I. Cover Page:

Enterprise Cloud Computing and Big Data (BUDT737)

Project Title: Company Bankruptcy Prediction using financial leverage-based features.

Team Members:

Sharjeel Nawaz
Tirth Shah
Siddhesh Mishra

(SIGN THE FOLLOWING STATEMENT AND INCLUDE IT ON THE COVER PAGE OF YOUR PROPOSAL)

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	**Signature**
	Sharjeel Nawaz	SN
	Tirth Shah	TS
	Siddhesh Mishra	SM

II. Executive Summary

Introduction

The ability to predict potential bankruptcy situations is crucial for investors, lenders, and other stakeholders in the financial sector. This project aims to develop a robust predictive model for company bankruptcy using financial leverage-based features. By leveraging the power of Python, PySpark, and logistic regression, we can analyze large datasets and provide valuable insights into the financial health of publicly traded companies.

Data and Features

The project utilizes a comprehensive dataset containing financial information for thousands of publicly traded companies. Key features considered for the analysis include:

1. Current Ratio: A liquidity ratio that measures a company's ability to pay its short-term obligations.
2. Quick Ratio: A more conservative liquidity ratio that excludes inventories from current assets.
3. Gross Margins: A profitability ratio that measures the percentage of revenue left after deducting the cost of goods sold.
4. Net Income: The bottom-line profitability metric that represents a company's total earnings or losses.
5. Leverage Ratio: A solvency ratio that measures the extent to which a company relies on debt financing.

Methodology

The project leverages the distributed computing capabilities of Apache Spark and the PySpark library to handle large-scale data processing and analysis. The following steps are involved:

1. Data Preprocessing: The dataset is cleaned, handled for missing values, and transformed into a suitable format for model training.
2. Feature Engineering: Additional relevant features may be derived from the existing data to enhance the predictive power of the model.
3. Model Training: Logistic regression, a powerful classification algorithm, is employed to train the predictive model using the financial leverage-based features.

4. Model Evaluation: The trained model is evaluated using appropriate metrics, such as accuracy, precision, to assess its performance.
5. Model Deployment: The final model can be deployed as a service or integrated into existing financial analysis workflows for real-time bankruptcy prediction.

Benefits and Applications

The developed bankruptcy prediction model offers several benefits and potential applications:

1. Risk Assessment: Investors and lenders can use the model to assess the risk of potential bankruptcies, enabling informed decision-making.
2. Portfolio Management: Financial institutions can leverage the model to monitor and manage their investment portfolios, mitigating potential losses.
3. Early Warning System: The model can serve as an early warning system, alerting stakeholders to potential financial distress situations, allowing for timely interventions.
4. Regulatory Compliance: Regulatory bodies can use the model to identify companies at risk of bankruptcy, ensuring proper oversight and protection for investors.
5. Industry Analysis: Various industries operate using different capital structures. This model can be used to anlayze S&P500 companies to define guidance for key features. These industries are:

Conclusion

By leveraging financial leverage-based features, Python, PySpark, and logistic regression, this project aims to develop a robust predictive model for company bankruptcy. The model's ability to analyze large datasets and provide accurate predictions can significantly benefit investors, lenders, and regulatory bodies in the financial sector, enabling informed decision-making and risk mitigation strategies.

III. Data Description (1 page)

Data source

We source data from Kaggle using Kaggle API.

```
# Downloading the company bankruptcy data from kaggle

!kaggle datasets download -d fedesoriano/company-bankruptcy-prediction
```

```
company-bankruptcy-prediction.zip: Skipping, found more recently modified loc
```

Data Description:

Kaggle dataset is provided as a 'company-bankruptcy-prediction.zip' file consisting of 'data.csv' file.

```
# Unzipping the dataset
!unzip 'company-bankruptcy-prediction.zip'
```

```
Archive:  company-bankruptcy-prediction.zip
replace data.csv? [y]es, [n]o, [A]ll, [N]one, [r]ename: Y
  inflating: data.csv
```

The dataset consists of over 96 variables/features/columns and 6829 observations/rows.

```
# from pyspark.sql.session import SparkSession
session=SparkSession.builder.appName("bankrupt").master("local[2]").getOrCreate()
data=session.read.csv("data.csv", header=True, inferSchema=True)
data.columns
```

```
' Total debt/Total net worth',
' Debt ratio %',
' Net worth/Assets',
' Long-term fund suitability ratio (A)',
' Borrowing dependency',
' Contingent liabilities/Net worth',
' Operating profit/Paid-in capital',
' Net profit before tax/Paid-in capital',
' Inventory and accounts receivable/Net value',
' Total Asset Turnover',
' Accounts Receivable Turnover',
' Average Collection Days',
' Inventory Turnover Rate (times)',
' Fixed Assets Turnover Frequency',
' Net Worth Turnover Rate (times)',
' Revenue per person',
' Operating profit per person',
' Allocation rate per person',
' Working Capital to Total Assets',
' Quick Assets/Total Assets',
' Current Assets/Total Assets',
' Cash/Total Assets',
' Quick Assets/Current Liability',
' Cash/Current Liability',
' Current Liability to Assets',
' Operating Funds to Liability',
' Inventory/Working Capital',
' Inventory/Current Liability',
' Current Liabilities/Liability',
' Working Capital/Equity',
' Current Liabilities/Equity',
' Long-term Liability to Current Assets',
' Retained Earnings to Total Assets',
' Total income/Total expense',
' Total expense/Assets',
' Current Asset Turnover Rate',
```

Sample size (n) = 6819

Variables (k) = 96

However, we are only interested into 6 variables, which are all numeric in nature, and bankruptcy output. Hence,

```
[ ]  # Selecting only necessary columns for model training

     clean_data = data.select(' ROA(C) before interest and depreciation before interest',' Operating Gross Margin',  ' Current Ratio', ' Quick Ratio', ' Debt ratio %', ' After-tax Net Profit Growth Rate', 'Bankrupt?')
     clean_data.show(5)
```

```
+-----------------------------------------------------+-------------------------+-------------------+-------------------+-------------------+-------------------------------+---------+
| ROA(C) before interest and depreciation before interest| Operating Gross Margin|      Current Ratio|        Quick Ratio|      Debt ratio %| After-tax Net Profit Growth Rate|Bankrupt?|
+-----------------------------------------------------+-------------------------+-------------------+-------------------+-------------------+-------------------------------+---------+
|                                          0.370594257300249|       0.601457213277793|  2.25896331056689|1.2077550852353198|0.207576261450555|              0.688979462807371|        1|
|                                          0.464290937454297|       0.610235085544617| 6.016205877672769|  4.03936682691048|0.171176346101521|              0.689692901209989|        1|
|                                          0.426071271876371|       0.601450006486113|11.542553689380101| 5.347560222436551|0.207515796474892|              0.689462676951537|        1|
|                                          0.399844001364988|       0.583541129160121|    4.194058676884|  2.89649114708036|0.151464764035432|              0.689109535618137|        1|
|                                          0.465022181055916|       0.598783493564335|  6.02244610781246|  3.72744660861623|0.106509054630105|              0.689696883030727|        1|
+-----------------------------------------------------+-------------------------+-------------------+-------------------+-------------------+-------------------------------+---------+
only showing top 5 rows
```

All variables are either %ages or ratios, and hence do not have any units. None of the variables are categorical variables.

We used these variables because leverage, growth, margins and returns signify how well a company can pay off its liabilities and return to its investors. If these indicators are not healthy, there is a risk that company might be considering bankruptcy in the future. Also, these indicators are not impacted by high volatility in stock price of market capitalization. Usually, these numbers are updated every 3 months, giving professional insights into the company's financial health.

III. Research Questions (1 page)

1. **Identifying Early Warning Signs**: Can we identify specific patterns or thresholds in the financial leverage-based features that serve as early warning signs of potential bankruptcy? For example, are there certain levels of current ratio, quick ratio, or leverage ratio that consistently indicate an increased risk of bankruptcy?
2. **Assessing the Impact of Profitability**: How do profitability metrics, such as gross margins and net income, contribute to the prediction of bankruptcy? Can we quantify the relationship between declining profitability and the likelihood of bankruptcy across different industries or company sizes?
3. **Evaluating Industry-Specific Factors**: Do the predictive capabilities of the model vary across different industries? Are there industry-specific nuances or additional features that need to be considered to enhance the accuracy of bankruptcy predictions within certain sectors?
4. **Determining the Significance of Leverage**: What is the relative importance of the leverage ratio compared to other financial ratios in predicting bankruptcy? Can we identify the optimal combination of leverage-based features that maximizes the predictive power of the model?
5. **Temporal Analysis**: How do the predictive patterns change over time? Can we detect early signs of financial distress by analyzing the trends and fluctuations in the financial leverage-based features leading up to bankruptcy events?
6. Risk Stratification: Can we stratify companies into different risk categories based on their financial leverage-based features? This would allow stakeholders to prioritize their attention and resources towards companies with higher bankruptcy risk.
7. **Model Generalization**: How well does the developed model generalize to new, unseen data? Can we validate the model's performance across a diverse range of companies, industries, and economic conditions to ensure its robustness and reliability?

By investigating these questions, we aim to gain a deeper understanding of the relationships between financial leverage-based features and bankruptcy risk. The insights derived from this analysis can inform decision-making processes, risk management strategies, and early intervention measures for investors, lenders, and regulatory bodies in the financial sector.

IV. Methodology (1 page)

In the "Company Bankruptcy Prediction using Financial Leverage-Based Features" project, we employed several techniques to develop a robust predictive model:

1. Logistic Regression:

Logistic regression is a widely-used machine learning algorithm for binary classification problems, such as predicting the bankruptcy or non-bankruptcy status of a company. It models the probability of an event occurring based on one or more independent variables. We chose logistic regression because it provides interpretable results, allowing us to understand the relative importance of each feature in the prediction process. Additionally, logistic regression is robust to multicollinearity among the independent variables, which is common in financial data.

2. PySpark:

PySpark, the Python API for Apache Spark, was chosen as the primary data processing and analysis framework due to its ability to handle large-scale datasets efficiently. With thousands of samples and multiple features, the dataset required a distributed computing environment to enable parallel processing and minimize computational bottlenecks. PySpark's scalability and integration with Python made it an ideal choice for this project.

3. Feature Engineering:

Effective feature engineering is crucial for building accurate predictive models. We employed techniques such as feature scaling, normalization, and transformation to ensure that the financial leverage-based features were in a suitable format for the logistic regression algorithm. Additionally, we explored the creation of derived features by combining or transforming existing features to capture more complex relationships and improve the model's predictive power.

4. Data Preprocessing:

Fortunately, the dataset was quite clean. We only scaled 2 columns and selected 6 features of our interest as sub-set.

5. Model Evaluation:

To assess the performance of the trained model, we employed accuracy score.

V. Results and Finding (varies considerably in length depending on study)

To test our model, we used yahoo finance's web scrapping API to scrape information about publicly trading companies. Then we used our model to predict their incline towards bankruptcy.

```
# data scrapping libraries
import yfinance as yf
```

```
[ ]  ticker = input ('Please enter your ticker name : ')
```

```
     Please enter your ticker name : GRMN
```

```
[ ]  # Storing all the information of the company
     data2 = yf.Ticker(ticker).info
```

```
# Creating a dataset with required variables
test_data = [str(data2.get('shortName')), float(data2.get('returnOnAssets')), float(data2.get('grossMargins')), float(data2.get('currentRatio')), float(data2.get('quickRatio')), float(data2.get('debtToEquity')), float(data2.get('ear

from pyspark.sql.types import StructType, StringType, DoubleType
from pyspark.sql.types import StructField, StructType

company_schema = StructType([
    StructField(' Company', StringType(), True),
    StructField(' ROA(C) before interest and depreciation before interest', DoubleType(), True),
    StructField(' Operating Gross Margin', DoubleType(), True),
    StructField(' Current Ratio', DoubleType(), True),
    StructField(' Quick Ratio', DoubleType(), True),
    StructField(' After-tax Net Profit Growth Rate', DoubleType(), True),
    StructField(' Debt ratio %', DoubleType(), True),
```

However, this comes with a risk of false-positive values, which can wrongly identify a healthy company as high-risk company. One way to do that is to adjust thresholds for class-A and class-B probabilities in logistic regression. At the current scope of the project, we have not employed this technique, but a more detailed study is required to find working threshold values. We relied on default threshold values of 0.5.

Our final accuracy of the model was 89.15%. However, with more fine tuning, we believe this can be further improved.

```
#Determining the accuracy
lrresults=lrresults.withColumn("compare",lrresults['Bankrupt?']-lrresults['prediction'])
correct=lrresults.filter(lrresults['compare']==0).count()
incorrect=lrresults.filter(lrresults['compare']!=0).count()
print(correct/(correct+incorrect))
```

```
0.8915254237288136
```

We tested over a dozen public companies through our model, and it was good to see that these companies were healthy.

1. **Identifying Early Warning Signs**: We tested our model for Garmin, which is a struggling company due to mobile phone based GPS systems. The company is also facing fierce competition for its wearables, and other electronics. Our model predicted this a high risk company

```
# Predicting the result of desired company
usecase_results=lrmodel.transform(usecase_newData)
usecase_results.select(' Company','prediction').show(2,truncate=False)
```

```
+-----------+----------+
| Company   |prediction|
+-----------+----------+
|Garmin Ltd.|1.0       |
+-----------+----------+
```

2/3. Assessing the Impact of Profitability and Evaluating Industry-Specific Factors: We found that profitability has a slight impact on bankruptcy of a company. The main features implying a future bankruptcy are quick ratio and current ratio. This can cause misleading results for tech companies, which are enjoying great profitability and growth of stock prices. We tested this hypothesis, and AMD was wrongly predicted by our model, as a high risk company.

```
+--------------------------+----------+
| Company                  |prediction|
+--------------------------+----------+
|Advanced Micro Devices, Inc.|1.0       |
+--------------------------+----------+
```

Determining the Significance of Leverage: We have found that leverage ratio directly impacts on the risk of company's probability of bankruptcy. However, overall weight of leverage is much lower than quick and current ratios.

Temporal Analysis: It is very unclear from current dataset to determine change in company's health over time. However, a more detailed data set with details of company's industry, geography, and other factors, and at least 20 years of data, we can do this analysis.

Risk Stratification: Yes, above points mention that we can perform risk stratification based on this study in 2 ways. By building a portfolio of companies with various risk levels. Secondly, by expecting more returns on high-risk companies, as a premium.

Model Generalization: It is not clear from this data. To do model generalization, we need audited data of at least 10k companies listed with SEC, which is currently not available.

VI. Conclusion

Leverage based risk assessment for a company's risk for bankruptcy is valuable for a higher view analysis. However, this does not capture macroeconomic industry trends. Any company classified as high risk should future be analyzed before making a final determination. However, we can use this model to perform early-stage study to build portfolio or find companies of our interest for investment or hedging.

VII. Appendix (Any additional information to be submitted):

Company	prediction
Garmin Ltd.	1.0

Company	prediction
Walt Disney Company (The)	0.0

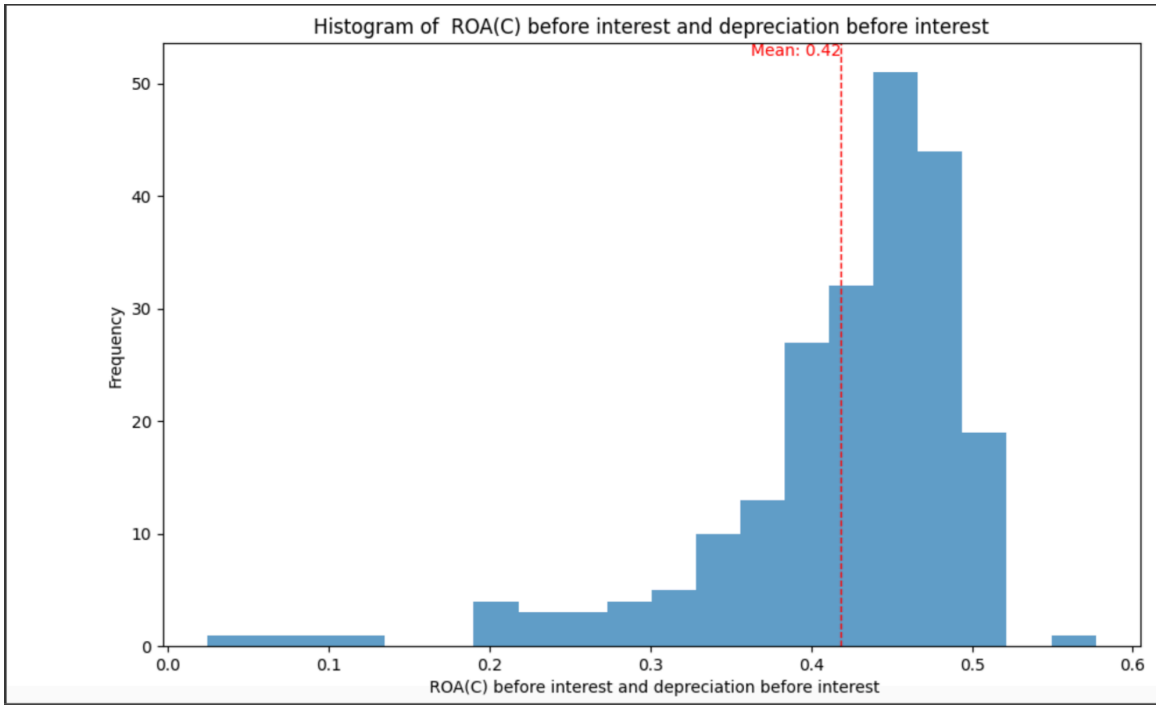
Company	prediction
Hewlett Packard Enterprise Comp	0.0

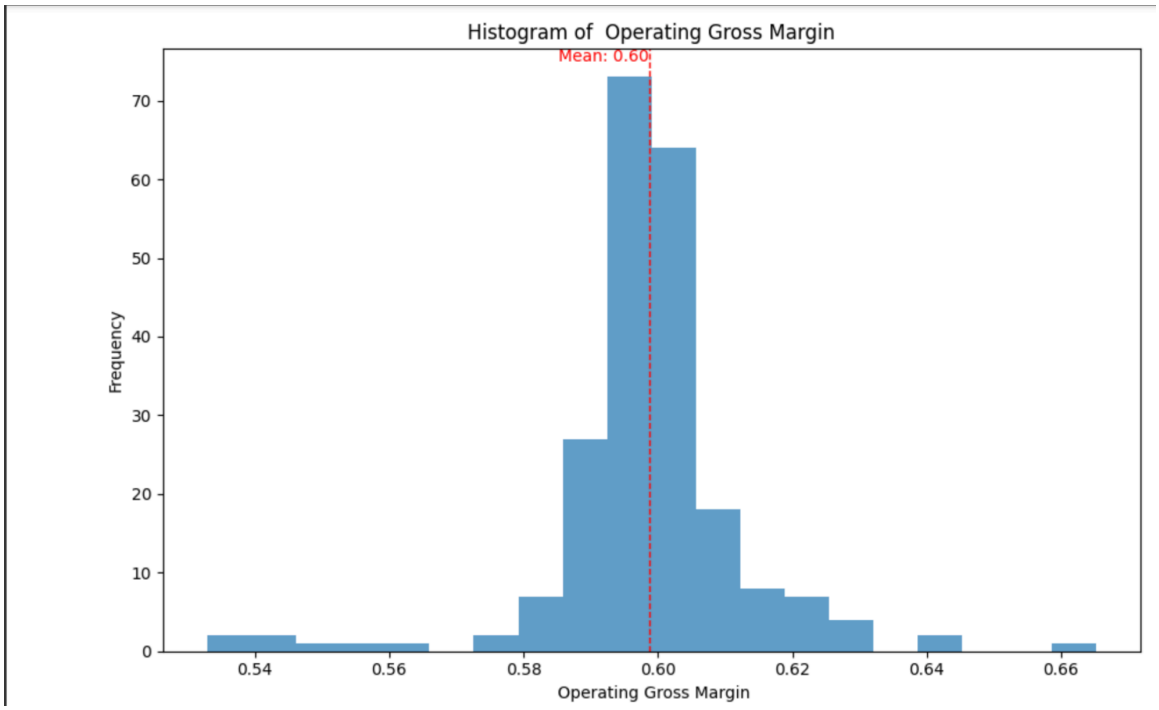
Company	prediction
Apple Inc.	0.0

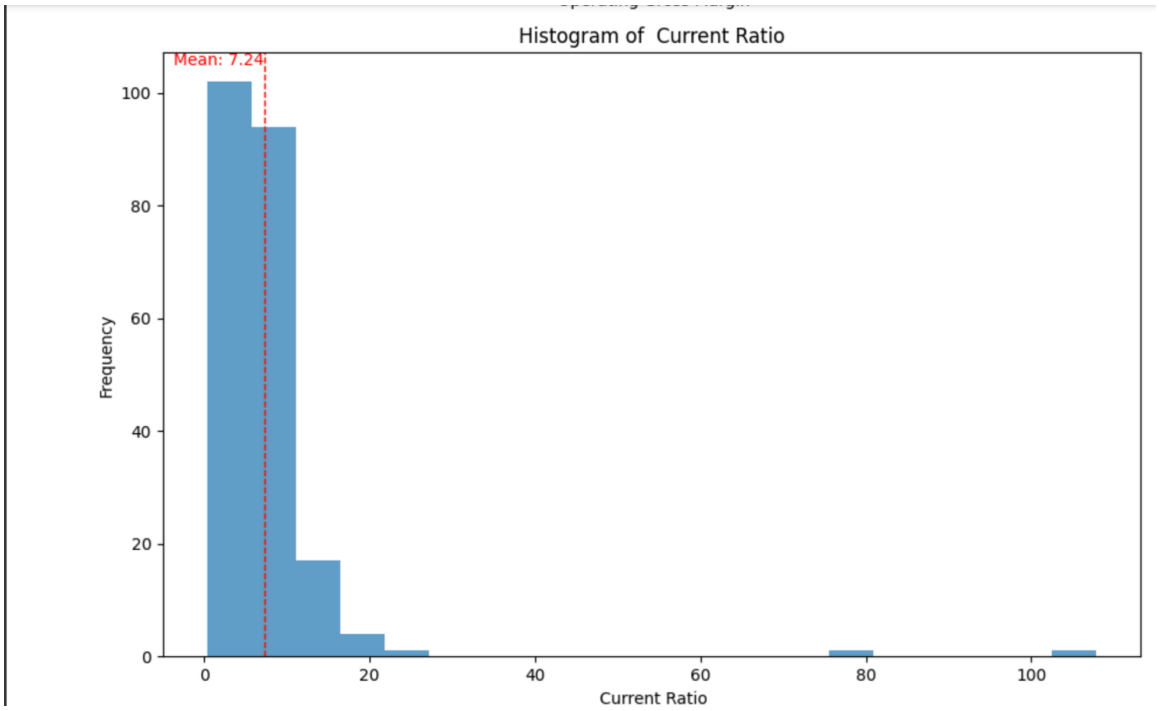
Company	prediction
Microsoft Corporation	0.0

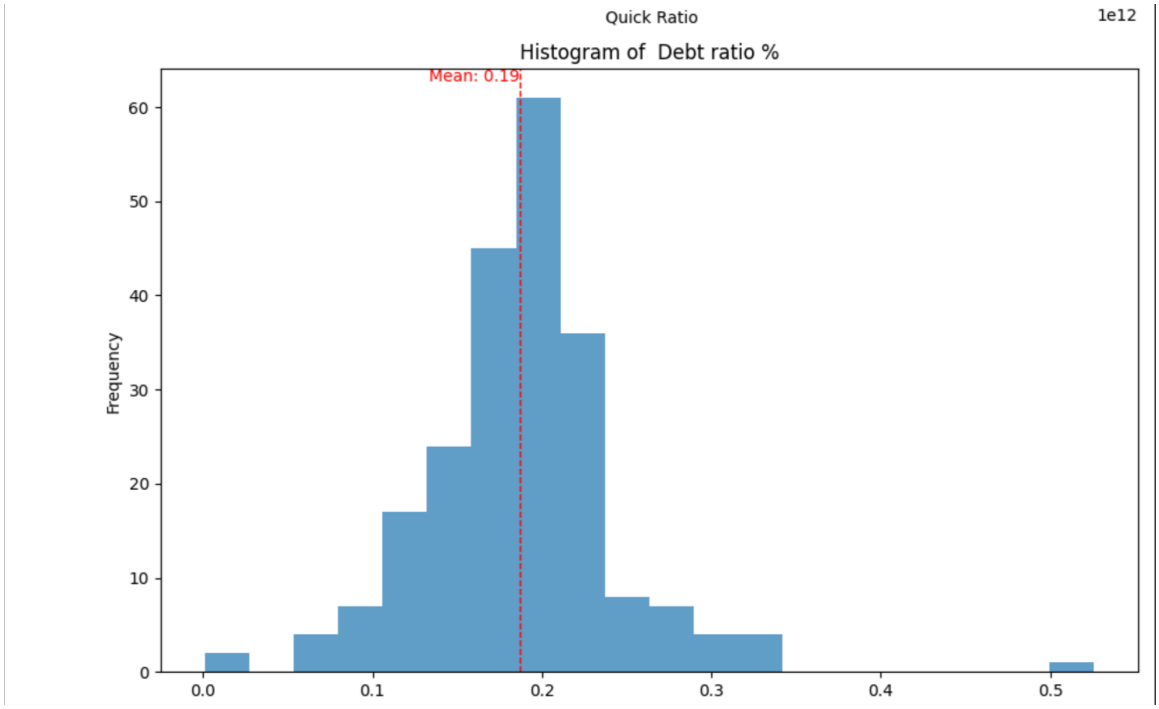
Company	prediction
NVIDIA Corporation	0.0

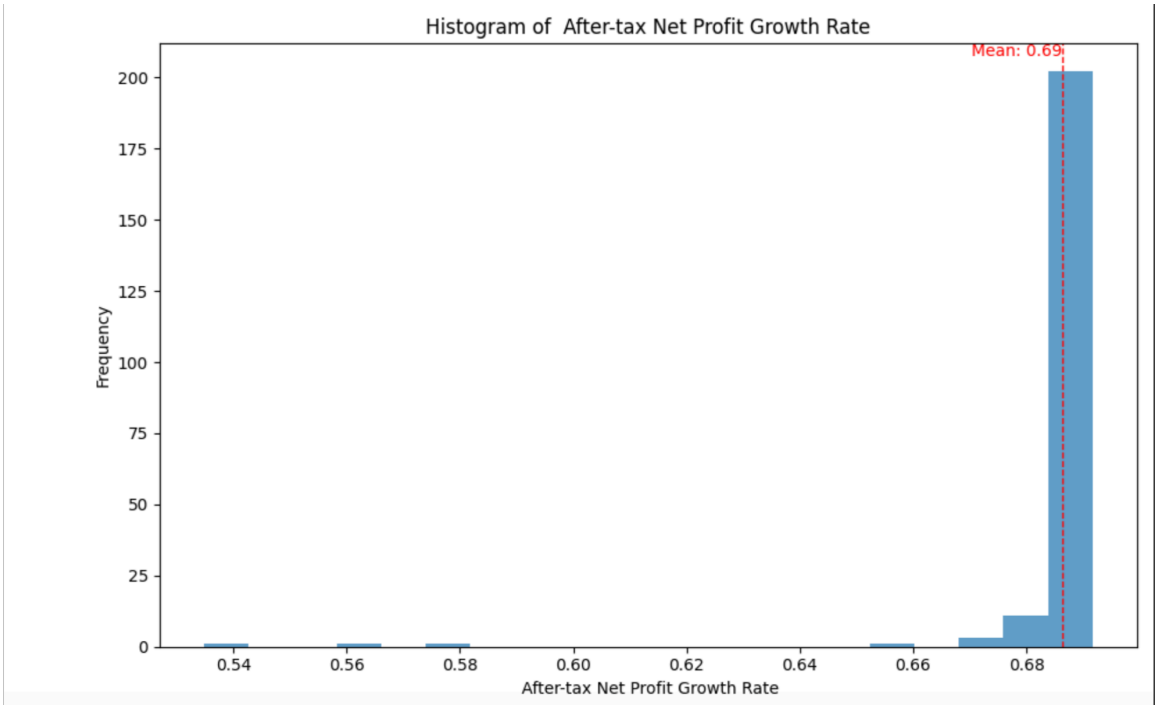
Company	prediction
Advanced Micro Devices, Inc.	1.0






Histogram of Current Ratio
Mean: 7.24
Frequency
Current Ratio
0
20
40
60
80
100


Quick Ratio
1e12
Histogram of Debt ratio %
Mean: 0.19
Frequency
0.0
0.1
0.2
0.3
0.4
0.5


Histogram of After-tax Net Profit Growth Rate
Mean: 0.69
Frequency
200
175
150
125
100
75
50
25
0
0.54
0.56
0.58
0.60
0.62
0.64
0.66
0.68
After-tax Net Profit Growth Rate