FINAL PROJECT

Course – Business Analytics

BUSI 650 (Spring’23 Campus – 41)

Professor – Sana Ramzan

Submitted by:

Trishla Tiwari (2214436)

Taranjeet Dhanjal (2223021)

Ali Hyder(2229596)

Dharmik Joshi (2219007)

Vaneet Bansal (2231679)

Mansi Sharma (2105793)

Date Due : June 17, 2023

Table of Content

Introduction

In today’s competitive business environment, companies need to make data-driven decisions to remain relevant and successful. Business analytics provides the tools and techniques to extract meaningful insights from data to support decision-making. This project aims to conduct an analysis of companies' data and different ratios to identify patterns and trends and provide actionable recommendations to improve sales performance.

The project discusses the data set of the financial standing of companies in an industry. The data set includes financial statistics such as financial ratios, growth rate, m score, EBITDA, gross profit and net profit margin, return on assets and more. The dependent variable selected is EBIDTA, earnings before interests, depreciation, tax and amortization. The independent variables are Enterprise Value, Market Capitalization, Revenue Growth, Net Income and Debt Growth. The scope of the project is to understand the variability of the EBIDTA and to understand what factors in what context affect it. for an organization to make decisions for the future, analyzing past data is crucial, especially in the state and to understand where the company is standing financially and to determine what affects its financial performance. It is crucial to understand the current status and predict the future according to current data and plan for the future. The main topic of the analysis is,

Does the financial performance affect the EBITDA of a company

Further we would also be answering the following questions under this report,

- What is the relationship between the enterprise value and EBIDTA?
- What is the relationship between market capitalization and EBIDTA?
- What is the relationship between revenue growth and EBIDTA?
- What is the relationship between the net income and EBIDTA?
- What is the relationship between debt growth and EBIDTA?

Methodology

In order to analyze the data and come to an understanding of the research's main question and sub-questions a few methods and analytic tools are used throughout the project. First, to understand the characteristics of the data set, to understand the nature of the EBITDA of the different companies in the industries the descriptive analysis will be used to find mean, median, mode and regression analysis. Regression analysis will be conducted on the five independent variables and EBIDTA to find the relationship between the five independent variables and the EBIDTA. For ease of calculations and understanding of the data, the data has been scrubbed and all null values have been removed. All data with outliers have been removed with respect to EBITDA and Enterprise Value using the interquartile range.

Secondly, to understand the patterns and nature of the different companies' financial standing based on various factors visualization will be used with an excel dashboard and Tableau to obtain a visual understanding of the companies.

Lastly, machine learning will be used for predictive analysis to predict how the financial performance of each company will be based on the given historic data.

In managing, business forecasting is a critical point as on the forecasting the budget, product manufacturing, sales and marketing and all the operations will be based. Also, in return, it gives a basic idea of how to lead the business according to the forecasting, as it provides a view for the future to plan for the financial status.

Different techniques used for analysis

Various techniques and methods are used for analyzing the data and coming up with suitable results. Some of the techniques used are as follows.

1. *Descriptive analysis* - These provide descriptive statistics which give us a summary of the data and help to identify patterns and trends. This technique can be used to calculate measures such as mean, median, mode, range, Standard deviation and correlation.
2. *Regression analysis* - Regression analysis is a statistical technique used to find the relationship between one or more independent variables and a dependent variable. It can be used to identify the factors that impact sales performance and predict future sales.
3. *Visualisation* - Visualisation techniques such as graphs, and charts can be used to present the data in a clear and understandable way. This can help everyone to identify the patterns and the trends to make informed decisions. In visualization various ways are there to visualize data–
    - Excel Dashboard
    - Tableau Dashboard

These techniques are used in business analytics to help companies make data-driven decisions and improve performance. Each technique has its own strengths and weaknesses and the choice of technique will depend on the specific research questions and data available for analysis.

4. *Predictive Analysis based on Machine Learning*- The main purpose of this section is to use machine learning techniques to make predictions. The focus here is on analyzing the predictions/findings from using the various models. The scope of this section is limited to the dataset given.

Furthermore, with the help of SMOTE code, Mscore, and Fraud/Non-Fraud variables have also been used to predict and analyze the companies' classification as fraud/non-fraud. Using google collab backed by python language, the predictions have been made with this model. Along with the SMOTE codes, the RMSE model, and the linear/logistic regression model have also been used.

Limitations of the report:

The report is subjected to the limitations of the data set provided and also to the writers being novices and a few external constraints which are beyond the writer's control. This report may also have observer bias as it is based purely on the understanding of the writers and their interpretation of the data.

Descriptive Analysis Report

A descriptive report is a document that offers a thorough overview of a certain topic or issue with the objective to explain and present information in a clear and succinct manner, utilizing tables, graphs, or other types of visual aids to enable the report readers to grasp the information being provided. Here, for the purpose of this report, we shall focus on the data of EBITDA and its measures of central tendency and measures of dispersion.

Measures of Central Tendency

A central tendency measure is a statistical metric that determines the position of a distribution's or data set's centre. It is used to define the middle or centre value of a data collection. There are three common forms of central tendency measurements:

- *Mean:* The mean is the mathematical average of a set of data. It is calculated by adding up all of the values in the data set and dividing the total by the number of values. The mean gives us a sense of what the centre value in the data set may be. It is significant because it considers every single observation in the data set to help locate the centre point. In our data set, the mean is 388,207,231.7. This is the average of the earnings of all companies before subtracting the interest, taxes, depreciation and amortization.
- *Median:* The median is the value central value of the dataset. It is computed by sorting the data from least to greatest and then choosing the intermediate value. If the number of values is even, the median is the average of the two middle values. Here, the median is 211,556,000. There exists a significant difference between the median and mean, which means that even though the outliers have been removed from the dataset, there still exist some that heavily influence the mean.
- *Mode:* The mode is the most often occurring value in a set of data, i.e., it is the most common value. For the EBITDA, the mode is 114,044,000. This means that the EBITDA of most companies in the data set is 114,044,000.

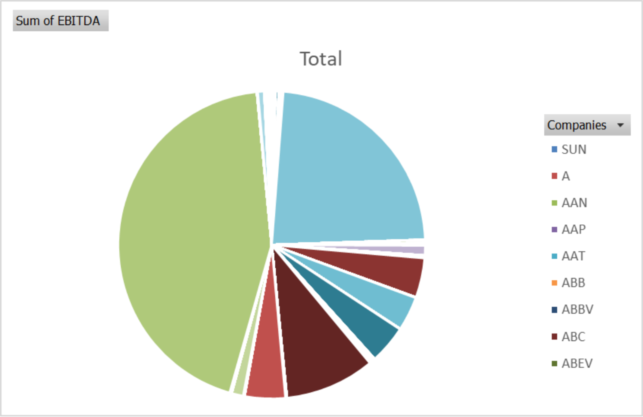
From the above data, we can also conclude that the data is positively skewed. It is evident from the values of mean, median and mode as, Mean>Median>Mode. This indicates positive skewness.

Measures of Dispersion

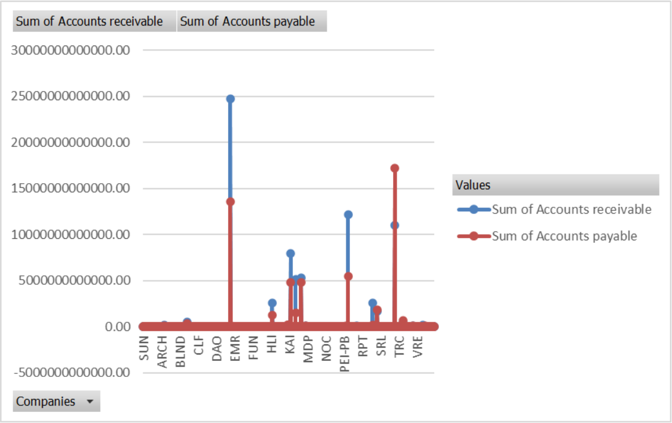
Measures of dispersion, also known as measures of variability, are statistical metrics that characterize how spread out or scattered a collection of data is. They give information on the distribution of the data and can aid in understanding the degree to which the data is clustered or spread out.

- *Range:* the most basic measure of dispersion is the range. It is the difference between the highest and lowest value in a dataset. The range of our data is $4,772,791,000. This gives us an idea of how vastly spread the earnings of the different companies are.
- *Variance:* The variance is a measure of how much the values in a data collection depart from the mean. It is determined by summing the squared deviations from the mean and dividing by the number of values in the data set. The variance of our data set is 347,240,188,835,514,000. Since it is expressed in square units of the original data, the variance looks like an extreme value.
- *Standard Deviation:* The standard deviation is defined as the square root of the variance. It computes the average amount that values in a data set deviate from the mean. The standard deviation here is equal to $589,270,896. This suggests that the data points deviate by $589,270,896 from the mean of the data.
- *Interquartile Range:* The difference between the third quartile (the value below which 75% of the data falls) and the first quartile (the value below which 25% of the data falls) is the interquartile range (IQR). It measures the distribution of the middle 50% of the data. In our report, we have used an interquartile range of EBITDA and Enterprise value to clean our data set to make it more comprehensible.

Excel Dashboard



The pie chart above shows a thorough picture of companies based on their respective EBITDA values. AAN has the highest share of over 1577 companies of EBITDA with 44% total share of 70625000000000 and HMC, SONY and TSM have the least share of 1% each. This could be a useful tool for investors, financial analysts, or anyone interested in comparing the financial performance of these companies.

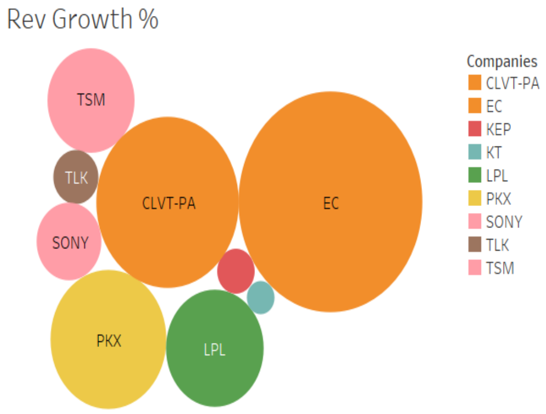


The line graph compares the top 10 companies based on their accounts payable and receivables. EC has the highest sum of accounts receivable and TLK has the highest accounts payable. This could be a useful tool for investors, financial analysts, or anyone who is interested to gain insights into the financial health of these companies.

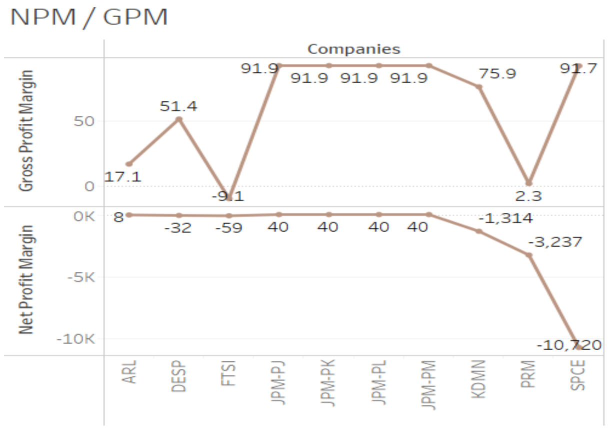
<u>Tableau Dashboard Report</u>

Data visualization: Individuals may develop interactive and visually appealing dashboards, charts, and graphs using the well-liked data visualization application Tableau. It offers a variety of personalization choices for visualization and makes it simple to transfer the insights to others.

Data analysis: Tableau includes strong data analysis features, such as data filtering, sorting, and grouping. Users are offered the ability to edit the data in an array of ways in order to obtain conclusions about the root cause of trends and patterns.



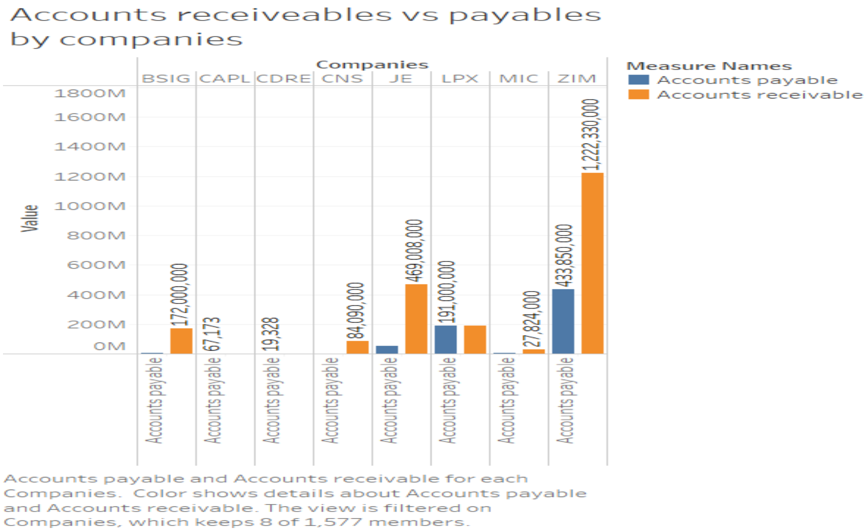
Companies. Color shows details about Companies. Size shows sum of Rev Growth. The marks are labeled by Companies. The view is filtered on Companies and sum of Rev Growth. The Companies filter keeps 10 of 1,577 members. The sum of Rev Growth filter includes greater than and or equal to 0.0000 and keeps Null values.



The trends of sum of Gross Profit Margin and sum of Net Profit Margin for Companies. The data is filtered on sum of Ebitda, which keeps non-Null values only. The view is filtered on Companies, which keeps 10 of 1,577 members.

Above graphs are made in Tableau for interpretation for data analysis. These are the top 10 companies in comparison with the REV Growth %. This has the highest revenue growth out of 1577 companies. EC, CLVT-PA. TSM and SONY are the topmost companies with 0.8 0.4, 0.3 and 0.1 respectively.

In the other Line Graph, we can see companies with high gross profit but very low declining Net profit margins. These are the analysis of the top 10 companies with high gross margins.


Accounts receiveables vs payables by companies

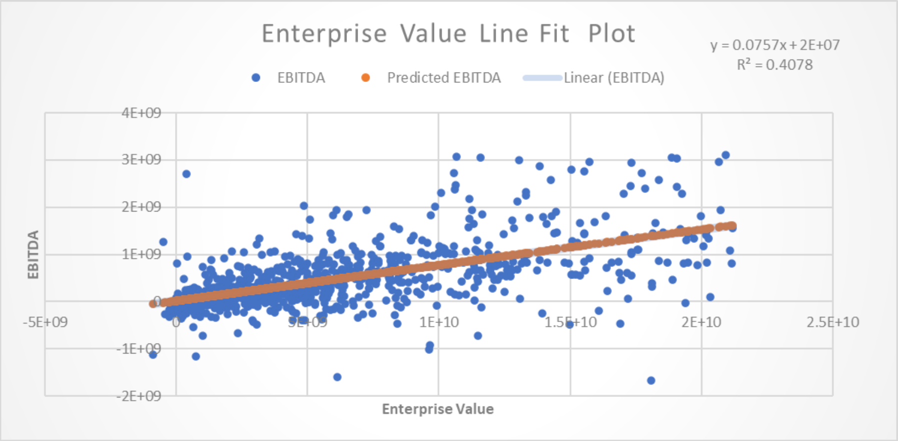
Accounts payable and Accounts receivable for each Companies. Color shows details about Accounts payable and Accounts receivable. The view is filtered on Companies, which keeps 8 of 1,577 members.

In the above graph, we can see the top 10 companies in the terms of Accounts Receivables. Account Receivables are shown with a comparison of accounts payable. The table shows how much the company is entitled to receive and payout the cash. This shows the cash flow of the company.

Tableau is a popular tool for data analysis and visualization in the business intelligence field. It provides clients access to real-time data and lets them create different types of reports to support decision-making.

Linear Regression Reports

EBITDA over Enterprise value regression model:

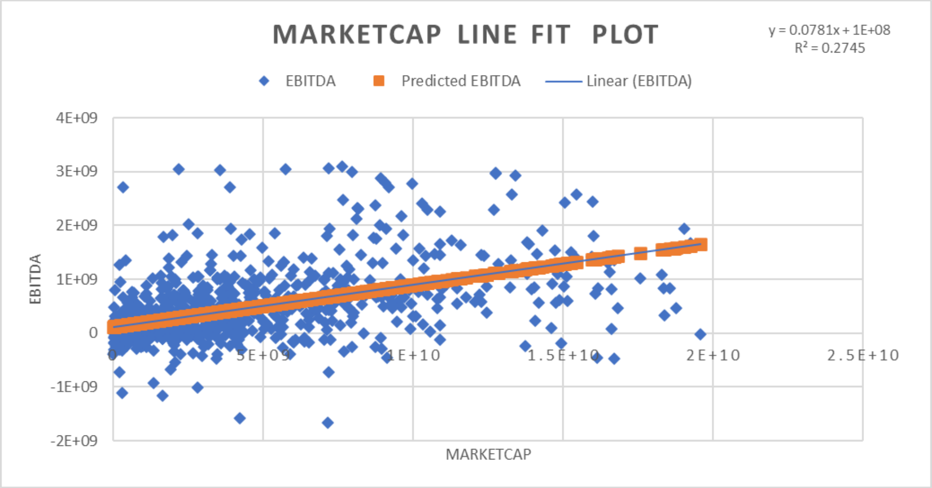


First we shall look at the correlation coefficient (r) to determine the relationship between the enterprise value and the EBITDA. The multiple r is equal to 0.64, which indicates a moderately strong linear correlation between the two variables, thus also indicating that the two are directly related to each other. By understanding this relationship, we can therefore anticipate the value of the second variable (y variable, EBITDA) on the basis of the first variable (x variable, enterprise value). Their relationship is mathematically expressed as:

$$y = 0.0757x + 2E{+}07, \text{ where } r^2 = 0.4078$$

The R-square value of 0.4078 evaluates the strength of the relationship between the variables, thus indicating that the model accounts for 40.78% of variability seen in the target variable. The constant 2E+07, is the value of the dependent variable, EBITDA, when the x variable is zero. The value of the constant is 17,179,240.93 and is written as 2E+07 for ease of interpretation mathematically. The regression coefficient of y on x is the coefficient of the independent variable. It reflects a change in the amount of the dependent variable, that corresponds to a change in the amount of the independent variable by one unit. This means that the EBITDA would increase by 0.0757 if the enterprise value increased by 1 unit.

EBITDA over Market Capitalization regression model:

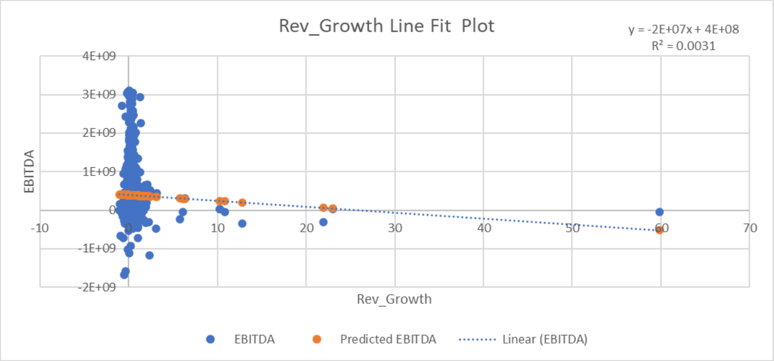


The correlation coefficient r is equal to 0.524. Similar to the correlation coefficient r of the previous model, the multiple R of this model also has a moderately positive relationship between the two variables and showcases a positive relationship between the two variables. It is mathematically represented as,

$$y = 0.0781x + 1E{+}08, \text{ where } r^2 = 0.2745$$

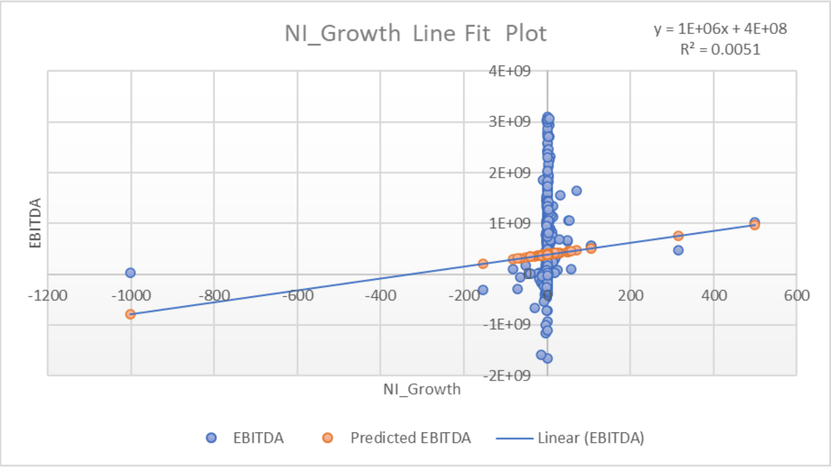
The R-squared of 0.27 indicates that a trend can be seen even in noisy, highly variable data. Thus, even though the data points deviate from the regression line, the trendline indicates that the independent variable can still be used to get insights into the dependent variable. The coefficient of x is equal to 0.0781, this proves that there is a positive relationship between the two variables showing that as the independent variable increases, the dependent variable also increases. The constant can be defined as the mean of the dependent variable when all independent values are 0. The constant in our case is 117,905,503.1 (written as 1E+08). This means that if the market capitalization was 0, the EBITDA would be 117,905,503.1.

EBITDA over Revenue Growth regression model:



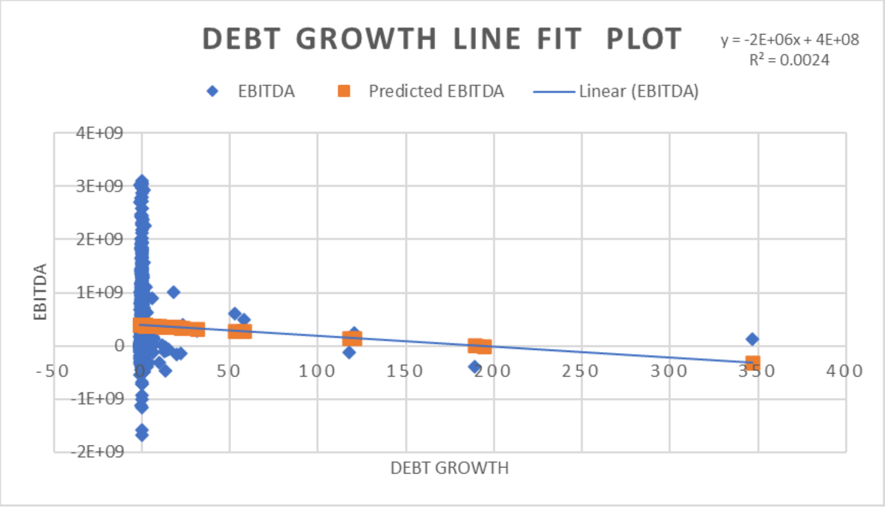
A correlation coefficient of 0.056 indicates that there is barely any correlation between the two variables and making predictions from the independent variable for the dependent variable is very difficult and would result in unreliable predictions. An R-squared of 0.0031 means the strength of the relationship between the variables accounts for a meagre 0.31% of variability seen in the target variable. This can thus be interpreted as there not being any significant relationship between the two. The coefficient of x of -15,204,291.25 suggests that there is a negative relationship between the two variables, meaning, a rise in the revenue growth would result in a fall in the EBITDA. The constant or intercept of 394136111.3 indicates that when there is no revenue growth, i.e., it is zero, the EBITDA would be 394136111.3.

EBITDA Over Net Income Growth Regression Model:



Like the revenue growth regression model, the correlation coefficient of the Net Income Growth model also has an r less than 0.1, i.e., 0.072. This is an indicator of a negligible correlation between EBITDA and net income growth. Furthermore, the R-squared of 0.005, also confirms this. R-squared less than 0.05 suggests that the model does not explain much of the data variance, yet it is substantial as compared to not having a model. When the net income is zero, the EBITDA is 387,621,715.3. The coefficient of x is 1,174,592.383, which suggests that there is a positive relationship between the two variables. A rise in Net income growth would result in a rise in EBITDA.

EBITDA Over Debt Growth Regression Model:



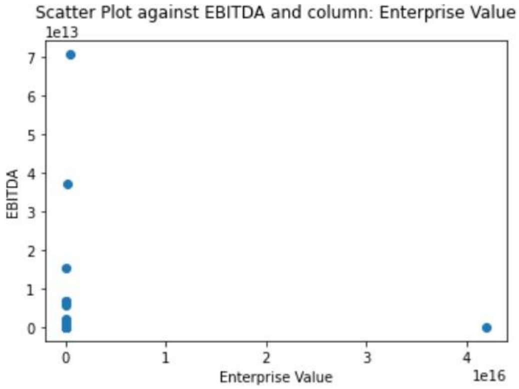
Lastly, we shall look at the EBITDA in relation to debt growth. Much like the previous two models. There is negligible relation between the two variables as evident from the correlation coefficient of 0.049 and an R-squared of 0.0024. The negative coefficient of x, i.e., -2012997.295 is an indicator of a negative relationship between the x and y variables. However, when the x variable is 0, there would be a minimum of 390,769,356.3 in terms of earnings before interest, tax, depreciation and amortization.

Predictive Analysis

Before diving deeper into the analysis and prediction, we cleaned up the data by removing outliers from the dataset in addition to removing n/a values from the columns.

1. Enterprise Value (X) and EBITDA (Y)

Visualization | Analysis | Training | Predictions:

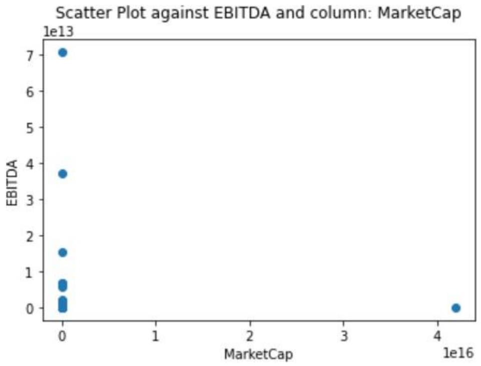


Here, the regression model is used to determine a relationship between enterprise value and EBITDA. The score of this model is 0.39 which indicates that the model explains approximately 39.28% of the variation in the dependent variable. The coefficient for this analysis is 0.089 which means that for every one-unit increase in the independent variable, the dependent variable is expected to increase by 0.089 units. Furthermore, the mean squared error was 5.819 which is a very big number and indicates that the model may not be very accurate in predicting the dependent variable. Moreover, from evaluating the R^2, we get a value of 0.392 which indicates a not-so-good relationship between the two variables taken and the accuracy is low.

Before making predictions based on hypothetical values, the model was trained using 0.25 as the test size. On defining a random enterprise value, the code generated a predicted EBITDA. To further substantiate the accuracy of the relationship, the RMSE (root mean squared) or r^2 was used, whose value was 94,308,935,286,976.14 which indicates that the model is not very accurate in predicting EBITDA and the relationship is also not so good. By using the logistic regression model it was proved that the model has a low accuracy of 0.1645.

2. Market Cap (X) and EBITDA (Y)

Visualization | Analysis | Training | Predictions:

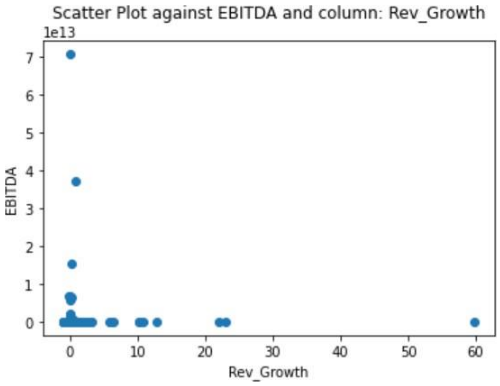


Here, the regression model is used to determine a relationship between the market capitalization of a company and EBITDA. The score of this model is -0.0032 which indicates that the model cannot explain the variation between the independent and dependent variables. The coefficient for this analysis is 0.205 which means that for every one-unit increase in the independent variable, the dependent variable is expected to change by 0.205 units. The mean squared error of 9.614 indicates that the model may not be very accurate in predicting the dependent variable. An ◢^{2} of -0.0032 indicates no or negative relationship between the two variables taken.

This model was also trained in a similar way to the previous one, and to understand the accuracy of the relationship, RMSE (root mean squared) or ◢^{2} is used, whose value was 94,326,586,938,646.86 indicating that the model is neither accurate in predicting EBITDA nor does it establishes any relationship. This has been proved by the low accuracy of the logistic regression model which is 0.1645.

3. Rev_Growth (X) and EBITDA (Y)

Visualization | Analysis | Training | Predictions:

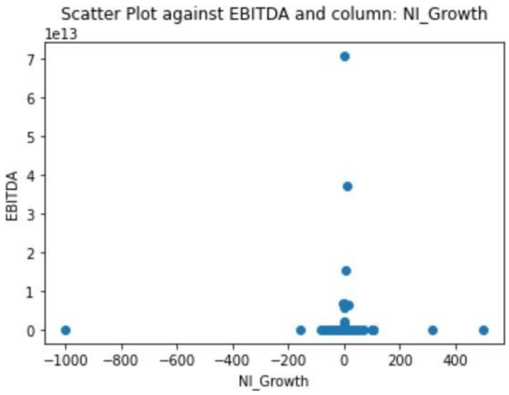


Here, we have evaluated a regression model to determine a relationship between the revenue growth of a company and EBITDA. The score of this model is -0.0032 which indicates that the model cannot explain the variation between the independent and dependent variables. The coefficient for this analysis is -1.732 which means that for every one-unit increase in the independent variable, the dependent variable is expected to change by -1.732 units. We also calculated the mean squared error which was 9.6104 which is a very big number and indicates that the model may not be very accurate in predicting the dependent variable. Moreover, we used one of the evaluation metrics of ^{2}and got a value of -0.00332 which indicates that the model does not explain much of the variation in the dependent variable.

Before making predictions based on hypothetical values, we trained the model using 0.25 as the test size. On defining a random revenue growth value, we got a predicted EBITDA. To further substantiate the accuracy of the relationship, we used RMSE (root mean squared) or ^{2} whose value was 2,074,091,336,610.3972 which indicates that the model is neither accurate in predicting EBITDA nor does it establishes any relationship. But on the flip side when we further used the logistic regression model I got an accuracy of 0.8354 which is very high and negates my earlier model's relationships.

4. NI_Growth (X) and EBITDA (Y)

Visualization | Analysis | Training | Predictions:

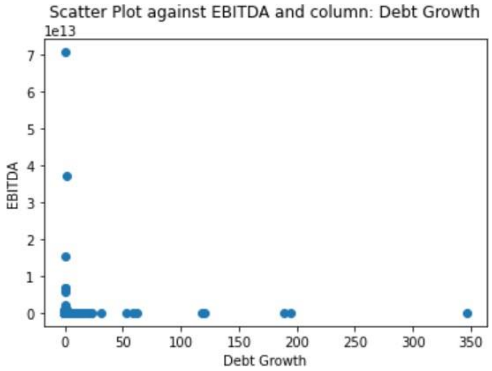


Here, we have evaluated a regression model to determine a relationship between the net income growth of a company and EBITDA. The score of this model is -0.0032 which indicates that the model cannot explain the variation between the independent and dependent variables. The coefficient for this analysis is 95,312,588.151 which means a very strong relationship between the two variables considered hereunder. We also calculated the mean squared error which was 9.6080 which is a very big number and indicates that the model may not be very accurate in predicting the dependent variable. Moreover, we used one of the evaluation metrics of ^{2}and got a value of -0.00332 which indicates that the model does not explain much of the variation in the dependent variable.

Before making predictions based on hypothetical values, we trained the model using 0.25 as the test size. On defining a random net income growth value, we got a predicted EBITDA. To further substantiate the accuracy of the relationship, we used RMSE (root mean squared) or ^{2} whose value was 2,074,091,336,610.384 which indicates that the model is neither accurate in predicting EBITDA nor does it establishes any relationship. But on the flip side when we further used the logistic regression model we got an accuracy of 0.8354 which is very high and negates the earlier model's relationships.

5. Debt Growth (X) and EBITDA (Y)

Visualization | Analysis | Training | Predictions:



Here, the evaluation of the regression model is to determine a relationship between the debt growth of a company and EBITDA. The score of this model is -0.0029 which indicates that the model cannot explain the variation between the independent and dependent variables. The coefficient for this analysis is -2.169 which means a very poor relationship between the two variables considered hereunder. The calculated mean squared error was 9.593 which is a very big number and indicates that the model may not be very accurate in predicting the dependent variable. Moreover, we used one of the evaluation metrics of ^{2} and got a value of -0.0029 which indicates that the model does not explain much of the variation in the dependent variable.

Before making predictions based on hypothetical values, we trained the model using 0.25 as the test size. On defining a random debt growth value, we got a predicted EBITDA. To further substantiate the accuracy of the relationship, we used RMSE (root mean squared) or ^{2} whose value was 2074091336610.3972 which indicates that the model is neither accurate in predicting EBITDA nor does it establishes any relationship. But on the flip side when we further used the logistic regression model we got an accuracy of 0.8322 which is very high and negates my earlier model's relationships.

6. SMOTE [Using EBITDA and Rating]

We used the SMOTE (Synthetic Minority Oversampling Technique) code to balance the dataset. To use SMOTE, we had to choose one categorical variable, i.e. Rating when compared with EBITDA. We split the dataset into a training size of 0.8 and a test size of 0.2, after which we trained the model using linear regression and checked the accuracy of the logistic regression model, which was 0.46. This accuracy is not good for predictions.

7. Fraud or Not Fraud Prediction [Using MScore values]

It was very intriguing to see a fraud/non-fraud column in the dataset, and hence we tested it using Mscore to predict which companies are fraud/non-fraud. Plugging in random values, the predicted score was 0.9525.

Findings and Conclusion

Table: 1

Description	R^2		R = Correlation Coefficient	
	Linear Regression	Predictive Analysis	Linear Regression	Predictive Analysis
EBITDA Vs. Enterprise Value	0.410	0.390	0.640	0.080
EBITDA Vs. Market Cap	0.270	-0.003	0.520	0.200
EBITDA Vs. Revenue Growth	0.003	-0.003	0.060	-1.730
EBITDA Vs. NI Growth	0.005	-0.003	0.070	9,53,12,588.151
EBITDA Vs. Debt Growth	0.002	-0.002	0.050	-2.169

On analyzing the companies based on the various aforementioned methodologies, and in conjunction with Table 1, we had a few takeaways. From the pivot and tableau analysis, we understood that companies like EC and CLVT-PA. TSM and SONY are the topmost companies, with 0.8, 0.4, 0.3 and 0.1, respectively, with the highest revenue growth.

Additionally, ZIM is the company with the highest account receivable and also has the highest difference between receivables and payables. Although this data could be useful from the financial aspect, it still has errors/null values, which reduce the accuracy of giving us a futuristic view based on historical data.

We then used the linear regression model with the help of excel and the predictive analysis using machine learning. These two methods gave us a lot of insights, but there were discrepancies found in the outcomes from either of the analysis. While using the excel tool, we had to clean a lot of data and get rid of the outliers. This was a challenging task using IQR (interquartile range), which left us with limited data to act upon. This method was time-consuming and tiresome.

On the other hand, the machine learning technique is effective and less time-consuming. One can easily feed in the entire data (population) and make predictions. Using codes, we first cleaned the data, removed the outliers and n/a values, then trained and tested the data, checked its accuracy, and made predictions. For some of the variables, we even used SMOTE code to balance the data, which is a result of oversampling. Although the end results from predictive analysis using machine learning were different from that of the excel tool based, machine learning as a base is still a preferable and more reliable tool.

References

Ozgon.D. (2021). Google Colab Tutorial for Beginners | Get Started with Google Colab.

https://youtu.be/RLYoEyIHL6A

Keith.M. (2022). Introduction to Google Colab.

https://youtu.be/WFvY3qgtMqM

Programming with Mosh. (2019). Python Machine Learning Tutorial (Data Science).

https://youtu.be/7eh4d6sabA0

Decisive Date. (2019). The Fundamentals of Predictive Analytics - Data Science Wednesday

https://youtu.be/4y6fUC56KPw

Contribution Table

Name	Contribution
Saher Ismail Dinware	Descriptive Analysis + Regression Analysis + Findings & Conclusion
Sagar Sanjay Goel	Predictive Analysis [Machine Learning] + Findings & Conclusion
Piyush Chopra	Tableau
Himanshi Rally	Excel Dashboard
Samankula Ravindi Siriwardhane	Introduction