**Group Name:** Teen Titans GO

**Team Members:**

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

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

**Project to Drive Folder:** <https://drive.google.com/drive/folders/1UHLYBmOjwAKpz99Zz8kMlziO1BFHrIth?usp=sharing>

**README-**

**Final Data folder name -> Raw\_Data**

**Final ML Models folder name -> ML\_Model**

**Final progress folder name -> Report**

**Final visualizations folder name -> Visualizations**

**Profit Prediction for Fortune 500 Companies**

**PROJECT INTRODUCTION:** Our project aims to provide precise predictive analysis tailored to Fortune 500 companies in the United States for 2024 based on the data from 2019 to 2023. We aim to predict the profit using financial data such as stock price, revenue, industry, sector, etc. With our analysis, we aim to help investors make informed decisions and allow companies to allocate resources efficiently to optimize profit. Also, these predictions play a big role in navigating the risks of an unpredictable market and economic crisis. This is important to all of us as these predictions affect employment, the economy, and our cost of living. Increased spending is correlated with a larger employment market and more opportunities. These companies are of great interest to investors, which makes precise prediction crucial.

**DATA:** The data collection process involved gathering rank, company name, tickers, revenue, and profit data from a GitHub repository. Web scraping techniques were utilized to extract sector and industry information from Yahoo Finance. Monthly stock price data was sourced from the API. Additionally, government data such as unemployment rates and GDP figures were incorporated into the dataset. The initial data and the cleaned data can be found in the Raw Data folder.

This information allows us to gain insights into the financial standing and market performance of these companies. By organizing companies into different sectors, we can examine trends and patterns in performance within various industries. This categorization allows for more in-depth analysis and helps us reach significant conclusions about dynamics unique to each sector. Moreover, we have integrated macroeconomic indicators such as the Unemployment Rate, Tax, Financial Ratio, and GDP to provide a holistic perspective. This broad contextualization enhances our grasp of the diverse factors impacting company performance and market trends.

**ML MODEL 1:** **RANDOM FOREST CLASSIFIER (JANKI)**

Using a RandomForestClassifier to evaluate financial data, the code illustrates a classification strategy. Data cleaning, categorical variable encoding, quantile-based revenue category definition, data splitting for training and testing, feature scaling, classifier training, prediction-making, and model performance evaluation are the many phases of this technique. The classification result produces good precision, recall, and F1-score values across many revenue categories. This suggests that using the provided features, the RandomForestClassifier model did a good job of forecasting revenue categories. The model's overall accuracy of 95% indicates that it is a solid method for assessing financial data because it can effectively classify companies into revenue groups. Because RandomForestClassifier manages complicated relationships in the data and is resistant to overfitting, as demonstrated by the balanced performance indicators in the classification report, it is appropriate for use in this situation.

**ML MODEL 2: Random Forest Regression for Predicting Profit (YASHWI)**

Constructed a Random Forest Regression model designed to predict profit based on the collected data set of features. I used the past four years (2020-2023) of data into a data frame and cleared the required columns accordingly. The dataset is then split into training and testing sets, with preprocessing pipelines for numerical and categorical data. The numerical pipeline imputes missing values with the median and scales the features using standardization, while the categorical pipeline handles missing values by labeling them as 'missing' and performs one-hot encoding. To optimize model performance, hyperparameter tuning is conducted using RandomizedSearchCV. After fitting the model, I evaluated the performance using the test set and computed the R2 score and mean squared error. The model achieved an R2 score of 0.7507, indicating a good fit to the data. The mean squared error is 9896903.0421, which is higher than normal, but after debugging and manipulating the data, I came to realize that the mean squared error could be improved with a larger dataset because after cleaning and filtering the data, the model doesn't have enough data to better predict.

**ML MODEL 3: Forecasting Fortune 500 Rankings (DONA)**

In this analysis, various machine learning techniques were employed to predict the rankings of Fortune 500 companies based on a diverse set of features including revenue, profit, market data (open and close prices for three months), and GDP data. Initially, a simple linear regression model was applied, resulting in a Mean Absolute Error (MAE) of 102.40. Subsequently, polynomial features were incorporated, and Ridge Regression was implemented, leading to a slight improvement in MAE to 90.34. Following this, a more advanced ensemble approach was adopted, combining RandomForestRegressor, GradientBoostingRegressor, XGBRegressor, and LGBMRegressor models. This ensemble model achieved a significantly lower MAE of 20.42, indicating a superior predictive accuracy. The MAE value signifies the average absolute difference between the predicted and actual rankings of Fortune 500 companies, suggesting that the ensemble model's predictions deviate from the actual rankings by approximately 20 positions on average. This level of accuracy underscores the effectiveness of the ensemble model in predicting the rankings of Fortune 500 companies, given the total number of companies in the list.

**ML MODEL 4: Economic Trend Forecast Long Short-Term Memory (LSTM) (PRATIK)**

Using Long Short-Term Memory (LSTM) neural networks, the offered code develops a machine learning model that evaluates economic patterns including a particular focus on the link between the GDP and the unemployment rate. Data on the unemployment rate and GDP from 2020 to 2023 are combined and cleaned up, allowing the model to be trained to recognize periodic trends and connections. The LSTM architecture is a great option for time series analysis since it can handle sequential data with ease. Users may predict future changes in the GDP and unemployment rate using the trained model which offers useful data for financial planning, policymaking, and investment decision-making. Because of the model's capacity to accurately forecast and capture complex historical interactions stakeholders and investors are better equipped to predict changes in the economy and reduce risks. The code makes it easier to create a strong and trustworthy tool for assessing and predicting economic trends which helps different industries make well-informed decisions and implement strategic planning.

**ML MODEL 5: Ridge Regression (Dimitar):**

Unfortunately we were working with limited data samples, but we did have a lot of features. The goal of this model was to predict the profit of a company for a given year, in terms of its characteristics such as, sector, industry, stock prices, gdp and unemployment rate. The reason behind choosing this model was because different machine learning literature referenced this model as being robust to low data samples. Since we found our selves in that situation, it seemed reasonable to give it a try. This model showed really good results. It achieved a high rate of prediction of the relationship between the dependent and independent variables, as well as resulting in a low rate of error.

**VISUALIZATION 1: Profit by sector(JANKI)**

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: Sector Performance over the years (DONA)**

We collected sector information for the years 2020-2023 for the Fortune 500 companies and created pie charts for each year to illustrate the distribution of companies that secured positions on the Fortune 500 list. The trend from 2020 to 2023 reveals a clear pattern: Industrials companies constantly dominate the Fortune 500 rankings and in the last year, consumer cyclical companies have grown in prominence. Real estate companies, on the other hand, appear to be struggling, indicating potential difficulties or limitations in the sector. This is interesting as the consistent presence of consumer cyclical companies at the top raises the possibility that these companies excel because of their ability to anticipate and respond to shifting consumer preferences and market conditions. Sectors that struggle to achieve top rankings, such as real estate, might be because of the impact of economic cycles, or shifts in demand on their performance. But, in the last two years, they have been doing comparatively well.

**VISUALIZATION 3: Revenue by sector (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: US GDP vs. Unemployment Rate (PRATIK)**

After mounting Google Drive, the code retrieves and imports the required information in terms of types of data needed which include the US monthly GDP data and the Unemployment Rate data for the period of 2011 to 2023. Data consistency and integrity were ensured by verifying the columns that have been included in both datasets. By transforming the data into date-time formation objects, the code integrates the date format and makes analyzing time easier. To get the datasets ready for combining; it then extracts the month and year from the date and time objects. The code combines the two datasets into a single coherent dataset, which is essential for comparison analysis, using the month and year as common identifiers. Data analysis relies heavily on visual representation, and the code uses Matplotlib to plot the combined data in an attractive manner. In order to throw light on any possible connections or patterns; the resulting graphic compares the US GDP trend over time with the unemployment rate. The comparison is made easier by the use of two axes: the unemployment rate is plotted on the right y-axis and the GDP is plotted on the left which is the x-axis. To ensure understanding, legends are used to differentiate between the two represented variables.

**VISUALIZATION 5: Comparison of Normalized US Monthly GDP and Mean Stock Price of Fortune 500 Companies (PRATIK)**

Comparison of Normalized US Monthly GDP and Mean Stock Price of Fortune 500 Companies. This visualization presents data regarding the US Monthly GDP and the average stock price of Fortune 500 businesses using the Pandas and Matplotlib libraries. The input for the function process\_data is an axis object and a year. It reads data that has been cleaned out of a CSV file that matches the specified year.

The mean stock price for each month is computed after the GDP data for each year is adjusted. The mean stock price and the adjusted GDP statistics are then shown on a subplot. The function process\_data is run for every year between 2020 and 2023, and the data is shown on a 2x2 grid of subplots. The month is shown on the x-axis, while the normalized value is shown on the y-axis. In every subplot, the normalized mean stock price of Fortune 500 businesses for a given year is compared with normalized GDP data. Which line reflects the mean stock price and which represents adjusted GDP data is shown by the legend.

Lastly, a graphic is presented to show the correlation between the US Monthly GDP and the average stock price of Fortune 500 businesses throughout the given time frame.

**VISUALIZATION 6: Yearly Mean Stock Price by Sector (2020 - 2023) (YASHWI)**

The chart shows the average stock prices for different sectors from 2020 to 2023, providing valuable information about how each sector performed over the four years. Each bar on the chart represents a specific sector, with the height of the bar indicating the average stock price for that sector in a particular year. The use of different colors for each year makes it easier to compare the performance of sectors over time, making it simple to identify any changes in how sectors are doing. By examining the variations in bar heights for each sector across the years, one can spot trends in how stock prices in specific sectors have fluctuated, pinpointing which sectors are consistently strong or unpredictable. When you compare the heights of bars in different sectors for a particular year, it helps to identify which sectors are performing relatively well or poorly compared to others. This visual representation is beneficial for investors, analysts, and decision-makers to comprehend how sectors are performing, evaluate potential investment opportunities, and make well-informed choices in the financial markets by considering factors such as sector resilience, adaptability, and vulnerability to external influences.

**VISUALIZATION 7: Monthly Average Stock Price (2020 - 2023) (YASHWI)**

Using data on stock prices from 2020 to 2023, our analysis reveals distinct patterns in monthly performance. Specifically, January and December stood out as very strong months for stock price performance. Nevertheless, this encouraging trend was short-lived, as stock values fell in the months that followed after January. It's interesting to note that stock prices seemed to bounce back near the end of the year, suggesting that late-year optimism is a recurring theme. The general outlook for 2020 remained low despite these rare upswings, indicating difficult market circumstances and unpredictable economic times. This in-depth analysis of stock price dynamics offers insightful information about the fluctuation and cyclicality of market patterns over the studied period.

**VISUALIZATION 8: Average Monthly Stock Price vs Monthly GDP (Dimitar)**

We saw in previous analyses that the GDP and the unemployment rate have an inverse relationship. Because of that we asked the question: Does GDP have any correlation with the stock prices? And our opinion from the data is that, yes, GDP and stock prices do have a relationship. Stock prices and GDP do share a strong correlation, especially in the long term. However, in the short term stocks do show volatility due to outside factors. Nevertheless, we conclude that GDP is a significant factor and we include it in the set of features along with stock prices.

**VISUALIZATION 9: Scatter Plot of Revenue vs. Profit over the Years (DONA)**

This is a scatter plot graphing revenue versus profit over the years 2020 to 2023. The horizontal axis represents revenue, while the vertical axis represents profit. Each dot on the graph corresponds to a data point, and the color of the dot indicates the year, with a color gradient ranging from dark purple for 2020 to yellow-green for 2023, shown in the legend on the right side of the graph. The distribution of dots suggests variability in both revenue and profit across the years. The concentration of dots near the origin implies that a substantial number of data points had relatively low revenue and profit. There are fewer dots as the revenue and profit values increase, indicating that fewer entities achieved higher revenue and profit. The plot does not show a clear trend between revenue and profit over the specified years. There are outliers which signifies that there are companies with unusually high or low revenue and profit compared to others in the dataset.

**VISUALIZATION 10: Comparing Stock Prices of Top Performing Companies (Dimitar)**

Since our ultimate objective is to predict a company’s profit based on stock prices, among other features, we want to see how the stock prices of the top performing companies behave, since after all they have the largest profit, hence being top performers. And the results were indeed surprising. The stock prices varied strongly among the companies. Higher ranked companies could be seen to have a lower stock price than lower ranked companies. But one thing they shared in common is that all of them, despite the price of a stock at a given month, were seeing continuous rises of their stock prices. Despite short term volatility, the stock prices of top performing companies were steadily increasing in the long term.

**ADDITIONAL WORK:**

**VISUALIZATION 11:** **Yearly Mean Stock Price and Unemployment Rate (2020 - 2023) (YASHWI)**

This visualization represents the yearly mean stock price and unemployment Rate from 2020 to 2023, using the unemployment and stock price on a monthly basis. From this analysis, the trend illustrates an inverse relationship between the average stock price and the unemployment rate, as the stock price rises and the unemployment rate falls. In 2020, the graph shows the mean stock price starting above $110 and increasing to above $150 by 2021. The average unemployment rate, in contrast, starts high in 2020 at around 8%, then decreases sharply to below 5.5% by 2021. Over the past four years, there appears an upward trend for the average stock prices, indicating the overall market is improving and positive economic conditions. Overall, the visualization offers valuable insights into the economy and highlights the relationship between the stock market and the economic market conditions.

**VISUALIZATION 12: Total Profitability over by Sector (DONA)**

The bar chart provides a visual representation of the total profitability of companies across various sectors. Each bar corresponds to a specific sector, with the sectors arranged in descending order based on their total profitability. The tallest bar, representing the "Financial Services” sector, indicates that companies within this sector collectively exhibit the highest level of profitability among all sectors analyzed. These companies are more efficient as they are generating a higher percentage of profit for each dollar of input expended. Following closely behind is the "Industrials'' sector, showing a substantial but slightly lower level of profitability. The "Real Estate" sector, represented by the shortest bar, signifies the sector with the least total profitability among the sectors examined. This visualization offers valuable insights into the relative performance of different sectors in terms of profitability, allowing stakeholders to identify sectors that are thriving financially and those that may require attention or strategic interventions to enhance profitability.

**VISUALIZATION 13: Sectoral Contributions to US GDP (PRATIK)**

Analyzing the various sectors' contributions to the US GDP offers important insights on the state and direction of the economy. Policymakers, investors, and companies may make well-informed decisions to promote sustainable development and reduce risks by knowing which sectors are driving economic growth. For example, if the GDP is regularly contributed to a large extent by specific industries, such as technology or healthcare, politicians may emphasize funding and policies that encourage innovation and research in these fields. In the same way, investors might strategically deploy capital, looking for growth prospects in industries that contribute significantly to the GDP. Additionally, companies may match their plans to the leading industries, which might boost their profitability and competitiveness. Furthermore, tracking shifts in sectoral contributions over time enables the detection of new trends and issues, directing preemptive actions to deal with changing economic dynamics. All things considered, examining the GDP's sectoral contributions is an essential tool for influencing economic policy, encouraging innovation, and guaranteeing the US will have sustained economic growth in the future.

**ML MODEL 6: Linear Regression for Predicting Sector (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 MODEL 7: RANDOM FOREST REGRESSOR (JANKI)**

The scatter plot graphically compares log-transformed predicted revenue against log-transformed actual revenue from a RandomForestRegressor machine learning model. On this plot, the x-axis and y-axis represent the actual and predicted revenues, respectively, suggesting both sets of figures have undergone a logarithmic transformation to mitigate data skewness and facilitate better interpretation. The plotted blue dots, denoting individual predictive data points, cluster tightly along a red dashed line which serves as the benchmark for perfect predictions, signifying a high degree of accuracy in the model's revenue forecasting ability. This accuracy is underscored by a legend distinguishing the blue dots as the "Predicted vs Actual" revenue and the red line as the "Ideal Prediction." The accompanying statistical results yield an R2 score close to 0.995, indicating an exceptional 99.5% variance explained by the model, and a Mean Squared Error (MSE) of roughly 3,323,187.77, which measures the average squared deviation between predicted and actual revenues, though this figure should be contextualized within the bounds of the logarithmic scale applied. Complementing the visualization, the provided code is part of a Python script that methodically carries out data procurement, cleansing, including outlier exclusion, and establishment of a preprocessing pipeline complete with hyperparameter tuning via randomized search, culminating in the performance evaluation of the model using R2 and MSE, and the display of optimized model parameters.

**ML MODEL 8: SVR (Dimitar)**

SVR is known for its ability to generalize well even with a small amount of data. This is crucial when you have limited samples because overfitting becomes a significant concern with more complex models.SVR can capture nonlinear relationships between input features and the target variable effectively, making it suitable for scenarios where the relationship is not linear, which is often the case with economic data like stock prices and GDP. It even performs well high-dimensional feature spaces, which is advantageous when dealing with multiple economic indicators. This model showed really good results. It achieved a high rate of prediction of the relationship between the dependent and independent variables, close to the Ridge Regression model, and even though it had a pretty low error rate, it ended up being an order of magnitude greater than the Ridge Regression model.

**RESULTS:**

An understanding of the dataset and its consequences was obtained by analyzing several machine-learning models and visualizations. With respect to revenue group classification, the Random Forest Classifier performs successfully, showing good accuracy, recall, and F1-score values. Its total accuracy of 95% highlights how useful it is for evaluating financial data. On the other hand, the Random Forest Regression model for profit prediction indicates difficulties brought on by small amounts of data and suggests ways to improve with a bigger dataset. On the other hand, the Fortune 500 ranking prediction combined model achieves higher accuracy, suggesting its effectiveness in ranking prediction.

In addition, the economic trend forecasting LSTM model demonstrates its potential to capture complicated historical relationships, supporting investors in making economic decisions. The model is further enhanced by visualizations that provide insights like trends in sector-specific performance and relationships between GDP and unemployment rates. Another ML Model stands out for its attractive scatter plot, which highlights the model's remarkable revenue forecasting accuracy and is backed by strong R2 scores and low Mean Squared Error. All things considered, these results help make well-informed decisions on investment strategies, financial planning, and policy creation.