Hochschule Fresenius University of Applied Sciences

Faculty of Economics and Media

International Business School

Industrial Engineering and International Management

Cologne Campus

Supervised Learning in Business: Evaluating and Explaining Regression-Based Methods for Pattern Discovery

Master's Thesis

in partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

Tanmay Ubhate.

Matriculation number: 400338253

1st examiner: Prof. Dr. Stephan Huber

2nd examiner: Amit Ray

Due Date: 14th July 2025

Abstract

Companies often struggle to understand the underlying patterns in their data, leading to sub-optimal strategic decisions. Machine learning algorithms based on regression techniques can help uncover these patterns. However, these methods are sometimes unfairly dismissed. The reasons for this are a lack of understanding and problems in interpreting the results. A comprehensive assessment and explanation of these techniques for business practice can help managers understand the effectiveness and interpretability of different regression methods to ultimately provide organizations with actionable insights for data-driven decisions.

I Table of Contents

II List of Tables

II Table of Figures

III List of Abbreviation

CRISP-DM	Cross-Industry Standard Process for Data Mining
CSV	Comma-Separated Values
EDA	Exploratory Data Analysis
HITL	Human-in-the-Loop
IDE.	Integrated Development Environment
KPI	Key Performance Indicator
L1	Lasso Regularization Penalty (Absolute Value Penalty)
L2	Ridge Regularization Penalty (Squared Value Penalty)
MAE	Mean Absolute Error
ML	Machine Learning
OLS	Ordinary Least Squares
R&D	Research and Development
R^2	Coefficient of Determination
RMSE	Root Mean Squared Error
SEC	Securities and Exchange Commission (U.S.)
SG&A	Selling, General, and Administrative Expenses
SHAP	SHapley Additive exPlanations

1. Introduction

In the era of abundant financial data and unmatched computing power, data-driven decision-making has become an imperative business advantage particularly in finance, where extremely significant decisions must be grounded in data-driven insights. S&P 500 quarterly reports generate massive amounts of revenue, cost, income, and investment data that are generally not tapped to their full potential using traditional means like spreadsheets and basic regressions because they cannot uncover complex, non-linear patterns. Kelly & Xiu (2023) addressed a considerable momentum to leverage the flexible modelling of supervised machines, learning specifically regression-based approaches as a means of achieving diagnostic, forecasting, and decision support value from past financial KPIs variables. This trend reflects an appreciation for recognizing that more sophisticated analytical models are required for useful interpretation and utilization of corporate financial metrics.

Supervised machine learning is the act of instructing algorithms to learn mappings of input attributes to target predictions to enable models to predict new data. In this area, regression remains a centerpiece of financial modeling since it can be utilized to predict continuous values like return on assets or net income using multiple financial indicators. However, traditional methods like Ordinary Least Squares (OLS) fail under multicollinearity, outliers, and non-linearity commonly found in financial data as James, Witten, Hastie et al. (2009) inferred in their statistical learning lessons with application of python. Accordingly, Tibshirani (1996) have preferred ridge regression, based on L_2 regularization to fight multicollinearity by shrinking the size of coefficients, and Lasso regression, further enhancing interpretability by performing variable selection using L_1 regularization. Moreover, Polynomial regression offers model flexibility in its ability to capture non-linear interaction, yet susceptible to overfitting when in need of control. Empirical evidence suggests that regularized regression methods always outperform the standard OLS under real-world financial data with more robust coefficient estimates and better prediction performance Hastie et al. (2009)

While highly promising, regression-based ML models are likely to be misinterpreted or underutilized in practical applications. The cause of the unease is what research scholars have termed the interpretability–accuracy trade-off (Bell et al. 2022). While more complex models will provide greater predictive accuracy, they will so at the cost of lucidity, making it increasingly challenging for non-technical stakeholders to fully understand or sign off on the findings. In domains like finance where the selection influences market stability, shareholder value and regulation, interpretability is not a concession but a necessity. To eliminate this Lipton (2018) implies stakeholders must be able to ask and answer, “Why did the model predict this?

This is the job of explainable AI (XAI), a topic of interest that is gathering momentum rapidly and one which makes complicated models more transparent and actionable. These are the problems tackled by this thesis in comparing four regression models from a real S&P 500 company dataset. Tested models are Linear regression (the control), ridge regression (which uses L2 regularization to handle multicollinearity), lasso regression (which uses L1 regularization to select features), and polynomial regression (which handles non-linear interaction using higher-order terms). Each model is not just tested for its predictability a measure in simple terms such as coefficient of determination R^2, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) but also interpretability using transparency of coefficients, feature importance, and SHAP (SHapley Additive ExPlanations) values.

The data, collected from quarterly financial reports of S&P 500 companies, includes variables such as Total Revenue, Operating Income, Net Income, Research & Development (R&D) expenditure, Cost of Goods Sold, and Interest Expense. These are critical firm performance and strategy metrics. Since each is intricately connected with the others, they make for an ideal testing ground for model regression comparison. Notably, such variables are often collinear (e.g., gross profit and operating income) or strategically controlled (e.g., R&D timing for earnings influence), and thus diligent preprocessing and human oversight is essential Chan et al. (2022)

1.1 Supervised Learning in Financial Forecasting

Kelly & Xiu (2023) discuss how even if machine learning has been very predictive in finance applications, its implementation in financial institutions at scale has been slow due to operational and interpretability challenges. They argue that ML's higher predicting capacity is appreciated for a variety of tasks such as forecasting returns, volatility, and even macroeconomic aggregates, but financial firms are yet hesitant to use these models in real-world decisions. This aversion is not motivated by model poor performance, but by demands for compliance with regulations, explainability, worker objections against automation, and legacy reliance on rule-based or linear models that are easier to audit and validate. The authors also point towards the reality that in the lack of interpretable tools and rightful organizational support structures, even high-performing ML models continue to be disconnected from strategy formation and risk management procedures.

1.2 The Interpretability–Accuracy Dilemma

The Black-box problem of ML models can yield high accuracy without intelligible explanations states Lipton (2018). In finance, the trade-off is further compounded. Depending on black-box algorithms can undermine confidence, generate compliance concerns, and lead to unstable financial systems. Interpretability therefore becomes a practical imperative.

Interpretability is of different kinds. Intrinsic interpretability is when models are transparent by nature, for example, linear regression. Post-hoc interpretability, however, means techniques like LIME Ribeiro et al. (2016) and SHAP Lundberg & Lee (2017) that explain model outputs after the model has been trained. Rudin (2019) argues for the preference of inherently interpretable models that both are essential: simple models are useful for audit and compliance, but post-hoc aids allow complex models to be used responsibly. This thesis benefits from both strategies. Basic models are analyzed straight out of their coefficients, while more advanced models are described with SHAP values, which allocate each feature's contribution to a particular prediction in a fair way. This ensures interpretability is not compromised at all, even when accuracy is prioritized.

1.3 The Domain Expertise Role

Another of the key philosophical underpinnings of this thesis is that human judgment must not be substituted by machine learning but rather supported instead, Choudhury et al. (2021) emphasizes the central role that domain expertise plays in ensuring that ML systems are contextually meaningful and strategically situated. Human understanding is needed from feature choice to evaluation. For instance, while a model could determine that Operating Income is a leading indicator of Net Income, a subject matter expert might very well question whether such recent earnings management practices could overstate this relationship.

This fusion approach machine precision with human context is the foundation of this thesis. Gunning et al. (2019) argue that explainable AI must be directed towards human purposes, including causal reasoning, fairness, and actionable decision-making. Without such alignment, even technically correct models may result in erroneous conclusions.

1.4 Methodological Alignment

This background is overtly informed by the methodological stance adopted in this study. The research involves a parallel comparison of four regression models Linear, Ridge, Lasso, and Polynomial Regression on the same data. The data includes real-world financial information extracted from multiple quarters of S&P 500 companies, offering width and depth for model comparison. All models are assessed on a combination of quantitative metrics, including the coefficient of determination R^2, RMSE, and MAE, to determine predictive performance. Qualitative assessments are used in tandem to adjust interpretability based on the legibility of model outputs and the usefulness of explanations. To enable interpretability, visual and numerical aids such as coefficient plots and SHAP values are used. The overall purpose of the methodological approach is to come up with an effective decision-support system that will enable practitioners and analysts to select appropriate regression models according to their business scenarios.

1.5 Thesis Objectives

The main objectives of this thesis are to address key frailties in the application of machine learning regression models in financial analysis. The first objective is to exhaustively evaluate the performance of various regression models in forecasting essential financial indicator net income. A second objective is to make a meticulous comparison of these models, not only on predictive performance, but also on how transparent and interpretable they are to end-users. A third objective is to ascertain the real-world usability of these models in organizational settings, where explainability, compliance, and usability are often just as important as accuracy. Finally, the thesis aims to offer a set of best-practice guidelines that can inform financial practitioners in selecting regression models based on the specific characteristics of their data and the decision-making needs of their company.

1.6 Structure of the Thesis

The structure of this thesis is designed to map a methodical exploration of the intersection of regression-based machine learning and business decision-making. It begins with a selective review of the literature on regression methods, interpretability frameworks, and machine learning applications in business analysis. This is followed by a concise explanation of the research approach, covering data acquisition, preprocessing, model implementation, and evaluation metrics. Results are then presented, giving a side-by-side comparison of model performance on both interpretability and accuracy measures. A lengthy discussion is given, exploring the managerial implications of the results and weighing the trade-offs between usability and complexity. The thesis is rounded out with a presentation of best practices, a discussion of research limitations, and recommendations for future research in the field of interpretable machine learning for finance.

2. Literature Review

The following literature review Table 1. support both the theoretical grounding and the practical orientation of this study. It begins with a critical look at the limitations of traditional financial modeling and the emerging necessity of machine lesarning in enabling more responsive, data-driven decisions. Readers are encouraged to interpret each section not in isolation, but as part of a broader narrative that connects model choice, human judgment, and organizational needs. The review moves systematically from foundational regression methods to modern considerations like interpretability and deployment frameworks, culminating in a holistic view of what makes machine learning usable, reliable, and aligned with business priorities.

Literature Source	Core application to thesis	Application in This Study
Kelly & Xiu (2023)	Advocated supervised ML in financial modeling, emphasized slow adoption due to interpretability & compliance	Justified the use of regression for net income prediction and need for transparency
Hastie et al. (2009)	Explained Linear, Ridge, Lasso, Polynomial regression with practical ML implications	Guided model selection, regularization rationale, and interpretation strategy
Lipton (2018)	Introduced the interpretability–accuracy dilemma in ML models	Framed the need for SHAP and interpretability alongside model accuracy
Rudin (2019)	Interpretable models in place of black-box ones for critical decisions	Supported the thesis's SHAP-based transparency layer and use of linear models
Choudhury et al. (2021)	Promoted human-in-the-loop ML, emphasizing domain expertise in model alignment	Inspired the human validation of features, feature pruning, and interpretation of SHAP values

Literature Source	Core application to thesis	Application in This Study
Montesinos López et al. (2022)	Defined model evaluation criteria (R^2, MAE, RMSE), warned against overfitting and leakage	Shaped evaluation design and led to removal of derived/correlated variables
Lundberg & Lee (2017)	Developed SHAP for model interpretability with local and global explanations	Applied SHAP to all models to ensure stakeholder-aligned transparency
Brzozowska et al. (2023)	Advocated CRISP-DM framework in business data mining	Used to guide data understanding, preprocessing, modeling, and evaluation structure

Table 1. Foundational Literature on Regression Methods and Business Usability.

This literature review begins by defining the constraints of traditional approaches and the growing importance of data-driven decision-making before analyzing how regression-based machine learning can aid strategic business decisions. It then describes key machine learning concepts, highlighting regression approaches commonly used in business forecasting. The review discusses the merits, drawbacks, and business use cases for several regression models, including ridge, lasso, polynomial and linear. To place these methods in the context of practical uses, the CRISP-DM model is a structured process for deploying machine learning models in real-world corporate settings. The review also deals with the importance of human involvement, and the interpretability and explainability of models, especially in applications involving decision-making that affects financial issues or compliance standards. Finally, it views the measurement of model performance to include not just accuracy, but interpretability, reliability, and alignment with organizational objectives, emphasizing the need for a technically sound model to enable open and reliable decision-making.

2.1 Understanding why business needs pattern recognition in data to make better strategic decisions

The rapid advancement of AI together with Machine Learning has caused an increased adoption of pattern recognition in business analytics operations. Businesses rely on pattern recognition to extract correlations and structures from data which enables automation while improving strategic decision-making and actionable insights. Pattern recognition in business environments relies on supervised learning models and unsupervised learning models to use historical data for classification and prediction of upcoming event Jakubik and colleagues (2024) addressed traditional model centric AI is overtaking data centric AI and implies business models that prioritize systematized data creation, refinement, and extension over model optimization since modern AI performance attained through traditional model-centric methodologies has peaked.

Modern high-dimensional situations demand decision-making approaches that traditional frameworks based on human expertise together with historical heuristics and linear models sometimes fall short of addressing their requirements. The thesis points out that these systems have natural constraints through human mental biases and excessive information and their inability to process large, complicated datasets instantly. As a result, this leads individuals to produce decisions that are not optimal and take longer to implement while reacting to situations instead of being proactive. The application of conventional methods depends on fixed regulations and unchanging models which prove insufficient for accommodating fresh or shifting data thus businesses face substantial risks in environments that cannot be predicted (Batz et al., 2025).

Systems based on machine learning present fresh methods because they track patterns while boosting predictions through large dataset analysis and automatic optimization. Machine learning algorithms uncover hidden relationships and intricate patterns that human beings are incapable of detecting. Machine learning serves dual purposes as both a technical framework and theoretical framework which modern organizations apply to change their information processing methods for decision-making purposes Batz and colleagues (2025) address current complex data-intensive environments reveal substantial restrictions in traditional human judgment-based decision-making along with rule-based logic systems.

Machine learning (ML) predictive analytics has emerged as a vital tool for modern financial forecasts and return on investment optimization. This study shows how predictive models work by evaluating past and present data to detect patterns while assessing risks which leads to generating future insights to support strategic financial planning. Through machine learning algorithms unstructured data volumes become manageable while these algorithms discover intricate non-linear connections that traditional models cannot detect. Organizations that use this capability detect irregularities while choosing optimal investments and produce more precise future revenue projections. Machine learning-based predictive analytics improves ROI measurement and dynamic budgeting through its continuous projection refinement with new data leading to enhanced short-term operational success and long-term value development (Aro, 2024).

Predictive analytics pattern discovery enables businesses to see patterns, identify correlations and forecast future occurrences by looking at historical and real-time data. Aro (2024) explains how in financial decision-making, such ease helps businesses shift away from reaction planning to forward-strategy planning. Machine learning methodologies, especially non-linear modeling, help streamline businesses to unpack enormous volumes of structured as well as unstructured data to uncover subtle patterns that linear statistical models hardly capture. This enables business organizations to make informed investment, budget, and customer strategy choices while, at the same time, eliminating uncertainties and inefficiencies across operations.

Besides, Aro (2024) observes that predictive analytics considerably improves the precision of financial predictions, operational responsiveness, and risk establishment. For example, by the recognition of patterns of customers’ behavior or market directions, companies can efficiently distribute resources, increase credit evaluation, and even predict fraudulent dealings. These anticipatory data are especially vital in transforming and complex markets where decisions should be made on time and based on facts. Lastly, the integration of pattern recognition in planning strategy allows organizations to not only streamline short-term conduct but also align long-term objectives with developing business conditions. The development of machine learning (ML) technology has transformed predictive analytics into an essential tool which analyzes ROI together with financial forecasts. Predictive analytics evaluate current data and previous trends to help organizations make well-informed predictions about their future financial outcomes for improved operational and investment choice safety. Broby (2022) points machine learning as critical algorithms excel at detecting complex patterns which operate outside linear relationships, and they require no human guidance to modify their operations according to market changes. The ML implementation brings maximum financial returns by enhancing credit risk assessment and budget allocation and asset risk management. Businesses that use predictive models can detect upcoming business risks and chances which results in better investment outcomes because they gain market leadership opportunities.

2.2 Overview of Machine Learning for Regression-Based Predictions

Machine learning (or ML) has strong predictive analytics techniques for KPI forecasting when traditional methods are not appropriate because they do not accommodate complex, nonlinear relationships. Regression is an important supervised learning technique that forecasts continuous outcomes and is often utilized to examine the relationships between input variables and the performance indicator or metric Diamantini Diamantini, Khan, Mircoli & Potena (2024) contend that regression provides a more accurate and interpretable forecast which is important in business contexts since business decisions must be transparent and understandable. Prediction systems leverage supervised machine learning models that typically use regression types of algorithms to map historical data to projected or anticipated performance outputs.

Diamantini and colleagues (2024) also compare seven regression algorithms from the FinTech and Insurtech domains: Linear Regression, Decision Trees, Random Forests, XGBoost, Support Vector Regression, Neural Networks, and the Multi-Horizon Quantile Recurrent Forecaster (MHQRF). They find basic models such as linear regression are still indispensable for bilinearly sound evaluation and intuitiveness, even as the more comprehensive back casting methodologies like MHQRF and XGBoost yield predictive accuracy (e.g., R2 up to 0.98). They establish a superior benchmark of good or excellent performance based on common measures or various datasets; R2, RMSE, MAE, MSE, etc. In the end, regression is ultimately exemplified more as a flexible and scalable methodology to help with real-time business forecasting and decision-making assistance than as a fundamental ML method.

Regression is especially useful for forecasting business KPIs because it provides a mathematically sound way to model and forecast continuous variables, like revenue, sales, or operational metrics, based on historical patterns and elements of interest. According to Diamantini and colleagues (2024), regression algorithms can model and handle independent variables in a way that captures both linear and nonlinear relationships, including time of year, item type, and seasonality concerning a business' key performance indicators. This allows a business to understand how multiple inputs would lead to outcomes and aid in proactive decision making and resource allocation.

Additionally, regression models (e.g., Linear Regression) tend to be very interpretable, while newer and more complex techniques (e.g., XGBoost, MHQRF) balance accuracy and flexibility for operationalizing real-time, multi-horizon forecasting initiatives. Regression models can also be assessed using common error metrics (e.g. MAE, RMSE, R^2) that provide familiar benchmarks for continuing iterating, modeling, and assessing the models for business purposes (Diamantini et al., 2024).

Regression techniques have a strong reputation as powerful tools for forecasting business KPIs due to their ability to quantify the relationships between multiple covariates and continuous measures of performance. Regression models are particularly appropriate for analyzing structured time series data, as stated by

Mohammed & Mandal (2022), in which organizations must accurately forecast KPIs regarding demand, order delivery times, and inventory, in supply chain and operational contexts. As such, many practitioners value linear regression because it is simple, interpretable, and efficient; it could be the first model one uses for forecasting, or it may be favored when an organization requires interpretability for decision making processes.

The paper goes on to state that nonlinear and more complex regression models such as polynomial regression or ensemble regression techniques that allow for more flexibility in identifying non-linear trends, which are prevalent in business situations, are viable approaches for KPI forecasting. Additionally, organizations can leverage regression methods in combination with big data architecture to increase the scalability and responsiveness of their KPI forecasting process. Mohammed and co-author's (2022) state that emphasize that these advanced models are especially relevant in KPI forecasting, where business data frequently exhibit non-linear behaviors and build more data-driven decisions to improve efficiency and competitiveness.

Even though machine learning (ML) models are capable of prediction, there are several issues that can restrict their usage in real-life business situations. There are a variety of issues, with one of the main ones being overfitting, which the NCBI chapter on model evaluation talks about in detail. This occurs when the model does not perform well on previously unseen data, or in other words, the model has become very used to the quirks of the training data, including noise and outliers (Montesinos López, Montesinos López, & Crossa, 2022).

Techniques to address overfitting involve regularization, cross-validation, and simply lowering the complexity of the model. Furthermore, poor interpretability is an important limitation in industries such as health care and finance as complex techniques (such as deep neural networks or ensemble methods) tend to have lower transparency, which reduces stakeholder confidence, model interpretation, and regulatory acceptance (Montesinos et al., 2022).

Finally, ML models assume fundamental data characteristics such as independence, stationarity, and representativeness; if any of those assumptions are broken, the results could be misrepresented and misled, especially when it comes to time-series forecasting for business-oriented KPIs. Rigorous evaluation procedures, domain-specific modifications, and a balance between foreknowledge accuracy and practical application should all be utilized (Montesinos et al., 2022).

Machine learning models are powerful predictors, but they still have significant challenges to face, which, in practice, affect their reliability and transparency. Overfitting is one problem in machine learning where models learn the noise in the training data rather than generalize to new data and subsequently underperform if the new input data's distribution shifts. A second reported challenge is interpretability (Murdoch, Singh, Kumbier, Abbasi-Asl & Yu, 2019).

More complex models, such as deep neural networks or ensemble approaches, often act and function like a “black box,” taking inputs and producing predictions without relative, real-time insight into the exact processes mapping from input transformations to predictions. Murdoch and colleagues (2019) define a Predictive, Descriptive, Relevant (PDR) framework to differentiate interpretability.

Furthermore, they highlight the necessity of clearly explaining the model for debugging and clarifying stakeholders, and ethical constraints. Lastly, models frequently operate with assumptions about the data, e.g., independence of features, stationarity, and representativeness both in training and test feature distributions. When the assumptions do not fit or are constraining to real-world scenarios, the model outputs could be untrustworthy or intentionally biased. Addressing these issues must strike a careful compromise between accuracy and transparency but understanding the data situations that allow effective machine learning applications call for methodology validated accuracy, design for interpretability of results and assumptions, and the awareness of data situations to provide assurances of trustworthy machine learning (Murdoch et al., 2019).

2.3 Core Regression Methods in Literature and Practice

Regression models are important tools for predictive analytics and decision support; this is true in the case for examples such as leading Financial KPIs to enabling resource optimization or energy consumption forecasting. Linear regression is a standard model describing the linear relationship between the independent variable and its dependent variable using the least squares method Mathotaarachchi et al. (2024) highlight linear regression as computationally efficient and easy to interpret but the accuracy of the model diminishes when multicollinearity or a non-linear relationship is present. This paper explores advanced techniques that are either regularization or non-linear based when the independence assumption is violated.

Ridge regression uses an L2 penalty to shrink the coefficients and addresses overfitting and multicollinearity effects, while still retaining all the coefficients. Lasso regression uses an L1 penalty to set some coefficients to zero, to reduce features Mathotaarachchi et al. (2024) experimented with high-dimensional property price data and demonstrated in the case of high-dimensional data containing irrelevant variables, lasso regression makes the model manageable and sparse.

Alternatively, polynomial regression builds on linear regression by using higher order features to account for nonlinear interactions. Like any regression approach, polynomial regression can be overfit if regularization is not applied.

Mathotaarachchi et al. (2024) compared linear regression, polynomial regression, Lasso, and Ridge models and showed that in terms of accuracy and robustness, Lasso and Ridge models perform better than linear regression, while polynomial regression only performs better than linear regression when nonlinearity is strong and well-regularized. Accordingly, the choice of regression approach is contingent, and mediated by context, with consideration given to accuracy, complexity, and interpretability.

Beyond model performance, Mathotaarachchi et al. (2024) highlight that interpretability continues to be a relevant factor in choosing regression techniques for business decision-making. Especially in business domains like finance, operations, and resource allocation, stakeholders often require models that perform well but also yield transparent explanations for their predictions on a consistent basis. The authors argue that simpler models such as linear regression and even Lasso without regularization give a clearer image of variable influence and causality and therefore are less complicated for non-technical decision-makers. In contrast, more complex models like polynomial regression with interaction terms can conceal the effect that individual features have on outcomes unless complemented with diagnostics or visualizations.

The paper also contrasts the generalizability of regression models across different dataset sizes and compositions. The authors conducted experiments on artificial and real-world datasets to evaluate how each type of regression reacts to noise, outliers, and redundancy. The authors determined that Ridge regression exhibited stable performance under varying conditions since it is immune to overfitting and is therefore suitable for real-time predictive systems. Meanwhile, Lasso performed optimally in situations with few data or where feature removal was required for computationally economic reasons. The article concludes that model choice in business settings is not a technical choice, but a strategic compromise between predictive capacity, interpretability, operational cost, and stakeholder trust Mathotaarachchi et al. (2024).

2.4 Business-Oriented Modeling Frameworks: CRISP-DM

Brzozowska and co-authors (2023) begins the CRISP-DM methodology with a critical phase of Business Understanding. This step sets the stage and ensures that the process is driven by practical requirements and not solely analytic considerations by translating organizational goals into data mining objectives. In the case study of manufacturing, specific data mining tasks using artificial neural networks (ANNs) were outlined because of understanding the problem of predicting machine assembly time without requiring excessive technological assumptions.

In addition, this phase determines data availability, resource constraints, and success metrics for modeling. Notably, this business-driven approach reduces risk and ensures insights are kept actionable by offering a clear point of reference for all later phases data understanding, preparation, modeling, and evaluation. The last step, deployment or converting analytical ideas into actionable tools and decisions. Model construction and evaluation may produce technically accurate outputs, but deployment guarantees that these results are incorporated into business processes, such as production schedule forecasting tools or real-time personalization systems. The research suggested a predictive ANN model to assist in quotation-stage planning, improving accuracy and competitiveness. In identifying deployment as an iterative learning stage instead of a one-off endpoint, post-deployment experience returns to business knowledge, closing the CRISP-DM loop. Especially in dynamic settings such as manufacturing or finance, this circular nature helps facilitate ongoing improvement and assists in mapping data mining projects to changing business requirements Brzozowska et al. (2023).

The CRISP-DM process, or Cross-Industry Standard Process for Data Mining, remains arguably the most well-known and widely used methodology in the business analytics and data science community. It is an industry-neutral, iterative model organized as a collection of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Schröer et al. (2021) recently performed a systematic literature review to find out how CRISP-DM continues to be applied in academic studies and real-world practice after its first release two decades ago. The authors examined 24 empirical studies during 2017-2019 to find best practices, common challenges, and emerging gaps in the use of CRISP-DM across industries.

Authors explain that CRISP-DM is still the de facto standard for data mining projects since it is well organized, easy to comprehend, and technology independent. Flexibility and simplicity were the main adoption reasons, as cited by numerous researchers. Despite that, eight of the researched studies gave no explanation for the choice of CRISP-DM, and only two compared it to other methodologies before adopting it. The review also highlights that the use of CRISP-DM cuts across multiple domains, including healthcare, education, and engineering. In the healthcare sector, it was employed in tasks such as cancer diagnosis, while in manufacturing, it was applied in quality classification for lithium battery production Schröer et al. (2021). Each phase of the CRISP-DM model was analyzed individually for the selected studies. The Business Understanding phase was the most diverse, with some studies giving brief formulations of project objectives and goals and others including these elements implicitly in introduction sections. The Data Understanding phase was more uniform in its application, often involving descriptive statistics, plots, and comprehensive documentation of data sources. Some utilized several data sources, and missing attributes were collected manually to enhance completeness and readability Schröer et al. (2021).

In the Data Preparation phase, processes of transformation such as feature engineering and normalization were discussed the most. Cleaning and selection were also discussed, albeit not universally in all research. In the Modeling phase, nearly all empirical research employed and compared more than one algorithm. Algorithms decision trees, support vector machines, and random forests were used depending on the business objective and data type. The Evaluation phase typically included model performance metrics, i.e., accuracy or confusion matrices, and typically involved graphical representations and discussion of results Schröer et al. (2021).

One of the most jarring outcomes of the review was the common neglect to address the Deployment phase. While a key component of CRISP-DM, 17 of the 24 studies did not mention or describe any activities related to deployment. Of the remaining, few addressed the actual implementation of models into working systems. Others justified non-deployment using the poor performance of models or by stating deployment would be implemented on subsequent phases of the project. This reflects an overwhelming flaw in data science projects conducted at the academic level, where models are normally developed and tested in a vacuum without becoming part of real applications Schröer et al. (2021).

The authors argue that while CRISP-DM remains a valuable methodological resource, its deployment advice is inadequate for modern data-driven businesses. Particularly, the challenges of integrating predictive models within production systems with infinitely shifting architectures raise questions regarding the comprehensiveness of the framework for modern machine learning implementation. Nevertheless, the formal nature of CRISP-DM still offers a good foundation for repeatable, transparent data projects especially when complemented with project-local technologies or tools Schröer et al. (2021) Overall, Schröer et al. (2021) picture that CRISP-DM is still at the center of data mining initiatives due to its domain independence, phase-based structure, and transparency. However, the under-exploitation of its final stage deployment means that it is a critical flaw in connecting data science research and real strategic deployment. For CRISP-DM to be totally effective as a business-oriented modeling methodology, future initiatives need to be directed toward further strengthening deployment guidelines and integrating them more directly within the overall lifecycle of data analytics projects Schröer et al. (2021).

2.5 Interpretability and Human-in-the-Loop (HITL) in Regression

In high-stakes applications such as banking and politics, the uptake and credibility of machine learning models rely significantly on the balance between model complexity and interpretability. Models that are more complex, such as ensemble methods, neural networks, and deep learning structures, may achieve better prediction performance, particularly when dealing with vast and unstructured data. Nonetheless Khan et al. (2025) relegated the models as ‘black boxes’ devoid of the transparency required for stakeholder trust, ethical responsibility, and regulatory compliance. The lack of clarity prevents their application whenever decisions should be backed by regulations that are conveyed to non-technical stakeholders. Although simple models such as linear regression can be less predictive in certain situations, they are often preferred due to their intrinsic interpretability and simplicity of verification.

Post hoc explainability methods, including SHAP, LIME, and counterfactual explanations, have become a viable way of resolving this trade-off. These tools seek to distill interpretable reasoning from intricate models without undermining their predictive performance. However, these tools add another level of abstraction and possible bias, and the quality of their outputs can be model and dataset dependent A measured approach is thus merited: simpler models will be adequate in applications where interpretability is foremost (e.g., loan approval or risk assessment), while black-box models complemented by explainability modules can be justified when prediction accuracy takes precedence. Ultimately, the authors posit that the pressure for transparency from decision-makers, the current regulatory environment, and the risk tolerance of the domain must all collectively guide the choice between interpretability and complexity Khan et al. (2025).

Human-in-the-loop (HITL) machine learning has gained increasing attention in recent years due to its potential in bridging the gap between algorithmic and expert-level reasoning, especially in mission-critical tasks such as regression tasks. According to Mosqueira-Rey et al. (2022), the inclusion of human feedback into the machine learning pipeline enhances the interpretability, transparency, and contextual fit of prediction models. This is particularly significant in regression issues where high-quality numerical predictions often drive high-risk business or operational decisions.

Interpretability, as the very central pillar of HITL systems, is paramount in regression as stakeholders must understand how and why a prediction has been made. The authors classify interpretability in two broad dimensions: global interpretability, which deals with understanding the overall structure and logic of the model, and local interpretability, which deals with understanding individual predictions. In regression contexts, especially in the case of linear or regularized regression (i.e., Lasso or Ridge), mathematical structure naturally provides some interpretability. HITL workflows further augment this by allowing users to validate model assumptions, remove unimportant features, and review the rationale behind model choices, as seen in Mosqueira-Rey et al. (2022).

The study further points out the importance of visual and interactive interfaces for HITL regression pipelines. Visual explanation methods like SHAP values or feature influence plots enable users to see the contribution of each variable to a specific output. This is particularly valuable in regression activities like financial forecasting, where decision-makers ought to know how features such as revenue, total assets, or

R&D expenses affect outputs like net profit. The authors point out that these interfaces enable real-time human involvement, hence propelling the modeling process from being a solely computational exercise to an interactive process Mosqueira-Rey et al. (2022).

Mosqueira-Rey et al. (2022) explains that interpretability in HITL systems is both a technical and strategic goal. Explainable models foster trust among business users, auditors, and customers, and are more likely to be deployed in regulated or sensitive setups. In regression-based applications, such interpretability allows not just facile understanding but also error detection, model validation, and ongoing improvement especially when used in dynamic systems prone to data drift. Finally, the article discusses how HITL systems can support iterative learning, whereby human corrections drive model revisions through cycles. This continuous human feedback process augments predictive validity and ethical appropriateness over time. In complex regression problems such as those involving behavioral, financial, or environmental prediction the human analyst is not only a supervisor but also a co-designer of the model logic, making it statistically sound and contextually meaningful simultaneously.

2.6 Model Evaluation in Business Contexts

Machine learning (ML) model assessment in the business domain demands a multi-faceted evaluation process beyond traditional performance metrics like accuracy, R^2, MAE, or RMSE. Models in business are not only statistically tested but also assessed for their strategic importance, financial viability, usability, and long-term operational impacts. Mizgajski et al. (2021) cite such complexity in their paper, which explores the misalignment between machine learning research conducted in academe and actual business needs particularly when it comes to demonstrating Return on Investment (ROI).

The authors present One of the strongest arguments that academia focuses primarily on optimizing predictive metrics, companies are more interested in tangible outcomes such as increased efficiency, cost savings, or enhanced customer satisfaction. A less predictive model in business decision-making situations may be preferable if it is easier to interpret, can be implemented more quickly, or can be easier to integrate into existing workflows. This points to the need for evaluation models that incorporate economic metrics such as ROI, payback, and implementation cost—considerations normally overlooked in academic requirements Mizgajski et al. (2021)

The paper proposes a multi-layered evaluation paradigm where models are tested on four interconnected layers: data readiness, model correctness, deployment feasibility, and return cost. The layered evaluation encompasses the entire ML lifecycle, ranging from preprocessing and feature extraction to operational impact and profitability. For example, a customer churn model can be very accurate but, in a situation where it costs more to retain a customer than the revenue it brings in, the model's worth in the real world is negative. Therefore, the use of cost-sensitive measures becomes critical while evaluating performance from a business point of view Mizgajski et al. (2021)

Another essential learning point from Mizgajski et al. (2021) is the importance of human-centered evaluation. Success in most business implementations of machine learning models hinges on how readily they can be consumed by business users, including analysts, managers, and executives. This requires a compromise between performance and interpretability, especially for regression models deployed in financial forecasting, where the stakeholders must understand contribution of inputs like revenue or assets to projected outcomes like Net Income. Utilities like SHAP and LIME enable this interpretability, but the authors are interested in hinting that such utilities must be embedded within the organizational feedback loops to create model trust and actionability. In business applications, where stakeholders require unambiguous explanations of artificial intelligence decisions to build trust and support decision-making, model interpretability is essential. **(murdoch2019?)** definitions offer the Predictive, Descriptive, Relevant (PDR) framework as one way of evaluating interpretability. This framework requires that models not only perform at high levels of predictive accuracy but also deliver descriptive accuracy, which involves transparency into how input features affect outputs, and relevance to human users, meaning that explanations should be framed in terms of stakeholder needs and domain-specific knowledge. This would imply, therefore, that characteristics used within models should have consistent and obvious interpretations, like “payment history” or “customer tenure,” so that business leaders can discern how these characteristics affect outcomes such as credit risk or customer churn. Without this alignment, even accurate models may be viewed suspiciously or misused, especially in industries that are regulated, where transparency and the ability for auditing are essential **(murdoch2019?)** definitions.

In short, business model assessment requires a holistic framework that not only focuses on statistical precision but also on strategic fit, financial performance, interpretability, and operability. Mizgajski et al. (2021) state the findings support the idea that ML solutions should be seen as strategic investments rather than being evaluated solely based on technical criteria. According to this perspective, the usefulness of machine learning is in its ability to provide quantifiable business results, pointing decision-makers in the direction of models that are not only technically sound but also in line with organizational objectives and actual application scenarios.

3. Methodology

The methodological technique employed to construct, verify, and interpret regression models on real financial data is carried out on. Built on jupyter notebook, google colab is a cloud-based Python Integrated Development Environment (IDE). It provides free access to computer resources, such as GPUs and TPUs, and doesn’t require any setup. Colab works especially well for data science, machine learning, and education (Google, n.d.). Access my GitHub repository Regression_analysis to reproduce, contribute or evaluate the exact results as directed in this thesis. Repository contains ‘Tanmay_Ubhate_400338253_Master_Thesis.pdf’ copy of this thesis, ‘regression_analysis.ipynb’ python jupyter notebook containing all regression pipeline syntax, financial data sp500 companies.csv, README.md stating overview of the project and guidance to replicate the results. The basic objective of this research is to model and predict a firm's net income using internal financial measures obtained from quarterly financial statements, income statement, balance sheet, and cash flow statement variables to provide data driven decision.

The methodology integrates key steps data selection, source validation, pre-processing, Exploratory data analysis (EDA), feature engineering and model construction through multiple regression techniques.

The research uses a quantitative approach where supervised learning regression models are tested on a well-structured financial data set. This approach aligns with the overall objective of the study assessing the balance between model accuracy and interpretability and development of a decision support tool suitable for business users with modest technical ability. The models are validated and trained using a rigorous evaluation protocol focused on predictive performance and model explainability.

This methodology section is authored to meet standards of academic integrity and replicability and sound empirical financial analysis standards. Whenever possible, open-source software, transparent processes, and ethical data practices are used to ensure that the findings are not only reliable but also practical.

3.1 Importing libraries

This study leverages a set of core Python libraries that represent best practices in modern data science and machine learning workflows. These libraries were selected not only for their robust functionality and integration but also for their extensive use in academic and industrial research, as validated in Raschka et al. (2020).

3.2 NumPy and Pandas

The backbone of numerical and tabular data handling in this project was provided by NumPy and pandas. NumPy offers highly efficient operations on large arrays and matrices, serving as the computational foundation for numerous higher-level libraries Harris et al. (2020). Addepalli et al. (2023) present in their article pandas as a widely used open-source python library for data manipulation and analysis. It facilitates effective management of organized data by means of filter, summary, and data transformation tools. Pandas offers basic visualizations like line and scatter plots for data exploration. Its comprehensive time series capabilities involve time-indexing, resampling, and rolling computation. The library is required for cleaning the data, including dealing with missing data and duplicates. Pandas works nicely with other libraries like numpy, matplotlib, and scikit-learn. Pandas is used extensively across industries like finance, healthcare, and social sciences.

3.3 Matplotlib and Seaborn

For data visualization, matplotlib and seaborn were used to explore data distributions, relationships, and model diagnostics. Matplotlib is a foundational plotting library, while Seaborn extends its functionality by providing preconfigured statistical plots Waskom (2021). Visualization is a crucial aspect of exploratory data analysis (EDA), and these tools are fundamental for interpreting complex financial data patterns.

3.4 Scikit-learn

All core modeling and evaluation tasks in this thesis were carried out using scikit-learn. As the paper details, scikit-learn provides an accessible and unified interface for regression algorithms (e.g., Linear, Lasso, and Ridge), data preprocessing (StandardScaler, PolynomialFeatures), model selection (train_test_split), and evaluation metrics (mean_absolute_error, mean_squared_error, r2_score). The ability to build pipelines using make_pipeline ensures that preprocessing and modeling steps are reproducible and streamlined.

3.5 SHAP (SHapley Additive Explanations)

To improve model transparency and interpretability, this thesis used SHAP, a library implementing game-theoretic approaches for explaining machine learning predictions. Although not explicitly discussed in Raschka et al. (2020) , SHAP complements the broader movement toward explainable AI highlighted in the paper, especially for black-box models. In this study, SHAP was applied to evaluate feature contributions and enable decision traceability in regression analysis. SHAP makes it possible to visualise and quantify both localised instance-based reasoning and global model behaviour by assigning individual predictions to input variables. This interpretability was also helpful in ensuring that the internal logic of the model was in line with knowledge of the financial domain, maintaining the transparency, credibility, and actionability of predictive insights for decision-makers.

3.6 Data Sources

Dataset financial data sp500 companies.csv is taken from open source Kaggle by a genuine author with credible background. As stated in source of dataset financial information was collected by web scraping from Yahoo Finance aggregator of U.S. Securities and Exchange Commission (SEC) filings. According to Boritz & No (2019) this public data platforms often add extra calculated metrics, derived fields, or transformed data that you won't find in raw official filings like 10-K filings, annual report submitted to SEC still very useful due to simplicity of variables extracted for investors, analyst, ML practitioners and Application developers.

The data file was in CSV format after importing data in google collab IDE it was read using pandas .read_csv() and upon loading, fundamental structural information was examined with numpy .info() to detect for parsing errors or their absence and check data types

3.7 Dataset Structure

The dataset has 2,012 observations across the two years as quarterly data with 4 different dates:2020-12-31, 2021-03-31, 2021-06-30, 2021-09-30. Each observation is a specific firm-quarter pair with the following 15 numeric variables (with data type integers for whole numbers and float for to capture decimal):

1. Total Revenue
2. Operating Income
3. Research & Development (R&D) Expenditure
4. Selling, General, and Administrative (SG&A) Expenses
5. Interest Expense

6. Income Before Tax
7. Income Tax Expense
8. Gross Profit
9. Earnings Before Interest and Taxes (EBIT)
10. Total Operating Expenses
11. Cost of Revenue
12. Total Other Income (Net)
13. Net Income from Continuing Operations
14. Net Income Applicable to Common Shares
15. Net Income (target variable)

These variables were in the selected data set, due to their straightforward application to the research question and their frequent occurrence in financial modeling literature. Standardized financial categories allow consistency across firms and quarters.

3.8 Temporal and Sectoral Granularity

Every company's data is picked up for four quarterly periods: March 31, June 30, September 30, and December 31. This degree of granularity enables one to identify trends, seasonality breakdown, and between-year performance comparison. While the dataset has companies from various sectors (e.g., technology, health care, consumer goods), the modeling activity is sector-agnostic unless specifically mentioned. The aim is to develop generalized regression models that can pick up broad patterns of profitability rather than sectoral nuances.

3.9 Target Variable Justification

This study models net income as the dependent variable, according to Holthausen & Watts (2001) recognizing its established relevance in financial reporting and market valuation. Net income reliably reflects a firm's ability to generate future cash flows and is strongly associated with investor decision-making and firm valuation. Its use ensures alignment with standard accounting practice and enhances the model's interpretability. Moreover, net income serves as a practical and measurable output that integrates operational, financial, and strategic performance. Employing solely internal financial metrics in modeling Net Income, this thesis avoids market-based variables confounding effects on share price, volatility, or trading volume. The research design to accomplish the same purpose of arriving at findings relevant to internal stakeholders such as CFOs, analysts, and operations managers.

3.10 Data Cleaning and Imputation

The first part of preprocessing involved two steps: the removal of structurally redundant columns (an unnamed index column) and the removal of duplicate records to prevent statistical bias. Missing values were identified by '.isnull( ).sum()' fuction from variables, research development, interest expense, SG&A, Net income from continuing operations and net income applicable to common shares contained null values. To offset these, median imputation was applied using the '.fillna()' function. Median rather than mean values were selected to mitigate the influence of skewed distributions and outliers, frequent characteristic of financial KPIs due to irregular earnings shocks and extreme transactions. This method preserves the dataset's central tendency while improving robustness and reducing distortion Little & Rubin (2019). Median imputation provides a more robust method for maintaining data integrity in the context of quarterly financial reporting, when variables such as R&D expenditure or one-time charges can vary greatly between organisations and time periods. Furthermore, it guarantees that during model training and assessment, the fundamental statistical presumptions of regression models particularly those that are sensitive to scale and variance remain true.

3.11 Exploratory Data Analysis (EDA)

Key metrics were derived through the '.describe()' function of pandas library to summarize key variables in the dataset. The results revealed a high degree of variability, particularly in the net income column, which serves as the dependent variable. Notably, total revenue and operating income also exhibited wide value ranges, spanning from substantial losses to significant profits. The standard deviation for these variables was in the billions, indicating that the dataset contains extreme outliers and large fluctuations, a common trait in financial data, especially across companies of different sizes.

To obtain an understanding of distributional characteristics and detect anomalies, histograms and boxplots were created using matplotlib and seaborn libraries. Histograms (sns.histplot) revealed right-skewed distributions for the features net income and total revenue, indicating that most companies fall in a mid-performing band, whereas a few of them have very high values. Boxplots confirmed the skewness and indicated outliers, especially for high capital companies. These visual aids gave preliminary insight into the nature of data and set the stage for smart removal of outliers.

3.12 Outlier Detection and Treatment

A revised Interquartile Range (IQR) approach was used to handle outliers. To aggressively eliminate mild and moderate outliers, the normal IQR boundaries (Q1–1.5IQR to Q3+1.5IQR) were tightened to Q1 = 42nd percentile and Q3 = 65th percentile. The goal of this modification is to identify small outliers that are known to affect model learning but would not be seen in conventional boxplots. By strengthening the boundaries, the model was protected against anomalous financial entries that disproportionately affect regression coefficients, remarkable income, valuation reductions & corporate restructuring.

According to Aggarwal (2016), financial data frequently show sparse high-magnitude anomalies and lengthy tails, which calls for outlier detection techniques that are specific to the distributional characteristics of the domain. The revised IQR filter was subsequently used on net income, total revenue, and operating income. When removed, distributions were rendered more even, reducing the excessive influence of outliers and improving model generalizability. Boxplots re-plotted following removal showed fewer outliers and a more even interquartile range, affirming the effectiveness of this preprocessing step.

3.13 Correlation Analysis and Feature Selection

A Pearson correlation matrix was generated to examine relationships between numeric variables, following the guidance of Chan et al. (2022), emphasizing its value as a vital tool for identifying patterns in data. The relationship of each feature with net income was presented graphically by a heatmap and ranked bar plot. Net income from continuing operations, income before tax and net income applicable to common shares were highly correlated (> 0.9) within themselves.

These variables were not in the model for redundancy risks and information leakage. Including them would have been manipulating the accuracy of the model because they are either mathematically or logically deducible from the target variable Montesinos López et al. (2022). Similarly, Interest Expense, despite being moderately correlated, was excluded due to negative correlation. This pruning process of correlation resulted in a feature matrix (X) with independent, non-redundant predictors, resulting in enhanced interpretability as well as predictive validity.

3.14 Feature Scaling and Splitting Dataset

As there were varying scales between the financial figures Total Revenue being in hundreds of billions and SG&A or R&D being in millions standardization was applied to avoid model bias on large-scale features. ‘Standardscaler()’ function from scikit libraries was utilized to normalize all the numerical features with z-score normalization, rescaling each feature to a mean of zero and standard deviation of one.

After scaling, data was split into training and test sets with a proportion of 70/ 30 using ‘train_test_split()’ function with ‘random_state=42’. The random state acts as a seed for the random number generator, which influences how data is split or how models initialize internal parameters, to maintain consistency and enable reproducibility of results, for functions that involve randomness, such as train test split. () Fitting the model was carried out with the training set (X_train, y_train), and the test set (X_test, y_test) was left for final evaluation.

3.15 Model Development and Evaluation Strategy

This thesis employs comparative modeling in the comparison of performance of four supervised regression techniques in the estimation of Net Income of companies. The models employed are Linear Regression, Ridge Regression, Lasso Regression, and Polynomial Regression, which are widely applied in financial analysis and forecasting modeling. These models were selected based on their respective advantages: Ridge and Lasso's ability to handle multicollinearity and carry out regularization; Polynomial Regression's ability to capture possible nonlinear relationships among financial variables; and Linear Regression's ease of use and interpretability Hastie et al. (2009).

The study attempts to uncover trade-offs between accuracy, generalizability, and transparency three important factors when using machine learning tools for finance by assessing different models on the same dataset using consistent metrics. The Linear Regression model is utilized in this research as the base model. It provides an open-ended structure where the connection among the dependent and independent variables is specified via a linear equation. The largest benefit to this method is interpretability and ease but it depends on the homoscedasticity, normality, and no multicollinearity assumptions that do not necessarily exist with real finance data Hastie et al. (2009). It is a fundamental baseline for assessing the incremental value of more complex models in spite of its limitations in complex or noisy data sets. Linear regression remains applied in financial practice because of its interpretability, particularly in audit and regulatory settings where model explainability matters.

Ridge Regression is a variation of linear regression in which an L2 penalty term is added to reduce the regression coefficients to deal with multicollinearity between predictors Hoerl & Kennard (1970). When predictors are highly correlated, which is a common feature in financial datasets, this regularisation strategy stabilises the estimate process by shrinking the coefficients of less useful characteristics. Ridge is especially useful when all variables have some predictive value since it divides weight more evenly among features rather than eliminating variables like Lasso exhibits Hastie et al. (2009). To determine whether this regularisation enhances prediction performance over the conventional linear model while maintaining interpretability and minimising overfitting, this study employs ridge regression.

Lasso Regression, another linear regression with regularization, employs an L1 penalty that not only shrinks coefficients but also sets some of them to zero and hence performs implicit feature selection Chan et al. (2022). This characteristic makes Lasso extremely useful for the scenario of high-dimensional data or when some predictors are of questionable importance. Polynomial Regression, on the other hand, allows the estimation of non-linear effects through interaction terms as well as higher-order terms. Although more flexible, the model can overfit if not properly regularized or tested on out-of-sample data Tibshirani (1996).

All these models were fit to the training subset of the data and tested on the holdout test set in a work to estimate generalization performance. Fitting was done with the ‘fit()’ function of the scikit-learn package and prediction was done with the ‘predict()’ function. The models performance was measured on three critical evaluation measures: the R^2, used to estimate the variance explained by the model; RMSE which punishes large errors more severely than small ones; and MAE, which is robust to outliers and provides currency interpretability. The outcomes of every model were merged into a tabular aggregate to enable facile comparison and interpretation.

3.16 Methods for Model Interpretability and Explanation

In addition to quantitative performance, interpretability is also a key component of model evaluation in this study. Interpretability is native to Linear, Ridge, and Lasso models. All these models output a set of coefficients that can be interpreted as the change in the target variable due to a one-unit change in the predictor with other variables held constant as Hoerl & Kennard (1970) illustrating the implications of nonorthogonality in two dimensions. These coefficients have immediate business use, where stakeholders get to understand what financial metrics contribute most to driving net income. Interpretability of polynomial regression is limited in nature by the inclusion of interaction terms and higher-order terms.

The resulting model becomes difficult to interpret without advanced statistical knowledge. To address this limitation, the study employed SHapley Additive exPlanations (SHAP), a post-hoc interpretability approach that analyzes model predictions to their additive contributions of each feature as well Lipton (2018) identified need of explanation of ML methods. For each model, SHAP summary plots and coefficient weight plots were generated to visually communicate feature importance. Such plots facilitate easy model behavior interpretation and ensure that business users can derive actionable insights. The use of SHAP thus bridges the gap between managerial interpretability and statistical complexity, which is especially necessary in high-stake decision-making environments where black-box models are often avoided in favor of more interpretable alternatives Lipton (2018)

3.17 Ethical Questions, Transparency and Replicability

The application of machine learning models to finance data needs close attention to ethical implications, data privacy, and reproducibility. As the kaggle dataset used in the research consists of publicly available data from Yahoo Finance, bias or implicit inference risk should still be acknowledged boritz2020significant. For instance, the dominance of large-cap firms within the S&P 500 can skew the model towards trends that are less applicable to mid-cap and small-cap firms. Feature scaling and outlier deletion techniques were applied consistently to normalize the influence of firm size by preventing large profitable businesses from dominating the regression coefficients, these pre-processing measures improved the model’s ability to generalise across a range of businesses, **(JMLR:v22:20-3032021reproducibility?)** addressing this as need of high reproducibility. Feature scaling enabled predictors that are scored on extremely different scales like revenue and R&D expenditure to contribute more fairly to the model’s learning process.

Transparency of the model was prioritized throughout the analysis. Each step in data processing was reproducible and well documented, ranging from missing value imputation to standardization as well as feature transformation. Splitting data was done with a fixed random seed for the purpose of ensuring reproducible results in experimental runs. Furthermore, the whole pipeline of modeling was written using open-source libraries in Python so that everything was completely reproducible for other practitioners and researchers.

(pedregosa2011?) scikit proposing scikit-learn allowed for model portability across contexts and adherence to industry best practices in machine learning research. This transparent and open-source approach not only increases academic rigor but also facilitates comparison, validation, and extension of the results in future studies.

To promote academic integrity and address the principles of responsible AI, the research also held out variables that could cause information leakage. These are mathematically derivable target variable fields such as Net Income from Continuing Operations and Interest Expense that also exhibited multicollinearity with several operational attributes. By holding them out, model performance reflects predictive capability rather than spurious overfitting to non-relevant or derivative features.

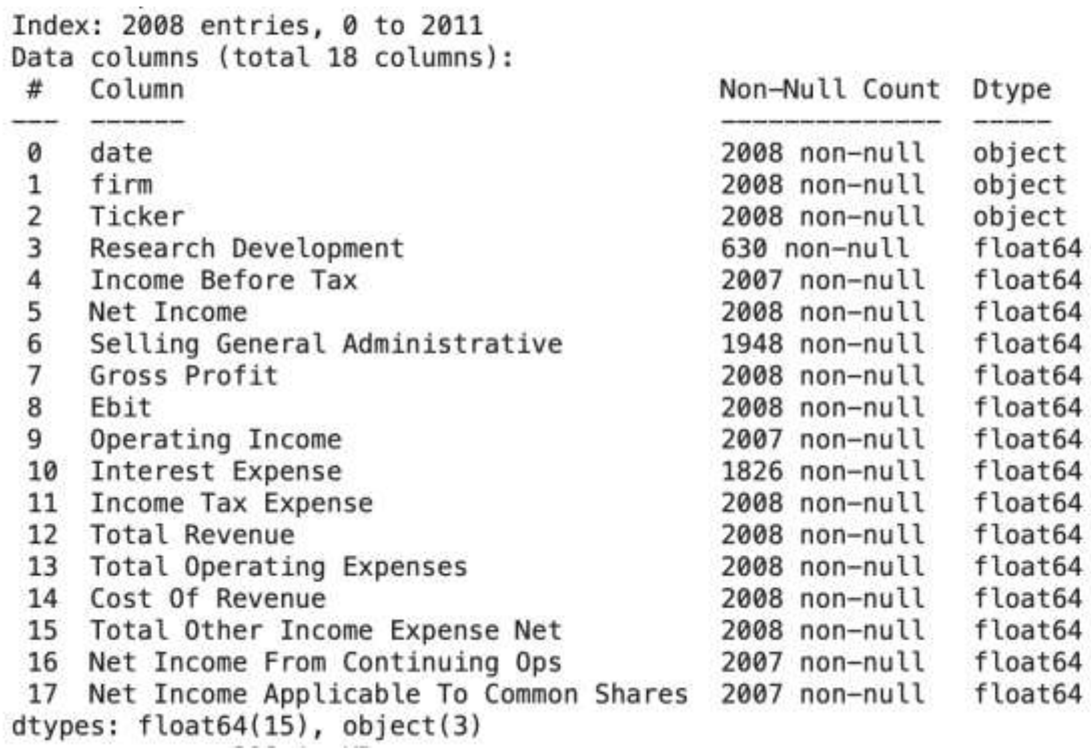
In conclusion, this research’s methodological design incorporates best practice in both machine learning and financial modeling. Through the incorporation of a broad spectrum of regression models, rigorously applied preprocessing guidelines, and interpretability tools such as SHAP, the thesis includes an integrated framework for uncovering the drivers of corporate profitability. Ethical standards of transparency, fairness, and data integrity have been embedded in the analytical pipeline throughout to ensure that not just results are statistically sound but also contextually appropriate.

4. Results and Analysis

This part presents findings from a quarterly financial KPI regression analysis of companies listed in the S&P 500 stock exchange index. The aim of this study was to identify the most influential financial variables that can account for net income and the resultant performance of different regression models to perform the task. Emphasis was put on the relationship between internal financial build-up and profitability, without regard to external market measures like share price. Distributional properties, correlation patterns, regression efficiency, and SHAP values based post-model explanation are addressed in the following sections.

4.1 Distribution Analysis and Preprocessing Results

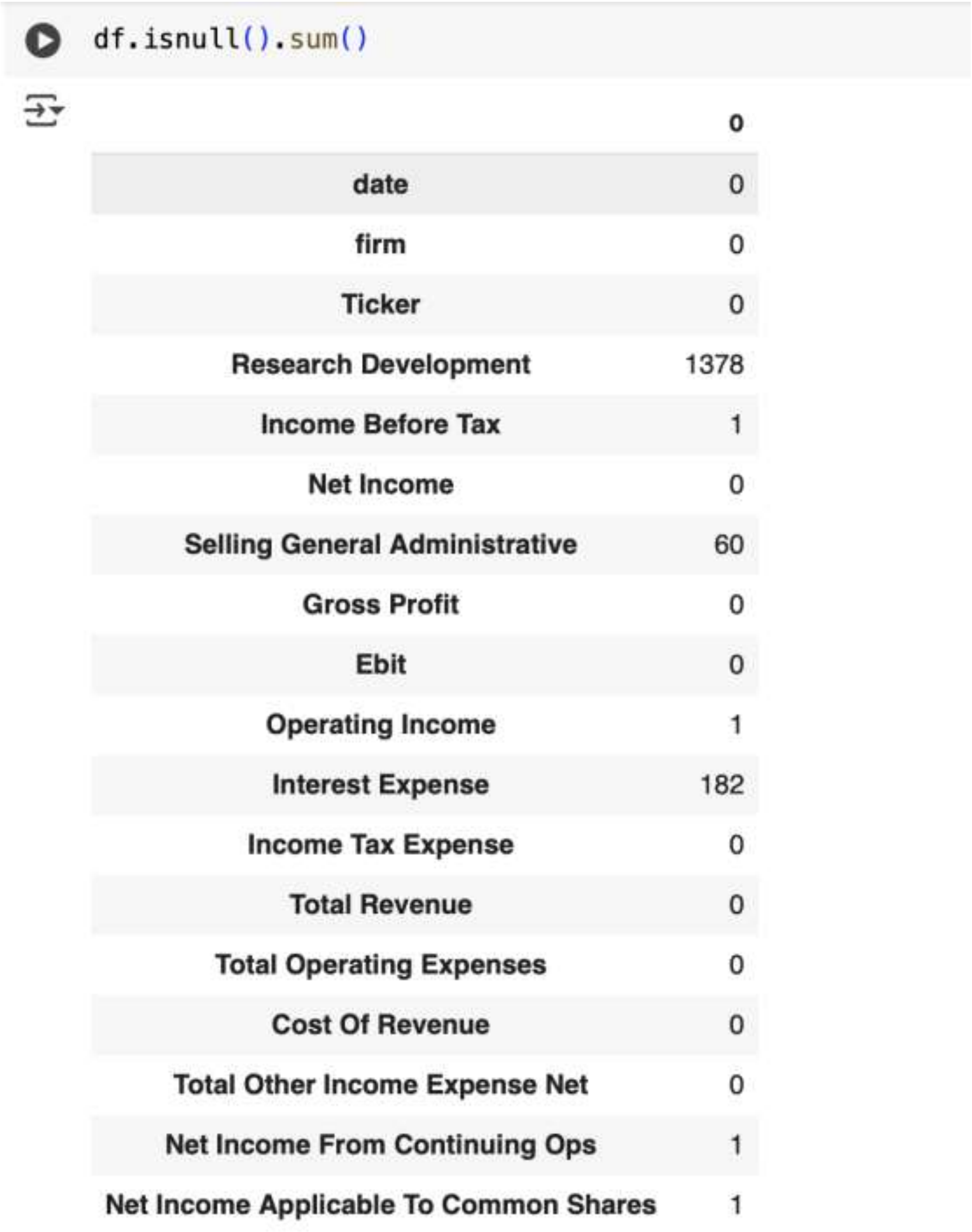
Figure 1. mentioned output of '.info()' function from python pandas library, confirmed the majority of variables in the dataset are of 'float64' data type that stores form of numeric value (especially decimal values). This is best suited for numerical analysis particularly in regression-based supervised learning ML algorithm. They enable the model to process and measure continuous variables that are critical for financial prediction operations. The 'object' columns firm and ticker, which store textual identifiers. These types of categorical variables cannot directly inform regression models without transformation, as they do not have inherent numerical meaning. Raschka et al. (2020) emphasizes the importance of preprocessing these kinds of data to enable compatibility with machine learning algorithms. The inability of regression models to process raw text underscores the importance of exclusion or encoding of these kinds of columns where the goal is to extract generalized, scalable insights from financial variables.



```
Index: 2008 entries, 0 to 2011
Data columns (total 18 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   date                                2008 non-null   object
 1   firm                                2008 non-null   object
 2   Ticker                              2008 non-null   object
 3   Research Development                630 non-null    float64
 4   Income Before Tax                   2007 non-null   float64
 5   Net Income                          2008 non-null   float64
 6   Selling General Administrative      1948 non-null   float64
 7   Gross Profit                        2008 non-null   float64
 8   Ebit                                2008 non-null   float64
 9   Operating Income                    2007 non-null   float64
 10  Interest Expense                    1826 non-null   float64
 11  Income Tax Expense                  2008 non-null   float64
 12  Total Revenue                       2008 non-null   float64
 13  Total Operating Expenses            2008 non-null   float64
 14  Cost Of Revenue                     2008 non-null   float64
 15  Total Other Income Expense Net      2008 non-null   float64
 16  Net Income From Continuing Ops      2007 non-null   float64
 17  Net Income Applicable To Common Shares  2007 non-null   float64
dtypes: float64(15), object(3)
```

Figure 1: Figure 1. Pandas '.info()' function output

Notes: Date, firm & ticker are object data types, and all other variables are float64 and best suitable for regression methods



```
df.isnull().sum()
```

	0
date	0
firm	0
Ticker	0
Research Development	1378
Income Before Tax	1
Net Income	0
Selling General Administrative	60
Gross Profit	0
Ebit	0
Operating Income	1
Interest Expense	182
Income Tax Expense	0
Total Revenue	0
Total Operating Expenses	0
Cost Of Revenue	0
Total Other Income Expense Net	0
Net Income From Continuing Ops	1
Net Income Applicable To Common Shares	1

Figure 2: Figure 2. Count of null/ missing values

Notes: Median imputation was statistically appropriate to preserve uniform model input dimensions across the dataset remain interpretable for firms that report them.

Figure 2. an output of sum of null values magnifies the columns missing or null values. Boritz & No (2019) indicate that inconsistencies in financial data, particularly for line items such as R&D, are common due to different treatments in the accounts, firm sizes, and the aggregation of data from sources. Because over 60% of its values were missing, the Research Development column stood out. It is mostly relevant to R&D-intensive industries like technology and pharmaceuticals, but many other businesses either do not invest in R&D or do not disclose it clearly, which accounts for its missingness.

Recognising that its presence or absence may carry sector-specific meaning, the variable was kept in place despite the sparsity to maintain potential domain signal. To prevent adding systemic bias or exaggerating its impact in models, careful median imputation was used, and interpretive prudence was maintained Little & Rubin (2019). Keeping such a variable in place is consistent with best practices for retaining analytical validity in diverse financial datasets while conserving structurally significant features. While it performs well in maintaining central tendency, it can underrepresent extreme but important patterns such as innovation expenditure bursts or R&D cutbacks that are extreme.

Exploratory data analysis (EDA) examined the underlying distributions of key financial KPIs: histogram for net income, total revenue, and operating income were highly positively skewed as observed in Figure 3. In practical terms, it implied that most companies tended to have minimal to low earnings, and a smaller subset of large-capitalization companies had obscenely high revenues and profits. Such skewness is not rare but rather usual for financial data in capital markets broby2022use, where there are asymmetrical distributions of firm size and revenue potential.

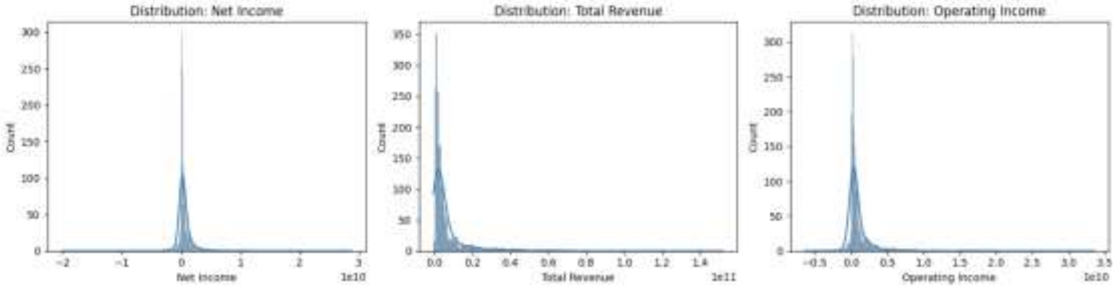


Figure 3: Figure 3. Histogram for outlier check

Notes: Higher affecting variables to labelled Net income in financial prospective might have highest outliers.

In the histograms presented in Figure 3, the vertical axis labeled as count refers to the number of firms whose reported values fall within each bin of the respective financial variable’s range. It enables visualization of the distribution density for variables, The high peaks near zero and long right tails highlight the positive skewness characteristic of financial data. Skewness statistically in this context refers to having a long right tail in the histogram, with the minority of firms significantly outdoing average performance. Most of the companies in the sample had net incomes between $1 billion and $10 billion, some giants recorded more than $100 billion as quarterly profit. The consequence is a big standard deviation and skewed mean value relative to the median. The impact on regression modeling is that outliers can skew coefficient estimation, leading to overfitting or reduced generalization to average-performance companies. As emphasized by Montesinos López et al. (2022), skewed distributions and the presence of extreme outliers in input variables can significantly affect the predictive stability of regression models.

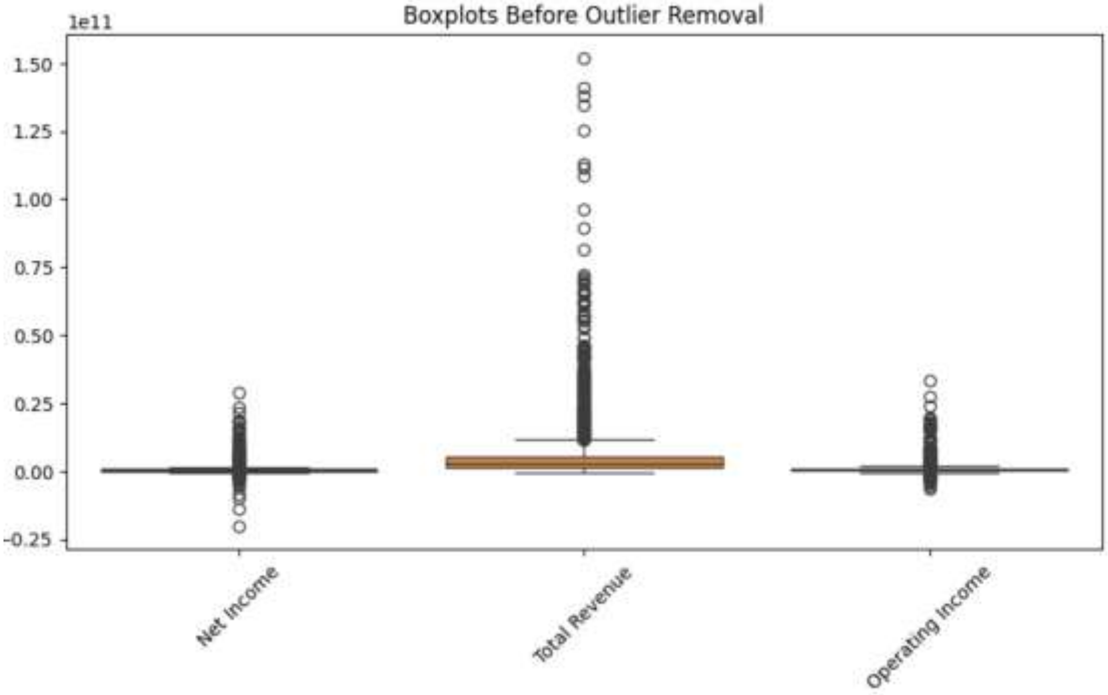


Figure 4: Figure 4. Box-plot for outlier check

Notes: The values in net income should be limit between the standard deviation for stable modelling outputs using interquartile limits.

To mitigate against this, an IQR-based outliers removal method highlighted by Aggarwal (2016) was used. Unlike the standard procedure (using 25th and 75th percentiles), this study made use of modified values at 42nd and 65th percentiles, limiting accepted interquartile range to observe both mild and outlier instances. This strict filtering was warranted considering the desire to have the regression models learn from central, representative monetary behavior rather than extreme corporate outliers.

The effect of this elimination of outliers could be compared in the post-filtered boxplots shown in Figure 5. The graphs depicted that the outliers had indeed been successfully removed from the data without significantly compromising the variety of observations. For example, volatility in operating income and total revenue decreased substantially, while net income retained its moderately skewed shape due to the natural volatility in profitability.

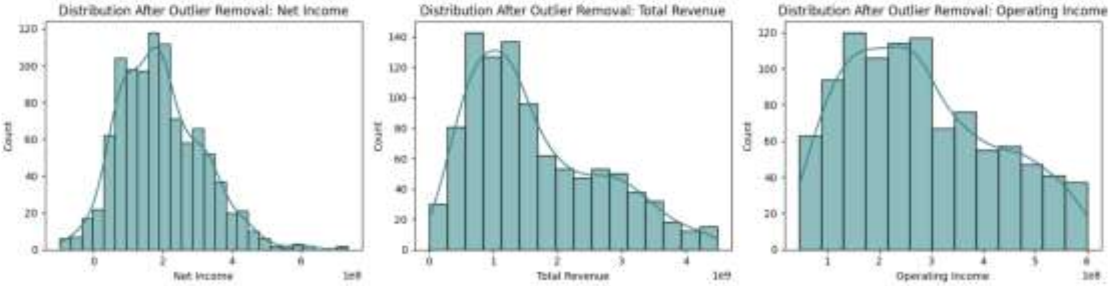


Figure 5: Figure 5. Distribution plots after outlier removal

Notes: Even with high degree of outliers removal, some values enables to permute extreme values to maintain fitting.

Notably, this preprocessing decision is a balancing act between two significant factors. On the one hand, the exclusion of too many outliers reduces sample richness and denies model the ability to generalize exceptional but possible business cases.

On the other hand, failure to exclude heavy skew induces skewness into the regression line in the direction of dominant but unrepresentative firms and leads to unsound predictions for the entire dataset. The method used here was consistent with best practice in financial data modeling as Hyndman & Koehler (2006) note, adhering to such best practices in time series and financial data modeling helps ensure that predictive models remain robust across varying data distributions and yielded a more stable and generalizable regression framework.

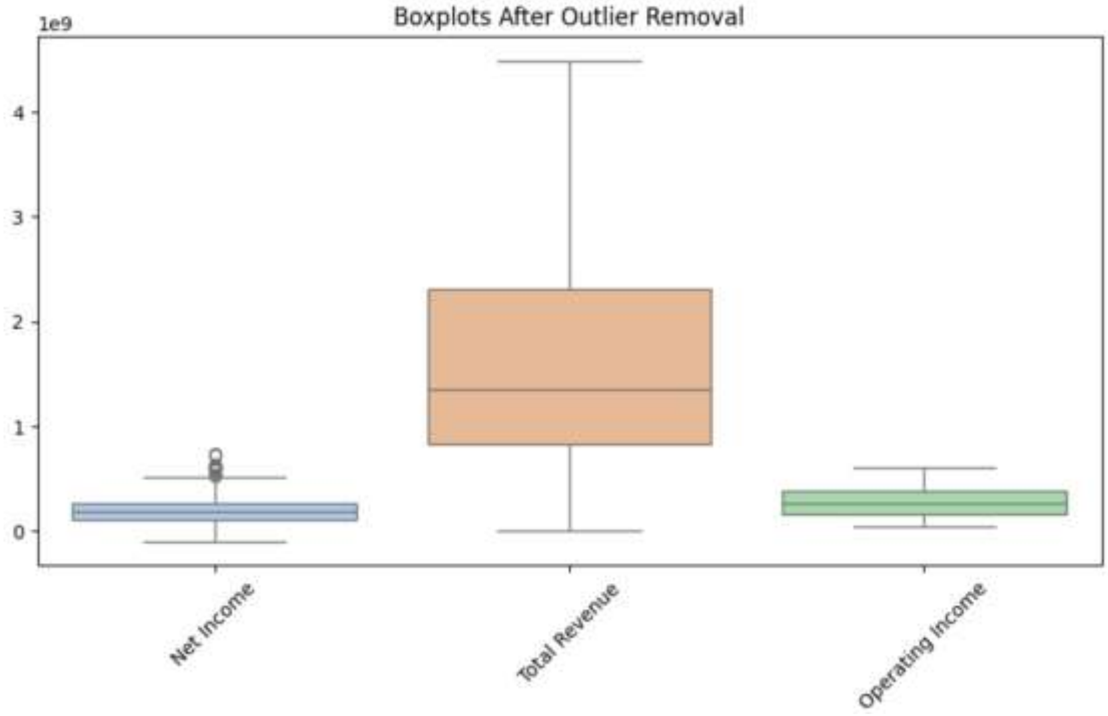


Figure 6: Figure 6. Box plot after outlier removal

Notes: Features should good fitted as all values could be impute within quartile limit.

Through these preprocessing steps, it was confirmed via the box plot in Figure 6 that outliers were successfully compressed within the interquartile range (IQR). The training dataset achieved a higher level of statistical homogeneity, reduced heteroscedasticity, and improved compatibility with linear regression assumptions. This provided a robust foundation for the subsequent stage of modelling and enhanced both interpretability and predictive accuracy.

4.2 Feature Relationships and Correlation

Correlation analysis was conducted to examine the direction and strength of linear associations between variables. Figure 7's heatmap revealed income before tax, operating income and gross profit to have the maximum acceptable positive correlations with net income, with coefficients approaching +0.90. This validates prior studies of conventional financial principle that earnings generated before tax obligations is a very good predictor of bottom-line profitability Broby (2022)).

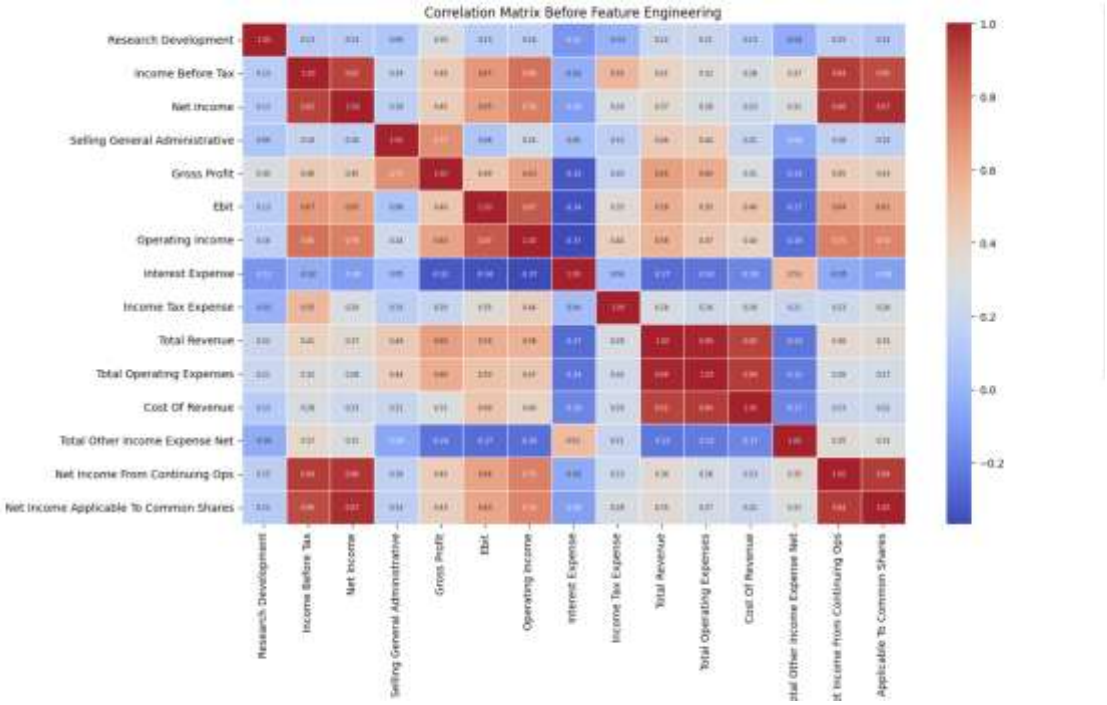


Figure 7: Figure 7. Correlation matrix

Notes: Variables have high to low varying correlation address to check labelled data specific features.

A bar plot of the top 15 highest correlated variables (Figure 8) gave further insight., SG&A (Selling, General and Administrative) expenses and cost of revenue were least correlated with net income. These variables represent core operational expenses, and their least correlation validates that higher spending in these areas reduces profitability by a considerable amount if not aligned with sustainable revenue generation. Therefore, balancing revenue expansion with cost control is critical for maintaining a healthy bottom line and achieving long-term financial stability. These trends validate the importance of cost-effectiveness in business financial planning.

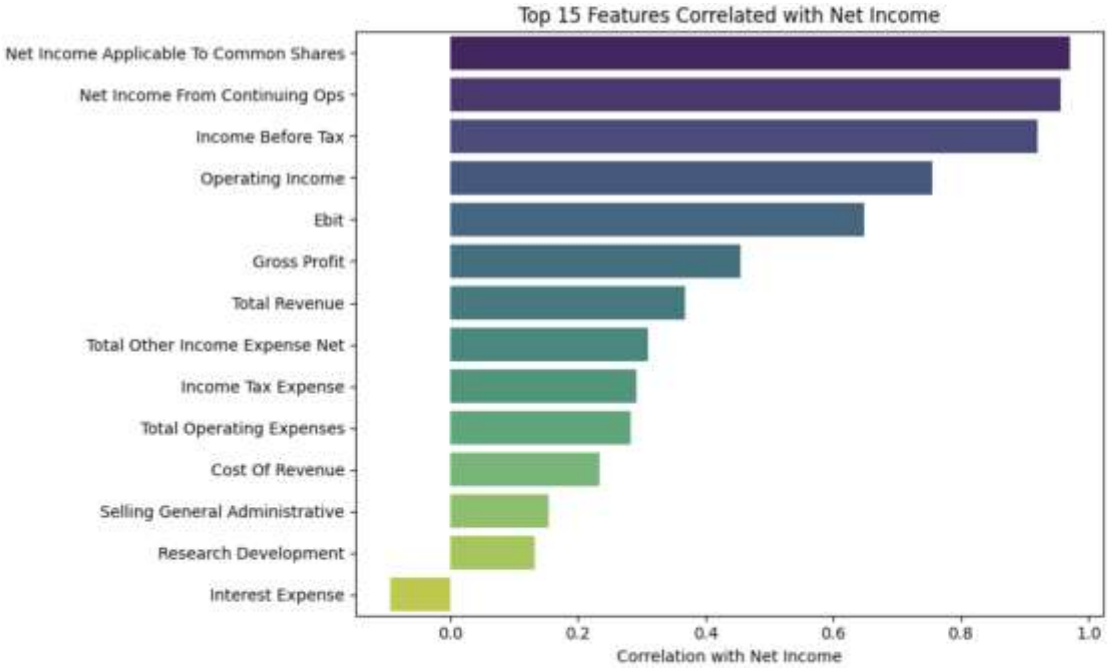


Figure 8: Figure 8. Top 15 correlated features.

Notes: Net income applicable to common shares , net income from continuing ops and negative interest expense , income before tax are highly correlation should be eliminated.

To avoid multicollinearity and data leakage, highly correlated features income before tax, net income from continuing ops and net income applicable to common shares were excluded from the regression models. This allowed the models to learn from independently informative predictors, which enhanced their capability to generalize. James et al. (2013).

4.3 Regression Model Evaluation

All four types of regression models were tested and compared, Linear Regression, Ridge Regression, Lasso Regression, and Polynomial Regression (degree=2). The three measures used to compare them were R², MAE and RMSE. These measures provide an equally weighted description of the capacity to forecast, accuracy in absolute terms, and sensitivity to very large prediction error, respectively Hyndman & Koehler (2006).

Model	R²	MAE	RMSE
Linear Regression	0.902	14.2 Millions	36.6 Millions
Lasso Regression	0.874	16.4 Millions	41.5 Millions
Ridge Regression	0.869	17.0 Millions	42.4 Millions
Polynomial Regression (d=2)	0.898	13.6 Millions	37.3 Millions

Table.2 Accuracy Comparison of Models

The linear regression model achieved the highest R² of 0.902, indicating that over 90% of net income variability was explained by the independent financial factors. A high relationship between the selected KPIs and profitability aligns with the principle that simple, interpretable models often perform well when the underlying relationships are mostly linear (james2013introduction). The polynomial regression model closely followed with an R² of 0.898 but recorded the lowest MAE (13.6 million), which reflects improved-average predictions, particularly in identifying non-linear feature interactions Hastie et al. (2009) like operating income and total revenue.

RMSE of linear regression (36.6 million) also reported the lowest value for all models and reflected minimum susceptibility to extreme errors, an important advantage considering the large range values and potential outliers used in financial data. The lasso and ridge regression reported slightly lower R² values (0.874 and 0.869, respectively) along with higher MAEs and RMSEs, suggesting their imposed bias–variance trade-off due to regularization. Although they yield better generalization and less overfitting probability, their poorer accuracy in this case means that the dataset being high-dimensional did not suffer from multicollinearity severe enough to require heavy penalty Tibshirani (1996).

Mostly, the findings show Net Income is extremely predictable using principal financial indicators, and highest performance was achieved by unregularized models, most likely because features were informative and relatively clean preprocessing. The strong numbers across models confirm the efficacy of financial KPIs in driving bottom-line outcomes and validate their use in analytical forecast models.

Mostly, the findings show Net Income is extremely predictable using principal financial indicators, and highest performance was achieved by unregularized models, most likely because features were informative and relatively clean preprocessing. The strong numbers across models confirm the efficacy of financial KPIs in driving bottom-line outcomes and validate their use in analytical forecast models.

4.4 Interpretability via SHAP Analysis

In addition to model accuracy measurement using traditional metrics such as R^2, MAE, and RMSE, this study also addressed the interpretability implied by Lundberg & Lee (2017) about regression models a key requirement for business applications. To bridge predictive performance and decision-making interpretability, SHAP values were employed. SHAP is a unified cooperative game theory framework that assigns each financial feature a contribution score for a particular prediction. The strength of SHAP lies in its additivity, whereby each prediction can be attributed to the contribution of each feature Lundberg & Lee (2017), allowing both global and local interpretability.

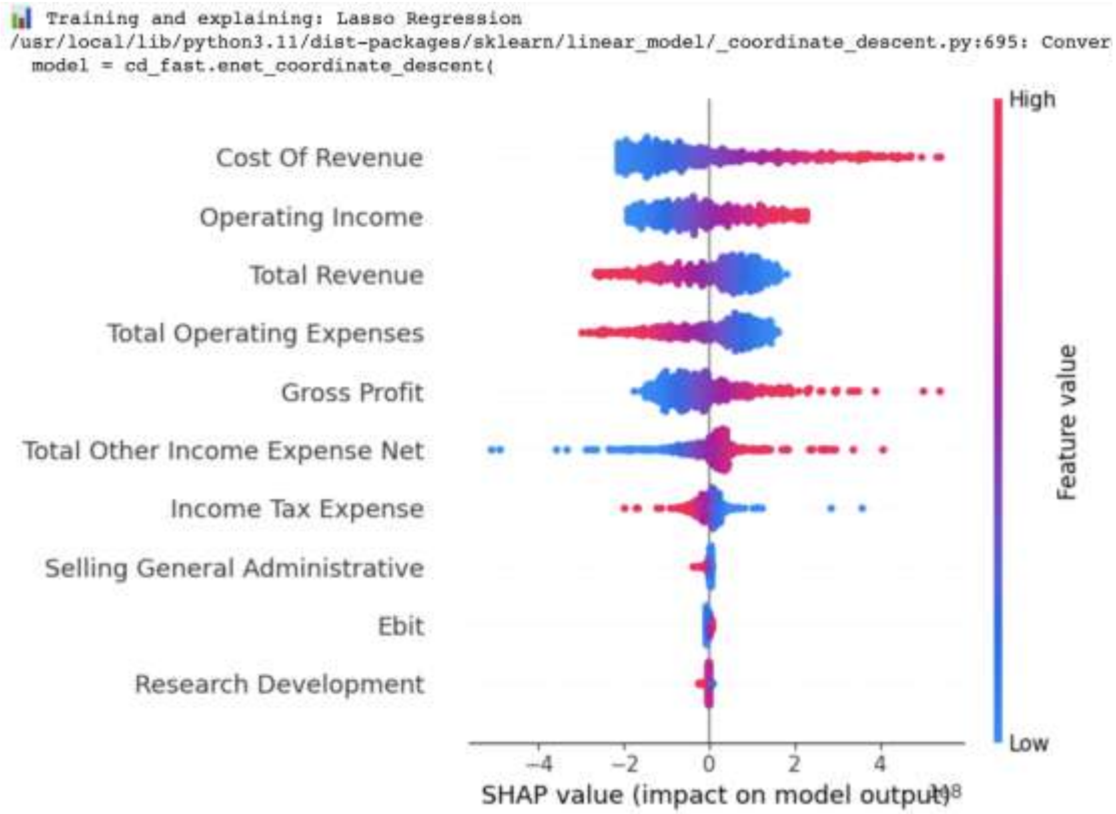
Along with feature rankings, illustrated a nice feature of SHAP is that it can offer local explanations. With summary plots, SHAP visually demonstrated how a particular company’s net income prediction was constructed from its own individual financial profile. For example, a company with high cost of revenue and SG&A but medium gross profit would have those adverse contributors powerfully decrease its prediction, irrespective of any revenue-based positives. These individualized explanations offer correct, case-level detail that is critical stated Kelly & Xiu (2023) for budgeting, forecasting, or audit justification.

Across all models, cost of revenue, total revenue consistently ranked as the most influential but total operating expense and operating income were interchanged for regularization model Lasso & Ridge. These set of features provide strong evidence that operational performance metrics are key drivers of net income. Polynomial models captured subtle interactions missed by linear approaches, while Lasso and Ridge highlighted feature sparsity and penalized complexity. The integration of

SHAP into model evaluation not only reinforced the numerical results but also enhanced transparency, bridging the gap between statistical performance and economic interpretability.

Linear regression, possessing the highest R² (0.902) and the smallest RMSE (36.6 million), also exhibited a strong SHAP profile (Figure 9). cost of revenue and total operating expenses again emerge as top drivers, followed closely by total revenue and gross Profit. The summary plot (Figure 10) discloses balanced SHAP value distribution around zero with equally distributed contribution by feature magnitudes (red vs. blue).

There is a transparent, stable correspondence between operations cost structures and profitability that is reflected by the model. This offers businesses reliable prioritization of margin-controlling levers for financial strategy. As reflected, SHAP makes the model more comfortable to work with so that managers and financial analysts can more readily adopt machine learning in high-stakes environments.



```
📊 Training and explaining: Lasso Regression
/usr/local/lib/python3.11/dist-packages/sklearn/linear_model/_coordinate_descent.py:695: Conver
  model = cd_fast.enet_coordinate_descent(
```

Figure 9: Figure 9. Summary plot of SHAP values for linear regression.

Notes: Highlights cost of revenue and operating income as dominant predictors, with clear direction and consistent influence on Net Income.

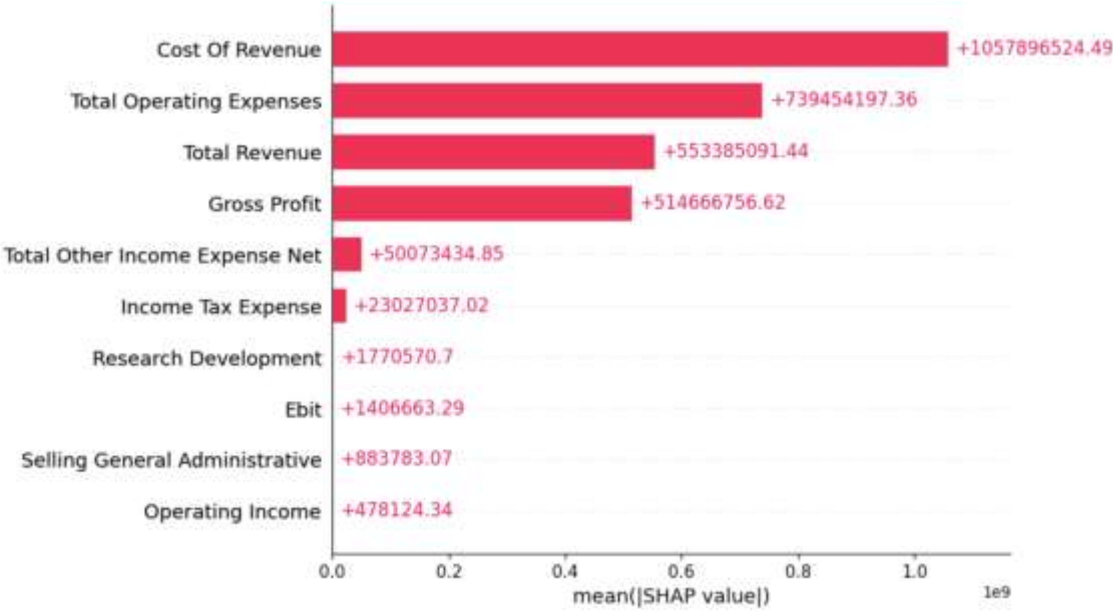
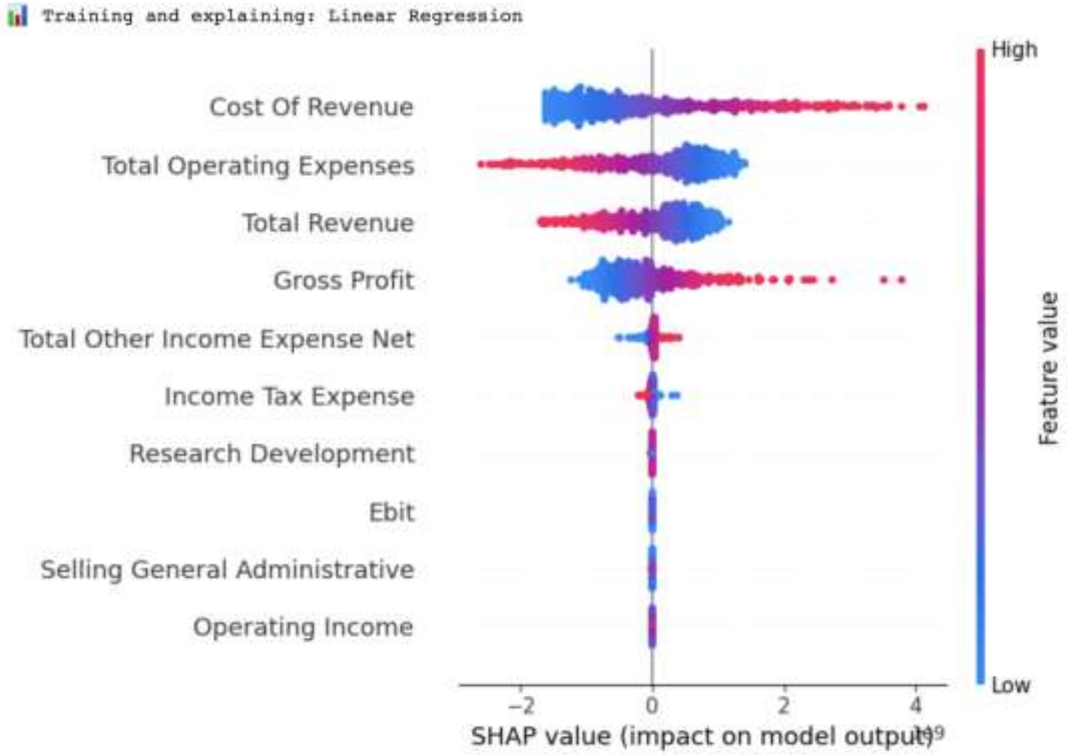


Figure 10: Figure 10. SHAP Bar Plot for linear Regression.

Notes: Ranks features by average contribution, confirming few key KPIs drive most of the predictive power in a stable and interpretable model.

Figure 11 shows that Cost of Revenue, Operating Income, and Total Revenue are the top three predictors of prediction outcomes in the Lasso model with SHAP values over 80–130 million. As shown by the summary plot (Figure 12), high values of Cost of Revenue negatively contribute to Net Income (points with a slope to the left), but high Operating Income and Revenue always push predictions high. This confirms the model’s ability to eliminate weaker signals, allowing cleaner feature selection. For business, this enhances model trust and simplifies insight delivery to financial decision-makers focused on lean KPIs.



```
📊 Training and explaining: Linear Regression
```

Figure 11: Figure 11. SHAP Summary Plot of Lasso Regression.

Notes: Visually sparse, showing Lasso zeroes out low-impact features and isolates high-impact KPIs like operating income.

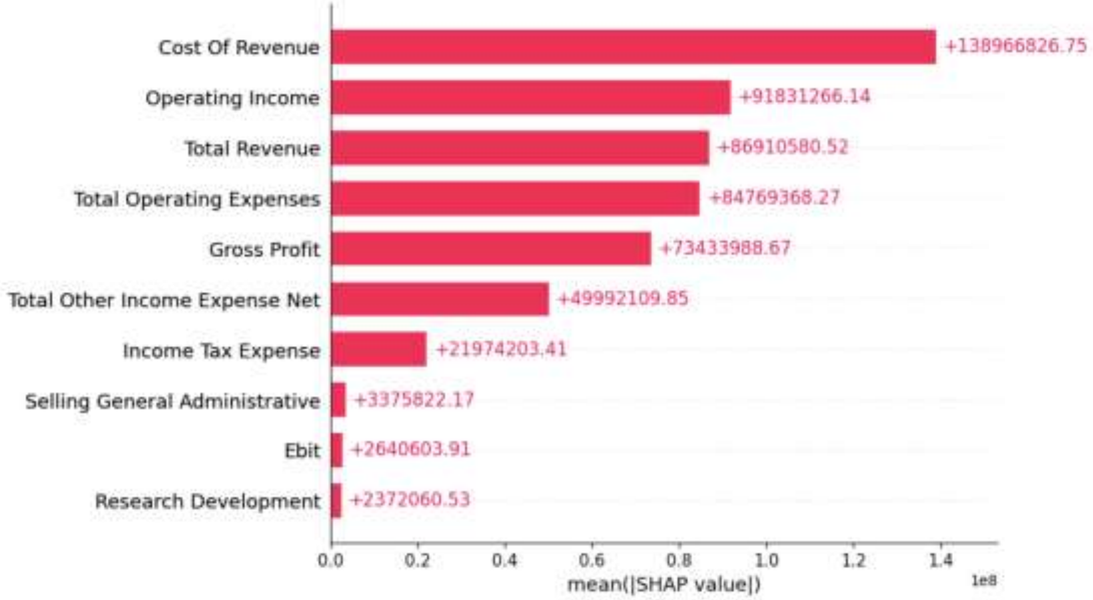
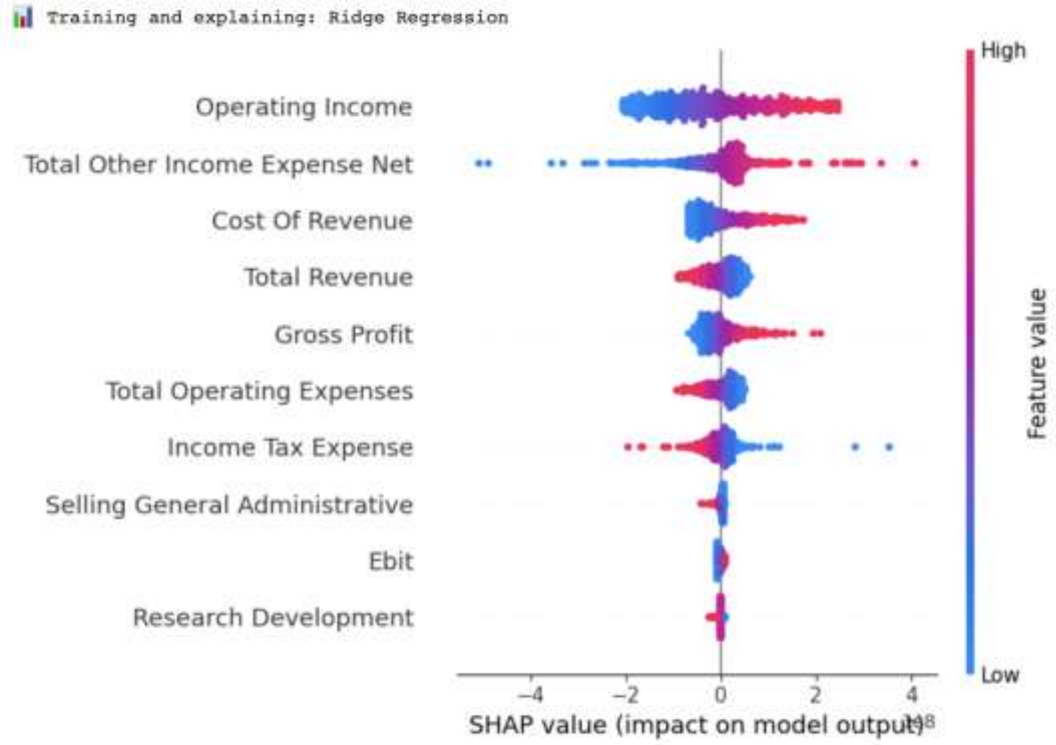


Figure 12: Figure 12. SHAP Bar Plot of Lasso Regression.

Notes: Only a few variables have non-zero influence, making Lasso ideal for simplified, decision-focused financial modelling.

The ridge regression model indicates operating income and other income/ expense (Net) as being leading influencers (Figure 5), which is consistent with its L2 regularization approach of shrinking but retaining coefficients. Its summary plot (Figure 6) shows a centered distribution of SHAP values, higher than Lasso, indicating less and smoother effects of single variable. Ridge kept a broader set of variables with medium effect which is indicative of its ability to capture secondary yet relevant signals. Despite a slightly lower R^2 (0.869) and higher RMSE (42.4 million), Ridge balances bias and variance, making it ideal for businesses requiring generalizable models across noisy financial environments. This model’s inclusiveness is suitable for exploratory use cases in which the goal is to discover the full spectrum of financial drivers.



```
📊 Training and explaining: Ridge Regression
```

Figure 13: Figure 13. Summary plot of SHAP values for Ridge Regression.

Notes: Displays smoother distribution across all features, showing Ridge retains broad input without favouring extreme coefficients.

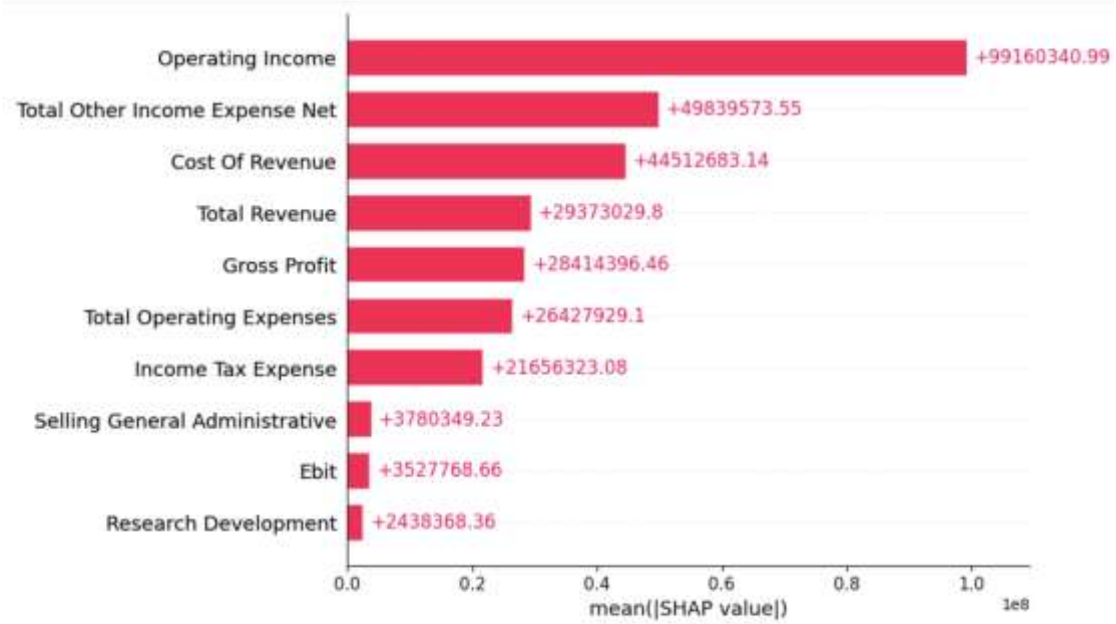


Figure 14: Figure 14. Bar plot of SHAP values for Ridge regressions.

Shows all features contribute moderately, supporting use when full-feature modeling is required for compliance or auditability.

Unlike linear models, polynomial regression can capture interaction terms and non-linear effects (e.g., Total Operating Expenses², Total Revenue × Cost of Revenue). As is evident from the bar plot (Figure 8), higher-order combinations are taking up most explanation space, meaning the model benefits from being able to capture complex relationships in firm-level financial structures. Although it has strong competitive R² (0.898) and lowest MAE (13.6 million), it lacks nuanced interpretability. This is especially useful in strategic simulation or scenario planning, where the collective effect of several KPIs can have concealed hidden profitability dynamics Diamantini et al. (2024).

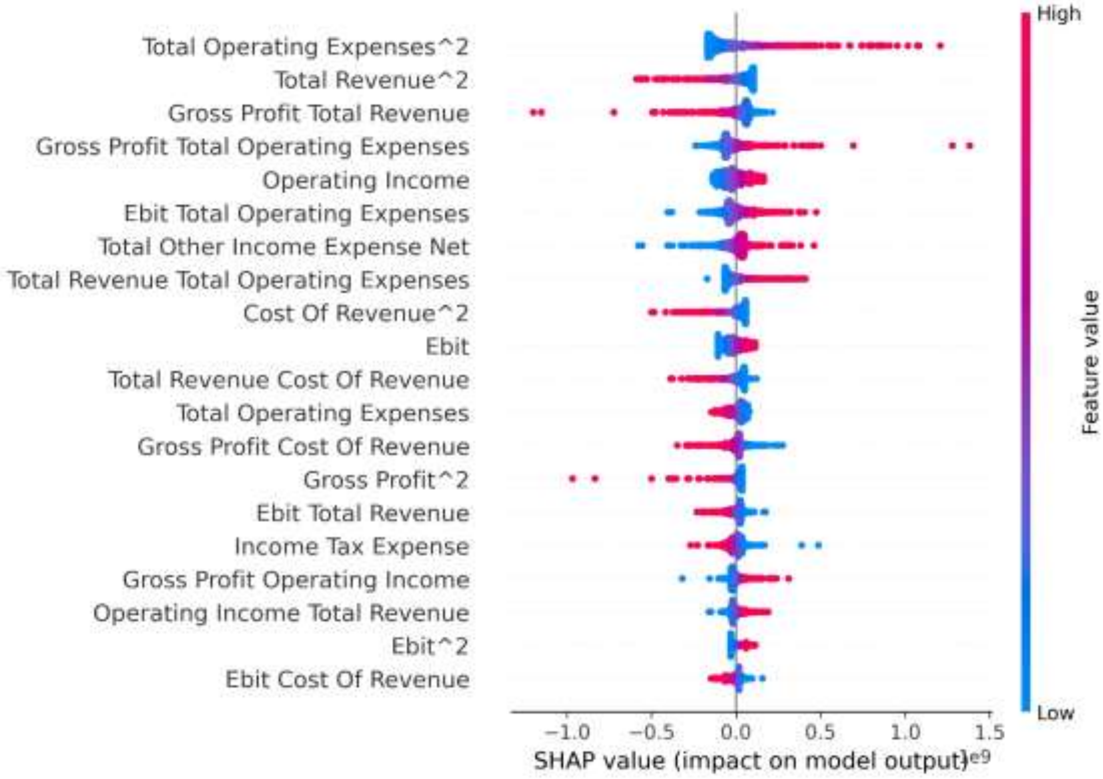


Figure 15: Figure 15. SHAP Summary Plot of Polynomial Regression.

Notes: Captures squared and interaction terms, confirming presence of non-linear effects in Net Income prediction.

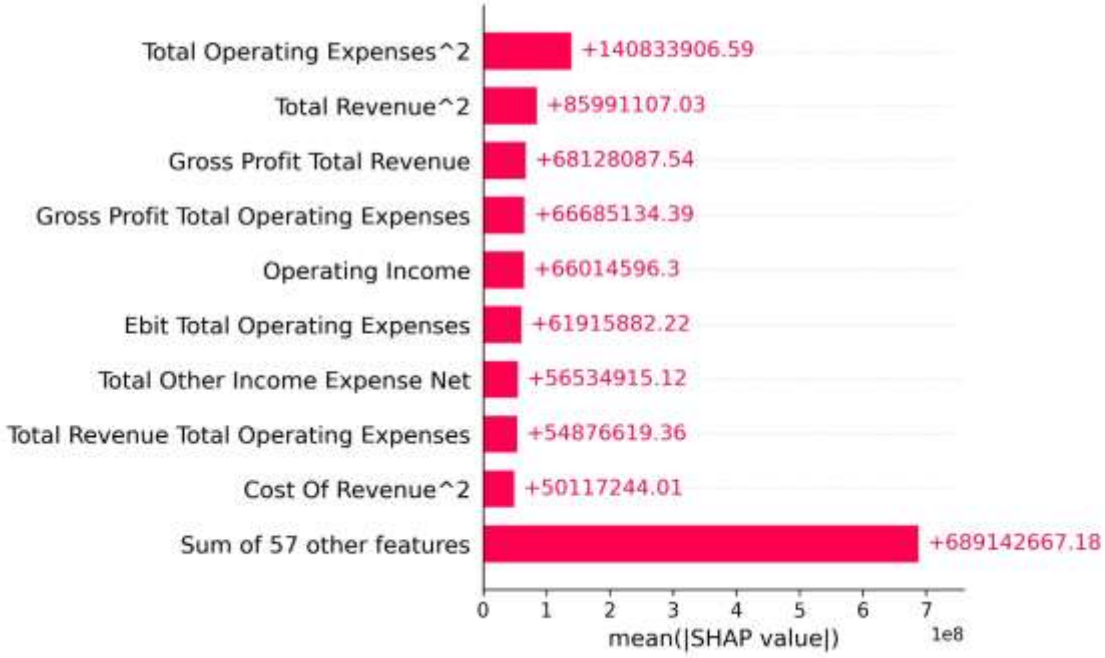


Figure 16: Figure 16. SHAP Bar Plot of Polynomial Regression.

Notes: Reveals complex dependencies; while powerful, interpretability decreases due to many transformed terms.

The incorporation of SHAP in this research not only added transparency but also met emerging regulatory and stakeholder demands for explainable AI. In domains like finance, where models impact investment, credit scores, or compliance, black-box models are often unacceptable as rudin2019stop questioning the usage of black box models. SHAP gives a principled way to open the black box and get specific, interpretable explanations. Interpretability ensures that predictions are not only accurate but also explainable and thereby makes the models suitable for integration into enterprise financial systems, strategic planning software, or investor reports.

4.5 Managerial Implications

The results of this research hold significant implications for financial strategy decision-making in business, for operations planning, and for more general utilization of explainable machine learning. Perhaps the most widely made finding across the regression models was that cost of revenue and SG&A costs were negatively correlated with Net Income prediction. These variables typically surfaced as the most important downward influences on profitability. Their effect reaffirms the significant role of cost management in contributing to profitability. Particularly in industries with razor-thin margins, modest fluctuations within these spending categories can result in extreme earnings fluctuations. The explanatory power of these factors in the SHAP analysis demonstrates unequivocally that profitability is not solely determined by revenue level, but by restraint of spending and structural performance.

Whereas total revenue was positively correlated with net income, SHAP explanations showed that revenue, by itself, was not sufficient to predict profitability. Where revenue was accompanied by out-of-proportion expenses, the net effect on predicted income was neutral or even negative. This indicates a key operational learning revenue expansion needs to be supported by cost-efficient scaling disciplines to create financial benefit. Revenue growth, when pursued alongside a simultaneous diminishment of focus on maintaining margins, can dilute profit and create misleading indicators of company health. This note is in line with a growing body of data that suggests performance needs to be considered both through growth and efficiency metrics, rather than solely top-line performance Rudin (2019). Besides basic drivers of finance, the research also identified industry-specific variation in placing importance on certain features. For example, the predictive ability of Research & Development (R&D) spending varied extensively between the sample. For technology, pharmaceutical, and software firms, R&D contributed significantly to net income forecasts. However, in sectors such as logistics, retail, or utilities, R&D SHAP values were minimal or zero. Similarly, Total Other Income impacted more for diversified holding companies or conglomerates than for specialist operating businesses. This means organizations must consider sectoral and structural context when creating financial models, rather than necessarily applying one-size-fits-all KPI frameworks across various industries. This pragmatism permits focused strategies and model customization that increase accuracy and applicability Lev & Gu (2016).

Moreover, explainability of machine learning models instills trust and adoption in executive environments, especially in financial planning and reporting frameworks where auditability is required. SHAP's ability to reverse-predict results to specific variables makes it more transparent and closes the gap between statistical modeling and successful financial management. In contrast to more traditional black-box approaches that may offer high predictive power but low interpretability, SHAP-enhanced models offer transparent causal narratives that are available to non-technical stakeholders. This alignment with explainable AI best practices is increasingly well-known as a requirement for high-stakes applications, particularly finance, health, and policy-making applications Rudin (2019). Finally, the outputs of the model have action-item implications for managerial decisions. Businesses can use such models to predict what the impact of adjustments in the significant financial levers reducing SG&A or shifting investment from operating expenses to R&D expenses, would be on net income under different assumptions. When embedded in strategic dashboards or budgeting packages, interpretable machine learning models are decision support systems complementing rather than replacing human insight. They offer probabilistic anticipation with human judgment retention, enabling data-driven yet responsive cycles of planning.

In summary, this study illustrates that machine learning when combined with interpretability techniques like SHAP possesses the ability to enhance financial forecast and resource allocation by identifying the strongest financial drivers, noting industry-specific variations, and translating complex model outputs into actionable business insights for decision-makers.

5. 5 Best Practices for Applying Regression Methods in Business Pattern Discovery

To provide transparency and critical reflection on the modeling process a tabular format represented as Table 2. was developed in which to map each methodological step of the thesis to benchmarked best practices, observed deviations, and human guidance intervention. This table is a diagnostic and methodological audit, offering line-by-line comparison of what was performed, what is convention in academic or industry practice, where were the potential risks emerging, and where was expert input essential. Reading across each row, the reasoning and implications of key decisions can be followed from data acquisition and preprocessing to model interpretability and reproducibility. Vertically, the table illustrates how these stages are linked with each other, emphasizing that methodological rigor in machine learning should be coupled not only with technical convention but also with domain-specific reason and business context. This reflective schema enhances the thesis's replicability and emphasizes the importance of human-in-the-loop thinking for applied financial modeling.

Methodol-ogical Stage	Implementation in This Thesis	Best Practice	Deviation / Risk	Human Guidance Implied
Data Acquisition & Validation	Used Kaggle dataset, source mentioned: Yahoo Finance (web-scraped).	Use authoritative, raw, auditable sources.	Risk of calculated fields, derived variables, and incomplete audit trails.	Finance domain expert to validate variable definitions and materiality.
Preprocessing	Median imputation, IQR-based	Use domain-specific imputation and outlier	Aggressive outlier removal may reduce rare	Finance expert should confirm which extreme

Metho dol- ogical Stage	**Implementation in This Thesis**	**Best Practice**	**Deviation / Risk**	**Human Guidance Implied**
g & Cleaning	outlier removal with stricter bounds (42–65 percentile).	strategies (e.g., visualization as per distribution).	but real signals; possible information loss.	values are anomalies vs. strategic cases.
Feature Selection & Multicollinearity Handling	Pearson correlation, manual removal of highly collinear or derivative features.	Use feature importance analysis, domain insight to remove mathematically dependent and logically redundant features.	Manually dropped variables risk data leakage or bias if not guided by business logic.	Human analyst must ensure removed features don’t represent important business signals.
Model Development & Selection	Used Linear, Ridge, Lasso, Polynomial (Degree=2).	Select models based on bias-variance and business use-case alignment.	Polynomial regression risks overfitting; Lasso may exclude marginally important features.	Human-in-the-loop should validate model fit against actual business cycles and causality.
Evalua- tion & Metrics	R², MAE, RMSE used across all models.	Combine statistical metrics with strategic metrics (e.g., ROI, cost of error, impact	Evaluation lacked explicit business cost discussion of misprediction or	Business analyst must translate accuracy metrics into

Methodological Stage	Implementation in This Thesis	Best Practice	Deviation / Risk	Human Guidance Implied
		analysis).	model deployment.	actionable decision thresholds.
Model Interpretability	SHAP applied to all models for local/global explanation.	Combine post-hoc explanations (e.g., SHAP) with interpretable models when stakes are high.	Polynomial model remains hard to interpret despite SHAP; linear model explanations might oversimplify.	Thorough examination of SHAP outputs to link intuition with model output by internal personnel.
Ethics, Bias & Generalizability	Standardized variables to reduce firm-size bias; excluded variables with high leakage risk.	Proactively document bias sources (firm size, sector, derived features).	Skewed distribution from large-cap firms may still bias model; excluded R&D may hide sector signal.	Ethical oversight needed to ensure fair model use across firm sizes and industries.
Reproduc-ibility & Transparency	Data, code and output can be accessed on GitHub and supporting literature.	Use open, version-controlled pipelines with documentation.	Potential reproducibility gap if underlying dataset versioning or licensing is unclear.	Data engineer or reviewer must validate pipeline completeness and license compliance.

Table 3. Best Practices for Applying Regression Methods in Business Pattern Discovery.

Best practices used and exercised here were the outcome of diligent interaction with machine learning (ML) literature as well as financial data modeling best practices. Such best practices were discovered during the critical reading of methodological books like Hastie et al. (2009), Tibshirani (1996), and Lipton (2018), and rigorously adhered to down the pipeline data cleaning, regularization choice, model choice, to SHAP-based interpretation. All such best practices whether involving domain-specific outlier thresholds of deletion or removing redundant variables on information leakage grounds, were guided by business acumen and regulatory caution, not just algorithmic convenience. To convey these to business stakeholders as essential safeguards for ensuring model validity and decision relevance in financial machine learning. Translating models not as enigmatic black boxes rooted in domain intuition or statistical precision but as decision-aiding instruments based on comprehensible logic, finance causality, and trade-off analysis is crucial. The negative impact of cost of revenue on forecasting net income was logical and presented using SHAP and could be taken directly by CFOs or operations planners.

The most accurate and critical mistake were ML users most likely to commit when handling financial datasets in supervised learning is information leakage through highly correlated or derived variables. Items like “net income from continuing operations or income before tax can quantitatively express parts of the target variable and fool model performance. Without human curation or domain knowledge, including it compromises the integrity and usability of the model something carefully avoided in this study in following Chan et al. (2022) and Montesinos López et al. (2022) guidelines.

For an entry-level machine learner interested in establishing financial analysis capabilities, this research provides an interpretable and reproducible framework. It not only indicates how to employ regression models but also why features are important, how to identify statistical traps such as multicollinearity, and how to bring SHAP values to business interpretation. The path from preprocessing to explainability not only gives students technical competence, but strategic insight a twofold talent increasingly demanded by careers in data-driven business.

The structured interpretation and rigorous evaluation conducted in this thesis reflect a progression from introductory understanding to applied competency in supervised learning within financial contexts. Through confronting practical challenges such as multicollinearity, information leakage, and the interpretability–accuracy trade-off, this work demonstrates how beginner-level knowledge can evolve into more analytical, method-driven engagement with machine learning models. It illustrates the critical importance of aligning model development with business context, ethical use, and stakeholder transparency. As such, this thesis serves as an example of how early-stage practitioners can transition toward more nuanced and responsible applications of machine learning in fields such as strategic finance, regulated analytics, and explainable AI.

6. Limitation

Even though it is founded on a rigorous methodological basis, this study was faced with numerous everyday challenges that governed both its scope and its intensity. Among the most basic and yet relentlessly obstinate of these challenges was data irregularity: missing values, conflicting financial reporting, and unbalanced distributions had a tendency to complicate assumptions in modelling and required incessant readjustments of preprocessing strategies. Time restrictions were another exacting limitation. The thesis was written according to a rigorous academic schedule, limiting experimentation of many modelling pipelines or intensifying cross-sectional comparisons between industries. Additionally, the intersection between business usability and machine learning depth vital for real-world practicability required ongoing tuning. Switching continually between financial reasoning and data science reasoning meant no step could be fully automated or outsourced; domain verification and interpretation conformity were required in each model decision. This cross-firm, interdisciplinary scope, as valuable as it was, necessarily constrained the ability of the thesis to project out to broader cross-firm or time-series analysis. Furthermore, given the real-world complexity of implementing interpretive models in the commercial environment, this study is more diagnostic than prescriptive.

7. References

Addepalli, L. et al. (2023). Assessing the performance of python data visualization libraries: A review. *International Journal of Computer Engineering in Research Trends*, *10*(1), 29–39. https://doi.org/10.22362/ijcert/2023/v10/i01/v10i0104

Aggarwal, C. C. (2016). An introduction to outlier analysis. In *Outlier analysis* (pp. 1–34). Springer. https://doi.org/10.1007/978-3-319-47578-3_1

Bell, A., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2022, June). It's just not that simple: An empirical study of the accuracy–explainability trade-off in machine learning for public policy. *Proceedings of the 2022 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 248–266. https://doi.org/10.1145/3531146.3533090

Boritz, J. E., & No, W. G. (2019). How significant are the differences in financial data provided by key data sources? A comparison of XBRL, compustat, yahoo ! Finance, and google finance. *Journal of Information Systems*, *34*(3), 47–75. https://doi.org/10.2308/isys-52618

Broby, D. (2022). The use of predictive analytics in finance. *The Journal of Finance and Data Science*, *8,* 145–161. https://doi.org/10.1016/j.jfds.2022.05.003

Brzozowska, J., Pizoń, J., Baytikenova, G., Gola, A., Zakimova, A., & Piotrowska, K. (2023). Data engineering in CRISP-DM process production data – case study. *Applied Computer Science*, *19*(3), 83–95. https://acs.pollub.pl/pdf/v19n3/6.pdf

Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., & Chen, Y.-L. (2022). Mitigating the multicollinearity problem and its machine learning approach: A review. *Mathematics*, *10*(8), 1283. https://doi.org/10.3390/math10081283

Choudhury, P., Allen, R., & Endres, M. G. (2021). Machine learning for pattern discovery in management research. *Strategic Management Journal*, *42*(1), 30–57. https://www.hbs.edu/ris/Publication%20Files/Machine%20learning%20for%20pattern%20discovery%20in%20management%20research_37653690-e05a-42a6-9bdc-46f408efe8a6.pdf

Diamantini, C., Khan, T., Mircoli, A., & Potena, D. (2024). Enhancing KPI forecasting through regression algorithms using historical data. In X. S. Yang, S. Sherratt, N. Dey, & A. Joshi (Eds.), *Proceedings of ninth international congress on information and communication technology* (pp. 1013–1022). Springer. https://doi.org/10.1007/978-981-97-3559-4_36

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Oliphant, T. E., et al. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Holthausen, R. W., & Watts, R. L. (2001). The relevance of the value-relevance literature for financial accounting standard setting. *Journal of Accounting and Economics*, *31*(1–3), 3–75. https://doi.org/10.1016/S0165-4101(01)00029-5

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r* (Vol. 103). Springer.

Kelly, B., & Xiu, D. (2023). Financial machine learning. *Foundations and Trends® in Finance*, *13*(3-4), 205–363. https://bfi.uchicago.edu/wp-content/uploads/2023/07/BFI_WP_2023-100.pdf

Khan, F. S., Mazhar, S. S., Mazhar, K., AlSaleh, D. A., & Mazhar, A. (2025). Model-agnostic explainable artificial intelligence methods in finance: A systematic review, recent developments, limitations, challenges and future directions. *Artificial Intelligence Review*, *58*(8). https://doi.org/10.1007/s10462-025-11215-9

Lev, B., & Gu, F. (2016). *The end of accounting and the path forward for investors and managers*. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/epdf/10.1002/9781119270041.fmatter

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 36–43. https://doi.org/10.1145/3233231

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Third Edition). John Wiley & Sons. https://content.e-bookshelf.de/media/reading/L-12468270-75fd0fbf69.pdf

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777. https://dl.acm.org/doi/10.5555/3295222.3295230

Mathotaarachchi, K. V., Hasan, R., & Mahmood, S. (2024). Advanced machine learning techniques for predictive modeling of property prices. *Information*, *15*(6), 1–35. https://doi.org/10.3390/info15060295

Mizgajski, J., Szymczak, A., Morzy, M., Augustyniak, Ł., Szymański, P., & Żelasko, P. (2021). Return on investment in machine learning: Crossing the chasm between academia and business. *Foundations of Computing and Decision Sciences*, *45*(4), 281–304. https://doi.org/10.2478/fcds-2020-0015

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate statistical machine learning methods for genomic prediction* (pp. 109–139). Springer. https://doi.org/10.1007/978-3-030-89010-0_4

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2022). A survey on human-in-the-loop machine learning: A scalable and interactive approach to machine learning. *Artificial Intelligence Review*, *55*, 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *Information*, *11*(4), 193. https://doi.org/10.3390/info11040193

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://www.jstor.org/stable/2346178

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Appendix a

Case Understanding & objective

This notebook contains methodology applied to evaluate Regression based Supervised Machine learning models in form of accuracy, interpreteablility & explainability. To address data driven decision when applied to real business dataset of S&P 500 of the largest publicly traded companies in the United States.

Main business question driving analysis: How the components of income statement, cash flow statements & balance sheet affect the labelled variable net income ?

```
# STEP 1: IMPORT LIBRARIES
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import Lasso
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
from sklearn.metrics import accuracy_score, confusion_matrix, ConfusionMatrixDisplay
from sklearn.linear_model import LinearRegression, Ridge, Lasso
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler, PolynomialFeatures
import shap
shap.initjs()
```

```
#LOAD DATA
file_path = '/content/financial data sp500 companies.csv'
financial_KPIs = pd.read_csv(file_path)
```

```
#to check variable properties
financial_KPIs.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2012 entries, 0 to 2011
Data columns (total 19 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Unnamed: 0                          2012 non-null   int64
 1   date                                2012 non-null   object
 2   firm                                2012 non-null   object
 3   Ticker                              2012 non-null   object
 4   Research Development                634 non-null    float64
 5   Income Before Tax                   2011 non-null   float64
 6   Net Income                          2012 non-null   float64
 7   Selling General Administrative      1952 non-null   float64
 8   Gross Profit                        2012 non-null   float64
 9   Ebit                                2012 non-null   float64
 10  Operating Income                    2011 non-null   float64
 11  Interest Expense                    1830 non-null   float64
 12  Income Tax Expense                  2012 non-null   float64
 13  Total Revenue                       2012 non-null   float64
 14  Total Operating Expenses            2012 non-null   float64
 15  Cost Of Revenue                     2012 non-null   float64
 16  Total Other Income Expense Net      2012 non-null   float64
 17  Net Income From Continuing Ops      2011 non-null   float64
 18  Net Income Applicable To Common Shares  2011 non-null   float64
dtypes: float64(15), int64(1), object(3)
memory usage: 298.8+ KB
```

Date ,firm & ticker are object data types and all other variables are float64 and best suitable for regression methods.

Data Preprocessing

```
#BASIC CLEANING
financial_KPIs.drop(columns=['Unnamed: 0'], errors='ignore', inplace=True)
financial_KPIs.drop_duplicates(inplace=True)
```

```
financial_KPIs.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 2008 entries, 0 to 2011
Data columns (total 18 columns):
 #   Column                                Non-Null Count  Dtype  
---  ------                                --------------  -----  
 0   date                                  2008 non-null   object 
 1   firm                                  2008 non-null   object 
 2   Ticker                                2008 non-null   object 
 3   Research Development                  630 non-null    float64
 4   Income Before Tax                     2007 non-null   float64
 5   Net Income                            2008 non-null   float64
 6   Selling General Administrative        1948 non-null   float64
 7   Gross Profit                          2008 non-null   float64
 8   Ebit                                  2008 non-null   float64
 9   Operating Income                      2007 non-null   float64
 10  Interest Expense                      1826 non-null   float64
 11  Income Tax Expense                    2008 non-null   float64
 12  Total Revenue                         2008 non-null   float64
 13  Total Operating Expenses              2008 non-null   float64
 14  Cost Of Revenue                       2008 non-null   float64
 15  Total Other Income Expense Net        2008 non-null   float64
 16  Net Income From Continuing Ops        2007 non-null   float64
 17  Net Income Applicable To Common Shares  2007 non-null   float64
dtypes: float64(15), object(3)
memory usage: 298.1+ KB
```

After dropping duplicates 4 entries are deleted.

```
#check missing count
financial_KPIs.isnull().sum()
```

```
                                          0
                           date           0
                           firm           0
                         Ticker           0
           Research Development        1378
              Income Before Tax           1
                     Net Income           0
  Selling General Administrative         60
                   Gross Profit           0
                           Ebit           0
               Operating Income           1
               Interest Expense         182
             Income Tax Expense           0
                  Total Revenue           0
       Total Operating Expenses           0
                Cost Of Revenue           0
 Total Other Income Expense Net           0
 Net Income From Continuing Ops           1
Net Income Applicable To Common Shares    1

dtype: int64
```

Impute median in all missing values

```
financial_KPIs['Research Development'] = financial_KPIs['Research Development'].fillna(financial_KPIs['Research Development'].median())
financial_KPIs['Interest Expense'] = financial_KPIs['Interest Expense'].fillna(financial_KPIs['Interest Expense'].median())
financial_KPIs['Selling General Administrative'] = financial_KPIs['Selling General Administrative'].fillna(financial_KPIs['Selling Genera
financial_KPIs['Net Income From Continuing Ops'] = financial_KPIs['Net Income From Continuing Ops'].fillna(financial_KPIs['Net Income Fro
financial_KPIs['Net Income Applicable To Common Shares'] = financial_KPIs['Net Income Applicable To Common Shares'].fillna(financial_KPIs
financial_KPIs['Operating Income'] = financial_KPIs['Operating Income'].fillna(financial_KPIs['Operating Income'].median())
financial_KPIs['Income Before Tax'] = financial_KPIs['Income Before Tax'].fillna(financial_KPIs['Income Before Tax'].median())
```

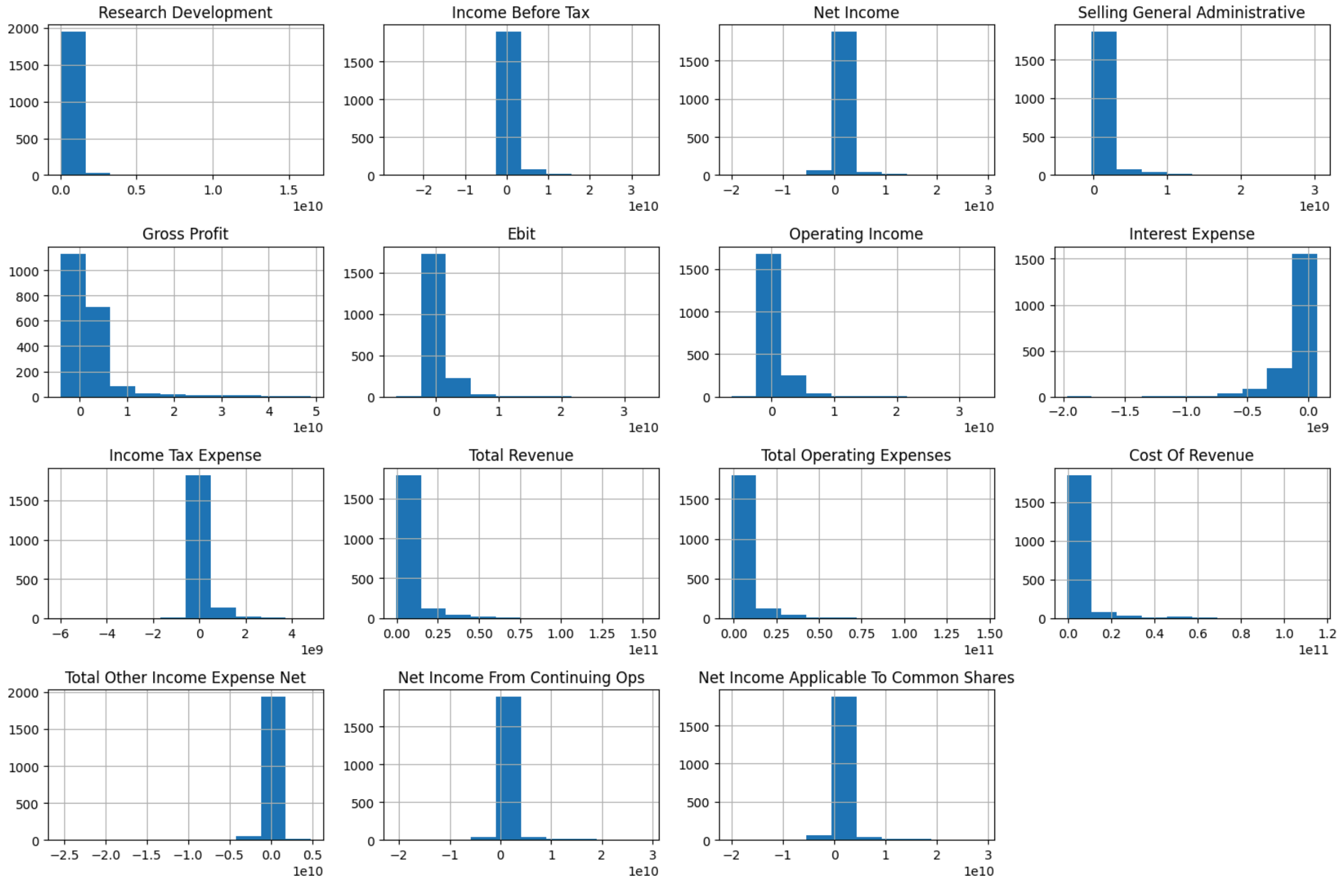
Explorartory Data Analysis

```
#to check spread of data
financial_KPIs.describe()
```

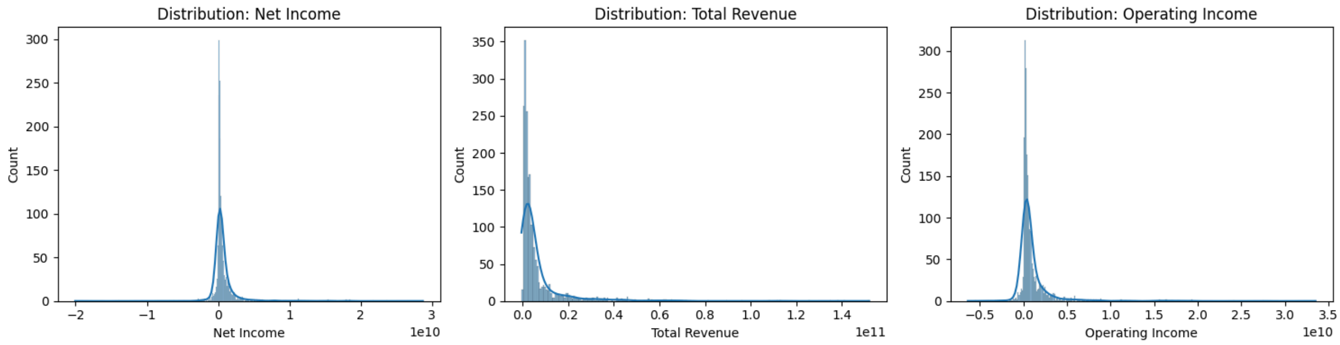
	Research Development	Income Before Tax	Net Income	Selling General Administrative	Gross Profit	Ebit	Operating Income	Interest Expense	Income Expe
count	2.008000e+03	2.008000e+03	2.008000e+03	2.008000e+03	2.008000e+03	2.008000e+03	2.008000e+03	2.008000e+03	2.008000e
mean	3.108597e+08	8.718564e+08	7.150870e+08	1.081071e+09	2.564063e+09	8.703134e+08	1.016716e+09	-1.084421e+08	1.567633e
std	8.717355e+08	2.488077e+09	2.110919e+09	2.486674e+09	5.050444e+09	2.110724e+09	2.277994e+09	1.654547e+08	4.436321e
min	-1.030000e+07	-2.661300e+10	-2.007000e+10	-3.613000e+09	-4.062000e+09	-6.389000e+09	-6.389000e+09	-1.972000e+09	-6.010000e
25%	1.625000e+08	1.475888e+08	1.197518e+08	1.581908e+08	5.377618e+08	1.607500e+08	2.003418e+08	-1.190000e+08	1.758150e
50%	1.625000e+08	3.450000e+08	2.816470e+08	3.793250e+08	1.050040e+09	3.557670e+08	4.230000e+08	-5.520000e+07	6.295000e
75%	1.625000e+08	8.327000e+08	6.710118e+08	8.660000e+08	2.131500e+09	8.152500e+08	9.100000e+08	-2.500000e+07	1.643280e
max	1.646600e+10	3.357900e+10	2.875500e+10	3.033100e+10	4.890400e+10	3.353400e+10	3.353400e+10	7.200000e+07	4.824000e

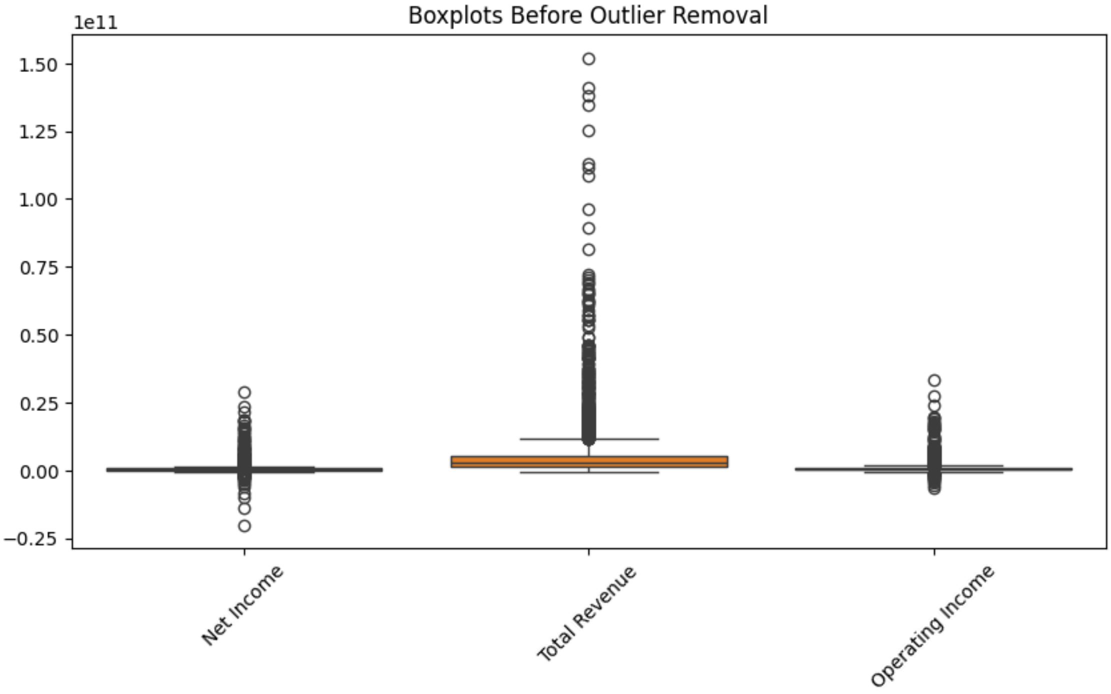
Total revenue and operating income might have highest outliers.

```
#plot histogram for visualization of data spread
financial_KPIs.hist(figsize=(15, 10))
plt.tight_layout()
plt.show()
```



```
#INITIAL EDA (Before Outlier Treatment)
plt.figure(figsize=(15, 4))
for i, col in enumerate(['Net Income', 'Total Revenue', 'Operating Income']):
    plt.subplot(1, 3, i+1)
    sns.histplot(financial_KPIs[col].dropna(), kde=True)
    plt.title(f'Distribution: {col}')
plt.tight_layout()
plt.show()

# Boxplots
plt.figure(figsize=(10, 5))
sns.boxplot(data=financial_KPIs[['Net Income', 'Total Revenue', 'Operating Income']])
plt.title("Boxplots Before Outlier Removal")
plt.xticks(rotation=45)
plt.show()
```



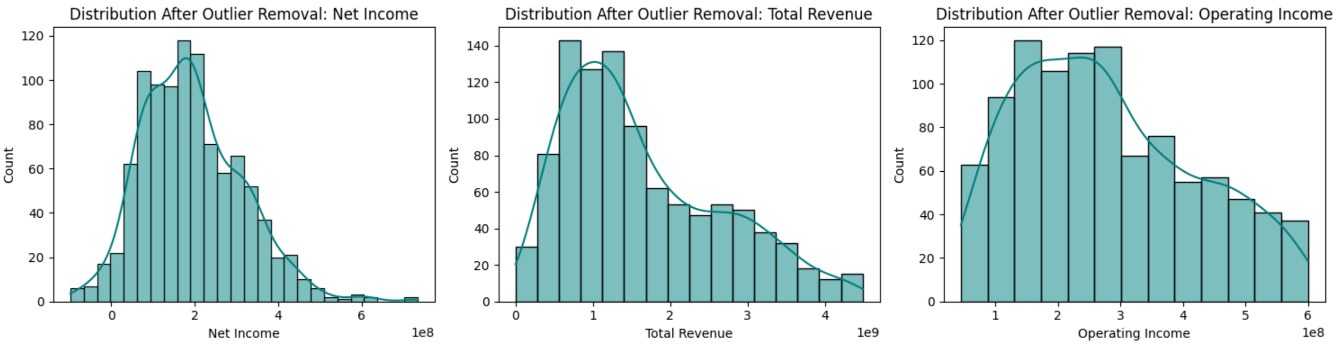


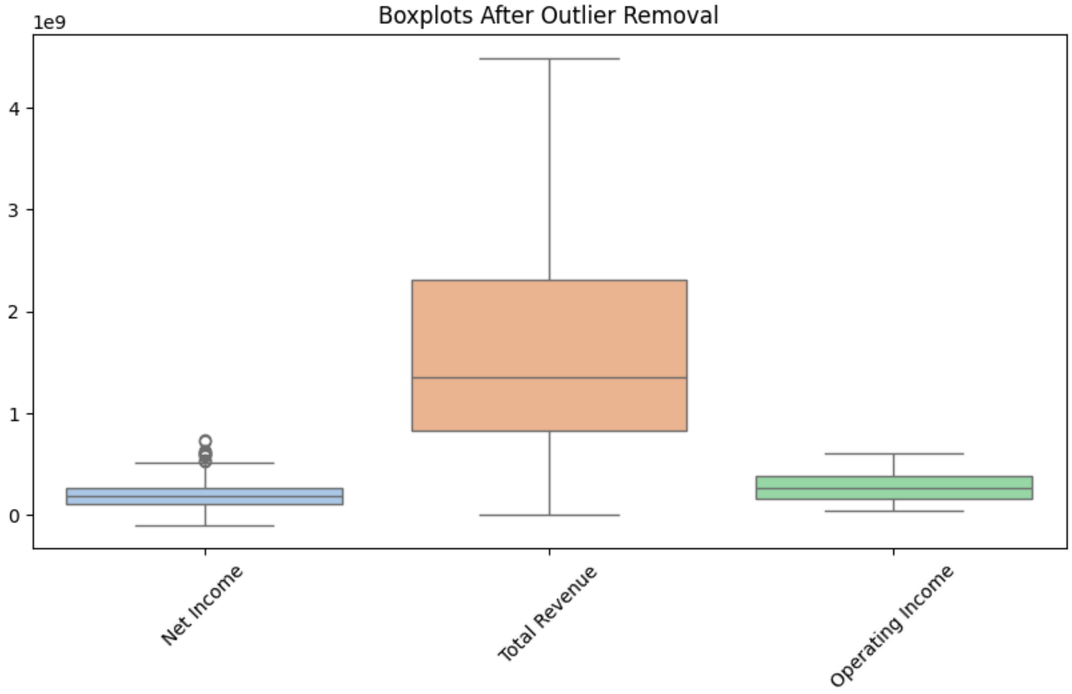
The values in net income should be limit between the standard deviation for stable modelling outputs using interquartile limits.

```
#OUTLIER REMOVAL USING IQR METHOD
def remove_outliers_iqr(dataframe, cols):
    for col in cols:
        Q1 = dataframe[col].quantile(0.42)
        Q3 = dataframe[col].quantile(0.65)
        IQR = Q3 - Q1
        lower = Q1 - 1.5 * IQR
        upper = Q3 + 1.5 * IQR
        dataframe = dataframe[(dataframe[col] >= lower) & (dataframe[col] <= upper)]
    return dataframe

financial_KPIs = remove_outliers_iqr(financial_KPIs, ['Net Income', 'Total Revenue', 'Operating Income'])
```

```
# EDA AFTER OUTLIER REMOVAL
plt.figure(figsize=(15, 4))
```

```
for i, col in enumerate(['Net Income', 'Total Revenue', 'Operating Income']):
    plt.subplot(1, 3, i+1)
    sns.histplot(financial_KPIs[col].dropna(), kde=True, color='teal')
    plt.title(f'Distribution After Outlier Removal: {col}')
plt.tight_layout()
plt.show()

plt.figure(figsize=(10, 5))
sns.boxplot(data=financial_KPIs[['Net Income', 'Total Revenue', 'Operating Income']], palette="pastel")
plt.title("Boxplots After Outlier Removal")
plt.xticks(rotation=45)
plt.show()
```

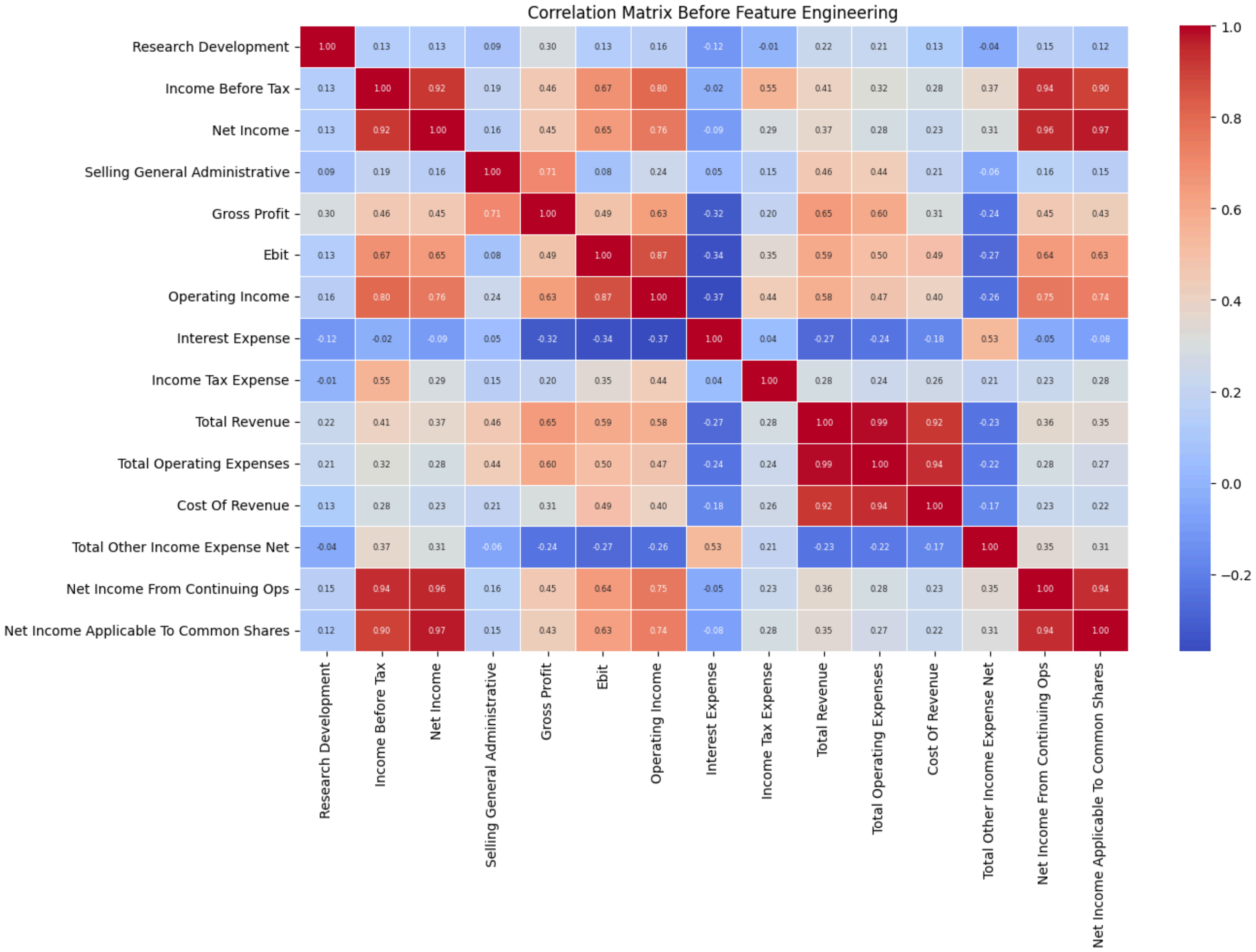




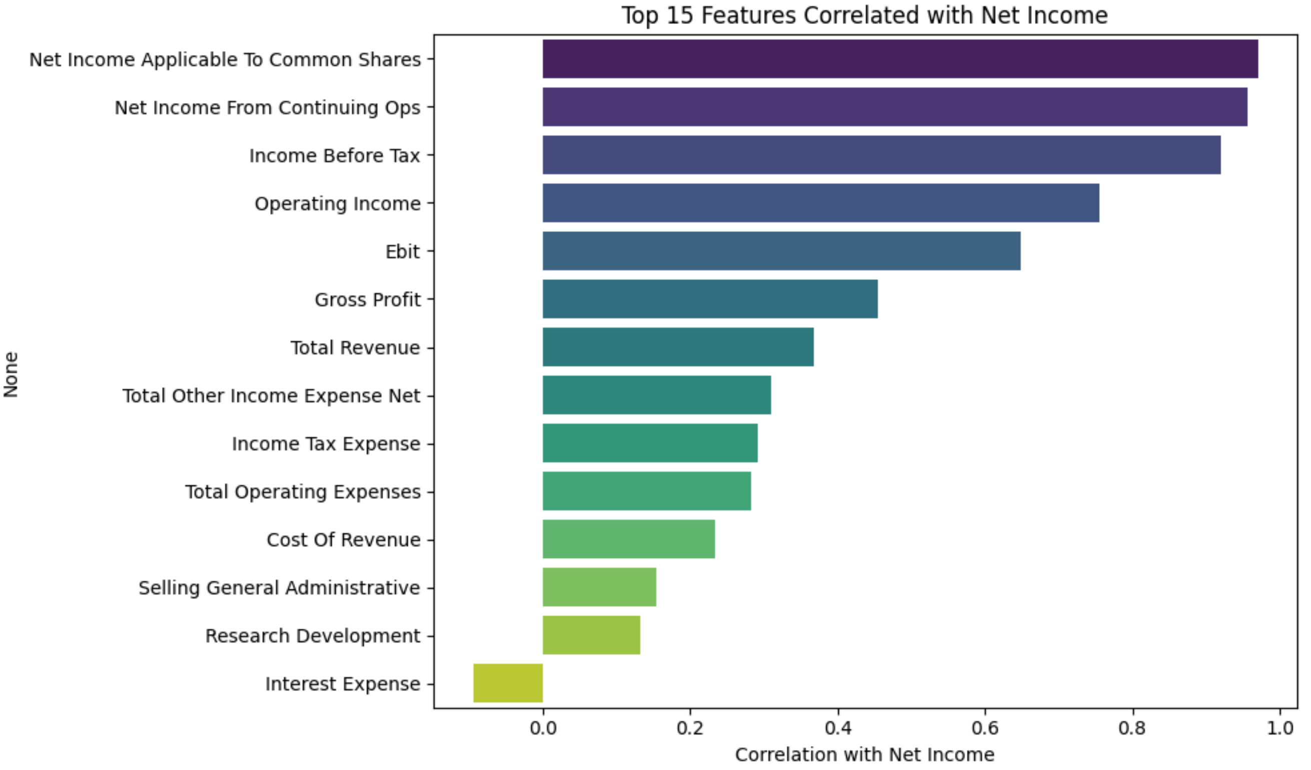
Net income still has outliers but to avoid overfitting, as values are already fitted between 0.35 to 0.62.

```
# Calculate the correlation matrix
correlation_matrix = financial_KPIs.select_dtypes(include=np.number).corr()

# Heatmap with readable annotations
plt.figure(figsize=(14, 10))
sns.heatmap(correlation_matrix, cmap='coolwarm', annot=True, fmt=".2f", linewidths=0.5, annot_kws={"size": 6})
plt.title("Correlation Matrix Before Feature Engineering")
plt.tight_layout()
plt.show()

# Top 15 features most correlated with Net Income
top_corr = correlation_matrix['Net Income'].drop('Net Income').sort_values(ascending=False).head(15)

# Plot Top 15 Correlations
plt.figure(figsize=(10, 6))
sns.barplot(x=top_corr.values, y=top_corr.index, palette='viridis')
plt.title("Top 15 Features Correlated with Net Income")
```

```
    plt.xlabel("Correlation with Net Income")
    plt.tight_layout()
    plt.show()
```



```
/tmp/ipython-input-14-719545562.py:16: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `le

  sns.barplot(x=top_corr.values, y=top_corr.index, palette='viridis')
```



Variables to eliminate: Net income applicable to common shares , net income from continuing ops and income before tax are highly correleated. Interest expense has negative correlation and decrease the predictive accuracy.

Feature Engineering, Spliting Data & Model Initiating.

```
#PREPARE FEATURES & TARGET
X = financial_KPIs.drop(columns=['Net Income','Net Income Applicable To Common Shares','Net Income From Continuing Ops','Interest Expens
y = financial_KPIs['Net Income']
```

```
#SCALE & SPLIT
# Drop non-numeric columns before scaling
X_numeric = X.select_dtypes(include=np.number)

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X_numeric)
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3, random_state=42)
```

Split of 70% training & 30% testing is implied. Standard scaler scale the high vaules near to zero taking mean.

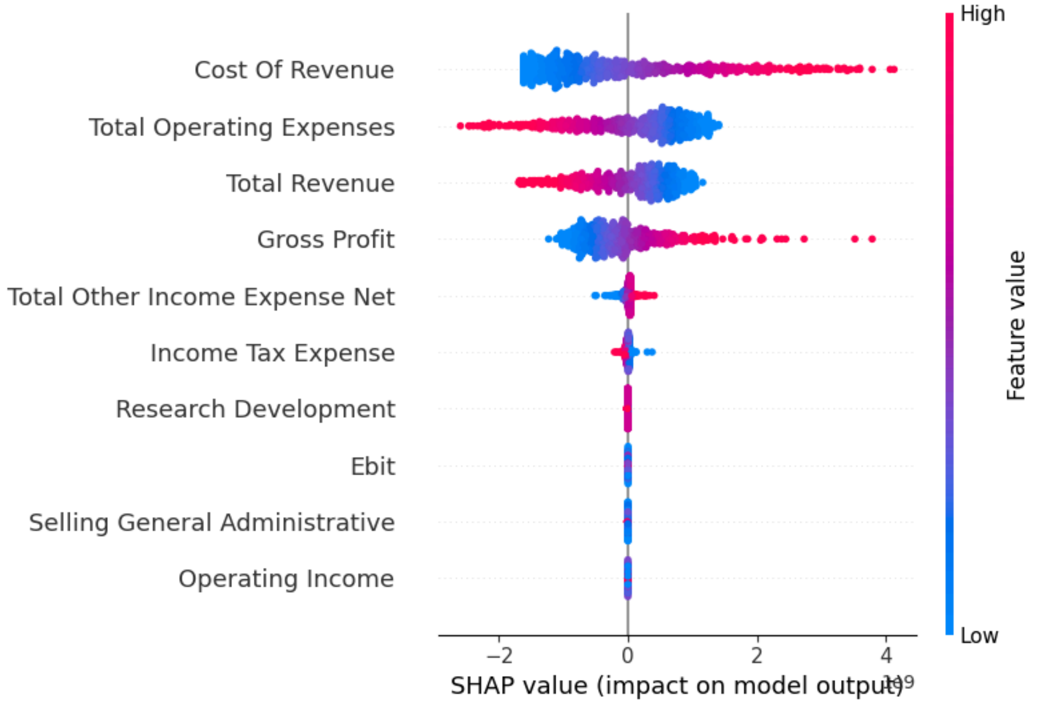
```
#DEFINE MODELS
models = {
    "Linear Regression": LinearRegression(),
    "Ridge Regression": Ridge(alpha=1.0),
    "Lasso Regression": Lasso(alpha=0.1),
    }
```

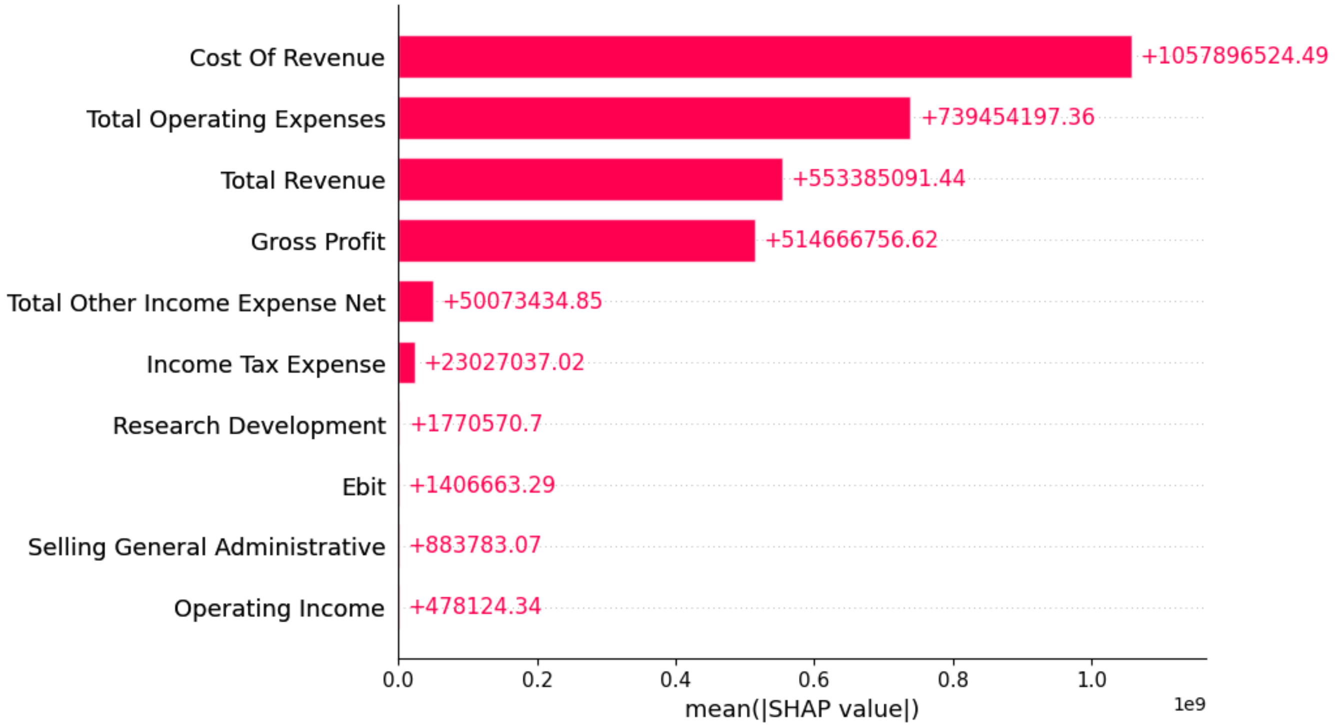
Selected white box models to evaulate. Alpha values indicate regularization parameter: L1 for lasso & L2 for ridge regression.

Model Evaluation & Explaination

```
results = []
shap_outputs = {} # Dictionary to store SHAP outputs

# Loop through each model
for name, model in models.items():
    print(f"\n📊 Training and explaining: {name}")

    # Fit model
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    # Store evaluation results
    results.append({
        "Model": name,
        "R²": r2_score(y_test, y_pred),
        "MAE": mean_absolute_error(y_test, y_pred),
        "RMSE": np.sqrt(mean_squared_error(y_test, y_pred))
    })

    # Create SHAP explainer with feature names
    # Ensure X_train is a DataFrame with correct column names
    X_train_df = pd.DataFrame(X_train, columns=X_numeric.columns)
    explainer = shap.Explainer(model, X_train_df)
    shap_values = explainer(X_train_df)

    # Store SHAP outputs
    shap_outputs[name] = (shap_values, X_train_df)

    # Plot SHAP summary and bar chart
    shap.summary_plot(shap_values, X_train_df)
    shap.plots.bar(shap_values, max_display=10)


    # Results DataFrame
    results_df = pd.DataFrame(results).sort_values(by="R²", ascending=False)
    print("\n📈 Model Comparison Results:\n")
    display(results_df)
```

```
📊 Training and explaining: Linear Regression
```

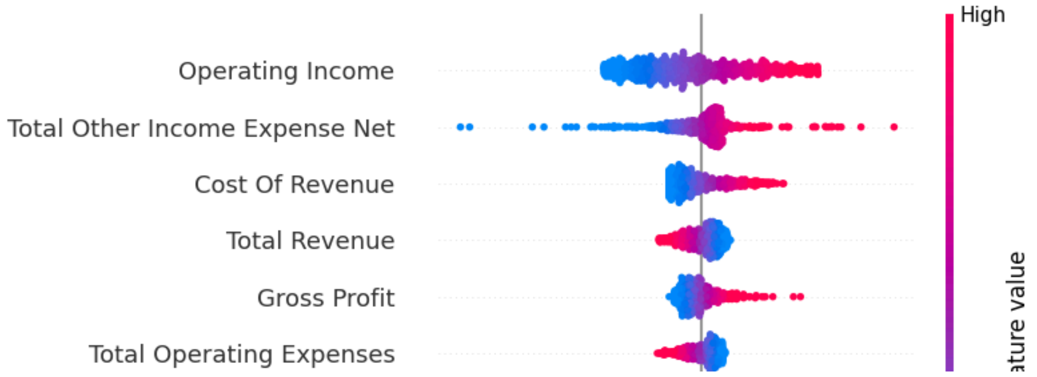


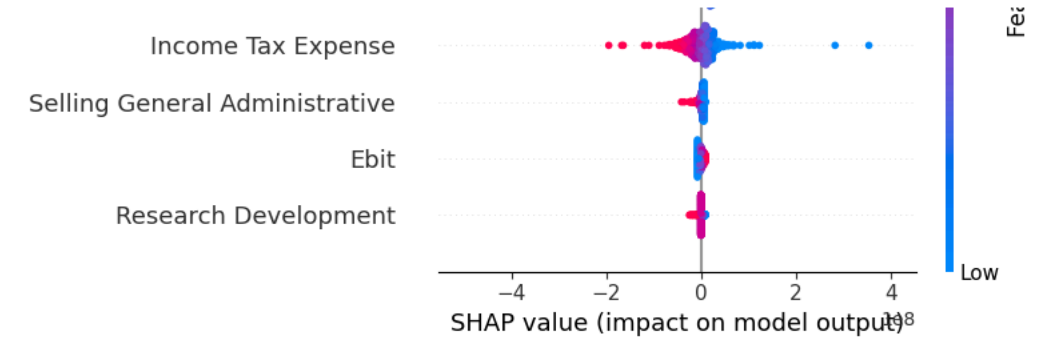


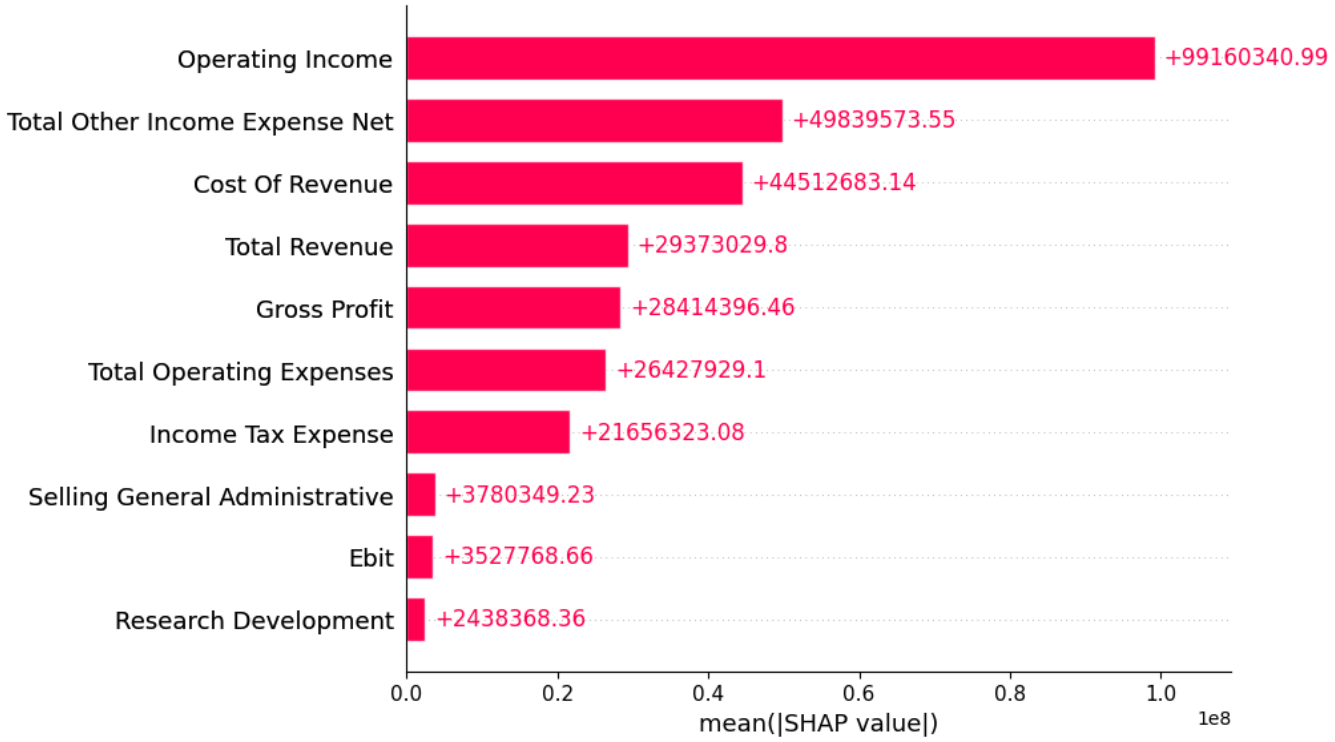
```
📈 Model Comparison Results:
```

	Model	R²	MAE	RMSE
0	Linear Regression	0.902071	1.422822e+07	3.667882e+07

```
📊 Training and explaining: Ridge Regression
```



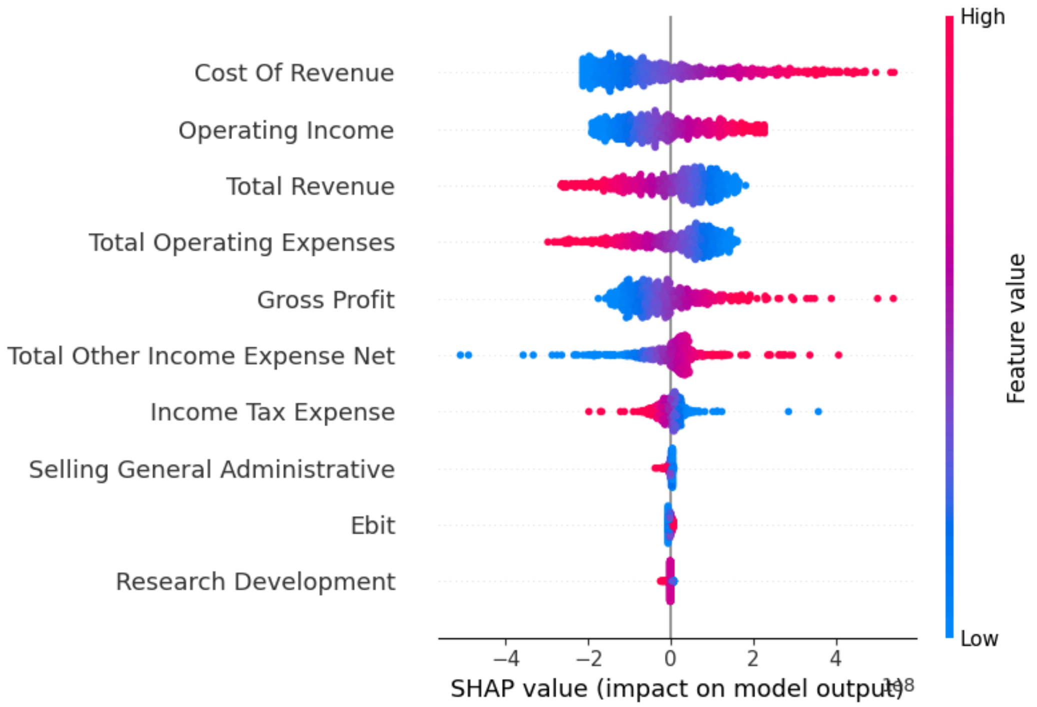


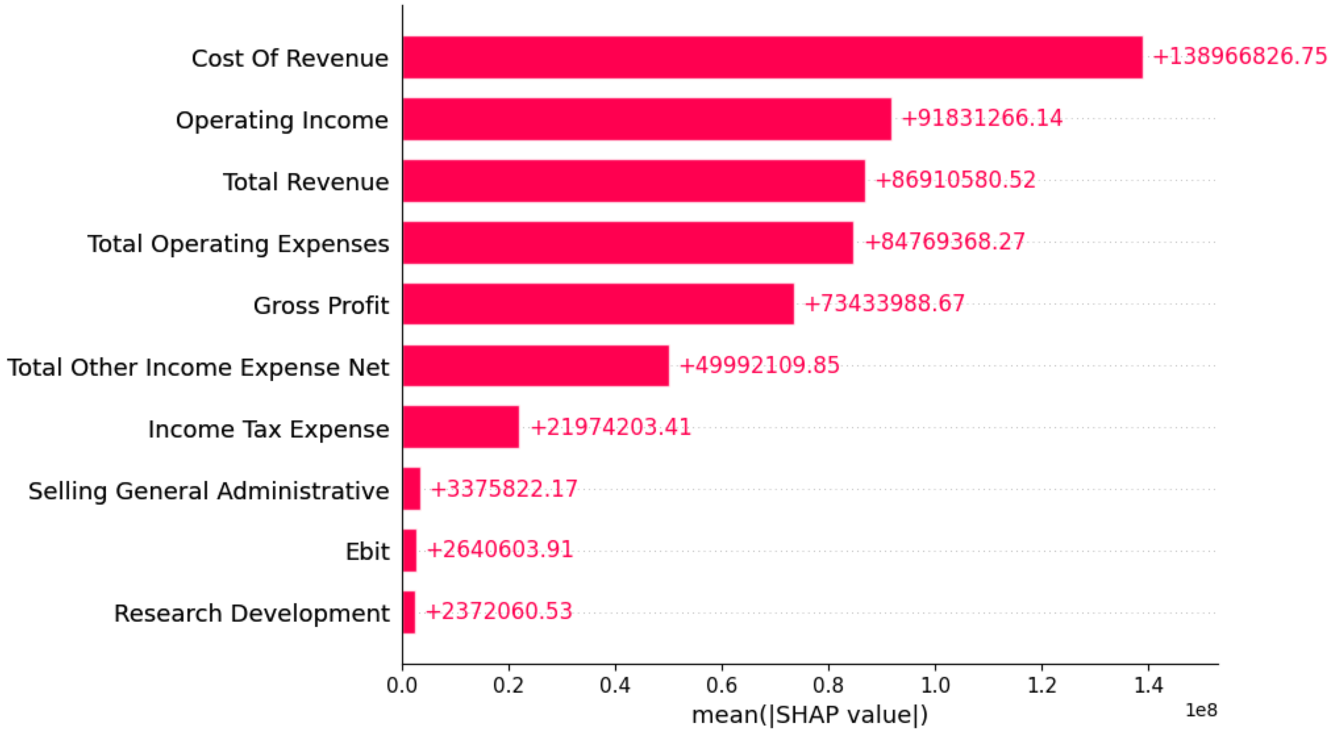


```
📈 Model Comparison Results:
```

	Model	R²	MAE	RMSE
0	Linear Regression	0.902071	1.422822e+07	3.667882e+07
1	Ridge Regression	0.869165	1.699837e+07	4.239577e+07

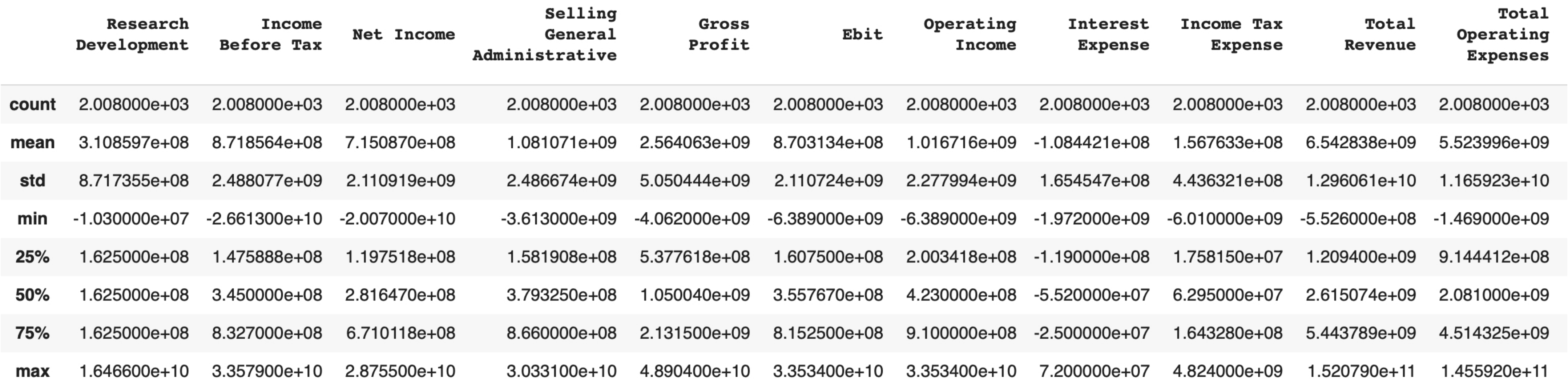
```
📊 Training and explaining: Lasso Regression
/usr/local/lib/python3.11/dist-packages/sklearn/linear_model/_coordinate_descent.py:695: ConvergenceWarning: Objective did not conve
  model = cd_fast.enet_coordinate_descent(
```



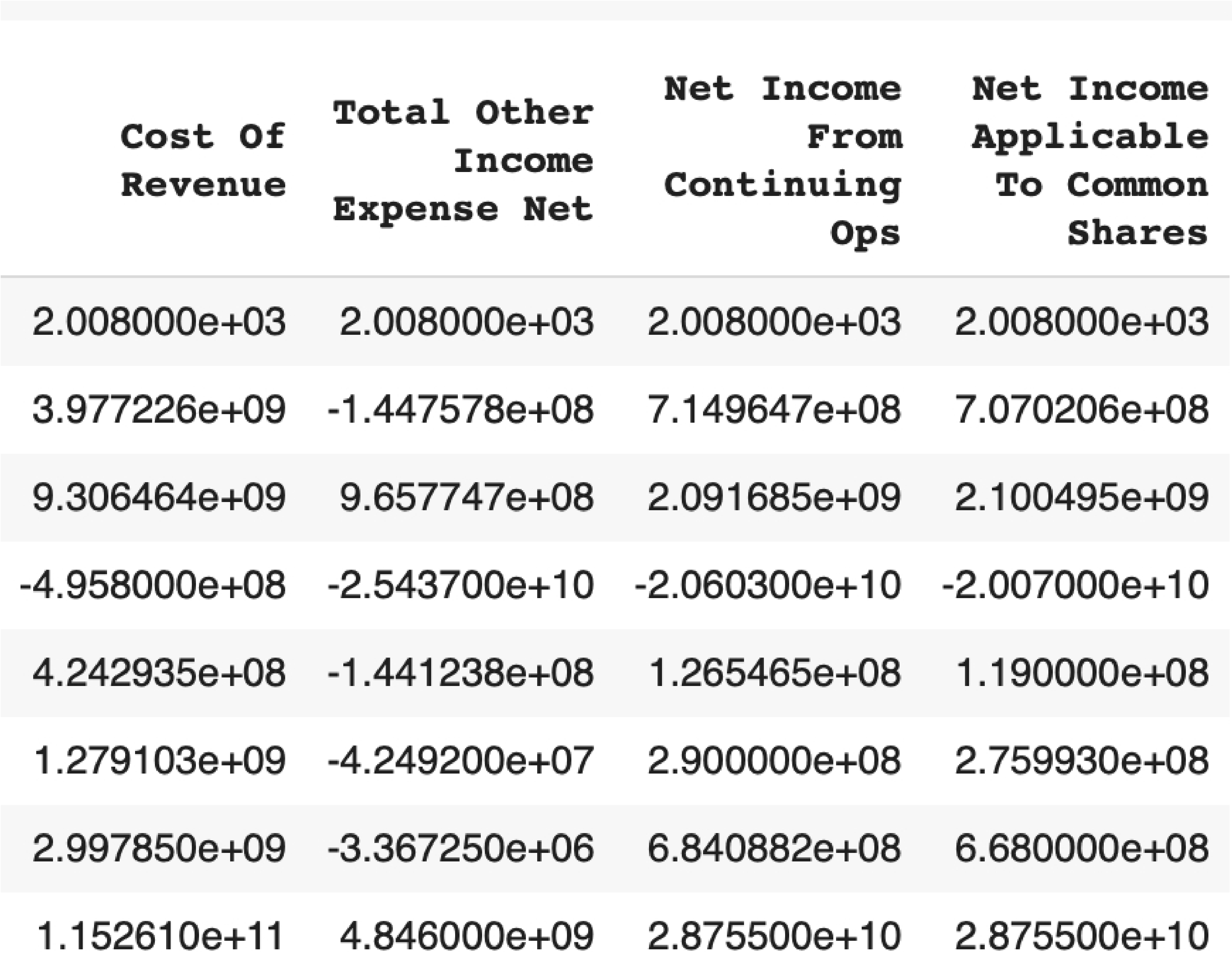


```
📈 Model Comparison Results:
```

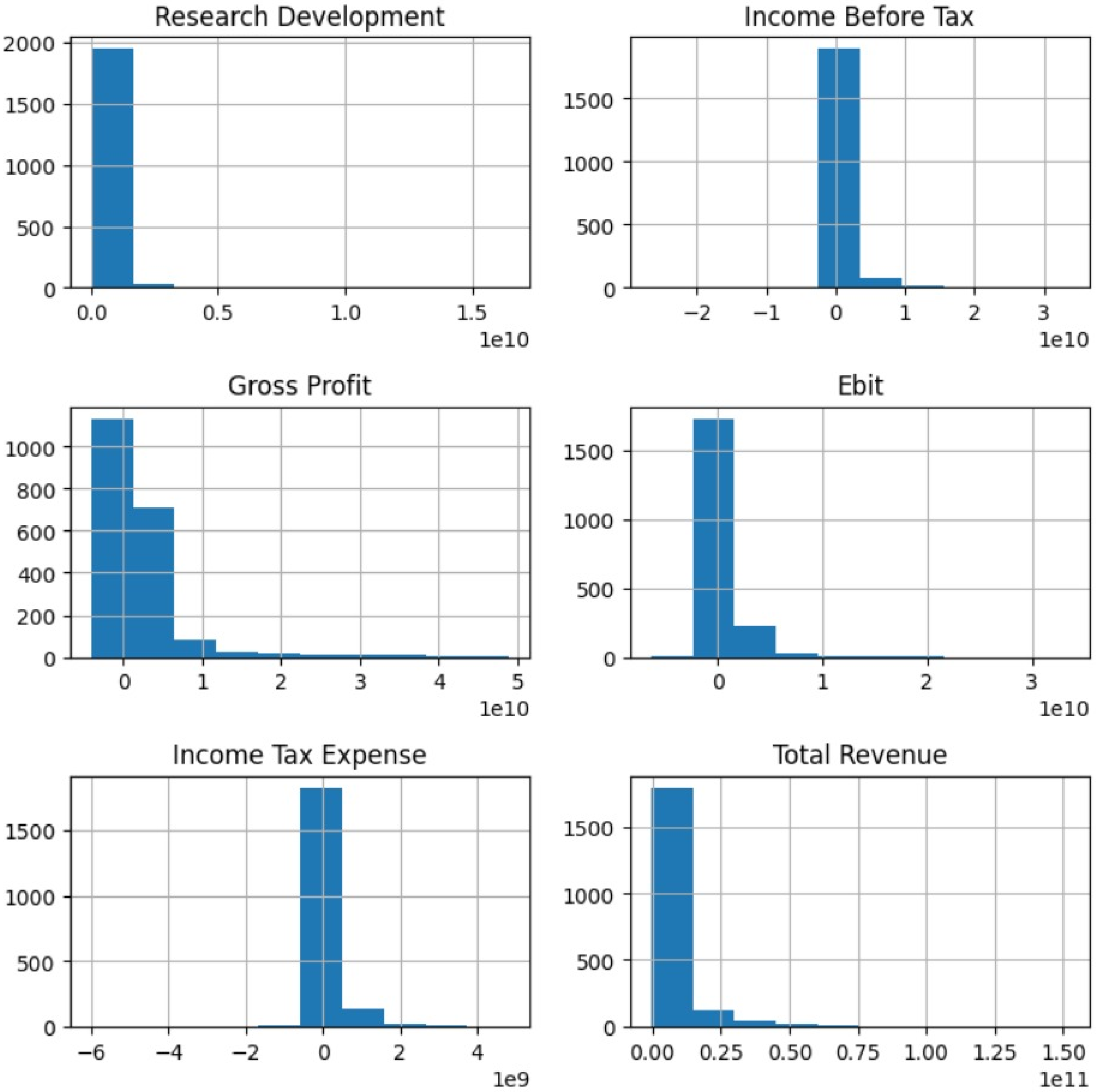
	Model	R^2	MAE	RMSE
0	Linear Regression	0.902071	1.422822e+07	3.667882e+07
2	Lasso Regression	0.874417	1.649845e+07	4.153620e+07
1	Ridge Regression	0.869165	1.699837e+07	4.239577e+07

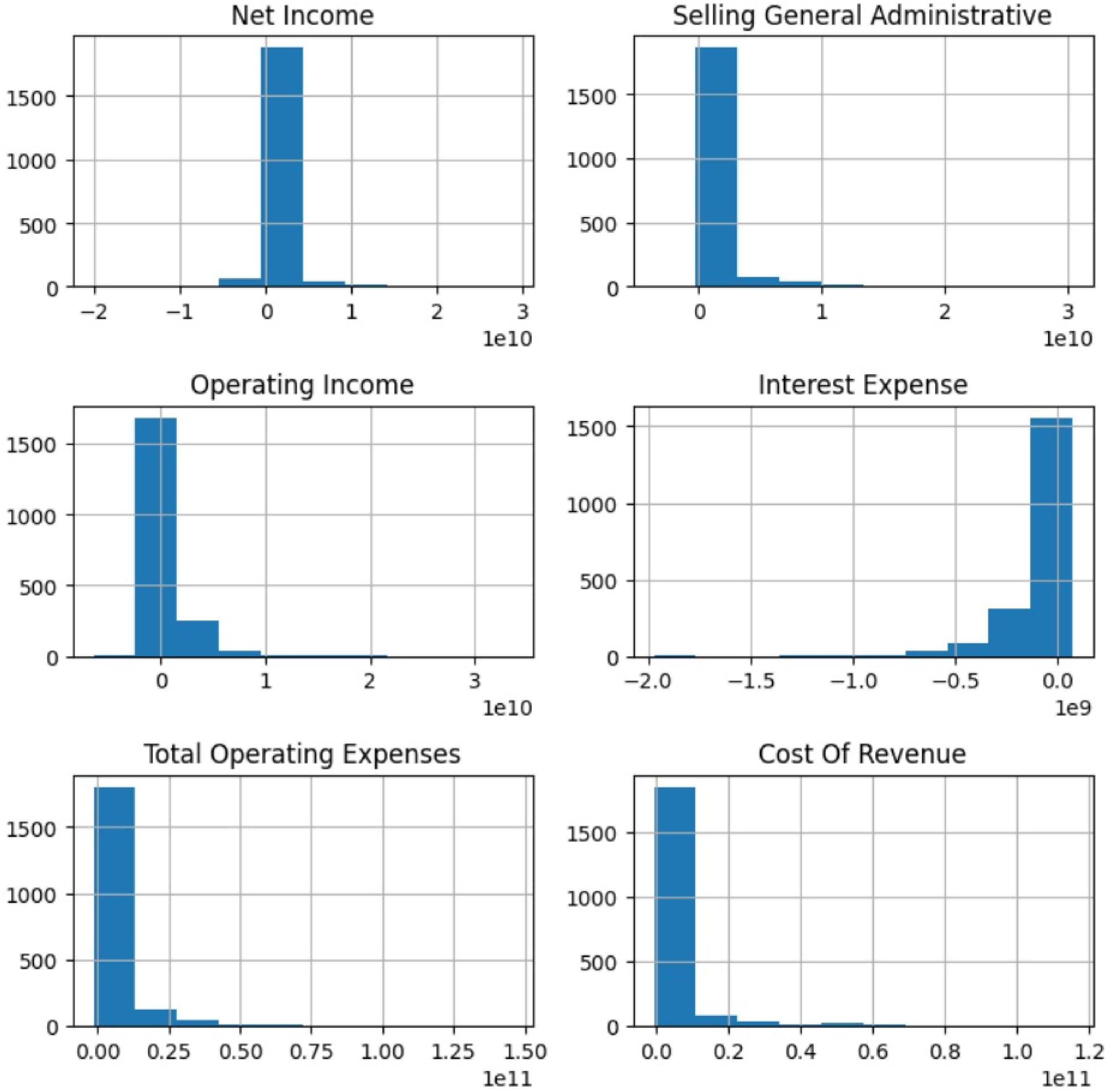


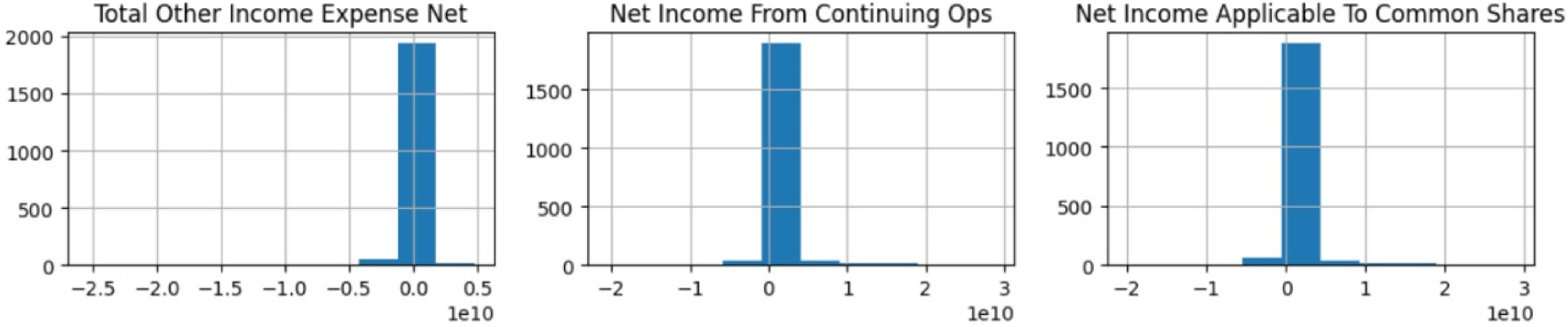
	Research Development	Income Before Tax	Net Income	Selling General Administrative	Gross Profit	Ebit	Operating Income	Interest Expense	Income Tax Expense	Total Revenue	Total Operating Expenses
count	2.008000e+03	2.008000e+03	2.008000e+03	2.008000e+03	2.008000e+03	2.008000e+03	2.008000e+03	2.008000e+03	2.008000e+03	2.008000e+03	2.008000e+03
mean	3.108597e+08	8.718564e+08	7.150870e+08	1.081071e+09	2.564063e+09	8.703134e+08	1.016716e+09	-1.084421e+08	1.567633e+08	6.542838e+09	5.523996e+09
std	8.717355e+08	2.488077e+09	2.110919e+09	2.486674e+09	5.050444e+09	2.110724e+09	2.277994e+09	1.654547e+08	4.436321e+08	1.296061e+10	1.165923e+10
min	-1.030000e+07	-2.661300e+10	-2.007000e+10	-3.613000e+09	-4.062000e+09	-6.389000e+09	-6.389000e+09	-1.972000e+09	-6.010000e+09	-5.526000e+08	-1.469000e+09
25%	1.625000e+08	1.475888e+08	1.197518e+08	1.581908e+08	5.377618e+08	1.607500e+08	2.003418e+08	-1.190000e+08	1.758150e+07	1.209400e+09	9.144412e+08
50%	1.625000e+08	3.450000e+08	2.816470e+08	3.793250e+08	1.050040e+09	3.557670e+08	4.230000e+08	-5.520000e+07	6.295000e+07	2.615074e+09	2.081000e+09
75%	1.625000e+08	8.327000e+08	6.710118e+08	8.660000e+08	2.131500e+09	8.152500e+08	9.100000e+08	-2.500000e+07	1.643280e+08	5.443789e+09	4.514325e+09
max	1.646600e+10	3.357900e+10	2.875500e+10	3.033100e+10	4.890400e+10	3.353400e+10	3.353400e+10	7.200000e+07	4.824000e+09	1.520790e+11	1.455920e+11



Cost Of Revenue	Total Other Income Expense Net	Net Income From Continuing Ops	Net Income Applicable To Common Shares
2.008000e+03	2.008000e+03	2.008000e+03	2.008000e+03
3.977226e+09	-1.447578e+08	7.149647e+08	7.070206e+08
9.306464e+09	9.657747e+08	2.091685e+09	2.100495e+09
-4.958000e+08	-2.543700e+10	-2.060300e+10	-2.007000e+10
4.242935e+08	-1.441238e+08	1.265465e+08	1.190000e+08
1.279103e+09	-4.249200e+07	2.900000e+08	2.759930e+08
2.997850e+09	-3.367250e+06	6.840882e+08	6.680000e+08
1.152610e+11	4.846000e+09	2.875500e+10	2.875500e+10


Research Development
2000
1500
1000
500
0
0.0
0.5
1.0
1.5
1e10
Income Before Tax
1500
1000
500
0
−2
−1
0
1
2
3
1e10
Gross Profit
1000
800
600
400
200
0
0
1
2
3
4
5
1e10
Ebit
1500
1000
500
0
0
1
2
3
1e10
Income Tax Expense
1500
1000
500
0
−6
−4
−2
0
2
4
1e9
Total Revenue
1500
1000
500
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1e11


Net Income
0
500
1000
1500
−2
−1
0
1
2
3
1e10
Selling General Administrative
0
500
1000
1500
0
1
2
3
1e10
Operating Income
0
500
1000
1500
0
1
2
3
1e10
Interest Expense
0
500
1000
1500
−2.0
−1.5
−1.0
−0.5
0.0
1e9
Total Operating Expenses
0
500
1000
1500
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1e11
Cost Of Revenue
0
500
1000
1500
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e11



Total Other Income Expense Net

Net Income From Continuing Ops

Net Income Applicable To Common Shares

HOCHSCHULE FRESENIUS
UNIVERSITY OF APPLIED SCIENCES

Affidavit

I hereby affirm that this submitted paper was authored unaided and solely by me. Additionally, no other sources than those in the reference list were used.

Parts of this paper, including tables and figures, that have been taken either verbatim or analogously from other works have in each case been properly cited with regard to their origin and authorship.

This paper either in parts or in its entirety, be it in the same or similar form, has not been submitted to any other examination board and has not been published.

Köln, 14.07.2025

Location, Date

I.N. Wob

Signature