FINAL PROJECT

Mian Muhammad Hasnain Ali

Student Number: 501144656  
Professor: Ceni Babaoglu

Course: CIND 820, Submission Date: December 05, 2022

Contents

[Abstract 3](#_Toc121176291)

[Answer to Additional Questions 4](#_Toc121176292)

[Summary of Research Papers 6](#_Toc121176293)

[Descriptive Dataset Statistics 9](#_Toc121176294)

[Initial Analysis 11](#_Toc121176295)

[Univariate Analysis: 11](#_Toc121176296)

[Visualizations 12](#_Toc121176297)

[Analysis 13](#_Toc121176298)

[Bivariate Analysis: 16](#_Toc121176299)

[Visualizations 17](#_Toc121176300)

[Analysis 18](#_Toc121176301)

[Exploratory Data Analysis (EDA) 20](#_Toc121176302)

[Description of Applied Methodology 22](#_Toc121176303)

[Graph Showing Overall Methodology 22](#_Toc121176304)

[Brief Explanation of Each Step 22](#_Toc121176305)

[Modelling, Evaluation, & Results 24](#_Toc121176306)

[Conclusions 31](#_Toc121176307)

[Shortcomings of the work and concluding remarks 32](#_Toc121176308)

[GitHub Repository Link 33](#_Toc121176309)

[Reference Page 33](#_Toc121176310)

# Abstract

Each year Fortune Magazine publishes a list of companies with highest revenues in the United States, it’s called “Fortune 1000”. This list has two subsets, Fortune 500, and Fortune 100 respectively. The most popular and widely used list out of the three is Fortune 500 (Sarath, 2022). According to the Fortune magazine’s website, the top 500 companies generated $16.1 trillion in revenues and $1.8 trillion in profits for the year 2021. Therefore, naturally any news on these companies attracts a wide range of audience. From an individual to a large corporation, a lot of decisions are made everyday based on the rankings of these companies. Whether it’s a potential employee, a supplier, or an investor, the rankings affect all its stakeholders. For the company itself, it’s a badge of honor to be included in the Fortune 1000 list.

As mentioned before, these rankings of corporations are based on highest revenues. But the problem with highest revenue generating companies is it does not necessarily mean they are the best performing companies. Revenue is the money brought into the company by its business activities, whereas profit is the money left after paying all its payables. Consequently, regardless of how much money a company makes, if it is unable to save enough for itself then this could mean it lags efficiency which means potential problems. This situation could question company’s going concern. As the business generates more revenue by selling its products and services, it requires more money to fund its working capital. A Working capital is calculated by subtracting current liabilities from current assets, as listed on the company’s Balance Sheet. Current assets include cash, accounts receivable and inventory. Current liabilities include accounts payable, taxes, wages and interest owed (Beaver, 2022). Therefore, more revenue requires more money to fund the working capital, inability to do so means severe problems. Small profit or a loss for the financial year could also repel a lot of potential investors and other stakeholders. But if a company is profitable, it could share that profit with its existing shareholders in the form of dividends or it can be kept as retained earnings of the company. With these points in mind, it can be argued that a better approach is to rank companies with the highest profits instead.

This project aims to predict profits of Fortune 1000 companies using supervised learning with regression models. To determine which model performed the best, a detailed evaluation will be carried out. Usually predicting profits are performed by Financial Analysts or the Accountants. This task involves taking a lot of factors into consideration and performing a detailed analysis on a company’s financial statements. But we are using a very different approach yet aiming for the same output, this project demonstrates predicting profits using machine learning models. In theory this can save a lot of time, effort, and resources. The dataset used is from Kaggle titled “Fortune 1000 Companies by Revenue 2022”. The data goes through the cleaning process followed by a detailed EDA and modelling to predict the profits. The tool used for the project is Python and the platform is Jupyter Notebook. Some of the libraries used include NumPy, Pandas, sklearn, Scikit-learn, Seaborn, matplotlib, etc.

# Answer to Additional Questions

**Question 1:** **What do you already know about the topic?**

After carrying out extensive research it became obvious that there has not been very much research done on this topic. The attributes contained in the dataset are related to Accounting and Finance. If one were to pick financial statements of any publicly traded company, they would find similar attributes in those documents as well. Account and Finance have been around way before computer science was invented let alone Machine learning. So, there are many techniques available to predict not just profit, but pretty much all the elements of financial statements. Companies forecast the profit figures not just for the next quarters but for many years to come, but that is usually for company’s internal use. Even if those figures are shared publicly, still it would have been a lot of work compiling a list of 1000 companies with their profits.  
Even though there has been a considerable amount of work done on related topics but surprisingly no research has been done in specifically predicting profit of fortune 1000 companies. Mostly researches on Fortune 1000 companies are related to predicting next year’s share prices but not profit.

**Question 2:** **What do you have to say critically about what is already known?**

Despite being so much research done especially in the areas of predicting continuous variables, but one area has been specifically ignored and that is predicting profit figures for Fortune 1000 Companies. Most of the studies are related to predicting next year’s share prices. This is somewhat understandable because company’s share price is the single most important figure and is probably a cause for a lot of decision making. Therefore, its not a surprise that the focus of most research work has been towards predicting the share price.

Predicting a company’s profit is primarily a core finance and accounting topic. These predictions are based on conventional techniques even though there has been a lot of development in the recent years in finance with the rising trend of data science. But despite so many techniques around, still the focus has been entirely towards predicting sales revenue of companies, cashflow, share price, return on investment and some other key variables. Profit as a separate target variable is also predicted but only for a single company or its related companies. But when we look at predicting profit for fortune 1000 companies or S&P 500 or any other top companies, then no field has yet produced good research on this topic as yet.

**Question 3:** **Has anyone else ever done anything exactly the same?**

As mentioned before no one has ever done anything exactly the same.

**Question 4: Has anyone else done anything that is related?**

There are many examples of projects and works that have been done in the past which are not exactly related but the idea is to predict a specific variable. The Fortune 1000 companies’ data has not been used. If we look at the other datasets that are related are S&P 500, Forbes, and Bloomberg. None of these datasets have been used for the same purpose of research as our research. When we talk about the technique to predict a specific variable in the dataset then yes there has been many such cases where a prediction is made using Machine Learning techniques. Some examples can be found in the research paper discussion area within this project.

**Question 5:** **Where does your work fit in with what has gone before?**

As mentioned, there has not been any similar work done before. There are uncountable examples where the company’s profit is predicted using Accounting or Finance techniques. Similarly, there are so many projects within the Machine Learning where a variable is predicted using different algorithms and techniques. But there is not even a single project, Accounting or Machine Learning based where profit is predicted for Fortune 1000 Companies. There is so much that needs to be carefully thought through to get the right figure. Starting from the sales revenue forecast followed by accounting for costs of sales, salaries, and other expenses to get to the gross profit. From here, to get to the Profit after tax figure, the gross profit is then adjusted for two of the biggest and complex figures, interest and taxation. Therefore, a lot of financial, nonfinancial information along with other factors are considered to come up with an estimate. If only relied on Accounting & Finance techniques, it is not possible to estimate profit with the limited information that we have in our dataset. And to forecast profit for 1000 companies it would not only take a lot of time but to gather all the data first would be an extremely difficult task.   
On the other hand, most of the research carried out within Data Science and Machine learning on similar datasets such as Fortune 1000, S&P 500, Forbes, and Bloomberg are to either predict the share price of a company or something else.

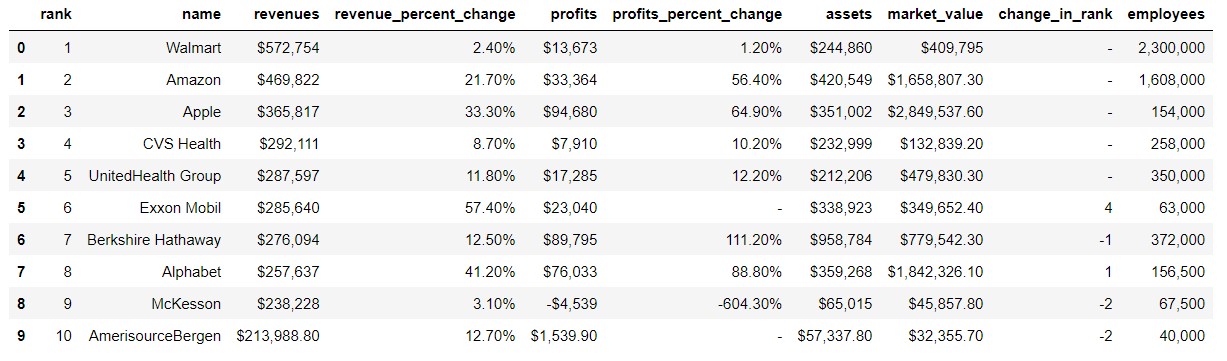
**Question 6:** **Why is your research worth doing in the light of what has already been done?**

This research is unique and is the first of its kind. No one has ever done anything similar. This research will pave the way for other future works of similar nature and act as a reference point. Same ideas can be applied to other subsets of Fortune 1000, Forbes, S&P 500, and Bloomberg Businessweek list.   
The Fortune magazine started publishing the list of highest companies by Revenue since 1955. From that time onward up till present, the general perception is that a company with highest revenue is the top ranked company.

# Summary of Research Papers

1. In this research regression models are evaluated over predicting used car prices. Out of the three models, Gradient Boosting, Random Forest, and Multiple Linear Regression, the first model produces the best result with Mean Square Error of 0.28 while the other two models gave a lower MSE. The purpose of this research was to develop the best price predicting model of used cars (Nitis Monburinon, 2018)  
   The approach taken in this research is somewhat like our prediction of the profit using almost the same models.
2. This is another example of using Machine Learning models to predict a variable ‘Price’ of used cars. This project also relates to our research in a sense that there is one target variable and there are four models used to predict the variable. These four techniques include Multiple Linear Regression, K-Nearest Neighbours (KNN), Decision Trees, and Naïve Bayes. Models were evaluated using 10-fold cross validation. The problem with the study is that the number of records were cut down from +400 to 97 which hampered the overall performance (Pudaruth, 2014)
3. In a paper published by National Library of Medicine same approach is applied in predicting healthcare costs. But the study is on a relatively larger dataset that had 90,000 individuals, 6.3 million medical claims and 1.2 million pharmacy claims. The models applied in our research are used here in addition to many other models, with Gradient Boosting providing the best prediction results. (Morid, 2018)
4. Another finding where four regression models are tested to predict the aggregate space and water thermal load in buildings in Sweden. The models used are SVM, Regression Tree, Feed Forward Neural Network, and Multiple Linear Regression. After a detailed analysis and evaluation of the models it is revealed that the best model in predicting the variable is SVM. Not to mention the other two models namely Neural Network and Multiple Linear Regression also performed well enough with an exception of Regression Tree (dowu, 2016)
5. One of the many more examples on predicting a target variable using regression techniques is this project where house prices are predicted using Multiple Linear Regression, Lasso, Ridge, SVM, and XGBoost are applied followed by evaluation of the models using RMSE, R-Square, and RMSLE. (J Manasa, 5 March, 2020)
6. In this paper by Huang Lei and Huang Cailan they adopted a very effective approach in predicting the business’s revenue data by support vector machine. The dataset attributes somewhat resemble the Fortune 1000 companies by revenue. The SVM outperformed other models such as Random Forest, Gradient Boosting Regression Tree. There was not too great of a difference between all three but former scored the best when evaluated by MAE, RMSE, and MAPE. (IEEE XPLORE)
7. This research is carried out on one of the major airline groups which consist of 26 airlines with a large data set to predict profitability. Models used for the task were three Decision Tree, and Logistic Regression. The Decision tree did outperform the Logistic Regression. As per the ASE values, the best predictive model was the 2-branch percentage value decision tree. (Choi, 2019)
8. In this particular paper, a financial statement analysis is performed on historical data to predict stock returns. The elements of financial statements which are used here are almost the same as what we are using for our project. Out of the models used: Random Forest, Neural Network, and Linear Methods, the first one performed the most accurate and highest abnormal return. (Amel-Zadeh, 2020)
9. The dataset used in this research has 41267 observations and is based on S&P 500 Companies, which are all part of the dataset of our research “Fortune 1000 Companies by Revenue”. Except our dataset has 1000 observations. Both researches are similar in a sense that they attempt to predict future values. The machine learning models used are Linear regression, Polynomial regression, and Support Vector regression. From the analysis it is evident that the latter (SVR) outperformed the other two models. The Average Standard Deviation and MSE for SVR is 5.6 and 0.14 whereas it is 25.59, 0.55 for Polynomial Regression, and 29.70, 0.66 for Linear Regression. (Sakhare, 2019)
10. This project is one of its kind where the researchers predicted the future gold rates based on the data which included interest rates of different countries, Commodities rates, Bonds, Currencies, and Shares. The two machine learning models; Linear Regression, and Neural Networks were used to predict the values. The reason why this project is relevant is because this dataset even though very diversified but does not have any categorical attributes, plus supervised machine learning models are used to predict the future gold rates. (Sami, 2018)
11. This research paper demonstrates which variable is the most important in predicting next period’s profit. It takes six years of data from 119 companies listed on Tehran Stock Exchange. Essentially four hypotheses were designed and tested. It concluded, that out of the four variables, the net operating profit after tax has the highest coefficient of determination (61%) in predicting next period’s profit. The techniques used are all purely finance related. It is common in Financial Management to forecast attributes of the financial statements for many useful reasons. (Shourvarzi, Analysis of the predictive ability of the components of economic value added in predicting next period’s operating profit: Evidence from Iran, 2011)

# Descriptive Dataset Statistics



The dataset “Fortune 1000 Companies by Revenue 2022” is taken from Kaggle. It has 10 attributes with 1000 observations. Below is a brief description of these attributes:

* **Rank:** This very first attribute in the dataset shows the ranking of Fortune 1000 companies. Number 1 being the highest and number 1000 is the lowest in the order. It has zero null values, and the data type of this attribute is ‘object’.
* **Name:** The names column lists down all the names of companies. The data type is ‘object’ and It has no missing values.
* **Revenues:** It represents sales revenues of all the companies, which is the money brought into the company by its business activities, be it selling products or services. It contains six figure numbers with special character dollar sign ‘$’. Data type of this attribute is float64 and it does not have any missing/null values.
* **Revenue\_percentage\_change:** This is the percentage change by which last year’s profit has changed. The positive figures mean that revenue has increased and negative means it has declined. The data type is float64 with special character ‘%’ and there are 992 non-null values.
* **Profits:** Profit is the difference between revenue earned for the year subtracted by total costs. If the difference is positive, it’s a profit, if negative then it’s a loss. This attribute is a float64 data type with 997 non-null values.
* **Profits\_percentage\_change:** It is a float64 data type and has 727 non-null values. The values are in percentage format and represent the amount from which the profits have increased/decreased compared to last year.
* **Assets:** It is the amount at which a company records its assets in the financial statements. It does not mention in the data set description if this attribute refers to noncurrent assets or current assets. The former are the ones which are held by the business for long term and latter for short term. We would assume that the Assets attribute is the sum of both current and noncurrent assets. It does not have any null values and is a float64 data type. Values are shown in millions.
* **Market\_Value:** This is the value at which the company’s share price is trading on the stock market. It is calculated by multiplying total number of shares issued with share price at which shares are traded on stock market. Almost always, the market value and the book value of a company is different. If this figure is higher than the book value it means the company is performing well, if its lower than the book value then it means that the shareholders have lost confidence in the company. The market\_value attribute has 955 non-null values with float64 data type.
* **Change\_in\_rank:** This column tells us by how many ranks the company has gone up or down in the ranking as compared to last year. It has 1000 non-null values with ‘object’ data type.
* **Employees:** With float64 data type and 999 non-null values, the employee attribute represents the number of people employed by the company.

With the help of describe() function, important mathematical information can be displayed to get a better understanding of the dataset at hand. The describe function was applied to the entire data set with 10 attributes but it only returned 7 with float64 data type and ignored the other data type ‘object’.

Table

Description automatically generated

# Initial Analysis

## Univariate Analysis:

As a starting point all the special characters are removed from the dataset. The data has different special characters included within. Removing them is necessary to be able to prepare the data for further use. It is common practice to deal with special characters in the cleaning phase. Failure to remove special characters can cause errors while processing commands that can lead to delays in completing the task. Therefore, dealing with them at the cleaning phase is most suitable.  
Two of the columns in our data set have special characters with percentage sign (%). These percentage values are converted to decimals to have identical formats. The ‘$’ sign along with ‘,’ and parenthesis are removed, leaving all float64 data type columns with decimal values.  
It is also important to rename the column names so that they don’t cause confusion later in the analysis especially if additional columns are added. Therefore, column names have been renamed.

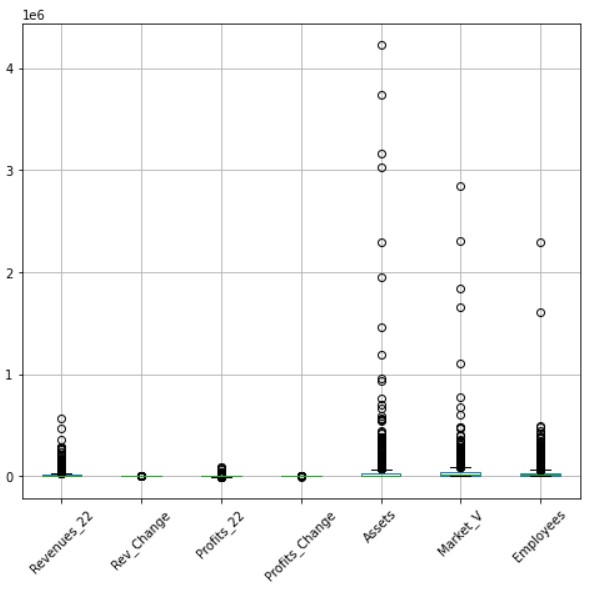
A detailed data dictionary is created, the code of which can be found in the relevant coding file. Instead of simply printing the data dictionary, a dictionary function is created which asks for user’s input for description of attributes. Creating a data dictionary is considered good practise and is recommended for one’s own ease as well as for the ease of others who work on the same data set.

Some of the attributes in the dataset have null values included. Ignoring null values can cause problems especially when modeling the dataset. Many libraries require the data to be cleaned first. A good example would be using sklearn, TensorFlow, or any other machine learning or deep learning packages, it is a requirement to clean/imputate null values before passing the data forward for additional commands or modelling. Otherwise, the output will be an error message.  
Another reason for treating the null values before modeling is that when an unknown value is compared to another value, the result is unknown, which affects the output.   
Our profits\_change attribute had the most null values of 273, followed by 45 for market\_value, and less than 10 each for rev\_change, Profits, or Employees. To treat these null values, we applied imputation method to replace null values with mean values of their respective columns.

Checking for outliers is very important at this stage as having them could mean bad data. The causes could be the result of poor data entry, processing errors, or poor sampling etc. There could also be true outliers in the data which represent natural variations and they could be left as is in the data. When we look for outliers in our dataset and examine each 7 of the float64 datatype attributes, Assets, Market Value, and Employees columns stand out the most in the boxplot shown below. But we cannot base our opinion only on the boxplot therefore we will also plot each attribute individually to see the spread of it and get a better understanding of our data. Furthermore, we are dealing with financial information of companies which requires more scruitny of the data to be fully understood.

The visualization of the attributes is as follow:

Visualizations:



Chart

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart

Description automatically generated

If we compare the individual attributes with their relevant boxplot representation, and the nature of that data itself, we see that each one of them can be understood in a better way. We will explore each column separately below:

Analysis:

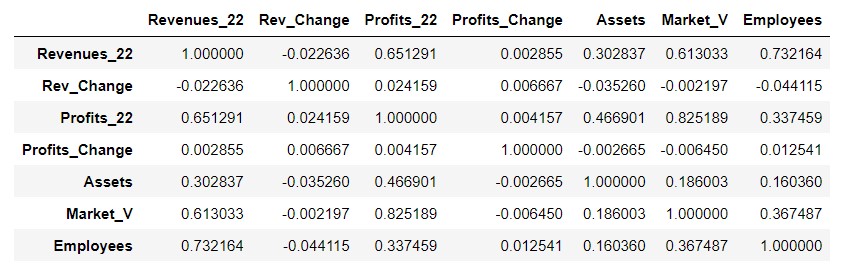
* **Revenues\_22:** The boxplot shows some values apart from the main clutter but when compared with the Revenue\_22 graph, it starts very high on the y axis but falls and continues to decrease till it reaches the last company in the list. This makes sense because these companies are ranked as per the highest revenue earned in the year. Therefore, there are no outliers in this column.
* **Rev\_Change:** For revenue change, which represents the percentage by which the revenue changed by last year apparently has no outliers. The boxplot does not show any evidence of it. But revenue percentage graph shows a few spikes. The data description table confirms this by indicating the max value of Rev\_Change column to be 21.99% which is not abnormal. It simply means that the company sold 21.99% more of its products or services this year. Throughout in the rest of the Rev\_Change column there are only two more instances where the revenue percentage increased between 10% to 15%. It is also noteworthy to see that these companies with increased revenue did not have the same kind of increase in their profits. This means the companies ended up paying more in costs as increased revenues increases the expenses as well.
* **Profits\_22:** There is no evidence of any outlier in this attribute. The boxplot confirms this, and the graph shows a decreasing curve with no outliers. The profit figure is directly proportional to revenue. But there are other factors which impact the profit as well. Profit is calculated by subtracting all expenses from revenue which includes cost of selling the products & services, salaries, administrative expenses, interest, taxes just to name a few. So, if a company generates high revenue, ideally its profits should be high as well, but it does not always happen like this because of aforementioned variables. If the costs are more than the revenue earned it could turn into the loss instead of profit. This can be observed in the profit graph where anything below 0 is a loss. Therefore, it really comes down to the company’s efficiency in saving enough money for itself after accounting for all the payables.   
  When we look at both the revenue and the profit graph side by side, the shape essentially for both look the same - with exception of some companies below 0 making loss in this financial year - the differences however are the costs, as some companies have higher costs where the other companies have lower costs.
* **Profits\_Change:** There are four occurrences where the values change significantly than the average values. Three out of four values are positive which means that the companies with these values have made more profits. But an important thing to note here is that these companies did not have the same kind of increase in their revenues. The only possible explanation to this would be the companies must have cut down in their costs to save more money in profits.  
  The fourth value which is the minimum (negative) value in the whole column indicates that this particular company has made significant loss for the year. This company is in the fourth quarter of the dataset meaning its ranking is almost at the bottom of the 1000 companies. Looking at its revenue change value it shows there is no significant change in its revenue compared to last year, however it made huge loss in current year.
* **Assets:** The boxplot shows some values further away from other data. This can also be seen in the Asset graph’s first quarter which shows high spikes. The columns min, max, mean, and other values confirm these differences. This should not be a surprise rather these differences should be expected given the nature of the Assets attribute. Companies operating in different industries have different kinds of assets. A company involved in large scale manufacturing would keep assets like plant, machinery, warehouses etc as compared to a service providing company which offers its services remotely. As briefly explained in data description, Assets values are recorded book values which are subject to accounting rules like depreciation, revaluation and disposing an asset at its residual value. Therefore, asset values can never be negative which is evident in the graph. This can be proved by looking at the minimum value in assets column which is a positive value.
* **Market\_Value:** We do see in the first quarter there are high values. In comparison to assets, it can be said that it is slightly better in terms of the spread. High values resulting in spikes in the graph are expected as this attribute is the value of the company at which its shares are traded in the stock market. The share prices change every day and so does the market value of the company. One of the major commonalities in market value and assets is that both contain positive values. There would be differences in these two values as one is recorded strictly by applying accounting rules and the other is calculated company’s share price. However both are used together to understand if the company is performing better or worse.
* **Employees:** It cannot have a negative value because of the nature of this attribute. There are two very high values right at the beginning of the data, but it should not come as a surprise because the companies are Walmart and Amazon with 2.3 million and 1.6 million employees. But if we look at the rest of the data there are differences, but nothing compared to the first two companies. The number of employees working in a company differs because of its size, industry, and business sector.

The column name Rank\_Change is of no significance and therefore dropped. Similarly, the attribute Rank is only sequential numbering of the companies, and it is also dropped because of having no significance value. This leaves us with a total number of 8 attributes including company names.  
There were additional columns generated with the help of existing columns for example the Profits\_21 column was generated with the help of Profits\_22 and Profits\_Change. Similarly, Revenue for the previous year was generated along with costs for two years. These columns were eliminated because of the reason explained in the shortcomings of the work and concluding remarks at the end.

After considering different aspects and examining our dataset, it can be said that the data is consistent, and it is ready to be explored further for Bivariate and Multivariate analysis. The target variable **Profit\_22** has been chosen to answer our question.

## Bivariate Analysis:

Now that we have explored each one of our attributes in detail, it is time to explore them in pairs and understand the relationships with one another. Firstly, we look at the correlation table. This is to find out if our data contain variables that are highly correlated and can be eliminated for a better analysis.



We can see from our correlation table that there are no relationships that are either perfectly positively or perfectly negatively correlated with each other. However, the relationships fall under three categories:

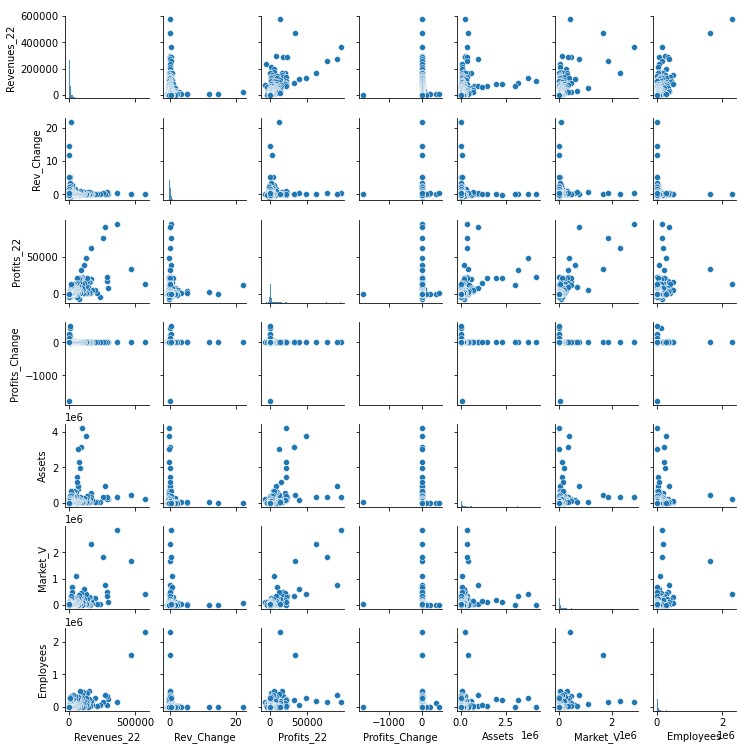
* Weak/Small for values between 0.1 to -0.1
* Moderate/Medium for values between 0.3 to -0.3
* Strong/Large for values between 0.5 to -0.5 (or higher than 0.5 but less than 1.0, and lower than -0.5 but not lower than -1.0)

To explore our correlations further and comment on their relationships we would use visualizations to dig deeper. For this a heatmap can be used.

Visualizations:



We can also plot pairwise scatterplot visualizations to help us with our analysis.



With these visualizations, let’s discuss some important correlations in our data.

Analysis:

* **Revenues\_22 & Profits\_22:** The correlation detected between Revenues\_22 and our target variable Profits\_22 is 0.65. This shows a strong positive correlation. Meaning, if one changes, the other changes in the same direction, not by the same value but it has a strong impact. Using a Spearman correlation test we get the coefficient value of 0.610. With alpha being 0.05 the p-value is p=0.00 therefore confirming the samples are correlated. The code of this can be found in the relevant code file. This also corroborates the fact that if a company generates more money by selling more of its products or services there is a possibility to make higher profits given costs are handled efficiently.
* **Revenues\_22 & Market\_Value:** The correlation value we see for these two attributes is 0.61 which is considered strongly positively correlated. The spearman test calculates the value of 0.577 with p-value p=0.000. It is true that a company’s revenue could impact its market value as well (HARPER, 2022). Based on our correlation coefficient of 0.61 we can say they do indeed affect each other.
* **Proftis\_22 & Assets:** According to Investopedia, “Return on assets (ROA) is an indicator of how profitable a company is relative to its assets or the resources it owns or controls” (MCCLURE, 2021). Therefore, these two figures are linked with one another from accounting’s perspective. The correlation between Profit and Assets is 0.47 which is almost a strong correlation. The spearman coefficient is calculated to be 0.647 and the p-value p=0.000 which confirms a correlation.

Similarly, for the rest of our data, there is a positive correlation between all the attributes in our data except Rev\_Change and Profits\_Change. The reason is that both these variables are percentages which have been converted to decimals. Their relation can be seen as mostly negatively correlated with the rest of the data. However, the category of negative correlation it falls under is low/weak.

When we look at the correlation of attributes with only the target variable Profit\_22, and the target set at greater than 0.3 we get the following results that show the Market\_V is the most highly positively correlated attribute followed by Revenues\_22.

|  |  |
| --- | --- |
| Revenues\_22 | 0.651291 |
| Profits\_22 | 1.000000 |
| Assets | 0.466901 |
| Market\_V | 0.825189 |
| Employyes | 0.337459 |

We will be dropping the two variables from the datasets which are of no significance. First is the ‘Rank’ column. It contains numbers starting from 1 to 1000. Python has the built-in feature to sequentially number the data so keeping the Rank column does not make any sense. The second column is the ‘Rank\_Change’ column. It contains values which are negative numbers, positive numbers, or null values. The negative numbers indicate the rank changed where the company has gone down in ranking. The positive numbers show the opposite.

# Exploratory Data Analysis (EDA)

Now that the data has been cleaned, modified, and explored with the help of Bivariate and Univariate analysis, its time to do the exploratory data analysis. Although the analysis performed previously during initial analysis includes a big chunk related to EDA. This section will perform additional EDA to take things further and to answer more questions about the data.

The dataset at this stage has 8 attributes including ‘Name’. By using the first two attributes ‘Revenues\_22’ and ‘Rev\_Change’, a new attribute ‘Revenues\_21’ can be produced. The Rev\_Change is the percentage change by which the revenue has changed compared to last year. If Rev\_Change and Revenues\_22 columns are multiplied, we can get the values for a third column. Then, by taking the difference between Revenues\_22 and the new (difference) column, values from last year’s revenue can be generated which can be named ‘Revenues\_21’.

Using the same approach, we can take the Profits\_22 and Profits\_Change columns and produce profits for last year and name it ‘Profits\_22’. The only problem here is that Profits\_22 has 3 null values whereas the Profits\_Change has 273. Even if we impute the null values with mean or using some other technique, the values produced for ‘Profits\_21’ will not be very reliable.

Another column can also be generated with the help of Revenues\_22 and Profits\_22 columns. As we know the profit is calculated by subtracting all the expenses from the revenue (AccountingTools, 2022). Based on this logic, if we took the difference of both Revenues\_22 and Profits\_22, the resulting amount would be the costs for the year and therefore it can be called ‘Costs\_22’.   
Now using the previously generated Revenues\_21 and Profits\_21 columns, a ‘Costs\_21’ column can be generated by simply taking their difference, and additionally a ‘Costs\_Change’ column can be produced by subtracting Costs\_22 from Costs\_21 and the result divided by ‘Costs\_21’.

Assets and Market value has been described previously and explained how assets are the book value which a company records as per its accounting rules, whereas the market value is what the company’s shares are traded for on the stock market multiplied by the total number of shares issued. If we take the difference of these two columns and make a new column ‘Asset\_Market\_Diff’, the values will either be positive or negative. If the market value of a company is more than its book value (assets) it indicates that the share price of the company is traded at a good price and the company is moving in the right direction. Whereas if the same value is lower than the assets, the company is under performing which is not a good sign. This also means that the shareholders have lost confidence in the company.

With these new additions, our total attributes would look like this:

Name, Revenues\_22, Revenues\_21, Rev\_Change, Costs\_22, Costs\_21, Costs\_Change, Profits\_22, Profits\_21, Profits\_Change, Assets, Market\_V, Employees

The correlation is calculated between the variables after the new additions, and the results show a strong positive correlation between some variables for example Revenue\_22 and Revenue\_21 is 0.9025, Costs\_22 and Revenues\_21 is also very strongly positive at 0.904.   
Similarly, the a high positive correlation is detected between Costs\_22 and Revenues\_22 at 0.9912.   
The reason there is a high correlation between these variables is because they are created with the help of one another, either by taking the difference or by adding.

The dataset after the modifications can produce a lot of new useful information, but it would go out of our scope of work. The market value and profit figures could be used to calculate the price earning ratio which is one of the best tools used to check the company’s stock price and earnings per share. This tells us if the company is undervalued or overvalued (Team, 2022)

Similarly, if Revenue is divided by the Assets, it would give us Asset turnover of the company which is the value of the company’s sales relative to the value of the assets. If we go on and compare Revenue with total number of employees, this will tell us how much money was made per employee of the company. This is very useful information but unfortunately it cannot be applied in answering the question we are trying to answer.

These newly added variables can contribute in many ways depending what we are looking for. In our case, these newly added columns turned out to be highly correlated and therefore are no longer included in our dataset. The current attributes excluding ‘Name’ are 7, and our Univariate and Bivariate analysis show that these attributes are good enough to train our models and predict the profit\_22 variable.

# Description of Applied Methodology

## Graph Showing Overall Methodology

## Brief Explanation of Each Step

**Abstract, Literature Review, & Data Description**

This is the first stage of the project. The first task was to select the theme, formulate problem questions and find supporting dataset to answer the questions. The theme chosen is Regression and the research question is to “Predict the Profit figure in the dataset using machine learning”. The dataset chosen for this project is “Fortune 1000 Companies in the US by Revenue” from Kaggle.   
The project kick starts with the project Abstract which details the background of the problem and then connects it with our research question. Then it goes on to explaining what it attempts to achieve followed by the strategy and the tools which will be used in the project.

Literature review was carried out to find and go through similar works that have been published in the past. This included looking for 10 research papers and writing a summary for all of them. The aim was to find the research papers which answered the same question as ours with using either the same or a similar dataset. The research papers found in the process were not exactly what this project aims to find but rather resembling ideas using same methodologies, programing language, models etc.

Data description and Statistics provided detailed information about the dataset before performing initial analysis. Each attribute is described so that the reader who is not familiar with the terms can easily understand the meanings. In addition, the dataset is explored further by looking into the math functions like min, max, mean etc

**Data Cleaning and Modification**

Data is cleaned by removing special characters, converting columns to same formats, changing column names followed by creating a detailed data dictionary. Furthermore, null values are imputed by mean and data is further examined for outliers and any other abnormalities to prepare it for further processing.

**Initial Analysis and EDA**

A detailed initial analysis was carried out on the data. This included univariate and bivariate analysis. During that, a portion of EDA was also performed, and the analysis answered some good questions about the data. Visualizations are used to better understand the attributes in both, Univariate, Bivariate analysis. Followed by analyzing each one of the attributes separately as well as in groups.   
Things are taken one step forward in our EDA. Where using existing attributes, more data is produced, and new questions are answered.

**Modelling, Evaluation, & Results**

There are four models that are used in this project and then evaluated to find out which one of them performed the best. The models include:

1. Multiple Linear Regression
2. Random Forest Regressor
3. Decision Tree Regressor
4. Support Vector Machine Regression

The measures used to evaluate the models are:

1. Mean Absolute Error (MAE)
2. Mean Square Error (MSE)
3. R- Squared Error
4. Root Mean Square Error (RMSE)
5. Cross Validation (5-fold)

The library used mainly for the modelling is sklearn with all its relevant packages to our project. The data is split 70% for training and 30% for testing in all four models. The results are displayed for predicted and actual values for all the models along with visualization and evaluation of the model with commentary. Based on our findings we will decide which model has performed the best by comparing the evaluation scores.

**Conclusions**

This is where a short summary is given of our analysis and the discussion is concluded by defining the usefulness of our analysis

**Shortcomings**

This section will discuss all the shortcomings of our project. Things that were not achieved, the obstacles, and how things could have done differently.

# Modelling, Evaluation, & Results

At this stage appropriate machine learning models are chosen to answer our research question. Given the nature of our problem, it is classified as a regression problem that can be solved using supervised machine learning models.   
Choosing the right models is challenging. But after doing extensive research only four machine learning models have been finalized.

1. Multiple Linear Regression
2. Random Forest Regressor
3. Decision Tree Regressor
4. Support Vector Machine

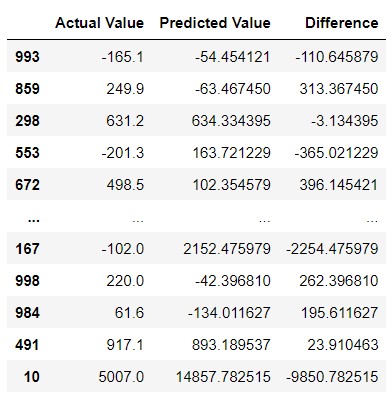
**Multiple Linear Regression**

The basis of this model is to assess and predict a dependent variable with the help of other independent variables. The dependent or the target variable is chosen based upon the problem question.   
To apply multiple linear regression the first thing we will have to do is to create two variables X and y. These two variables contain the following attributes:

Variable X = Revenues\_22, Rev\_Change, Profits\_Change, Assets, Market\_V, Employees

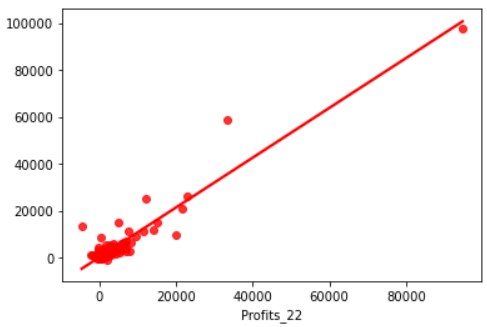
Variable y = Profits\_22.

The dataset is then split into training set and test set. It is common practise to split the data 80% for training and 20% for testing. Since there are no hard and fast rules related to this, we will split our data 70% for training and 30% for testing. This is done by the Python library sklearn.model\_selection and the package used is train\_test\_split.   
Using the same library and with a help of the package LinearRegression the model is trained using the X and y variables. The model is then tested on the remaining 30% of test data and results of the prediction are displayed along with actual values.



As it can be seen in the table above, the model has predicted the values with certain amount of accuracy. If the ‘Difference’ column is observed, which is the difference between the actual and predicted values, it shows that the differences are not too big. Most values in the ‘Difference’ column are three-digit values and the target variable ‘Profits\_22’ has a mean of four digits. It can be said that the prediction is not accurate but its not too bad either.

It is useful to understand the linear relationship between two of our parameters by plotting a scatterplot of these data points. A regression plot is used below to understand our results.



It can be seen in the plot that our data points form a clutter at the bottom left side of the plot with a straight line bifurcating the data points. These data points include both, our actual values as well as predicted values. This visual representation of our model result shows good results but we cannot form an opinion without evaluating the model. For this, we will test the accuracy of our model along with cross validation which will be compared with the results of other models to see which one of them performed the best.

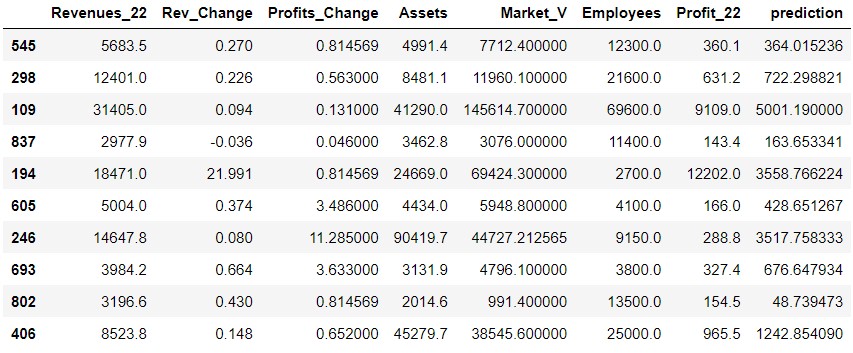
The accuracy of the model is calculated by r2\_score package from sklearn.metrics library. The code of this can be found in the relevant Python file. The cross validation with 5 folds is also used to evaluate the model. The results are as follow:

Accuracy = 86.06  
Cross Validation = 0.00898

**Random Forrest Regressor:**

This is the second model chosen for our problem mainly because of its accuracy and popularity of being able to perform both regression as well as classification tasks. Random forest is basically an ensemble of decision trees, meaning many trees, produced in a certain random way that it forms a random forest.

Firstly, the data needs to be split between training and test set just like any other model. For this purpose we will use sklean.model\_selection library and import the train\_test\_split function, and from sklearn.ensemble package use the RandomForestRegressor function. The data is split 70% for training and 30% for testing. The splitting of the data is the same for other models as well so to have an apples-to-apples comparison. The model is trained and then tested to show the predicted outcome in the following table:



At first glance the actual and the predicted values are much closer than multiple linear regression. Visualizing the results may give a better insight to our prediction.

Chart, scatter chart

Description automatically generatedChart

Description automatically generated

Both of these plots are for the random forest. It can be seen that most of our values lie at the bottom left side of the plot. That is because the actual data for Proftis\_22 lie within that range and our predicted values are also in the same range. There are of course some exceptions as we see some values further away from the rest. These profit values belong to companies with significantly higher profits.

To evaluate the model following tests are conducted. Their relevant coding can be found in the source code Python file:

|  |  |
| --- | --- |
| Mean Absolute Error | 1278.83 |
| Mean Squared Error | 22257402.58 |
| R-squared scores | 0.65 |
| Root Mean Squared Error | 4717.775 |
| Cross Validation (5-fold) | 0.056 |

Our Mean Absolute Error (MAE) is 1278.83. This means the magnitude of difference between the predicted observation and the true value of that observation is 1278, which is considered high. The Mean Squared Error (MSE) looks even higher with a value of 22257402. But the thing with MSE is that there is no correct value for it. Usually the lower value is considered better but since there is no exact measure, it all comes down to comparing the MSE with other models to select the one with lower value.   
The R-Saquared score of 0.65 is generally seen as a high level of correlation between the predicted and actual values. Therefore, the higher R-Squared the better it is.   
The Root Mean Squared Error is very high at 4717 for a dataset which has 1000 observations. This can be lowered down by going back to feature selection and make some changes there.  
Then lastly the cross validation with 5-fold has a value of 0.056. This value will also be compared with other models to see which has a higher value.

**Decision Tree Regressor**

Decision tree regressor is used as the third model for the project. This is also a common and widely used model for regression as well as classification problems. The way the decision tree works with continuous variable is that they work by principle of reduction variance, meaning it looks at different splits and calculates total weighted variance of every split only to choose the one with the lowest variance.

Using the same sklearn library with relevant packages, the data is split 70% for training and 30% for testing. The model is trained and then tested with the following results:

Table

Description automatically generated

Chart, scatter chart

Description automatically generated

Comparing the Profit\_22 and the prediction column we get the idea that the values have not been predicted to what the actual Profit\_22 values are or anything close.

The Root Mean Squared (RMSE) is calculated at 5979.816 the code to which can be found in thePython file. This value is very high, higher than RMSE of Random Forest Regressor. The 5-fold cross validation is in negative at -0.7182. These two values show that the decision tree regressor has not produced anything noteworthy.

**Support Vector Machine Regressor**

Support Vector Machine for regression is primarily used to predict continuous numerical variables. This model is also very popular for predicting of regression as well as classification problems. Usually, where there is a comparison between machine learning regression models, SVM is almost always found in that list.

Same strategy is used in splitting the data 70% for training and 30% for testing. The library used to train and model the data is sklearn, like what is used for other models. Once the model is trained, it is tested on the test set to produce the predicted values.

Table

Description automatically generated

The predicted and the actual columns show quite a bit of difference. We can analyze our predicted outcome with the help of evaluation measures to understand our results properly. In addition to this, a scatterplot of our predicted results along with actual values is plotted with straight line showing the performance of our model.

Chart, scatter chart

Description automatically generated

Looking at the scatterplot, it indicates that the predicted outcome is not what we expected. Failure of this model to predict the values with good accuracy needs to be investigated further to understand more. To evaluate the model, some of common evaluation measure are used which can be found below in the table. The coding can be found in the relevant Python source code file.

|  |  |
| --- | --- |
| Support Vector Regression Accuracy | -0.0482 |
| R2 square | -0.0482 |
| MAE | 2026.357 |
| MSE | 40409663.130 |
| Cross Validation (5-fold) | -0.3311 |

The accuracy and the R2 Square are both identical negative values. This shows that the SVM regression model fitted the data extremely bad. The other measures MAE, MSE, and Cross validation all indicate that this model is a failure.

# Conclusions

After using four machine learning models for regression and evaluating their results we can decide which model has performed the best. Out of the four models, the best performing model is the Random Forest Regressor with highest cross validation of 0.056 followed by Linear regression on second with cross validation 0.008.

# Shortcomings of the work and concluding remarks

The first mistake made was not choosing the right research question or choosing the wrong dataset to answer the question. The research question was, "Predict profit 2023 for the global 1000 companies and rank them as per highest profit earned". The question clearly asks for the future profit or the next year’s profit. The main idea behind was that the profit is a better measure then revenue, and therefore companies should be ranked according to the most profit made for the year. As anyone would understand that the Revenue is money earned; and profit is the money saved.

So, to answer this question extensive research was carried out to find similar works but no material was found on the subject matter. Out of all the materials found, 10 research papers were shortlisted where there was some sort of prediction involved. Still, nothing was found anything on exactly what was searched. Except something that was revealed to me later that most of the search results were projects on time series analysis. Even though the time-series was explored but it never clicked that this problem question is purely a time series problem, and it can easily be solved given the right dataset. For example, if we were to predict the profit of these 1000 companies and we had the dataset which contained profits for the last 10 years then we would've been in business! But as the data we have is profit for the year, even though I was able to generate the profit figures of last year with the help of "Profit\_Percentage\_Change" column. But I still felt that it cannot get the job done.

At that point I just went with the flow and wanted to apply the regression models to get somewhere at least. Therefore, I applied four regression models, Multiple Linear regression, Random Forest Regressor, Decision Tree Regressor, and Support Vector Machine regressor. These models were applied properly with a 70/30 split, generated the predicted values, evaluated the models with RMSE, MSE, R Squared scores, and with Cross validation (5fold) with Random Forest regressor outperforming the other three.

Even though I predicted the values, even compared the actual and the predicted but of course they were for present year. But it still did not answer my question, in fact I was no way near it. I could've predicted the profit for next year with the help of time series and appropriate dataset, but I was late for it. Therefore, this project was completed by answering a different question altogether.

# GitHub Repository Link

<https://github.com/hasnain-aly/CIND-820-Project-Repository/blob/main/Project%20Roughwork-checkpoint.ipynb>

# Reference Page

AccountingTools. (2022, May 14). *What is the formula for calculating profit?* Retrieved from Accounting Tools: https://www.accountingtools.com/articles/profit-formula.html

Amel-Zadeh, A. (2020). Machine Learning-Based Financial Statement Analysis. *University of Oxford - Said Business School*, 69.

Beaver, S. (2022, August 22). *What Is Working Capital? How to Calculate and Why It’s Important*. Retrieved from Net Suite: https://www.netsuite.com/portal/resource/articles/financial-management/working-capital.shtml

Choi, W. (2019). Predicting the U.S. Airline Operating Profitability using Machine Learning Algorithms. *International Journal of Aviation, Aeronautics, and Aerospace*, 8.

dowu, S. (2016, December 16). *Applied machine learning: Forecasting heat load in district heating system.* Retrieved from ScienceDirect: https://www.sciencedirect.com/science/article/abs/pii/S0378778816310155

HARPER, D. R. (2022, July 22). *Forces That Move Stock Prices.* Retrieved from Investopedia: https://www.investopedia.com/articles/basics/04/100804.asp

*IEEE XPLORE.* (n.d.). Retrieved from https://ieeexplore.ieee.org/: https://ieeexplore.ieee.org/abstract/document/9407627

J Manasa, R. G. (5 March, 2020). Machine Learning based Predicting House Prices using Regression Techniques. *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)* (p. 14). Bangalore, India: IEEE. Retrieved from IEEE Xplore.

MCCLURE, B. (2021, December 29). *How to Use ROA to Judge a Company's Financial Performance.* Retrieved from Investopedia: https://www.investopedia.com/articles/fundamental/04/012804.asp#:~:text=Return%20on%20assets%20(ROA)%20is,its%20assets%20to%20generate%20profits.

Morid, M. A. (2018). Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. *National Library of Medicine*, 18.

NA. (NA, NA NA). *Working Capital*. Retrieved from bdc.ca: https://www.bdc.ca/en/articles-tools/entrepreneur-toolkit/templates-business-guides/glossary/working-capital

Nitis Monburinon, P. C. (2018, May 18). *Prediction of prices for used car by using regression models.* Retrieved from IEEE Xplore: https://ieeexplore.ieee.org/abstract/document/8391177

PEREZ, C. (2017). Forecasting Firm Performance with Machine Learning: Evidence from Japanese firm-level data. *RIETI Discussion Paper Series 17-E-068*, 28.

Pudaruth, S. (2014). Predicting the Price of Used Cars using Machine Learning Techniques. *International Journal of Information & Computation Technology*, 12.

Sakhare, N. (2019). Performance analysis of regression based machine learning techniques for prediction of stock market movement. *International Journal of Recent Technology and Engineering*, 8.

Sami, I. u. (2018). Predicting Future Gold Rates using Machine Learning Approach. *International Journal of Advanced Computer Science and Applications*, 8.

Sarath. (2022, June 03). *FORTUNE 500 COMPANIES IN THE US*. Retrieved from eqvista: https://eqvista.com/fortune-500-companies-in-the-us/#:~:text=The%20Fortune%20500%20is%20more%20widely%20used%20than,yearly%20revenues%20during%20the%20fiscal%20years%20in%20question.

Shourvarzi, M. R. (n.d.).

Shourvarzi, M. R. (2011). Analysis of the predictive ability of the components of economic value added in predicting next period’s operating profit: Evidence from Iran. *African Journal of Business Management*, 8.

Swami, D. (2020). *Predicting Future Sales of Retail Products using.* California: University of Southern California.

Team, C. (2022, Nov 26). *Price Earning Ratio*. Retrieved from CFI: https://corporatefinanceinstitute.com/resources/valuation/price-earnings-ratio/

Xi Chen, T. (. (2022). Predicting Future Earnings Changes Using Machine Learning and Detailed Financial Data. *Journal of Accounting Research, Forthcoming*, 91.