Predicting Stock Prices using Linear Regression - An Analysis of S&P 500 Data

The objective of this project is to predict the future prices of the S&P 500 index using linear regression based on historical data. The dataset contains daily stock prices for the S&P 500 index in several years which can be easily downloaded in R using

("https://query1.finance.yahoo.com/v7/finance/download/%5EGSPC?period1=1467657600&period2=1627353600&interval=1d&events=history&includeAdjustedClose=true")

You will perform data cleaning, data preprocessing, and exploratory data analysis to gain insights into the data. You will then build a multiple linear regression model to predict the stock prices.

Steps to follow:

Load the dataset in R and preprocess the data by removing any missing values and normalizing the data.

Data Cleaning and Preprocessing: Load the dataset and remove any missing or irrelevant data. Preprocess the data by converting any non-numeric values to numeric.

Exploratory Data Analysis: Analyze the data to identify trends and patterns. Calculate summary statistics, generate visualizations, and perform correlation analysis.

Linear Regression: Build a linear regression model to predict future stock prices based on historical data. Train the model on a subset of the data and evaluate its performance using the remaining data.

Variable Significance: Analyze the significance of each variable and identify the most important factors affecting the stock prices.

Refine the Model: Refine the model based on the analysis of variable significance and evaluate its performance again.

Predict Future Stock Prices: Use the final model to predict the future prices of the S&P 500 index and visualize the results using plots and graphs.

Conclusion: Create a confusion matrix to perform binary classification of price increase/decrease and calculate evaluation metrics like accuracy, precision, recall, and F1-score.

Please attempt to solve the project on your own before referring to the provided project tips. However, if you encounter any difficulties, feel free to use the tips to assist you in completing the project.

Tips:

Load the necessary libraries such as tidyverse and lubridate.

Convert the date column to the Date format using the ymd() function from the lubridate package.

Create new columns for previous day's closing price and price difference using mutate() function from the dplyr package.

Clean the data by removing missing values using the drop_na() function from the tidyr package.

Split the data into training and test sets using the sample() function and subset() function.

Build a linear regression model to predict price difference using lm() function.

Refine the model using the summary() function to get information about the model's goodness of fit.

Predict the price difference for the test set using predict() function.

Calculate the predicted closing price for the test set by adding the predicted price difference to the test set's Adj.Close price.

Calculate the mean absolute error using the mean() and abs() functions.

Calculate the mean percentage error by dividing the difference between the predicted closing price and true closing price by true closing price and then calculating the mean.

Create a confusion matrix for binary classification of price increase/decrease using table() function.

Calculate evaluation metrics like accuracy, precision, recall, and F1-score using the confusion matrix.

Always check the data for missing values, outliers, and other anomalies before building the model.

Use different statistical and graphical tools to explore the data and identify patterns, trends, and relationships.

Split the data into training and test sets to evaluate the model's performance on unseen data.

Use different evaluation metrics to measure the model's performance, as one metric may not be sufficient.

Note: To ensure that you receive the full score for your project, it is important to fully describe your solution and provide clear explanations of your code. Please be thorough in your documentation and demonstrate a solid understanding of the concepts covered in the project

Design a project similar to Case Study One, in which you build a linear regression model to predict a variable such as the risk of heart attack or the price of cars, using online datasets.