

# Finance Project

Submitted to:-  
Dr. Nidhi Tanwar

## Cash flow forecasting and Cost estimation for Risk Analysis by using Machine Learning



Submitted by:-

Abhipsit Bajpai (23110011)

Rishit Bansal (23110023)

Pranay Pranshu (23110003)

Hanan Ahmed (23110030)

Pavitr (23110007)

Adarsh Chaturvedi (23110024)

## ABSTRACT

Cash flow forecasting and cost estimation are crucial elements of project risk analysis, directly influencing an organization's ability to manage financial stability and allocate resources effectively. Traditionally, these processes have relied on historical data and expert judgment, offering insights that may be limited by static assumptions and insufficient adaptability to changing economic landscapes. As project environments become more dynamic and market conditions increasingly volatile, traditional methods often fall short in accurately predicting the financial viability of complex projects. This has led to a growing need for more sophisticated tools and techniques to manage financial risks and ensure precise forecasting.

The rise of Machine Learning (ML) has introduced transformative capabilities in both cash flow forecasting and cost estimation. ML models, through their ability to learn from large datasets, are particularly adept at uncovering hidden patterns and relationships between numerous variables. Unlike traditional methods, ML algorithms such as regression models, decision trees, and neural networks adapt continuously to new data, offering forecasts that can reflect current and anticipated market dynamics. These models are capable of incorporating a broader range of factors, including project timelines, economic trends, and fluctuating market conditions, resulting in more accurate and flexible financial predictions. This improvement in accuracy not only enhances financial planning but also optimizes resource allocation, leading to more efficient project execution.

In addition to improving the precision of cash flow forecasts and cost estimates, ML also plays a pivotal role in project risk analysis. By analyzing historical project data, ML models can assess the likelihood of cost overruns, delays, and other potential disruptions. This provides project managers with a dynamic risk assessment framework that evolves with the project's progress and external market changes. The use of supervised learning algorithms allows for the continuous refinement of risk predictions, making it possible to implement proactive risk mitigation strategies. This significantly reduces the level of uncertainty in project outcomes, giving businesses a powerful tool for managing risk in a more data-driven, responsive manner.

The integration of ML into cash flow forecasting and cost estimation thus presents a substantial competitive advantage for businesses. With more reliable financial predictions and enhanced risk management strategies, organizations are better equipped to make informed decisions that minimize financial risks and improve overall project outcomes. This paper explores the application of ML techniques in detail, highlighting how they reduce uncertainty, optimize decision-making, and ultimately lead to improved project performance and profitability. By leveraging ML, businesses can move beyond static, traditional models and embrace a more dynamic approach to financial management that aligns with today's rapidly changing economic landscape.

## INTRODUCTION

In the rapidly evolving domain of financial forecasting, machine learning models have gained prominence as powerful tools for predicting critical financial metrics such as cash flows and project costs. Accurate predictions of these metrics are essential for informed capital budgeting, robust risk assessment, and effective strategic planning. While traditional forecasting methods offer reliability, they often rely heavily on structured data and conventional statistical techniques, limiting their ability to capture the complex, non-linear relationships frequently observed between financial variables. This gap becomes particularly evident when analyzing volatile financial environments or multifaceted projects.

Recent advancements in machine learning have introduced more dynamic and flexible approaches to financial forecasting. Algorithms such as Random Forest Regressor, Support Vector Regressor (SVR), and XGBoost have proven adept at managing high-dimensional datasets and capturing intricate relationships between financial features. Despite their strengths, however, challenges remain in optimizing model performance. Specifically, selecting the most relevant features and ensuring that models generalize well to unseen data without overfitting are critical hurdles in developing robust financial models. These challenges necessitate a careful balance between complexity and generalization to achieve accurate and reliable predictions.

This study aims to enhance the accuracy of cash flow predictions by utilizing advanced feature selection techniques such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) to reduce dimensionality and remove redundant data. By focusing on structured financial data, the research evaluates the performance of several machine learning models, with an emphasis on high-performing algorithms like XGBoost. In addition, this study explores various data preprocessing techniques and hyperparameter tuning to further optimize model performance, ultimately seeking to develop a model that is both reliable and interpretable in financial contexts.

By integrating advanced feature selection methods and fine-tuning machine learning algorithms, this study contributes to the expanding application of machine learning in finance. It provides valuable insights into the optimization of structured financial data for enhanced predictive performance, addressing critical issues such as model interpretability and the risks of overfitting, thus advancing the field of financial forecasting.

## **Literature Review**

### **1. The relation between cash flows and economic performance in the digital age : An empirical analysis**

Financial statements, including balance sheets, income statements, and cash flow statements, are essential for evaluating a company's financial health, risks, and future prospects. While balance sheets provide a snapshot of a company's assets and liabilities, and income statements reflect flow quantities based on economic activities, cash flow statements reveal the sources and uses of cash across different management areas. The adage "cash flow doesn't lie" highlights its importance, as cash flows are generally less susceptible to manipulation than other financial metrics.

In financial theory, cash flows are likened to the lifeblood of a company, vital for sustaining operations and enabling informed decision-making. According to the International Accounting Standards Board (IASB), financial statements should help users assess a firm's ability to generate sustainable future cash flows. This study investigates the extent to which cash flows can predict future economic performance, addressing the mixed findings in prior research regarding the predictive superiority of earnings versus cash flows.

The literature review reveals conflicting views on whether accrual-based earnings or operating cash flows serve as better predictors of future performance. The Financial Accounting Standards Board (FASB) posits that earnings are superior predictors due to their ability to reflect delayed effects through accrual accounting. However, numerous studies argue in favor of cash flows, citing their objectivity and reduced susceptibility to accounting manipulation.

Empirical research has produced varied results. Some studies uphold the FASB's assertion that earnings are superior predictors (e.g., Greenberg et al., 1986; Dechow, 1994), while others advocate for the predictive capabilities of cash flows (e.g., Farshadfar et al., 2008; Al-Debi'e, 2011). This study contributes to the ongoing debate by focusing on the predictive ability of cash flow components with respect to ROA in Italian companies.

The primary hypothesis posits that ROA in year  $N$  is influenced by cash flow components from year  $N-1$ . This hypothesis is based on the premise that cash flow generation is positively correlated with future profitability.

The analysis indicates that global net cash flow is positively correlated with ROA, confirming its significance as a predictor of future profitability. In contrast, cash flow for investments and corporate dynamics shows negative correlations with ROA. Notably, cash flow for extraordinary components does not exhibit a significant correlation with ROA. The regression analysis reinforces these findings, emphasizing the positive impact of global net cash flow on future operating income.

## 2. A study on cash flow analysis

Cash flow analysis is a fundamental tool used to assess the financial health of a company. It involves tracking the movement of cash into and out of the business over a specific period, helping determine a company's ability to meet its short-term and long-term obligations. Cash flow statements typically categorize cash flows into three sections: operating, investing, and financing activities. This structured approach ensures that businesses have an accurate picture of their liquidity, solvency, and overall financial health.

### Methods for Cash Flow Analysis

- **Direct Method:** The direct method involves tracking actual cash transactions related to operations, such as cash received from customers and payments made to suppliers. It offers a straightforward and clear picture of cash inflows and outflows. This method is particularly useful for businesses that want to understand their day-to-day cash operations.

$$\text{Cash Receipts} = \text{Cash Sales} + \text{Cash Collected from Credit Sales}$$
$$\text{Cash Payments} = \text{Payments to Suppliers} + \text{Payments to Employees} + \text{Operating Expenses}$$

- **Indirect Method:** The indirect method begins with net income and adjusts for non-cash transactions (e.g., depreciation) and changes in working capital to convert accrual-based figures into cash-based figures. This method is commonly used because it links the cash flow statement to both the income statement and balance sheet.

$$\text{OCF} = \text{Net Income} + \text{Non-Cash Expenses} \pm \text{Changes in Working Capital}$$

### Importance of Cash Flow Management

- **Liquidity**  
Proper cash flow management ensures that a company has sufficient liquidity to meet its short-term obligations, such as paying suppliers, wages, and debts. A company with poor liquidity can face serious financial difficulties even if it is profitable on paper.
- **Solvency**  
Solvency refers to a company's ability to meet its long-term obligations and sustain operations. Consistent positive cash flow from operating activities is a key indicator of a company's solvency.
- **Decision-Making**  
Cash flow analysis plays a critical role in decision-making. It helps management assess whether the company has enough cash to invest in future projects, pay off debts, or return capital to shareholders through dividends.

### 3. A nonlinear dynamic approach to cash flow forecasting

Cash flow forecasting is vital for the survival and valuation of corporations. Accurate predictions enable firms to make informed investment decisions and ensure transparency in earnings. Traditional forecasting models often fail to account for the dynamic nature of cash flows, leading to a need for more sophisticated approaches that reflect the complexities of financial operations.

Historically, cash flow estimation has been indirect, resulting in significant measurement errors. Early forecasting models primarily used lagged earnings and cash flows as predictors. The Dechow et al. (DKW) model introduced a framework that incorporates accruals alongside earnings and cash flows, emphasizing that univariate time-series models are insufficient for capturing the complexities of cash flows.

The Barth et al. (BCN) model further decomposed earnings into cash flows and accruals, employing pooled regression estimation. However, these traditional models often assume homogeneity across firms, which can fail to account for the diversity of business operations in real-world scenarios.

The proposed grey-box model effectively merges elements from both white-box (linear) and black-box (nonlinear) models. It retains a simple structure while allowing parameters to vary dynamically, using a Padé approximant function to capture the complexities of cash flow dynamics. This model incorporates firm-specific characteristics, such as sales growth rates and firm age, as exogenous variables.

The study employs panel data methods to handle heterogeneity across firms. This approach combines time-series and cross-sectional analyses, providing a more nuanced understanding than traditional pooled regression. Key methods include:

- **Demean Method:** This technique removes individual effects by subtracting group means from each variable, which clarifies the relationships being studied.
- **First Difference Method:** This method eliminates intercept differences by differencing variables over time, thereby enhancing accuracy in estimation.

Additionally, the Arellano-Bond estimator is utilized to address potential endogeneity issues in the dataset.

The grey-box model incorporates time-varying parameters modeled as functions of sales growth rate or firm age. These parameters are estimated using a Padé approximant, allowing for nonlinear relationships without imposing strict linearity.

The grey-box model is tested using data from U.S. firms spanning from 1957 to 2013. The empirical analysis yields several insights:

- Cash flow patterns vary significantly among firms, highlighting the need for tailored forecasting approaches.
- Survivorship bias affects observed cash flow trends; adjustments for this bias suggest lower growth expectations than previously believed.
- The dynamic parameters of the grey-box model demonstrate superior predictive performance compared to static models.

#### **4. Cash Flow management and its effect on firm performance: Empirical evidence on non-financial firms of China**

The study explores the impact of changes in cash flow measures and metrics on firm performance, using generalized estimating equations (GEEs) to analyze a panel dataset of Chinese firms. This method is robust for analyzing data with repeated measurements, allowing for more accurate insights into how cash flow management affects financial performance over time.

##### **Cash Flow Measures and Metrics**

Cash flow management is primarily concerned with the effective control of three key components: accounts receivables, inventories, and accounts payables. These components determine the cash flow measures, which are quantified through several key metrics:

- **Accounts Receivables Turning Days (ARTD):**  
This measures the average number of days it takes a firm to collect payments after a sale.

$$\text{ARTD} = (\text{Accounts Receivables} / \text{Sales}) * 365$$

- **Inventory Turning Days (ITD):**  
ITD assesses how efficiently a company manages its inventory by measuring how long it takes to sell and replace it.

$$\text{ITD} = (\text{Inventory} / \text{Cost of Goods Sold}) * 365$$

Efficient inventory management is crucial for maintaining liquidity, as holding excess inventory ties up cash.

- **Accounts Payables Turning Days (APTD):**  
This metric measures the average time a firm takes to pay its suppliers.

$$\text{APTD} = (\text{Accounts Payable} / \text{Purchases}) * 365$$

Longer payment terms with suppliers can improve a firm's cash position, but excessively delayed payments may strain supplier relationships.

These metrics are components of two broader financial management tools:

**Cash Conversion Cycle (CCC) and the Operating Cash Cycle (OCC).**

- **Cash Conversion Cycle (CCC):**

The CCC is a comprehensive measure of a firm's liquidity, indicating how long it takes to convert resource investments into cash inflows from sales.

$$CCC = ARTD + ITD - APTD$$

- **Operating Cash Cycle (OCC):**

The OCC reflects the time during which cash is tied up in the operating cycle, from the purchase of inventory to the collection of receivables.

$$OCC = ARTD + ITD$$

The study also examined the influence of leverage on the relationship between cash flow management and performance. Firms with lower leverage benefit more from improvements in cash flow measures and metrics, as they are less burdened by fixed interest payments and financial distress. Low-leverage firms, by reducing CCC and managing their working capital efficiently, can reinvest earnings and minimize the need for external financing, enhancing profitability. In contrast, high-leverage firms face greater financial constraints, which can limit their ability to optimize cash flow metrics. These firms are more likely to experience liquidity issues, as they must prioritize debt repayments, reducing the flexibility to invest in operations or take advantage of early payment discounts from suppliers.

## **5. Estimating project cash flows**

Estimating project cash flows is a critical process in financial management, particularly in capital budgeting and project evaluation. It involves predicting both the inflows and outflows of a project to determine its viability and overall impact on a company's financial health. This comprehensive overview will discuss the core components of cash flow estimation, including mathematical approaches, and the considerations involved in making accurate assessments.

### **Key Components of Cash Flow Estimation**

- **Initial Investment**

The initial investment encompasses all upfront costs required to start a project. These include the costs of acquiring fixed assets, such as machinery and equipment, as well as any installation expenses. This is the foundation upon which the project's operations are built.

$$\text{Initial Investment} = \text{Cost of Fixed Assets} + \text{Installation Expenses}$$

Accurate estimation of the initial investment is essential because it represents a major cash outflow that needs to be justified by the project's future cash inflows.



- **Working Capital**

Working capital represents the funds needed to manage the project's daily operational activities. It includes investments in current assets like inventory and accounts receivable, minus current liabilities. Proper working capital management ensures that the project maintains adequate liquidity to continue operating smoothly.

$$\text{Working Capital} = \text{Current Assets} - \text{Current Liabilities}$$

Changes in working capital affect cash flow because increasing current assets (like inventory) or reducing current liabilities can tie up cash, reducing the funds available for other uses.

- **Cash Inflows from Operations**

Cash inflows from operations are the revenues generated from the core business activities of the project. This includes income from sales, minus the operating expenses needed to sustain these sales. Predicting these inflows accurately is essential for determining whether the project will be profitable in the long run.

$$\text{Cash Inflow} = \text{Sales Revenue} - \text{Operating Expenses}$$

This figure should be estimated based on realistic sales forecasts, market conditions, and pricing strategies.

- **Terminal Cash Flow**

Terminal cash flow, also known as salvage or residual value, is the cash inflow received at the end of the project's life. This includes proceeds from selling off project assets or recovering any unused working capital. The terminal cash flow can have a significant impact on the overall cash flow, particularly for long-term projects.

$$\text{Terminal Cash Flow} = \text{Scrap Value} + \text{Recovery of Working Capital}$$

## **6. Cash Flow Prediction: MLP and LSTM compared to ARIMA and Prophet**

This research paper focuses on predicting accounts receivable cash flows, particularly for companies with numerous transactions and customers, such as e-commerce and transportation firms. Accurate cash flow predictions are essential for increasing returns, optimizing capital allocation, and preventing financial distress. The authors compare traditional forecasting methods like ARIMA and Prophet with more advanced techniques, such as neural networks, including multi-layer perceptrons (MLP) and Long Short-Term Memory (LSTM) networks. While LSTM is commonly used for time series forecasting, this paper highlights its novel application to cash flow prediction. The study finds that as the methods become more sophisticated, their flexibility and accuracy in forecasting improve. Additionally, the paper introduces a new performance metric called Interest Opportunity Cost (IOC), which considers interest rates and capital costs, providing a financially beneficial

approach to cash flow forecasting. This comprehensive exploration of various techniques offers practical insights for businesses managing large-scale financial operations.

The paper compares the performance of ARIMA, Prophet, MLP, and LSTM using both MSE and IOC metrics. The results showed that:

- **LSTM optimized with IOC** outperformed all other models in terms of financial impact, as it produced fewer costly errors (overdrafts or excessive surplus cash).
- **MLP optimized with IOC** also performed well but was slightly less accurate than LSTM.
- **Prophet** performed better than ARIMA in both MSE and IOC metrics, especially in handling holiday-related variations.
- **ARIMA** was the least effective due to its inability to incorporate external variables like holidays.

The paper demonstrates that deep learning methods, particularly LSTM, are superior to traditional time series models for cash flow prediction in complex business environments. The introduction of IOC as a performance measure further improves the practical value of these models by focusing on minimizing the financial impact of prediction errors. This allows finance managers to choose models that align with their specific business goals, such as optimizing working capital or minimizing interest costs.

The paper introduces **Interest Opportunity Cost (IOC)** as a financial evaluation metric for cash flow predictions. Traditional metrics like **Mean Squared Error (MSE)** do not consider the financial implications of prediction errors. In contrast, IOC evaluates the cost of over- or under-predicting cash flows.

## **7. Financial Risk Prediction and Management using Machine Learning and Natural Language Processing**

"Financial Risk Prediction and Management using Machine Learning and Natural Language Processing" explores the integration of machine learning (ML) and Natural Language Processing (NLP) in predicting and managing financial risk. The research highlights how traditional financial risk management methods, which primarily rely on historical data and financial statement analysis, are limited in handling large-scale unstructured data like social media content or news articles. The authors propose a novel method that combines NLP and a deep learning model (DeepFM) to improve the accuracy and efficiency of financial risk predictions.

In the modern financial landscape, the ability to predict and manage financial risk is essential for corporate stability. Traditionally, financial risk management relied on historical data and financial statements to forecast potential risks. However, these methods struggle with unstructured data, such as news reports and social media content, which can provide valuable insights into emerging risks. To address this, the study focuses on utilizing ML and NLP technologies, which have shown great potential in predicting risks by analyzing vast amounts of structured and unstructured data in real time.

The first part of the paper focuses on measuring financial risk tendencies using NLP. This involves analyzing unstructured textual data, including news articles, financial reports, and social media posts, to assess risk. NLP technologies can extract sentiment and key financial insights from these texts, allowing for a deeper understanding of risk dynamics.

The first part of the paper focuses on measuring financial risk tendencies using NLP. This involves analyzing unstructured textual data, including news articles, financial reports, and social media posts, to assess risk. NLP technologies can extract sentiment and key financial insights from these texts, allowing for a deeper understanding of risk dynamics.

The financial risk tendency of a document is predicted by determining whether the text suggests an increasing or decreasing risk. Using binary classification, the texts are categorized into positive or negative financial sentiment. The output is represented in two-bit one-hot encoding: [0, 1] for rising financial risk (negative sentiment) and [1, 0] for falling financial risk (positive sentiment). This analysis allows for the creation of financial risk propensity labels for each document, which are used to train the prediction model.

The study highlights the importance of combining machine learning techniques with NLP for financial risk prediction. The DeepFM model proves to be a robust tool for capturing both linear and non-linear interactions in complex financial data. By incorporating structured and unstructured data, the model offers a comprehensive approach to risk prediction, surpassing traditional models in accuracy and efficiency. Future work may focus on expanding the model's applicability to different types of financial risks and improving its ability to handle even more complex, large-scale datasets.

## **8. The analysis of financial market risk based on machine learning and particle swarm Optimization algorithm**

The research paper explores the critically important role of the financial industry in sparking national economic development against the background of its intrinsic risks. In order to mitigate those risks, the study discusses the potential of BT in optimizing the functions of the financial market. The authors applied clustering method with optimized principle for forming a decision tree using financial market risks modeled via ML and RF models. At the same time, the regional economy's industrial structure is transforming from primary to tertiary industry because of BT, acting as a change factor in improvement of financial efficiency, cost reduction, and avoidance of intermediaries. BT is also deemed to be a mechanism that fosters business development while further increasing mutual interplay between companies along a supply chain. The study therefore gives a theoretical grounding that allows BT to be applied in encouraging collaboration and innovation in financial services.

### **Feature Selection Techniques**

Three primary feature selection methods are used:

- **Filtering Method:** Removes noisy and irrelevant features before model training.
- **Wrapping Method:** Iteratively trains the model and evaluates subsets of features, selecting the best-performing subset. This method is more accurate but computationally intensive.
- **Embedding Method:** Integrates feature selection directly into the model training process, optimizing both the feature selection and learning process simultaneously.

## Results and Analysis

- **Financial Risk Analysis Based on Maximum Likelihood and Blockchain Technology:** The study evaluates regional financial performance by analyzing gross domestic product (GDP) and growth rates of primary, secondary, and tertiary industries. It finds that blockchain technology can significantly improve financial risk detection and control.
- **Financial Risk Analysis Based on Blockchain and PSO:** The model developed using PSO and blockchain technology shows promising results in optimizing regional financial development indicators. The study analyzes key factors such as GDP growth rates, asset investments, and fiscal regulations, highlighting the role of PSO improving financial decision-making processes.

The integration of ML, DL, PSO, and blockchain technology provides a robust framework for addressing financial market risks. While ML and DL techniques facilitate accurate classification and clustering of financial data, PSO optimizes the overall risk management process by finding optimal solutions in complex, multi-dimensional financial environments. Blockchain technology further enhances these models by providing secure and transparent data handling. This methodology provides a solid theoretical and practical foundation for future research on financial risk mitigation using advanced computational techniques.

## 9. Machine learning for financial forecasting, planning And analysis: recent developments and pitfalls

Financial planning and analysis (FP&A) is central to modern corporations, helping management make informed decisions regarding resource allocation, investment strategies, and risk management. It plays a crucial role in ensuring that a company can reach its financial objectives, balancing short-term and long-term goals such as profitability, liquidity, and growth. Accurate financial forecasts are vital, particularly in volatile markets. FP&A departments gather data from various sources—internal systems like accounting and HR and external sources like macroeconomic indicators—to support their forecasts and plans. The increasing digitization and availability of big data have created opportunities to improve the accuracy, speed, and insights generated from these forecasts. Machine learning (ML) techniques are especially suited for this transformation, providing predictive power that can streamline processes and deliver high-quality forecasts in complex, data-rich environments.

One of the major challenges discussed in the document is the "pitfall" of confusing forecasting and planning. Traditional machine learning models excel in predicting future outcomes based on past data (forecasting), but they do not inherently account for causal

relationships, which are necessary for planning and intervention tasks. Forecasting models focus on finding correlations, while planning involves causal inference, where an understanding of the underlying mechanisms is required to guide decision-making. For example, a sales forecast may indicate that an increase in price is correlated with higher sales, but a causal analysis may reveal that price increases were only a response to higher demand, and raising prices further would reduce sales. To address this issue, the authors introduce **causal machine learning** methods, particularly **double machine learning**, which mitigates model specification errors and enables more reliable causal inference.

The document includes a simulation study comparing machine learning techniques like lasso regression with traditional methods like ordinary least squares (OLS). The aim was to evaluate their performance in both forecasting and planning tasks. The study uses a simulated dataset representing five years of monthly sales data (60 observations) with 40 potential predictive factors and one intervention variable (promotional activity). Key findings include:

- **Forecasting:**
  - **Post-lasso regression** outperformed OLS in out-of-sample forecasting, with a lower root mean squared error (RMSE). This reflects the ability of lasso to handle high-dimensional data by selecting only the most relevant features.
  - OLS tended to overfit the data, resulting in poor performance when applied to unseen data.
- **Planning (Causal Inference):**
  - The naive approach of using lasso for causal inference yielded biased results, overestimating the impact of promotional activities.
  - The **double machine learning approach** provided unbiased estimates, accurately determining that the promotional activities had no real effect on sales in the simulation. This method corrected for the confounding effects of other variables, which the naive approach failed to do.

## 10. Predicting cash holdings using supervised machine learning algorithms

The document titled "Predicting Cash Holdings Using Supervised Machine Learning Algorithms" by Şirin Özlem and Omer Faruk Tan delves into the application of machine learning techniques for predicting cash holdings, a critical element of corporate finance. Focusing on firms listed on the Borsa Istanbul (BIST) between 2006 and 2019, the study explores how different supervised learning regression methods can forecast corporate cash holdings more effectively than traditional models. The use of machine learning represents a novel approach to this area of financial forecasting, which historically relied on linear models that often struggled to capture the complex dynamics involved in cash management.

The document titled "Predicting Cash Holdings Using Supervised Machine Learning Algorithms" by Şirin Özlem and Omer Faruk Tan delves into the application of machine learning techniques for predicting cash holdings, a critical element of corporate finance. Focusing on firms listed on the Borsa Istanbul (BIST) between 2006 and 2019, the study explores how different supervised learning regression methods can forecast corporate cash holdings more effectively than traditional models. The use of machine learning represents a

novel approach to this area of financial forecasting, which historically relied on linear models that often struggled to capture the complex dynamics involved in cash management.

An analysis of feature importance revealed that cash flow, current ratio, pretax margin, and net margin were the most significant predictors of cash holdings. These features were dominant across all models, particularly in tree-based algorithms like XGBoost, which effectively captured the impact of these variables on corporate cash reserves.

This study demonstrates that machine learning algorithms, particularly ensemble models like XGBoost, outperform traditional regression methods in predicting corporate cash holdings. The ability of these models to capture non-linear relationships and correct errors across iterations makes them highly suitable for financial forecasting.

For corporate managers, the application of machine learning models presents a powerful tool for optimizing cash management. The findings suggest that firms can benefit from more accurate cash holding predictions, leading to better decision-making in investment, liquidity management, and strategic planning. Future research could expand on these findings by incorporating macroeconomic factors and cross-country analyses to better understand cash holdings behavior in a global context.

## **11. Machine Learning Algorithms for Free Cash Flows Growth Rate Estimation**

In the realm of financial forecasting, traditional models such as ARIMA and linear regression have long been used to predict key metrics like cash flow growth rates. These models, while effective for certain applications, have limitations when applied to datasets that exhibit non-linear relationships or structural changes over time. As highlighted by **Brooks and Buckmaster (1979)**, traditional models require that financial data adhere to strict assumptions, such as stationarity and homoscedasticity, which are frequently violated in practice. Financial time-series data, especially fundamental indicators like Free Cash Flows (FCF), often display non-linearities that challenge these models.

The shift toward machine learning (ML) in financial forecasting has addressed some of these limitations. ML algorithms, unlike traditional statistical methods, can capture non-linear relationships and identify complex patterns within large datasets. **Goodfellow et al. (2016)** emphasize that machine learning models like Random Forests and Support Vector Machines (SVMs) can bypass assumptions of linearity and stationarity, making them more suitable for financial data with irregularities. Furthermore, ML models have the ability to handle high-dimensional datasets, enabling more accurate predictions even in the presence of a large number of explanatory variables.

Several studies have demonstrated the superiority of machine learning models over traditional methods in financial forecasting tasks. **Zhu et al. (2022)**, for instance, employed a Backpropagation Neural Network optimized with a Genetic Algorithm to predict Free Cash Flow and found that their model outperformed ARIMA in accuracy. Similarly, **Kumbure et al. (2022)** reviewed a range of ML techniques and concluded that ensemble models like

Random Forest and XGBoost consistently achieved higher predictive accuracy in stock market forecasting tasks compared to classical approaches.

An important consideration in the application of ML models to financial forecasting is the issue of overfitting. Overfitting occurs when a model is too complex and captures noise in the training data, leading to poor generalization on unseen data. **Bishop and Nasrabadi (2006)** argue that regularization techniques and cross-validation are essential to mitigating overfitting in ML models. In the context of financial forecasting, where data points are often limited due to quarterly reporting, avoiding overfitting is crucial to ensure reliable predictions. To address this, techniques like Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) have been widely used to reduce the dimensionality of financial data, thereby improving model generalization.

The application of machine learning in financial forecasting also extends to feature selection. **Guyon and Elisseeff (2003)** emphasize the importance of selecting the most relevant features from high-dimensional datasets to improve the accuracy and interpretability of ML models. Feature selection is particularly challenging in financial datasets, where many variables may be collinear or redundant. By incorporating techniques like PCA, researchers can reduce the dimensionality of the data, focusing the model's attention on the most informative variables. This not only enhances prediction accuracy but also makes the models more interpretable. Moreover, recent studies have focused on the comparison between different machine learning algorithms for specific financial tasks, such as predicting Free Cash Flow growth rates. **Vayas-Ortega et al. (2020)** compared various ML algorithms, including Bayesian Ridge Regression and K-Nearest Neighbors, for FCF forecasting. Their study found that K-Nearest Neighbors consistently produced the lowest error rates, especially when applied to smaller datasets, which is a common scenario in financial data due to quarterly reporting. **Evdokimov et al. (2023)** also benchmarked several ML algorithms against ARIMA and found that K-Nearest Neighbors and Bayesian Ridge Regression outperformed traditional models on Free Cash Flow data, highlighting their potential for more accurate financial forecasting in environments with limited data availability.

In conclusion, the literature suggests that machine learning techniques offer substantial improvements over traditional methods in the context of financial forecasting, particularly for non-linear and high-dimensional datasets. However, the challenges of overfitting and feature selection remain critical considerations. Future research should focus on refining these models, with an emphasis on balancing complexity and generalization to ensure robust financial predictions. As the adoption of ML in finance continues to grow, these models will likely become indispensable tools for both researchers and practitioners aiming to optimize investment decisions and manage financial risks effectively.

## **12. Forecasting project cashflow using Machine Learning**

In recent years, financial forecasting has experienced significant advancements, particularly through the integration of machine learning (ML) techniques. Traditional financial forecasting methods, grounded in statistical models and historical data, have often been limited by their linear approach, making them inadequate for addressing the complexities of financial markets. Early studies by **Makridakis et al. (1993)** and **Chatfield (2000)** highlighted the strengths of traditional statistical methods such as autoregressive integrated moving average (ARIMA) and linear regression but also pointed out their limitations when dealing with non-linear relationships and vast, high-dimensional datasets. These conventional approaches often failed to capture intricate market dynamics, leading to less accurate predictions, particularly in uncertain environments.

The introduction of machine learning models such as **Random Forests**, **Support Vector Machines (SVMs)**, and **Gradient Boosting algorithms** has transformed the field of financial forecasting. **Zhang et al. (2005)** demonstrated the superiority of machine learning models in handling high-dimensional data and uncovering non-linear relationships, which are common in financial datasets. These models, by learning from vast amounts of data, can predict outcomes based on complex patterns that traditional models often overlook. **Breiman (2001)**, in his seminal work on Random Forests, established that ensemble methods outperform single estimators by reducing variance and improving accuracy, a finding that has been widely applied in financial forecasting research.

The use of advanced algorithms, such as **XGBoost** and **Support Vector Regressor (SVR)**, has further refined forecasting techniques. **Chen and Guestrin (2016)** introduced XGBoost, a highly efficient implementation of gradient boosting, which has been proven effective in prediction tasks with large, sparse datasets, a common characteristic of financial data. Studies by **Tao et al. (2018)** and **Li et al. (2019)** found that XGBoost outperformed traditional machine learning models in predicting stock market trends and cash flow forecasts, emphasizing its capacity to handle complex, non-linear dependencies in financial variables. However, despite their predictive power, machine learning models are not without challenges. One of the significant concerns is the issue of overfitting, where models perform exceptionally well on training data but fail to generalize to new, unseen datasets. This problem was highlighted in **Goodfellow et al. (2016)**, who stressed the importance of regularization techniques and cross-validation in model training to mitigate overfitting. Moreover, the choice of relevant features in financial forecasting remains a critical challenge. **Guyon and Elisseeff (2003)** emphasized the need for robust feature selection techniques to improve model accuracy and reduce the dimensionality of data, allowing models to focus on the most informative variables.

In response to these challenges, dimensionality reduction techniques such as **Principal Component Analysis (PCA)** and **Singular Value Decomposition (SVD)** have been employed to enhance the efficiency of machine learning models. **Jolliffe (2002)** demonstrated the effectiveness of PCA in reducing data redundancy while preserving the essential patterns needed for accurate predictions. **Halko et al. (2011)** further explored SVD's role in improving computational efficiency in large datasets, making it an essential tool in financial forecasting where datasets are often high-dimensional.

The evolution of financial forecasting, driven by machine learning, also introduces new discussions about model interpretability. While models like Random Forests and XGBoost offer high accuracy, their "black-box" nature makes them less interpretable compared to



traditional statistical models. This concern was raised by **Lipton (2016)**, who argued for the development of more transparent ML models, especially in high-stakes fields like finance, where interpretability is crucial for decision-making. Efforts to create explainable machine learning models are ongoing, with techniques such as **SHAP (Lundberg and Lee, 2017)** being explored to provide clearer insights into model behavior and predictions.

In summary, the integration of machine learning into financial forecasting has opened new avenues for more accurate and dynamic predictions. However, challenges such as feature selection, overfitting, and model interpretability remain central to ongoing research. Future studies will likely focus on addressing these issues while further refining machine learning algorithms to make them more robust and transparent in financial contexts.

## **NEED OF THE STUDY**

The continuous development and changes in the global financial markets has led to an increase in importance of financial risk management for the stable operation of enterprises. Traditional financial risk management systems rely heavily on financial statement analysis and historical data statistics which shows clear limitations when dealing with large scale data.

The development in the machine learning and development of models like XGBRegressor, support vector regressor and neural learning models like multi-layered perceptrons and NLP in the recent years offers new perspective and methods for financial risk analysis. Machine learning models like multi layered perceptron can analyze vast data that are beyond the human capacity and give us a fair estimate.

These models can handle complex data containing all sorts of financial and other parameters like market trends, business cycles, and macroeconomic conditions. These models can also improve the accuracy of predictions by learning from historical data and continuously adjust their predictions based on new data. They are automated and are cost efficient which is the need of the future while continuously adopting to the new market conditions. Scalability is not an issue as these models can be scaled easily in a cost-efficient manner without sacrificing any sort of performance, making them suitable for every enterprise.

## **RESEARCH GAP**

Previous research in the field of cash flow prediction and cost estimation has shown significant promise through the application of machine learning models, yet several critical gaps remain. One prominent issue is the overwhelming focus on structured financial data, such as balance sheets, income statements, and cash flow statements, while largely neglecting the potential of unstructured data sources. Unstructured data, such as textual information from news articles, social media sentiment, earnings call transcripts, and other qualitative reports, contain a wealth of insights that are often overlooked. These sources can offer essential indicators of external risks, market sentiment, and broader economic factors that may impact a company's financial performance. The reliance on structured data alone limits the ability of models to fully capture these risks, potentially reducing the accuracy and robustness of predictions.

Another key gap lies in the lack of explainability and interpretability of the machine learning models used for financial forecasting. Advanced models, like XG Boost and other ensemble techniques, are highly accurate due to their ability to model complex, nonlinear relationships in high-dimensional financial data. However, these models are often treated as "black boxes" due to their opaque internal workings, making it difficult for users, including financial analysts and decision-makers, to understand how the predictions are made. This lack of transparency is especially problematic in the financial domain, where explainability is crucial for regulatory compliance and gaining stakeholder trust. Without clear explanations of which factors are driving cash flow predictions, it becomes challenging to assess the model's reliability and usefulness in real-world decision-making.

While machine learning models can process vast amounts of financial data and provide relatively accurate forecasts, the exclusion of unstructured data and the opacity of these models in prior research limit their applicability for comprehensive risk analysis. Therefore, addressing these gaps by incorporating unstructured data and enhancing model interpretability is essential for creating more robust and reliable financial forecasting tools.

## RESEARCH METHODOLOGY

The dataset used for predicting cashflows was obtained from the 2014 Financial Dataset on Kaggle, containing 225 financial attributes across multiple organizations. Given the complexity and high dimensionality of the data, feature reduction techniques like Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) were employed. These methods helped identify correlations between features, reducing redundancy and ensuring that only the most relevant and independent financial attributes were retained. Feature selection not only simplified the modeling process but also minimized the risk of overfitting due to an excessive number of variables.

The selected features included important financial metrics such as 'Cost of Revenue', 'R&D Expenses', 'SG&A Expense', 'Operating Expenses', 'Interest Expense', 'Fixed Asset Turnover', 'Asset Turnover', 'Current-Ratio', 'Quick-Ratio', 'Debt-Ratio', 'Debt to Equity Ratio', 'Market Cap', and 'Enterprise Value Multiple'. These variables are key indicators of a company's operational efficiency, financial health, and market performance, making them critical for estimating revenue and cashflows.

To prepare the data for machine learning models, it was first preprocessed. Missing and null values were removed rather than imputed, as their presence was minimal and the dataset size was large enough to mitigate any significant loss of information. This approach avoids potential biases introduced by imputation. Subsequently, the data was normalized using min-max scaling, which ensures that all features are on a comparable scale, preventing features with larger magnitudes from disproportionately influencing the regression models.

Multiple machine learning algorithms were tested to find the best model for cashflow estimation. These models were evaluated based on their predictive accuracy, measured by the mean absolute percentage error (MAPE). Here's a detailed breakdown of the algorithms and their performance. The relevant code snippets and the equations screenshot mentioned in the Jupyter notebook have also been mentioned:

**1. Random Forest Regressor (MAPE: 10.97%):** Random Forest is an ensemble learning method that builds multiple decision trees during training and averages their predictions. This algorithm tends to handle high-dimensional data well, making it suitable for financial datasets with complex feature interactions. Its moderate error rate could be due to overfitting, a common issue with random forests, especially when the number of features or trees is too high. Hyperparameter tuning, such as adjusting the number of trees or maximum depth, might further reduce the error.

**2. Support Vector Regressor (RBF Kernel) (MAPE: 1498.15%):** The Support Vector Regressor (SVR) aims to find a hyperplane in a high-dimensional space that fits the data points while minimizing prediction errors. The Radial Basis Function (RBF) kernel was used here, which is effective in capturing non-linear relationships. However, the extremely high MAPE suggests poor performance, possibly due to inappropriate kernel parameters, the high dimensionality of the dataset, or a lack of sufficient tuning. SVR models can also struggle with large datasets, especially when the data is not properly scaled or when there are complex patterns that RBF cannot capture without significant tuning.

**3. Multilayer Perceptron Classifier (MAPE: 83.12%):** Although a neural network model, the multilayer perceptron (MLP) was used in a regression setting. Neural networks can capture complex non-linear relationships, but they require careful tuning of hyperparameters such as the number of layers, neurons, learning rate, and epochs. The relatively high MAPE could be attributed to the absence of optimal tuning or overfitting due to the complexity of the model. In financial data, where patterns may not always be deeply non-linear, simpler models sometimes outperform neural networks unless a vast amount of training data is available and properly handled. This model was given a hidden layer of 500 neurons with RELU activation to capture non-linearity. 500 iterations were performed with the ADAM solver algorithm.

**4. XG Boost Regressor (MAPE: 6.38%):** Extreme Gradient Boost (XG Boost) is a powerful gradient boosting algorithm known for its high performance on structured/tabular data. It iteratively adds models to correct the errors of previous ones, leading to a strong learner. The low MAPE demonstrates that XG Boost captured patterns in the data well, benefiting from its regularization features that prevent overfitting. Its effectiveness may stem from its ability to handle missing data and to focus on minimizing loss functions like MAPE, which are critical in financial forecasting.

**5. XG Boost Random Forest Regressor (MAPE: 6.61%):** This variant combines the strengths of XG Boost and Random Forest. While slightly less accurate than the standard XG Boost model, its ensemble nature still yields competitive results. The relatively small increase in error suggests that the added complexity of combining multiple trees in an XG Boost framework may not always improve performance, but still provides a solid baseline.

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N f_i(x)$$

Where:

- $\hat{y}$  is the predicted value,
- $N$  is the number of trees,
- $f_i(x)$  is the prediction from the  $i$ -th tree for input  $x$ .

```
from sklearn.ensemble import RandomForestRegressor
```

```
model1=RandomForestRegressor()
```

```
model1.fit(training_data.iloc[:,0:13],training_data.iloc[:,13])
```

```
▼ RandomForestRegressor ⓘ ⓘ  
RandomForestRegressor()
```

```
prec1=model1.predict(testing_data.iloc[:,0:13])
```

```
mae1=mean_absolute_percentage_error(testing_data.iloc[:,13], prec1)
```

```
mae1
```

```
7.394424323949912
```

$$\hat{y} = \sum_{i=1}^n \alpha_i K(x_i, x) + b$$

Where:

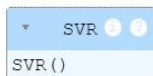
- $\hat{y}$  is the predicted value,
- $\alpha_i$  are the support vector coefficients,
- $K(x_i, x)$  is the kernel function between the support vectors  $x_i$  and the input  $x$ ,
- $b$  is the bias term.

With Support Vector Regressor

```
from sklearn.svm import SVR
```

```
model2=SVR(kernel="rbf")
```

```
model2.fit(training_data.iloc[:,0:13],training_data.iloc[:,13])
```



```
prec2=model2.predict(testing_data.iloc[:,0:13])
```

```
mae2=mean_absolute_percentage_error(testing_data.iloc[:,13], prec2)
```

```
mae2
```

```
394.03586771274325
```

$$\hat{y} = \sigma \left( W^{(L)} \cdot \sigma \left( W^{(L-1)} \cdot \sigma \left( \dots \sigma \left( W^{(1)}x + b^{(1)} \right) + b^{(L-1)} \right) + b^{(L)} \right) \right)$$

Where:

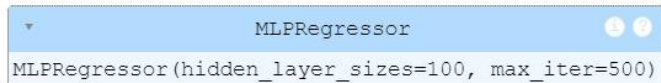
- $\hat{y}$  is the predicted output,
- $W^{(i)}$  are the weight matrices for the  $i$ -th layer,
- $b^{(i)}$  are the bias vectors for the  $i$ -th layer,
- $\sigma(\cdot)$  is the activation function, typically ReLU or Sigmoid.

With Neural Network

```
from sklearn.neural_network import MLPRegressor
```

```
model3=MLPRegressor(hidden_layer_sizes=100, max_iter=500, solver='adam', activation='relu')
```

```
model3.fit(training_data.iloc[:,0:13],training_data.iloc[:,13])
```



```
MLPRegressor(hidden_layer_sizes=100, max_iter=500)
```

```
prec3=model3.predict(testing_data.iloc[:,0:13])
```

```
mae3=mean_absolute_percentage_error(testing_data.iloc[:,13], prec3)
```

```
mae3
```

```
153.48385562824973
```

$$\hat{y} = \sum_{k=1}^K f_k(x), \quad f_k \in \mathcal{F}$$

Where:

- $\hat{y}$  is the predicted value,
- $K$  is the number of trees,
- $f_k(x)$  is the  $k$ -th tree's output for input  $x$ ,
- $\mathcal{F}$  is the space of all possible decision trees.

With Extreme Gradient Boosting Regressor based on decision trees

```
from xgboost import XGBRegressor
```

```
model4=XGBRegressor()
```

```
model4.fit(training_data.iloc[:,0:13],training_data.iloc[:,13])
```

```
► XGBRegressor ⓘ
```

```
prec4=model4.predict(testing_data.iloc[:,0:13])
```

```
mae4=mean_absolute_percentage_error(testing_data.iloc[:,13], prec4)
```

```
mae4
```

```
6.437865744402411
```

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T f_t(x)$$

Where:

- $\hat{y}$  is the predicted value,
- $T$  is the number of random forests built,
- $f_t(x)$  is the prediction from the  $t$ -th random forest for input  $x$ .

With Extreme Gradient Boosting Random Forest

```
from xgboost import XGBRFRegressor
```

```
model5= XGBRFRegressor()
```

```
model5.fit(training_data.iloc[:,0:13],training_data.iloc[:,13])
```

```
► XGBRFRegressor ⓘ
```

```
prec5=model5.predict(testing_data.iloc[:,0:13])
```

```
mae5=mean_absolute_percentage_error(testing_data.iloc[:,13], prec5)
```

```
mae5
```

```
6.419306708318433
```

The performance discrepancies between models arise due to differences in how each algorithm handles the intricacies of financial data. For instance, tree-based models like Random Forest and XG Boost excel in capturing hierarchical and nonlinear relationships, which are prevalent in financial systems. On the other hand, kernel-based methods like SVR might struggle without careful tuning, especially on datasets with high dimensionality or complex inter-feature correlations. Similarly, deep learning models like MLP are prone to overfitting when the dataset is not large enough or the feature engineering isn't sophisticated enough to capture subtle patterns.

In this context, models like **XG Boost** and **XGB Random Forest** outperformed others, likely due to their robustness in handling complex data distributions and their ability to generalize well, even with a reduced feature set.

The error of 6% still was something that could have been because of non-optimal preprocessing methods. Therefore, the scaling method was changed to the z-score method which ensures a Gaussian distribution of data and effective handling of the outliers.



```
from sklearn.metrics import mean_absolute_percentage_error
✓ 0.0s

mae=mean_absolute_percentage_error(y_test, prediction)
✓ 0.0s

mae
✓ 0.0s

0.1334454483543111
```

The XG Boost model was again fitted over the same data and resulted in a MAPE of just 0.13% which was a significant jump in the effectiveness of the model.

These cashflow predictions can then be used to find the Capital Budgeting Metrics in an even more effective way which can be used to estimating the risk associated with the project.

## OBJECTIVES

The primary objective of this study was to develop an effective machine learning model for predicting project cashflows, both inflows and outflows, based on historical financial data. Given the complexity and high dimensionality of financial datasets, the goal was to explore different regression algorithms to identify the most accurate model for forecasting cashflows. Furthermore, the study aimed to determine the optimal preprocessing and feature selection methods to improve the performance of these models. By achieving accurate cashflow predictions, the research sought to provide a valuable tool for estimating Capital Budgeting Metrics, such as Net Present Value (NPV), thereby supporting better decision-making in project evaluation and risk assessment.

## RESEARCH DESIGN

The research followed a structured approach involving data preprocessing, feature selection, and model evaluation. The 2014 Financial Dataset from Kaggle, consisting of 225 financial attributes from various organizations, served as the primary data source. Feature reduction techniques like Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) were applied to reduce redundancy and enhance model performance. Subsequently, multiple machine learning algorithms—including Random Forest, Support Vector Regressor, Multilayer Perceptron, XG Boost, and XG Boost Random Forest—were tested on the preprocessed dataset. Mean Absolute Percentage Error (MAPE) was used to assess the performance of these models. Additionally, the study explored different scaling techniques (min-max scaling and z-score normalization) to evaluate their impact on model accuracy.

The following were the metrics used: -

- **Cost of Revenue**

Definition: The total direct costs incurred in producing goods or services sold by a company. It includes raw materials, labor, and production costs but excludes indirect expenses like marketing and R&D.

Purpose: It helps assess how much it costs a company to generate revenue and is used to calculate gross profit.

- **R&D Expenses (Research and Development Expenses)**

Definition: The costs a company incurs in the process of developing new products, services, or technologies, or improving existing ones.

Purpose: These expenses are important for innovation and future growth, especially in tech, pharmaceutical, and engineering sectors.

- **SG&A Expense (Selling, General, and Administrative Expense)**

Definition: The combined costs related to selling products (salesforce salaries, marketing expenses) and general administrative costs (executive salaries, office rent, legal fees).

Purpose: It gives insight into the operating efficiency of the company and is a component of operating expenses.

- **Operating Expenses**

Definition: All expenses a company incurs from its normal business operations, including SG&A, R&D, and other costs like rent, utilities, and maintenance.

Purpose: These are non-production-related expenses that are crucial for running the company, and help in determining the company's profitability at the operational level.

- **Interest Expense**

Definition: The cost incurred by an entity for borrowed funds, like loans or bonds, typically expressed as a percentage of the loan.

Purpose: It is used to assess the cost of debt and how much a company spends on financing its operations through borrowing.

- **Fixed Asset Turnover**

**Definition:** A financial ratio that measures how efficiently a company is using its fixed assets (like property, plant, and equipment) to generate revenue.

**Purpose:** This helps evaluate how productive a company is in utilizing its investments in long-term assets.

## **Asset Turnover**

**Definition:** A ratio that measures the efficiency of a company in generating revenue from its total assets.

**Purpose:** It shows how well a company uses its assets to generate sales, with higher ratios indicating better efficiency.

- **Current Ratio**

**Definition:** A liquidity ratio that measures a company's ability to pay short-term obligations with its current assets.

**Purpose:** It indicates the company's short-term financial health. A ratio above 1 suggests the company can cover its liabilities.

- **Quick Ratio**

**Definition:** Also known as the "acid-test ratio," it is a stricter measure of liquidity than the current ratio, as it excludes inventory from current assets.

**Purpose:** This is used to evaluate a company's ability to meet short-term obligations without relying on inventory sales.

- **Debt Ratio**

**Definition:** A leverage ratio that shows the proportion of a company's total assets that are financed by debt.

**Purpose:** It helps assess the financial risk of the company. A high debt ratio implies more leverage and higher financial risk.

- **Debt to Equity Ratio**

**Definition:** A ratio that measures a company's financial leverage by comparing its total liabilities to shareholders' equity.

**Purpose:** It shows how much debt the company is using to finance its assets relative to the equity. A higher ratio indicates more risk.

- **Market Cap (Market Capitalization)**

Definition: The total value of a company's outstanding shares of stock.

Purpose: It gives an overall indication of the company's size and market value. Investors use it to determine the company's market standing and investment potential.

- **Enterprise Value (EV) Multiple**

Definition: The enterprise value multiple, such as EV/EBITDA, is used to measure a company's value and compare it to the earnings generated. EV is the company's market value, including both equity and debt, minus cash and equivalents.

Purpose: This multiple is often used to assess the company's overall valuation, particularly in the context of mergers and acquisitions. It shows how many times EBITDA the company is valued at.

These financial metrics are crucial for analyzing a company's performance, financial health, operational efficiency, and valuation

## **SOURCES OF DATA**

2014 Financial Data from *Kaggle* <https://www.kaggle.com/code/prayankkul/complete-financial-analysis/input>

This dataset (.csv) collects 200+ financial indicators for all the stocks of the US stock market. The financial indicators have been scraped from [Financial Modeling Prep API](#), and are those found in the 10-K filings that publicly traded companies release yearly.

## **LIMITATIONS OF STUDY**

While this research gives valuable insights about the models used for financial risk prediction and their accuracy, several limitations are there that can be acknowledged

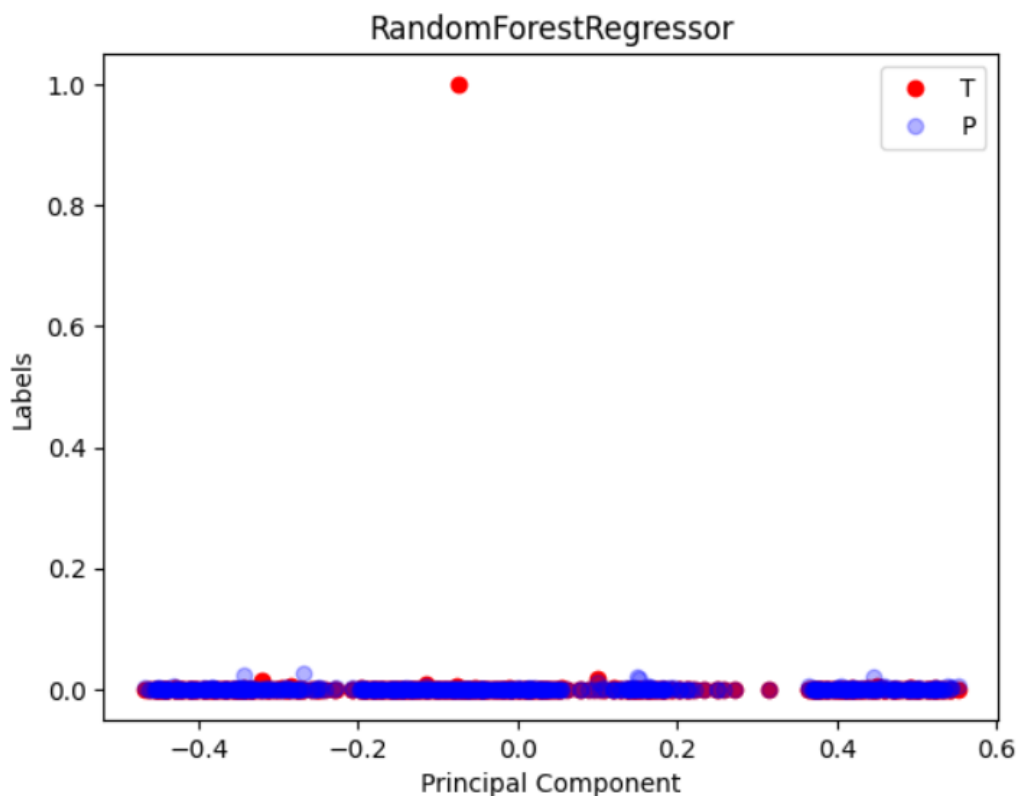
- **Data Dependency-** the prediction of the model is highly dependent on the data it is provided with. Incomplete or inconsistent may adversely affect the performance and reliability of the model. This suggests that the model's performance is constrained when there is an unavailability of the high-quality data.
- **Geographic and Market Constraints-** The study's scope is limited to certain markets and geographic region. It also does not include any government policy and market constraints. For this, further researched is needed.

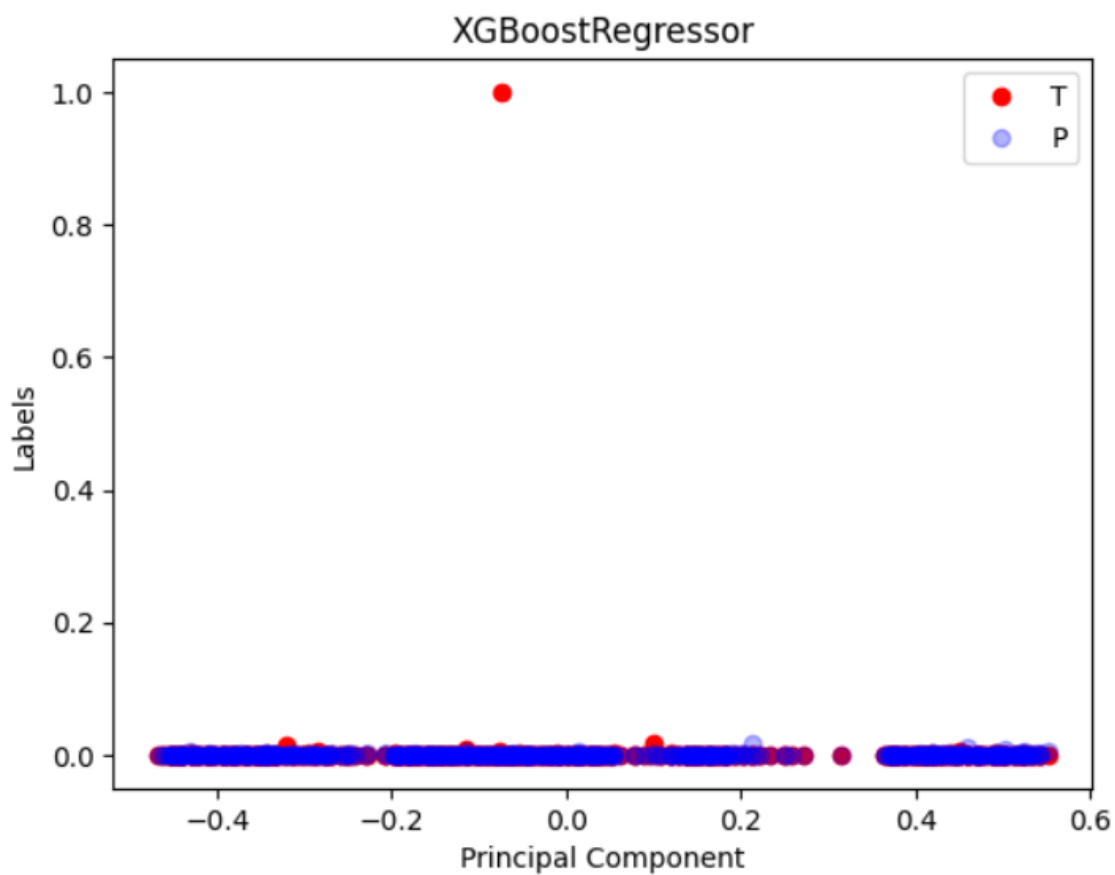
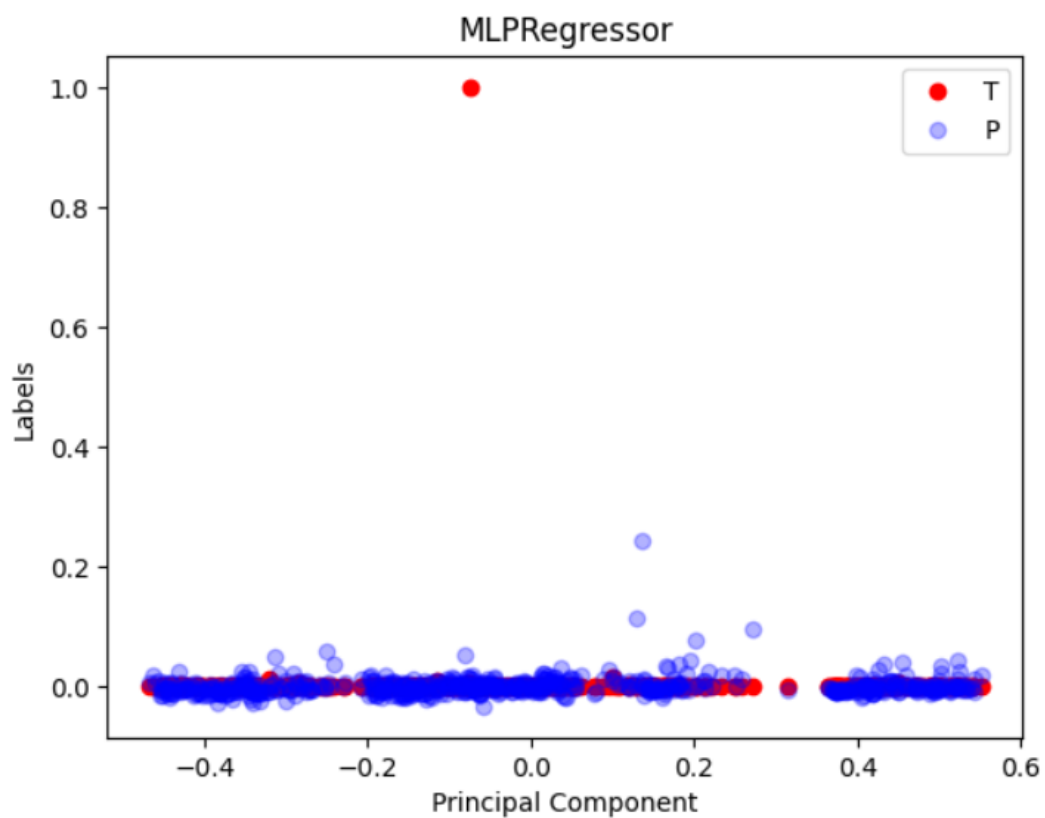
- **Real-time Prediction Limitations-** Although the models provide accurate results. But the financial status of a firm changes with time. This needs to be accounted using better data acquisition methods or by usage of algorithms like Recurrent Neural Networks, ARIMA or Prophet.

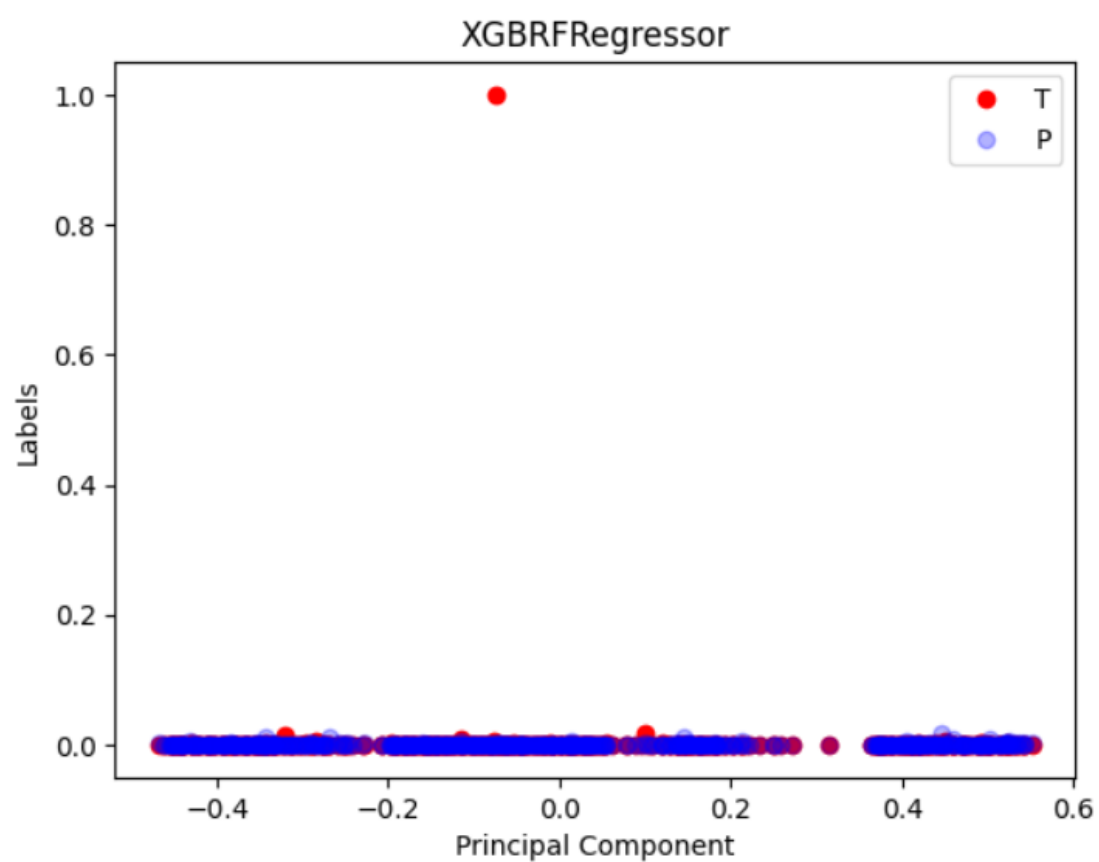
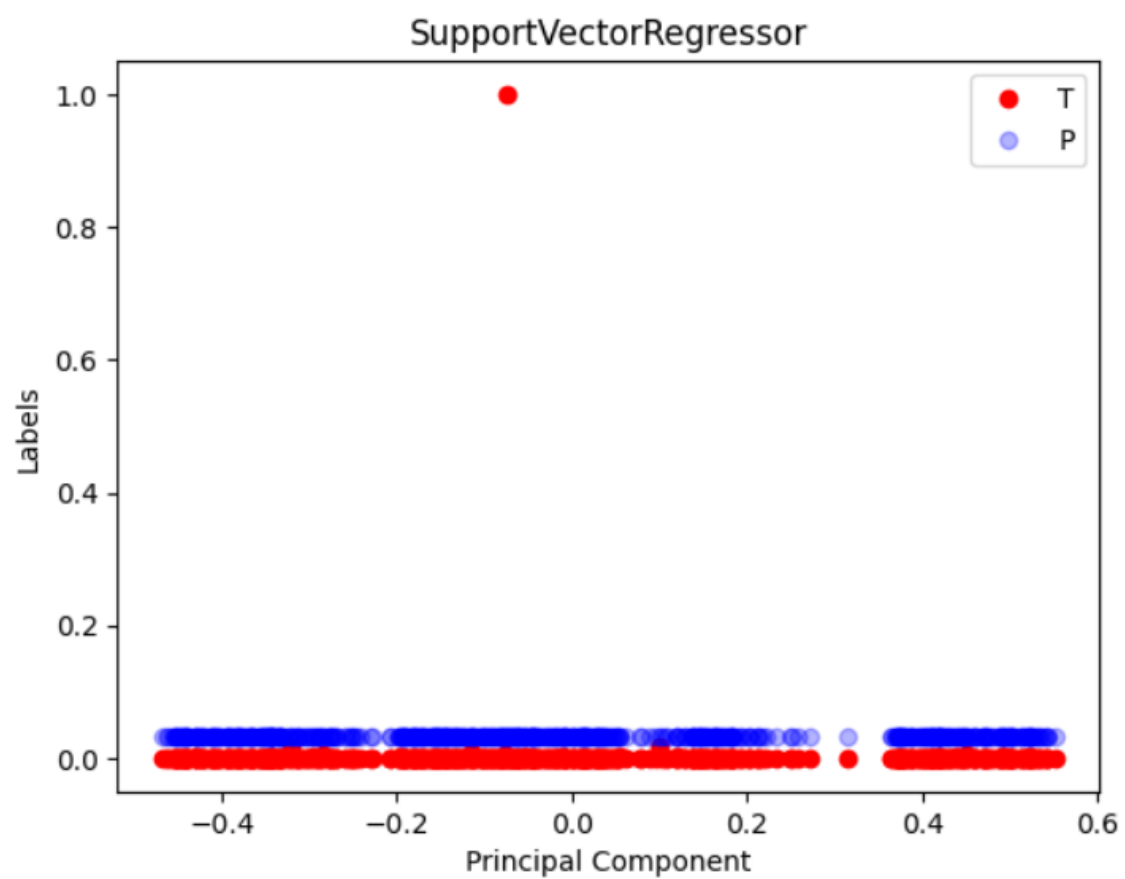
## DATA ANALYSIS AND INTERPRETATION

The data analysis revealed substantial differences in model performance based on the algorithms used. Tree-based models, particularly XG Boost and its Random Forest variant, exhibited the highest accuracy, with MAPE values of 6.38% and 6.61%, respectively, after initial testing. The results highlighted the effectiveness of XG Boost in handling complex financial datasets, benefiting from its ability to capture nonlinear relationships and its regularization capabilities that prevent overfitting. Support Vector Regressor, on the other hand, showed poor performance with an exceptionally high MAPE of 1498.15%, indicating that it was not well-suited for this dataset without significant tuning. A change in the data normalization method to z-score scaling led to a substantial improvement in XG Boost's performance, reducing MAPE to just 0.13%, underscoring the importance of appropriate preprocessing techniques. Overall, the analysis indicated that XG Boost, in combination with feature selection and z-score normalization, provided the most reliable and precise cashflow predictions.

The following are the plots depicting the comparison of different models in prediction power:







## RECOMMENDATIONS AND IMPLICATIONS

Based on the findings, it is recommended that future studies explore more recent financial datasets to validate the applicability of the models in contemporary settings. Additionally, a more thorough exploration of hyperparameter tuning for models like SVR and MLP could yield improved performance. The use of z-score normalization demonstrated a significant impact on model accuracy, suggesting that careful selection of preprocessing techniques is critical for financial forecasting. From a practical standpoint, the results of this study have significant implications for corporate financial planning, where accurate cashflow predictions are essential for capital budgeting and risk management. Organizations could implement the developed XG Boost model to enhance their financial forecasting capabilities and improve decision-making regarding long-term investments.

## CONCLUSION

This study successfully identified XG Boost as the most accurate and reliable model for predicting project cashflows based on the 2014 Financial Dataset, achieving a remarkable improvement in accuracy after switching to z-score normalization. While tree-based models outperformed other machine learning algorithms, the study highlighted the critical role of data preprocessing and feature selection in improving model performance. The results provide a strong foundation for future research and practical application in financial forecasting, particularly in estimating Capital Budgeting Metrics. Despite certain limitations, the study's methodology offers valuable insights into optimizing machine learning techniques for cashflow prediction in corporate finance.



## BIBLIOGRAPHY

1. Predicting cash holdings using supervised machine learning algorithms. <https://jfin-swufe.springeropen.com/articles/10.1186/s40854-022-00351-8>
2. Financial Risk Prediction and Management using Machine Learning and Natural Language Processing. [https://thesai.org/Downloads/Volume15No6/Paper\\_23-Financial\\_Risk\\_Prediction\\_and\\_Management.pdf](https://thesai.org/Downloads/Volume15No6/Paper_23-Financial_Risk_Prediction_and_Management.pdf)
3. The analysis of financial market risk based on machine learning and particle swarm optimization algorithm. <https://jwcen-urasipjournals.springeropen.com/articles/10.1186/s13638-022-02117-3>
4. Machine learning for financial forecasting, planning and analysis: recent developments and pitfalls. <https://link.springer.com/article/10.1007/s42521-021-00046-2>
5. Estimating Project Cash Flows. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3815383](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3815383)
6. A study on cash flow analysis. <https://www.jetir.org/papers/JETIREDO6067.pdf>
7. Cash flow management and its effect on firm performance: Empirical evidence on non-financial firms of China. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0287135>
8. The relation between cash flows and economic performance in the digital age: An empirical analysis. [https://www.researchgate.net/publication/340177986\\_The\\_relation\\_between\\_cash\\_flows\\_and\\_economic\\_performance\\_in\\_the\\_digital\\_age\\_An\\_empirical\\_analysis](https://www.researchgate.net/publication/340177986_The_relation_between_cash_flows_and_economic_performance_in_the_digital_age_An_empirical_analysis)
9. A nonlinear dynamic approach to cash flow forecasting. <https://link.springer.com/article/10.1007/s11156-022-01066-8>
10. Cashflow prediction: MLP and LSTM compared to ARIMA and Prophet. <https://link.springer.com/article/10.1007/s10660-019-09362-7>
11. Application Of Machine Learning Algorithms to Free Cash Flows Growth Rate Estimation. <https://www.sciencedirect.com/science/article/pii/S1877050923009560>
12. Forecasting project's cashflow using machine learning. <https://thesis.eur.nl/pub/51895/Buter.pdf>