National College of Ireland

# Investigating the Application of Tree-Based Machine Learning Techniques to Predict the Margin of Safety in Potential Stock Investments

MSc Research Project
Data Analytics

Keith Scully
Student ID: X22186344

School of Computing
National College of Ireland

Supervisor: Mohammed Hasanuzzaman

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | |
|---|---|
| **Student Name:** | Keith Scully |
| **Student ID:** | X22186344 |
| **Programme:** | Data Analytics **Year:** 2024 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Mohammed Hasanuzzaman |
| **Submission Due Date:** | 12th December 2024 |
| **Project Title:** | Investigating the Application of Tree-Based Machine Learning Techniques to Predict the Margin of Safety in Potential Stock Investments |
| **Word Count:** | 8,386 **Page Count:** 21 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** 

**Date:** 12/12/2024

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Investigating the Application of Tree-Based Machine Learning Techniques to Predict the Margin of Safety in Potential Stock Investments

Keith Scully
X22186344

**Abstract**

The stock market provides generous conditions for increasing and preserving wealth over the long term. Value investing is a strategy focussed on using fundamental analysis to estimate the intrinsic value of a company, then seeking to take advantage of divergences between market price and intrinsic value to minimise downside risk and maximise upside potential. This mispricing delta is regularly referred to in the investing world as the "margin of safety". Valuing a business is a complex task requiring specialised knowledge and consequently is routinely neglected by many investors. This leads to buying decisions that are based more on intuition and market momentum rather than by informed decision-making, which is often at an increased risk of financial loss. This research proposes a novel approach to directly estimate the "margin of safety" available on US company stocks by leveraging machine learning techniques to predict the degree of mispricing based on current market prices and publicly available financial accounting data. A number of tree-based learning algorithms were selected and appropriate predictive models were developed. The final XGBoost model was found to be most performant with an RMSE of 8.7%. Using SHAP explainable AI techniques it was also determined that elevated levels of Free Cash Flow were strongly associated with enhanced margin of safety values. Further experimentation with the final model using unseen historical data revealed that a portfolio of stocks selected based on model outputs had significantly outperformed the broader stock market over a recent 10-year period. This study demonstrates that machine learning can be applied successfully to value investing approaches, through the selection of investments that are grounded on current price alongside fundamental business performance. The results are encouraging for retail investors who can benefit from this research by being more informed on the risk of investing given current market pricing, while simultaneously increasing the potential of positive investment returns into the future.

***Keywords*** – Value Investing, Fundamental Analysis, Margin of Safety, Machine Learning, Tree-Based Algorithms, Explainable AI.

# 1 Introduction

The stock market is utilised as a vehicle to preserve and appreciate wealth, with a variety of strategies employed based on investor preference. Growth investing generally follows high-growth companies. The price for stock in such companies can appear unreasonably high by many commonly-used valuation metrics, and the momentum of increasing price levels is often a strong indicator of investor expectations for future growth. It is often the case that such companies have not yet matured and may have little history in the context of financial performance, with many popular growth companies yet to produce a net profit. The lack of historical business performance data, along with the difficulties associated with predicting the

future of these companies, means that an approach based solely on growth is a high-risk path for investors. An alternative strategy is known as value investing, where investors study the fundamentals of a company to determine its financial health and to estimate an intrinsic value for the business before comparing this against the current market price for the stock. To understand this approach, it is necessary to recognise the distinction between price and value. The value of a company is commonly calculated as the present value of all estimated future cash flows of the business, while the price is simply the amount at which the stock is currently trading. The trading price often deviates from the value, in both positive and negative directions, and value investors commonly look for buying opportunities where the price falls below the company's estimated value, when it is deemed to be undervalued. This discount in price against value is known as the 'margin of safety', a term coined by Graham and Dodd (1934), further emphasised by Graham (1949), embraced by Buffett (1984) with phenomenal success, and discussed extensively by Klarman (1991) in his book of the same name. The goal of achieving a margin of safety on a stock investment is to allow some degree of protection from potential miscalculation of future value along with limiting downside risk during other market events. As a consequence, it also increases the chances of improved future returns as the market price re-adjusts over time to re-align with a business's value.

Machine Learning (ML), a subfield of Artificial Intelligence (AI), has been used extensively in academic literature at the intersection of stock markets and ML, but has predominantly concentrated on price prediction without reference or consideration of intrinsic value. Some studies have used ML for predicting business values but with varying degrees of success and most often using relative valuation techniques that value businesses relative to the value of their peers in the market. Whilst there is some merit in this approach, it allows increased risk to investors given the influence of broader market effects. In an overvalued market it can lead to buying already overvalued investments.

Modern online stockbroking platforms are accessible to almost anyone, but without the requisite knowledge of business accounting and valuation processes it is difficult for the average investor to make informed investment decisions. The novel approach taken in this research is to leverage the strengths of modern ML methods, combined with fundamental value investing approaches, to predict the margin of safety available on company stocks and thus promote low-risk, high-return stock selections. It is hoped that this study will also encourage further research into the application of ML techniques to intrinsic valuation, thereby expanding the tools available to investors that support sensible investment decisions.

## 1.1 Research Objective

In the context of the research problem detailed above, the research objective is concerned with the following question: How can tree-based machine learning techniques leverage financial and macroeconomic data to predict the margin of safety in stock investments? Secondary to this question it is also intended to identify and understand the most important data points that strongly impact successful predictions to better inform investment decisions.

## 1.2 Limitations and Assumptions

In order to proceed with the application of ML methods it will be necessary to engineer the target variable in advance by utilising a conventional financial model. The parameter values for this model can be highly subjective, often subject to the preferences of the appraiser. A major component of this model involves estimating future cash flows at an appropriate growth rate. As this project focusses on generalising across a large number of companies, rather than on a single company, it is essential to define some conditions around the selection of this growth rate so that it can be broadly applied. The growth rate chosen will consider the annualised growth rate from recent fiscal years, before applying a decay factor to

estimated future cash flows in order to reflect the normal business cycle and ensure that this financial model remains slightly conservative in relation to future estimates. An upper boundary will be introduced to ensure that companies who have experienced exceptionally high growth rates in recent years are not expected to maintain such levels of high growth in future years and instead revert to more reasonable levels of growth. Additionally, given that all of the companies being considered in the dataset are a mixture of growing and mature companies, a lower bound on growth rates will also be set that matches historical GDP rate. Finally, a limitation of this financial model is that it is not appropriate for use on all types of businesses. This is mostly relevant to industries within the financial sector, predominantly banks and insurance companies, meaning these businesses will be excluded from this project.

### 1.3 Report Structure

The following outlines the structure of this report. Section 2 critically surveys related literature from the same field of study. Section 3 presents the proposed research methodology, with detailed subsections on all key stages. The design specification is provided in Section 4. In Section 5 a comprehensive summary of the project implementation is exhibited, focussing on each of the main phases of activity. Section 6 reveals the results of data mining under the chosen evaluation metrics, while also explaining feature impact in the final model. A back-test on unseen historical data is also detailed here. Finally, Section 7 provides some concluding remarks along with recommendations on future work.

## 2 Literature Review

Research at the intersection of the stock markets and ML is plentiful, though significant levels of research have contributed to the area price prediction which is unrelated to this research. Instead, research relating to company valuation at this intersection is more limited, though some interesting research has been carried out at this juncture that is more related to this study. There are also numerous research articles in which significant capabilities of tree-based ML models are evident.

### 2.1 Machine Learning in Prediction of Business Valuation and Market Returns

Several research articles have been identified that directly relate to the research question in this study. These articles are critical to the understanding of the problem domain and to critically assess the strengths and weaknesses of the proposed solutions.

Huang, Capretz and Ho (2021) used fundamental analysis of stocks with a variety of ML algorithms to predict future returns relative to a market index, and construct portfolios of the most promising stocks. Researchers extracted 21 features from quarterly accounting information for 70 companies in the S&P500 index covering the years of 1996 to 2017, with a target variable being the returns on the stock relative to the Dow Jones Industrial Average (DJIA) index. Their proposed aggregate model outperformed the DJIA benchmark. This research was useful as it identified a number of accounting items that could be considered for the current study. Using returns relative to the market strengthened their approach as it removed factors impacting the broader market, however there are also some limitations present. Their methodology implies a relationship between the financial position of a company and the returns relative to the market, without any consideration of the intrinsic value of a business. It is also unclear why returns were measured relative to the DJIA rather than the S&P500, particularly given the S&P500 is considered as the predominant indicator

of the US stock market in general. The researchers also proposed the use of cross-validation in future work, a technique that perhaps should be been employed in the initial work.

Research by Huang (2022) involved a deep learning (DL) framework for prediction of the delta between estimated intrinsic value and market price, based on accounting data from financial statements for 90 semiconductor companies on the Taiwan Stock Exchange (TWSE) over an 11-year period. A neural network is used to extract common patterns found throughout the market before comparing this with feature values for individual companies to determine a differential. All data is fed through a LSTM recurrent neural network to produce estimates of price-value gaps. This study was particularly beneficial given the researcher had consulted with accounting experts to establish an extensive list of accounting items for inclusion, which was utilised in the current research. However, there are some drawbacks with this research. Companies are not valued on their own merits and instead computes price-value caps from relative valuations comparing one company to the rest of the market. The risk is that when the market in general is overpriced it increases the possibility of stocks being recommended even when they too are overpriced, thus the increasing risk of loss to investors. Additionally, the use of DL techniques complicates explainability issues with a model, which is critical in the financial sector, therefore minimising the potential use cases for this framework. The use of so-called “black box” methods such as DL models are also advised against in other literature (Buczynski, Cuzzolin and Sahakian, 2021) as the lack of explainability violates common investing regulations.

An alternative approach to estimation of price-value gaps was undertaken by Hanauer, Kononova and Rapp (2022) who used publicly-available accounting data from European stocks to predict the fair values of stocks on a monthly basis. The mispricing percentage is then calculated and used to predict relative returns. Their methodology is heavily influenced by other literature, using 21 accounting variables to predict fair values by means of various ML algorithms. The final ensemble method promises substantially higher returns when compared with the original study that influenced their work. This research was useful as it provided further supporting evidence of accounting data points commonly found to be useful as predictors in ML models. Furthermore, it supported the decision for using tree-base methods. Their approach however also retains the common weakness associated with relative valuation approaches.

Peng and Lee (2024) take a more traditional approach to valuation by using a DCF model in their framework. This is alongside well-regarded ML algorithms to classify stocks as either overvalued or undervalued. The DCF model takes Free Cash Flow (FCF) and Weighted Average Cost of Capital (WACC) as inputs, consistent with tradition, but the unique aspect of their study is the use of reinforcement learning to fit FCF chronologically to the Levy distribution in an attempt to remove appraiser bias from estimation of DCF inputs. Results indicated that this approach produced high returns on back tests, though there is a limitation relating to accurately fitting FCF growth rates over time to a common distribution, a difficulty acknowledged by the authors. For average retail investors, approaches based on generation of large portfolios that are regularly adjusted are unfeasible due to lower investment capital and the negative impact of transaction fees.

Yang, Liu and Wu, (2018) developed a stock recommendation system using 20 financial accounting items from S&P500 companies, along with stock price data, to predict future returns. Stocks were ranked before selecting the top 20% to buy, and the bottom 20% to sell. The target variable was the forward quarter return, and the researchers tested multiple ML algorithms which were evaluated on Mean Squared Error (MSE). The proposed system employed an equally-weight stock allocation strategy to portfolios developed, and generated an annualised 16.12% return against the 7.12% return observed from the S&P500 over the same period. A similar approach was followed by Tsai, Gao and Yuan, (2023) who predicted

stocks that will have the best returns over the next quarter by ranking predictions and building portfolios. However, in this instance they focussed on all companies that made up the TW50 benchmark in Taiwan. Their proposed method promised potential for 50% excess returns against the market. This study was another that benefited the present research due to identification of accounting items as model inputs. However, large portfolios that are adjusted at higher frequencies are impractical for the average retail investor. Additionally, any study using the forward price for determination of the target variable also results in greater emphasis on price rather than value, as price action is heavily influenced by broader market sentiment that's not consistent with a company's intrinsic value.

The prediction of future FCF growth rates is critical to fundamental valuation, as a business's value is commonly considered to the present value of future cash flows. Evdokimov, Kampouridis and Papastylianou (2023) experimented with multiple ML algorithms to predict future FCF growth rates based on accounting data. An ARIMA time-series model was formulated as a benchmark against which all other models were evaluated. It was interesting that a K-Nearest Neighbours (KNN) model provided the best results based on Root Mean Squared Error (RMSE). This paper was also interesting because the authors rightly acknowledge that there's a general consensus around the calculation and application of a discount rate in a DCF model, but not against the FCF growth rate. However, this research is limited in that only 100 companies were included. The models developed were also highly restricted. For example, the RF model was trained with only 5 trees. Limitation on such hyperparameters can assist in minimising potential overfitting, but there was no discussion on whether their model may have been negatively impacted by such severe limitations.

One final article on stock valuation using ML techniques was from Geertsema and Lu (2023) who focussed on relative valuation methods, seeking to predict how much over- or under-valued a stock was relative to its' peers in the market. Researchers comment that relative valuation can act as a good proxy to fundamental value, and conclude that predictions evaluated from their model indicate strong predictive ability with regard to future returns. It was refreshing to read this paper given the acknowledgement of the differences between relative valuation and fundamental valuation. Many other studies presented their work as fundamental valuation simply because it used accounting data. However, this research also contains the weaknesses of other studies using relative valuation, in the context of the potential risk for investors. It was also interesting that the authors suggest ML models should complement human judgement and intuition in the selection of stocks, something important to consider particularly for retail investors with smaller portfolio sizes who can less afford to place their money on potentially riskier stocks.

## 2.2 The Effectiveness of Tree-Based Models in Financial Markets

In research articles addressed already, some have reported promising performance from tree-based methods. Huang, Capretz and Ho (2021) experimented with Random Forest (RF), Feed-Forward Neural Network (FNN) and Adaptive Neural Fuzzy Inference System (ANFIS) techniques, and while they ultimately proposed an aggregate model, the RF algorithm outperformed more complex algorithms on an individual basis. Hanauer, Kononova and Rapp (2022) also found that RF and Gradient Boosted Regression Trees (GBRT) models performed significantly better than linear regression, and emphasised the importance of employing learning techniques that can facilitate the complex non-linear relationships between accounting variables in the fundamental analysis of stocks. Peng and Lee (2024) found value from the inclusion of RF in their prediction framework. Yang, Liu and Wu (2018) found their RF model outperformed all other models in their study, which also included a tree-based Gradient Boosting Machine (GBM). Geertsema and Lu (2023) echoed other researchers in

emphasising the importance of modelling complex relationships in accounting data, which they feel they were able to achieve quite well with their GBM model.

Khartik et al. (2023) studied the use of DL for price prediction in the form of a classification task. In addition to Multi-Layer Perceptron (MLP), the researchers also employed XGBoost, RF and Decision Tree (DT) algorithms and while the MLP outperformed all other methods, it was noted that the DT algorithm displayed strong performance with XGBoost and RF models also noted for their predictive accuracy. Authors noted however that XGBoost and RF were resource intensive and while this may be the case with considerably large datasets, it would be of little concern in the present study given the more moderate volumes of data. The use of DL for the present research is also neglected due to challenges associated with explainability.

To predict stock market bubbles in the S&P500 market index, Başoğlu Kabran and Ünlü (2021) employed various Support-Vector Machine (SVM) models in a classification task to estimate a binary class based on whether the market was in a bubble or not. They used financial time-series data alongside macroeconomic data such as GDP, unemployment rates, inflation rates and interest rates, seeking real-time predictions. The SVM model performed at extremely high levels of accuracy, and while the DT model performed slightly lower than the SVM and Artificial Neural Network (ANN) models, it achieved an accuracy score of 96.55%, further supporting the predictive power and versatility of tree-based methods.

## 2.3 Summary and Research Gaps

Throughout the relevant literature associated with this problem domain, a number of common themes appear. Much of the research focussed on business valuation using ML are using relative valuation strategies, rather than a fundamental valuation strategy. Risk in an inflated market is not minimised in an overinflated market just because a stock is valued below its' peers. This presents a research gap involving fundamental valuation of stocks, with businesses separated from both peers and the market, to be valued on their own merits where risk is better managed, a gap that this research intends to address. Some studies also make extensive use of "black box" DL techniques, which violate financial industry regulations, placing significant constraints on the use cases for these solutions. The importance of solution being capable of modelling complex, non-linear characteristics of financial accounting data is another theme in related research, which supports the proposed use of tree-based methods in this research.

Finally, through various articles there existed a common cautionary observation. Strader et al., (2020) reviewed research directions in the domain of stock market predictions and observed that success in the market was a zero-sum game for investors, warning that wide adoption of a reliable solution would result in a scenario where no investor is better off. Khoa and Huynh (2023) echoed this in stating that improving the application of ML may result in the reduced future returns, while Buczynski, Cuzzolin and Sahakian (2021) suggest that the proliferation of algorithms that currently indicate unparalleled returns would eventually lead to decreasing future returns as value opportunities become harder to find given reduced levels of mispricing in the market. While these observations do not necessarily support any particular algorithm or approach, they do present a reasonable cautioning on the potential impact of ML in this particular problem domain.

# 3 Research Methodology

This research was executed in accordance with established KDD (Knowledge Discovery in Databases), which is ideally suited to studies involving data modelling and prediction. It consists of a series of dependent phases for data collection, exploratory data analysis (EDA)

and data pre-processing, transformation of data to prepare for application of ML algorithms, the development of ML models, the evaluation of model performance according to appropriate metrics, and finally the interpretation of the results of the study. The standard flow of activities through this methodology is shown in the figure below, with each phase detailed in the following subsections.


**Figure 1: KDD Research Methodology.**

## 3.1 Data Selection

It was identified early in the process the lack of suitable and freely-available datasets, requiring that a bespoke dataset be produced, which was a considerable undertaking. The Tickerdata[1] service was identified as a suitable data source for company financial accounting items, given the extensive history of financial data and wide-range of attributes for retrieval. Tickerdata provides most of the accounting items commonly seen in Income Statements, Balance Sheets and Cash Flow Statements from company filings, along with other metrics on business performance, management efficiency, and valuation ratios. It operates via an add-on to Google Sheets. Through astute spreadsheet preparation it was possible to collect a large amount of data at any given time.

In determining potentially valuable accounting items to retrieve data for, the related studies by Huang (2022), Yang, Liu and Wu (2018), Tsai, Gao and Yuan, (2023), Koklev (2022), and Quah (2008) were referred to. The collected data was also defined by limitations with the data provider, and informed judgement following a general study of the subject topic. Given the requirements for calculating the target variable, additional data points were collected. This included US Treasury Bond rates, commonly referred to as the Risk-Free Rate, which were obtained from the Wall Street Journal market data[2], as well as a figure

[1] https://tickerdata.com/
[2] https://www.wsj.com/market-data/quotes/bond/BX/TMUBMUSD10Y/historical-prices

representing the average historical market rate of return. This rate was calculated from data made available by Aswath Damodaran[3], and as of the end of 2023 the 50-year average market return was found to be 8%. This enabled the calculation of two more important variables: the Equity Risk Premium, which is essentially the difference between the US bond rate and the average market return at the time; and the Weighted Average Cost of Capital (WACC), which uses multiple variables to determine the average cost of debt and equity to a company. Additional 3-year Compound Annual Growth Rates (CAGRs) were calculated for revenue, operating income, various cash flow measures and outstanding share counts.

This research is focussed on the US stock market, where the S&P 500 index is often used to measure the performance of the US market in general. A list of companies to include in the project dataset was compiled. This consisted of all companies in the S&P 500 index, which were obtained from the Stock Analysis website[4] before being supplemented with a small number of additional large US companies not currently part of this index. These companies represent a very wide range of industries and sectors. For each company, accounting data was retrieved for the ten most recent fiscal years, and from full fiscal year accounts. As it is typically the case that stock price will vary throughout any given year, even with little variance in business performance, a number of pricing samples were taken for each fiscal year. Stock prices were also retrieved in Google Sheets through the dedicated function for Google Finance[5].

## 3.2 Pre-Processing

The collected data was loaded into a Python environment using Jupyter Lab. The pre-processing phase typically involves removal of unwanted data, treatment of missing values and outliers, etc., along with EDA. In this study the pre-processing phase also required considerable effort due to required calculation of the target variable. This was achieved using a Discounted Cash Flow (DCF) model. A DCF model estimates the present value of all predicted future cash flows. Because some financial services firms, such as banks and insurance companies, are generally not suitable for the DCF method due to their complex cash flow structures, 49 companies were removed from the prepared dataset. Analysis of missing values showed 843 rows missing stock price data, due to the lack of price history. As pricing data is critical to calculation of the target variable these rows were dropped.

An intrinsic valuation was completed for each observation using the DCF model. The DCF requires a suitable growth rate, a suitable discount rate for discounting future cash flows to present value, and the number of years to project into the future. A simple DCF model can be seen in the formula below, with growth rates necessary for estimating cash flow in each future period and discount rate necessary to reduce these to present value:

$$DCF = \frac{CF_1}{(1+r)^2} + \frac{CF_2}{(1+r)^2} + \cdots + \frac{CF_n}{(1+r)^2}$$

A starting Free Cash Flow amount matches the current accounting year, with current year values for Cash and Short-Term Investments as well Total Debt accounted for in the final step of the valuation. The growth rate used was the 3-year Revenue CAGR, a rate deemed to be most consistent and predictable having carried out some analysis including other cash flows measures such as Operating Income, Operating Cash Flow and Free Cash Flow. This growth rate was also controlled in various ways. Many companies, particular less mature

[3] https://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/histretSP.html
[4] https://stockanalysis.com/list/sp-500-stocks/
[5] https://support.google.com/docs/answer/3093281?hl=en

high-growth companies, experience extremely high levels of revenue growth in early years which stabilised later at more reasonable levels of growth. To remain conservative, and ensure that excessively high growth rates were not assumed to continue indefinitely, an upper bound was established. Outlier analysis of revenue growth rates indicated an upper outlier threshold of 28.85%, and for companies who experienced growth rates beyond this threshold it was found that the median growth rate for the following year was 18%. This 18% rate was used as the starting growth rate for future year projections in such cases. To further ensure the model remained conservative, a nominal decay factor of 0.9 was applied annually to gradually reduce the growth rates over time in the projections. This would mimic to some degree the normal business lifecycle where growth rates deteriorate over time. A final control measure was to set a lower bound for growth rates, to handle instances where companies suffer an occasional drop in revenue, which was set at 2.5%, the average rate of US GDP growth over the last 30 years according to the Federal Reserve Bank of St. Louis[6], a well-known source for US economic data. The discount rate also had controls applied. The WACC value is commonly used as a discount rate for DCF models, but some investors apply discount rates consistent with their desired rate of return while seeking to ideally outperform the market. Using the WACC initially, but setting a lower bound equal to the market's average rate of return and an upper bound equal to the upper threshold from outlier analysis on WACC values, which ensured all discount rates were between 8% and 12.7%.

It was also decided to include a rate of change for outstanding share counts. Public companies typically increase share counts through additional share offerings and stock-based compensation, and reduce outstanding share counts by buying back shares. A 3-year CAGR was calculated on each observation to determine this rate of change. The model was also set to project five years into the future. Once an estimated intrinsic value was determined for each observation, the target variable was calculated. This was the percentage difference between the current stock price and the DCF value, where a positive value indicates the stock is trading at a discount to its estimated value. The DCF modelling and calculation of the target was all carried out in a custom Python function, and once the target variable was generated the estimated intrinsic value was dropped to ensure ML models were not exposed to this value.

EDA is a critical phase in any data mining project, necessary to gain better understanding of the data. All features of the data were numerical. A small number of missing values were either calculated explicitly where possible, or alternatively imputed using the KNN Imputer from Scikit-Learn with a small value for $k$ ($k$=3) to ensure imputed values remained consistent with other very similar samples. A small number of rows without a 3-year CAGR value were dropped, due to the absence of at least 4 years in revenue data. The remaining dataset contained 10,146 rows and 53 predictors, along with the target variable. The ydata_profiling library was used to generate a report of descriptive statistic on each variable, with further analysis carried out with Matplotlib and Seaborn data visualisation libraries. Given the inherent nature of the data, many features displayed considerable numbers of outliers. A boxplot and histogram for Revenue is shown in Figure 2, with similar characteristics present in other variables. Very large companies in the data, such as Amazon, Apple, Alphabet (Google), Microsoft, etc., have revenues, cash flows, capital expenditures, etc., considerably larger than regular smaller companies.

Correlation analysis of all variables in the data indicated that there was little or no correlation between all data points, including between any given predictor and the target variable. A correlation matrix with heatmap is show in Figure 3.

[6] https://fred.stlouisfed.org/series/A191RL1A225NBEA

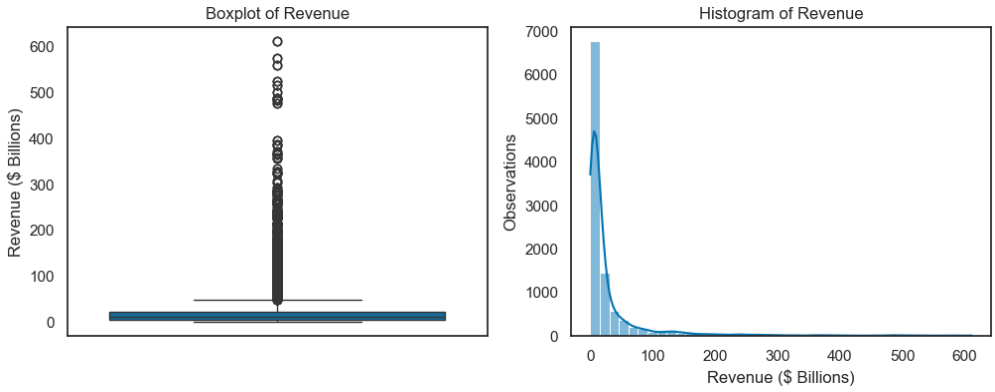


**Figure 2: Boxplot and Histogram of Revenue.**


**Figure 3: Correlation Matrix Plot of the processed dataset.**

Exploration of the target variable identified some extreme in both directions. Values greater than 1 indicate a 100% discount, deemed unreasonable. Values less than -2 indicate the stock was so overvalued it was trading at three times its value, which is possible but still quite extreme. Therefore, outliers on the target variable were treated, capping at 1 on the positive side, and on the negative side capping at the outlier threshold of 1.5 times the Inter-Quartile Range (IQR), which was -1.5. This brought 7.9% of observations in line with more reasonable values. A boxplot and histogram in Figure 4 show the distribution of the target variable after outlier treatment.

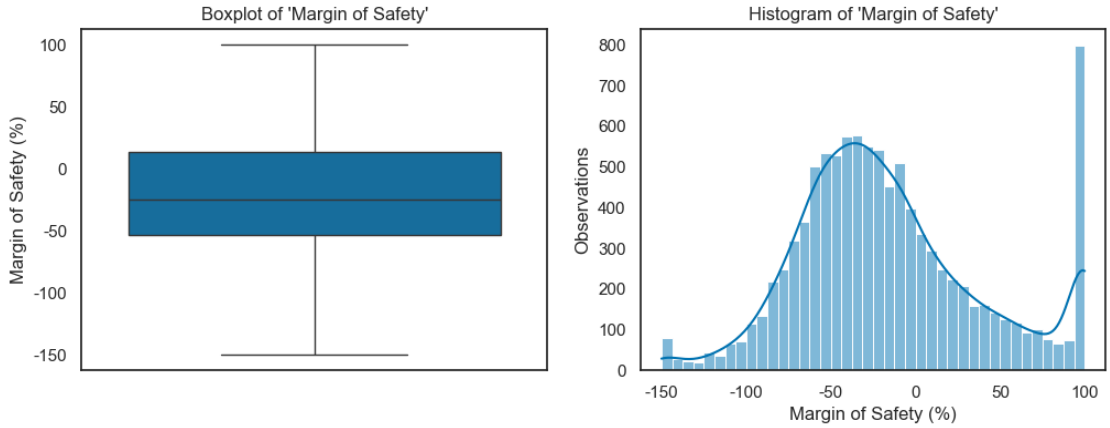


**Figure 4: Boxplot and Histogram of the target variable following treatment.**

## 3.3 Transformation

Data transformation involves preparing the data for input to ML algorithms. All numerical features were used for model training, while all categorical variables such as ticker symbol, company name, sector and industry, formed a multi-level index on the dataframe. Therefore, it was not necessary to dummy encode any input features. The possibility of treating outliers was considered, either by removing or capping observations, but it was ultimately decided to retain these features without any treatment as they were considered genuine values which could be represented in model training. All inputs were scaled using the robust scaling technique, which is more robust to outliers than other scaling methods.

## 3.4 Data Mining

Data mining takes place to make predictions and discover patterns and relationships in large datasets. Tree-based methods were used in this research due to their strengths in estimating non-linear functions, a characteristic of the prepared training data. The chosen supervised ML algorithms were as follows:

- **Random Forest Regressor[7]:** An ensemble algorithm that builds decision trees using the bagging technique. Randomised samples of data records and input features help prevent overfitting. This model was also used for feature selection and as an initial baseline model.
- **AdaBoost Regressor[8]:** This algorithm builds decision trees in series. Samples weights start with a default weight relative to the number of samples present, with adjustment made with each new tree. Incorrect predictions result in increased weights to concentrate subsequent trees on correcting those errors.
- **XGBoost Regressor[9]:** Uses a more extreme variation of gradient boosting, training weak learners in series. The first tree starts with a simple mean for each feature and each subsequent tree in the execution of the algorithm is focussed on predicting the errors of the previous tree. This algorithm is well-known for its accuracy. The SciKit-Learn API version was used due to the flexibility with other features of this library.

[7] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
[8] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html
[9] https://xgboost.readthedocs.io/en/stable/python/python_api.html

- **LightGBM Regressor[10]:** Similar to XGBoost, but utilises an alternative form of tree growth. It uses leaf-wise tree growth where only one leaf from any tree will be grown. It's extremely fast, and less memory intensive compared with XGBoost.

Feature selection was accomplished by extracting feature importance values, and permutation importance values, from the initial trained RF model. Feature importances are impurity-based and measure reduction in uncertainty associated with each feature, while permutation importances measure the significance of features by changing values to understand how model performance degrades, with larger decreases performance indicating greater importance. Selected features were the union of all features with feature importance or permutation importance values greater than or equal to 0.1. This resulted in a selection of ten features, shown in Figure 5 with their feature importance values.

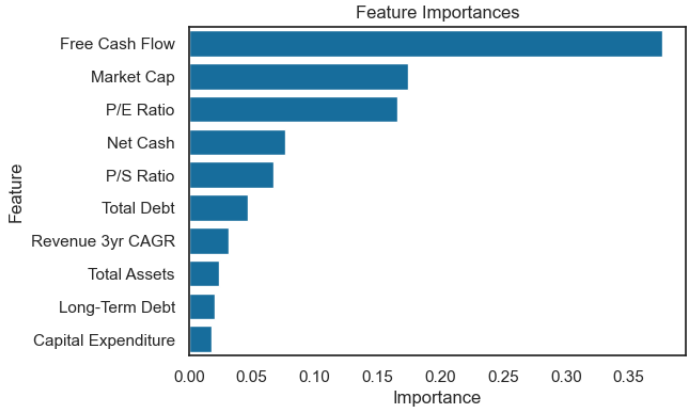


**Figure 5: Feature importance scores for all selected features.**

The data was split into train and test sets, using a test set size of 20%, with a customised train-test split function to ensure a reflective distribution of samples from each sector and company in both training and testing datasets. This split is show in Figure 6 below.

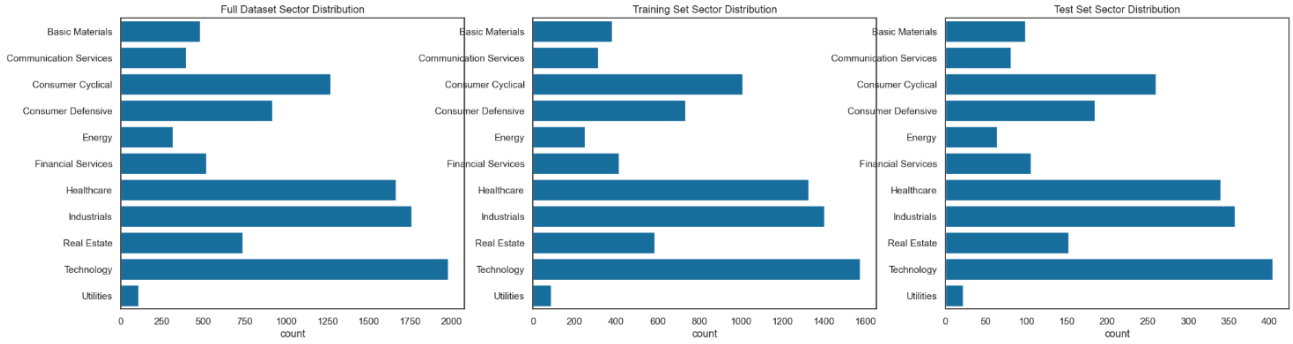


**Figure 6: Sample distribution by sector following customised train-test split.**

All chosen algorithms were trained using a grid search for hyperparameter tuning along with 5-fold cross-validation. The final models for each algorithm were compared with one another for best performance on test data.

[10] https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMRegressor.html

## 3.5 Evaluation

Evaluation of all model was carried out using common regression metrics to determine which model provided the best fit and most accurate predictions based on test data.

- **Root Mean Squared Error (RMSE):** RMSE provides the average error between actual and predicted values. A key benefit is that its easily interpreted given this measure is provided in the same units as the target variable. It's also known for outlier sensitivity, therefore inherently penalising larger errors. This metric can be calculated using the formula show below:

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}}$$

- **Mean Absolute Error (MAE)**: MAE is the mean of absolute errors in predictions. Like RMSE, error values are also on the same scale as the target variable. MAE can be calculated using the formula below:

$$MAE = \frac{\sum| y_i - \hat{y}_i |}{n}$$

- **R-Squared ($R^2$)**: Also known as the coefficient of determination, $R^2$ indicates the proportion of the dependent variable that can be explained by the independent variables, on a scale of 0-100%. It provides an estimate of the goodness of fit but doesn't support understanding the precision of error. It may indicate of how well unseen data may be predicted by the model, using the formula below:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Models were evaluated primarily by RMSE, with MAE and $R^2$ used as secondary metrics. Further analysis was then completed to better understand model fit using learning curve plots to ensure the chosen model was not overfitting the training data. For explainability purposes, SHAP (SHapely Additive exPlanations) values were calculated and plots generated to comprehend the degree to which model predictions were impacted by input features, which would help fulfil the secondary research objective.

# 4 Design Specification

This research was completed using a combination of tools, initially the aforementioned Google Sheets API for data collection, then primarily using the Python programming language in the Jupyter Lab environment for all data preparation and model development. The core Python libraries used were: Pandas, for data loading, cleaning, EDA, feature and target engineering, and data transformations; Ydata-profiling, Matplotlib and Seaborn for EDA and other plotting tasks; SciKit-Learn, for RF and AdaBoost regressor algorithms as well as functions for missing value imputation, scaling, hyperparameter tuning, pipeline creation and evaluation metrics; and SHAP model explainability. The standard Python pickle library was also used to store prepared dataframes and models for later use.

Having selected the final model, a pipeline was developed to include both fitted scaling transformer and the trained model, which could be leverage on new data. This pipeline, which is demonstrated in Figure 7, could also be utilised in model deployment.

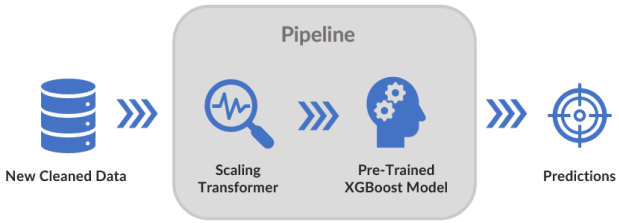


**Figure 7: Prediction pipeline.**

# 5 Implementation

The principal outputs from this study are the final compiled dataset and the developed models, particularly the best-performing model, both of which are detailed in this section.

## 5.1 Project Dataset

The project dataset was a key output from this study. The dataset is unique and the preparation of this dataset was a considerable undertaking. Given that this dataset was generated due to the lack of a similar publicly-available dataset, I have submitted a cleansed version of this dataset to the Kaggle platform[11] under the Creative Commons licence CC0 to benefit other studies and hopefully contribute to ML research in this topic area.

## 5.2 Models Developed

Four ML models were developed before comparing all of them to determine the final selected model. Details of training hyperparameter, along with some contextual information on system specifications, are provided in the following subsections.

### 5.2.1 Hardware and Software Specifications

All data collection, preparation and model training took place on a personal computer with system specifications detailed in Table 1. Without a suitable GPU, all models were developed on the local CPU.

**Table 1: Hardware & Software Specifications**

| | |
|---|---|
| **OS** | Windows 11 |
| **CPU** | 12th Gen Intel Core i5-12400 |
| **GPU** | n/a |
| **RAM** | 32GB |
| **Python** | V3.11.5 |

### 5.2.2 Model Training

Once the data was transformed and split into train and test sets, each algorithm was trained using a hyperparameter grid search to optimise the model, along with 5-fold cross-validation. Due to compute resource constraints, only hyperparameters considered most impactful were searched.

#### 5.2.2.1 Random Forest Algorithm

The RF algorithm was trained with tuning of n_estimators, max_dept & max_features hyperparameters. The most performant model was found with n_estimators=500,

[11] https://www.kaggle.com/datasets/keithscully/us-stock-valuation-analysis

max_dept=7 & max_features=1.0. Overall, model performance was reasonably good, but it was hoped that better models with lower loss metrics could be developed with through the various boosting algorithms.

#### 5.2.2.2 AdaBoost Algorithm

The AdaBoost algorithm was trained with tuning of n_estimators, learning_rate & loss hyperparameters. A decision tree with max_depth=6 was used as the base estimator. The most performant model was found with n_estimators=500, learning_rate=1.5 & loss='square'. Overall, model performance was better than the RF model.

#### 5.2.2.3 XGBoost Algorithm

The XGBoost algorithm was trained with tuning of the max_depth, eta, gamma, subsample & colsample_bytree hyperparameters. The number of boosting rounds was set to 1000 initially, though the early_stopping_rounds parameter was also set to avoid overfitting as more boosted trees were added to the model. The most performant model was found with max_depth=4, eta=0.2, gamma=0, subsample=1.0 & colsample_bytree=0.8. This model outperformed prior RF and AdaBoost models.

#### 5.2.2.4 LightGBM Algorithm

The LightGBM algorithm was trained with tuning of the max_depth, num_leaves, n_estimators & learning_rate hyperparameters. The most performant model was found with max_depth=7, num_leaves=31, n_estimators=500 & learning_rate=0.05. This model was quicker to train than the XGBoost model, but performance was slightly inferior.

# 6 Evaluation

In this section the key findings of these study will be presented including the evaluation scores of developed models, an analysis of model fit, a summary of model explainability, and the result from back-testing the model on new and unseen historical data and comparing performance of model outputs against the wider stock market.

## 6.1 Primary Evaluation Metrics

The primary evaluation metric utilised in this study was RMSE given its heavier weighting to larger errors. Further evaluation used MAE and R2, to help identify the strongest performing model overall. Performance measured for all models are shown in Table 2.

**Table 2: Comparison of Evaluation Metrics for Trained Models.**

| Algorithm | RMSE | MAE | R2 |
|---|---|---|---|
| Random Forest | 0.241 | 0.181 | 0.799 |
| AdaBoost | 0.198 | 0.165 | 0.866 |
| XGBoost | 0.087 | 0.060 | 0.974 |
| LightGBM | 0.110 | 0.076 | 0.959 |

The XGBoost model we deemed to be most performant, with the lowest RMSE, lowest MAE and highest R2 scores overall. A scatter plot of actual values versus predicted values is

shown in Figure 8, where it is evident that the chosen model is capable of making reasonably accurate predictions on the test data.

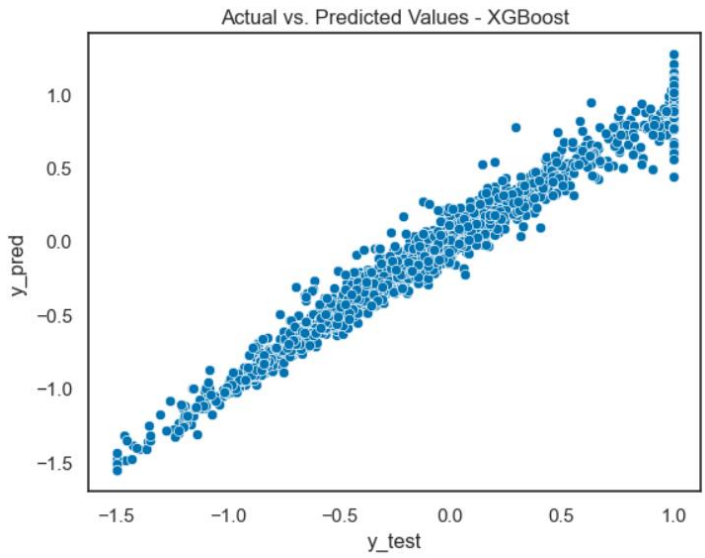


**Figure 8: Actuals vs Predicted for XGBoost Model.**

## 6.2 Analysis of Model Fit

Analysis of the model fit was carried out through the production of learning curve plots, shown in Figure 9, to identify the most suitable point in the training of the model at which the performance on the test set reaches its peak. The number of boosting rounds was incremental learning parameter of interest here, and based on the hyperparameters identified earlier the model was trained up to 1,000 boosting rounds. To continue training the algorithm beyond this point would result in very little improvement in the error metric on test data, and would only serve to overfit the model on the training data.

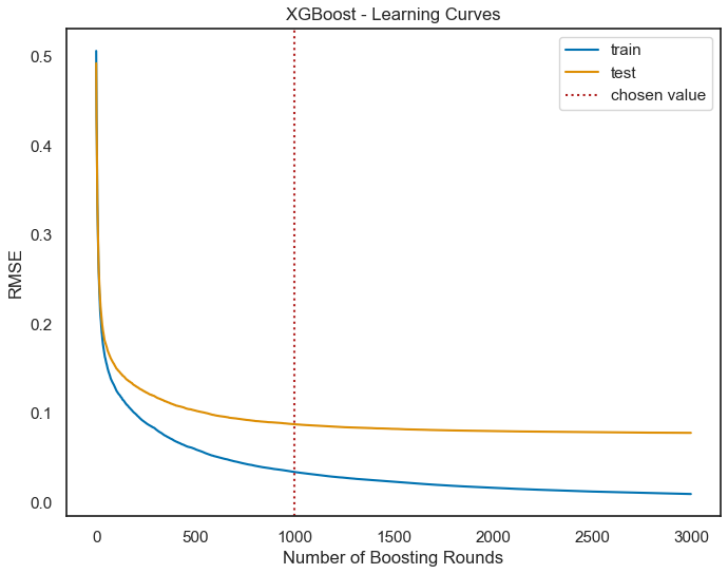


**Figure 9: Learning Curves for Training Loss Metric.**

## 6.3 Explainable AI with SHAP

SHAP values were used to help understand how input features impact model predictions. The ‘Free Cash Flow’ (FCF), ‘Market Cap’ and ‘Net Cash’ features have the most impact on model predictions. The bee-swarm summary plot shown in Figure 10 indicates that the FCF feature plays a very dominant role in the determination of predicted values. This is somewhat expected given how the DCF model used to determine target variables. Higher FCF values tend to encourage a higher degree of safety margin in a stock, as do ‘Net Cash’ and ‘Revenue 3yr CAGR’ features. The market capitalisation for a company is a measure of the price of its stock, so it is positive that the model was capable of estimating its relationship with the margin of safety target variable. However, the most interesting characteristic of this finding was that a lower value for this particular feature contributes to a higher safety margin, indicating that perhaps smaller companies are more likely to displaying higher levels of growth which would be consistent with the concept that as companies get bigger it becomes more difficult to maintain high levels of growth.

These SHAP values provide substantial interpretability for the model and highlight that high FCF, low Market Cap and high Net Cash values are significant drivers for obtaining a higher margin of safety. Considering these key features in conjunction, including the sign and magnitude of their values, can direct investors to choosing companies that have better future prospects and lower risk. By leveraging these insights, more informed and confident investment decisions can be made.

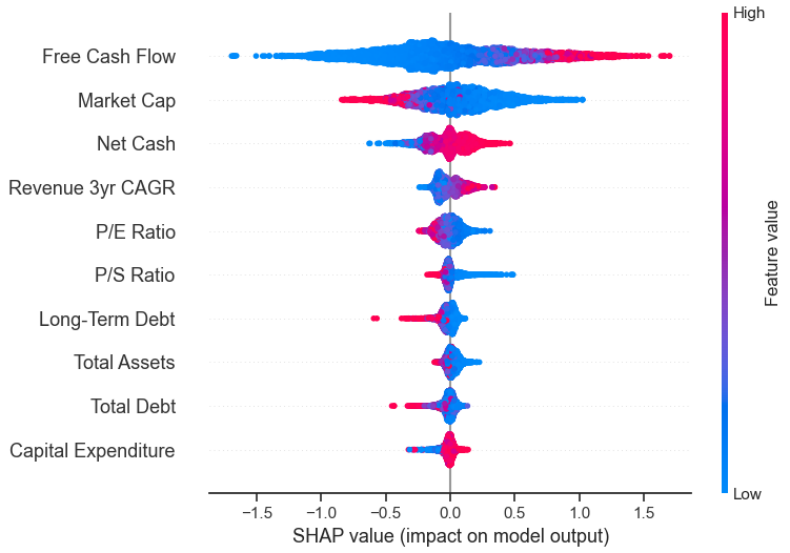


**Figure 10: SHAP Bee-Swarm Summary Plot of Feature Impact.**

## 6.4 Back Test on Unseen Historical Data

Once the final model had been chosen, the model was back tested on new and unseen data to make predictions on a cohort of US stocks given financial data available at a particular point in time. The intention here was to select a portfolio of stocks based on model predictions before measuring the performance of this portfolio against the market benchmark across a 10-year period.

A new dataset was generated for prediction with accounting data for fiscal year 2012, along with pricing data for all stocks at 1st April 2013, which pre-dates the original dataset. Having utilised the pipeline to carry out transformations on the new input data, predictions were then made to generate margin of safety values for each stock. The 20 stocks with the

highest predicted margin of safety values were selected to form a portfolio. Monthly pricing data was then obtained for each of these stocks using Google Finance, from 1st April 2013 to 1st April 2023. A theoretical $10,000 investment was divided equally across these 20 stocks. This portfolio was then compared with the S&P 500 index, and results indicate that the portfolio selected based on model predictions had significantly outperformed the market benchmark over both a 5-year and 10-year period, with considerable improvements in annualised rates of return. The back-test results are shared in Figure 11 and Table 3.

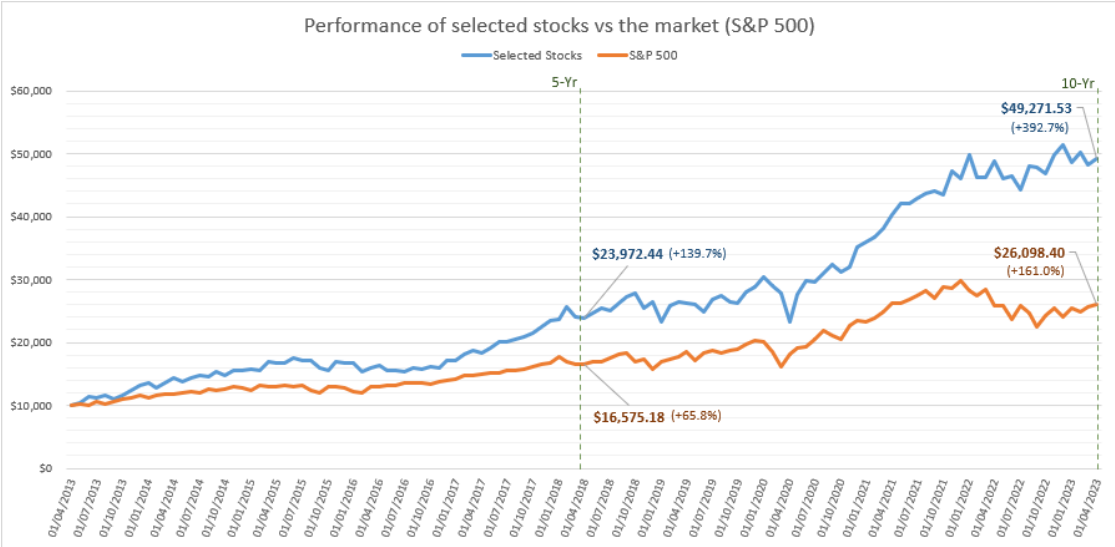


**Figure 11: Performance of Top 20 Stocks vs S&P 500.**

**Table 3: Performance Comparison of Selected Portfolio vs S&P500.**

| | 5 Years | | 10 Years | |
|---|---|---|---|---|
| **Benchmark** | **% Change** | **Annualised** | **% Change** | **Annualised** |
| S&P 500 | 65.8% | 10.6% | 161.0% | 10.1% |
| Project Portfolio | 139.7% | 19.1% | 392.7% | 17.3% |

## 6.5 Discussion

The findings indicate that ML algorithms can be successfully applied to predict the margin of safety available given the required accounting data and the stock price. The accuracy of predictions was reasonably good considering the complexity associated with the interaction of many distinct data features, a common theme in related literature.

The design of this study was generally robust, though a potential modification could be considered in how target values were generated. A limitation of the current implementation is that DCF modelling was applied generally, taking the same approach with all companies. The process could be customised to account for the diverse characteristics of various business sectors when estimating future growth rates, as well as ensuring that future company growth rates include some consideration of historical or expected sector growth rates. Given that FCF amounts can vary considerably year-on-year, the valuation process could possibly be improved by applying some smoothing technique to annual FCF values, perhaps with an exponential moving average, from which a potentially more reliable starting FCF value could be calculated and fed into the DCF model. Further limitations exist with the approach taken that are common to DCF valuations, where it was necessary to exclude certain companies

from the study, and because the DCF cannot model the potential impact to stock prices of market effects due to changing market sentiment.

The results for models developed in this study cannot be compared directly to models developed in related literature, as research objectives are quite dissimilar. While some related studies have also promised potential returns far greater than the market benchmark, most studies that focus on stock valuations have taken a relative valuation approach which does not consider an investor's risk of loss. The fundamental valuation approach taken in this research does not follow a strategy that exposes investors to wider market risk, instead proposing buy indicators only when the stock trades at a discount to its' estimated value. Other studies propose a portfolio optimisation strategy requiring frequent trading or many stock which is impractical to the average retail investor. Finally, the use of tree-based methods along with explainable AI techniques support the potential use of the developed model in the financial industry.

# 7 Conclusion and Future Work

The goal of this project was the investigate the potential of tree-based ML algorithms to accurately predict the margin of safety available on the current stock price. Extensive efforts in data collection were made to prepare a bespoke dataset suitable for this study, which is the first major output that has also been made publicly available to the ML community. However, it must be noted that the data retrieved from the Tickerdata API could not be validated for accuracy due to difficulties associated with the lack of detailed standards and the consistency of terminology used in financial statements. The other major output was the developed XGBoost model, which enabled the selection of a portfolio of stocks that significantly outperformed the market on back-testing. Using explainable AI techniques, the impact of input feature on model predictions was understandable, answering the secondary research question and providing insight on which accounting items drive higher margin of safety values.

This work contributes to research in the domain by proposing a novel approach to stock valuation by combining fundamental valuation strategies with modern ML methods to estimate the scale of price-value gaps and encourage low risk, high return investments. This represents a more risk-averse strategy when compared with many related studies, where stocks are valued on their own merit with wider market effects removed. It is hoped that this research will encourage further exploration of fundamental stock valuation using ML.

Future work may focus on the timeliness of accounting data. US public companies are required to file quarterly accounting data, meaning that quarterly financial values are publicly-available. However, it was not yet possible to retrieve quarterly data from the Tickerdata platform during the data collection phase of this project. Including quarterly data would allow the analysis of accounting data on a trailing-twelve-months basis, retaining a 12-month perspective on the business but at more regular intervals which would remove the excessive time lag between fiscal year-end accounts and some of the pricing dates that fall during the subsequent year. Further work could be also done carried out by expanding the intrinsic value calculation with alternative valuation methods such as the Dividend Discount Model so that more stocks from the financial sector, which are not suitable for DCF valuation, could also be included in the process to determine margin of safety target values. More extensive hyperparameter tuning, using techniques such as Bayesian Optimisation or with libraries such as Optuna or Hyperopt, may be useful in extracting further model performance gains. Further back-testing could also be carried out across other time periods to obtain a greater perspective of the model's potential.

While the proposed model can provide benefit to investors with technical skill using the developed pipeline, there is commercial potential in such a model where online stockbrokers could integrate valuation models in their platforms, potentially in stock screening features. This could involve estimation of intrinsic values or margin of safety values, to support retail investors in making investment decisions with greater understanding of the associated risk.

# Acknowledgements

I would like to sincerely thank my supervisor Dr. Mohammed Hasanuzzaman for his guidance and insightful advice during the execution of this project, as well as my wife and son for their support and understanding throughout the process.

# References

Başoğlu Kabran, F. and Ünlü, K.D. (2021) 'A two-step machine learning approach to predict S&P 500 bubbles', *Journal of Applied Statistics*, 48(13–15), pp. 2776–2794. Available at: https://doi.org/10.1080/02664763.2020.1823947.

Buczynski, W., Cuzzolin, F. and Sahakian, B. (2021) 'A review of machine learning experiments in equity investment decision-making: why most published research findings do not live up to their promise in real life', *International Journal of Data Science and Analytics*, 11(3), pp. 221–242. Available at: https://doi.org/10.1007/s41060-021-00245-5.

Buffett, W.E. (1984) 'The Superinvestors of Graham-and-Doddsville'. Columbia Business School, May. Available at: https://web.archive.org/web/20180814163147/https://www8.gsb.columbia.edu/articles/columbia-business/superinvestors.

Evdokimov, I., Kampouridis, M. and Papastylianou, T. (2023) 'Application Of Machine Learning Algorithms to Free Cash Flows Growth Rate Estimation', *Procedia Computer Science*, 222, pp. 529–538. Available at: https://doi.org/10.1016/j.procs.2023.08.191.

Geertsema, P. and Lu, H. (2023) 'Relative Valuation with Machine Learning', *Journal of Accounting Research*, 61(1), pp. 329–376. Available at: https://doi.org/10.1111/1475-679X.12464.

Graham, B. (1949) *The Intelligent Investor*. New York: Harper & Brothers.

Graham, B. and Dodd, D.L. (1934) *Security Analysis*. New York & London: McGraw-Hill Book Company.

Hanauer, M.X., Kononova, M. and Rapp, M.S. (2022) 'Boosting agnostic fundamental analysis: Using machine learning to identify mispricing in European stock markets', *Finance Research Letters*, 48, p. 102856. Available at: https://doi.org/10.1016/j.frl.2022.102856.

Huang, K.J. (2022) 'DeepValue: A Comparable Framework for Value-Based Strategy by Machine Learning', *Computational Economics*, 60(1), pp. 325–346. Available at: https://doi.org/10.1007/s10614-021-10151-4.

Huang, Y., Capretz, L.F. and Ho, D. (2021) 'Machine Learning for Stock Prediction Based on Fundamental Analysis', in *2021 IEEE Symposium Series on Computational Intelligence (SSCI). 2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, Orlando, FL, USA: IEEE, pp. 01–10. Available at: https://doi.org/10.1109/SSCI50451.2021.9660134.

Khartik, K. et al. (2023) 'A Survey of Price Prediction using Deep Learning Classifier for Multiple Stock Datasets', in *2023 Second International Conference on Electronics and Renewable Systems (ICEARS). 2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*, Tuticorin, India: IEEE, pp. 1268–1275. Available at: https://doi.org/10.1109/ICEARS56392.2023.10085095.

Khoa, B.T. and Huynh, T.T. (2023) 'The value premium and uncertainty: An approach by support vector regression algorithm', *Cogent Economics & Finance*, 11(1), p. 2191459. Available at: https://doi.org/10.1080/23322039.2023.2191459.

Klarman, Seth A. (1991) *Margin of Safety: Risk-Averse Value Investing Strategies for the Thoughtful Investor*. HarperBusiness.

Koklev, P.S. (2022) 'Business Valuation with Machine learning', *Finance: Theory and Practice*, 26(5), pp. 132–148. Available at: https://doi.org/10.26794/2587-5671-2022-26-5-132-148.

Peng, Y.-L. and Lee, W.-P. (2024) 'Valuation of Stocks by Integrating Discounted Cash Flow With Imitation Learning and Guided Policy', *IEEE Transactions on Automation Science and Engineering*, pp. 1–12. Available at: https://doi.org/10.1109/TASE.2023.3246590.

Quah, T. (2008) 'DJIA stock selection assisted by neural network', *Expert Systems with Applications*, 35(1–2), pp. 50–58. Available at: https://doi.org/10.1016/j.eswa.2007.06.039.

Strader, T.J. et al. (2020) 'Machine Learning Stock Market Prediction Studies: Review and Research Directions', *Journal of International Technology and Information Management*, 28(4), pp. 63–83. Available at: https://doi.org/10.58729/1941-6679.1435.

Tsai, P.-F., Gao, C.-H. and Yuan, S.-M. (2023) 'Stock Selection Using Machine Learning Based on Financial Ratios', *Mathematics*, 11(23), p. 4758. Available at: https://doi.org/10.3390/math11234758.

Yang, H., Liu, X.-Y. and Wu, Q. (2018) 'A Practical Machine Learning Approach for Dynamic Stock Recommendation', in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, New York, NY, USA: IEEE, pp. 1693–1697. Available at: https://doi.org/10.1109/TrustCom/BigDataSE.2018.00253.