Decision Analytics Journal 7 (2023) 100212

Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

An exploratory data analysis approach for analyzing financial accounting data using machine learning

Potta Chakri [a], Saurabh Pratap [b,*], Lakshay [b], Sanjeeb Kumar Gouda [b]

[a] Department of Mechanical Engineering, Indian Institute of Information Technology, Design & Manufacturing, Jabalpur, India
[b] Department of Mechanical Engineering, Indian Institute of Technology (BHU), Varanasi, 221005, India

ARTICLE INFO

Keywords:
Exploratory data analysis
Financial accounting
Linear regression
K-nearest neighbor
Support vector regressor
Decision tree

ABSTRACT

Analyzing financial accounting transactions is essential for gaining valuable hidden insights, optimizing performance, reducing expenses by identifying more efficient methods of conducting business operations, and improving profitability and bottom line. This study uses exploratory data analysis to analyze financial accounting data, including balance sheets, income statements, and cash flow statement data. Such descriptive analytics considers various parameters, such as the Debt-to-Equity Ratio, Current Ratio, Return on Capital Employed, Net Profit Margin, and Inventory Turnover Ratio, to determine profitability for investment decisions. Afterward, predictive analytics is used to predict total revenue as a dependent variable. Four supervised machine learning models are employed: Linear Regression, K-Nearest Neighbor, Support Vector Regressor, and Decision Tree. The results suggest that the decision tree is the most valuable model for performance analytics. The hyperparameter is the maximum depth of the tree, and its optimal value of nine is determined using a grid search.

1. Introduction

Financial accounting refers to the preparation of quarterly and annual financial statements, including balance sheets, income statements, and cash flow statements, for a given period. Financial statements (FSs) provide information to investors, regulators, financial analysts, and other users in a standardized and reliable format for use in making economic decisions regarding the financial condition, performance, and changes of a business [1].

Financial accounts act as key for current and future investors to make investment decisions to sell, buy or hold equity or debt instruments for financial return [2]. Financial accounting enables businesses to keep track of their quarterly or annual financial reports and transactions [3] It also enables us to make sound decisions regarding resource allocation and investment opportunities [2]. By combining AHP, TOPSIS, and Grey Relational Analysis, Guru and Mahalik [4] attempts to determine which Indian PSU banks are the most efficient and then compares those findings. The level of development of capital markets is linked to how companies perform after issuing securities in those markets [5]. There are two ways for a business to raise capital: Marsh [6] and Zhang et al. [7] equity financing and debt financing. In equity financing, the company sells company shares, which the general public, i.e., investors, can buy to invest money. It is one method of raising funds, the other being to borrow from banks, or creditors [5]. Bhatia et al. [8] discusses empirical methodological progress and new bank performance analysis dimensions. A debt-financing company must pay back the principal amount plus interest after a period of time, whereas an equity-financing company is not obligated to pay its investors; if the company profits, investors receive their profit share immediately. Included in quarterly and annual reports are the company's liabilities and equity, as well as its total revenue, total assets, and net income or loss.

Overall, the contribution of our study is to provide a systematic analysis of financial data from various sources, including the balance sheet, the income statement, and the cash flow statement. This type of analysis can assist business owners and managers in making better data-driven decisions as opposed to intuitive ones. This study also identifies trends in the critical financial parameters of a company, making it possible to predict the future performance of the company using a predictive model. It is needed to develop a predictive model for companies' revenue from assets, equity, liabilities, and income, taking into account the heterogeneity of different companies. Risk management should also be included in annual reports to help investors understand the company's future cash-flows and predict stock returns [9]. These reports can be shared with the firm's business investors and creditors. The general public can also view these reports and decide whether or not to invest in a company's shares. Therefore, financial accounting is essential.

* Corresponding author.
E-mail address: s.pratapiitkgp@gmail.com (S. Pratap).

https://doi.org/10.1016/j.dajour.2023.100212
Received 16 January 2023; Received in revised form 21 March 2023; Accepted 27 March 2023
Available online 29 March 2023

Table 1a
Literature review summary.

Sr. No.	Researcher (Year)	Problem type	Model	Dataset
1	Khaddafi et al. [18]	Finding impact between ROE, ROA, D/E ratio, NPM & current ratio towards growth income.	Assumption testing using F-test & T-test	Indonesia Stock Exchange
2	Nuryani and Sunarsi [19]	Finding influence between CR and D/E Ratio on Dividing Growth.	Descriptive analysis with regression testing and hypothesis testing.	Trucking company PT. Gajah Mas.
3	Nariswari and Nugraha [20]	Finding the effect of NPM, GPM and TAT.	Multiple linear regression algorithms are used by processing data using software Eviews version 8.	Packaging and Plastic industry companies were listed on Indonesia Stock Exchange (ISE) from 2014 to 2018.
4	Sunjoko and Arilyn [21]	Examine the effect of fixed assets turnover, total assets turnover, inventory turnover, and average collection period, current ratio on profitability	Empirical Techniques are used to find relations.	The pharmaceutical companies were listed on Indonesia Stock Exchange (ISE) period 2007 to 2013.
5	Jufrizen and PHS [22]	Finding influence of Effect of CR, D/E ratio, NPM and TAT on EPS.	Linear and Multiple regression analysis methods and the classical hypothesis test, partial Test, and simultaneous Test with SPSS 18 were used.	The chemical company was listed on the Indonesia Stock Exchange (ISE) from 2013 to 2017.

Financial reporting includes external or audited financial statements (income statement, statement of comprehensive income, balance sheet, statement of cash flows, and statement of stockholders' equity). Timeliness of audit reports is recognized by the accounting profession, users of accounting information, and regulatory and professional agencies as a crucial aspect of financial accounting information. Companies must attempt to reduce audit delays and improving timeliness of audit reports [10].

1.1. Need of data analytics

Currently, there is intense competition on the market among businesses. Companies striving to achieve the top position. Today, technology is a tremendous asset for businesses and plays a crucial role in their success ([11,12]). One such method is data analytics, which enables us to extract insights from data. By utilizing these helpful tips, businesses can reduce expenses and increase profits. It offers a new method of conducting business operations [13,14]. It also aids in comprehending customer behavior and preparing customer-centric strategies [15,16]. Data analytics encompasses both exploratory and prescriptive analysis [17] . Therefore, data analytics is crucial. In the proposed paper, data analytics are performed on annual and quarterly report data from 2016 to 2021 for 25 companies worldwide. This data set contains the balance sheet, income statement, and cash flow statement information for all 25 businesses. These businesses operate in numerous fields, including finance and banking, telecommunications, biotechnology, beverages and brewing, electronics, and pharmaceuticals. Consequently, our dataset is a collection of diverse businesses.

Exploratory data analysis (EDA) or descriptive analysis is performed. Various profitability indicators, such as the Debt-to-Equity Ratio, Current Ratio, Return on Capital Employed, Net Profit Margin, and Inventory Turnover Ratio, are plotted against time to reveal whether a company is profitable or not. Additionally, predictive analysis is performed to predict the dependent feature, a company's total revenue. The parameters Total Assets, Total Liabilities, Net Income, and Stockholders' Equity are distinct. Four techniques for supervised machine learning are employed: Linear Regression, K Nearest Neighbor, Support Vector Regressor, and Decision Tree. This is how the proposed data set is analyzed using data analytics.

The knowledge we gain about various companies helps us determine which companies are suitable for investment, whether it is advantageous to purchase a company's stock, and what the company's business growth is. It has performed well throughout the years. What is the profit of the company? These insights allow us to make a more informed decision, as investing in a losing company is a terrible idea. Moreover, analytics can help us here. Investment is significant because it increases our wealth over time and generates returns that are resistant to inflation. These are the types of benefits we obtain from analyzing data using data analytics. The novelty of the proposed paper is that it examines data analytics in financial accounting; prior to that, a comprehensive study of accounting terminology is conducted as follows:

- Determining different accounting parameters and their importance.
- Interpret the parameters and know their mathematical formulation.
- Perform data analytics to gain insights and managerial implications.

Overall, this study allows us to comprehend accounting, which is essential before all else. In addition, various machine learning models, as stated earlier, are used in the proposed paper, that are successfully implemented.

The paper is organized as follows. Section 2 provides a literature review of different research works in the field of financial accounting. The proposed framework of data analytics approach for financial accounting is presented in Section 3. Section 4 presents the results and discussions discussing on how the dataset is generated and the results of the descriptive, predictive and prescriptive analyses are presented in detail. Section 5 discusses the managerial implications of the study performed in this paper. Finally, some concluding remarks along with future scope are presented in Section 6.

2. Literature review

This section begins with a review of prior work on accounting's essential parameters, followed by a discussion of prior work on the data analytics of accounting data, and concludes with a discussion of the need for the current study as well as previous techniques used. Table 1a provides a review summary of various research papers, including their problem type, the model used, and the Dataset. In the past, various quantitative and qualitative methods have been utilized for financial and accounting research. Kothari et al. [23] describes the strategic models for overcoming financial challenges in pharmaceutical supply chains. Bánhidi and Dobos [24] describes the connection between financial performance and digitalization. On the other hand, Özdemirci et al. [25] describes the financial system and risk associated with social banking alternatives.

Moise Ikegwu [26] explains the approaches from machine learning (ML) to predict the annual direction of profitability for firms. Their

results conclude that the random forest does not find the accrual profitability measures (Return on common equity (ROE), Return on assets (ROA), Return on net operating assets (RNOA)) as useful as the cash flow measures (Cash flow from operations (CFO) and Free cash flow (FCF)) in predicting the future. However, this work does not include regression setting to make predictions about the magnitudes of profitability. The purpose of Lv et al. [27,28] is to enhance the efficiency and precision of data processing in the finance sector by enhancing the trustworthiness of information acquisition. It also conducts statistical analysis on the evaluation indicators of AI research papers and looks into current publications and norms pertaining to financial intelligence system. Sultan Almansoori et al. [29] discusses that profitability ratios Return on common equity (ROE), Return on assets (ROA), Profit margin ratio have a positive impact on the profitability of Abu Dhabi National Oil Company (ADNOC) whereas the inventory turnover ratio has a negative impact on the profitability of ADNOC. In the past, machine learning and artificial intelligence have been utilized in a variety of financial contexts. For instance, Chhajer et al. [30] provides finance firms with artificial neural networks for stock market prediction. Similarly, Bhattacharjee et al. [31] provides an ML and DEMATEL-based approach for modeling the purchase intention of Indians, as well as modeling purchase intention. Afriyie et al. [32] provides a supervised machine learning algorithm for detecting fraud. Moreno-Garcia et al. [33] provides automated abstract screening based on machine learning for systematic reviews. On the other hand, Perla et al. [34] provides a neural network-based method for analyzing forex rates on the financial market. Data analytics, both descriptive and predictive, have proven to be a business enabler. In this direction, Pagano [35] provides a model for the combined use of LSTM and neural networks for predictive plant maintenance. ML techniques have since been implemented in a variety of industries; for instance, Durai and Shamili [36] provides data analytics techniques for smart farming. Hassan et al. [37] provides a comprehensive strategy for social inclusion that emphasizes equal opportunities and minimal disparities between individuals. Das et al. [38] presents a study on Bitcoins in which the K-means clustering method is used to determine the returns taking volatility and extreme values into account.

Khaddafi et al. [18] examines a hypothesis indicating that Return on Asset (ROA), Return on Equity (ROE), Net Profit Margin (NPM), Debt to Equity, and Current Ratio (CR) directly impact growth income and concludes that ROA, ROE, and NPM have a positive impact on growth income, whereas the debt-to-equity ratio and current ratio have a negative impact on growth income. Nuryani and Sunarsi [19] presents a hypothesis which states that the cash flow and debt-to-equity ratio have a direct impact on dividend return and infers that the current ratio and debt-to-equity ratio have a substantial impact on dividend growth. The objective of Nariswari and Nugraha [20] is to determine the relationship between NPM, Gross Profit Margin (GPM), and Total Asset Turnover (TAT) and profit growth, and to conclude that NPM has a greater impact on profit growth than GPM and TAT. Sunjoko and Arilyn [21] explains how to examine the effect of fixed assets turnover, total assets turnover, inventory turnover, and average collection period on profitability, as well as how extrapolating profitability influences fixed assets turnover and current ratio. In contrast, neither total asset turnover nor inventory turnover nor average collection period impact profitability. Sanchís-Pedregosa et al. [39] analyses the effects on financial performance of one of the dimensions of the services outsourcing strategy, namely depth, which entails determining whether each activity is completely or partially delegated. Cost, productivity, and profitability indicators assessed financial effects.

Jufrizen and PHS [22] examines the impact of the net profit margin, total asset turnover, and current ratio on earnings per share, concluding that the D/E ratio and CR have no significant impact on EPS. Still, TAT and NPM have a substantial effect on Earning per Share (EPS). Schneider et al. [40] determines the influence of data analytics in financial accounting, discusses challenges and research opportunities, and uses Ruchala and Mauldin's meta-theory of accounting information systems to determine current data analytics use. Oberholzer [41] describes a balance sheet-based data envelopment analysis model and an income statement-based model and compares these models based on the technical efficiencies and scale of the companies. Furthermore, it implies that although the balance sheet and income statement appear to be correlated. To determine if radical innovation can lead to better financial performance, Xin et al. [42] tracked the financial performance of publicly traded manufacturing firms in the United States that introduced radical innovations. Dutta et al. [43] discusses generating four types of data namely Asset Returns, Premium Inflows, Policy Related Outflows and Operating Expenses for performing stochastic linear programming (SLP) for life insurance companies.

2.1. Data analytics in stock price prediction

Ramesh et al. [44] describes a technique for resolving a stock market problem by predicting intraday stock returns using a back-propagation neural network algorithm. Sismanoglu et al. [45] present the prediction of the future value of a specific stock by analyzing its historical financial data on the market exchange process using an LSTM Deep Learning model on a dataset of IBM from the New York Stock Exchange from 1968 to 2018. Attempts to identify hybridized soft computing techniques, such as a neuro-fuzzy system, for automated stock market forecasting and trend analysis on Nasdaq Stock Market companies. Shrivastav and Kumar [46] described the prediction of Coca Cola stock dataset listed in New York Stock Exchange (NYSE) using machine learning techniques including Random Forest, Gradient Boosting Machine and Deep learning. Daradkeh [47] discusses the prediction of stock trends for Dubai Financial Market (DFM) using a hybrid data analytics framework that combines both convolutional neural networks and bidirectional long short-term memory (CNN-BiLSTM). Boyacioglu and Avci [48] describe the prediction of stock market return using an Adaptive Network-Based Fuzzy Inference System (ANFIS) for companies listed on the Istanbul Stock Exchange (ISE). Morris et al. [49] intends to identify a token-based ensemble algorithm that uses nonlinear and linear estimators, i.e., LSTM and Kalman filter, to forecast financial time series data. Lin et al. [50] discusses the support vector machine-based method for predicting the stock price trend of Taiwan Stock Exchange-listed companies. Wang and Leu [51] presents an RNN-based model for predicting stock price trends. RNN training utilizes features extracted from ARIMA analyses.

Ananthi and Vijayakumar [52] describes a method for stock price prediction based on candlestick and regression pattern detection. Consider stock price graphs as images and model them using deep learning neural networks [53]. The data comes from the Chinese stock exchange. He and He [54] analyzes the problems of information asymmetry between farmers and financial institutions in the current traditional and innovative financing models of farmers and financial institutions, high thresholds, and inability to account for knowledge asymmetry and inadequate collateral for agriculture and proposes solutions. Iyke and Ho [55] explains how commodity returns from energy, meat markets, livestock, and agriculture can be used to predict stock price. The dataset is comprised of data from the UK, Netherlands, and US stock markets spanning four centuries. Ghosh et al. [56] present random forests and LSTM as training methods to examine their effectiveness in predicting Standard and Poor's 500 stocks for intraday trading from 1993 to 2018. Using various machine learning algorithms, Leippold et al. [57] seek to identify an exhaustive set of return prediction factors, such as the impact of transaction costs. The data comes from the Chinese stock exchange. To examine the effect of short-term vendors on share prices and liquidity, Berkman and Eleswarapu [58] implemented forward trading on the BSE. Amin [59] presented artificial bee colony (ABC) algorithm for extracting the most suitable feature subset of accounting ratios to detect fraud on a sample of 83 Egyptian companies from 2010–2017. A novel portfolio building method uses machine learning algorithms for

stock return prediction and a mean-VaR model for choosing a portfolio [60]. To predict stock prices, machine learning regression models like Random Forest, Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), Support Vector Machine Regression (SVR), k-Nearest Neighbors (KNN), and Artificial Neural Network (ANN) are used.

2.2. Techniques

Khaddafi et al. [18] utilized F-test and T-test to test their hypotheses and validated them using Indonesia Stock Exchange data. Nuryani and Sunarsi [19] utilized statistical analysis, research analysis for explanation, hypothesis testing, and regression testing. Used information obtained from PT. Gajah Mas. Nariswari and Nugraha [20]. Algorithms for multiple linear regression are utilized by processing data with Eviews version 8 software. Companies in the plastics and packaging industries listed on the Indonesia Stock Exchange provided the data used. Sunjoko and Arilyn [21]. To determine correlations, empirical methods are used. Multiple linear regression analysis methods, t-test, the classical assumption test, the coefficient of determination, and f-test were utilized with software SPSS version 18 [22]. During the 2013–2017 timeframe, the data used came from a chemical company listed on the Indonesia Stock Exchange. Olowokudejo and Akindipe [61] utilized F-test to test their hypotheses validated from Nigeria Success Digest for the years 2011 to 2021. Moise Ikegwu [26] used a large sample of US firms over the period 1963–2011 for implementing and training random forests in a classification setting. Dospinescu and Dospinescu [62] utilized R Square, t test, F test and p-values for implementing regression model regarding the profitability of Romanian Stock Exchange's companies.

In order to mine trading patterns of pyramid schemes from financial time series data, Lv et al. [27,28] propose a quantitative approach. The system is made up of two sub-parts, the Contrast Trading Pattern Mining (TPM) algorithm and the Long-Range Sequence De-noising (LoRSD) algorithm. From 2005–2018, Yu et al. [63] analyze data from a sample of 30 Chinese provinces to examine the nonlinear relationship between economic growth and energy usage from the viewpoint of spatial spillover. Three viewpoints—direct finance, indirect finance, and financial openness are used to evaluate financial development. Chen et al. [64] proposes a hierarchical attention network for mining information and the stock market for valuable data for stock prediction, with the help of attentive multi-view news learning (NMNL). A news encoder and a market information encoder make up the fundamental components of the strategy. Based on gated orthogonal recurrent units (GORU) and variational auto-encoder (VAE), Leng et al. [65] proposes a new model for predicting the direction of market prices. Market information is created by concatenating encoded text information with normalized historical price information, which is then decoded by VAE. To make his forecasts about the stock market gauge, Wang et al. [66] uses the most recent iteration of the deep learning framework known as Transformer. In the beginning, Transformer was designed to solve the problem of natural language processing, but more recently, it has been utilized in the field of time series forecasting. By utilizing an encoder–decoder architecture and a multi-head attention mechanism, Transformer is able to more accurately characterize the fundamental principles that govern the dynamics of the stock market. Teng et al. [67] presents an LSTM model (MLCA-LSTM) that incorporates information from multiple scales of local cues and a hierarchical attention structure in order to capture these fundamental price trend patterns. The first benefit is the induction of distinctive local descriptors, which enriches multi-scale data, by applying sliding windows of various scales to stock price series. In addition, it utilizes a multi-branch LSTM structure and a hierarchical attention strategy to simultaneously focus on an aggregate temporal dependency and multi-scale interactions.

Most of the researchers have used linear and multiple regression to predict any financial accounting parameters in predictive analyses. Most studies have not conducted a comprehensive analysis of financial accounting. Also, in most studies, data analytics are performed on companies from a single industry, such as the electronics industry or the pharmaceutical industry, whereas our proposed paper analyzed heterogeneous industries (total 17 types of industries) to determine the most influential parameters for business performance. Further, Table 1b presents the comparative literature on financial studies conducted using data analytics and machine learning approaches. It shows the type of models and techniques applied in the past work and shows the need of current study.

3. Proposed research framework

The framework presented in this paper outlines a systematic approach for conducting data analytics in accounting. This serves as the basis of our work. By following this framework, we can better understand and analyze accounting data to make informed decisions. In this paper, we provide a detailed description of each step and demonstrate how the framework can be applied in practice.

Four main steps which are included in the framework are:
(i) Recognizing the Accounting Concepts
(ii) Descriptive Study
(iii) Predictive Study
(iv) Prescriptive Study

In the first step, we examine key accounting parameters, which is necessary for understanding the problem's context. The second step is to conduct a descriptive analysis. Descriptive analysis consists of the following steps: removing outliers; determining which company has the highest debt-to-equity ratio; EPS (earnings per share) vs. year; ROCE (return on capital employed) vs. year; Net Margin vs. year; Inventory Turnover Ratio vs. year; and visualizing company locations. Fig. 1 illustrates the basic structure of the paper, allowing us to visualize each step.

Thirdly, a predictive study is conducted. Net Income, Total Assets, Total Liabilities, and Shareholders' Equity were treated as independent features, while Total Revenue served as the output or target feature. In the proposed paper, four different models are used. Linear Regression, KNN (K Nearest Neighbors), SVR (Support Vector Regression), and Decision Tree are these models. All of the aforementioned machine learning algorithms are supervised machine learning algorithms. We employ the supervised learning algorithm because we already know the dataset's inputs and outputs. Using training and testing data, we train our models and then predict total revenue. The third step involves prescriptive analysis. We compared the performance metrics of our four models and determined which model was superior. Pandas and Sklearn are utilized for data manipulation, matplotlib and Seaborn for data visualization, and the machine learning library Sklearn for prediction. These libraries are a fundamental component of the Python programming language. Therefore, Python is used for all of the code in the proposed paper.

4. Analysis and result

4.1. Dataset

The analysis was conducted on the financial data of 25 global companies from 2016 to 2021. Companies' financial information includes balance sheets, cash flow statements, and income statements. Numerous businesses release their financial statements quarterly and annually. They are available on websites such as Money Control and Screener. Each financial report contains a variety of economic indicators that describe a company's performance on the market, its profit and loss, its suitability for investment, and its assets, liabilities, revenue, and expenses, among others. Companies' financial information includes balance sheets, cash flow statements, and income statements. Our raw data contains 32 columns including *mapcode, amount, compnumber, reporttype, consolidated, longname, auditorstatus, currency, shortname, mic, countrycode, region, ticker, address, indicator, statement, website*.

Table 1b
Comparative literature review of previous studies.

S. No	Authors	Industry	Deterministic/stochastic	Heterogeneous	Descriptive	Predictive	Prescriptive	Machine learning	Features	Techniques
1	Nielsen [68]	Insurance		✓	✓	✓	✓	Semi-supervised	Revenue, costs, and profit	Clustering, PCA, FA
2	Olowokudejo and Akindipe [61]	Accounting		×	×	✓	×	×	PAT, SHF, TA	OLSLR
3.	Yang et al. [69]	Stock Market		×	×	✓	×	Supervised	Financial news, financial reports, and tweets	CNN, RNN
4	Tripathi et al. [70]	Stock Market		×	×	✓	×	Supervised	Stock market prediction	FU-WOA, NN
5	Dutta et al. [43]	Insurance	Stochastic	×	×	✓	✓	×	ARs, PIs, PROs, OEs	SLP
6	Moise Ikegwu [26]	US firms	Deterministic	✓	✓		✓	✓	ROE, ROA, RNOA, CFO, FCF	RF, CVTS
7	Rakshit et al. [71]	Banking Sector		×	×	✓	×	Supervised		RBF, ALSD
8	Pechlivanidis et al. [72]	Greek related companies	Deterministic	✓	✓	✓	×	Unsupervised	ROE, ROA, ROCE, EBITDA	LSTM
9	Zhao and Yang [73]	Stock Market		✓	×	✓	×	Supervised	SA-DLSTM, LSTM	Stepwise logit regression, RF, Elastic Net
10	Heikal et al. [74]	Automotive		×	✓	✓	×	×	ROA, ROE, NPM, DER, CR	MLR
11	Current study	Various Industries	Deterministic	✓	✓	✓	✓	Supervised	EPS, ROCE, NPM, ITR, D/E ratio	LR, KNN, SVR, DT

PCA = Principal Component Analysis, FA = factor analysis, PAT = Profit after Tax, SHF = Shareholders' Fund, TA = Total asset of the companies, OLSLR = Ordinary Least Square Linear Regression, SLP = Stochastic Linear Programming, ROE = Return on common equity, ROA = Return on assets, RNOA = Return on net operating assets, CFO = Cash flow from operations, FCF = free cash flow, AR = Asset Returns, PI = Premium Inflows, PRO = Policy Related Outflows OEs = Operating Expenses, RF = Random Forest, CVTS = Cross Validation for Time Series, RBF = Radial Basis Functions, ALSD = Active Learning Through Sequential Design, LSTM = Long short-term memory, ROCE = return on capital employed, EBITDA = earnings before interest, taxes, depreciation and amortization, DEPR_PPE = Depreciation of fixed assets, OPMBD = Operating income per total sales, NPM = Net Profit Margin, CR = Current Ratio, NN = Neural network, FU-WOA = Fly Updated Whale Optimization Algorithm, SA-DLSTM = Denoising autoencoder Long short-term memory.
MLR = Multiple linear Regression, EPS = Earnings per share, D/E ratio = Debt to Equity ratio, ITR = Inventory turnover ratio, LR = Linear Regression, KNN = K-nearest neighbor, SVR = Support Vector Regressor, DT = Decision Tree.

Out of which, only six columns are relevant to our work; the others we dropped. Those six useful columns are: *amount, report date, report type, shortname, indicator, and statement*. This information is provided in Table 2, which describes the significance of each column in the dataset. Finance and banking, telecommunications, biotechnology, beverages and brewing, electronics, and pharmaceuticals are among the 17 industries used to determine the most influential parameters for business performance. There are several heterogenous companies in our dataset including Finance and Banking (3), Telecommunications company (3), Biotechnology company (2) and Insurance, Media, Transportation company as well, this shows the variety of companies in our dataset. We can see that the dataset contains businesses from various industries. Consequently, our dataset is a collection of diverse businesses. Due to confidentiality, we cannot disclose the names of the companies involved in this work.

As mentioned in the introduction, we will now discuss financial accounting in depth. Balance Sheets, Income Statements, and Cashflow

Table 2
Explanation of columns.

Column name	Explanation
Shortname	Name of company
Statement	Out of four financial statements, i.e., Balance Sheets, Cash Flow Statements, Income statements and Derived Statements, whose data is given.
Indicator	Each financial statement includes different parameters or indicators, and those data are given in this column.
Reporttype	Whether a report is an annual report or a Quarterly report.
Reportdate	On which day the report is published.
Amount	The value of different indicators or parameters is given.

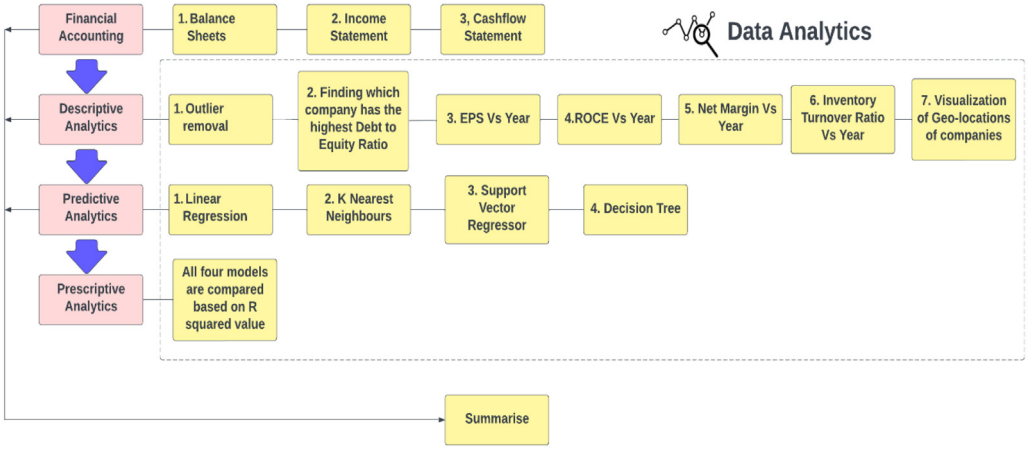

Fig. 1. Research framework.

Table 3
Explanation of different types of financial reports.

Name of report	Explanation
Balance Sheet	A balance sheet tells us about a company's liabilities, assets and shareholder equity.
Income Statement	An income statement tells us about a company's income and expenses. It also tells whether a company is making a profit or loss for a given period.
Cash flow Statement	A cash flow statement tells us about a company's aggregated data about all cash inflows a company acquires from its continuing operations and external investments. It also includes all cash outflows the company is doing for its business and investments during a given period.
Derived	The derived sheet includes parameters that are calculated from parameters from the above three reports. Mainly this report includes Ratio Analysis parameters.

Statements are just a few of the various types of financial reports that are described in Table 3. Further, Appendix A.3 depicts the Balance Sheet for XYZ Company. This is the actual format that companies use to generate their Balance Sheets. Due to privacy policies, the company name has been masked.

In company's balance sheet, there is information about its liabilities, assets, and shareholder equity in the form of quarterly or annual reports. Cash flow statements tells us about a company's all cash inflows and outflows. It also has information of different parameters such as cash from financing, cash from operations, cash from investing, present in companies Cash flow Statement. Appendix A.1 shows the Cash Flow Statement of Company XYZ. This is the actual structure in which companies produce their Cash Flow Statement. The company name is masked due to privacy policy. Further, it also has information of the different parameters such as total revenue, net income, earning per share, total expenses, present in companies income statement. Income statement tells us about a company's income and expenses. It also tells whether a company is making a profit or loss for a given period.

Appendix A.1 shows the Income Statement of Company XYZ. This is the actual structure in which companies produce their Income Statement. The company name is masked due to privacy policy.

4.2. Descriptive analysis

This section describes the descriptive analytics performed on the dataset for the 25 selected companies. EPS (earnings per share) vs. year; ROCE (Return on Capital Employed) vs. year; Net Margin vs. year; Inventory Turnover Ratio vs. year; and visualization of company geolocations are the steps for descriptive analytics, as stated in Section 3.

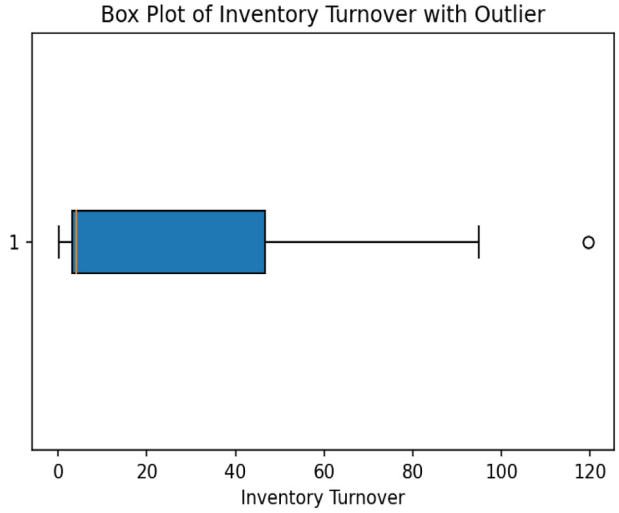

Fig. 2. Boxplot with outlier.

4.2.1. Outlier removal

The outlier must be eliminated from the dataset. We observed disparities in inventory turnover, debt-to-equity ratio, and net margin within our dataset. We utilized Inter quartile range method to eliminate these datasets. The steps required to eliminate such outliers are described next.

Inter quartile range (IQR) method:

Each dataset can be divided into quartiles.

Q1 = first quartile, i.e., 25% of the data points are below Q1
Q2 = Second quartile or Median, i.e., 50% of the data points are below Q2
Q3 = Third quartile, i.e., 75% of the data points are below Q3
$IQR = Q3 - Q1$
$Lower\ Limit(L) = Q1–1.5*IQR$
$Upper\ Limit(U) = Q3+1.5*IQR$

Any data point outside this range (L, U) is considered an outlier and should be removed for further analysis. Box plot can be used to visualize Outliers. Table 4 contains the mean, standard deviation, quartiles, and minimum and maximum value of Inventory Turnover with and without Outliers. We can see these parameters change when outliers are removed. Fig. 2 shows the boxplot of Inventory Turnover with an Outlier. In the box plot, we can see an outlier 120, which is indicated as a circle.

Fig. 3 shows the boxplot of Inventory Turnover without Outliers. We can see there is no outlier present. The Outlier 120 is removed

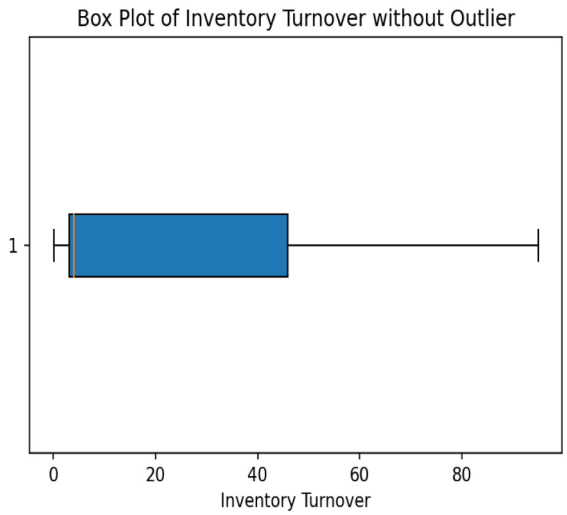

Fig. 3. Boxplot without Outlier.

Table 4
Inventory turnover with & without outlier.

Inventory turnover with outlier		Inventory turnover without outlier	
Count	91	Count	90
mean	23.493846	mean	22.427333
std	29.719353	std	28.080118
min	0.08	min	0.08
25%	3.1	25%	3.095
50%	4.03	50%	3.985
75%	46.585	75%	45.855
max	119.48	Max	94.89

here. Python programming is used here in the project to remove outliers using the IQR method.

4.2.2. Balance sheet plots

4.2.2.1. Finding the company with highest debt to equity ratio. It tells the degree to which a company is financing its operations through debt versus wholly owned funds.

Formula:

Debt to Equity = Total Debt/ Total Equity

Important Points-

• Ideally, D/E < 1, i.e., Only then will a company be in a comfortable position, so if the debt is less than your total equity capital, at least the company will have money to pay, and it will be in a comfortable position. In this case, it is more than 1. Hence, it is an uncomfortable position. A debt ratio of zero would tell that the firm does not finance its operations through borrowing at all.

• Tougher Times - Low D/E companies survive better; debt is such a liability that you have to pay if your profits are being eroded, then also you have to pay principal and interest every month. So, debt can also bankrupt the company.

• Compare companies in the Same Sector

• Higher D/E - E.g., Power Companies, Banks etc.

• Lower D/E - E.g., Retail, Software etc.

As depicted in Fig. 4, Company 6, 1 has a D/E ratio of zero for the year, indicating that it does not finance its operations through borrowing. The increasing D/E ratio of Company 11 over the past year is undesirable. Initially, the D/E ratio of Company 24 is approximately two, but it is decreasing over time, which is positive. The optimal D/E ratio for other companies fluctuates around one over the course of a year.

4.2.3. Income statement plots

4.2.3.1. Eps (basic) vs. year. Earnings per Share indicate a company's profitability or Financial Strength, and EPS is the part of companies Net Income or Profit distributed to each common Share during the year.

EPS(Basic) is calculated using the company's common shares. In contrast, diluted EPS considers all convertible securities, such as convertible preferred stock and convertible bonds, which can be converted into common stock or equity.

Formula:

Earnings Per Share = (Net Income - Preferred Dividend)/ Average Number of Shares Outstanding

Important Points->

• EPS of a company over the year should have a steady growth.

• The higher a company's EPS value, the more profitable it is.

• A negative EPS means the company has negative earnings or loses money.

Fig. 5 shows Company 24 having negative EPS means this company is at a loss. Company 9 has the highest EPS implies the company's shareholders will get the highest profit per Share. Company 1, 3, 15, 25 showing steady EPS over year, which is good.

4.2.4. Derived plots

4.2.4.1. Return on capital employed vs. year. Return on Capital Employed, a profitability ratio, tells how efficiently a company uses its capital to generate profits.

Formula:

ROCE = Earnings Before Interest & Tax (EBIT)/ Capital Employed
Capital Employed = Equity + Non-Current Liabilities

Important Points->

• A higher ROCE is always more favorable, indicating that more profits are generated per capital employed.

• At least 15%. A good ROCE value indicates the strength of the company. So, we can combinedly see the current ROCE and Average ROCE for 5 or 10 years to see how the company is doing.

• ROCE gives overall returns on the total capital employed in the business

• ROCE can be compared to returns from other investments, e.g., FD, MFs, Bonds etc. Invest if ROCE > Cost of Capital

Fig. 6 shows Company 1 is having highest ROCE, i.e., 20%–25% over the year. Most companies have ROCE values between 0 and 5%. Company 3, 11, 13 is having 15% ROCE initially, but it is decreasing over the year, which is not good.

4.2.4.2. Net margin vs. year. Net Profit Margin tell us How much profit the company is making on Total Revenue(or) Total Sales. In other words, what percentage of revenue is converted to profit.

We generally compare the net profit margin of different companies in the same sector/industry. (For example, the Paints Industry)

Formula:

Net Profit Margin = (Net Income/Total Revenue) *100
Net Income and Total Revenue are also called Net Worth and Total Sales.

Important Points->

• A high net profit margin means that a company is pricing its products correctly and is exercising good cost management.

• All operating and non-operating expenses deducted from Net Profit

• Gives an overall picture of a company's profitability

• Compare the Net Profit Margin of one company with other companies in the same sector.

• A negative profit margin is when your production expenses exceed your total revenue/sales for a specific period.

Fig. 7 shows Company 24 is having negative Net Margin means the production cost is more than total sales, which is terrible. Company 25 is having highest Net Margin, i.e., approx. 25%.

Most companies have Net Margins in the range of 5%–15%.

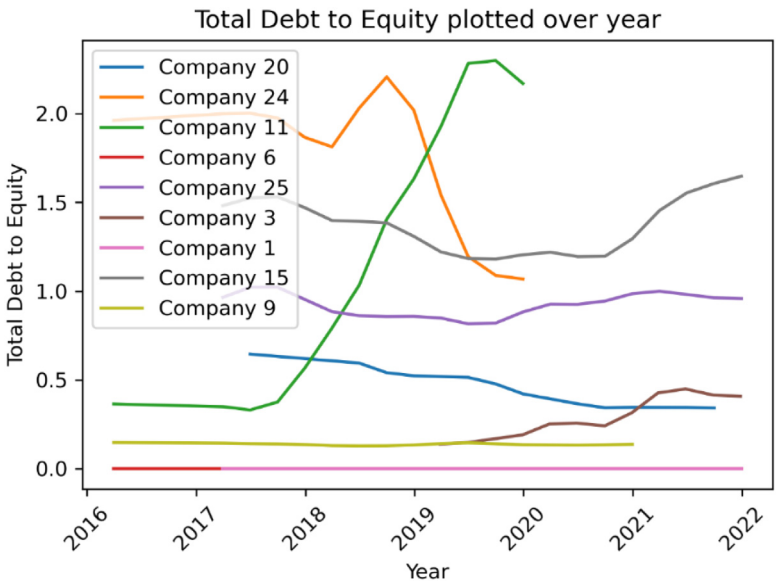

Fig. 4. Debt to equity vs. year.

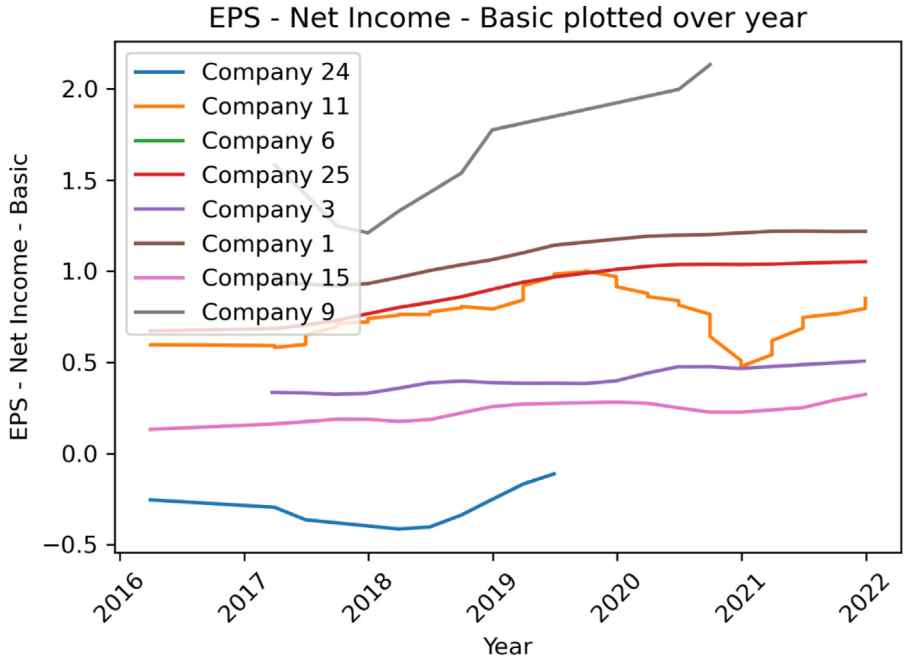

Fig. 5. EPS(basic) vs. year.

4.2.4.3. Inventory turnover ratio vs. year. The Inventory Turnover Ratio examines how quickly a company replaces its current Inventory and converts it into sales. A higher ratio means the company's product is in high demand and sells quickly, resulting in fewer inventory management costs and more profits.

Suppose the Inventory Turnover Ratio is two, then the company process the Inventory two times during the year. The Inventory processing means that you purchase the raw material, then work on it, which is the work-in-progress product, and then it becomes a final product, and then we sell it.

Formula:

Inventory Turn over Ratio Formula = Cost of Goods Sold/Average Inventory

Important Points->

• High the inventory turnover ratio is better.

• High Inventory Turnover Ratio — This indicates that the company has done an excellent job managing its Inventory, with lower holding costs and fewer obsolescence risks.

• Low Inventory Turnover Ratio — This implies that the company's products are not frequently sold in the market. Its Inventory becomes slow-moving, resulting in higher inventory costs and lower earnings.

Fig. 8 shows that most companies have an Inventory Turnover Ratio in the range of 2 to 5 means the company processes the Inventory two to five times during the year. Company 3 is having highest Inventory Turnover Ratio, which is good.

4.2.5. Geological visualization

In this section, for the chosen 25 companies, we first plotted distribution of companies across different country. Also, we visualized companies' locations using their longitude and latitude data on the world map.

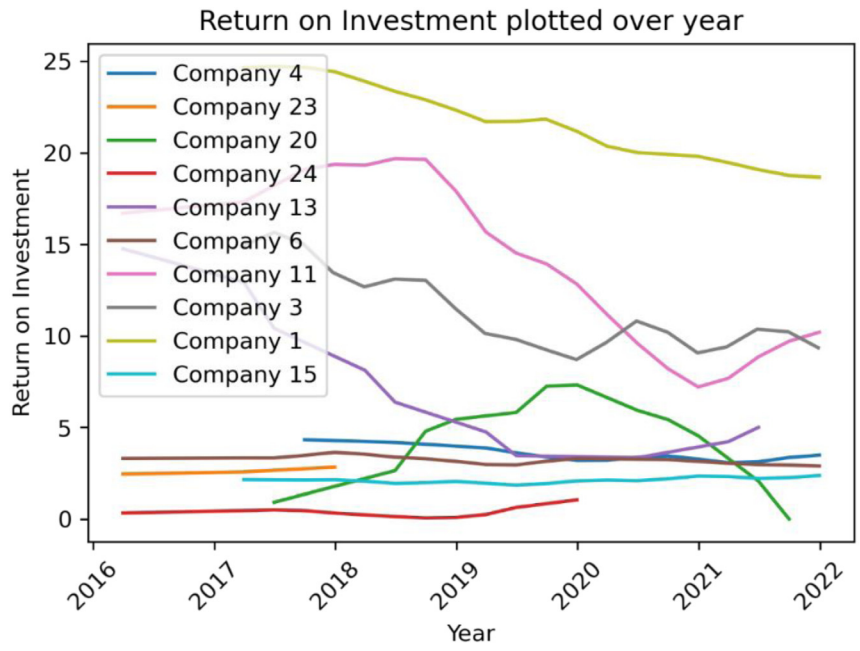

Fig. 6. ROCE vs. year.

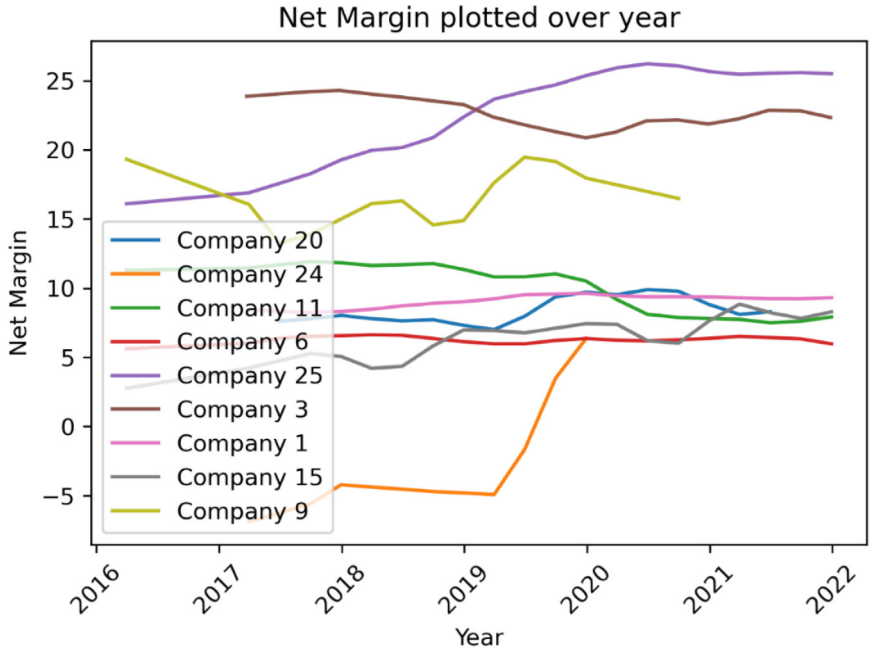

Fig. 7. Net margin vs. year.

Fig. 9 shows the distribution of companies across different countries, which tells us that more than 40% of companies are from Europe. Table 5 represents how many companies are present in which country. We can see most companies, i.e., 11 out of 25 companies, present in Europe. Fig. 10 visualizes companies' locations on the world map using latitude and longitude data. Latitude and longitude data are not given in the dataset, and we found those using Python API. Here we can see most of the companies are from Europe. Also, this thing is shown as a heatmap.

Table 5
Frequency table of geolocations.

Country	Count
Europe	11
Asia Pacific	4
North America	4
Latin America	2
Oceania	2
Middle East	1

4.3. Predictive analysis

Scikit-learn is an open-source Python library for machine learning. Scikit-learn is utilized here for Predictive Analysis. Machine Learning models typically reveal the relationship between independent and dependent features. Here, Net Income, Shareholders' Equity, Total Assets, and Total Liabilities serve as independent features, while Total Revenue

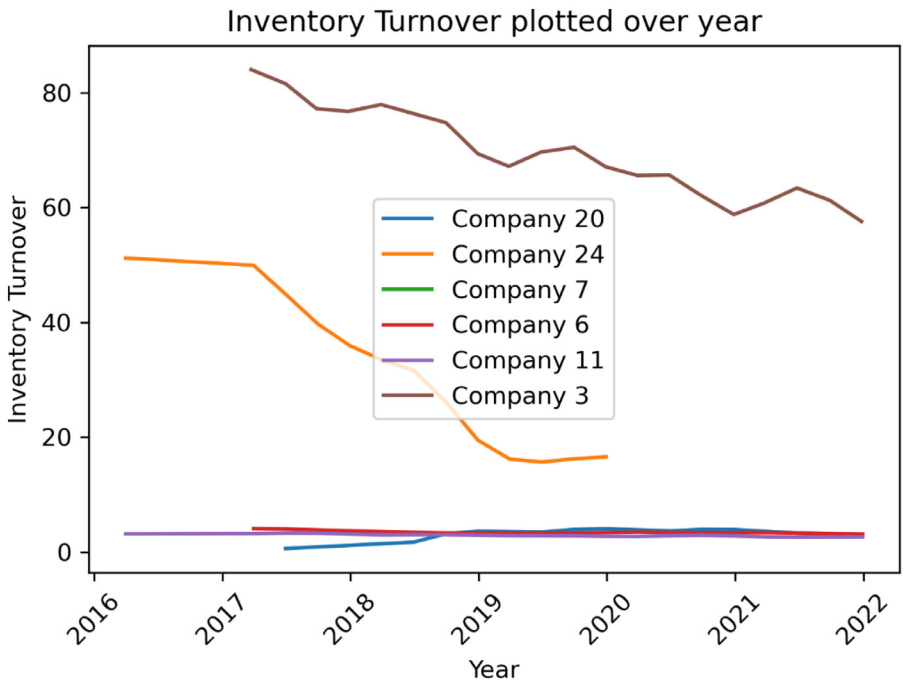

Fig. 8. Inventory turnover vs. year.

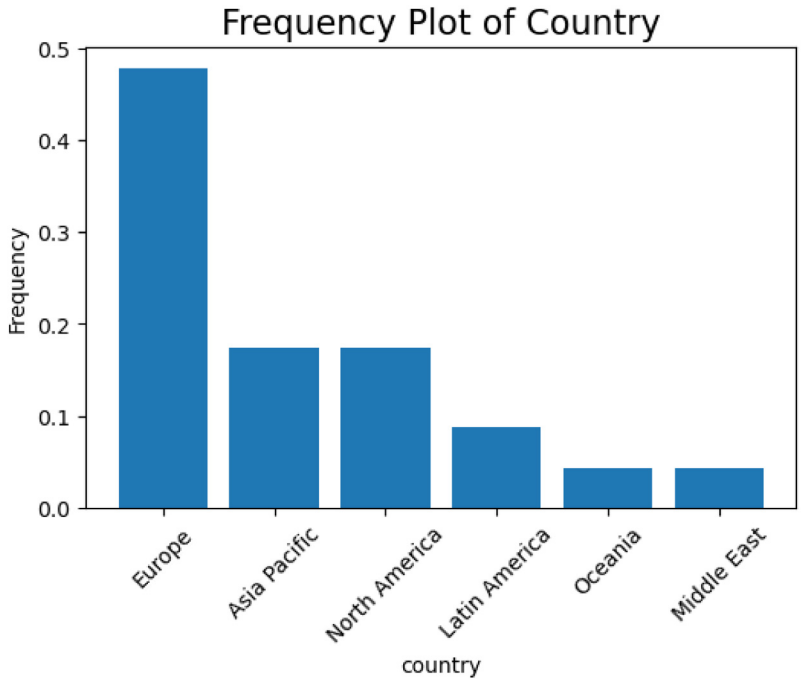

Fig. 9. Frequency plot of countries.

serves as the dependent or target feature. First, we train our model based on the values of Net Income, Shareholders' Equity, Total Assets, Total Liabilities, and the desired output Total Revenue. There are supervised machine learning models for every model. As previously mentioned, we employed four ML models: Linear Regression, KNN (K Nearest Neighbors), SVR (Support Vector Regression), and Decision Tree.

Data is preprocessed into the required format, i.e., all characteristics are in column format. A portion of these preprocessed data is displayed in Table 6.

4.3.1. Linear regression

It is the most basic machine learning prediction model. It finds a hyperplane in case of multi-dimensional data and a line in case of 2D data, which best fits our data.

It Minimizes the Sum of error of all points according to the hyperplane.

Mathematical Model of Linear Regression,

$$argmin_{(w,b)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = (w^T x_i + b)$$

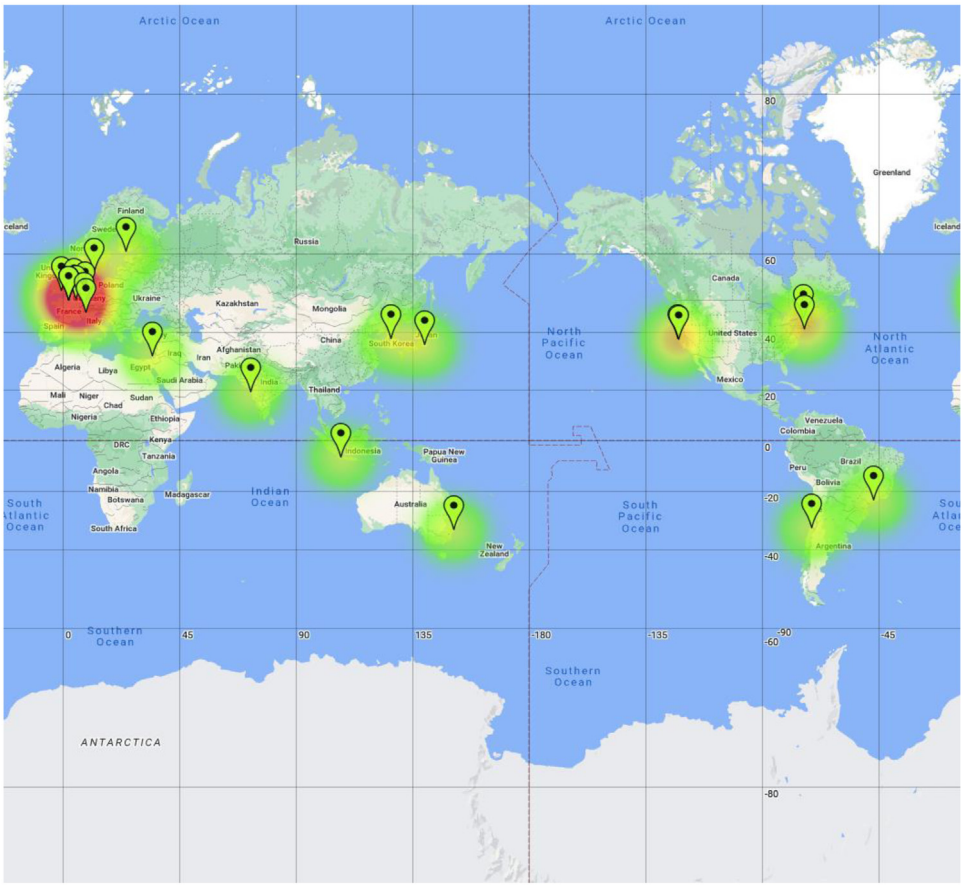

Fig. 10. Latitude & longitude plot on the world map.

Table 6
Glimpse of data on which model trained.

Net income	Stockholders' equity	Total assets	Total liabilities	Total revenue
−6.878000e+06	2.634630e+08	8.237970e+08	5.603340e+08	5.699900e+07
−5.371000e+06	2.634630e+08	8.237970e+08	5.603340e+08	2.304450e+08
8.099000e+07	4.872120e+08	1.232220e+09	7.450080e+08	7.383560e+08
2.905330e+08	4.872120e+08	1.232220e+09	7.450080e+08	2.734226e+09
5.395900e+10	2.650000e+12	3.980000e+12	1.340000e+12	1.070000e+12

$$argmin_{(w,b)} = \sum_{i=1}^{n} \left(y_i - \left(w^T x_i + b\right)\right)^2 + \lambda_1 \|w\| + \lambda_2 \|w\|^2$$

$Where, \lambda_1 \|w\| \Rightarrow for\ L1\ Regularization$

$\lambda_2 \|w\|^2 \Rightarrow for\ L2\ Regularization$

$\lambda_1\ \&\ \lambda_2\ are\ Hyperparameters$

$W\ is\ Normal\ Vector\ to\ the\ Hyperplane$

$b\ is\ Intercept\ of\ Hyperplane\ from\ Origin$

Linear regression is a distance-based ML model, so outliers significantly affect model performance. So, we need to remove the Outlier before training the model. Outlier is removed by calculating Inter Quartile Range (IQR) which is discussed in Section 4.2.1, i.e., everything outside IQR is removed. Then Pair plot is plotted to visually show any relation between any pair of independent features & Target feature.

From Fig. 11, we can see that Total Revenue and Net Income have a linear relation between them, i.e., when Net Income increases, then Total Revenue increases and vice versa. Also, it agrees with our hypothesis, which also says that revenue increases when income increases. From Fig. 12, we can see many points lies on $x = y$ line means most predicted and actual outputs are almost equal in magnitude, which is good. However, we can improve with other models as not all points are perfectly aligned on $x = y$ line.

Table 7
Performance table of LR.

Performance parameter	Value
MAE	1 413 009 905.091229
MSE	4.00763832762097e+18
R^2	0.7125066694210084

Then Standardization is performed on data to avoid the effect of different scales of different independent features. So, Standardization makes scales of all independent features into one scale. After Standardization average or mean of all features is converted to Zero, and the standard deviation is One. Then whole data is split into two parts, i.e., Train & Test Data. Train data is 80% of total data, and Test is the rest 20% of total data. This train test splitting is done using Sklearn Machine learning library in Python. Then the model is trained on train data then, tested using test data, and then the model's performance is calculated. Table 7 contains the value of performance metrics, i.e., MAE, MSE, and R squared for the LR model. From the above table, we can see R^2 value is approximately 0.71, which means $SS_R < SS_T$ means our Linear regression model is performing better than Simple Mean Model but not the best model as best R^2 value is 1. "model.coef" tells us coefficients for the linear regression, and "model.intercept" tells us to intercept

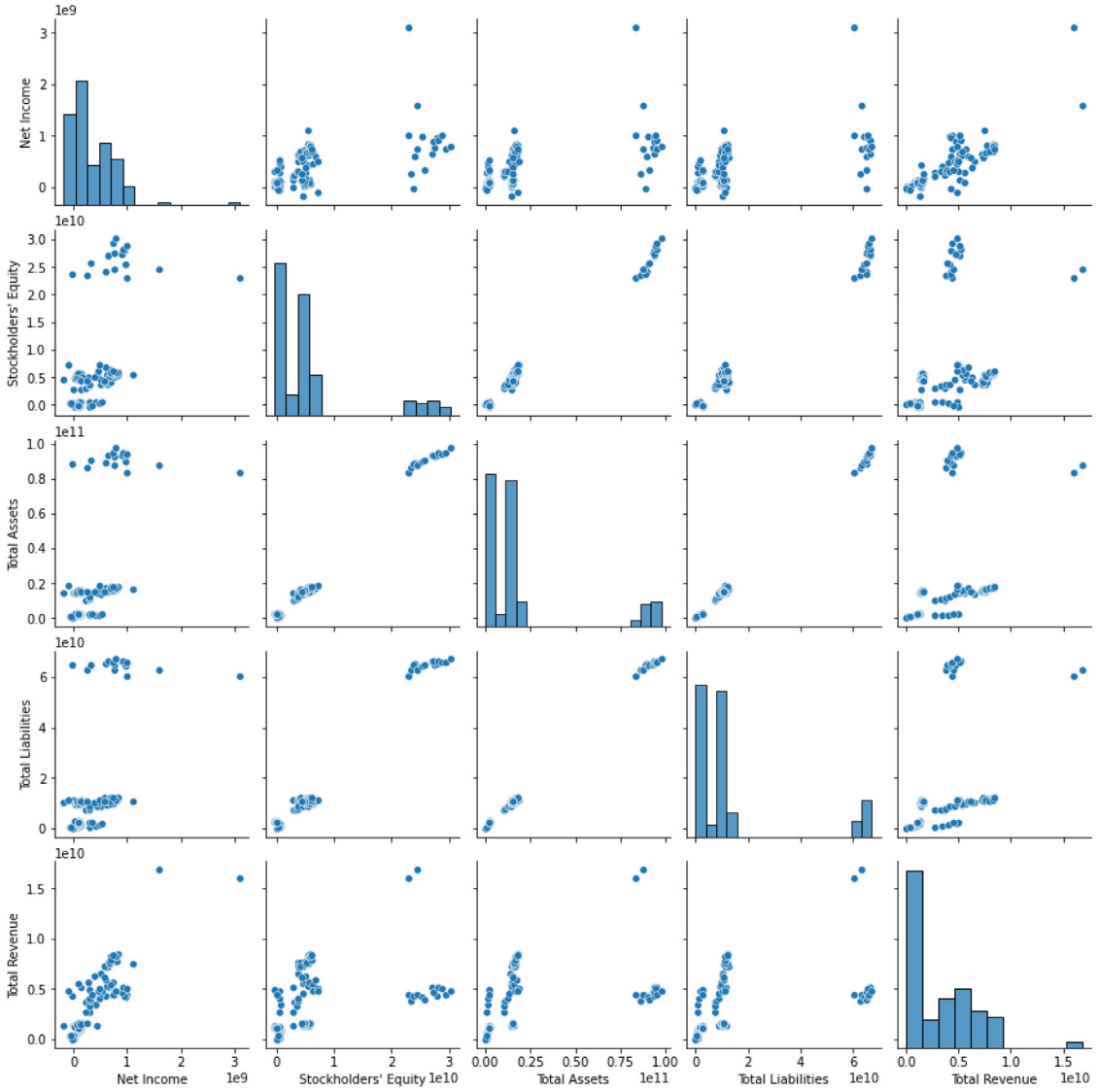

Fig. 11. Pair plot of features.

Table 8
Coefficient and intercept table of LR.

	Net income	Stockholders' equity	Total assets	Total liabilities
Coefficients	2.79557018e+09	8.23744319e+08	−3.84936879e+08	−8.75387743e+08
Intercept from origin		3 266 670 104.47		

hyperplane from the origin in the linear regression. Table 8 contains the coefficients and intercept value of the hyperplane. Our LR model fits this decision-making linear hyperplane.

4.3.2. KNN (K Nearest Neighbors)

KNN identifies the K nearest points to the test point, and then uses majority voting to determine or predict the test point's class. In the case of regression, it computes the mean of K nearest points and predicts this value as the output of the test data. K is a hyperparameter in KNN. If K is assumed to be one, then the model is overfit, i.e., it has a high variance. The outlier has a significant impact on the biased model. A single outlier can completely alter the performance of a model. The overfit model performs well with training data, but poorly with new data. If we assume K to be extremely large, the model underfits, i.e., it has a large bias. The Underfit model is unable to determine the relationship between the independent and dependent variables with precision. It performs poorly on both training and test data, with high error rates. Therefore, the ideal K value should neither be too low nor too high. Multiple K values are collected for model training, and the optimal K is observed. Thus, the K with the closest R-2 value to 1 is deemed optimal. As KNN is also a distance-based ML model, an outlier has a substantial impact on the performance of the model. Therefore, we must eliminate outliers prior to training the model. Calculating the inter quartile range (IQR) eliminates outliers, i.e., everything outside the IQR is eliminated. The entire data set is then divided into train data and test data. Eighty percent of all data consists of train data, while twenty percent consists of test data. This training test splitting is performed with the Python Sklearn Machine Learning library. Then, the data are normalized to eliminate the effect of different scales for different independent features. Thus, standardized measurement unifies the scales of all independent characteristics. After standardization, the mean or average of all characteristics is converted

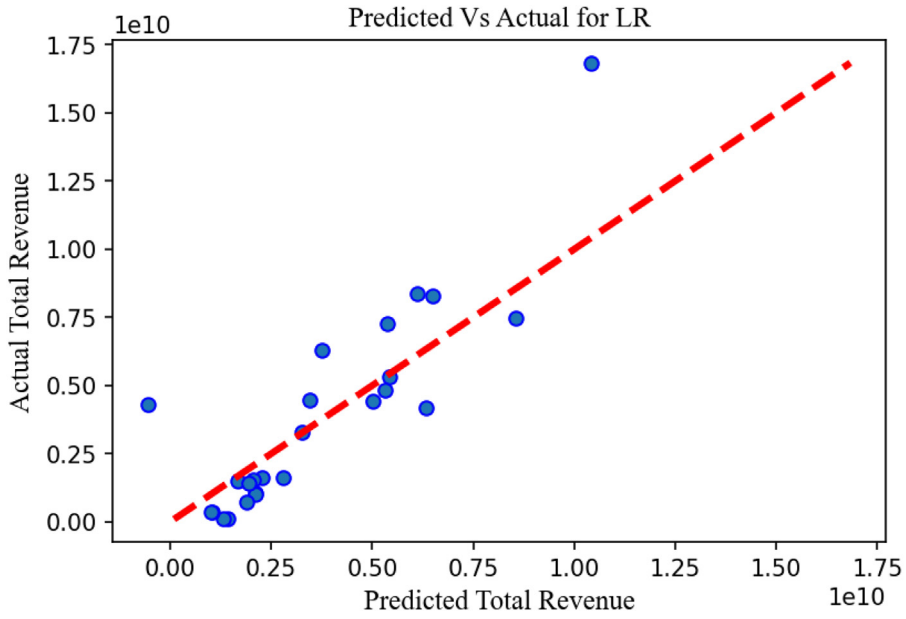

Fig. 12. Predicted vs. actual for LR.

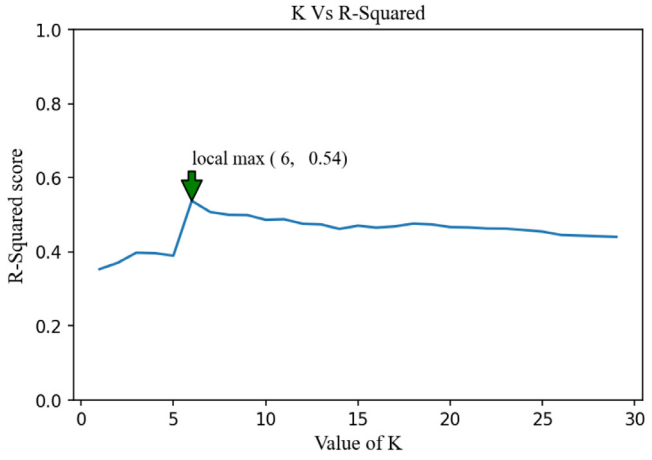

Fig. 13. K vs. R2 for KNN.

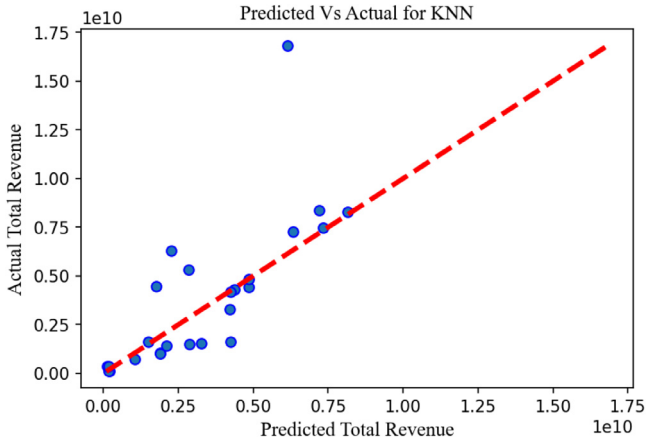

Fig. 14. Predicted vs. actual for KNN.

to zero and the standard deviation is converted to one. The data are then normalized to eliminate the effect of different scales for independent features. Normalization therefore converts the scales of all independent features to a single scale. Following normalization, the initial data is transformed into the unit hypercube. Both standardization and normalization perform similar functions, but we must determine which is superior for use in model training.

Without Normalization and Standardization, how the model is performed for different values of hyperparameter K is shown below. From Fig. 13, we can see that for K = 6; our model has the highest R^2 value of 0.54. So, $K = 6$ is taken for model training. From Fig. 14, we can see many points lie on $x = y$ line which means most predicted and actual outputs are almost equal in magnitude, which is good. Nevertheless, we can improve with other models as not all points are perfectly aligned on $x = y$ line.

With Standardization, how the model is performed for different values of hyperparameter K is shown below, From Fig. 15, we can see that for K = 19; our model has the highest R^2 value of 0.63. So, $K = 19$ is taken for model training. From Fig. 16, we can see many points lies on $x = y$ line means most predicted and actual outputs are almost equal in magnitude, which is good. However, we can improve with other models as not all points are perfectly aligned on $x = y$ line.

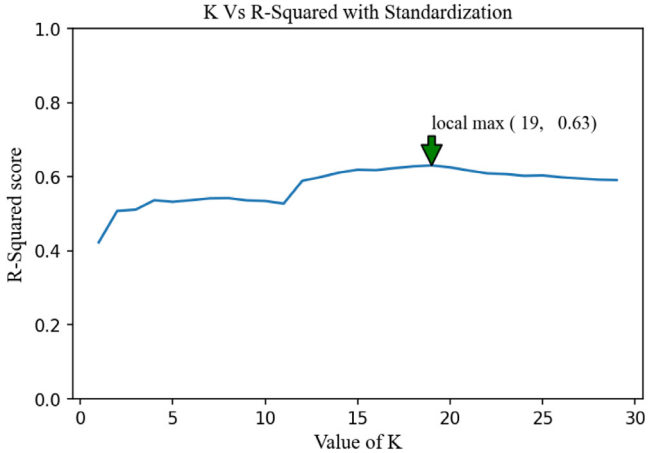

Fig. 15. K vs. R2 for KNN with standardization.

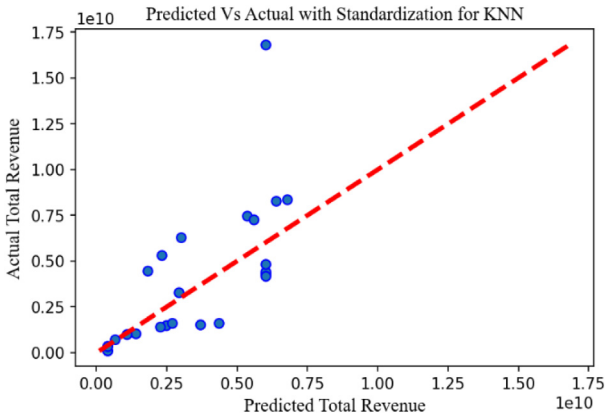

Fig. 16. Predicted vs. actual for KNN.

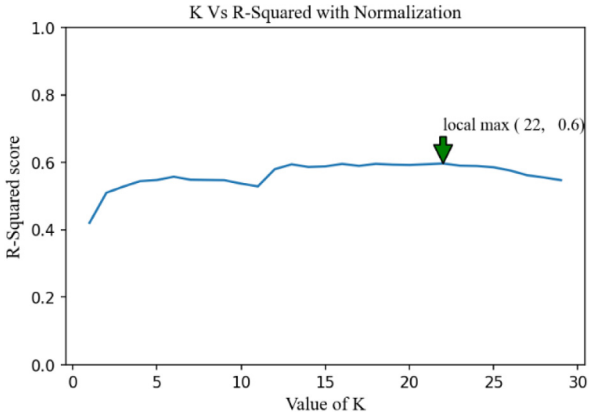

Fig. 17. K vs. R2 for KNN with normalization.

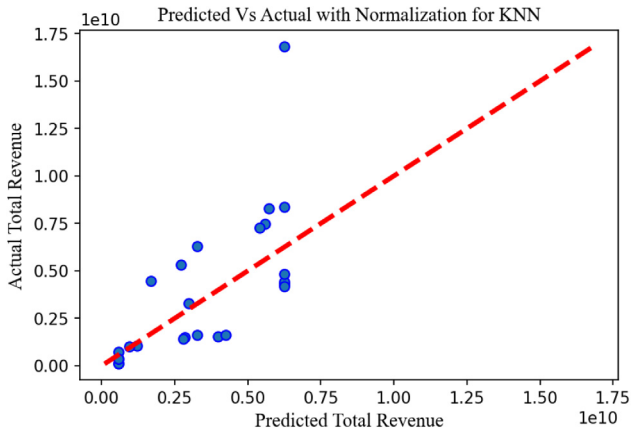

Fig. 18. Predicted vs. actual for KNN.

With Normalization, how the model is performed for different values of hyperparameter K is shown below, from Fig. 17, we can see that for K = 22; our model has the highest R^2 value of 0.597. So, $K = 22$ is taken for model training. From Fig. 18, we can see many points lies on $x = y$ line means most predicted and actual outputs are almost equal in magnitude, which is good. Nevertheless, we can improve with other models as not all points are perfectly aligned on $x = y$ line. From Table 9, we can see that KNN with Normalization is doing best among all three types of KNN models with a K value of 19. So KNN with Normalization is considered, compared with other models like Linear regression, SVR, and Decision Tree models. From Table 10, we can see

Table 9
KNN Model comparison with different pre-processing.

Model	Best K value	R^2 Value
KNN	6	0.54
KNN with normalization	19	0.63
KNN with standardization	22	0.597

Table 10
Performance table of KNN.

Performance parameter	Value
MAE	1 706 474 402.1052628
MSE	7.332226257164882e+18
R^2	0.63

R^2 value is approximately 0.63, which means $SS_R < SS_T$ means our Linear regression model is performing better than Simple Mean Model but not the best model as best R^2 value is 1.

4.3.3. SVR (Support Vector Regression)

Support Vector Machine finds a hyperplane so that the points are far, not too close to the hyper decision plane (π). In SVR, we take two more parallel hyperplanes (π^+, π^-) equidistance from the principal plane on both sides. All the positive & negative points should be either on π^+ or left of π^+ and either on π^- or right of π^- respectively. So, we need a plane (π^+, π, π^-) such that the margin is as high as possible. If the margin is low, there is an equal chance that a point belongs to any class.

The data is split into two parts, i.e., Train & Test Data. Train data is 80% of total data, and Test is the rest 20% of total data. This train test splitting is done using Sklearn Machine learning library in Python. Then the model is trained on train data then, tested using test data and then the model's performance is calculated. Then Standardization is performed on data to avoid the effect of different scales of different independent features. So, Standardization makes scales of all independent features into one scale. After Standardization average or mean of all features is converted to Zero, and the standard deviation is One. This model is robust to Outlier, so, no need to remove Outlier. That is why the model is trained using SVR without removing Outliers.

Mathematical Model of Support Vector Regression,

$$argmin_{(w,b)} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$Subject\ To: y_i - (w^T x_i + b) \le \xi$$

$$(w^T x_i + b) - y_i \le \xi$$

$$for\ all\ \xi \ge 0$$

$$W\ is\ Normal\ Vector\ to\ the\ Hyperplane$$

$$b\ is\ Intercept\ of\ Hyperplane\ from\ Origin$$

In SVR, C & ξ are hyperparameter. If we take C to be very high, then the model overfits, i.e., the model has high variance. Outlier highly impacts the high bias model. Single Outlier can completely change model performance. Overfit model performs well on Training data but performs poor for unseen data. If we take C to be very low, the model under fits, i.e., the model has a high bias. Underfit models cannot accurately find the relationship between the independent and dependent variables. It performs high errors on both train & test data. So Ideally C value should be not too low or not too high. Multiple C values are taken for model training, and which C is performing well is observed. i.e., the C, which has R^2 value nearer to 1 is considered best. ξ hyperparameter behaves exactly opposite to C, i.e., high ξ makes the model underfit, and low ξ makes the model overfit.

Our model is trained for different values of C & ξ

$$C = [10^0, 10^1, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}]$$

$$\xi = [10^0, 10^1, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}]$$

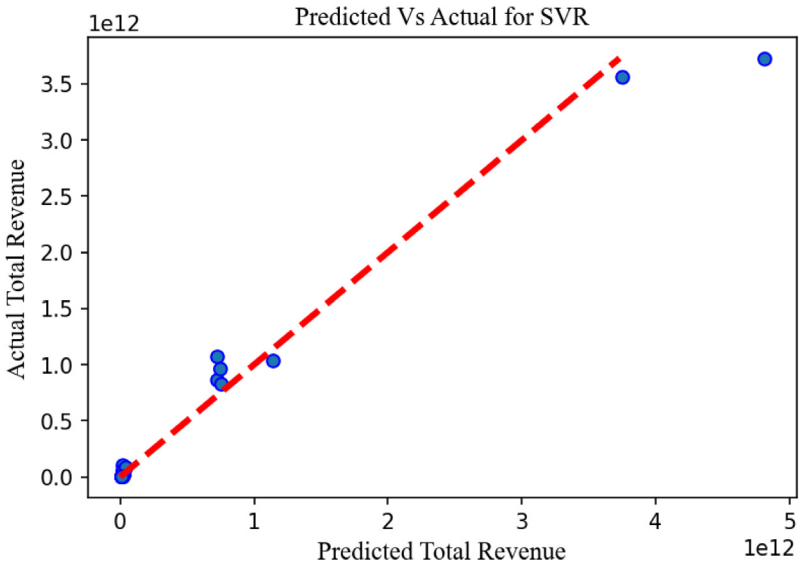

Fig. 19. Predicted vs. actual for SVR.

Table 11
Best hyperparameters for SVR.

Hyperparameter	Best value
C	10^9
ξ	10^{10}

Table 12
Performance table of SVR.

Performance parameter	Value
MAE	62 604 237 838.188194
MSE	3.320604842934192e+22
R^2	0.948

Table 13
Best hyperparameters for DT.

Hyperparameter	Best value
Maximum Depth of Tree	9

Table 14
Performance table of DT.

Performance parameter	Value
MAE	14 040 375 085.847776
MSE	1.2934393212586914e+21
R^2	0.9779765013123891

The most crucial idea in SVM is Kernel Trick. SVR uses different kernels such as RBF kernel, Polynomial kernel and domain-specific kernel for training data. By default, the RBF kernel is used. However, here, we used Polynomial Kernel.

Iterating over each pair of C & ξ we trained our model is performing best for below C & ξ values. Table 11 contains the best values of hyperparameters C and ξ for the SVR model, which we get by grid-search. From Table 12, we can see R^2 value is approximately 0.948, which means $SS_R < SS_T$ means our SVR model is performing better than Simple Mean Model but not the best model as best R^2 value is 1. From Fig. 19, we can see many points lie on the x = y line, which means most predicted and actual outputs are almost equal in magnitude, which is good.

4.3.4. Decision tree

The Decision Tree model is also called the Nested If Else model. If the model can be written as a nested if-else condition, this idea is called a decision tree. We can represent the model with the nested if-else condition in a tree diagram (as a tree structure). Every Tree has a single root node, multiple intermediate nodes & leaf node. Leaf nodes are present at the end of the Tree. Everything which is not root & leaf nodes are intermediate nodes.

In the decision tree, Information Gain is calculated for each independent feature, and whose information gain is maximum, that feature is first used to split the data. This is an iterative process. Also, to calculate information gain, we need another parameter calculated beforehand. That parameter is either Entropy or Gini Impurity. The data is split into two parts, i.e., Train & Test Data. Train data is 80% of total data, and Test is the rest 20% of total data. This train test splitting is done using Sklearn Machine learning library in Python. Then the model is trained on train data then, tested using test data and then the model's performance is calculated. This model is robust to Outlier, so, no need to remove Outlier. So, the model is trained using SVR without removing Outliers. In the Decision Tree, the Maximum Depth of the Tree is a hyperparameter. If we take Maximum Depth to be very high, then the model overfits, i.e., the model has high variance. Outlier highly impacts the high bias model. Single Outlier can completely change model performance. Overfit model performs well on Training data but performs poor for unseen data. If we take Maximum Depth to be very low, the model under fits, i.e., the model has high bias. Underfit model cannot accurately find the relationship between the independent and dependent variables. It performs high errors on both train & test data. So Ideally Maximum Depth value should be not too low or not too high. Multiple Maximum Depth values are taken for model training, and C is performing well. i.e., the Maximum Depth which has R^2 value nearer to 1 is considered best.

Iterating over Multiple Maximum Depth values, we trained our model is performing best for below Maximum Depth values. Table 13 contains the best value of the hyperparameter Maximum Depth of Tree for the DT model, which we get by linear search. From Table 14, we can see R^2 value is approximately 0.97, which means $SS_R < SS_T$ means our Decision Tree model is performing better than Simple Mean Model but not the best model as best R^2 value is 1. From Fig. 20, we can see many points lie on the x = y line, which means most predicted and actual outputs are almost equal in magnitude, which is good. Fig. 22 is the Decision Tree for our problem, generated using Python Programming Language. This is only a tiny portion of the big decision tree.

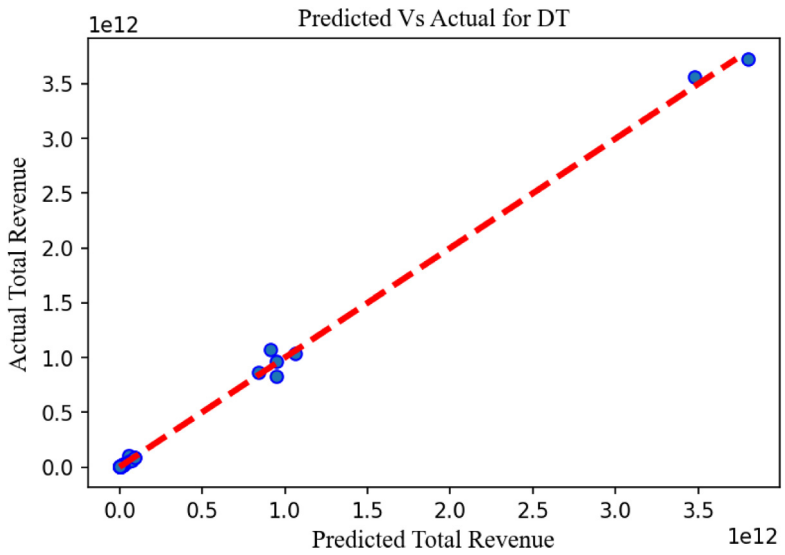

Fig. 20. Predicted vs. actual for DT.

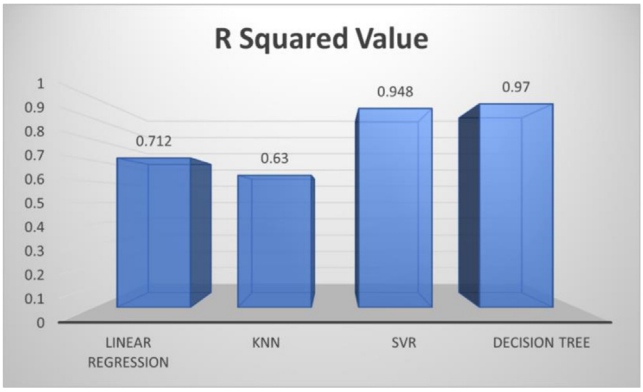

Fig. 21. Model comparison.

Table 15
Model comparison.

Model	R^2 Value	Hyperparameter value
Decision tree	0.97	Maximum Tree Depth = 9
SVR	0.948	C and $\xi = 10^9$ *and* 10^{10}
Logistic regression	0.712	None
KNN	0.63	K = 19

4.4. Prescriptive analysis

Out of all four models, i.e., (Linear Regression, KNN, SVR and Decision Tree) Decision Tree is performed best with R^2 value of 0.97. This decision tree model is having hyperparameter Maximum Tree Depth of 9. Then the second-best model is SVR with R^2 value of 0.948. This SVR model is having hyperparameters C and ξ of $10^9 and 10^{10}$ respectively. The third best model is Linear regression with R^2 value of 0.712. Then the worst model is KNN with R^2 value of 0.63. This KNN model has hyperparameter K of 19. Table 15 contains $\boldsymbol{R}^2$ Value of all models used along with their hyperparameter values in one place. From Fig. 21, we can see Decision Tree is having highest R^2 value, which means the target feature is perfectly explained with a minimum error by independent features.

5. Discussion and managerial insight

As seen from the previous studies on accounting, most of the researchers [18–20] had not done the descriptive analysis. However, we did the descriptive analysis in this paper and got the managerial insights for effective decision-making in heterogeneous firms.

The result highlighted that company 6 has a debt-to-equity ratio of zero. A debt-to-equity ratio of zero indicates that the company's operations are not financed by borrowing, which is desirable. Nonetheless, company 11 has the highest debt-to-equity ratio, and it has been rising over the past year, which is not a good sign. According to the EPS plot, company 9 has the highest EPS, while company 24 has the lowest EPS. EPS is proportional to a firm's profit; if profit is negative, so will EPS. Thus, company 24's negative EPS is due to its negative earnings, i.e., it is losing money. According to the ROCE plot, company 1 has an annual ROCE of between 20 and 25 percent. The reason for a company's high returns is its efficient use of capital to generate profits. Therefore, a higher ROCE is always preferable. According to the net margin plot, company 25 has the highest net margin at 25 percent.

The net margin represents the proportion of revenue that is converted into profit. Therefore, company 25 has the highest net margin because its production expenses are lower than its total revenue. According to the inventory turnover plot, company 3 has the highest inventory turnover. This is because a company processes its inventory numerous times throughout the year. Buying the raw materials and turning them into a work-in-progress product is part of inventory processing. The work-in-progress product is then turned into a finished product, which is then sold.

Jufrizen and PHS [22] finds the relation between CR, D/E ratio, NPM, and TAT on EPS. Sunjoko and Arilyn [21] finds the effect of fixed asset turnover, total asset turnover, inventory turnover, average collection period, and current ratio on profitability. Nariswari and Nugraha [20], finding the effects of NPM, GPM, and TAT. However, none of the above research finds a relation between Net Income, Stockholders' Equity, Total Assets, and Total Liabilities on a company's Total Revenue.

This paper used machine learning models to find a relationship between them. As we know, machine learning models train and perform well when the dataset is big, so a small-sized dataset is a limitation to our predictive study. The below paragraph summarizes the results we got in our predictive analysis.

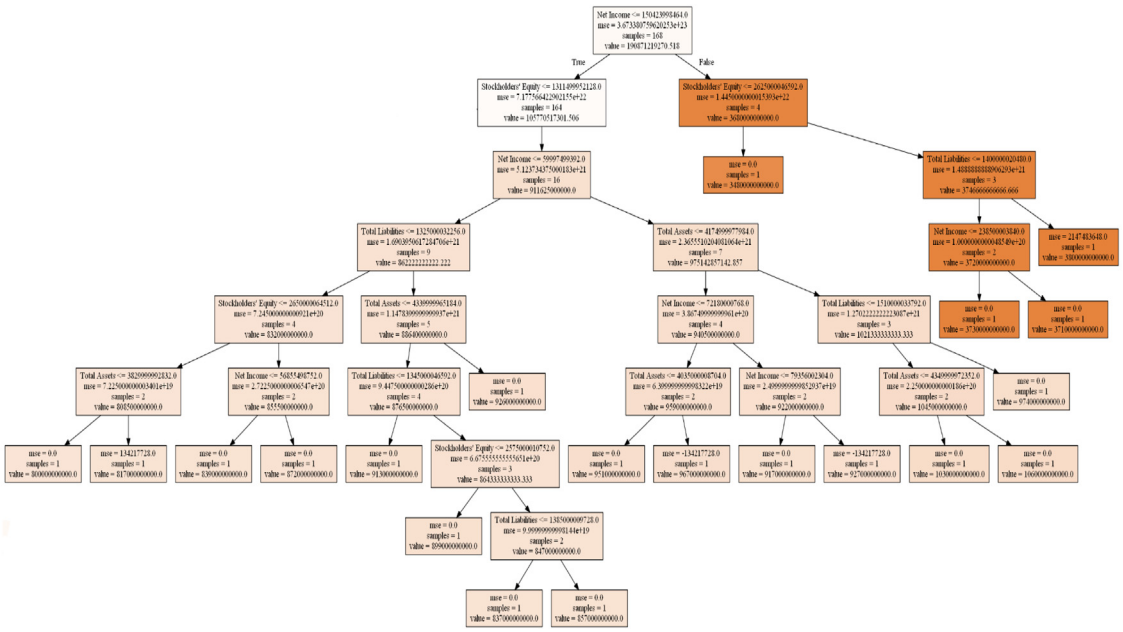

Fig. 22. Decision tree.

From predictive analysis, we learned that companies' total assets, total liabilities, net income, and stockholders' equity can be used to predict their total revenue. With an R^2 value of 0.97, we discovered that the decision tree is the most effective supervised machine learning model. This model has a maximum tree depth hyperparameter of 9. The worst model is then the KNN model, with an R2 value of 0.63 and a K value of 19. The performance of the decision tree model is satisfactory because the decision tree performs better when the dataset has a small number of dimensions, and our data has only four dimensions or features. Therefore, this is the reason for its success. Moreover, the KNN model performs well when the dataset has a high dimension. However, our dataset is small and has a low dimension, so KNN is not superior.

5.1. Managerial implications

This proposed paper may aid business owners and managers in making better data-driven, as opposed to intuitive, decisions. As a result of the fact that, in data-driven decision making, we examine vast amounts of data, analyze it to identify patterns, uncover valuable hidden insights, and use them to make business decisions. Moreover, we make intuitive decisions in the absence of data and sufficient time for analysis. Here, we rely solely on the knowledge and experience of the decision maker, which is still a risky basis for making decisions. It is difficult to obtain intuitive results, which we did using a data-driven approach, as described in Section 5.

Consequently, our data-driven analysis will assist managers in identifying trends in the critical financial parameters of a company, thereby enabling us to predict the company's future performance using a predictive model. A descriptive can help managers compare their companies' performance with that of their competitor companies with respect to revenue, operating cost, and profit. Also, companies in one sector can compare themselves with the market average using our study. Implementation of this study is easy. Managers need a small team of 3 to 4 people as financial risk analysts, data engineers, data scientists, and DevOps engineers. The initial cost of implementing may be a little bit costlier, but the return on investment we will be getting is worth investing in. Predictive analysis can help managers predict companies' revenue from assets, equity, liabilities, and income. Revenue is an important parameter to understand because it shows how much money the company makes by selling goods and services.

6. Conclusions and future research scope

To gain business insights, in this study financial accounting data is analyzed. Various financial metrics are provided to determine the relationship between financial characteristics using both descriptive and predictive analysis. For instance, debt to equity, earnings per share, inventory turnover, return on capital employed, and net margin are analyzed over the years to obtain significant insights about organizations characteristics. We discovered a correlation between a company's net income, stockholders' equity, total assets, and total liabilities in our prediction analysis. Unlike previous studies that have mostly performed empirical analysis on financial data, in our study an exhaustive investigation of financial accounting is done to not only determine hidden insights but also to enhance the reader's financial literacy. The three components of financial literacy are financial awareness, financial planning, and financial decision-making. Specifically, a company's balance sheet, income statement, and cash flow statement are reviewed. Using these parameters, the derived statement's other parameters are generated, including the Debt-to-Equity Ratio, Current Ratio, EPS (Basic), P/E Ratio, P/B Ratio, Return on Capital Employed, Net Margin, and Inventory Turnover Ratio. By assessing the aforementioned characteristics and charting them against time, we were able to establish which business is performing well.

The knowledge we receive about various firms enables us to assess whether companies are suitable for investment, whether it is beneficial to purchase a company's stock, what its business growth is, how well it has performed over the years, and what its profit is. In addition, there are other insights, such as how to analyze financial reports published by firms, the many techniques businesses use to generate income, how the company generates an asset, revenue, and profit from the cash it has raised, and how to calculate a company's operational expenses. Using the P/E ratio, it is possible to establish whether a company's shares are overpriced or underpriced in comparison to other companies in the same industry. What is the company's market value? And how do you evaluate businesses in the same industry? These insights enable

us to make better decisions. Four machine learning models, namely Linear Regression, KNN, SVR, and Decision Tree, were employed, with KNN, SVR, and DT having been infrequently employed by other researchers in the past. Therefore, this prediction model is highly useful for projecting the overall revenue of a corporation. For future work, we can analyze massive data sets containing hundreds of thousands of company records spanning 20 to 30 years. With this, we will also be able to analyze and predict real-time data and deliver real-time results and insights. We made predictions using a machine learning model, but deep learning and big data are equally feasible choices.

While we provided a systematic analysis of financial data from various sources, including the balance sheet, the income statement, and the cash flow statement, an important extension of the current study is to conduct financial risk analysis based on financial and non-financial indicators using Machine Learning Techniques. Financial risk analysis to identify and assess potential risks associated with an investment, allowing them to make informed decisions, thereby reducing the likelihood of financial losses. Another potential opportunity is to make use of data mining techniques for dynamic maintenance of time series data. This can help investors in identifying patterns and trends in financial data, which can be used for forecasting and early-warning mechanisms.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: SAURABH PRATAP reports was provided by Indian Institute of Technology BHU Varanasi. SAURABH PRATAP reports a relationship with Indian Institute of Technology BHU Varanasi that includes: employment.
NA

Data availability

The authors do not have permission to share data.

Appendix

A.1. Income statement format

Company XYZ income statement			
Standalone income statement	———————- in Rs. Cr. ———————-		
	Mar 22 12 mths	Mar 21 12 mths	Mar 20 12 mths
Income			
Interest/Discount on advances/Bills	98XXX	94XXX	91XXX
Income from investments	26XXX	23XXX	20XXX
Interest on balance with RBI and Other inter-bank funds	25XX	23XX	18XX
Others	6XX	4XX	5XX
Total interest earned	127XXX	120XXX	114XXX
Other income	29XXX	25XXX	23XXX
Total income	157XXX	146XXX	138XXX
Expenditure			
Interest expended	55XXX	55XXX	58XXX
Payments to and Provisions for employees	12XXX	10XXX	9XXX
Depreciation	0	13XX	11XX
Operating expenses (excludes Employee Cost & Depreciation)	25XXX	21XXX	19XXX
Total operating expenses	37XXX	32XXX	30XXX
Provision towards income Tax	12XXX	11XXX	98XX
Provision towards deferred tax	0	−11XX	5XX
Other provisions and contingencies	15XXX	15XXX	12XXX
Total provisions and contingencies	27XXX	26XXX	22XXX
Total expenditure	120XXX	114XXX	111XXX
Net Profit/Loss for the Year	36XXX	31XXX	26XXX
Net Profit/Loss after EI & Prior year items	36XXX	31XXX	26XXX
Profit/Loss brought forward	0	57XXX	49XXX
Total Profit/Loss available for appropriations	0	88XXX	75XXX

A.2. Balance statement format

Company XYZ balance sheet			
Standalone Balance Sheet	———————- in Rs. Crore ———————-		
	Mar-22 12 mths	Mar-21 12 mths	Mar-20 12 mths
Equities and liabilities			
Shareholder's funds			
Equity share capital	5XX	5XX	5XX
Total share capital	5XX	5XX	5XX
Reserves and surplus	239XXX	203XXX	170XXX
Total reserves and surplus	239XXX	203XXX	170XXX
Total shareholders funds	240XXX	203XXX	170XXX
Deposits	1559XXX	1335XXX	114XXX
Borrowings	184XXX	135XXX	144XXX
Other liabilities and provisions	84XXX	72XXX	67XXX
Total capital and liabilities	2068XXX	1742XXX	153XXX
Assets			
Cash and balances with reserve bank of India	129XXX	97XXX	72XXX
Balances with banks money at call and short notice	22XXX	22XXX	14XXX
Investments	455XXX	443XXX	391XXX
Advances	1368XXX	1132XXX	993XXX
Fixed assets	60XX	49XX	44XX
Other assets	85XXX	45XXX	53XXX
Total assets	2068XXX	1746XXX	153XXX

A.3. Cash flow statement format

Company XYZ cash flow statement		
Standalone cash flow statement	in Rs. Crore	
	Mar-22 12 mths	Mar-21 12 mths
Net Profit/Loss Before extraordinary items and tax	41XXX	36XXX
Net cashflow from operating activities	41XXX	−16XXX
Net cash used in investing activities	−1XXX	−1XXX
Net cash used from financing activities	−7XXX	22XXX
Foreign exchange gains/Losses	−1XX	2XX
Net Inc/Dec in cash and cash equivalents	32XXX	5XXX
Cash and cash equivalents begin of year	86XXX	81XXX
Cash and cash equivalents end of year	119XXX	86XXX

References

[1] A.A. Olayinka, Financial statement analysis as a tool for investment decisions and assessment of companies' performance, Int. J. Financ. Account. Manag. 4 (1) (2022) 49–66, http://dx.doi.org/10.35912/ijfam.v4i1.852.

[2] J.A. Nicholls, Integrating financial, social and environmental accounting, Sustain. Account. Manag. Policy J. 11 (4) (2020) 745–769, http://dx.doi.org/10.1108/SAMPJ-01-2019-0030.

[3] M.E. Barth, Financial accounting research, practice, and financial accountability, Abacus 51 (4) (2015) 499–510, http://dx.doi.org/10.1111/ABAC.12057.

[4] S. Guru, D.K. Mahalik, A comparative study on performance measurement of Indian public sector banks using AHP-TOPSIS and AHP-grey relational analysis, Opsearch 56 (4) (2019) 1213–1239.

[5] T. Didier, R. Levine, R. Llovet Montanes, S.L. Schmukler, Capital market financing and firm growth, J. Int. Money Finance 118 (2021) http://dx.doi.org/10.1016/j.jimonfin.2021.102459.

[6] P. Marsh, The choice between equity and debt: An empirical study, J. Finance 37 (1) (1982) 121, http://dx.doi.org/10.2307/2327121.

[7] L. Zhang, S. Zhang, Y. Guo, The effects of equity financing and debt financing on technological innovation: Evidence from developed countries, Balt. J. Manag. 14 (4) (2019) 698–715, http://dx.doi.org/10.1108/BJM-01-2019-0011.

[8] V. Bhatia, S. Basu, S.K. Mitra, P. Dash, A review of bank efficiency and productivity, Opsearch 55 (3–4) (2018) 557–600.

[9] M. Yousefinejad, T.A. Hakami, A. Abdulkareem, S. Al-Bazi, J. Othman, A. Alhadadi, The effect of risk management reporting on financial efficiency, in: Finance, Accounting and Business Analysis, Vol. 4, 2023, http://faba.bg.

[10] S. Rathore, Financial reporting and business communication, in: Communication Perspectives in Modern Businesses, 2022, p. 260.

[11] I. Martincevic, The correlation between digital technology and digital competitiveness, Int. J. Qual. Res. 16 (2) (2022).

[12] C. Shao, Y. Yang, S. Juneja, T. GSeetharam, IoT data visualization for business intelligence in corporate finance, Inf. Process. Manage. 59 (1) (2022) http://dx.doi.org/10.1016/j.ipm.2021.102736.

[13] P. Mikalef, J. Krogstie, I.O. Pappas, P. Pavlou, Exploring the relationship between big data analytics capability and competitive performance: The mediating roles of dynamic and operational capabilities, Inf. Manage. 57 (2) (2020) 103169, http://dx.doi.org/10.1016/J.IM.2019.05.004.

[14] N. Sayedahmed, S. Anwar, V.K. Shukla, Big data analytics and internal auditing: A review, in: Proceedings - 2022 3rd International Conference on Computation, Automation and Knowledge Management, ICCAKM 2022, 2022, http://dx.doi.org/10.1109/ICCAKM54721.2022.9990045.

[15] S.K. Jena, Impact of customer-centric approach and customer dissatisfying cost on supply chain profit under price competition, J. Bus. Ind. Mark. (2023) http://dx.doi.org/10.1108/JBIM-02-2022-0111.

[16] H.T. Tseng, Customer-centered data power: Sensing and responding capability in big data analytics, J. Bus. Res. 158 (2023) 113689, http://dx.doi.org/10.1016/J.JBUSRES.2023.113689.

[17] D. Kalaitzi, N. Tsolakis, Supply chain analytics adoption: Determinants and impacts on organisational performance and competitive advantage, Int. J. Prod. Econ. 248 (2022) 108466, http://dx.doi.org/10.1016/J.IJPE.2022.108466.

[18] M. Khaddafi, M. Heikal, A. Ummah, Influence analysis of Return on Assets (ROA), Return on Equity (ROE), Net Profit Margin (NPM), Debt to Equity Ratio (DER), and current ratio (CR), against corporate profit growth in automotive in Indonesia stock exchange, Int. J. Acad. Res. Bus. Soc. Sci. 4 (12) (2014) http://dx.doi.org/10.6007/IJARBSS/v4-i12/1331.

[19] Y. Nuryani, D. Sunarsi, The effect of current ratio and debt to equity ratio on deviding growth, JASa (J. Akunt. Audit Sist. Inf. Akunt.) 4 (2) (2020) 304–312, http://dx.doi.org/10.36555/JASA.V4I2.1378.

[20] T.N. Nariswari, N.M. Nugraha, Profit growth : Impact of net profit margin, gross profit margin and total assests turnover, Int. J. Finance Bank. Stud. 9 (4) (2020) 87–96, http://dx.doi.org/10.20525/IJFBS.V9I4.937, (2147-4486).

[21] M.I. Sunjoko, E.J. Arilyn, Effects of inventory turnover, total asset turnover, fixed asset turnover, current ratio and average collection period on profitability, J. Bisnis Akunt. 18 (1) (2016) 79–83, http://dx.doi.org/10.34208/JBA.V18I1.40.

[22] D. Jufrizen, D. PHS, Effect of current ratio, debt to equity ratio, net profit margin, and total asset turnover on earning per share, in: International Conference on Global Education, 2019, Retrieved May 5, 2022, from https://ojs.pasca.unespadang.ac.id/index.php/ICGE/article/view/55.

[23] C. Kothari, W.J. O'Brien, N. Khwaja, Transportation programs under financial uncertainty: Identification of needs for future research, Transp. Res. Interdiscip. Perspect. 16 (2022) 100684.

[24] Z. Bánhidi, I. Dobos, A Data Envelopment Analysis model for ranking digital development in the countries of the European Union without explicit inputs and common weights analysis, Decis. Anal. J. (2023) 100167.

[25] F. Özdemirci, S. Yüksel, H. Dinçer, S. Eti, An assessment of alternative social banking systems using T-Spherical fuzzy TOP-DEMATEL approach, Decis. Anal. J. (2023) 100184.

[26] K. Moise Ikegwu, Machine Learning and Information-theoretic Approaches for Financial Applications (Doctoral dissertation), 2021.

[27] F. Lv, W. Wang, L. Han, D. Wang, Y. Pei, J. Huang …, M. Pechenizkiy, Mining trading patterns of pyramid schemes from financial time series data, Future Gener. Comput. Syst. 134 (2022) 388–398.

[28] Z. Lv, N. Wang, X. Ma, Y. Sun, Y. Meng, Y. Tian, Evaluation standards of intelligent technology based on financial alternative data, J. Innov. Knowl. 7 (4) (2022) 100229.

[29] M. Sultan Almansoori, M. Hadef Almansoori, M. Mohamed Almansoori, A. Rashed Almansoori, A. Abdulla Alhammadi, S. Mubarak Alnuaimi, Financial Analysis of ADNOC Supervised By: Professor Haitham Nobanee. https://ssrn.com/abstract=3895246.

[30] P. Chhajer, M. Shah, A. Kshirsagar, The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction, Decis. Anal. J. 2 (2022) 100015.

[31] D. Bhattacharjee, K. Ramesh, E.S. Jayaram, M.S. Mathad, D. Puhan, An integrated machine learning and DEMATEL approach for feature preference and purchase intention modeling, Decis. Anal. J. (2023) 100171.

[32] J.K. Afriyie, K. Tawiah, W.A. Pels, S. Addai-Henne, H.A. Dwamena, E.O. Owiredu ., J. Eshun, A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions, Decis. Anal. J. 6 (2023) 100163.

[33] C.F. Moreno-Garcia, C. Jayne, E. Elyan, M. Aceves-Martins, A novel application of machine learning and zero-shot classification methods for automated abstract screening in systematic reviews, Decis. Anal. J. (2023) 100162.

[34] S. Perla, R. Bisoi, P.K. Dash, A hybrid neural network and optimization algorithm for forecasting and trend detection of forex market indices, Decis. Anal. J. (2023) 100193.

[35] D. Pagano, A predictive maintenance model using Long Short-Term Memory Neural Networks and Bayesian inference, Decis. Anal. J. 6 (2023) 100174.

[36] S.K.S. Durai, M.D. Shamili, Smart farming using machine learning and deep learning techniques, Decis. Anal. J. 3 (2022) 100041.

[37] Z. Hassan, W. Khreich, I.H. Osman, An international social inclusion index with application in the organization for economic co-operation and development countries, Decis. Anal. J. 3 (2022) 100047.

[38] D. Das, P. Kayal, M. Maiti, A K-means clustering model for analyzing the Bitcoin extreme value returns, Decis. Anal. J. 6 (2023) 100152.

[39] C. Sanchís-Pedregosa, M.J. Palacín-Sánchez, M.D.M. González-Zamora, Exploring the financial impact of outsourcing services strategy on manufacturing firms, Oper. Manag. Res. 7 (2014) 77–85.

[40] G.P. Schneider, J. Dai, D.J. Janvrin, K. Ajayi, R.L. Raschke, Infer, predict, and assure: Accounting opportunities in data analytics, Account. Horiz. 29 (3) (2015) 719–742.

[41] M. Oberholzer, A non-parametric comparison among firms income statement-based and balance sheet-based performance, Int. Bus. Econ. Res. J. 12 (11) (2013) 1467–1478.

[42] J.Y. Xin, A.C. Yeung, T.C.E. Cheng, Radical innovations in new product development and their financial performance implications: An event study of US manufacturing firms, Oper. Manag. Res. 1 (2008) 119–128.

[43] G. Dutta, H.V. Rao, S. Basu, M.K. Tiwari, Asset liability management model with decision support system for life insurance companies: Computational results, Comput. Ind. Eng. 128 (2019) 985–998, http://dx.doi.org/10.1016/J.CIE.2018.06.033.

[44] V.P. Ramesh, P. Baskaran, A. Krishnamoorthy, D. Damodaran, P. Sadasivam, Back propagation neural network based big data analytics for a stock market challenge, Commun. Stat. - Theory Methods 48 (14) (2019) 3622–3642.

[45] G. Sismanoglu, M.A. Onde, F. Kocer, O.K. Sahingoz, Deep learning based forecasting in stock market with big data analytics, in: 2019 Scientific Meeting on Electrical-electronics & Biomedical Engineering and Computer Science (EBBT), IEEE, 2019, pp. 1–4.

[46] L.K. Shrivastav, R. Kumar, An ensemble of random forest gradient boosting machine and deep learning methods for stock price prediction, J. Inf. Technol. Res. 15 (1) (2021) 1–19, http://dx.doi.org/10.4018/jitr.2022010102.

[47] M.K. Daradkeh, A hybrid data analytics framework with sentiment convergence and multi-feature fusion for stock trend prediction, Electronics (Switzerland) 11 (2) (2022) http://dx.doi.org/10.3390/electronics11020250.

[48] M.A. Boyacioglu, D. Avci, An adaptive network-based fuzzy inference system (ANFIS) for the prediction of stock market return: the case of the Istanbul stock exchange, Expert Syst. Appl. 37 (12) (2010) 7908–7912.

[49] K.J. Morris, S.D. Egan, J.L. Linsangan, C.K. Leung, A. Cuzzocrea, C.S. Hoi, Token-based adaptive time-series prediction by ensembling linear and non-linear estimators: a machine learning approach for predictive analytics on big stock data, in: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2018, pp. 1486–1491.

[50] C.S. Lin, L.F. Chen, C.T. Su, T.C. Kon, Stock price impact on the Taiwan national quality award, Total Qual. Manag. Bus. Excell. 24 (1–2) (2013) 158–170.

[51] J.H. Wang, J.Y. Leu, Stock market trend prediction using ARIMA-based neural networks, in: Proceedings of International Conference on Neural Networks (ICNN'96), Vol. 4, IEEE, 1996, pp. 2160–2165.

[52] M. Ananthi, K. Vijayakumar, Retracted article: stock market analysis using candlestick regression and market trend prediction (CKRM), J. Ambient Intell. Humaniz. Comput. 12 (5) (2021) 4819–4826.
[53] Q. Liu, Z. Tao, Y. Tse, C. Wang, Stock market prediction with deep learning: The case of China, Final. Res. Lett. 46 (2022) 102209.
[54] Q. He, J. He, Analysis of the path to solve the rural financing problem based on agricultural big Data+ Financial Technology, Optik (2022) 170340.
[55] B.N. Iyke, S.Y. Ho, Stock return predictability over four centuries: The role of commodity returns, Final. Res. Lett. 40 (2021) 101711.
[56] P. Ghosh, A. Neufeld, J.K. Sahoo, Forecasting directional movements of stock prices for intraday trading using LSTM and random forests, Final. Res. Lett. 46 (2022) 102280.
[57] M. Leippold, Q. Wang, W. Zhou, Machine learning in the Chinese stock market, J. Financial Econ. 145 (2) (2022) 64–82.
[58] H. Berkman, V.R. Eleswarapu, Short-term traders and liquidity: a test using Bombay Stock Exchange data, J. Financial Econ. 47 (3) (1998) 339–355.
[59] E. Amin, Financial fraud detection of the Egyptian companies annual reports using artificial bee colony algorithm, Int. J. Bus. Data Anal. 1 (2) (2019).
[60] J. Behera, A.K. Pasayat, H. Behera, P. Kumar, Prediction based mean-value-at-risk portfolio optimization using machine learning regression algorithms for multi-national stock markets, Eng. Appl. Artif. Intell. 120 (2023) 105843.
[61] F. Olowokudejo, O.E. Akindipe, An exploratory study of asset liability management in the insurance industry in Nigeria: A panel approach, Acta Universitatis Danubius, Economica 18 (4) (2022).
[62] N. Dospinescu, O. Dospinescu, Title: A profitability regression model in financial communication of Romanian stock exchange's companies a profitability regression model in financial communication of Romanian stock exchange's companies citation style: A profitability regression model in financial communication of Romanian stock exchange's companies, Ecoforum 8 (1) (2019) 1:0-0.
[63] X. Yu, Y. Zhou, X. Liu, Impact of financial development on energy consumption in China: A spatial spillover analysis, Energy Strategy Rev. 44 (2022) 100975.
[64] X. Chen, X. Ma, H. Wang, X. Li, C. Zhang, A hierarchical attention network for stock prediction based on attentive multi-view news learning, Neurocomputing 504 (2022) 1–15.
[65] J. Leng, W. Liu, Q. Guo, Stock movement prediction model based on gated orthogonal recurrent units, Intell. Syst. Appl. 16 (2022) 200156.
[66] C. Wang, Y. Chen, S. Zhang, Q. Zhang, Stock market index prediction using deep transformer model, Expert Syst. Appl. 208 (2022) 118128.
[67] X. Teng, X. Zhang, Z. Luo, Multi-scale local cues and hierarchical attention-based LSTM for stock price trend prediction, Neurocomputing 505 (2022) 92–100.
[68] S. Nielsen, Management accounting and the concepts of exploratory data analysis and unsupervised machine learning: a literature study and future directions, J. Account. Organ. Chang. (2022).
[69] X. Yang, M.A. Loua, M. Wu, L. Huang, Q. Gao, Multi-granularity stock prediction with sequential three-way decisions, Inform. Sci. 621 (2023) 524–544.
[70] D.K. Tripathi, S. Chadha, A. Tripathi, Metaheuristic enabled intelligent model for stock market prediction via integrating volatility spillover: India and its Asian and European counterparts, Data Knowl. Eng. 144 (2023) 102127.
[71] S. Rakshit, B. Aslani, E. Rajah, N.R. Vajjhala, Exploring the role of artificial intelligence and machine learning for financial intelligence in Nigeria, in: Proceedings of the International Conference on Innovative Computing & Communication (ICICC), 2021.
[72] E. Pechlivanidis, D. Ginoglou, P. Barmpoutis, Can intangible assets predict future performance? A deep learning approach, Int. J. Account. Inf. Manag. 30 (1) (2022) 61–72.
[73] Y. Zhao, G. Yang, Deep learning-based integrated framework for stock price movement prediction, Appl. Soft Comput. 133 (2023) 109921.
[74] M. Heikal, M. Khaddafi, A. Ummah, Influence analysis of return on assets (ROA), return on equity (ROE), net profit margin (NPM), debt to equity ratio (DER), and current ratio (CR), against corporate profit growth in automotive in Indonesia Stock Exchange, Int. J. Acad. Res. Bus. Soc. Sci. 4 (12) (2014) 101.