


UNIWERSYTET WARSZAWSKI
Wydział Nauk Ekonomicznych

University of Warsaw

Faculty of Economic Sciences

Długa 44/50, 00-241 Warsaw

Postgraduate Studies

Data Science in Business Applications – Practical Workshops

PREDICTING KEY INPUTS TO THE DISCOUNTED CASH FLOW MODEL USING MACHINE LEARNING APPROACHES

Author: Jan Pabiszczak

Album Number: P-33666

Thesis prepared under the supervision of:
dr hab. Piotr Wójcik, prof. ucz.
Department of Data Science
Faculty of Economic Sciences UW

Warsaw, 2025

Supervisor's Declaration

I hereby declare that the present thesis has been prepared under my supervision and I confirm that it meets the requirements specified for diploma theses.

Date: 25.09.2025 (signature) Piotr Wojcik

Author's Declaration

Being aware of my legal responsibility, I declare that this diploma thesis has been written by me independently and does not contain any content obtained in a manner inconsistent with the applicable regulations.

I also declare that the submitted thesis has not previously been subject to procedures related to obtaining a postgraduate certificate or a professional title at a higher education institution.

Furthermore, I declare that this version of the thesis is identical to the attached electronic version.

Date: 25.09.2025 (signature) Jan Polowczyk

Abstract

This thesis presents a data-driven approach to corporate valuation that integrates traditional financial methods with modern data science techniques. The study develops a multi-phase Discounted Cash Flow (DCF) model, enhanced with predictive modeling of key drivers and Monte Carlo simulation to incorporate uncertainty. Using Adobe Inc. as a case study, the research combines financial statement data and macroeconomic indicators with machine learning algorithms such as ElasticNet, Lasso, Decision Trees, and XGBoost to forecast revenue growth. The resulting valuation framework generates probability distributions of firm value rather than a single deterministic outcome (point estimate), providing a more transparent and robust view of corporate worth. The findings demonstrate how data science can enhance the reliability and interpretability of valuation models, bridging the gap between academic finance and practical business applications.

Streszczenie

Niniejsza praca przedstawia podejście do wyceny przedsiębiorstw, łączące tradycyjne metody finansowe z nowoczesnymi technikami data science. W badaniu opracowano wieloetapowy model zdyskontowanych przepływów pieniężnych (DCF), uzupełniony o modele statystyczne służące do prognozowania kluczowych zmiennych. Powyższy model DCF połączono następnie z symulacją Monte Carlo, uwzględniającą element niepewności w predykcjach. Studium przypadku dotyczy spółki Adobe Inc., dla której w prognozowaniu dynamiki przychodów wykorzystano dane ze sprawozdań finansowych oraz wskaźniki makroekonomiczne, a także algorytmy uczenia maszynowego (ElasticNet, Lasso, drzewa decyzyjne i XGBoost. Opracowane podejście generuje rozkład prawdopodobieństwa wartości przedsiębiorstwa, w odróżnieniu do pojedynczej deterministycznej wyceny, co pozwala uzyskać bardziej wiarygodny obraz wartości spółki. Wyniki wskazują, że zastosowanie data science może zwiększyć skuteczność i interpretowalność modeli wyceny, łącząc teorię finansów z praktycznymi zastosowaniami biznesowymi.

TABLE OF CONTENTS

Introduction

Valuation is one of the most important tasks in finance, investment analysis, and corporate decision-making. Among different approaches, the Discounted Cash Flow (DCF) model remains the main method for conducting company valuations. By forecasting future free cash flows and discounting them at the cost of capital, the model provides an estimate of the intrinsic value of a company[1]. Despite its popularity, DCF is highly sensitive to underlying assumptions regarding revenue growth, operating margins, reinvestment, and discount rates. Even small changes in these inputs can lead to significant discrepancies in estimated value. This makes the model appear untrustworthy and controversial.

The fast development of data science creates new opportunities to address the issues mentioned above. Predictive modeling allows one for the systematic estimation of financial variables, using historical company data and macroeconomic indicators rather than relying solely on analyst judgment or historical averages (which are often oversimplified). At the same time, Monte Carlo simulation enables analysts to limit, but not entirely eliminate uncertainty by replacing fixed assumptions with probability distributions. Together, these methods allow valuation models to move beyond point estimates and produce ranges of potential outcomes, taking care of both expected performance and risk.

The fundamental goal of this thesis is to design and implement an empirical valuation framework consisting of three key elements:

1. Predictive modeling of key inputs using machine learning techniques.
2. A multi-phase DCF model implemented in Python and designed for flexibility across companies.
3. Monte Carlo simulation, to generate distributions of potential firm values under uncertainty.

[1] Damodaran, A., 2012. Damodaran on Valuation: Security Analysis for Investment and Corporate Finance. 3rd ed. Hoboken, N.J.: John Wiley & Sons.

In my project, I focused on Adobe Inc. as a case study, though the framework can be generalized to other companies in a similar sector. The thesis follows a structured workflow: gathering financial and macroeconomic data, cleaning and preprocessing, selecting variables, building predictive models, and conducting a valuation enhanced with simulation.

This compound approach seeks not only to produce a robust valuation of Adobe but also to demonstrate how combining machine learning with the fusion of statistics and financial theory can improve the realism and practical relevance of corporate valuation.

Ultimately, the thesis aims to combine methods from finance, econometrics, and machine learning into a complete, data-driven valuation framework. By diligently testing predictive models and simulation-based valuation, this project aims to evaluate whether data science methods can improve both the accuracy and the robustness of corporate valuation compared to traditional approaches.

To achieve this, the following hypotheses are formulated:

H1. Monte Carlo simulation produces a valuation distribution that better captures valuation uncertainty than a single deterministic DCF estimate.

Justification: The distribution of outcomes communicates risk more effectively than a single deterministic number.

H2. Integrating macroeconomic variables (e.g., GDP growth, interest rates) into predictive models enhances forecasting accuracy compared to using firm-level financials alone.

Justification: Company performance is partly driven by macroeconomic conditions.

The remaining part of this thesis is structured as follows. **Chapter I** describes the data sources, the sectoral classification of companies, and the preprocessing steps applied to prepare the dataset for modeling. **Chapter II** introduces the machine learning models used to forecast next-year revenue growth, including both linear and non-linear

approaches, and evaluates their predictive performance against baseline models. **Chapter III** develops the multi-phase Discounted Cash Flow (DCF) model, explains its integration with machine learning forecasts, and embeds Monte Carlo simulation to reflect uncertainty in key valuation inputs. **Chapter IV** presents the empirical case study of Adobe Inc., applies the forecasting and valuation framework, and discusses the results in the context of sector peers and prevailing market valuations. The thesis concludes with a summary of findings, and implications for valuation practice.

CHAPTER I Data and Research Framework

1.1. Data Sources

The empirical foundation of this study is built on two complementary categories of data: firm-level financial statements and macroeconomic indicators. Both dimensions are material. Financial statements capture company-specific fundamentals such as profitability, leverage, and reinvestment patterns, while macroeconomic indicators reflect the environment in which firms operate. Together, they provide a coherent dataset for forecasting revenue growth and for constructing a valuation framework that connects micro- and macro-level drivers[2].

1.1.1. Financial Data (FMP API)

Firm-level data were obtained from the Financial Modeling Prep (FMP) API, which provides access to the three principal financial statements:

- **Income Statement**: revenues, cost of goods sold (COGS), operating expenses, EBIT (operational income), interest expense, net income, EPS;
- **Balance Sheet**: total assets, current assets, liabilities, equity, debt structure, working capital;
- **Cash Flow Statement**: operating cash flow, investing cash flow, financing flows, capital expenditures (CAPEX).

The dataset for Adobe Inc. (ticker: ADBE) initially covered the period 2010–2024, that is a 15-year horizon. To prevent overfitting to a single firm, the dataset was expanded to include approximately 100 companies in the IT and AI business sector, including software, SaaS, cloud computing, and enterprise AI firms. At first, this full period was considered for modeling, as longer histories typically improve the stability of statistical forecasts. However, during the cleaning and exploratory phases, it became obvious that a large proportion of companies in the broader training set were relatively young, with financial statements only available for recent years. Including the full 15 year span would therefore introduce more missing values and bias the dataset toward older, more established firms with longer financial histories.

[2] Damodaran, A., 2024. The Little Book of Valuation: How to Value a Company, Pick a Stock and Profit. Updated edition. Hoboken, N.J.: John Wiley & Sons

Adobe Inc. was selected as the case study because it combines strategic relevance as a leading SaaS and creative software provider, reliable and transparent financial reporting, and strong applicability to modern valuation frameworks. These characteristics make it an ideal company for testing the integration of machine learning forecasts into a Monte Carlo–enhanced DCF model.

For this reason, the modeling horizon was restricted to the 2019–2024 period. This choice has both practical and economic justification:

- **Data completeness**: it maximizes the number of firms with full, consistent reporting across financial statements, ensuring balanced observations for machine learning;
- **Economic relevance**: the years 2019–2024 better capture a dynamic period for the global economy and the technology sector. The dataset reflects the effects of the COVID-19 pandemic, the recovery that followed, rising interest rate, and geopolitical shocks (such as the Russia–Ukraine war). For SaaS and cloud companies in particular, these years serve as a natural experiment in demand shifts, cost structures, and capital allocation[3].

Thus, while the broader 2010–2024 dataset was used for descriptive analysis, the 2019 - 2024 window was selected as the modeling sample. In the context of IT and SaaS firms, where business models evolve rapidly and macroeconomic shocks significantly alter revenue trajectories, this shorter horizon is arguably more meaningful for forecasting.

1.1.2. Macroeconomic Data (FRED and Others)

Macroeconomic variables were incorporated to provide context not captured by financial data. As Investopedia explains, "fundamental analysis also looks at macroeconomic factors … Analysts may consider gross domestic product, inflation, interest rates …[4]"

[3] Reuters, 2025. Five years on, the economic impact of COVID-19 lingers. [online] Available at: https://www.reuters.com/business/healthcare-pharmaceuticals/five-years-economic-impact-covid-19-lingers-2025-03-08/ [Accessed 19 August 2025].
Kenton, W., 2022. The long-term impacts of the COVID-19 K-shaped recovery. Investopedia. [online] Available at: https://www.investopedia.com/long-term-impacts-of-the-covid-19-k-shaped-recovery-5200711 [Accessed 19 August 2025].

[4] Investopedia (2023) Fundamental analysis: What it is and how to use it. Investopedia. Available at: https://www.investopedia.com/terms/f/fundamentalanalysis.asp.

(Investopedia, 2023). These variables reflect cyclical and structural conditions shaping corporate performance and valuation. Data were collected from the Federal Reserve Economic Data (FRED) database and other public sources. The selected indicators and their relevance are:

- **Real GDP (GDPC1):** A measure of aggregate economic activity and demand (proxy for aggregate demand). Expansions in GDP typically stimulate enterprise and consumer spending, leading to increased demand for technology and creative software solutions such as Adobe's products;
- **CPI Inflation (CPIAUCSL):** Inflation affects both costs and revenues. Rising inflation increases expenses (e.g., wages, infrastructure), potentially compressing margins. At the same time, moderate inflation enables firms to raise prices, boosting nominal revenues[5];
- **Federal Funds Rate (FEDFUNDS):** This is the primary short-term interest rate set by the Federal Reserve, shaping overall credit conditions and liquidity in the economy. Higher rates raise borrowing costs, discourage investment, and feed directly into the weighted average cost of capital (WACC), lowering DCF valuations[6];
- **10-Year Treasury Yield (GS10):** A benchmark for long-term financing costs and a standard input for discount rates in valuation models. Rising yields increase the cost of debt and equity capital (risk-free benchmark), compressing valuation multiples, slowing growth expectations and overall investment[7];
- **Unemployment Rate**: An indicator of overall economic health and consumer demand. In the software sector, employment levels also drive enterprise software licensing and subscription demand, as corporate spending on digital tools correlates with workforce expansion;
- **Consumer Sentiment (UMCSENT):** A forward-looking measure of household and corporate confidence. Strong sentiment often precedes higher

[5] U.S. Bureau of Labor Statistics. (n.d.). Consumer Price Index for All Urban Consumers: All Items in U.S. City Average [CPIAUCSL]. Retrieved August 19, 2025, from FRED, Federal Reserve Bank of St. Louis: https://fred.stlouisfed.org/series/CPIAUCSL

[6] Board of Governors of the Federal Reserve System (U.S.). (n.d.). Federal Funds Effective Rate [FEDFUNDS]. Retrieved August 19, 2025, from FRED, Federal Reserve Bank of St. Louis: https://fred.stlouisfed.org/series/FEDFUNDS

[7] Board of Governors of the Federal Reserve System (U.S.). (n.d.). Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity, Quoted on an Investment Basis [GS10]. Retrieved August 19, 2025, from FRED, Federal Reserve Bank of St. Louis: https://fred.stlouisfed.org/series/GS10

IT and marketing budgets, which are directly relevant to Adobe's enterprise customer base.

By incorporating these macro variables into the dataset, the model accounts for differences in growth outcomes across economic environments (e.g., post-crisis recovery versus periods of high interest rates). This enhances both predictive accuracy and economic interpretability, which will be demonstrated.

1.2. Sectoral Categorization of Companies

To avoid overfitting the forecasting model to Adobe alone and to construct a more representative training set, the dataset was extended to include approximately 100 companies across different sectors of the software and AI ecosystem. Firms were grouped into categories based on their primary business model, revenue structure, and role in the broader technology value chain (e.g., vertical or horizontal orientation). The resulting categorical variable ("**label")** was added to the dataset as a one-hot encoded feature, allowing the machine learning model to distinguish sector-specific growth dynamics while still leveraging common patterns across firms.

This heterogeneity is reflected in differences in revenue drivers, cost scaling, and the intensity of AI adoption.

The sectors were chosen deliberately, based on financial comparability to Adobe and their relevance for predicting future growth rates. The categorical variable therefore serves two complementary purposes:

- **Feature engineering**: it embeds sector-level economic structure into the ML pipeline as a categorical feature (later one-hot-encoded);
- **Economic control**: it prevents the model from forcing a single growth pattern across fundamentally different sub-segments of the software industry

The choice of sectors reflects both financial comparability and economic relevance to Adobe's business model:

- **Enterprise SaaS (Enterprise SaaS label)**

This category captures Adobe's closest peers. Firms such as Microsoft, Salesforce, and ServiceNow deliver subscription-based enterprise software platforms with high margins and recurring revenue streams. They are also leaders in integrating AI into productivity and workflow solutions. Their growth dynamics, driven by subscription expansion, cross-selling, and AI-driven upselling, mirror Adobe's historical and projected trajectory, making them the core benchmark group[8];

— **Cloud Infrastructure, AI Analytics, and Cybersecurity (Cloud & Data label)**
This group provides the foundational infrastructure that enables enterprise SaaS adoption, including cloud-native databases, observability tools, and cybersecurity platforms. Companies like Snowflake, and Datadog face similar enterprise sales cycles and usage-based pricing structures. Although their performance may be more volatile, their scaling dynamics closely relate to SaaS demand;

— **AI-Integrated Developer and Productivity Platforms (AI DevOps label)**
This category includes AI-first developer tools and automation platforms such as Palantir, C3.ai, and GitLab. These firms demonstrate how AI integration accelerates user adoption, monetization (higher revenue per user), and workflow automation. Their niche but high-growth profiles provide **forward-looking comparables** for Adobe's ongoing rollout of AI-enriched creative and marketing tools;

— **Collaboration and Customer Engagement SaaS (Collab & CX label)**
Firms like Zoom, RingCentral, and Five9 fall into this group. While not directly comparable in product scope, their business models share structural similarities with Adobe: subscription licensing, high scalability, and enterprise-oriented customer acquisition. Their revenues are tightly linked to enterprise IT budgets, making them a secondary but correlated signal group for Adobe's forecasts;

— **Vertical SaaS and Workflow Automation (Vertical SaaS label)**

[8] ServiceNow What is enterprise SaaS? Available at: https://www.servicenow.com/products/it-asset-management/what-is-enterprise-saas.html.

Vertical SaaS platforms are tailored for specific industries, such as construction (Procore), legal (Clio), or insurance (Guidewire)[9]. So they go "deep" into one industry vertical, instead of going "horizontal" across all industries. While niche, they share SaaS fundamentals: recurring revenue, scalable cost structures, and R&D reinvestment, that are comparable to Adobe's model and enrich modeling across varied industry contexts (in order to generalize better);

— **Mature Software Firms Pivoting to AI (Mature Pivots label)**

Companies such as SAP, Oracle, and Accenture provide useful benchmarks for long-term growth normalization. They represent firms that, after a phase of high expansion, transitioned into more stable, lower-growth regimes while integrating AI. This category informs the DCF's terminal phase assumptions, highlighting how AI adoption eventually converges to slower but sustainable growth;

— **Consumer SaaS-like Platforms with AI Integration (Consumer SaaS label)**

While consumer-facing, companies like Shopify and Duolingo are increasingly adopting AI to enhance personalization and user engagement. Including them adds alternative monetization models to the dataset, giving contrast to enterprise SaaS but still showing AI-driven revenue growth patterns relevant to Adobe's Creative Cloud.

— **Mega-Cap Technology Anchors (Mega-Cap Tech label)**

The largest technology firms: Apple, Google, Meta, Amazon, Nvidia serve as anchors and boundary cases. They set benchmarks for AI adoption strategies, enterprise cloud penetration, and long-term IT spending trends. Microsoft is both an anchor and a direct competitor to Adobe in enterprise SaaS, while the others provide context for macro-level technology adoption;

— **Excluded Category: AI-Driven Electronic Design Automation (AI EDA)**

Initially, firms such as Cadence, and Synopsys were considered. However, their business model differs substantially from SaaS peers: long-term, high-ticket licensing agreements, hardware-linked customer bases, higher capital intensity, and growth drivers tied

[9] Pouladian, B. (2023) Software is dead? Not so fast — vertical SaaS is thriving. Medium, 31 July. Available at: https://medium.com/@ben_pouladian/software-is-dead-not-so-fast-vertical-saas-is-thriving-b6218207b97c (Accessed: 19 August 2025).

to semiconductor cycles rather than enterprise IT demand. For these reasons, the EDA segment was excluded from the machine learning pipeline, avoiding distortions in model training[10].

Finally, a categorical mapping was applied by assigning each ticker to its sector label. This transformation ensured that the model could establish sector-specific heterogeneity while maintaining comparability across the broader SaaS and AI software landscape.

1.3. Data Cleaning and Preprocessing

The raw dataset, as constructed from firm-level financial statements and macroeconomic indicators, required systematic cleaning and preprocessing to ensure that it was both statistically robust and suitable for machine learning models. The following steps were undertaken in a reproducible and structured pipeline[11];

1. **Handling Missing Values**

A completeness check revealed that a subset of variables, especially for younger firms or earlier years, contained missing values. Observations with missing target values (next-year revenue growth, revenuegrowth_t+1) were removed to preserve the validity of the supervised learning environment. For explanatory variables, missing numeric values were imputed using the mean calculated on the training set only. The same imputer was then applied to the test set to prevent data leakage from future observations into the training phase. This procedure ensured that model evaluation remained unbiased[12].

2. **Encoding of Categorical Variables**

The principal categorical feature in the dataset was the sectoral label, which groups firms into categories such as Enterprise SaaS, Cloud & Data, or Mega-Cap Tech. This label was first cast into categorical data type and subsequently recoded into dummy variables

[10] Data Gravity (2023) Synopsys and Cadence: The $160B Unsung Heroes of Silicon Valley. Available at: https://www.datagravity.dev/p/synopsys-and-cadence-the-160b-unsung.
[11] Géron, A., 2023. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. 3rd ed. Sebastopol, CA: O'Reilly Media.
[12] Grus, J., 2019. Data science from scratch: first principles with Python. 2nd ed. Sebastopol, CA: O'Reilly Media

with a reference level (called also one-hot encoding). One-hot encoding produced a binary vector for each sector, ensuring compatibility with algorithms that require purely numerical inputs while preserving interpretability of sectoral effects. Encoding was performed consistently across training and test subsets.

3. Train-Test Split

To ensure robust model validation and avoid overfitting, the dataset was partitioned into training and test sets. In machine learning practice, it is common to stratify the split when the target variable is unbalanced (e.g., when the "No" class is much smaller than the "Yes" class). In this case, however, the split was **stratified by sectoral label**, maintaining proportional representation of different categories in both subsets. This design reflects realistic economic diversity: the machine learning model is trained on a balanced cross-section of firms and evaluated on equally representative out-of-sample observations.

4. Final Preprocessed Dataset

The resulting datasets were exported as serialized objects, containing:

- cleaned and imputed financial variables;
- macroeconomic indicators aligned to firm-year observations;
- one-hot encoded sectoral labels, and
- next-year revenue growth as the supervised target variable.

This preprocessing pipeline ensures that the dataset is model-ready: it balances the requirements of machine learning (complete, numerical, standardized inputs) with the constraints of financial research (economic interpretability and sectoral comparability). The careful handling of missing values, categorical encoding, and stratified splitting provides a robust foundation for the forecasting approaches presented in Chapter II.

1.4. Feature Engineering and Variable Selection

To transform raw financial statements and macroeconomic series into effective predictors of revenue growth, a systematic feature engineering process was applied. The aim was to construct variables that capture the key drivers of firm performance while ensuring interpretability in both financial and machine learning contexts.

1. Construction of Financial Ratios

From the raw statements, a set of derived financial ratios was engineered to serve potential predictors of next-year revenue growth. These variables reflect established dimensions of corporate analysis:

- **Profitability metrics:** EBIT margin, net margin, return on assets (ROA), and return on invested capital (ROIC), which measure the efficiency of turning revenues and invested capital into profits;
- **Efficiency ratios:** asset turnover, sales-to-capital ratio, and R&D intensity (R&D/revenues), capturing how effectively firms deploy resources to generate sales;
- **Leverage:** debt-to-equity ratio and interest coverage, indicating the extent of financial risk and debt sustainability;
- **Investment and Reinvestment:** CAPEX-to-revenue ratio and reinvestment rate, highlighting growth orientation and capital allocation policies;
- **Growth Dynamics:** year-over-year revenue growth, reflecting momentum effects that often persist in firm performance.

Together, these variables provide both level information (e.g., size of revenues) and dynamic performance (e.g., growth rates), making them useful inputs for forecasting models.

2. Integration of Macroeconomic Indicators

Macroeconomic variables (GDP growth, inflation, interest rates, unemployment, and consumer sentiment) were merged with firm-level observations by year. This alignment ensured that each firm-year observation reflected not only internal

fundamentals but also the external economic environment in which the company operated.

3. Encoding of Sectoral Labels

The categorical variable representing sectoral classification was transformed into a set of binary indicators (via one-hot encoding). This allowed the machine learning models to capture systematic differences between sectors while still estimating pooled relationships across firms.

Since the dependent variable is quantitative (revenuegrowth_t+1) and the predictor is qualitative (categorical) the ANOVA (Analysis of Variance) test was applied to assess whether mean growth rates differ significantly across sectors:

- **Null hypothesis (H_0)**: All eight sectoral means of the dependent variable are equal (no systematic effect);
- **Alternative hypothesis (H_1)**: At least one sectoral mean differs compared to the others.

The F-statistic F = 2.198906, represents the ratio of the between-group variance (explained by sectoral differences) to the within-group variance (residual).

The corresponding p-value, PR(>F) = 0.0336, is the probability of observing an F-statistic as extreme as 2.1989 (or more extreme) if the null hypothesis were true:

- At the conventional 5% significance level, H_0 is rejected (0.0336 < 0.05);
- At the stricter 1% level, H_0 would not be rejected.

Thus, the overall F-test indicates that the sectoral label is informative and adds explanatory power: there is statistical evidence that next-year revenue growth differs across at least one pair of sectors.

	df	sum_sq	mean_sq	F	PR(>F)
C(label)	7.0	1.288392	0.184056	2.198906	0.033626
Residual	387.0	32.393245	0.083703	NaN	NaN

Table 1. ANOVA results for the sectoral categorical variable and the target variable..

Figure 1 presents boxplots of next-year revenue growth (t+1) across company categories. The distributions highlight clear sectoral differences: some categories display higher medians and greater dispersion (e.g., AI DevOps, Consumer SaaS), while others display more stable but lower growth patterns (e.g., Mature Pivots, Mega-Cap Tech). This heterogeneity, together with the ANOVA F-statistic results, provides empirical justification for including the sectoral label variable in the regression framework.

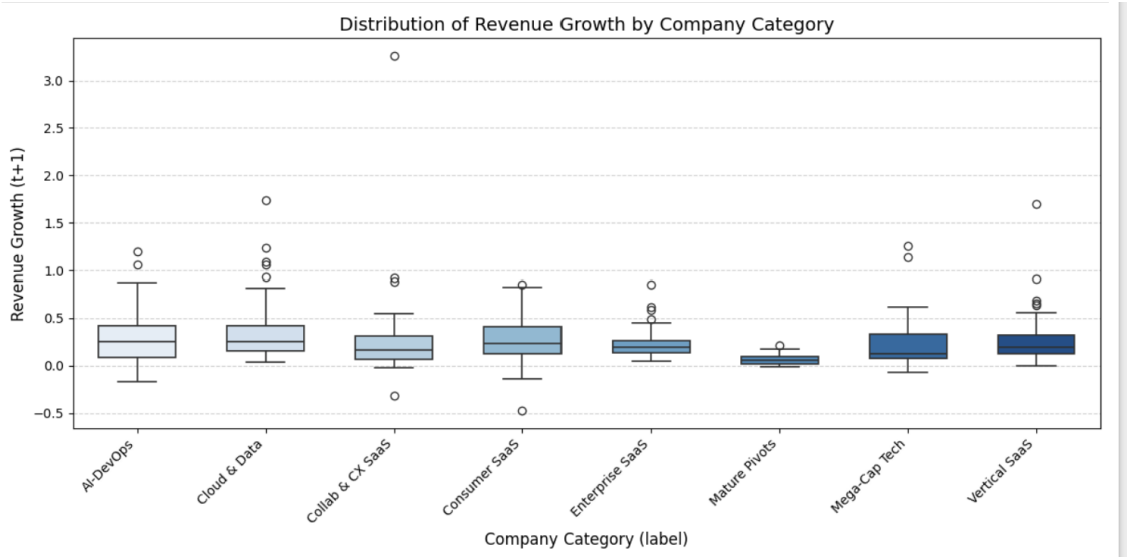

Figure 1. Boxplots representing the next-year revenue growth (t+1) across sectorial categories.

4. Variable Selection

The feature set was deliberately constructed to be broad, capturing the different drivers of revenue growth. However, to mitigate the risks of overfitting and multicollinearity, feature importance diagnostics (e.g., correlation analysis and tree-based feature ranking) were applied in subsequent stages. This approach ensured that the forecasting models balanced predictive accuracy with economic interpretability.

As a result, the final dataset included financial ratios, growth metrics, macro variables, and sector controls. Together, these features constitute the explanatory variables (X, with next-year revenue growth (rev_growth_t+1) as the target variable (y).

Figure 2 presents a correlation matrix in the form of heatmap to help visualize;

— which variables are strongly correlated with each other (indicating potential multicollinearity);

— which variables show unique relationships with the target variable, reflecting their predictive influence.

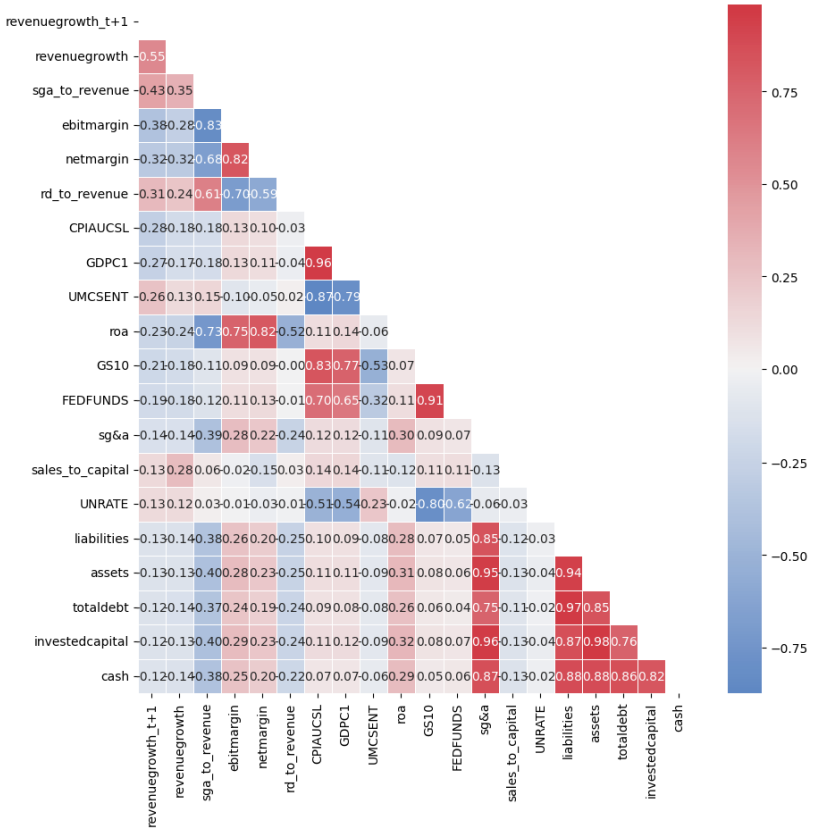

Figure 2. Correlation matrix of features and the target variable.

Figures 3–5 illustrate the pairwise correlations between the target variable (*revenuegrowth_t+1*) and three of the most correlated independent variables.

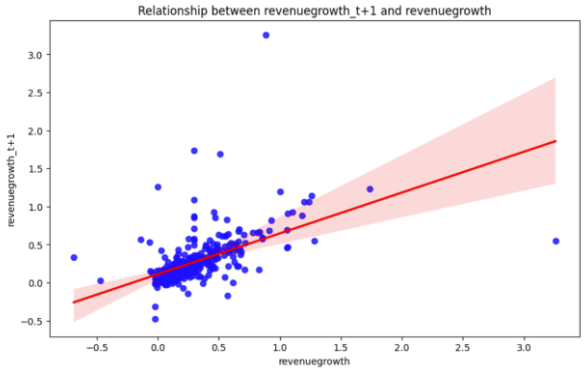

Figure 3. Correlation of Current Revenue Growth with Next-Year Revenue Growth.

Description (target vs revenue growth):

Figure 3 shows the relationship between current revenue growth and next-year revenue growth. The results reveal a clear positive linear trend, with the regression line indicating that past revenue growth is a strong predictor of future growth. The confidence interval (shaded area) is tight in the central range, suggesting a good model fit for the majority of observations. While a few high-growth outliers are visible, they do not dominate the overall pattern.

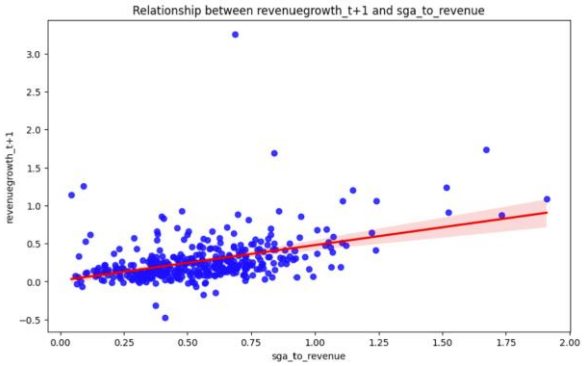

Figure 4. Correlation of Selling, General, and Administrative Expenses to Revenue ratio with Next-Year Revenue Growth.

Description (target vs Selling, General, and Administrative Expenses to Revenue ratio):

Figure 4 illustrates the correlation between the SG&A-to-revenue ratio and next-year revenue growth. The relationship is positive, although weaker than in the previous case, indicating that firms with leaner and more efficient cost structures tend to achieve more sustainable growth. This finding aligns with the notion that cost efficiency improves scalability in SaaS and AI-driven business models, where controlling overhead is a critical determinant of long-term expansion capacity.

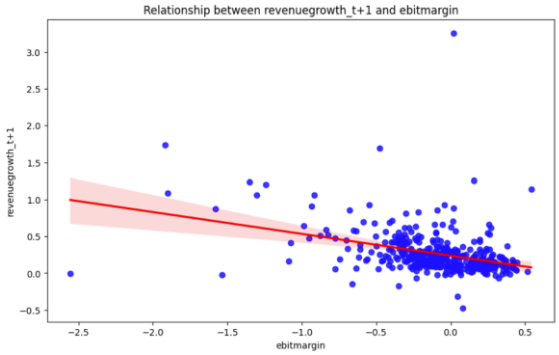

Figure 5. Correlation of EBIT margin with Next-Year Revenue Growth.

Description (target vs ebitmargin):

Figure 5 presents the correlation between EBIT margin and next-year revenue growth. Here the relationship is negative: firms with higher current profitability tend to display lower future growth. Most observations cluster around low margins and low growth, while the regression line remains stable despite some negative outliers. This reflects a common trade-off in the technology sector, where fast-growing firms often sacrifice short-term profitability for reinvestment, while mature, profitable firms exhibit slower but more stable growth.

Financial interpretation of the above plots is that in the tech and SaaS sectors, fast growing firms often sacrifice margins for growth (investing in expansion, burning cash). By contrast, more profitable companies tend to be more mature firms with slower growth rates. This trade-off is common in SaaS, IT, and AI industries, where companies prioritize scaling over short-term profitability[13]. Such firms often reinvest aggressively in marketing, R&D, or infrastructure to win market share and acquire users, even at the cost of negative earnings or margins.

1.5. Chapter Summary

This chapter established the empirical foundation of the study. First, the dataset was constructed from two complementary sources: firm-level financial statements retrieved via the FMP API and macroeconomic indicators from FRED. Second, to prevent overfitting to Adobe alone, the dataset was extended to approximately 100 SaaS, cloud,

[13] Boston Consulting Group (2025) Rule of 40: Lessons from Top Performers in Software. Available at: https://www.bcg.com/publications/2025/rule-of-40-lessons-from-top-performers-software.

and AI-related firms, grouped into economically meaningful sectors. Third, a state-of-the-art data cleaning and preprocessing pipeline was then applied. This included missing value handling, categorical encoding, and stratified train-test splits to preserve the distributional characteristics of the dataset across subsamples. Fourth, feature engineering produced a comprehensive set of financial ratios, growth indicators, and macroeconomic controls that capture both firm-specific fundamentals and broader economic conditions.

These features serve as the explanatory variables, with next-year revenue growth as the target (dependent) variable.

The resulting dataset is both machine learning–ready and economically interpretable, providing a robust basis for the predictive modeling approaches presented in Chapter II. By structuring the dataset in this way, Adobe's revenue growth forecasts can be placed in the context of sector peers and broader macroeconomic conditions, improving both the credibility and interpretability of the valuation exercise.

CHAPTER II. Machine Learning Models for Forecasting

2.1. Predictive Modeling Approach

The forecasting problem was formulated as a supervised regression task, with next-year revenue growth (rev_growth_t+1) as the dependent variable. This target was chosen for its direct relevance to valuation. Revenues determine the scale of future cash flows in a Discounted Cash Flow (DCF) model and strongly influence reinvestment, profitability, and firm value.

The explanatory feature set combined three types of information:

- **Firm-level financial ratios**: profitability (EBIT margin, ROIC), efficiency (asset turnover, sales-to-capital, R&D intensity), leverage (debt-to-equity, interest coverage), investment (CAPEX/revenues, reinvestment rate), and growth dynamics (revenue growth);
- **Macroeconomic indicators**: GDP growth, inflation, interest rates, unemployment, and consumer sentiment, capturing cyclical and structural drivers of technology-sector revenues;
- **Sectoral categories**: encoded as one-hot variables, enabling the model to account for systematic differences between enterprise SaaS, cloud infrastructure, collaboration tools, and other software sub-segments.

The modeling pipeline followed three guiding principles:

1. **Comparative approach** – estimating both linear and tree-based models to balance interpretability and predictive power;
2. **Cross-validation** – applying multiple validation schemes (K-Fold, Repeated K-Fold) to ensure robustness;
3. **Out-of-sample testing** – evaluating all models on a hold-out dataset to measure genuine predictive performance.

This dual focus on accuracy and interpretability ensures that the forecasts not only improve statistical performance but also remain meaningful in a corporate finance context.

2.2. Linear Models: ElasticNet and Lasso

2.2.1. Motivation

Ordinary Least Squares (OLS) regression is a natural starting point for forecasting, but faces limitations when predictors are numerous and correlated, as is typical of financial ratios. Regularization techniques address these issues by penalizing large coefficients, which improves generalization and reduces overfitting[14]:

- Lasso Regression (L1 penalty): shrinks some coefficients exactly to zero, performing implicit feature selection;
- ElasticNet (L1 + L2 penalties): balances sparsity (L1) with stability (L2), which is useful when variables (such as profitability and reinvestment) are correlated.

2.2.2. **Implementation**

Prior to estimation, all explanatory variables were standardized to zero mean and unit variance. This step ensured that the regularization penalties in Lasso and ElasticNet operated consistently across features. It prevented variables with larger scales from dominating the optimization process.

Prior to estimation, all explanatory variables were standardized to zero mean and unit variance. **To avoid data leakage**, the standardization parameters (mean and variance) were computed exclusively on the training set and subsequently applied to the test set. This procedure ensured that the regularization penalties in Lasso and ElasticNet operated consistently across features while preventing variables with larger scales from dominating the optimization process

Model complexity was controlled through hyperparameter tuning. The penalty strength (alpha) and mixing parameter (l1_ratio) were optimized using a cross-validated randomized search. This approach balances efficiency with robustness by sampling from the parameter space rather than relying on an exhaustive grid search.

Model performance was evaluated on both the training and hold-out test sets using multiple error metrics: Mean Absolute Error (MAE) for interpretability, Root Mean

[14] Géron, A. (2023) Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 3rd edn. Sebastopol, CA: O'Reilly Media.

Squared Error (RMSE) to penalize large deviations, and R^2 to assess explained variance. This multi-metric approach provided a comprehensive view of forecasting accuracy and robustness.

2.2.3. Interpretation

The results highlighted the growth–profitability trade-off typical of SaaS firms. High reinvestment ratios correlated positively with growth, whereas high current profitability (EBIT margin) correlated negatively. Lasso helped identify the most influential variables, while ElasticNet provided stable estimates in the presence of multicollinearity.

2.3. Tree-Based Models: Decision Trees and XGBoost

2.3.1. Motivation

Linear models assume proportional, additive effects, which may not capture the nonlinear dynamics of revenue growth. For example, growth may accelerate once a certain threshold of R&D intensity is surpassed, or sectoral effects may interact with reinvestment rates. Tree-based models naturally capture such nonlinearities and interactions.

- Decision Trees partition firms into groups based on rules that minimize prediction error;
- Random Forests aggregate multiple trees, reducing variance and improving robustness;
- XGBoost and LightGBM use gradient boosting to iteratively improve weak learners, often achieving state-of-the-art performance on structured financial data.

2.3.2. Implementation

Hyperparameter optimization was conducted to enhance predictive accuracy while preventing overfitting. For tree-based ensembles such as Random Forests and XGBoost, parameters including tree depth, learning rate, and the number of estimators were tuned using cross-validated randomized search. In addition to these global parameters, structural constraints were applied to regularize the trees: minimum samples per leaf and minimum samples required for a split were set above very low values, thereby reducing the risk of overly complex trees capturing noise rather than signal.

Following training, feature importance scores were extracted to quantify the relative contribution of each predictor. The rankings provided both statistical and economic interpretability. From a statistical perspective, variables linked to current revenue growth (momentum), and reinvestment intensity (e.g., R&D-to-revenue) emerged as dominant. Profitability ratios (e.g., EBIT margin) also played a notable role. Sectoral classifications contributed meaningfully to the forecasts, though not as strongly as reinvestment intensity or momentum effects.

From a financial perspective, this aligns with growth theory, which holds that firms can only expand sustainably if they reinvest efficiently, and industry context largely determines how much growth is possible. Macroeconomic variables also appeared among the influential predictors. This insight reinforced the hypothesis that growth forecasting is sensitive not only to firm fundamentals but also to the broader economic environment.

All models were estimated on the stratified training sample (preserving sectoral representation) and evaluated on the out-of-sample test set, ensuring that reported metrics reflected genuine predictive power rather than artifacts of model complexity.

Note: tree-based models do not require feature standardization; scaling might be applied only to linear models.

2.3.3. Interpretation

Tree-based models captured richer dynamics than linear regression. Feature importance rankings consistently placed momentum and reinvestment intensity among the strongest predictors of growth.

2.4. Model Validation and Performance Metrics

Validation relied on K-Fold cross-validation as the primary scheme and Repeated K-Fold to reduce split randomness. Repeated K-Fold Cross-Validation runs the K-Fold procedure multiple times with different random splits, which reduces sensitivity to any single partition and produces a more stable estimate of model performance, at the cost of higher computation[15].

[15] Wikipedia (2025) Cross-validation (statistics). Available at: https://en.wikipedia.org/wiki/Cross-validation_(statistics).

For reproducibility, the author used K=5 folds; robustness checks used Repeated K-Fold (K=5, repeats=3) with a fixed random state. Leave-One-Out cross-validation was considered but not implemented due to its high computational cost and limited incremental benefit for this dataset.

Model performance was evaluated using complementary error metrics, each capturing a distinct dimension of predictive accuracy:

- Mean Absolute Error (MAE): average absolute difference between predicted and actual next-year revenue growth, reported in the same units as the target (e.g., percentage points if growth is expressed as %);
- Root Mean Squared Error (RMSE): penalized larger deviations more heavily, useful for identifying tail risks;
- R^2 (Coefficient of Determination): variance explained by the model.

These metrics were consistently reported on the hold-out test set to ensure honest assessment of generalization ability.

2.5. Comparison with Baseline Forecasts

All advanced models were benchmarked against a baseline OLS regression. This simple specification assumed linear, additive relationships between features and growth, serving as a reference point;

- Regularized models (Lasso, ElasticNet) improved upon OLS by reducing overfitting and shrinking irrelevant variables. Lasso achieved sparser solutions, while ElasticNet performed best when features were correlated;
- Tree-based models (Random Forest, XGBoost) consistently outperformed OLS in terms of MAE and RMSE. They captured nonlinear relationships and sector interactions that OLS could not represent.

While the baseline OLS served as a useful benchmark, the results demonstrated that regularization and ensemble methods yielded superior predictive accuracy, particularly in capturing the heterogeneous growth dynamics of SaaS and AI-driven firms.

2.6. The Duolingo Case (data error and hypergrowth outlier)

A notable outlier in the sample was Duolingo (DUOL). In the raw FMP feed, Duolingo appeared with an extraordinary next-year revenue growth (>1,300%) and a set of unusual accounting ratios:

- Negative equity coexisting with substantial cash holdings (i.e., strongly negative net debt);
- R&D and SG&A intensity ~ 44–45% of revenue each, indicative of an aggressive reinvestment strategy;
- Derived ratios (ROE, debt-to-equity, liabilities-to-equity) taking −1.0 placeholders due to negative denominators, and invested capital rendered negative, which together flip conventional leverage/return logic.

Economically, this pattern is consistent with venture-style hypergrowth:

1. **Equity financing**: such firms fund themselves mostly by issuing equity (shares) rather than debt;
2. **Aggressive R&D and customer acquisition**: growth is prioritized over profitability;
3. **Operating losses and write-offs**: expenses exceed revenues for extended periods;
4. **Depressed book equity despite high cash buffers**: negative retained earnings (from cumulative losses) reduce book equity, while repeated equity raises create large cash reserves.

However, closer inspection showed that the extreme revenue growth figure was not an economically meaningful signal but rather the result of a data integrity issue in the FMP history:

- The income statement history incorrectly included a 2001 record (revenue = USD 51k), followed by the first legitimate record in 2019 (revenue = USD 70.76m);
- Since Duolingo was founded in 2011, a 2001 statement is implausible;
- The calculation of year-over-year growth between these two points artificially treated an 18-year gap as adjacent years, inflating the revenue growth rate to nonsensical levels (>1,300%).

Modeling impact

— **OLS sensitivity**: The squared-error loss function gave disproportionate weight to the corrupted outlier, inflating RMSE and driving R^2 deeply negative;
— **Robustification attempts**: Alternative specifications, including log-transformations of the target and features as well as robust regressions, did not stabilize the fit because the corrupted point lay far outside the training distribution;
— **Trees/ensembles**: Split points and feature thresholds were also skewed when the outlier was retained.

Decision and rationale

Given the clear data error, the Duolingo outlier was removed from the modeling dataset (test set). This single exclusion materially improved RMSE, MAE and R^2, and restored stable, realistic out-of-sample performance. It is important to note, that the exclusion did not compromise representativeness: the dataset still contained a broad range of high-growth firms, ensuring that the diversity of growth dynamics was preserved.

The Duolingo case underscores two points: (i) API data can contain corrupted history, and (ii) a single extreme observation can destabilize financial forecasting unless the end user carefully checks and evaluates the results. Transparent diagnostics, targeted cleaning, and principled exclusion restored both economic realism and statistical robustness of the forecasting pipeline.

Figure 6 displays the predicted vs. actual plot for the baseline linear regression (scikit-learn API) with the Duolingo outlier included. The single extreme observation forces the y-axis to rescale (predictions > 500%), compressing the remaining points near the origin and effectively flattening the fit. This results in a large RMSE and negative R^2, i.e., worse than predicting the sample mean.

Figure 7 reports the same model estimated after removing the corrupted Duolingo observation. The scatter now aligns visibly along the 45° identity line (dashed), with markedly tighter dispersion and reduced heteroscedasticity. Correspondingly, MAE/RMSE drop and R^2 turns positive, indicating that the model recovers genuine out-of-sample explanatory power once the data error is removed.

After removing the corrupted Duolingo record, one atypical observation remained visible in the upper-left corner of the scatterplot. This is not a data error but rather reflects a firm with unusual financial metrics. Such cases are deliberately retained in the dataset, as the objective is not to fine-tune the model to a perfectly "clean" sample, but to ensure robustness across a heterogeneous universe of companies. Retaining firms with atypical profiles increases the external validity of the model by testing its performance under diverse real-world conditions.

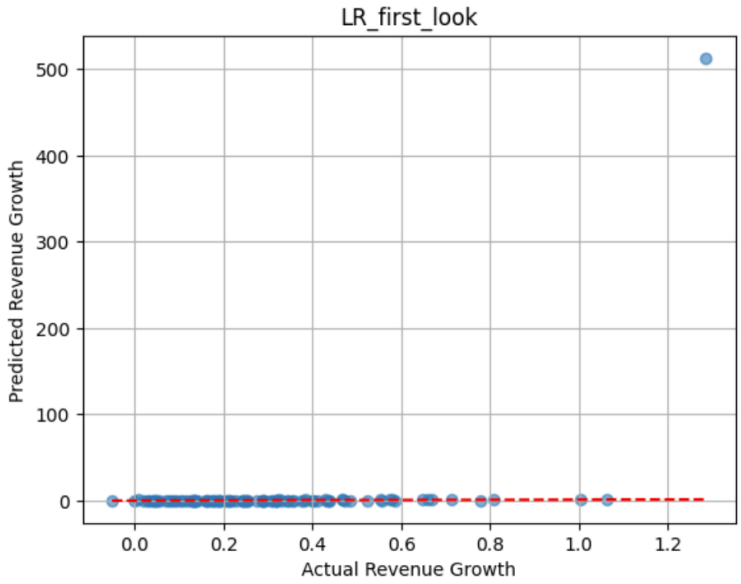

Figure 6. Predicted vs. actual next-year revenue growth (OLS) with Duolingo outlier included.

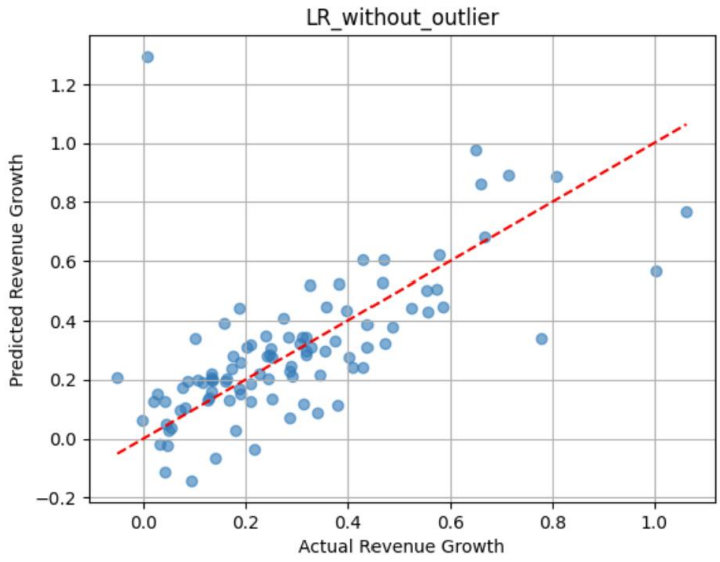

Figure 7. Predicted vs. actual next-year revenue growth (OLS) after removing the corrupted Duolingo observation.

2.7. Chapter Summary

This chapter introduced a machine learning framework for forecasting next-year revenue growth. Linear models (Lasso, ElasticNet) provided interpretability and highlighted the trade-off between growth and profitability, while tree-based models (Decision Trees, Random Forests, XGBoost) captured nonlinear interactions and complex sectoral dynamics.

Validation exercises confirmed that these approaches improved predictive accuracy relative to a baseline OLS regression. Importantly, the inclusion of regularization and ensemble methods enhanced both robustness and interpretability of the forecasts.

These results provide the essential input for Chapter III, where predicted revenue growth rates are embedded into a multi-phase DCF valuation model enhanced with Monte Carlo simulation.

CHAPTER III. Discounted Cash Flow Model with Monte Carlo Simulation

3.1. Structure of the Multi-Phase DCF Model and its key assumptions

The valuation framework applied in this study builds on the classical Discounted Cash Flow (DCF) model, but extends it into a multi-phase structure designed to capture the life cycle of a technology firm. Instead of assuming a single, constant growth trajectory, the model distinguishes between three distinct stages of corporate development followed by a terminal phase valued in perpetuity.

This multi-phase DCF structure is consistent with the framework presented in Professor Aswath Damodaran's lectures[16] and writings[17], and the McKinsey Valuation textbook[18].

Core Assumptions

- **Free Cash Flow to the Firm (FCFF)**: defined as Net Operating Profit After Tax (NOPAT) less reinvestment. Reinvestment requirements are estimated using the sales-to-capital ratio, following Damodaran's principle that growth must be "funded" through incremental capital;
- **Operating margins**: Initiated from Adobe's current operating profitability (~ 31%) and allowed to vary stochastically in Monte Carlo simulations. In deterministic runs, margins are assumed to converge smoothly to long-term sector averages;
- **Reinvestment and growth linkage**: Growth is determined by the product of the reinvestment rate and the return on invested capital (ROIC):

$$g = \text{Reinvestment Rate} \times \text{ROIC} \quad (1)$$

[16] Damodaran, A., n.d. *Teaching: Valuation*. [online] Stern School of Business, New York University. Available at: https://pages.stern.nyu.edu/~adamodar/.
[17] Damodaran, A., 2012. *Damodaran on valuation: security analysis for investment and corporate finance*. 3rd ed. Hoboken, NJ: John Wiley & Sons.
[18] Koller, T., Goedhart, M. and Wessels, D., 2020. *Valuation: measuring and managing the value of companies*. 7th ed. Hoboken, NJ: John Wiley & Sons.

This ensures consistency between financial economics and statistical forecasts: higher growth is only sustainable if reinvestment earns returns above the cost of capital;

— **Cost of capital (WACC)**: Computed dynamically each year, reflecting the evolving capital structure. The cost of equity is derived from the Capital Asset Pricing Model (CAPM), while the cost of debt is based on company-specific borrowing spreads. In simulation, borrowing costs are drawn from a triangular distribution to reflect uncertainty;

— **Discounting convention**: Cash flows are discounted using the mid-year convention, acknowledging that inflows occur throughout the year rather than exclusively at year-end.

Phase 1: High Growth

In the early years, growth rates can exceed the economy's baseline. This is driven by innovation, underpenetrated markets, or competitive advantages. For Adobe, this phase is anchored in the machine learning forecast of revenue growth (11.44% for FY2025) derived in Chapter II.

Assumptions in the model:

— Revenue growth begins at an elevated rate and decelerates modestly;

— Operating margins may still be evolving - companies in this stage often reinvest heavily (R&D, SG&A) and sacrifice short-term profitability for expansion;

— Horizon typically 3–5 years, depending on sector maturity.

Phase 2: Transition

Firms cannot sustain extraordinary growth forever. Competition, market saturation, and diminishing marginal returns on capital drive convergence toward more moderate levels.

Assumptions in model:

- Growth gradually decelerates toward the economy's nominal growth ceiling (real GDP + inflation ≈ 2–3%);
- Operating margins converge toward sustainable levels (sectoral averages). For software companies, Damodaran notes that margins typically improve as fixed costs are spread, but eventually stabilize once economies of scale are exhausted;
- Sales-to-capital ratio normalizes, reflecting higher reinvestment needs per unit of revenue;
- Horizon typically 3–7 years. Smooth convergence is enforced using linear functions to avoid unrealistic "cliffs.". This is why the code uses linear convergence functions for margins, taxes, and STC ratios.

Phase 3: Stable Maturity

In the long run, every firm converges toward a mature profile. Growth can no longer exceed the economy's capacity (GDP + inflation) without implying it eventually overtakes the economy itself. The latter is impossible.

Assumptions in model:

- Growth stabilizes between 2–3% (Adobe set to 3%);
- Margins stabilize at industry averages (here ~31% for Adobe, consistent with sector economics);
- ROIC converges toward WACC, reflecting the erosion of competitive advantage in efficient markets (unless a defensible moat justifies excess returns);
- Reinvestment shrinks, as only modest reinvestment is required to sustain low growth;
- WACC, beta, and leverage converge toward stable, market-level averages.

Phase 4: Terminal Value

The terminal value captures the present value of all cash flows beyond the explicit forecast horizon. It often contributes more than half of enterprise value in practice, making it very sensitive, and critical point of each valuation. It is calculated using the Gordon Growth formula.

Safeguards:

- g_terminal < WACC, otherwise the model produces infinite or negative values;
- Firms can't grow at a rate equal to or greater than their cost of capital forever. If WACC−g is less than epsilon (0.0001), the code **raises an error** instead of calculating a nonsensical terminal value;
- Excess returns in perpetuity are allowed only with explicit justification (e.g., defensible competitive advantages).

3.2. Integration of Machine Learning Forecasts

A key innovation of the valuation framework presented in this thesis is the direct implementation of machine learning predictions into the DCF model. Conventional DCF approaches typically rely on analyst judgement or simple extrapolations of historical averages. Such methods carry a high risk of bias and may fail to capture the complexity of firm-level and sector characteristic dynamics.

However, the model developed here employs a tree-based XGBoost algorithm to predict Adobe's revenue growth for FY2025 (11.44%), as derived in Chapter II. This value serves as the anchor point for the high-growth phase of the DCF, ensuring that the projection horizon begins with an empirically validated, out-of-sample estimate rather than a mostly subjective assumption.

Implementation:

- The predicted growth rate (11.44%) is mapped into the DCF parameters revenue_growth_rate_cycle1_begin and revenue_growth_rate_cycle1_end. This anchors the first phase of the projection;
- Operating margins, sales-to-capital ratios, and reinvestment rates then evolve endogenously through the DCF's linear convergence functions, maintaining consistency between growth, profitability, and capital efficiency;
- In the Monte Carlo extension (see Section 3.3), this point estimate serves as the mean of the lognormal growth distribution, allowing uncertainty to be modeled around the central machine learning forecast.

The implementation of this novel mechanism creates a bridge between financial valuation and statistical learning. Compared to traditional valuation methods, it offers a potential competitive advantage by not only reducing subjectivity but also increasing transparency in company valuations.

3.3. Embedding Monte Carlo Simulation into the DCF Framework

Rather than relying on fixed point estimates for key inputs to the DCF approach such as revenue growth, operating margins, or financing costs, the model incorporates Monte Carlo simulation to explicitly account for uncertainty. In this setup, parameters are repeatedly sampled from calibrated probability distributions, generating a large set of alternative valuation paths.

The outcome is not a single intrinsic value, but a distribution of equity values, which reflects both the expected outcome and the full range of potential risks. This approach improves robustness, highlights downside exposures as well as upside potential, and provides a more realistic foundation for decision-making than traditional deterministic DCF models.

This design offers three advantages:

— **Transparency**: uncertainty in inputs is modeled explicitly rather than embedded in hidden judgment calls;
— **Realism**: asymmetries and bounds in financial variables are respected (e.g., growth cannot fall below –100%, margins cannot exceed 100%);
— **Decision relevance**: results are communicated through confidence intervals and probability bands, rather than a single number that hides risk.

3.3.1. Choice of Probability Distributions (Lognormal, Triangular)

The probability distributions were selected based on both the economic characteristics of each variable and the degree of uncertainty faced by the analyst.

— **Revenue Growth**: Modeled as a lognormal distribution around the machine learning forecast (mean ~ 11.4%). This choice reflects the empirical asymmetry of growth outcomes. Upside surprises (e.g., new product adoption, market expansion, AI-driven demand) are more likely and potentially large, heavy-tailed, whereas downside surprises are limited. Adobe's revenues are unlikely to collapse far below zero. The lognormal specification guarantees that growth never falls below −100%, because it models the logarithm of the growth multiplier (1+g) as normal. Since exponentials are strictly positive, the resulting growth rate satisfies $g > -1$. This property aligns with the economic reality that firms can lose all revenues but not more than that;

— **Operating Margin**: Modeled as a triangular distribution with minimum, most likely, and maximum values (~ 26%–31%–36%). The anchor point of 31% corresponds to Adobe's reported EBIT margin in 2024, while the lower and upper bounds reflect plausible variation derived from both Adobe's historical 5-year range (~29%–36.7%) and sectoral evidence. According to Damodaran's industry dataset for Software (Entertainment)[19], the median operating margin across 81 firms was approximately 32.4%**.** Importantly, Adobe's margins have been relatively stable across the past five years (mean ~ 33.6%, standard deviation ~ 2.8

[19] Damodaran, A., Margins by Sector (US). Stern School of Business, New York University. Available at: https://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/margin.html.

percentage points), which makes it appropriate to use the most recent observation as the central anchor for the distribution.

This reflects analyst judgment and cross-sectional peer data, while capping extreme outcomes. It also prevents implausible results such as –150% margins or +200% margins and centers the distribution at Adobe's observed profitability.

— **Cost of Debt (Pre-Tax)**: Modeled as a triangular distribution bounded between 4.4% and 5.5%, with a mode of 5.25%. In credit markets, the spread is defined as the excess yield over Treasuries required by lenders to compensate for default risk, and is computed as:

Spread = Average Corporate YTM (by rating class) – Treasury Yield.

For Adobe (rated A/A2), the pre-tax cost of debt was estimated by adding:

- The U.S. 10-year Treasury yield in 2025, taken as the average of daily observations from FRED (4.40%)[20], and
- The average default spread for A-rated corporates in January 2025 (~0.85%) from Damodaran's dataset[21].

This produces a central estimate of ~5.25%. The triangular distribution captures uncertainty around financing conditions while remaining symmetric and bounded. Borrowing costs have natural floors (risk-free rate + spread) and ceilings (a plausible stress case). This specification mirrors Professor Damodaran's principle of "simplicity over complexity" when a most likely range is known but detailed statistical estimation is unnecessary.

3.3.2. Simulation Design and Implementation

[20] Board of Governors of the Federal Reserve System (US), 2025. *Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity (monthly average)*. FRED, Federal Reserve Bank of St. Louis. Available at: https://fred.stlouisfed.org/series/GS10

[21] Damodaran, A., 2025. *Ratings, Interest Coverage Ratios and Default Spreads by Rating Class*. Stern School of Business, New York University. Available at: https://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/ratings.html

The Monte Carlo simulation was structured to integrate probabilistic reasoning into the DCF framework. It helped the model to be more robust and transparent. The design focused only on those variables that are both:

- essential to value creation and
- subject to the greatest uncertainty.

Other parameters remained deterministic in order to prevent unnecessary complexity and preserve interpretability.

Each uncertain variable was mapped directly to a corresponding DCF input parameter. For example, sampled values of revenue growth replaced the deterministic growth assumption in Phase 1 of the multi-phase DCF, while stochastically drawn operating margins substituted for the baseline margin input. This provided that each simulation run generated a coherent valuation scenario, consistent with the financial model.

In practice, financial variables are rarely independent. Correlations between them can range from marginal to significant, depending on context. In the present case:

- higher revenue growth often implies stronger margins (operating leverage), but also;
- higher margins and growth tend to reduce borrowing costs (credit spreads tighten).

To capture these interactions, a correlation matrix was implemented on the stochastic draws via Cholesky decomposition. This ensured that simulated scenarios reflected economically meaningful joint outcomes rather than raw and, in the result unrealistic combinations of independent inputs.

The modelling process involved conducting 1,000 trials. This scale is sufficient for the empirical distribution of equity values to converge toward stability (by the law of large numbers) and remain computationally transparent. Each trial generated a complete valuation scenario, producing a distribution of equity values rather than a single point estimate. Results are reported using the median (as a robust measure of central tendency), and percentile bounds to capture a 90% confidence interval.

From the executive point of view, such probabilistic framework allows analysts to discover not only highly subjective point estimate but also the range of plausible outcomes. It offers board. members and investors a more realistic and actionable view of intrinsic value.

3.4. Chapter Summary

This chapter has outlined the technical aspects of the study, consisting of the valuation framework that extends the classical Discounted Cash Flow model into a multi-phase structure. Instead of assuming constant growth path, the model computes three stages of corporate development (high growth, transition, and maturity) followed by a terminal value in perpetuity. Such design mirrors the economic life cycle of technology firms and provides that key parameters, such as revenue forecasts or margins, evolve realistically over time.

A distinctive unique feature was the integration of machine learning forecasts into the DCF engine. The model did not rely only on analyst judgment but anchored its first-phase revenue growth assumption in out-of-sample statistical predictions. The trained machine learning model was applied to new data it has not seen before (Adobe's 2024 fundamentals) to forecast the future revenue growth.

Uncertainty was addressed by embedding Monte Carlo simulation into the DCF. Appropriate distributions (lognormal for growth, triangular for margins and cost of debt) were chosen to reflect the economic characteristics of each uncertain variable. In addition, Cholesky decomposition was implemented to account for internal relationships between financial drivers. This created economically meaningful joint outcomes.

The simulation generated a distribution of equity values rather than a single point estimate. This probabilistic approach avoids the illusion of precision connected with deterministic models and instead provides investors and executives with a transparent view of both central expectations and tail risks. It leads to adopting more aware decisions during business processes taken upon the modeled results.

Taking everything into consideration, the methodology outlined above more informed, reasonable and data-driven decision-making. Boards and investors are no longer

dependent on purely subjective analyst inputs, but instead gain tools to derive their own findings, grounded on clearly presented ranges of outcomes.

CHAPTER IV. Empirical Case Study: Adobe Inc.

4.1. Overview of Adobe Inc.

Adobe Inc. (NASDAQ: ADBE) is one of the world's leading software companies and a key player in the Software-as-a-Service (SaaS) ecosystem. Founded in 1982 and headquartered in San Jose, California, Adobe has evolved from a traditional packaged software provider into a diversified subscription-based platform. Its operations are structured around three main business segments[22]:

- Digital Media, which includes flagship creative and document tools such as Photoshop, Illustrator, and Acrobat, offered primarily through Creative Cloud and Document Cloud subscriptions;
- Digital Experience, covering analytics, content management, and marketing automation solutions targeted at enterprise clients;
- Publishing and Advertising, a smaller legacy segment that complements the core offerings.

Adobe has carried out one of the most successful business model transitions in the software industry, moving from one-time license sales to recurring subscription revenues. This shift has not only stabilized revenues but also enhanced predictability and visibility of cash flows (characteristics highly valued in equity markets). The SaaS model has further enabled scalability and margin expansion, allowing Adobe to reinvest heavily in research and development while keeping strong profitability[23].

In recent years, Adobe has positioned itself at the forefront of artificial intelligence adoption within creative and enterprise workflows. Its proprietary AI engine, Adobe Sensei, powers generative features across Creative Cloud and marketing tools, aligning the company with broader industry trends in generative AI and creating new opportunities for monetization through premium subscriptions.

[22] Wikipedia. (2025). Adobe Inc. Retrieved from https://en.wikipedia.org/wiki/Adobe_Inc.
[23] Adobe Business. (2024). Generative AI Overview – Adobe Firefly & GenStudio. Retrieved from https://business.adobe.com/ai/adobe-genai.html.

Financially, Adobe ranks among the largest software companies worldwide, with revenues surpassing USD 19 billion in FY2023[24]. Profitability metrics remain strong, supported by gross margins above 85% and operating margins consistently exceeding 30%[25]. However, Adobe operates in a highly competitive SaaS and enterprise software landscape, facing increasing competition from both established technology giants (e.g., Microsoft in productivity software) and disruptive AI-native entrants.

Positioned against this peer group, Adobe reflects a dual identity: it remains a dominant creative software provider with extensive market share, while simultaneously transforming itself into a SaaS platform oriented toward AI-driven monetization. This duality makes Adobe a particularly relevant case for machine learning–based growth forecasting in valuation, as both historical comparable and forward-looking peers provide meaningful benchmarks.

Adobe was selected as the empirical case study in this research for three reasons:

1. **Strategic positioning**: As one of the market leaders in enterprise SaaS and creative software, Adobe serves as a benchmark for the broader SaaS and AI-enabled ecosystem;
2. **Data availability**: The company has an extensive and reliable reporting history across income statements, balance sheets, and cash flow statements, making it well-suited for both machine learning forecasts and valuation modeling;
3. **Relevance to valuation frameworks**: Adobe's growth trajectory, reinvestment policies, and AI adoption strategy illustrate the dynamics of modern SaaS firms, making it an ideal subject for testing the integration of machine learning forecasts into a multi-phase DCF model enhanced with Monte Carlo simulation.

4.2. Dataset Statistical Overview and Descriptive Statistics

4.2.1. Overview

The empirical foundation of this study is a panel dataset covering firm-level financial and macroeconomic variables for a broad set of SaaS and technology companies over the period

[24] Adobe Inc. (2023). Fiscal Year 2023 Annual Report. Retrieved from https://www.adobe.com/cc-shared/assets/investor-relations/pdfs/a56y5trgw.pdf

[25] MacroTrends. (2025). Adobe Gross Margin 2010–2025. Retrieved from https://www.macrotrends.net/stocks/charts/ADBE/adobe/gross-margin

2019–2024. This dataset served as the input for model training and validation. It combines profitability, efficiency, leverage, and reinvestment ratios with macroeconomic indicators (GDP growth, inflation, interest rates, unemployment, and consumer sentiment) as well as categorical variables encoding sectoral segments.

The aggregated sample includes both mature incumbents and younger high-growth firms. It ensures heterogeneity in growth patterns, which is a key requirement for training machine learning algorithms. This design exposes models to both stable, low-volatility profiles and venture-style hypergrowth outliers. As discussed in Chapter II, challenges involved missing data (particularly for companies with limited reporting history) and occasional inconsistencies in the FMP API feed. After the cleaning, the balanced panel covered the years 2019–2024, with an average of around 100 firm-year observations annually. While Chapter I detailed the construction of these datasets, this section provides descriptive statistics to situate Adobe's fundamentals within the broader sample.

For a more focused analysis, a separate Adobe-only dataset covering 2019–2023 was constructed. This subset was not used for training the machine learning models but instead served two complementary purposes:

1. **Benchmarking**: positioning Adobe's growth, profitability, and reinvestment metrics against the broader SaaS universe;
2. **Valuation input**: providing Adobe-specific features to the forecasting pipeline, which were then integrated into the discounted cash flow (DCF) model enhanced with Monte Carlo simulation.

4.2.2. Comparative descriptive statistics

Revenue and Profitability

The full dataset shows highly skewed revenue distributions, with mean annual revenues of USD 17.1 billion but a median of only USD 799 million, reflecting the presence of a small number of mega-cap firms alongside numerous smaller SaaS entrants. In contrast, Adobe's revenues averaged USD 15.4 billion between 2019 and 2023, placing it close to the cross-sectional mean yet far above the sectoral median. This confirms its position as a mature large-cap SaaS company. Adobe's revenue trajectory was stable, ranging from a minimum of USD 11.2 billion to a maximum of USD 19.4 billion, consistent with sustained growth.

Profitability indicators highlight Adobe's structural advantage. The firm's mean EBIT margin stood at 33.6% and its mean net margin at 30.6%, substantially exceeding the sample-wide medians (−6.5% and −8.1%, respectively). Across the dataset, average profitability was negative, driven by early-stage firms with heavy reinvestment outlays. Adobe's consistently high margins therefore place it among the most profitable actors in the SaaS ecosystem.

Reinvestment Ratios

R&D and SG&A intensity further underscore the contrasts. Adobe's R&D-to-revenue ratio averaged 17.0% and SG&A intensity 35.3%, both below the sector averages of 25.0% and 55.5%, respectively. This reflects economies of scale and established market dominance: while smaller competitors spend a majority of revenues on development and customer acquisition, Adobe has already amortized much of these costs across a global subscriber base. Nevertheless, its continued R&D spending demonstrates a commitment to innovation (e.g., Adobe Sensei AI).

Capital expenditures (CAPEX) relative to revenue also highlight scalability. Adobe's mean CAPEX-to-revenue ratio was 2.5%, compared to a sector average of 7.3%. This gap illustrates the software model's efficiency, as Adobe requires relatively limited physical reinvestment compared to infrastructure-heavy SaaS or cloud peers.

Return Metrics

Adobe also outperformed peers on return ratios. Its mean return on assets (ROA) reached 17.9%, compared to a dataset mean of −4.2%, while its return on equity (ROE) averaged 33.4%, well above the sample's distorted mean of −12.4%. Return on invested capital (ROIC) for Adobe was 29.6%, versus a highly noisy dataset mean of 22.8% with extreme outliers. These figures confirm Adobe's ability to generate strong shareholder value relative to its asset base.

Growth Dynamics

Revenue growth marks a final contrast. The dataset's median annual growth rate was 23.6%, with a mean of 30.1% but substantial variance (standard deviation ≈ 29.9%). Adobe's average growth over 2019–2023 was 16.7%, slower than the cross-sectional mean but stable (consistent with its maturity). Its growth ranged from 10.2% to 23.7%, compared to extremes of −68.9% to over +325% in the dataset. This underscores Adobe's maturity and stability relative to volatile smaller SaaS firms.

Overall, Adobe's descriptive profile reflects its status as a mature, profitable SaaS incumbent. Compared to the broader sample, it features:

- Revenues above the sector median but aligned with large-cap peers;
- Margins and return metrics well above sector averages;
- More moderate reinvestment intensities due to economies of scale;
- Growth rates below smaller firms, but stable and sustainable.

This positioning defines Adobe's current identity: while no longer in the hypergrowth phase typical of smaller SaaS firms, it continues to deliver superior profitability and efficiency metrics, making it a robust benchmark for valuation. These features suggest that machine learning forecasts for Adobe are less likely to be influenced by extreme reinvestment–growth trade-offs and more by stability, and scalability. Those are factors directly relevant for the DCF-based valuation in Chapter III.

SUMMARY STATISTICS - ADOBE INC.

STATISTICS	REVENUE	EBITMARGIN	RD_TO_REV	SGA_TO_REV	CAPEX_TO_REV	ROE	ROIC	REVENUEGROWTH
count	5.00	5.000	5.000	5.000	5.000	5.000	5.000	5.000
mean	15,367.86	0.336	0.171	0.353	0.025	0.334	0.296	0.167
std	3,368.03	0.028	0.007	0.010	0.006	0.042	0.069	0.062
min	11,171.30	0.293	0.161	0.343	0.019	0.280	0.215	0.102
25%	12,868.00	0.329	0.170	0.349	0.021	0.326	0.248	0.115
50%	15,785.00	0.343	0.170	0.351	0.025	0.329	0.293	0.152
75%	17,606.00	0.346	0.173	0.354	0.030	0.339	0.333	0.227
max	19,409.00	0.368	0.179	0.369	0.033	0.397	0.390	0.237

Table 2. Descriptive statistics for key financial ratios and performance indicators for Adobe (Revenue in millions of USD).

SUMMARY STATISTICS - FULL DATASET

STATISTICS	REVENUE	EBITMARGIN	RD_TO_REV	SGA_TO_REV	CAPEX_TO_REV	ROE	ROIC	REVENUEGROWTH
count	493.00	493.000	493.000	493.000	493.000	493.000	493.000	479.000
mean	17,103.13	-0.115	0.250	0.555	0.073	-0.124	0.228	0.301
std	65,196.82	0.375	0.126	0.284	0.309	3.137	3.740	0.299
min	46.47	-2.555	0.018	0.044	0.000	-16.401	-19.984	-0.689
25%	422.18	-0.269	0.162	0.363	0.017	-0.517	-0.149	0.132
50%	798.71	-0.065	0.232	0.526	0.031	-0.115	-0.005	0.236
75%	2,551.00	0.144	0.310	0.707	0.061	0.095	0.156	0.399
max	574,785.00	0.541	0.790	1.972	5.148	57.800	65.231	3.258

Table 3. Descriptive statistics for key financial ratios and performance indicators for the broader SaaS/ technology dataset (Revenue in millions of USD).

4.3. Log-Transformations of Variables: Experiments and Results

Logarithmic transformations are widely used in financial econometrics to mitigate skewness, stabilize variance, and approximate normality[26]. Given the strong asymmetry in several raw financial variables and in the dependent variable (revenuegrowth_t+1), two approaches were carefully tested:

1. Log-transforming independent variables (revenues, net income, R&D, SG&A)
2. Log-transforming the dependent variable (revenuegrowth_t+1).

As shown in Figure 8, the raw firm-level predictors (e.g., revenue, net income, R&D) show strong right-skewness with long tails and extreme outliers. This pattern is typical in technology firms, where a few large incumbents coexist with numerous smaller "players" (start-ups or rapidly growing young companies). Applying log-transformations yielded more symmetric, bell-shaped distributions for most predictors. Taking log(1+g) is standard in finance (for growth/returns), as it transforms multiplicative growth into additive form.

[26] Wooldridge, J. M., 2016. Introductory econometrics: A modern approach. 6th ed. Boston: Cengage Learning

However, variables that can take negative or near-zero values (e.g., net debt, free cash flow, invested capital, equity) remained irregular even after transformation. In such cases, the log-signed approach produced bimodal or clustered distributions, reflecting structural differences between firms with positive versus negative values. Thus, while the transformation generally improved normality, it could not fully eliminate non-standard distributional shapes.

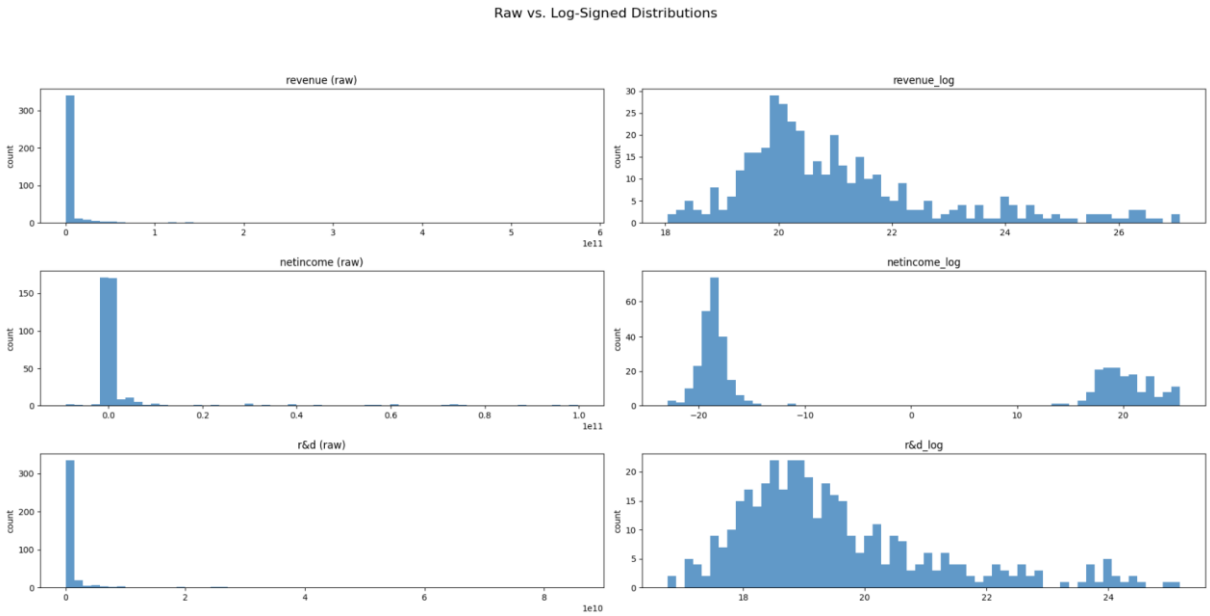

Figure 8. The comparison of the choosen independent variables distributions.

The raw distribution of next-year revenue growth (revenuegrowth_t+1) was highly skewed (3.70) with extreme kurtosis (28.02), largely driven by hypergrowth outliers (Figure 9). A log-transformation substantially normalized the distribution, reducing skewness to −0.96 and excess kurtosis to 3.21[27].

[27] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer.

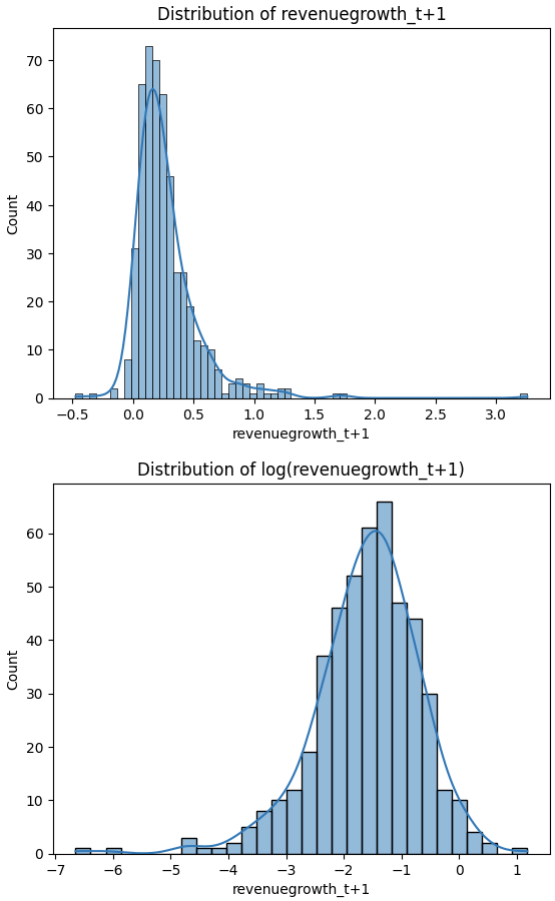

Figure 9. The raw distribution of the orignal target variable (revenuegrowth_t+1) and log target variable distribution.

Although the distributional properties were enhanced, predictive performance showed no material gains after transforming the target variable and subsequently back-transforming predictions to the original scale.

Figure 10 compares the target distributions between the training (blue) and test (orange) datasets. Several observations arise:

- **Overall shape**: In both sets, the target variable is positively skewed, with most firms exhibiting moderate next-year revenue growth between 0% and 50%. A sharp concentration appears around 10–20%, while a small number of outliers in the training set reach values above 200–300%;
- **Train vs. test comparison**: The training set has a larger sample size and hence a higher density across the distribution. Both sets overlap closely in the central region, suggesting that the main dynamics of growth are consistently represented. However, the test set appears slightly narrower, with fewer extreme observations in the right tail (capped near 100%). The training set includes rare but very high growth outliers (>300%).

Implications for modeling:

— From an econometric perspective, the skewness of the dependent variable implies potential heteroskedasticity and non-normal residuals, which may bias inference and reduce the efficiency of OLS-type models;

— From a financial perspective, the presence of extreme positive growth outliers reflects venture-style dynamics (hypergrowth firms. While financially relevant, their rarity can distort average error metrics (e.g., RMSE));

— From a machine learning perspective, the train–test mismatch in the tail behavior may reduce robustness. A model trained on extreme outliers may not be evaluated against them in testing, leading to optimistic out-of-sample metrics.

For these reasons, a log-transformation of the target variable was tested. While it reduced skewness and improved normality, it did not materially enhance predictive accuracy.

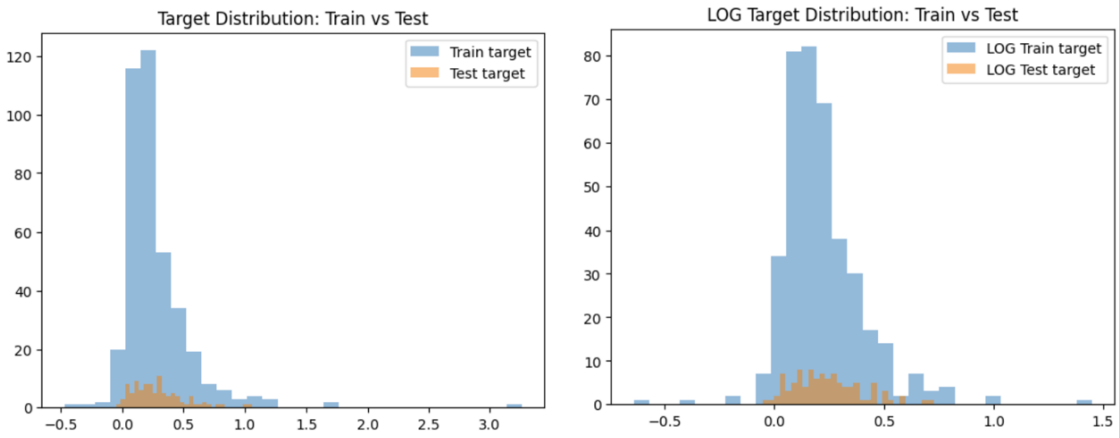

Figure 10. The distributions of the target variable in the training and test sets, presented separately for the original scale (left) and the log-transformed scale (right).

The comparison below summarizes the results:

— **Elastic Net (no log transform)**: Train $R^2 = 0.39$, Test $R^2 = 0.60$; Train MAE = 0.119, Test MAE = 0.095;

— **Elastic Net (log target)**: Train $R^2 = 0.37$, Test $R^2 = 0.58$; Train MAE = 0.116, Test MAE = 0.096;

— **XGBoost (no log transform)**: Train $R^2 = 0.63$, Test $R^2 = 0.61$; Train MAE = 0.085, Test MAE = 0.090 (test),

For XGBoost, which is robust to skewed target distributions, the log transformation had no effect. For ElasticNet, performance slightly deteriorated after transformation.

Although log-transformations improved the statistical normality of the distributions, they did not enhance predictive accuracy. Both ElasticNet and XGBoost achieved superior out-of-sample results on the original target scale, with higher R^2 and lower MAE. Consequently, the final modeling pipeline retained raw values of revenuegrowth_t+1.

This decision also supports interpretability: growth forecasts expressed in percentage-point terms can be embedded directly into the Discounted Cash Flow (DCF) valuation framework without requiring additional back-transformations or bias adjustments.

4.4. Results of Machine Learning Forecasts

4.4.1. Baseline Linear Regression (OLS)

Before evaluating regularized and tree-based approaches, the study implemented two baseline Ordinary Least Squares (OLS) regressions to provide benchmark performance levels. These models serve two purposes: first, to illustrate the limitations of naïve specification, and second, to demonstrate how careful feature selection and preprocessing can improve generalization leading to notable improvements.

Naïve OLS with all predictors:

The first baseline applied a simple linear regression on the full set of available predictors, without filtering for relevance or multicollinearity:

— **Training set**: $R^2 = 0.475$, MAE = 0.111, RMSE = 0.212, MAPE = 197%;

— **Test set**: $R^2 = 0.214$, MAE = 0.119, RMSE = 0.190, MAPE = 264%.

These results reveal clear overfitting: while the model explains nearly half of the variance in training data, its predictive power collapses on the hold-out (test) set. Moreover, the exceptionally high MAPE values further highlight instability, as observations with revenue growth close to zero cause the denominator in the percentage error to shrink,

artificially inflating the metric. While unsuitable for prediction, this specification provides a useful lower benchmark.

It is also important to note that subsequent modeling will place less emphasis on MAPE. Because of its construction, that is sensitivity to small denominators, MAPE can report disproportionately high errors even when predictions are reasonably accurate in absolute terms.

The baseline OLS regression was also used to create importance ranking of predictors identified during the data-cleaning and feature-engineering phase. The author implemented a univariate linear regression for each independent variable and computes the chosen performance metrics. The results can be seen in Table 4. Each row therefore corresponds to a separate regression fitted on the training set only. For transparency, multiple goodness-of-fit and error metrics are presented: R^2 and Adjusted R^2 (higher = better), AIC and BIC (lower = better), and error measures (MAE, MedAE, RMSE, and MAPE). The ranking itself is based solely on R^2, as it provides a straightforward measure of explanatory power in this simple specification. Because MAPE can be unstable when revenue growth values are close to zero, it is not used for ranking.

	variables	R2	Adj_R2	AIC	BIC	MAE	MedAE	RMSE	MAPE_pct
0	revenue	0.371072	0.362988	-22.678916	1.194399	0.121278	0.076758	0.231579	145.031819
1	revenuegrowth	0.301500	0.299723	10.763889	18.721661	0.124458	0.073917	0.244051	184.950788
2	sga_to_revenue	0.186328	0.184258	71.049534	79.007306	0.154629	0.106592	0.263404	185.128142
3	ebit	0.145095	0.140733	92.575829	104.512486	0.155089	0.103797	0.269996	248.749690
4	ebitmargin	0.145062	0.142886	90.591104	98.548876	0.155127	0.104138	0.270001	248.648014
5	netmargin	0.104818	0.102540	108.760385	116.718157	0.162521	0.112856	0.276283	260.527118
6	rd_to_revenue	0.094582	0.092278	113.251417	121.209188	0.164367	0.116061	0.277858	243.227461
7	CPIAUCSL	0.079194	0.076851	119.908165	127.865937	0.167014	0.106602	0.280209	280.861529
8	GDPC1	0.073383	0.071026	122.392884	130.350655	0.168074	0.112759	0.281092	281.259335
9	UMCSENT	0.066554	0.064178	125.293629	133.251400	0.169610	0.113565	0.282126	273.776724
10	roa	0.052978	0.050568	130.997093	138.954865	0.168632	0.122235	0.284170	249.704502
11	GS10	0.044715	0.042284	134.428413	142.386184	0.171043	0.123139	0.285407	285.670540
12	label	0.038252	0.020856	149.091783	180.922869	0.174286	0.125091	0.286371	248.263697

Table 4. OLS-based predictor importance screening - variables ranked by R^2 (higher = better).

Firm-level predictors such as revenue, revenue growth, and operating margins explain a considerably larger share of next-year revenue growth than macroeconomic indicators. This is expected, because company performance is primarily driven by firm-specific factors. Macroeconomic variables display relatively low variance, that is a single value applies to a broad set of firms within the sample. Such aggregation is the main reason macro variables exhibit low explanatory power in univariate regressions. However, their role may increase drastically when modeled jointly with sectoral or firm-level interactions.

Building on this observation, the study hypothesizes that while macroeconomic indicators are weak predictors on their own, they can still add **incremental forecasting value** when combined with firm-level financials. Variables such as GDP growth, interest rates, and consumer sentiment capture systemic forces (e.g., aggregate demand, financing conditions, and household spending), that accounting ratios alone cannot fully reflect. By implementing these broader economic drivers, predictive models are expected to deliver more accurate and robust forecasts of revenue growth.

Refined OLS with selected predictors

The second baseline introduced feature filtering based on Pearson correlation coefficients with the target variable, combined with the exclusion of highly collinear variables. This procedure, described in Chapter I, resulted in a more fitted specification including: **revenuegrowth, sga_to_revenue, ebitmargin, rd_to_revenue, CPIAUCSL, GDPC1, liabilities, together with sectoral categorical dummies (label_*)**.

To ensure comparability across features, predictors were standardized to zero mean and unit variance before estimation:

- **Training set**: $R^2 = 0.418$, MAE ≈ 0.118, RMSE ≈ 0.223, MAPE ≈ 179%;
- **Test set**: $R^2 = 0.609$, MAE ≈ 0.095, RMSE ≈ 0.134, MAPE ≈ 135%.

Compared to the naïve specification, the refined model delivered substantial gains in out-of-sample accuracy: the test R^2 increased from 0.214 to 0.609 (by almost 40pp) and MAE dropped from 0.119 to 0.095 (~20%). This demonstrates that careful variable selection, guided by both statistical diagnostics and economic reasoning, can meaningfully improve

model robustness. Nevertheless, persistent high MAPE values confirm that OLS regressions remain sensitive to percentage-growth targets and outliers.

The contrast between the naïve and refined OLS specifications highlights two key insights:

- **Naïve OLS demonstrates overfitting**: despite reasonable training fit, generalization to unseen data is poor, illustrating the disadvantages of including all available predictors without taking into consideration their collinearity;
- **Feature selection improves robustness**: the refined model achieved significantly better out-of-sample accuracy, highlighting the importance of statistically and economically motivated variable screening.
- **OLS has structural limits**: even with improved specification, linear regression struggles with the skewed and heavy-tailed distribution of revenue growth, motivating the transition toward regularized linear models (ElasticNet, Lasso) and nonlinear tree-based methods (XGBoost).

4.4.2. Regularized Linear Models: ElasticNet and Lasso

Regularized regressions were then applied to address predictor redundancy in a more systematic and data-driven way. Both ElasticNet and Lasso extend the classical OLS framework by introducing penalty terms that shrink coefficients toward zero, thereby reducing variance and mitigating overfitting:

- ElasticNet achieved a test R^2 of 0.601, MAE ≈ 0.095, and RMSE ≈ 0.133;
- Lasso regression produced nearly identical results, confirming the robustness of sparse, regularized models.

Compared with naïve OLS, these methods substantially improved predictive stability. While their performance was broadly similar to the refined OLS, the advantage lies in their automated feature selection, reducing reliance on manual correlation-based screening.

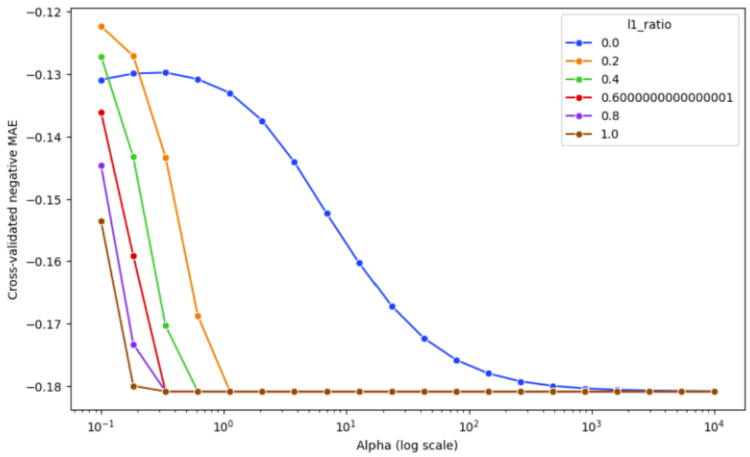

Figure 11. ElasticNet Performance Visualization: explore the interaction between regularization strength and mixing penalty (L1 vs L2).

Finally, the Lasso model highlighted a set of independent variables as the most influential drivers of next-year revenue growth. These variables are presented in **Table 5**.

	predictor	coefficient	abs_coef
5	revenuegrowth	0.118022	0.118022
9	sga_to_revenue	0.060322	0.060322
34	UMCSENT	0.026328	0.026328
6	ebitmargin	-0.025904	0.025904
41	label_Mega-Cap Tech	0.018913	0.018913
21	roa	0.018411	0.018411
29	GDPC1	-0.015791	0.015791
28	fcf_to_netincome	0.011008	0.011008
35	negative_equity_flag	0.010215	0.010215
8	rd_to_revenue	0.006288	0.006288

Table 5. Independent variables selected by the Lasso regression as the most predictive drivers of next-year revenue growth, ranked in descending order by the absolute value of their standardized coefficients.

4.4.3. Non-Linear Machine Learning Models

In order to complement linear approaches, a series of non-linear models were implemented. Decision Trees, Random Forests, and Extreme Gradient Boosting (XGBoost). These models are particularly well suited to financial forecasting, where predictor interactions are complex and relationships between variables are often non-linear. In contrast to OLS or regularized regressions, tree-based methods can naturally capture threshold effects (e.g., small changes

in interest rates may have little impact until they cross a certain threshold), and interaction terms across firms, without requiring explicit specification.

Decision Trees

The tuned Decision Tree Regressor (depth = 3, min_samples_split = 10, min_samples_leaf = 10) achieved a test R^2 of 0.485, with an MAE of approximately 0.097 and RMSE of 0.154. This performance outperformed the naïve OLS baseline but remained limited by the single tree's shallow. The result illustrates the bias–variance trade-off, that is deeper trees risk overfitting to noise in the training data, whereas shallower trees may lack the flexibility to capture the full complexity of firm-level revenue dynamics.

Random Forests

By averaging predictions across 400 trees, the Random Forest model improved generalization. On the holdout set, it achieved an R^2 of 0.587, with MAE of approximately 0.090 and an RMSE of 0.138.

Feature importance analysis (Figure 12) indicated that historical revenue growth overwhelmingly was the most influential predictor, followed by SG&A-to-revenue and capital intensity metrics such as CapEx-to-revenue. This ranking aligns with financial intuition, that firms with strong recent growth, leaner cost structures, and stable reinvestments are more likely to sustain revenue momentum.

Compared to a single shallow tree, the ensemble approach reduced variance and delivered materially stronger predictive accuracy.

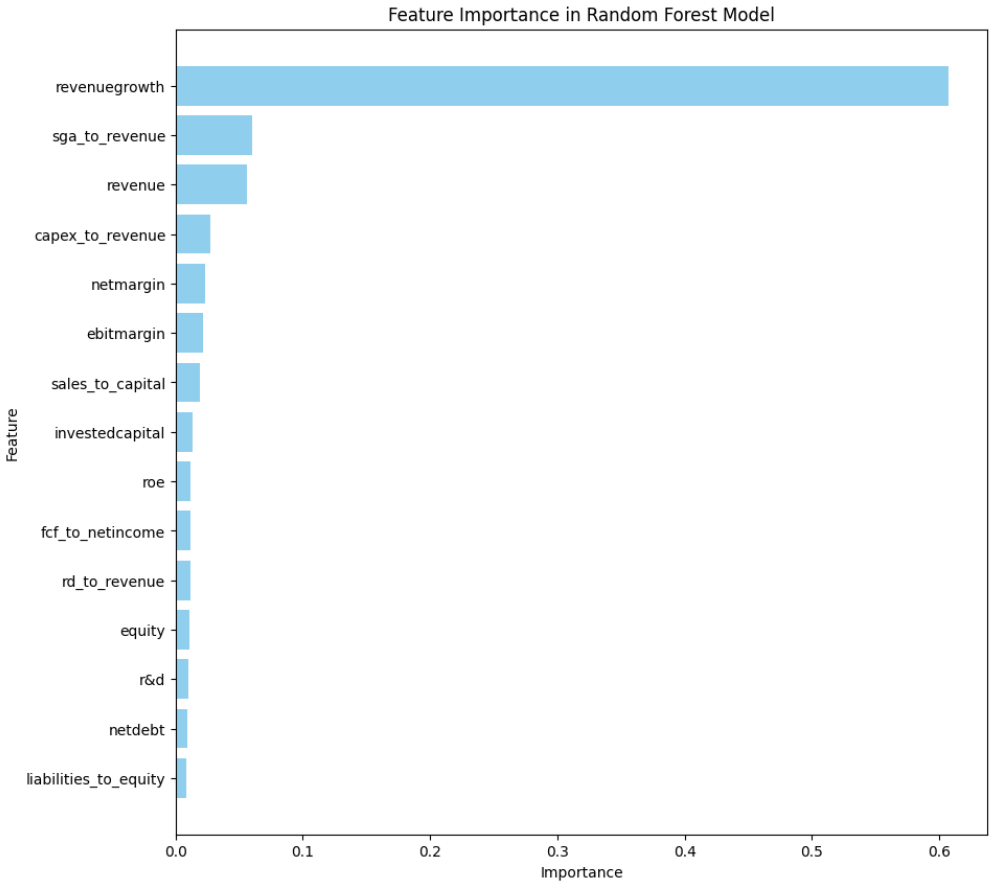

Figure 12. Random Forest Feature Importance (Mean Decrease in Impurity, normalized).

Extreme Gradient Boosting (XGBoost)

The best-performing non-linear model was XGBoost, which achieved a test R^2 of 0.612, RMSE ~ 0.133 and MAE ~ 0.090. Compared with Random Forests, XGBoost's boosting mechanism, where each tree is trained sequentially on the residuals of the previous one, captured subtler feature interactions and improved predictive efficiency.

It is worth mentioning that during training the parameters such as xgb__min_child_weight were deliberately constrained to avoid overly small splits, in order to reduce the risk of overfitting.

To better understand the drivers of XGBoost predictions, a full SHAP analysis was conducted:

— Global feature importance confirmed that revenuegrowth was the strongest predictor (mean SHAP value ~ +0.09), followed by CapEx-to-revenue (+0.04), SG&A-to-revenue (+0.03), and capital efficiency ratios (sales-to-capital, ROIC);
— Beeswarm plots revealed directional heterogeneity: high SG&A ratios consistently reduced predicted growth, whereas CapEx-to-revenue had a dual efect, enhancing growth when associated with productive reinvestment, but reducing it when linked to inefficiency (inefficient spending);
— Waterfall plots provided firm-level (local) explanations. For insatnce, in one case the model forecasted ~0.36 growth (vs. baseline ~0.26), driven positively by strong historical revenue growth (+0.11) and low SG&A (+0.05), but offset by low ROIC (–0.03) and high CapEx (–0.04).

The SHAP insights mirror established corporate finance theory:

— **Momentum effect**: past revenue growth is the most reliable indicator of short-term growth[28];
— **Cost efficiency**: low SG&A indicates scalable operations, which the model consistently rewarded;
— **Capital allocation**: CapEx and ROIC effects were context-dependent, reflecting that reinvestment can either fuel growth when productive or erode value when inefficient. Growth alone is not enough, it must be profitable growth. This aligns with corporate finance theory that reinvestment only drives long-term growth when it earns ROIC above the cost of capital. As Professor Damodaran states: *"Growth is determined by how much you reinvest and how well you reinvest: Growth rate = Reinvestment Rate × Return on Invested Capital (ROIC)[29]"*.

[28] Jegadeesh, N. and Titman, S., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. Journal of Finance, 48(1), pp.65–91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x.
[29] Damodaran, A., n.d. Terminal value. Stern School of Business, New York University. Available at: https://pages.stern.nyu.edu/~adamodar/pdfiles/country/TerminalValue.pdf.
Damodaran, A., 2012. Investment valuation: Tools and techniques for determining the value of any asset. 3rd ed. Hoboken, NJ: John Wiley & Sons.

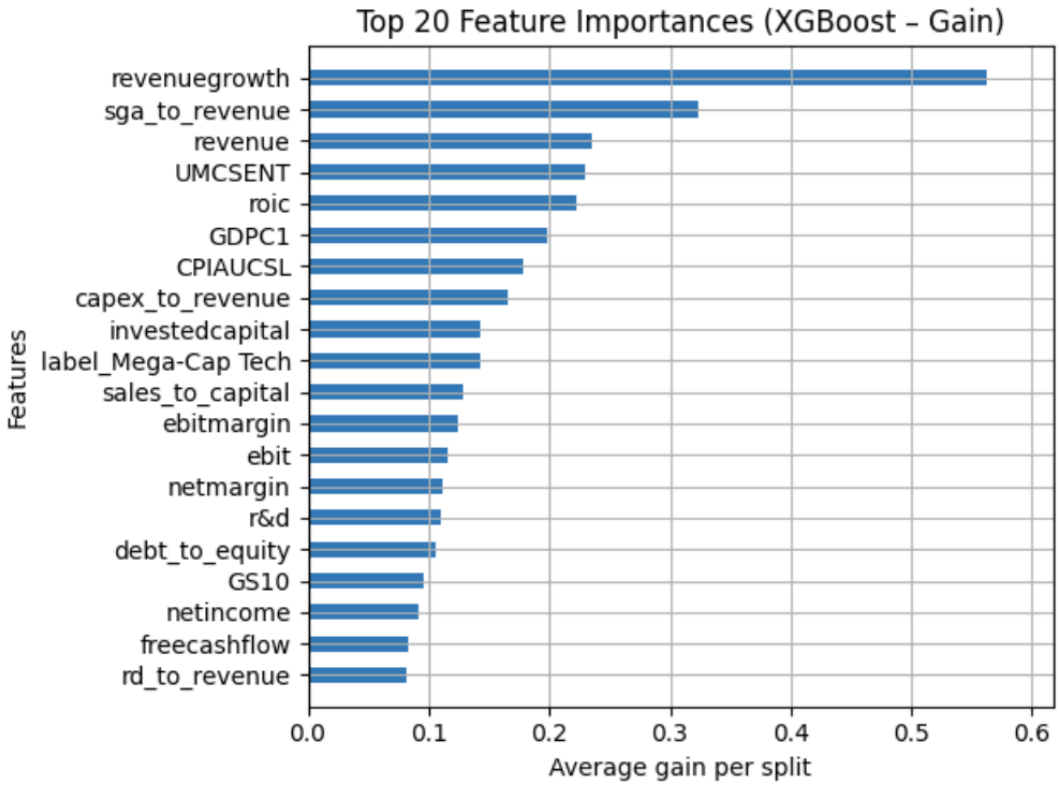

Figure 13. Comparison of feature importance in the XGBoost model: Gain-based importance.

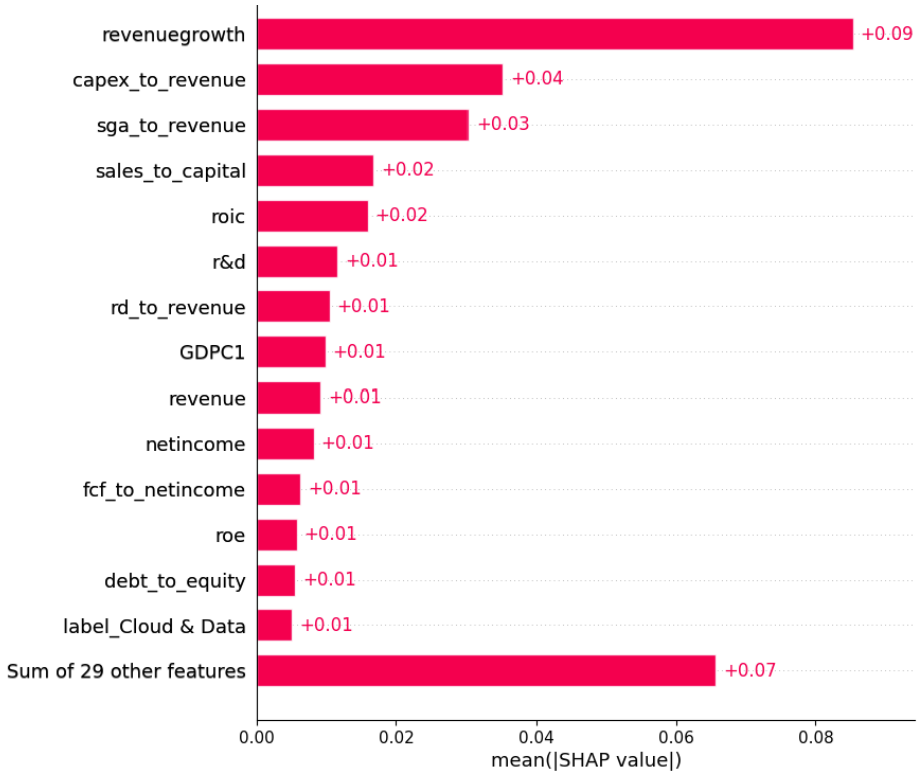

Figure 14. Comparison of feature importance in the XGBoost model: SHAP global importance, measured as mean absolute SHAP values per feature.

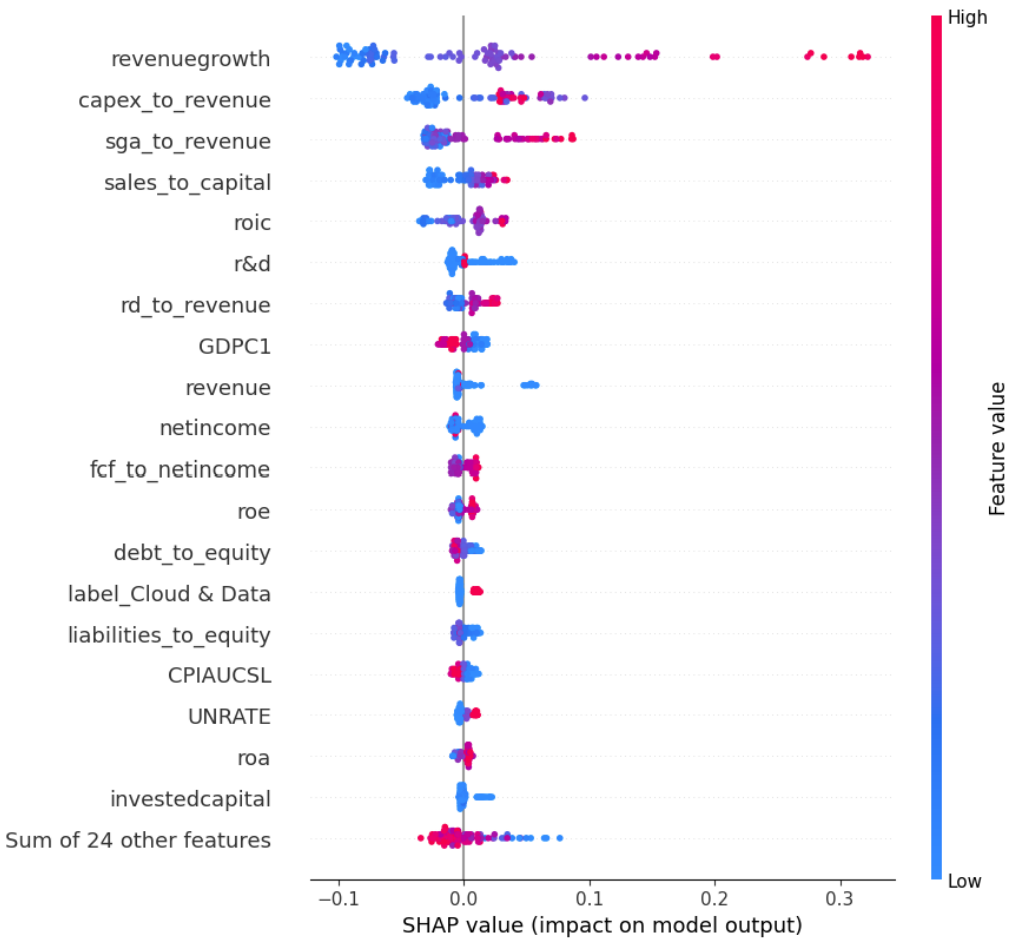

Figure 15. SHAP beeswarm plot for the XGBoost model, showing the distribution of feature impacts on revenue growth forecasts. Each point represents a single firm-year observation, with color indicating the feature value (red = high, blue = low).

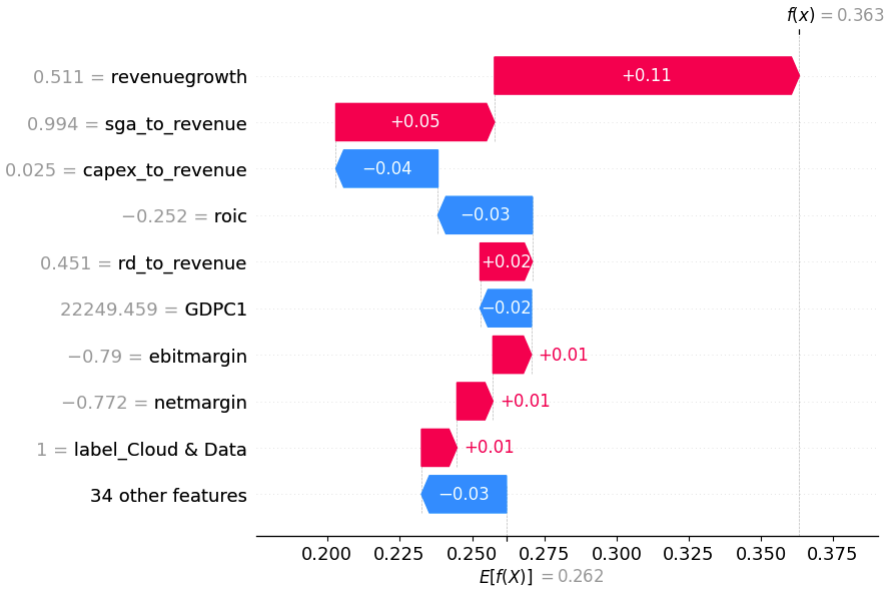

Figure 16. SHAP waterfall plot for a single firm-year observation, providing a local explanation of one prediction. Positive (red) and negative (blue) bars indicate the contribution of each feature to shifting the prediction away from the model's baseline value.

4.4.4. The final prediction result

The final forecasting step applied the best-performing model (the tuned **XGBoost regressor**) to Adobe's most recent financial and macroeconomic data for fiscal year 2024. The model incorporated both firm-level ratios (e.g., profitability, reinvestment intensity, efficiency) and macroeconomic drivers (GDP growth, interest rates, consumer sentiment), together with sectoral classifications.

Based on these inputs, the model predicted Adobe's revenue growth for fiscal year 2025 at approximately 11.44%. This estimate fits with Adobe's historical growth pattern in the SaaS ecosystem. It shows that Adobe keeps expanding steadily thanks to innovation, but its growth is also shaped by the broader macroeconomic environment. The gap between Adobe growth and the cross-sectional median of approximately 21% observed in the broader dataset of SaaS and AI-enabled firms (2019–2023) reflects Adobe's position as a mature platform. It no longer achieves the rapid expansion typical of younger AI-native competitors but instead delivers stable growth connected with high margins and robust free cash flow generation.

From a statistical perspective, this result represents an out-of-sample prediction from the model trained on a diverse panel of SaaS and AI-enabled firms between 2019–2023.

The forecast is backed by an R^2 of 0.612 and MAE ~ 0.090 (see Section 4.4.3), suggesting that the model achieves both explanatory strength and reliable accuracy. The use of cross-validation and out-of-sample testing ensured robustness, while SHAP interpretability confirmed that the drivers of Adobe's growth align with established financial theory.

Overall, the model's final forecast of 11.44% revenue growth for 2025 serves as the key input for the subsequent multi-phase DCF valuation with Monte Carlo simulation (Chapter 3). This integration of machine learning forecasts into valuation not only provides a data-driven estimate but also embeds statistical approaches into forward-looking corporate finance analysis.

4.5. Valuation Results from the DCF with Monte Carlo

To integrate uncertainty into the valuation, the baseline multi-phase DCF was extended with a Monte Carlo simulation. This approach replaces single-point estimates of key drivers with probability distributions, generating a full distribution of potential equity values rather than a single deterministic figure.

The simulation explicitly accounted for uncertainty in:

- **Short-term revenue growth**, modeled with a lognormal distribution ($\mu = 0.097$, $\sigma = 0.15$ in log-space, consistent with an expected mean growth of about 11.4%);
- **Operating margin**, assumed to vary around a mean of 31.3% with $\sigma = 2$ percentage points (normal distribution);
- **Pre-tax cost of debt**, modeled with a triangular distribution reflecting plausible spreads (4.40%–5.50%, mode 4.78%).

A correlation structure was imposed to reflect economic realism: higher growth correlates positively with margins (+0.6). There are three main reasons for that:

1. **Economies of scale**[30]:

[30] Investopedia (2015) *Which industries tend to have the greatest EBITDA margins?* Available at: https://www.investopedia.com/ask/answers/052015/which-industries-tend-have-greatest-ebitda-margins.asp

— As firms grow revenues, fixed costs (e.g., overhead, infrastructure) get spread over a larger base;
— This often improves margins, especially in SaaS businesses where software delivery costs are low once the product is built;

2. **Innovation-driven growth**:

— High R&D investment may first depress margins, but once it pays off, it can drive both revenue growth and higher profitability.

On the other hand, margins and growth are both inversely related to borrowing costs (–0.4 and –0.15 respectively). This ensures that simulated outcomes are consistent with how fundamentals move together in practice.

Across 1,000 simulations, the equity value distribution was right-skewed, with a long right tail reflecting occasional upside scenarios where high growth coincides with strong margins and favorable financing costs.

— **Mean equity value**: $108.8 billion;
— **Median equity value**: $106.3 billion;
— **Standard deviation**: $22.3 billion.

The 90% confidence interval (P5–P95) ranges from $77.7 billion to $149.5 billion, while the broader 95% confidence interval (P2.5–P97.5) extends from $72.5 billion to $160.8 billion. The downside risk (P5) indicates that in 1 out of 20 scenarios, Adobe's equity could fall below $77.7 billion.

Statistic	Value
Mean	108,812.75
Median	106,288.25
90% CI (P5–P95)	77,656.80 → 149,538.90
95% CI (P2.5–P97.5)	72,505.35 → 160,767.77
Downside Risk (P5)	77,656.80

Table 6. Adobe equity value table (Units in USD millions).

Figure 16 shows the resulting distribution of equity values. The histogram and kernel density estimate highlight the asymmetric, positively skewed shape of the distribution. The

median (black line) is flanked by the 5th and 95th percentiles (red dashed lines), which form the 90% confidence interval.

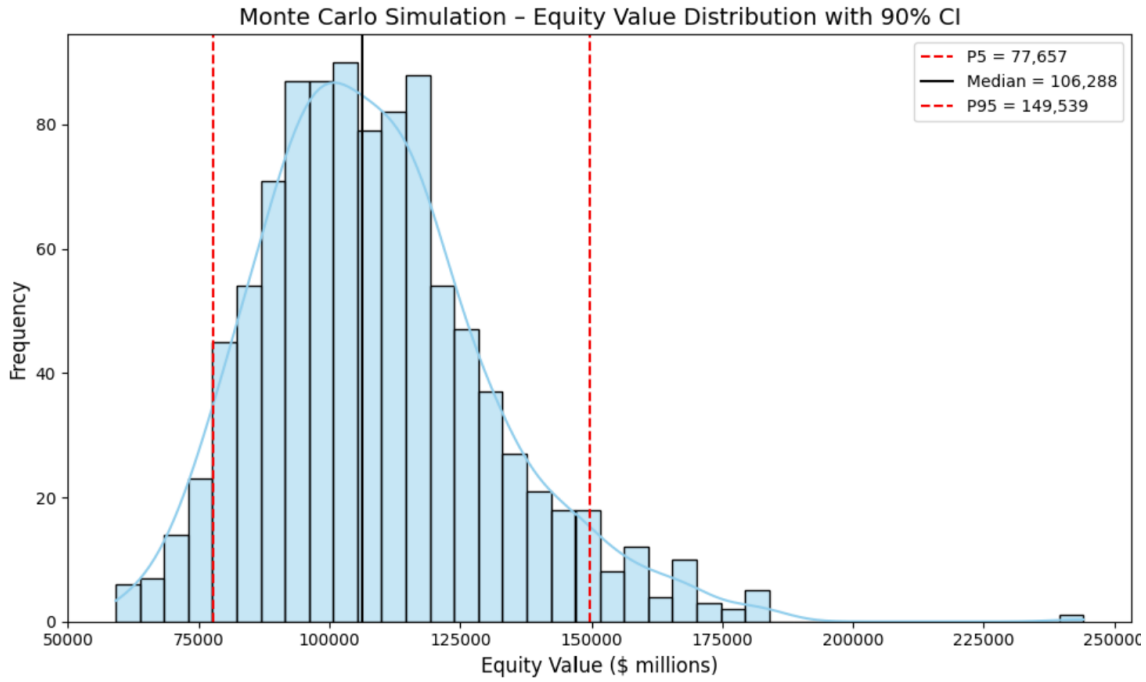

Figure 17. Monte Carlo Simulation - Equity Value Distribution with 90 % Confidence Interval.

4.6. Interpretation of Findings

Collectively, the non-linear models demonstrated clear improvements over linear baselines, with XGBoost providing the best balance between accuracy and interpretability. The integration of SHAP analysis ensured not only predictive accuracy but also transparency, reinforcing the model's value in decision-making contexts. These results confirm that growth forecasting in the software sector benefits from approaches capable of handling non-linearities, interactions, and heterogeneous firm behavior.

Bringing the results together, the Monte Carlo DCF was anchored in the machine learning–predicted revenue growth for 2025 (11.44%). This forecast, derived from an XGBoost regression, served as the starting point for the high-growth phase of the DCF model. By embedding this prediction within a probabilistic valuation framework, the analysis ensures that uncertainty around growth, profitability, and financing costs is fully reflected in the final equity distribution.

The simulation results indicate a median equity value of $106B, with a 90% confidence interval spanning $78–150B. This means that in nine out of ten scenarios, Adobe's equity value is expected to lie within this band, while more extreme values are possible in the long tails of the distribution. The right skew highlights the possibility of occasional strong upside cases (e.g., breakthrough product adoption), while the left tail is shorter, consistent with the limited downside inherent in a subscription-based business model.

This analysis emphasizes two key points:

1. **Single-point estimates can be misleading.** While the baseline DCF suggested an equity value near $106 billion, the Monte Carlo results demonstrate that plausible equity values span a wide range (roughly $70–160 billion);
2. **Skew and tail risk matter.** The distribution's right skew reflects that upside surprises can generate significant additional value, whereas downside is limited. For risk-averse investors, the 5th percentile downside ($78B) is more relevant than the extreme right-tail outcomes;
3. **Market comparison.** Relative to Adobe's prevailing market capitalization (approximately $152 billion[31]), the median simulated valuation suggests the stock is trading toward the upper end of its 90% confidence interval, implying limited upside (since most simulations fall below the current price) and non-negligible downside risk (since the stock could revert toward the median or lower percentiles under more moderate assumptions).

Furthermore, the results also confirm **Hypothesis 2** of this thesis, that macroeconomic conditions improve predictive accuracy when integrated into the machine learning framework. While firm-level drivers such as revenue momentum and SG&A efficiency ranked highest in importance, the feature-importance diagnostics (OLS screening, Random Forest, and SHAP analysis) consistently showed that macroeconomic variables, especially GDP growth, interest rates, and consumer sentiment, made meaningful contributors to the forecasts. These factors capture systematic variation that firm-specific ratios alone cannot explain. Firm performance cannot be fully understood in isolation, as the external economic environment play an integral role in shaping short-term growth outcomes

[31] As of August 19, 2025, Adobe's market capitalization was $151.5 billion, while its enterprise value was $152.4 billion. Retrieved from https://finance.yahoo.com/quote/ADBE/key-statistics/

Taking everything into consideration, this methodology demonstrates that, while Adobe is a high-quality business, at current levels (at $152 B), Adobe trades close to the 95th percentile of simulated outcomes, meaning investors at this price are effectively paying for optimistic assumptions to materialize.

4.7. Chapter Summary

This chapter integrated machine learning–based forecasting with a probabilistic valuation framework to assess Adobe Inc. within the SaaS and AI-enabled software sector. The analysis began with a statistical overview situating Adobe against its peers, demonstrating its status as a mature large-cap incumbent with superior profitability, efficiency, and cash flow stability but slower revenue growth relative to younger, high-volatility firms.

Applying the tuned XGBoost model to Adobe's 2024 financials yielded a forecasted revenue growth of 11.44% for 2025. This value served as the anchor for the DCF model, which was enhanced with Monte Carlo simulation to capture uncertainty in growth, margins, and financing costs.

The simulation produced a right-skewed equity distribution with a median of $106B and a 90% confidence interval of $78–150B. Compared with Adobe's prevailing $152B market capitalization, the results suggest the stock trades near the upper end of fair-value estimates, with limited upside and non-negligible downside risk. Overall, the findings demonstrate the value of integrating machine learning forecasts with probabilistic valuation to capture both central expectations and tail risks in corporate finance.

Bibliography

Adobe Inc., 2023. Fiscal Year 2023 Annual Report. Retrieved from https://www.adobe.com/cc-shared/assets/investor-relations/pdfs/a56y5trgw.pdf

Adobe Business, 2024. Generative AI Overview – Adobe Firefly & GenStudio. Retrieved from https://business.adobe.com/ai/adobe-genai.html.

Board of Governors of the Federal Reserve System (U.S.). (n.d.). Federal Funds Effective Rate [FEDFUNDS]. Retrieved August 19, 2025, from FRED, Federal Reserve Bank of St. Louis: https://fred.stlouisfed.org/series/FEDFUNDS.

Board of Governors of the Federal Reserve System (U.S.). (n.d.). Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity, Quoted on an Investment Basis [GS10]. Retrieved August 19, 2025, from FRED, Federal Reserve Bank of St. Louis: https://fred.stlouisfed.org/series/GS10.

Boston Consulting Group, 2025. Rule of 40: Lessons from Top Performers in Software. Available at: https://www.bcg.com/publications/2025/rule-of-40-lessons-from-top-performers-software.

Damodaran, A., 2012. Damodaran on valuation: security analysis for investment and corporate finance. 3rd ed. Hoboken, N.J.: John Wiley & Sons.

Damodaran, A., 2012. Investment valuation: Tools and techniques for determining the value of any asset. 3rd ed. Hoboken, NJ: John Wiley & Sons.

Damodaran, A., 2024. The little book of valuation: how to value a company, pick a stock and profit. Updated ed. Hoboken, N.J.: John Wiley & Sons.

Damodaran, A., 2025. Ratings, Interest Coverage Ratios and Default Spreads by Rating Class. Stern School of Business, New York University.

Damodaran, A., n.d. *Teaching: Valuation*. [online] Stern School of Business, New York University. Available at: https://pages.stern.nyu.edu/~adamodar/.

Data Gravity, 2023. Synopsys and Cadence: The $160B Unsung Heroes of Silicon Valley. Available at: https://www.datagravity.dev/p/synopsys-and-cadence-the-160b-unsung.

Federal Reserve Bank of St. Louis (FRED), 2025. Federal Reserve Economic Data (FRED).

Financial Modeling Prep (FMP), 2025. Financial Modeling Prep API.

Grus, J., 2019. Data science from scratch: first principles with Python. 2nd ed. Sebastopol, CA: O'Reilly Media.

Géron, A., 2023. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. 3rd ed. Sebastopol, CA: O'Reilly Media.

Hastie, T., Tibshirani, R., & Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer.

Investopedia, 2023. Fundamental analysis: definition, how it's used, and example. Investopedia.

Jegadeesh, N. and Titman, S., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. Journal of Finance, 48(1), pp.65–91.

Koller, T., Goedhart, M. and Wessels, D., 2020. Valuation: measuring and managing the value of companies. 7th ed. Hoboken, NJ: John Wiley & Sons.

Kenton, W., 2022. The long-term impacts of the COVID-19 K-shaped recovery. Investopedia. [online] Available at: https://www.investopedia.com/long-term-impacts-of-the-covid-19-k-shaped-recovery-5200711.

MacroTrends., 2025. Adobe Gross Margin 2010–2025. Retrieved from https://www.macrotrends.net/stocks/charts/ADBE/adobe/gross-margin

Pouladian, B., 2023. Software is dead? Not so fast — vertical SaaS is thriving. Medium, 31 July. Available at: https://medium.com/@ben_pouladian/software-is-dead-not-so-fast-vertical-saas-is-thriving-b6218207b97c (Accessed: 19 August 2025).

Reuters, 2025. Five years on, the economic impact of COVID-19 lingers. [online] Available at: https://www.reuters.com/business/healthcare-pharmaceuticals/five-years-economic-impact-covid-19-lingers-2025-03-08/.

ServiceNow What is enterprise SaaS? Available at: https://www.servicenow.com/products/it-asset-management/what-is-enterprise-saas.html.

U.S. Bureau of Labor Statistics. (n.d.). Consumer Price Index for All Urban Consumers: All Items in U.S. City Average [CPIAUCSL]. Retrieved August 19, 2025, from FRED, Federal Reserve Bank of St. Louis: https://fred.stlouisfed.org/series/CPIAUCSL

Wikipedia, 2025. Adobe Inc. Retrieved from https://en.wikipedia.org/wiki/Adobe_Inc.

Wikipedia, 2025. Cross-validation (statistics). Available at: https://en.wikipedia.org/wiki/Cross-validation_(statistics).

Wooldridge, J. M., 2016. Introductory econometrics: A modern approach. 6th ed. Boston: Cengage Learning

List of Figures

List of Tables