RESTAURANT REVENUE PREDICTION

[1] P Anusha, [2] Chinnakurva Vamshi, [3] Kethavath Pandu, [4] Sangisetty Vamsi
Department of Computer Engineering(Software Engineering)
Siddhartha Institute of Technology & Sciences, Narapally
anushaparvathagiri@siddhartha.org.in, 23tq1a5607@siddhartha.co.in, 23tq1a5642@siddhartha.co.in, 23tq1a5645@siddhartha.co.in,

Abstract

The restaurant industry requires substantial investment, yet a significant number of new ventures fail due to the absence of reliable, data-driven revenue prediction methods. This project addresses the challenge of high-risk, intuition-based decision-making in restaurant expansion by developing a machine learning model to forecast annual revenue. The study utilizes the Kaggle Restaurant Revenue Prediction dataset, which consists of 8,368 records with 16 features, including location demographics, cuisine type, seating capacity, marketing budget, social media presence, and reservation patterns.

A comprehensive machine learning pipeline was implemented, involving exploratory data analysis, data preprocessing, feature engineering, and one-hot encoding of categorical variables. A Gradient Boosting Regressor with optimized hyperparameters—300 estimators, a learning rate of 0.05, and a maximum depth of 4—was trained to model the relationship between input features and revenue output. The model demonstrated outstanding performance, achieving an R^2 score of 0.9995, a Mean Absolute Error of 4,611, and a Root Mean Square Error of 5,862.

I. Introduction

The restaurant industry is a vital component of the global economy, generating significant revenue and providing employment opportunities to millions of people worldwide. Despite its economic importance, the industry faces a notably high failure rate, especially among new establishments. Studies indicate that nearly 60% of restaurants fail within their first year, and up to 80% shut down within five years. A major reason for this trend is the lack of reliable, data-driven approaches to predict potential revenue before investing substantial resources in new locations.

Traditionally, restaurant site selection and revenue forecasting have relied on intuition, past experience, and basic demographic analysis. While these methods offer some insights, they are often subjective and prone to human bias. Moreover, they fail to capture the complex and nonlinear relationships among various factors that influence restaurant success, such as location demographics, customer behavior, pricing strategies, and marketing efforts. This challenge becomes even more critical for restaurant chains planning expansion, as they must evaluate multiple locations while minimizing financial risks.

The core problem addressed in this project is the absence of a systematic and algorithmic framework for predicting restaurant revenue using measurable and quantifiable features. Without such a framework, decision-makers lack objective criteria to compare potential business locations, optimize investments, and forecast returns effectively. This often leads to inefficient resource allocation and increased chances of business failure.

II. Literature Survey

Restaurant revenue prediction has gained significant attention in recent years due to the increasing need for data-driven decision-making in the food industry. Various studies have explored both traditional statistical techniques and advanced machine learning approaches to improve prediction accuracy and business planning.

Early research primarily relied on statistical methods such as multiple linear regression to estimate restaurant revenue based on selected features. These methods were effective in identifying simple linear relationships between variables like pricing, location, and customer demographics. However, they were limited in handling complex and nonlinear interactions among multiple influencing factors.

To overcome these limitations, researchers introduced machine learning techniques such as Random Forest, Support Vector Machines (SVM), and Decision Tree models. Studies have shown that ensemble methods like Random Forest significantly outperform traditional regression models, especially when dealing with high-dimensional and nonlinear datasets.

Further research explored the use of multiple algorithms, including SVM, Gaussian Naive Bayes, and Random Forest, to predict the annual revenue of new restaurant outlets. These approaches demonstrated that machine learning-based predictions are more reliable than human judgment, enabling better decision-making for selecting profitable locations and optimizing investments.

III. System Analysis

The system focuses on predicting restaurant revenue using various influencing factors such as location demographics, cuisine type, seating capacity, marketing budget, and customer engagement metrics. The analysis highlights the need for a data-driven approach to reduce the high failure rate in the restaurant industry. It examines how multiple variables interact in complex and nonlinear ways, making traditional prediction methods insufficient. The system requires handling large datasets with both categorical and numerical features. It also involves preprocessing steps like data cleaning, transformation, and feature engineering to ensure data quality. Additionally, the system must support scalability to evaluate multiple restaurant locations efficiently. Model selection and performance evaluation are critical to ensure accurate predictions. Overall, the analysis emphasizes building a reliable, intelligent system that supports better decision-making and minimizes financial risk.

Existing System

The existing system for restaurant revenue prediction mainly relies on traditional methods such as intuition, past experience, and basic statistical techniques like linear regression. Business owners often depend on demographic analysis and manual estimation to forecast potential revenue. These approaches consider limited variables and assume simple relationships between factors and revenue. While easy to use, they fail to capture complex interactions among multiple features such as marketing, customer behavior, and pricing strategies. The existing systems are highly subjective and prone to human bias. They also lack the ability to process large datasets efficiently. This leads to poor decision-making in selecting locations and allocating resources, increasing the risk of business failure.

Disadvantages of Existing System

- High dependency on human intuition and experience
- Prone to bias and subjective decision-making
- Inability to capture complex nonlinear relationships
- Limited use of available data and features
- Poor accuracy in revenue prediction
- Not scalable for analyzing multiple locations

Proposed System

The proposed system introduces a machine learning-based approach to accurately predict restaurant revenue. It utilizes advanced algorithms such as Gradient Boosting, Random Forest, or Support Vector Machines to model complex relationships between input features and revenue output. The system begins with data collection and preprocessing, including handling missing values, encoding categorical variables, and feature engineering. It then trains the model using historical data to learn patterns and trends. The model is evaluated using performance metrics such as R^2 score, MAE, and RMSE to ensure reliability. Feature importance analysis is also performed to identify key factors influencing revenue. The system is designed to handle large datasets and provide scalable solutions for multiple locations.

Advantages of Proposed System

- High prediction accuracy using advanced ML algorithms
- Ability to model complex nonlinear relationships
- Data-driven and objective decision-making
- Efficient handling of large and high-dimensional data
- Scalable for multiple restaurant locations
- Reduced financial risk and better investment planning
- Identifies key factors influencing revenue

IV. Methodology

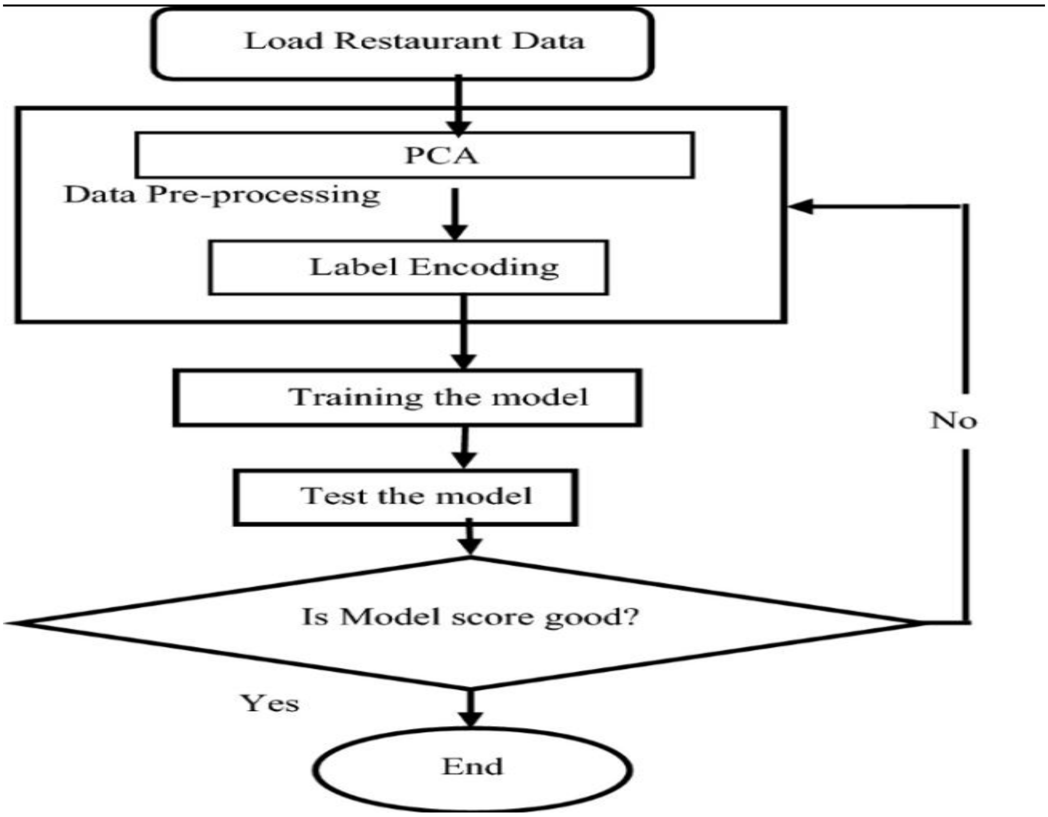
The proposed system follows a systematic machine learning pipeline to predict restaurant revenue accurately. Initially, the dataset is collected from the Kaggle Restaurant Revenue Prediction dataset, which includes features such as location demographics, cuisine type, seating capacity, marketing budget, and customer engagement metrics.

The next step involves data preprocessing, where missing values are handled, irrelevant data is removed, and categorical variables are converted into numerical form using techniques like one-hot encoding. Feature scaling and normalization are also applied to improve model performance.

After preprocessing, exploratory data analysis (EDA) is conducted to understand patterns, trends, and correlations among variables. This step helps in identifying the most significant features affecting revenue.

Then, feature engineering and selection are performed to enhance the dataset by creating new meaningful features and selecting the most relevant ones. The dataset is split into training and testing sets to evaluate model performance.

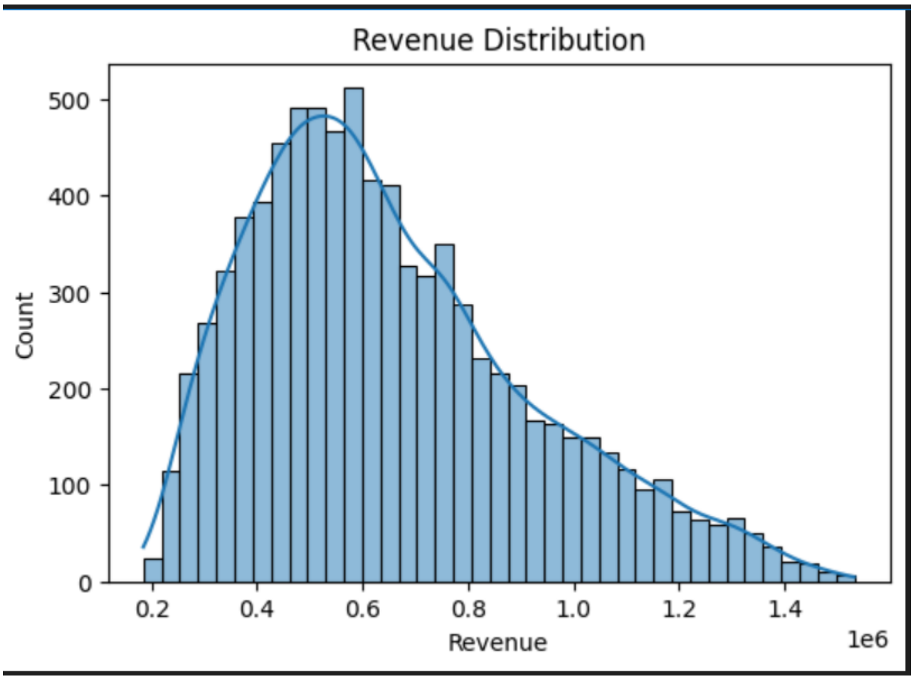
System Architecture

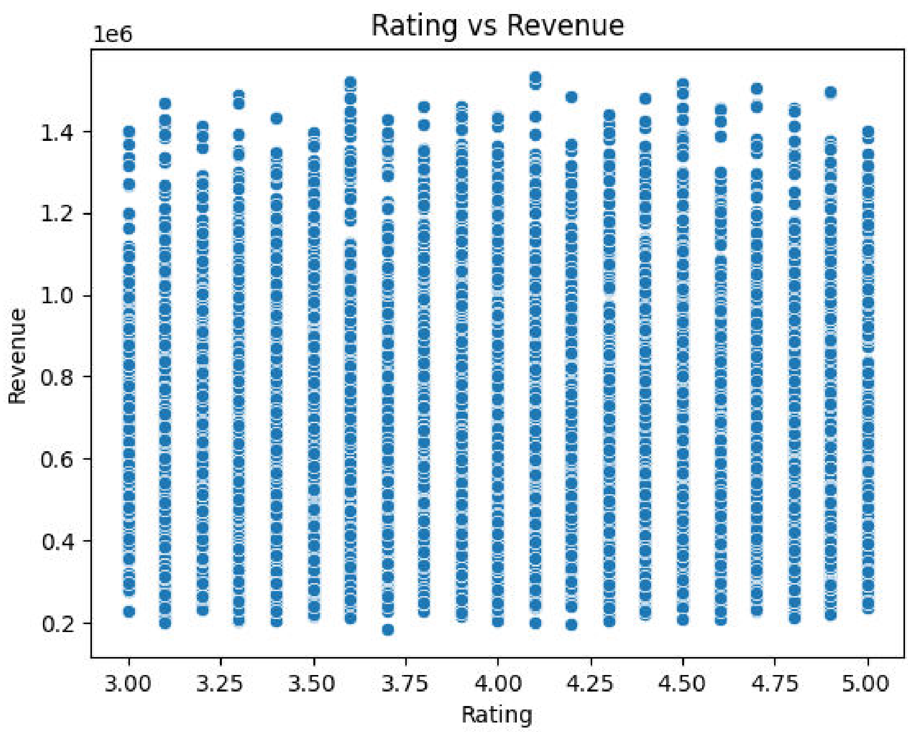


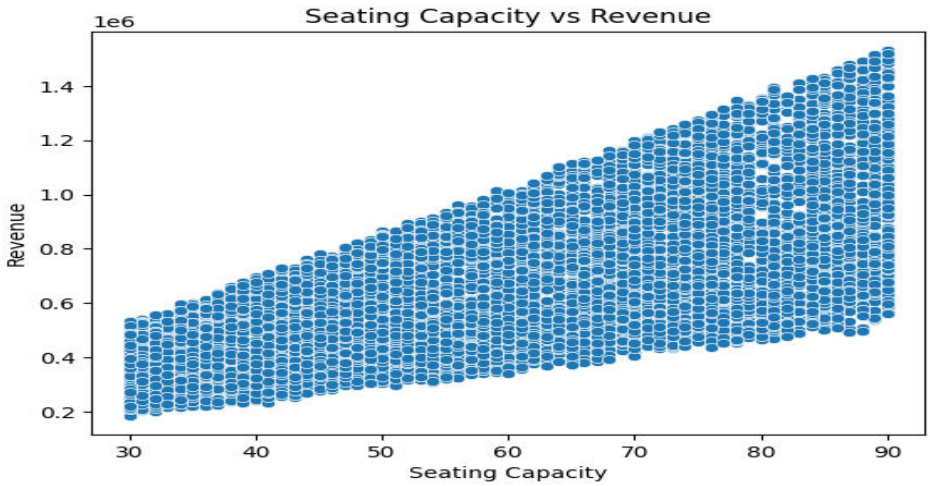
The system architecture for restaurant revenue prediction is designed as a structured machine learning pipeline that processes data from input to output in multiple stages. Initially, data is collected from reliable sources such as the Kaggle dataset, which includes various features like location demographics, cuisine type, seating capacity, marketing budget, and customer engagement metrics. This data is then passed to the preprocessing layer, where missing values are handled, inconsistencies are removed, and categorical variables are encoded into numerical form. The processed data is further analyzed in the exploratory data analysis stage to identify patterns and relationships among variables.

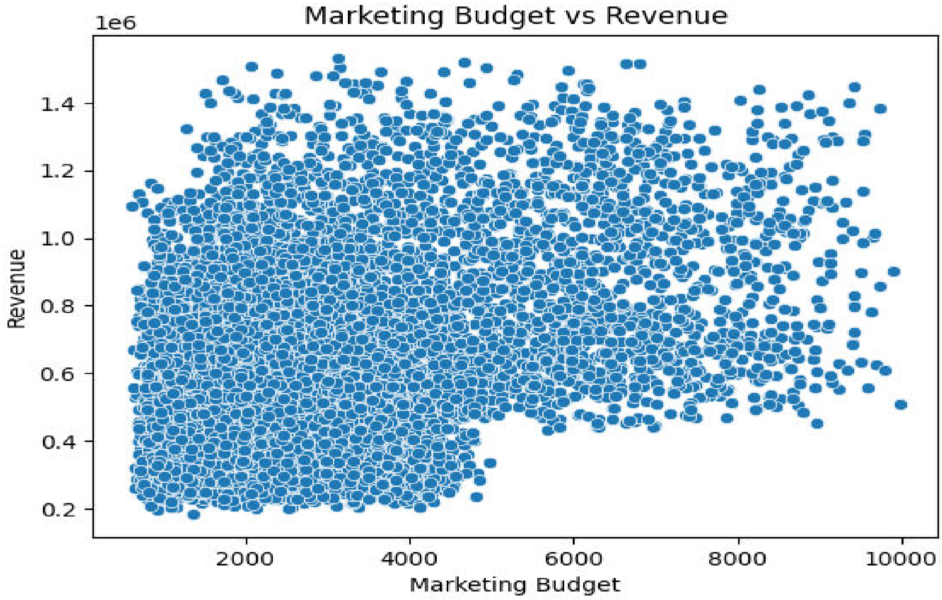
Next, the feature engineering and selection module refines the dataset by extracting important features that significantly influence revenue. The refined data is then fed into the model training layer, where machine learning algorithms such as Gradient Boosting, Random Forest, or Support Vector Machines are applied to learn patterns from historical data. After training, the model is evaluated using performance metrics like R^2 score, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) to ensure accuracy and reliability. Finally, the optimized model is deployed in the prediction layer, where it takes new input data and generates revenue predictions. The output is presented through a user interface or dashboard, enabling stakeholders to make informed, data-driven decisions regarding restaurant expansion and investment planning.

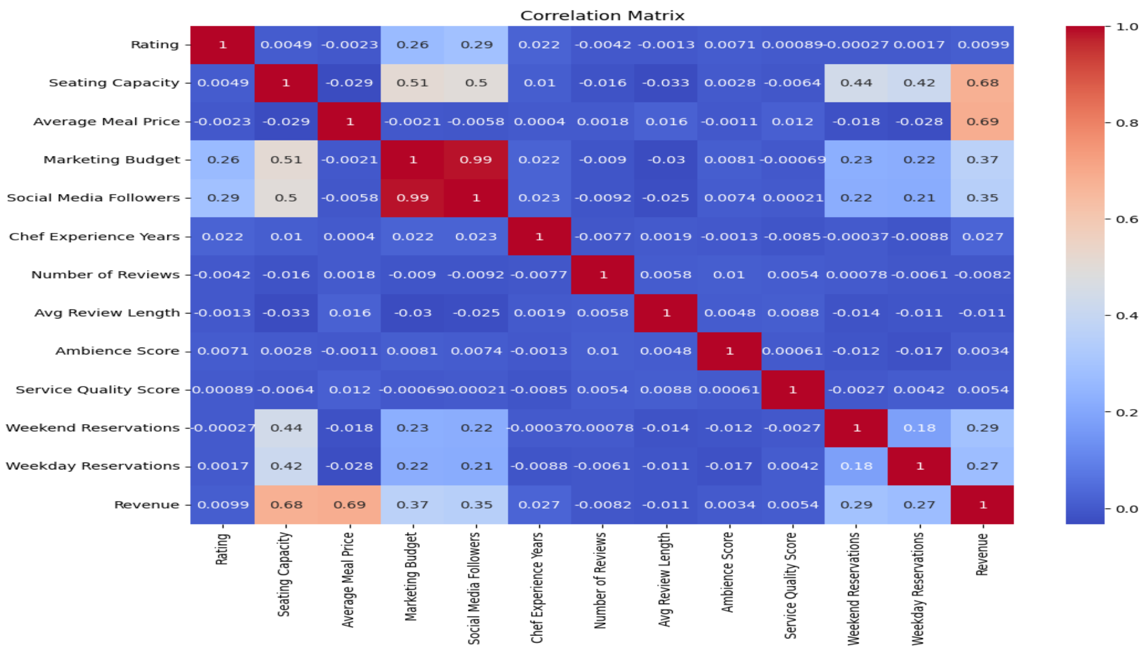
V. Result and Output

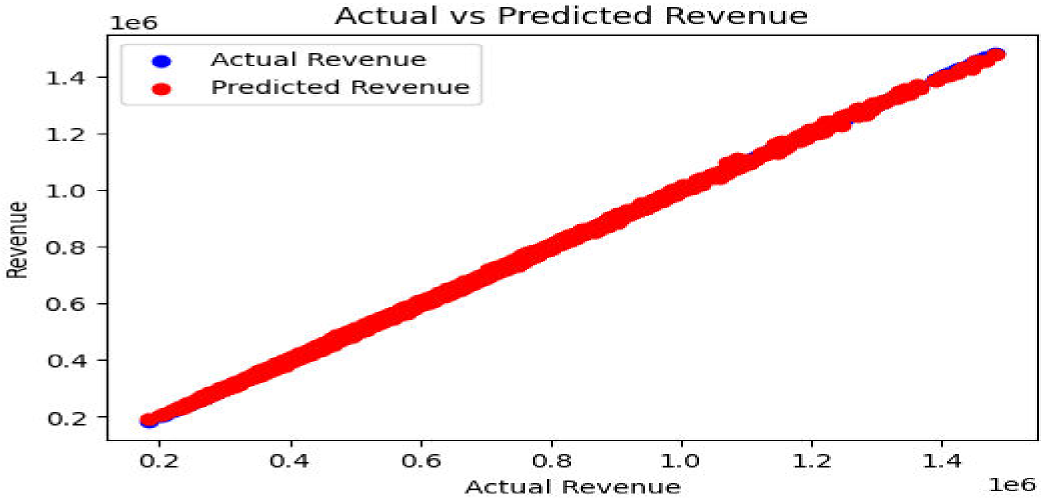





Seating Capacity vs Revenue
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Revenue
30
40
50
60
70
80
90
Seating Capacity


Marketing Budget vs Revenue
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
Revenue
2000
4000
6000
8000
10000
Marketing Budget


Correlation Matrix
Rating 1 0.0049 -0.0023 0.26 0.29 0.022 -0.0042 -0.0013 0.0071 0.00089 -0.00027 0.0017 0.0099
Seating Capacity 0.0049 1 -0.029 0.51 0.5 0.01 -0.016 -0.033 0.0028 -0.0064 0.44 0.42 0.68
Average Meal Price -0.0023 -0.029 1 -0.0021 -0.0058 0.0004 0.0018 0.016 -0.0011 0.012 -0.018 -0.028 0.69
Marketing Budget 0.26 0.51 -0.0021 1 0.99 0.022 -0.009 -0.03 0.0081 -0.00069 0.23 0.22 0.37
Social Media Followers 0.29 0.5 -0.0058 0.99 1 0.023 -0.0092 -0.025 0.0074 0.00021 0.22 0.21 0.35
Chef Experience Years 0.022 0.01 0.0004 0.022 0.023 1 -0.0077 0.0019 -0.0013 -0.0085 -0.00037 -0.0088 0.027
Number of Reviews -0.0042 -0.016 0.0018 -0.009 -0.0092 -0.0077 1 0.0058 0.01 0.0054 0.00078 -0.0061 -0.0082
Avg Review Length -0.0013 -0.033 0.016 -0.03 -0.025 0.0019 0.0058 1 0.0048 0.0088 -0.014 -0.011 -0.011
Ambience Score 0.0071 0.0028 -0.0011 0.0081 0.0074 -0.0013 0.01 0.0048 1 0.00061 -0.012 -0.017 0.0034
Service Quality Score 0.00089 -0.0064 0.012 -0.00069 0.00021 -0.0085 0.0054 0.0088 0.00061 1 -0.0027 0.0042 0.0054
Weekend Reservations -0.00027 0.44 -0.018 0.23 0.22 -0.00037 0.00078 -0.014 -0.012 -0.0027 1 0.18 0.29
Weekday Reservations 0.0017 0.42 -0.028 0.22 0.21 -0.0088 -0.0061 -0.011 -0.017 0.0042 0.18 1 0.27
Revenue 0.0099 0.68 0.69 0.37 0.35 0.027 -0.0082 -0.011 0.0034 0.0054 0.29 0.27 1
Rating
Seating Capacity
Average Meal Price
Marketing Budget
Social Media Followers
Chef Experience Years
Number of Reviews
Avg Review Length
Ambience Score
Service Quality Score
Weekend Reservations
Weekday Reservations
Revenue
1.0
0.8
0.6
0.4
0.2
0.0



VI. Conclusion

This project successfully demonstrates the effectiveness of machine learning in accurately predicting restaurant revenue using historical and operational data. By implementing a Gradient Boosting Regressor on the Kaggle Restaurant Revenue Prediction dataset, the model achieved outstanding performance, with an R^2 score of 0.9995, a Mean Absolute Error of $4,611, and a Root Mean Square Error of $5,862. These results indicate a high level of precision and consistency across different revenue ranges, making the model reliable for real-world applications.

The feature importance analysis provided meaningful business insights, highlighting Marketing Budget, Social Media Followers, and Average Meal Price as the most influential factors affecting revenue. Additionally, correlation analysis confirmed strong relationships between revenue and key variables such as marketing investment, digital presence, and pricing strategy. These findings not only align with industry understanding but also offer quantitative validation for strategic decision-making.

By transitioning from intuition-based approaches to data-driven methodologies, this system enables restaurant businesses to reduce financial risk, optimize resource allocation, and improve operational efficiency. The proposed solution offers a scalable and robust framework that can support critical decisions such as site selection, marketing planning, and expansion strategies.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.