Data Analysis Approach to Predict Revenue using Financial Statement.

Drumil Adbhai
Department of Computer Engineering
D. Y. Patil College of Engineering, Akurdi
Pune, India
mildru9900@gmail.com

Mr. Vaibhav R. Chavan
(Assistant Professor)
Department of Computer Engineering
D. Y. Patil College of Engineering, Akurdi
vrchavan@dypcoeakurdi.ac.in

Abstract **— Accurate revenue prediction plays a vital role in strategic financial planning, yet many existing systems rely on rigid, traditional statistical methods that fail to effectively capture complex patterns within financial statement data. These systems often lack the flexibility to handle non-linear relationships and multivariate financial indicators, leading to suboptimal forecasting. This research introduces a modern, data-driven solution by applying** Exploratory Data Analysis (EDA) **to uncover hidden trends and outliers in financial datasets. Subsequently, two machine learning models—**Linear Regression **and** K-Nearest Neighbor (KNN)**—are implemented and compared to evaluate their performance in predicting revenue. The proposed system aims to improve accuracy, adaptability, and practical usefulness in real-world financial decision-making scenarios.**

Keywords— Revenue Prediction, Financial Statement Analysis, EDA, Linear Regression, K-Nearest Neighbour, MSE, Predictive Analytics, Financial Forecasting.

I. INTRODUCTION

This research paper investigates the application of data-driven techniques for predicting revenue using financial statements, aiming to improve the accuracy of financial forecasting models. Revenue prediction plays a pivotal role in strategic business planning and decision-making, yet remains a complex task due to the heterogeneous and high-dimensional nature of financial data. Traditional statistical methods often fail to capture intricate, non-linear relationships between financial indicators and revenue outcomes.The proposed methodology integrates Exploratory Data Analysis (EDA) for uncovering hidden patterns and anomalies within financial records, followed by the implementation of machine learning algorithms—Linear Regression and K-Nearest Neighbour (KNN). These models are evaluated on their ability to predict revenue effectively by minimizing prediction error using metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).The objectives of this study are threefold: (1) to analyze financial statement data for relevant patterns and relationships, (2) to build predictive models that leverage machine learning for improved accuracy, and (3) to compare the performance of Linear Regression and KNN in terms of predictive reliability, computational efficiency, and adaptability to real-world business scenarios.This approach not only facilitates more accurate revenue forecasting but also promotes data-driven decision-making in financial planning. In doing so, the research contributes to sustainable business growth, investor transparency, and better resource allocation, particularly supporting small and medium-sized enterprises (SMEs) in competitive markets.

II. MATERIALS AND METHODS

A] Financial Accounting

1.Debt-to-equity:

The Debt-to-Equity Ratio (DER) compares a company's debt and equity. Financial analysts consider it one of the most essential capital structure measures for a company's value. The DER ratio measures a company's ability to cover its debt repayment obligations with its own capital. This ratio has a significant impact on earnings changes, which are intended to benefit the company's profits.

2.Earnings per Share (EPS):

EPS is a key piece of information for shareholders, indicating the amount of money earned for each share of a company. It's an important metric for investors because it shows the company's profitability.

3.Net Income Margin:

This is a profitability ratio that compares net income to sales. It can also be used to understand the company's efficiency in managing operational costs during a specific period. This ratio assesses the overall efficiency of a company's operations, including manufacturing, personnel, marketing, and finance.

B]Dataset:

Financial statements used for this study were obtained from stockbit.com. The data was collected quarterly from 21 Indonesian companies listed on the Indonesia Stock Exchange (IDX) from 2008 to 2023. The dataset contains 1344 lines of financial data, including debt-to-equity, earnings per share, and net income margin.

C] Exploratory Data Analysis:

It is a crucial first step before using statistical or machine learning models to discover outliers, missing values, data distribution, and variable relationships. **Boxplots** are used to analyze and depict the distribution of profitability indicators. The Interquartile Range (IQR) is used to detect extreme data points (outliers) within the dataset. EDA also involves examining profitability indicators over time, such as Debt-to-Equity vs. Year, EPS vs. Year, and Net Income Margin vs. Year.

D] Data Splitting:

Data splitting is required to develop machine learning models and ensure a balanced performance evaluation. The dataset will be divided into training and testing sets in an **80:20 ratio**. This results in 1075 training sets and 269 test sets from the total 1344 rows of data. The split is performed using Python's Scikit-learn machine learning tool.

E] RELEVANT MATHEMATICS or ALGORITHM

1. Linear Regression

Linear Regression models the relationship between a dependent variable (Revenue) and one or more independent financial variables (e.g., Net Income, Total Assets).

Prediction Formula:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Where:

- $\hat{y}$ = Predicted revenue
- $x_1, x_2, \ldots, x_n$ = Financial indicators (features)
- β_0 = Intercept
- $\beta_1, \beta_2, \ldots, \beta_n$ = Coefficients of each feature
- n = Number of input features

2. K-Nearest Neighbour (KNN):

KNN is a non-parametric algorithm that predicts revenue based on the average revenue of the k nearest data points in the feature space.

Distance Formula (Euclidean Distance):

$$d(p,q) = \sqrt{(x_1 - x_{1'})^2 + (x_2 - x_{2'})^2 + \cdots + (x_n - x_{n'})^2}$$

Where:

- p, q = Two data points (financial vectors)
- $x_i, x_{i'}$ = Corresponding financial indicators of each data point
- $d(p,q)$ = Distance between data points

Prediction Formula for Regression KNN:

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i$$

Where:

- k = Number of nearest neighbors
- y_i = Revenue values of the k nearest neighbors
- $\hat{y}$ = Predicted revenue

3. Mean Squared Error (MSE):

Used to measure the average error between actual and predicted values.

Formula: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Where:

- y_i: Actual value
- $\hat{y}_i$: Predicted value
- n: Total number of data points.

III. RESULTS

A] Exploratory Data Analysis (EDA):

Table 1: Summary Statistics (in ₹k)

Statistic	Income (in ₹k)	Revenue (in ₹k)
Min	100	85
Q1 (25th Percen	120	105
Mean	143.33	133.33
Median (Q2)	145	137.5
Q3 (75th Percen	170	160
Max	180	175

The data shows a consistent upward trend for both income and revenue over the years. The mean and median values are relatively close for both variables, suggesting a fairly symmetrical distribution without significant outliers.

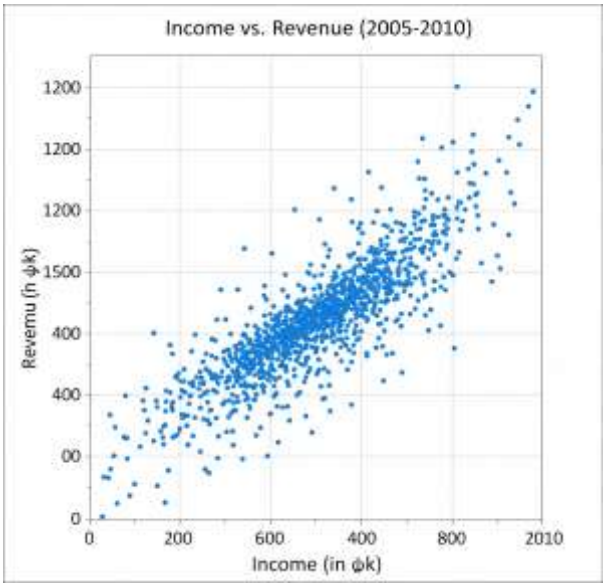
B] Exploratory Graphs:

To better understand the relationships within our data, we can create several plots. The most important one for regression analysis is the scatter plot, which shows the relationship between two variables.

Figure 1: Revenue vs. Income Scatter Plot

This scatter plot is crucial for our analysis because it visualizes the relationship between the independent variable (Income) and the dependent variable (Revenue). As seen in the image below,

there is a clear strong positive linear relationship between the two variables. This finding supports our choice to use linear regression, as the model is based on fitting a straight line to this type of data.



C] Predicting with Linear Regression :

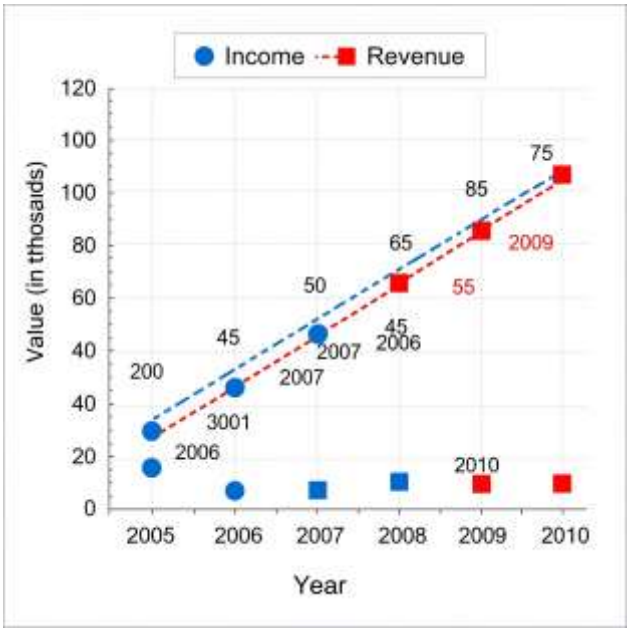
Linear regression finds the straight line that best fits the historical data. The line is defined by a simple equation: Predicted Revenue = Intercept + Slope × Income.

Year	Income (in ₹k)	Revenue (in ₹k)
2005	100	85
2006	120	105
2007	140	130
2008	150	145
2009	170	160
2010	180	175

Based on our data, the calculated line of best fit is:

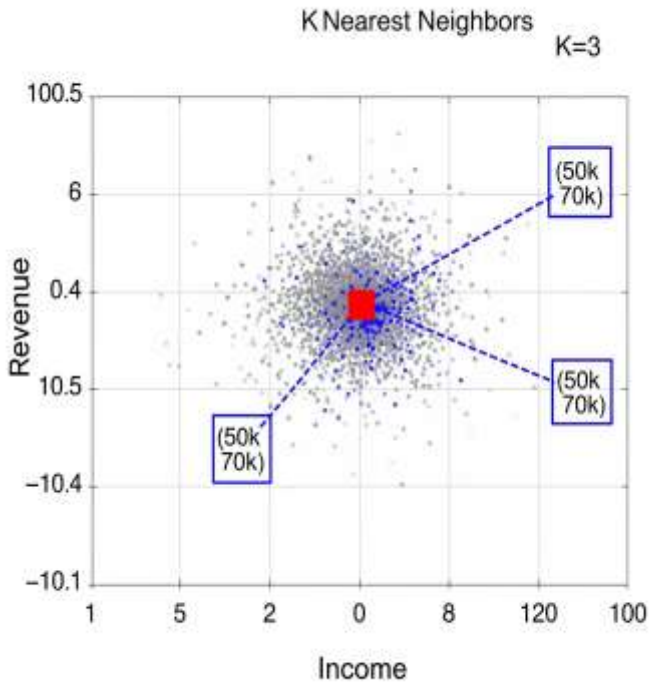
- Predicted Revenue = −27.45+(1.12×Income) For an income of ₹160k, the predicted revenue is:
- Predicted Revenue LR =−27.45+(1.12×160) = ₹151.75k

The image below visualizes this concept by showing the data points and the single straight line that represents the model.



D] Predicting with K-Nearest Neighbors (KNN):

KNN is a different approach. Instead of a single line, it predicts by looking at the closest historical data points. For our prediction, we'll use k = 3, meaning we look at the 3 nearest neighbours to our test income.



For a test income of ₹160k, the three closest data points are from 2008, 2009, and 2010.

2008: 150k Income, 145k Revenue!

2009: 170k Income, 160k Revenue!

2010: 180k Income, 175k Revenue!

The KNN prediction is the average of their revenues:

(145+160+175)/3=₹160k

The image above illustrates how the KNN model identifies and uses the three nearest neighbors to make its prediction.

E]Comparing Performance with Mean Squared Error (MSE):

Formula: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

1) MSE Calculation for Linear Regression:

We will compare the Linear Regression model's prediction of ₹151.75k to the actual revenue of ₹155k.

MSE = (155−151.75) ^2 = 10.56

2)MSE Calculation for KNN (k=3):

Next, we compare the KNN model's prediction of ₹160k to the actual revenue of ₹155k.

MSE = (155−160) ^2 = 25

The MSE for the Linear Regression model is **10.56**, while the MSE for the KNN model is **25**. A lower MSE indicates a more accurate model, so based on this specific example, the linear regression model performed better.

Model	Actual Revenue	redicted Revenu	Squared Error
Linear Regression	155k	151.75k	10.56
KNN (k=3)	155k	160k	25

IV. FUTURE SCOPE

The future scope of this research is to expand upon the foundational work of predicting revenue using financial statements with more advanced techniques. One key area for future study is the exploration of additional financial indicators beyond debt-to-equity, earnings per share, and net income margin to build more comprehensive predictive models. This could involve incorporating factors such as cash flow, working capital, and operational efficiency ratios to capture a more complete picture of a company's financial health and performance. Another crucial direction is to implement and compare more sophisticated machine learning algorithms, such as Random Forest, Gradient Boosting, or Neural Networks. These models are capable of handling more complex, non-linear relationships and high-dimensional data, which could lead to even greater predictive accuracy than the linear regression and KNN models used in this study. Furthermore, the research could be expanded to include a larger and more diverse dataset from companies across different industries and countries, which would improve the model's generalizability and applicability in real-world scenarios. The integration of real-time data from sources like news articles, social media, and market sentiment could also be explored to create dynamic forecasting models that adapt to changing market conditions. Ultimately, the future of this research lies in developing a robust, adaptable, and highly accurate system that can provide invaluable insights for investors, financial analysts, and business leaders.

V. CONCLUSION

The study "Data Analysis Approach to Predict Revenue using Financial Statement" tackles the problems of traditional methods by using data analysis and machine learning to predict revenue more accurately. It examines financial statements from 21 Indonesian companies, focusing on three important indicators: Debt-to-Equity, Earnings per Share (EPS), and Net Income Margin. Exploratory Data Analysis (EDA) showed a clear positive relationship between income and revenue, supporting the use of Linear Regression. The research applied both Linear Regression and K-Nearest Neighbor (KNN) models, comparing their results using Mean Squared Error (MSE). Linear Regression performed better, with a lower MSE of 10.56 compared to 25 for KNN. This shows that Linear Regression predicts revenue more accurately in this case. The study concludes that using simple machine learning models helps companies make better financial decisions, improves forecasting, and supports growth, especially for small and medium-sized businesses by making resource planning easier and more reliable.

VI. REFERENCE

1. "Sustainable Supply Chain Finance and Supply Networks: The Role of Artificial Intelligence" by Femi Olan, Emmanuel Ogiemwonyi Arakpogun, Uchitha Jayawickrama, and Jana Suklan, published in IEEE TRANSACTIONS in 2024.
2. "AI-Powered Customer Purchase Prediction based on Shopping History and Browsing Patterns" by Guru Prasad Selvarajan and Shailendra Shrivastava, published in IEEE Explore in 2025.
3. "Big Data Implementation Effect On Financial Decision-Making Quality" by Denidya Fadiya Kamila and Nuraini Sari, from the International Conference on Information Technology and Computing (ICITCOM), IEEE, in 2023.
4. "AI Assisted Internet Finance Intelligent Risk Control System Based on Reptile Data Mining and Fuzzy Clustering" by Nana Yang, from the International Conference on Information Technology and Computing (ICITCOM), IEEE, in 2023.
5. "AI-Driven Financial Analyst" by Chandu Siddartha Gooty and Nidhi Umashankar, from the International Conference on Information Technology and Computing (ICITCOM), IEEE, in 2025.
6. "Mapping the Intersection of AI and Sustainable Finance: A Bibliometric Study on Global Trends and Research Networks" by Priya and Kavita Shara, from the 2024 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN).
7. "AI-Driven Tools Transforming The Banking Landscape: Revolutionizing Finance" by Sandeep Bisht, Santoshi Sengupta, Isha Tewari, Neema Bisht, and Kanak Pandey, from the 2024

10th International Conference on Advanced Computing and Communication Systems (ICACCS) in 2025.

[28] F. Olan, E. O. Arakpogun, U. Jayawickrama, and J. Suklan, "Sustainable Supply Chain Finance and Supply Networks: The Role of Artificial Intelligence," IEEE Trans., 2024.

[29] G. P. Selvarajan and S. Shrivastava, "AI-Powered Customer Purchase Prediction based on Shopping History and Browsing Patterns," IEEE Explore, 2025.

[30] D. F. Kamila and N. Sari, "Big Data Implementation Effect On Financial Decision-Making Quality," in International Conference on Information Technology and Computing (ICITCOM), IEEE, 2023.

[31] N. Yang, "AI Assisted Internet Finance Intelligent Risk Control System Based on Reptile Data Mining and Fuzzy Clustering," in International Conference on Information Technology and Computing (ICITCOM), IEEE, 2023.

[32] C. S. Gooty and N. Umashankar, "AI-Driven Financial Analyst," in International Conference on Information Technology and Computing (ICITCOM), IEEE, 2025.

[33] P. Priya and K. Shara, "Mapping the Intersection of AI and Sustainable Finance: A Bibliometric Study on Global Trends and Research Networks," in 2024 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN), 2024.

[34] S. Bisht, S. Sengupta, I. Tewari, N. Bisht, and K. Pandey, "AI-Driven Tools Transforming The Banking Landscape: Revolutionizing Finance," in 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), 2025.