNs assume).

For classification, minority oversampling techniques[10] can be used to gain a more uniform distribution of classes, thus combating the problem of imbalanced distribution.

[9] Zhu, Y., Brettin, T., Xia, F. et al. Converting tabular data into images for deep learning with convolutional neural networks. Sci Rep 11, 11325 (2021). https://doi.org/10.1038/s41598-021-90923-y

[10] Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal Of Artificial Intelligence Research, 16, 321-357. doi: 10.1613/jair.953