Business Case: S&P 100 Stock Price Prediction and Market Sentiment Analysis

Context:

A fintech company is seeking to develop an AI-powered tool to predict stock prices and analyze market sentiment for the top 100 companies in the S&P 100 index. Using the **Yahoo Finance API** via the yfinance Python library, the company wants to gather historical stock price data for these companies, perform exploratory data analysis (EDA), and create predictive models. These models will help investors make informed decisions on whether to buy, hold, or sell stocks. Additionally, the tool will incorporate neural networks (LSTM) for advanced forecasting and sentiment analysis to add a deeper layer of insight into how public opinion influences stock movements.

The project will focus on collecting data from the past 10 years for each of the S&P 100 companies, conducting detailed analysis and visualization, and building machine learning models for future price prediction. For the advanced stages, students will implement sentiment analysis using Hugging Face's NLP models and create a predictive system that accounts for both quantitative and qualitative data.

Python Library: yfinance

pip install yfinance

import yfinance as yf

 # Example: Downloading data for Apple (AAPL)
data = yf.download('AAPL', start='2010-01-01', end='2020-12-31')
print(data.head())

Documentation:

- Official Documentation: yfinance on GitHub
- Functionality: yfinance allows you to fetch real-time and historical market data, access company financials, and retrieve data on dividends, splits, and institutional holdings.

Business Objectives

1. Data Collection for the S&P 100 Companies (Last 10 Years)
Your first task is to collect historical stock price data for the past 10 years for each of the top 100 companies in the S&P 100 index. You will use the **Yahoo Finance API** via the yfinance Python library.

List of S&P 100 Tickers

You will work with the following S&P 100 tickers, which represent major companies across various industries:
AAPL, MSFT, AMZN, GOOGL, GOOG, FB, BRK.B, JNJ, NVDA, JPM, UNH, V, PG, HD, MA, PFE, BAC, DIS, VZ, KO, NFLX, INTC, MRK, CSCO, T, CMCSA, CVX, XOM, PEP, ABT, ADBE, WMT, NKE, PYPL, TMO, CRM, ORCL, MCD, MDT, COST, AXP, LLY, BMY, QCOM, DHR, TXN, UNP, UPS, LIN, SBUX, HON, AVGO, AMGN, CAT, AMT, GILD, GS, SCHW, BKNG, MS, ISRG, SPGI, ZTS, INTU, FIS, USB, RTX, DE, C, BLK, PLD, MMM, IBM, NOW, SYK, CB, MO, EL, BA, ADP, CI, CL, SO, MRNA, LMT, TGT, ADI, GE, MDT, ABBV, WFC, CVS, LRCX, WM, PGR, EW, ITW, CME, NEE, AON, FISV, TRV

Data Collection Steps:

      1.   **Use** yfinance: Retrieve stock data for each ticker (open, high, low, close, volume, adjusted close).

      2.   **10-Year Historical Data**: Collect daily stock data for each ticker from the last 10 years.

      3.   **Data Cleaning**: Handle missing values, outliers, and ensure that the dataset is ready for further analysis and modeling.

Deliverable: A structured dataset containing historical price data (open, high, low, close, volume) for the S&P 100 companies over the last 10 years.

2. Exploratory Data Analysis (EDA) and Interactive Visualization Using Tableau

Once the data is collected, you need to perform a detailed **Exploratory Data Analysis (EDA)** to understand the patterns, trends, and insights within the stock price data. This analysis will help the company make better decisions regarding which stocks to focus on and how to analyze their price movements.

Key EDA Focus Areas:

      •   **Price Trends**: Visualize long-term trends for stock prices.

      •   **Volatility**: Analyze which stocks have the highest and lowest volatility over the 10-year period.

      •   **Sector Performance**: Group stocks by sector (technology, healthcare, financials, etc.) and compare their performance.

      •   **Correlation Analysis**: Explore how different stocks correlate with one another and with the overall S&P index.

Tableau Dashboard:

      •   **Objective**: Build an interactive Tableau dashboard that allows users to explore the stock data visually. Users should be able to filter by company, sector, and time range.

• **Features**: Visualize daily, monthly, and yearly price movements, and allow for comparison between companies and sectors.

Deliverable: An interactive Tableau dashboard with filters that presents key insights about the stock data, including price trends, sector comparisons, and volatility.

3. Build a Supervised Machine Learning Model for Stock Price Prediction

The next step is to create a machine learning model that predicts future stock prices based on historical data. This predictive model will help investors make informed decisions on when to buy or sell stocks.

Predictive Modeling Steps:

1. **Feature Engineering**: Create features from the stock data, such as moving averages (50-day, 200-day), price momentum, volatility measures, and technical indicators like RSI or MACD.
2. **Supervised Learning Models**: Use regression models like:
• **Linear Regression**
• **Decision Trees**
• **Random Forest**
• **Gradient Boosting (XGBoost, LightGBM)**
3. **Evaluation**: Measure model performance using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2.

Prediction Target:

• **Closing Price Prediction**: Your target variable will be the closing price for the next day based on historical data and the engineered features.

Deliverable: A machine learning model that predicts the next day's closing price for stocks in the S&P 100, along with a performance evaluation report.

4. Extra: Implement Neural Networks (LSTM) for Advanced Time Series Forecasting

For an advanced challenge, you will implement **Long Short-Term Memory (LSTM)** neural networks, a type of Recurrent Neural Network (RNN) designed to handle time series data effectively. LSTMs can capture long-term dependencies and patterns in stock price movements, making them ideal for price prediction.

LSTM Model Implementation:

1. **Select 5 Stocks**: Choose 5 stocks (e.g., AAPL, MSFT, AMZN, GOOGL, JPM) and build an LSTM model for each.
2. **Input Features**: Use historical price data (open, high, low, close, volume) as input features.

3.   **Training and Evaluation**: Train the LSTM model on past stock prices and predict future closing prices. Evaluate the model using time series-specific metrics like RMSE and MAE.

Deliverable: An LSTM-based time series forecasting model that predicts stock prices for selected companies and evaluates accuracy against real price movements.

5. Extra: Market Sentiment Analysis Using Hugging Face Transformers

In addition to quantitative price predictions, sentiment analysis can provide valuable insights into how market sentiment impacts stock prices. For this task, you will perform sentiment analysis using **Hugging Face's Transformers** to analyze financial news, social media posts, or earnings reports related to the selected stocks.

Sentiment Analysis Steps:

1.   **Data Collection**: Gather news articles, social media data (e.g., Twitter), or earnings reports for the selected stocks.
2.   **Sentiment Classification**: Use a pre-trained sentiment analysis model from Hugging Face to classify the sentiment as positive, neutral, or negative.
3.   **Correlation with Price**: Analyze how sentiment trends correlate with stock price movements over time.

Deliverable: A sentiment analysis report showing how public sentiment influences stock prices, integrated into the Tableau dashboard or used as a feature for stock price prediction.

Deliverables Summary

1.   **Data Collection**: A clean, well-structured dataset containing 10 years of stock price data for the S&P 100 companies.
2.   **EDA and Tableau Dashboard**: An interactive dashboard displaying key insights such as price trends, volatility, sector performance, and correlations.
3.   **Predictive Model**: A supervised learning model that predicts future stock prices based on historical data and technical indicators.
4.   **LSTM Model**: A time series forecasting model using LSTM for selected stocks.
5.   **Sentiment Analysis**: A sentiment analysis report that correlates public sentiment with stock price movements.

Business Challenge for Students

As data scientists, your task is to help this fintech company build a comprehensive tool that collects, analyzes, and predicts stock prices for the S&P 100 companies. You will need to perform exploratory data analysis, build predictive models, and

integrate advanced techniques like neural networks and sentiment analysis. Your work will provide investors with actionable insights to improve their trading strategies.

Key Questions to Consider:

- How will you ensure the stock data is clean and reliable for modeling?
- Which features will be most important for predicting future stock prices?
- How can sentiment analysis enhance stock price predictions and investor decisions?