**Group Name:** Teen Titans GO

**Team Members:**

* Dimitari Gjorgievski - dgjor2
* Yashwi Shah - yshah33
* Dona Maria - dmari21
* Janki Patel - Jpate305
* Pratik Patel - ppate460

**Project Repo -** <https://github.com/ppate460/CS418_Group11>

**Progress Report**

**Stock Prediction For Fortune 500**

**Project introduction:**

Our project centers on delivering precise predictive analysis tailored for Fortune 500 companies in the United States. Our focus encompasses forecasting essential metrics such as stock prices, revenues, profits, and overall rank within this prestigious list of corporations. We recognize the critical role accurate predictions play in guiding the decisions of investors and businesses alike, ultimately influencing the stability of the economy. Through transparent delineation of our methodologies, challenges, and ongoing progress updates, our aim is to provide stakeholders with dependable forecasts. By empowering stakeholders with reliable insights, we seek to facilitate informed decision-making and contribute to economic prosperity for all involved parties.

In the dynamic realm of finance, the ability to anticipate the trajectories of Fortune 500 companies holds profound significance. These predictions serve as invaluable tools for investors, offering insights that can inform strategic decisions regarding investments. Moreover, they play a pivotal role in shaping the strategies of businesses, influencing hiring practices, and overall economic vitality. By striving for accuracy, transparency, and continuous improvement in our predictive analysis endeavors, we aspire to contribute to a landscape where stakeholders can navigate market uncertainties with confidence. Through collaborative efforts and a commitment to excellence, we aim to foster a future characterized by stability, growth, and prosperity for all.

**Any changes since the proposal:**

A significant addition to our plan is the inclusion of sector and industry data. Recognizing the influence of broader market trends and sector-specific dynamics, we deemed it essential to incorporate this information for more accurate predictions. Additionally, we have shifted from collecting stock price data on a quarterly basis to a monthly basis. This adjustment allows for more granular analysis and timely insights into market fluctuations. By broadening our analysis and increasing the frequency of data collection, we aim to enhance the accuracy and relevance of our predictions, better serving the needs of investors and businesses.

Another important change that we made was the span of the data. Initially, we had acquired data ranging from 1986-2023, but later we had to adjust it to 2011-2023. The reason for this was the APIs we are using to acquire additional data do not allow us to go far back in history.

**Data Collection:**

Our dataset forms the foundation of our analysis for Fortune 500 companies. Our data contains Fortune 500 companies for 13 years from 2011 to 2023 to ensure robust analysis and prediction. Each file contains essential components such as Rank, Company Name, Sector, Industry, Revenue, Profit, and Stock Price on a monthly basis, providing insights into financial standing and market performance. These variables provide invaluable insights into the financial standing and market performance of these companies. By categorizing companies into sectors, we're able to analyze trends and patterns in performance across different industries. This segmentation facilitates deeper analysis and enables us to draw meaningful conclusions about sector-specific dynamics. Additionally, we've incorporated macroeconomic indicators like Unemployment Rate, Tax, Financial Ratio, and GDP to offer a comprehensive view. This broader contextualization offers a comprehensive view, enriching our understanding of the multifaceted factors influencing company performance and market trends.

Data Description:

* Rank (numerical): The numerical position of the company within the Fortune 500 list.
* Company Name (Text): The name of the Fortune 500 company.
* Sector (Categorical): The broad category or industry sector to which the company belongs.
* Industry (Categorical): The specific field or niche within the sector that the company operates in.
* Revenue(numerical): The total income generated by the company over a specific period.
* Profit (numerical): The net income after deducting expenses from revenue.
* Stock Price (numerical): The monthly closing price of the company's stock over the course of a year (monthly m1 - m12 with opening and closing price).
* GDP (numerical): The Gross Domestic Product, a measure of the economic performance of a country.
* Unemployment Rate (numerical): The percentage of the labor force that is unemployed and actively seeking employment.
* Producer Price Index (numerical): A measure of the average change in selling prices received by domestic producers for their output.
* Consumer Price Index (numerical): A measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services.

Source:

We retrieved rank, company name, revenue, and profit data from a GitHub repository through web scraping techniques, leveraging publicly available datasets for Fortune 500 companies. Sector and industry information was obtained from Yahoo Finance, utilizing web scraping methods to extract relevant data. Additionally, stock price data spanning monthly intervals (m1 - m12) was sourced from the Polygon API, providing comprehensive insights into the market performance of these companies. By integrating data from multiple sources and employing diverse data acquisition methods, we ensure a robust and comprehensive dataset for our predictive analysis of Fortune 500 companies.

**Exploratory data analysis:** explain what your data looks like (visualizations are often

better). Include any interesting issues or preliminary conclusions you have about your data. (remove this after done with EDA)

**Exploratory Data Analysis:**

After reading the data into pandas DataFrames, the code then focuses on data cleaning to handle missing values and formatting inconsistencies, particularly in the 'Profit' and 'Revenue' columns where commas are removed and data types are converted to float for numerical computation. The exploratory analysis includes the aggregation of profit and revenue by sector, generating insights into sector-wise financial health over four years. Visualizations such as bar charts and violin plots are used to showcase the distribution and trends of profit by sector, revealing which sectors are more profitable over time. Additionally, a line graph presents the trends of profit by sector across the years, illustrating changes and potentially highlighting sectors with steady growth or volatility.

**At least five visualizations** that show an interesting hypothesis, along with an

explanation about why you thought this was an interesting hypothesis to

Investigate. Write the name of the member(s) who is responsible for each of them while

explaining it. (remove this after done with visualizations)

**Visualizations:**

**Visualization 1 (Janki & Dimitari):**

**Profit**

A trend analysis of earnings across sectors from 2020 to 2023, makes a strong case for the Technology sector's growth and resilience in the face of varying economic situations that may have been impacted by the pandemic and market changes. This theory is especially noteworthy since it suggests that society and economic institutions are shifting and becoming more reliant on technology and digital services. This is being driven by the widespread trend of remote labor, digital transformation, and e-commerce. Examining this theory is essential because it could provide information about long-term strategic investments and sustainable growth in the IT sector, which could have an impact on policy-making, workforce development, and education. As the analyst in charge, a comprehensive evaluation of the sector's performance, growth sustainability, and its inter-sectoral impacts would be imperative, necessitating a blend of data analytics, sector expertise, and economic forecasting to fully understand the implications for the future economy.

**Visualization 2 (Yashwi & Dona):**

The trend from 2011 to 2023 underscores a notable pattern: the consistent dominance of either consumer cyclical or financial services companies in securing positions within the top 500 rankings of the Fortune List. Notably, recent years have witnessed a notable ascendancy of consumer cyclical companies, indicating their increasing prominence and market strength. Conversely, real estate companies seem to be facing challenges, as evidenced by their limited representation within the top 500, suggesting potential struggles or limitations within the sector.

One hypothesis arising from this visualization is that sectors with higher levels of adaptability and innovation may be more likely to secure top positions in the rankings. For instance, the consistent presence of consumer cyclical companies at the top suggests a hypothesis that these companies may excel due to their ability to anticipate and respond to changing consumer preferences and market conditions. Conversely, sectors facing challenges in achieving top rankings, such as real estate, may prompt hypotheses about the impact of regulatory constraints, economic cycles, or shifts in demand on their performance. By testing these hypotheses against historical data and contextual insights, stakeholders can gain a deeper understanding of the drivers of sectoral success within the Fortune 500 rankings**.**

**Visualization 3 (Janki):**

The revenue trend for each sector from 2020 to 2023. This graphic raises an intriguing theory: in 2021, the Energy sector saw an abnormally high revenue peak, in contrast to other sectors that did not see such a sharp increase. This peak may be the result of a number of causes, including variations in the price of energy globally, modifications to energy policy, a rise in demand, or a mix of these.

Examining this theory is intriguing because it may provide light on the underlying political and economic forces that will influence the energy industry in particular in 2021. For the purpose of making future decisions about energy investments, sustainability initiatives, and energy production plans, it is imperative that investors, policymakers, and companies operating in the sector have a thorough understanding of these dynamics. Additionally, this research may aid in predicting such patterns or becoming ready for future market volatility. It's a useful study that illustrates how outside variables can have a rapid and significant impact on industry-specific economics, telling the story of how market performance and world events interact.

**Visualization 4 (Pratik):**

The relationship between GDP (Gross Domestic Product) and the Unemployment Rate is crucial for understanding the overall health of the economy.

This visualization shows the relationship between the US GDP in terms of value and the US Unemployment Rate over the period 2011-2023. By analyzing the graph, we can make a conclusion that there is an inverse relationship between these two variables. When the economic growth is robust and businesses are expanding rapidly, it results in a decline in the unemployment rate. If the GDP is increasing steadily while the unemployment rate is decreasing, it suggests a healthy economy with strong job creation which is exactly the case in our last 13 years of data represented. There was a clear spike in the unemployment rate during the pandemic which can be noticed just by looking at the graph and in the exact time frame, the GDP experienced a steep decline in terms of value which suggests an inverse relationship. So these two variables are crucial to predict the overall health of the economy moving forward.

**Visualization 5 ():**

**Machine Learning Analyses:**

**ML Analyzes 1 (Yashwi & Dona):**

We conducted a machine learning analysis on a dataset spanning from 2011 to 2023, focusing on the sector performance of Fortune 500 companies, and predicted which sector has the most contribution to the Fortune 500 companies in 2024. From the data collected, we gathered Fortune 500 companies' sectors from 2011 to 2023 and created a pie chart for each year from 2011 to 2023. The data is initially concatenated across years, and logistic regression is employed to predict the number of companies in each sector for 2024 based on previous years' data. Additionally, the code generates a pie chart visualizing the predicted number of companies in each sector for 2024, providing a graphical representation of the sector-wise performance outlook. This approach could serve as a baseline model for predicting sector performance in the absence of 2024 data. The scaled predictions ensure that the sum of predicted companies across all sectors for 2024 remains constant. The pie chart visualization then presents the predicted distribution of companies across sectors for 2024. We tested our accuracy and we got 100 percent for all training set accuracy, validation set accuracy, and test set accuracy.

**ML Analyzes 2 (Janki):**

The projected 2024 profit chart, which is probably based on an examination of data from 2020 to 2023, indicates a positive future for the Consumer Cyclical and Technology sectors. Thanks to increased consumer spending and a possible post-pandemic economic rebound, Consumer Cyclical is expected to achieve a market-leading profit margin. The IT sector's optimistic outlook is consistent with both its well-established growth trajectory and the growing societal acceptance of digital transformation. By comparing these projections to past growth rates and more general economic indicators, a baseline comparison would be made to determine whether or not these projections reflect the long-term trends that have been noticed.

Even though this prediction is speculative, it serves as a crucial benchmark for businesses and investors, providing a way to compare expected market performance to sector-specific and global economic trends. However, it is important to treat such forecasts cautiously, combining them with other analytical tools and economic theories for a more sophisticated decision-making process. This sample of Python code shows how sectoral profit data from the past can be used to create such predictive insights using linear regression modeling.

**ML Analyzes 3 (Janki):**

These forecasts most likely originate from a machine learning model—more precisely, a linear regression model—that was trained on historical data. Given the high revenue projection for the Real Estate industry as well, the Financial Services sector has a big lead, implying a prediction of significant financial transactions, possibly as a result of higher investments, recovery from economic downturns, or a thriving real estate market. This may be a sign of a strengthening economy and a favorable investment environment, both of which encourage increased activity in the financial services sector. Ideally, the model would analyze historical data to determine average growth rates for a baseline comparison. It would then project these rates into the future, accounting for any known market conditions or anticipated changes. This baseline would be used to compare the revenue forecast in order to determine how optimistic or conservative the projections were. Planning investments, developing policies, and making strategic decisions can all benefit greatly from understanding these results. Forecast increases in revenue for industries like healthcare and manufacturing could potentially be a sign of developments in those areas or a reaction to rising demand. Lower projections for industries like utilities and basic materials, on the other hand, can point to a maturation of those markets or a move toward new sources of energy and materials.

**Reflection (Pratik):**

* **What is the most challenging part of the project that you’ve encountered so far?**

The most difficult part of the project so far has been collecting and cleaning the data. Additionally, ensuring that the data is consistent, and free of errors or missing values requires careful attention to detail and strict data-cleaning procedures.

* **What is your initial insight?**

The initial insights for the projects are that we have completed the most important and very difficult part of the project is data gathering and analysis. Now that we have multiple visualizations analyzed using Machine Learning, we are confident that we are on the track to achieve the goal of our project as promised.

* **Are there any concrete results you can show at this point? If not, why not?**

Yes, as shown and described above multiple visualizations are the result of our project so far. We believe that these visualizations are sufficient evidence that our goal for the project is going to be achieved and very similar to these visuals. These visulas just did not appear out of nowhere, there are Machine Learning algorithms that are applied to the model and create visuals which are explained above under the “Machine Learning Analysis” section thoroughly.

* **Going forward, what are the current biggest problems you’re facing?**

We believe that our biggest challenge is going to be the news that is coming out everyday about all of those Fortune 500 companies which directly-indirectly affect the price of the stock of that company. We still do not have a proper solution of how we are going to tackle that. As per say, we do not know of any AI model that can automatically read all of the news and tell us whether the news is positive or negative for that particular company. So there are two options to corporate some sort of API that gathers news of all of the company or totally eliminate the news from our predictions model and purely shows the ending stock price based on the available data. Another thing that we need to work on is gathering stock data for the companies in our database from 2011-2019, since at the moment we only had the means to gather data from the years 2020-2023.

* **Do you think you are on track with your project? If not, what parts do you need to dedicate more time to?**

Overall, we are on track with our project. However, there is one aspect that requires additional attention: obtaining stock price data from 2011 to 2019. This time period is critical for our analysis because it covers a large portion of our dataset and provides important historical context for understanding market trends and company performance over time. As a result, spending more time collecting and incorporating stock price data into our analysis will be critical to ensure the completeness and accuracy of our predictive models.

* **Given your initial exploration of the data, is it worth proceeding with your project, why? If not, how will you move forward (method, data etc)?**

We believe that we have sufficient data and deep understanding now about the goal of our project and how we are going to complete it. We believe that we are still going to face some hiccups on the way when it comes to the ML model that predicts stock prices for fortune 500 companies by gathering all of the variables together. Right now, they are independently shown by a few graphs but combining one by one is going to be our approach moving forward.