Stock evalulation analysis

(COMP3125 Individual Project)

Teck Zin Tan  
*dept. data and computer science*

# Introduction (*Heading 1*)

In today’s fast-paced market, new brokerage apps have made it easier than ever for the average person to invest in stocks. While this democratization of investing is empowering, it has also contributed to increased market volatility. Identifying undervalued companies early can be highly valuable for achieving strong returns. As a result, many hedge funds and professional investors rely on advanced algorithms and large datasets to uncover these opportunities often using data that’s out of reach for the average investor, especially for someone like me, a student with limited resources.

However, I believe that even with just a few key publicly available data points, we can still evaluate a company’s stock effectively and identify if it’s undervalued. Thanks to the growing availability of open data sources, platforms like Yahoo Finance and Interactive Brokers (IBKR) now offer APIs that allow individuals to collect and analyze market data. Additionally, access to historical datasets can be found on Kaggle and other website, enables back testing of strategies, allowing us to refine algorithms and improve accuracy over time. This project aims to explore that potential using simple, accessible metrics and API data, and even the possibility of the ROI will beat the market index like the S&P 500 or NASDAQ 100.

# Datasets

## Source of dataset (Heading 2)

## The dataset sourced from Kaggle, along with data retrieved through APIs such as Yahoo Finance, provides the foundation for analyzing stock valuation—specifically whether a stock is overvalued or undervalued. The key financial metrics used in this analysis include the price-to-earnings (P/E) ratio, price/earnings-to-growth (PEG) ratio, price-to-sales (P/S) ratio, and debt-to-equity (D/E) ratio. Additional indicators such as three-year and five-year profit growth rates, as well as five-year cash flow yield, are also incorporated. These metrics are essential in evaluating a company's financial performance and long-term potential.

## Before using any dataset, it is important to verify that it is reliable and accurately reflects both historical and current market conditions. For this reason, all Kaggle datasets considered for this project are reviewed for credibility, source transparency, and data freshness to ensure the analysis remains valid and trustworthy.

## Character of the datasets

The stock data used in this project includes both static datasets from Kaggle and dynamic data fetched via the Yahoo Finance API (yfinance). The Kaggle datasets provide foundational information such as historical high, low, open, close, and volume prices for individual stocks. These are useful for time series analysis and trend detection. To supplement this, Yahoo Finance’s API is used to fetch additional financial metrics that are not available in the Kaggle dataset such as the PEG ratio, debt-to-equity (D/E) ratio, and other valuation indicators.

The Kaggle dataset is generally clean and well-structured, requiring minimal preprocessing. Most of the values are ready for use in model training and evaluation, as they come from a trusted and curated source. For the Yahoo Finance data, selective querying allows for precise control over which fields are retrieved, reducing the chance of noise or irrelevant information. Basic cleaning steps, such as standardizing data formats and handling any missing or null values, are applied as needed to ensure consistency between the two data sources.

The Kaggle dataset I will be using will be one for S&P 500 one for NASDAQ 500, and the last one for all the stocks listed in the NYSE for a few years.

The column from Kaggle of the S&P 500 and NASDAQ will include open, high, low, close, volume, avg\_vol\_20d and change percent. This will need some clean to remove avg\_vol\_20d. While the other dataset for all the stocks will have the column of date, open, high, low, close, volume and Name, this won’t need any cleaning.

To complement the static data from Kaggle, I will use the Yahoo Finance API (yfinance) to fetch additional key financial variables that are not included in the original dataset. These include the PEG ratio (price/earnings-to-growth), debt-to-equity (D/E) ratio, price-to-sales (P/S) ratio, and updated P/E ratio values. These variables provide deeper insight into a company's valuation, financial health, and long-term growth potential. Using yfinance allows me to retrieve the most recent data directly from Yahoo Finance and integrate it seamlessly with the historical dataset for a more complete analysis.

By combining both static and dynamic data, the model benefits from a rich set of features while maintaining flexibility and up-to-date accuracy.

# Methodology

To determine whether a stock is overvalued or undervalued, we use earnings surprise—the difference between a company's actual earnings per share (EPS) and the expected EPS from analyst forecasts. If a company beats expectations (i.e., actual EPS > expected EPS), it is labeled as undervalued (1). If it misses or meets expectations (actual EPS less than expected EPS), it is labeled as overvalued (0). This label serves as the target variable in our classification models.

The earnings surprise formula is:

Earnings Surprise = Actual EPS − Expected EPS

We then apply two supervised machine learning methods: Logistic Regression and Linear Regression, each used for different purposes but both leveraging financial metrics such as P/E, PEG, D/E, P/S ratios, profit growth, and cash flow yield as input features.

## Method A: Logistic Model (Classification Model)

Logistic Regression is a statistical method for binary classification.

Purpose:

To classify stock if it overvalues (1) or undervalue (0).

Assumptions:

### P(y=1∣x): probability that the stock is undervalued

### : input features (e.g., P/E, PEG, D/E, etc.)

### ​: intercept (bias term)

### , : learned model coefficients

Equations:

= P( = 1) = probability of no O-ring damage at temperature

= E [ |]

## Mehtod B: Linear Regression Model

Purpose:

To predict the EPS surprise as a continuous variable using financial ratios.

Formula:

Where:

### is the predicted EPS surprise

### ​ are coefficients learned by the model

### ϵ: is the error term

# Results

In this section, present your findings using an appropriate method, such as equations, numerical summaries, or visualizations like charts and graphs. Clearly explain all results and provide guidance on how to interpret them. If any unexpected results arise, discuss possible reasons or contributing factors. To improve clarity and organization, consider using subsections (e.g., A, B) to separate different aspects of your results.

My knowledge from the advanced statistics course, I will analyze the model by comparing the predictors and the R-squared value for the linear regression model, as well as the accuracy for the logistic regression model. I will also calculate the p-values for each explanatory variable to determine their statistical significance. If needed, I will use a Chi-square test to compare both models and assess which one better fits the dataset. Additionally, I may use measures like Somers’ D test to evaluate the strength of association in the model, as an alternative to relying solely on p-values. Below I will discuss the result I found for both model (method A and B).

Data collected

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Ticker** | **Current Price** | **PE Ratio** | **PEG Ratio** | **DE Ratio** | **PS Ratio** | **Profit Growth** | **Cash Flow Yield** | **EPS Actual** | **EPS Expected** | **EPS Surprise** | **EPS Surprise %** |
| **AAPL** | 202.38 | 30.75684 | 394.318462 | 146.994 | 7.54986 | 0.048 | 0.032174 | 1.65 | 1.63 | 0.02 | 1.226994 |
| **MSFT** | 524.11 | 38.36823 | 161.891266 | 32.661 | 13.828423 | 0.236 | 0.015607 | 3.46 | 3.22 | 0.24 | 7.453416 |
| **GOOGL** | 189.13 | 20.163113 | 0.417547 | 11.481 | 6.17256 | 0.194 | 0.021717 | 2.27 | 2.01 | 0.26 | 12.935323 |
| **AMZN** | 214.75 | 32.736282 | 52.630678 | 51.641 | 3.505807 | 0.642 | 0.017226 | 1.59 | 1.36 | 0.23 | 16.911765 |
| **TSLA** | 302.63 | 182.30724 |  | 16.823 | 10.527588 | -0.163 | 0.001372 | 0.27 | 0.39 | -0.12 | -30.769231 |

FIgure 1


A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

## Result A

To evaluate which stocks are undervalued, a logistic regression model is used. The model produces a classification result of either **1** or **0**, indicating whether a stock is undervalued or overvalued, respectively. This classification is based on the **EPS (Earnings Per Share)** and whether the stock met or missed earnings expectations, which helps determine its valuation in a given period.

### Performance Metrics:

#### The model achieved an overall accuracy of 66.7% on the test set.

#### The precision and recall scores were balanced across classes, each with an F1-score of 0.67, despite the small sample size.

#### Confusion matrix results: the model correctly predicted 1 undervalued stock and misclassified 1 overvalued stock.

### Feature Coefficients:

#### The most influential predictor was the Debt-to-Equity Ratio (–0.8264), followed by Profit Growth (–0.4152) and Cash Flow Yield (–0.4130), all negatively correlated with undervaluation.

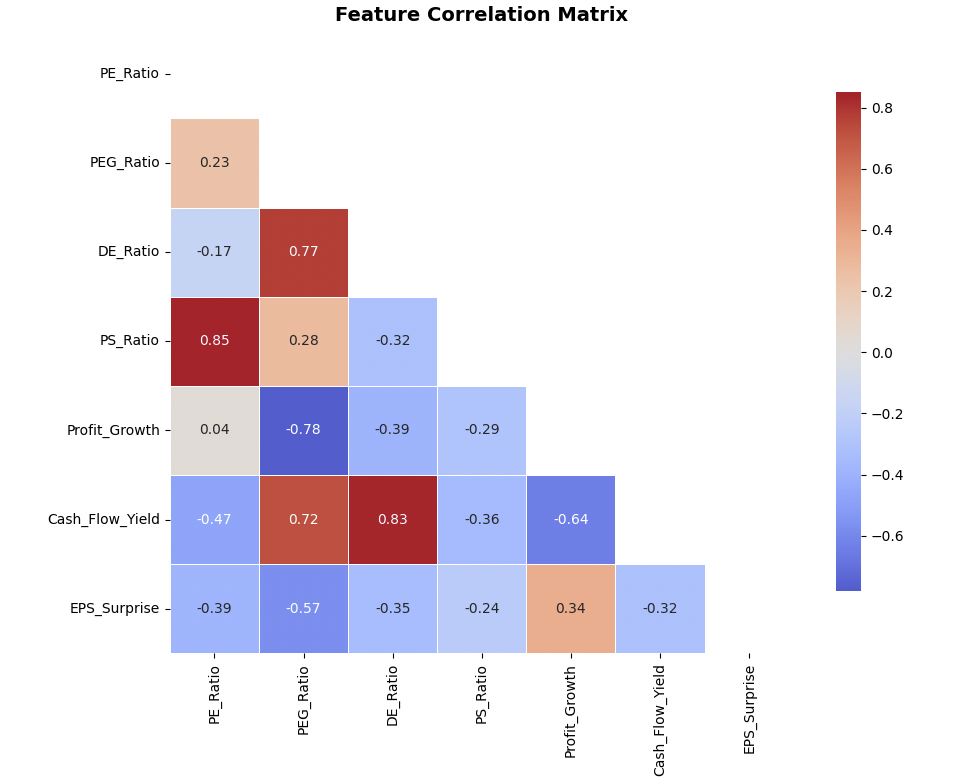
#### The only feature with a positive coefficient was PS Ratio (0.1313), though its magnitude was lower.

### Prediction Insights:

#### Stocks such as GOOGL (0.805), NVDA (0.857), META (0.720), and MSFT (0.690) had the highest probabilities of being undervalued.

#### AAPL (0.005) and V (0.223) were predicted to be overvalued with high certainty.

#### The classification largely aligned with actual earnings surprises, suggesting that the model captured some real valuation signals.



The correlation matrix shows strong multicollinearity between some explanatory variables. To improve the linear regression model and reduce redundancy, highly correlated variables such as P/E Ratio and P/S Ratio (r = 0.85)

A screenshot of a graph

AI-generated content may be incorrect.

Shows which stock is undervalue or overvalue, MSFT, GOOG, NVDA and META are all under value right now with the recent EPS.

A screen shot of a black screen

AI-generated content may be incorrect.

## A screenshot of a computer AI-generated content may be incorrect.

## B. Results B

## A linear regression model was used to predict the EPS Surprise as a continuous variable using the same set of predictors.

### Performance:

#### The model had a negative R² of –0.211, indicating that the model performs worse than a baseline that always predicts the mean EPS surprise.

#### This implies that the linear regression approach failed to capture meaningful patterns in the data.

### Feature Impact:

#### All feature coefficients were small in magnitude, with Profit Growth and Cash Flow Yield being the only two features with positive influence on EPS surprise.

#### Features such as P/E Ratio, PEG Ratio, PS Ratio, and D/E Ratio had negative coefficients, suggesting slight inverse relationships.

### Model Diagnostics:

#### The residuals were widely spread and inconsistent, especially for META, which had a large prediction error.

#### The predicted EPS surprises were clustered close to zero, showing the model’s inability to capture variance across stocks.

A graph and diagram with red and green squares

AI-generated content may be incorrect.

As you can see we are able to predict most of the stock with the EPS an earnings report, however the R squared is still way too low for any real interpretation to take from this.

# Discussion

From this, we can conclude that the logistic regression model is significantly more accurate while also being less complex—an advantage further supported by the Pairwise Test discussed earlier.

Like all models, the approaches used in this project had their limitations. The logistic regression model, while achieving a moderate accuracy of 66.7%, was trained and tested on a very small dataset. This limited sample size can introduce bias, reduce generalizability, and lead to unstable predictions. Additionally, some stocks were misclassified such as AMZN highlighting the model's difficulty in handling borderline cases or complex financial profiles.

# The linear regression model performed poorly, with an R² value of –0.211, suggesting it failed to capture the underlying relationship between the selected features and EPS surprise. This indicates that linear assumptions may not be suitable for modeling financial outcomes like earnings surprise, which are likely influenced by nonlinear and external market factors.

# For future improvements, a larger dataset with more companies and historical data across multiple quarters would likely improve both model accuracy and robustness. Furthermore, incorporating more variables to better determine the financial characteristics of a stock would enhance the model's predictive accuracy. Statistical techniques such as the Wald test can be used to assess the significance of individual predictors, while pairwise comparison tests and the Chi-square test can help identify associations between categorical variables and classification outcomes. Additionally, conducting model comparisons using a reduced vs. full model framework would allow for testing whether the inclusion of additional variables significantly improves model performance. These approaches would provide more rigorous validation of the explanatory variables used in the model.Conclusion

##### This project aimed to evaluate the undervaluation of stocks using both logistic and linear regression techniques. The logistic regression model showed moderate success in identifying undervalued stocks based on earnings surprise and financial ratios. In contrast, the linear regression model failed to effectively predict EPS surprises, suggesting the relationship is likely nonlinear.

##### The most important result was identifying D/E Ratio, Profit Growth, and Cash Flow Yield as the most influential factors in undervaluation classification. This insight can help investors focus on these metrics when screening for potentially undervalued companies.

##### In the real world, the findings demonstrate how data-driven models when paired with solid financial metrics—can offer valuable support in investment decision-making. However, they must be applied cautiously, considering data quality, model limitations, and the complexity of market .

##### References

1. [1] T. Bajaj, “Yahoo Finance All Stocks Dataset (Daily Update),” Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/tanavbajaj/yahoo-finance-all-stocks-dataset-daily-update
2. [2] A. Mvd, “S&P 500 Stocks,” Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks
3. [3] Yahoo Finance API, “Yahoo Finance API for Stock Market Data,” Python Library, Accessed: Aug. 2, 2025. [Online]. Available: https://pypi.org/project/yahoo-finance/
4. [4] Stock API (yfinance), “yfinance: Yahoo! Finance market data downloader,” Python Library, Accessed: Aug. 2, 2025. [Online]. Available: https://github.com/ranaroussi/yfinance