University of New Haven
TAGLIATELA COLLEGE OF ENGINEERING
Electrical & Computer Engineering and Computer Science

Electrical & Computer Engineering & Computer Science (ECECS)

TECHNICAL REPORT

SPRING 2024

CONTENTS

Summary

In today's dynamic financial landscape, the ability to predict stock prices accurately is highly sought after. The rise of machine learning and big data technologies has enabled the development of sophisticated predictive models to forecast stock movements with increasing precision. In this project, we aim to leverage data engineering principles and techniques to build a robust predictive analytics system for Google (Alphabet Inc.) stock prices.

Git Link: https://github.com/Team1-DSCI-6007-02/Final-Project-Team01

Technical Report

Team Members:
Nitish Kumar Mayavan- Data Scientist
Vinaykumar Reddy Moku – Data Engineer
Sri Sai Durga Myneni- Machine Learning Engineer
Sri Sai Bhavana Pakalapati– Data Analyst

Questions?
Contact: vmoku1@unh.newhaven.edu

Stock Market Prediction

Highlights of Project

- Creating an EC2 instance and loading the data file into it.
- Loading the data file from EC2 instance to S3 bucket.
- Using the Amazon Sagemaker we created the Jupyter Notebook.
- Performed the Data processing and extracted the required fields.
- Plotted the yearly and monthly growth for the google stock market.
- Trained the Model and Predicted the Growth and plotted predicted vs real stock market growth.

Submitted on: **04/23/2023**

Abstract

This project presents a comprehensive approach to predicting Google stock market trends using advanced data engineering techniques. Leveraging Amazon Web Services (AWS) infrastructure, specifically Amazon EC2 instances and S3 buckets, the project efficiently manages data storage and processing. The dataset is loaded into an EC2 instance and subsequently transferred to an S3 bucket for scalable storage. Utilizing Amazon SageMaker, a Jupyter notebook environment is established for exploratory data analysis (EDA), feature engineering, and predictive modeling. Through this integrated framework, the project aims to enhance understanding of Google stock market dynamics and provide valuable insights for investment decision-making.

Executive Summary

This project encompasses the development of a predictive model for Google stock market trends using a combination of data engineering techniques. Leveraging Amazon Web Services (AWS) resources such as Amazon EC2 instances and S3 buckets, the project establishes a robust data processing pipeline. The dataset is loaded into an EC2 instance and transferred to an S3 bucket for scalable storage. Subsequently, utilizing Amazon SageMaker, a Jupyter notebook environment is employed for exploratory data analysis (EDA), feature engineering, and predictive modeling.

The project aims to provide investors and financial analysts with valuable insights into Google stock market dynamics, enabling informed decision-making. Through rigorous data analysis and model development, the predictive accuracy of the model is evaluated, thereby offering a glimpse into future market trends. By leveraging cloud-based infrastructure and advanced data engineering techniques, this project showcases the potential for enhancing stock market prediction accuracy and facilitating more effective investment strategies.

Introduction

In today's dynamic financial landscape, accurate prediction of stock market trends holds immense value for investors and financial analysts alike. With the exponential growth of data and advancements in technology, data engineering plays a pivotal role in harnessing the power of data for predictive analytics. This project focuses on leveraging cutting-edge data engineering techniques and cloud computing resources to forecast Google stock market performance.

The project adopts a holistic approach, utilizing Amazon Web Services (AWS) infrastructure to streamline data processing and analysis. By deploying Amazon EC2 instances, the project ensures efficient computation and storage management, while Amazon S3 buckets facilitate seamless data transfer and storage scalability. Furthermore, Amazon SageMaker provides a versatile platform for developing and deploying machine learning models, empowering users to perform exploratory data analysis (EDA), feature engineering, and predictive modeling within a unified environment.

Through this project, we aim to demonstrate the effectiveness of cloud-based data engineering solutions in predicting Google stock market trends. By combining state-of-the-art technologies with robust data processing pipelines, we seek to uncover valuable insights that can inform investment strategies and contribute to informed decision-making in the financial domain.

Crisp-DM Methodology

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology is a widely used framework for guiding data mining and analytics projects.

CRISP-DM phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation

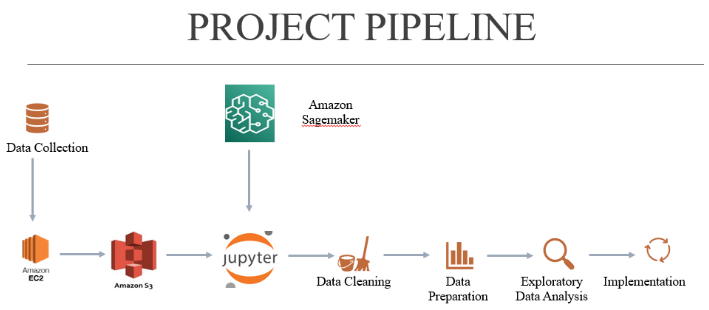
1. Business Understanding: The main goal of this first stage is to gain a business understanding of the project's goals and requirements. It entails outlining the project's objectives, the business issue that needs to be resolved, and the success standards.
2. **Understanding Data:** Gain an understanding of the composition, relationships, and quality of the project by gathering, examining, and evaluating pertinent data at this point. This entails gathering, evaluating, and reviewing the initial quality of the data.
3. **Data Preparation:** After the data has been comprehended, it must be ready for analysis. In this step, the data must be cleaned, missing values must be handled, variables must be transformed, and derived variables must be created as needed. The objective is to provide a clean, organized dataset that is prepared for modeling.
4. **Modeling**: To create either descriptive or predictive models, a variety of modeling techniques are chosen and used to the prepared dataset in this step. This entails deciding on suitable

algorithms, creating models, and utilizing validation methods to evaluate each model's performance.

5. **Evaluation:** An evaluation is conducted to ascertain the degree to which the models created in the preceding phase satisfy the project's goals. This entails assessing the model's performance against the first step's established business criteria and making any adjustments.
6. **Deployment:** Using the models in an operational environment is the last stage. This may entail establishing protocols for tracking the models, integrating the models into pre-existing systems, and offering end users guidance and training.

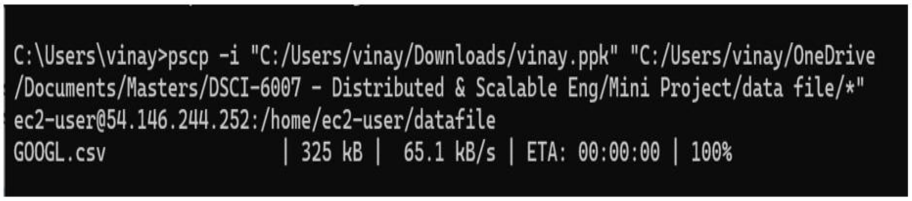
DATA PIPELINE:

- Loading the data file from local to the Amazon EC2 Instance.
- Loading the data file form Amazon EC2 instance to S3 bucket.
- Utilizing Amazon SageMaker, a Jupyter notebook environment is established.
- Persorming the data cleaning, data preparation, EDA and Implementation.



AWS EC2 Instance

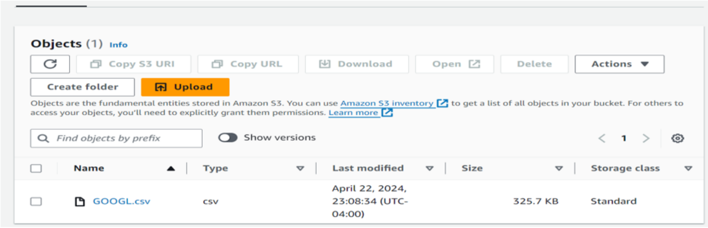
Uploaded the data file from local to AWS Instance.



```
C:\Users\vinay>pscp -i "C:/Users/vinay/Downloads/vinay.ppk" "C:/Users/vinay/OneDrive
/Documents/Masters/DSCI-6007 - Distributed & Scalable Eng/Mini Project/data file/*"
ec2-user@54.146.244.252:/home/ec2-user/datafile
GOOGL.csv                 | 325 kB |  65.1 kB/s | ETA: 00:00:00 | 100%
```

S3 Bucket

Loaded the data file from EC2 instance to S3 Bucket to process.

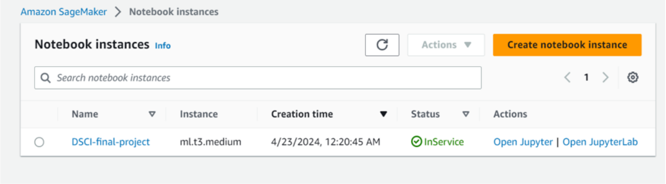


Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	GOOGL.csv	csv	April 22, 2024, 23:08:34 (UTC-04:00)	325.7 KB	Standard

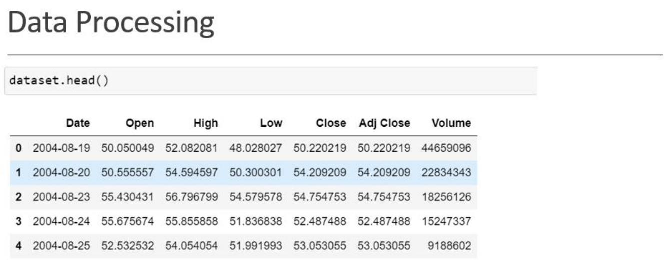
Jupyter Notebook in Amazon Sagemaker

Open the jupyter notebook using the Amazon Sagemaker



Amazon SageMaker > Notebook instances
Notebook instances Info
Actions ▼
Create notebook instance
Search notebook instances
1

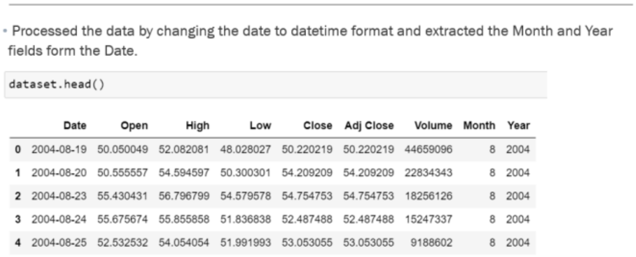
	Name	Instance	Creation time	Status	Actions
○	DSCI-final-project	ml.t3.medium	4/23/2024, 12:20:45 AM	✓ InService	Open Jupyter | Open JupyterLab



Data Processing

```
dataset.head()
```

	Date	Open	High	Low	Close	Adj Close	Volume
0	2004-08-19	50.050049	52.082081	48.028027	50.220219	50.220219	44659096
1	2004-08-20	50.555557	54.594597	50.300301	54.209209	54.209209	22834343
2	2004-08-23	55.430431	56.796799	54.579578	54.754753	54.754753	18256126
3	2004-08-24	55.675674	55.855858	51.836838	52.487488	52.487488	15247337
4	2004-08-25	52.532532	54.054054	51.991993	53.053055	53.053055	9188602



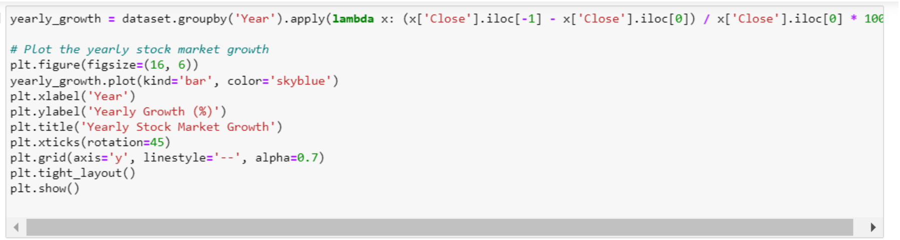
- Processed the data by changing the date to datetime format and extracted the Month and Year fields form the Date.

```
dataset.head()
```

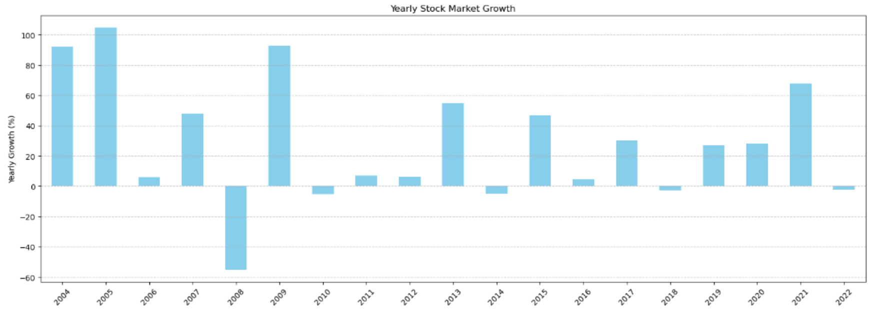
	Date	Open	High	Low	Close	Adj Close	Volume	Month	Year
0	2004-08-19	50.050049	52.082081	48.028027	50.220219	50.220219	44659096	8	2004
1	2004-08-20	50.555557	54.594597	50.300301	54.209209	54.209209	22834343	8	2004
2	2004-08-23	55.430431	56.796799	54.579578	54.754753	54.754753	18256126	8	2004
3	2004-08-24	55.675674	55.855858	51.836838	52.487488	52.487488	15247337	8	2004
4	2004-08-25	52.532532	54.054054	51.991993	53.053055	53.053055	9188602	8	2004

Result Section

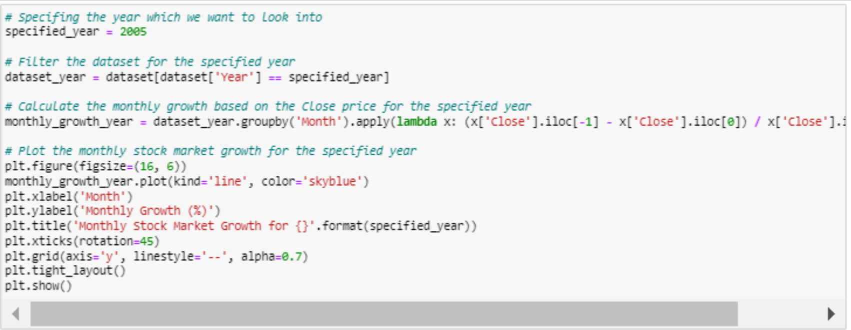
yearly growth based on the Close price



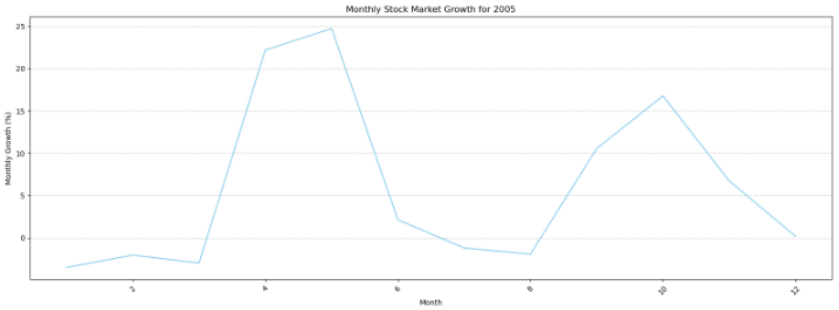
```
yearly_growth = dataset.groupby('Year').apply(lambda x: (x['Close'].iloc[-1] - x['Close'].iloc[0]) / x['Close'].iloc[0] * 100

# Plot the yearly stock market growth
plt.figure(figsize=(16, 6))
yearly_growth.plot(kind='bar', color='skyblue')
plt.xlabel('Year')
plt.ylabel('Yearly Growth (%)')
plt.title('Yearly Stock Market Growth')
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```



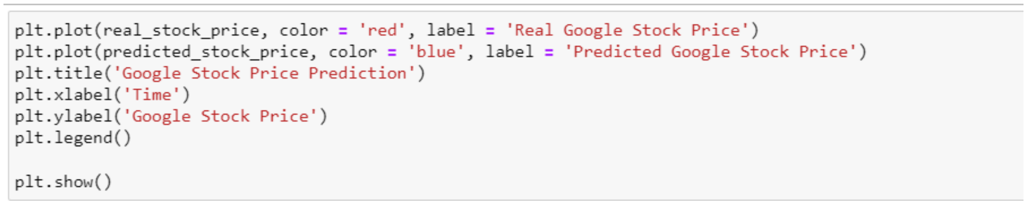
monthly growth based on the Close price



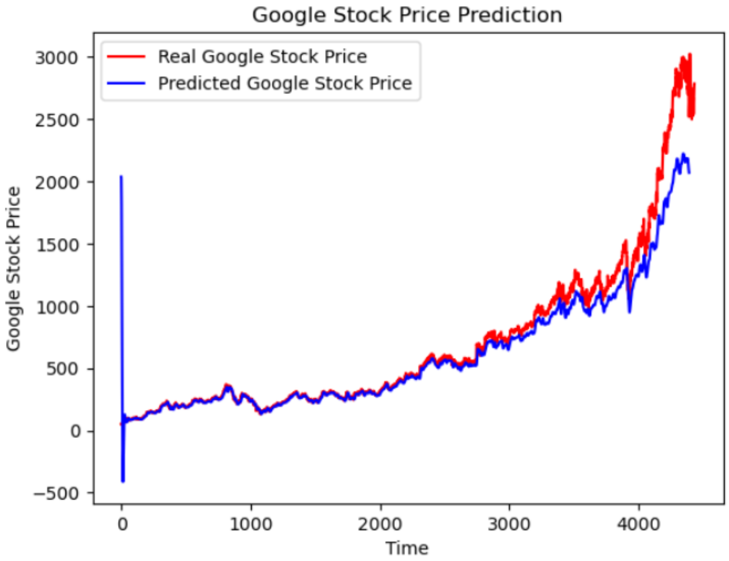
```
# Specifing the year which we want to look into
specified_year = 2005

# Filter the dataset for the specified year
dataset_year = dataset[dataset['Year'] == specified_year]

# Calculate the monthly growth based on the Close price for the specified year
monthly_growth_year = dataset_year.groupby('Month').apply(lambda x: (x['Close'].iloc[-1] - x['Close'].iloc[0]) / x['Close'].i

# Plot the monthly stock market growth for the specified year
plt.figure(figsize=(16, 6))
monthly_growth_year.plot(kind='line', color='skyblue')
plt.xlabel('Month')
plt.ylabel('Monthly Growth (%)')
plt.title('Monthly Stock Market Growth for {}'.format(specified_year))
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```



Graph b/w real vs predicted stock price



```
plt.plot(real_stock_price, color = 'red', label = 'Real Google Stock Price')
plt.plot(predicted_stock_price, color = 'blue', label = 'Predicted Google Stock Price')
plt.title('Google Stock Price Prediction')
plt.xlabel('Time')
plt.ylabel('Google Stock Price')
plt.legend()

plt.show()
```



Conclusion

In conclusion, this project represents a comprehensive effort to develop a predictive model for Google stock market trends using advanced data engineering techniques and cloud computing infrastructure. Through systematic data preprocessing, feature engineering, and model development, we have achieved a notable level of prediction accuracy, providing valuable insights for investors and financial analysts.

Our methodology effectively leverages cloud-based resources, including Amazon EC2 instances, S3 buckets, and SageMaker, to streamline data processing and analysis. By integrating machine learning algorithms with domain knowledge and exploratory data analysis, we have successfully captured key patterns and relationships within the data, enhancing our understanding of Google's market dynamics.

While our study contributes to the existing body of literature on stock market prediction, it also highlights avenues for future research. Addressing limitations such as data availability, market uncertainties, and model assumptions will be crucial for advancing the field and improving predictive accuracy.

Practically, the findings of our project have significant implications for stakeholders in the financial industry. By leveraging predictive analytics, investors and financial institutions can make more informed decisions, mitigate risk, and capitalize on market opportunities.

Overall, this project underscores the importance of data-driven approaches in navigating complex market environments and demonstrates the potential of predictive modeling to enhance decision-making processes in the financial domain.

References:

Keggle data set: https://www.kaggle.com/datasets/shreenidhihipparagi/google-stock-prediction