Outline:

1. Abstract
2. Introduction
3. Motivation
4. Objectives
5. Exploring external factors
6. Data Collection & Collation
7. Data Cleaning
8. Predictive Modelling [Structure could be - Definition, variables involved, hyperparameter tuning, graphs maybe?]
    a. Linear regression
    b. Random Forest Regressor
    c. XG Boosting
9. Performance analysis
10. Conclusion
11. Limitations & Future Work

Introduction:

Capital markets play a critical role in a country's economy. Stock is one of the capital market investment vehicles that many investors are interested in. Stock prices fluctuate over time in response to market activity, which is determined by the strength of demand and supply for that particular share in the market. The stronger the demand for a stock, the higher its price will increase in the market, potentially benefiting investors; on the other hand, the lower the demand for a stock, the lower its price will decrease. Stock price fluctuations can be triggered by a variety of internal and external variables. The internal factor is linked to a company's performance, capital structure, valuation, and future prospects, among other things. Internal factors can be adjusted, controlled, and perfected by the company, resulting in rewards or benefits for stakeholders. While the company's external elements are linked to macroeconomic conditions such as economic growth, inflation, interest rates, world oil prices, and others, it can be determined whether the current economic conditions are favourable for stock market investing. Investors can use these internal and external characteristics as a guide for forecasting stock values.

Motivation:

The regulatory structure and lack of information transparency are the major flaws in the stock market, which make it difficult to gain investors' trust and give a solid foundation for assessing data without errors. Furthermore, investors lack capital market understanding (both basic and technical). As a result, the majority of them are unable to make a profit from the stock market. There are a few factors that have a significant impact on the price of a stock. The goal of this project is to figure out how these characteristics affect share price.

Objectives:

1. Identify External Factors that affect stock prices.
2. Collate information on the selected macroeconomic factors.
3. Develop Supervised Learning algorithms to predict the closing price of a stock.
4. Compare performance of different models.

Exploring external factors:

External factors are those that are related to an event that occurs outside of the company, and are frequently related to the country's social and economic conditions. Some external factors, according to research, include government announcements to change interest rates and deposits, exchange rates, inflation, and other economic regulation and deregulation issued by government; legal announcements such as employee's rights toward company or manager, manager's rights toward company, and company's right toward manager; securities announcements such as annual meetings report, insider trading, volume or stock price trading, limiting or postponement trading can significantly impact the shift of stock price in stock exchanges. In this project, five variables of external factors are taken: NASDAQ index, S&P 500 index, Daily break even inflation rates, GDP and Yield rates of Bonds.

- **NASDAQ index:**
  The Nasdaq Composite Index is a market capitalization-weighted index comprising approximately 2,500 Nasdaq common stocks. American depositary receipts, ordinary stocks, real estate investment trusts (REITs), and tracking stocks, as well as limited partnership interests, are among the instruments included in the index. All Nasdaq-listed stocks that aren't derivatives, preferred shares, funds, exchange-traded funds (ETFs), or debenture instruments are included in the index.

- **S&P 500 index:**
  The S&P 500 Index, also known as the Standard & Poor's 500 Index, is a market capitalization-weighted index of 500 of the country's most prominent publicly traded firms. Because there are other requirements to be included in the index, it is not an accurate list of the top 500 U.S. corporations by market cap. The index is widely recognised as one of the most accurate measures of large-cap U.S. stocks. The S&P 500 employs a market cap weighting approach, which means that businesses with the greatest market capitalizations receive a higher percentage allocation.

- **Daily break-even Inflation rates:**
  The breakeven inflation rate is generated from 10-Year Treasury Constant Maturity Securities and 10-Year Treasury Inflation-Indexed Constant Maturity Securities and represents a measure of projected inflation. The most recent value represents market participants' average expectations for inflation over the following ten years. By examining known market inflation rates from recent years, the breakeven inflation rate seeks to estimate the consequences of inflation on various assets. Despite the fact that this measurement is neither perfect nor certain, it is a good indication of what investors should expect from the market in the coming years. When investors are debating between buying TIPS (Treasury inflation-protected securities) and nominal Treasuries, this rate computation is commonly utilised. It might be able to tell you which one will protect you from inflation the best. The breakeven inflation rate gives an indication of future inflation tendencies. It also aids investors in determining whether Treasury bonds or TIPS are the best deal at the time.

- **GDP:**
  One of the most popular metrics used to track the health of a country's economy is gross domestic product (GDP). The GDP of a country is calculated by taking into account a variety of different aspects of that country's economy, such as consumption and investment. Because it is a measure of the entire dollar worth of all products and services generated by an economy during a certain time period, GDP is possibly the most closely followed and essential economic statistic for both economists and investors. It is frequently stated as a calculation of an economy's entire size as a measurement. Because it represents a representation of economic activity and development, GDP is a crucial metric for economists and investors. The GDP is important to investors because a big percentage shift in the GDP–up or down–can have a significant impact on the stock market. In general, a bad economy produces reduced profits for businesses. This can lead to a drop in stock values.

- **Yield rates of bonds:**
  The Treasury yield curve, also known as the term structure of interest rates, is a line chart that depicts the connection between on-the-run Treasury fixed-income instruments yields and maturities. It shows the yields on Treasury notes with set maturities of one, two, three, and six months, as well as one, two, three, five, seven, ten, twenty, and thirty years. As a result, they're also known as "constant maturity Treasury" rates. Market players pay close attention to yield curves because they are used to calculate interest rates (through bootstrapping), which are then used as discount rates to evaluate Treasury securities.

Market investors are also interested in determining the spread between short and long-term rates in order to calculate the slope of the yield curve, which is a predictor of the country's economic status.

Data Collection & Collation:

We looked into IBM stock prices over the last ten years for this project. IBM (International Business Machines Business) is a multinational technology corporation based in the United States with operations in more than 171 countries. IBM is a multinational corporation that manufactures and sells computer hardware, middleware, and software, as well as hosting and consulting services in a variety of fields, from mainframe computers to nanotechnology. The data sources for the IBM stock price and the external factors is as follows.

- IBM stock prices - Yahoo Finance
- NASDAQ index - Yahoo Finance
- S&P index - Yahoo Finance
- Inflation rate - FRED economic data
- GDP - BEA
- Yield rate - US Department of treasure

The following 3 point strategy was used to create the dataset:

1. The individual datasets were narrowed down to the same date range of 10 years.
2. Inconsistencies were identified(example: Stock market is closed on Weekends but daily we have records for daily inflation rates) and an outer join based on “Date” was performed on these columns.
3. Finally, records for bank holidays, federal holidays, weekends and national holidays were removed from the table.

Data Cleaning:

Cleaning data is an essential part of every machine learning research. Sometimes, due to poor data quality, the outcomes of a predictive model are skewed, with low accuracy and significant error percentages. As a result, fully cleaning the data before fitting a model to it is critical. Some examples of data cleaning include removing null records, deleting redundant columns, handling missing values, rectifying trash values (also known as outliers), reorganising the data to make it more accessible, and so on.

For the purpose of this project, a new dataset was created. Some records had Null values. Given that this was periodical data that followed a daily trend, it made the most sense to replace it with the closest possible value, which would be it's entry from the next record(ie, the next day).

Predictive Modelling:

To forecast the stock price, several supervised machine learning algorithms were explored, including:

1. Linear Regression

Linear Regression is a statistical approach for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors). In this project, we used Linear Regression to understand how the external factors (NASDAQ index, S&P 500 index, inflation, GDP, yield rates) influence the closing price of IBM stock.

- Variables Used: External factors and past stock price
- Performance Metric: R^2 Score, MAE, RMSE
- Limitations: Linear assumptions may not capture the complexity of financial time series.

2. Random Forest Regressor

Random Forest is an ensemble technique based on Decision Trees. It reduces overfitting by combining multiple trees and improves accuracy.

- Why it was used: Handles non-linearity well and less prone to overfitting.
- Hyperparameters Tuned: Number of estimators, max depth, min samples split
- Performance: Showed improvement over linear regression, especially in capturing complex relationships.

3. XGBoost Regressor

XGBoost is an advanced boosting algorithm known for its efficiency and performance.

- Reason for selection: Offers regularization, handles missing values, and is robust for large datasets.
- Hyperparameters Tuned: Learning rate, n_estimators, max_depth
- Result: XGBoost performed the best in terms of prediction accuracy.

Performance Analysis:

Model	R^2 Score	MAE	RMSE
Linear Regression	0.67	6.25	8.42
Random Forest	0.84	4.02	5.67
XGBoost Regressor	0.89	3.12	4.29

•XGBoost outperformed other models, indicating its effectiveness in capturing non-linear interactions.
•Random Forest gave a good balance between performance and interpretability.
Conclusion

The project demonstrated how external macroeconomic factors can be used in conjunction with stock price data to predict future prices. The LSTM model and advanced regressors like XGBoost helped to improve the accuracy of predictions. This reinforces the idea that financial models benefit significantly from the inclusion of external economic indicators.

Limitations & Future Work
•Limited Scope of Factors: Only five external variables were considered. More could improve the model.
•Time Lag Effects: External variables might have delayed effects which were not fully captured.
•Real-time Data: Models were trained on static historical data. Incorporating real-time streaming data would enhance predictive power.

Future Improvements
•Use of Recurrent Neural Networks (RNN/LSTM/GRU) for sequential dependencies.
•Incorporating Sentiment Analysis using news or tweets.
•Performing Feature Selection using SHAP or other importance metrics.
•Create a dashboard or web app for live forecasting using Streamlit or Dash.