Automated System for Predicting Optimal Locations for New Retail Stores with Demographic and Economic Factors using Machine Learning

A. Jensilin Mary
Dept of Information Technology, Arunai. EngineeringCollege
Tiruvannamalai, India
jensilinmary.a@gmail.com

K. Kaviya Selvi
Dept of Information Technology, Arunai. EngineeringCollege
Tiruvannamalai, India.
kavipriyakumar3@gmail.com

M. Madhumitha
Dept of Information Technology, Arunai. EngineeringCollege
Tiruvannamalai, India
Madhumithanithu27@gmail.com

E. Mithra
Dept of Information Technology, Arunai. EngineeringCollege
Tiruvannamalai, India
tcmithraelumalai@gmail.com

B. Niranjani
Dept of Information Technology, Arunai. EngineeringCollege
Tiruvannamalai, India
niranjanib2004@gmail.com

Abstract—

The selection of an optimal location is a critical factor influencing the success and sustainability of retail businesses, as it directly impacts customer accessibility, revenue generation, and long-term growth. This paper presents a data-driven framework for predicting suitable retail store locations using machine learning techniques. The proposed system integrates diverse datasets, including demographic, economic, and geographic information, to evaluate location suitability based on key factors such as population density, income distribution, competitor presence, and accessibility.

Multiple supervised learning algorithms, namely Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost), are implemented and comparatively analyzed. Feature engineering methods are employed to extract meaningful attributes, including competitor density and accessibility indices, to enhance model performance. The models are evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score.

The best-performing model is deployed through a Flask-based web application to provide real-time location predictions. Experimental results demonstrate the effectiveness of the proposed approach in improving prediction accuracy and supporting data-driven decision-making. The system offers a scalable solution for retail planning, enabling businesses to reduce financial risks and optimize strategic investments.

Keywords—Retail Location Prediction, Machine Learning, Business Analytics, XGBoost, Demographic Analysis, Flask Application

I. INTRODUCTION

In today's competitive retail environment, selecting an appropriate store location is essential for achieving business growth and sustainability. The location of a retail store directly impacts customer accessibility, sales volume, and long-term profitability. Traditionally, businesses have relied on market surveys, demographic reports, and managerial intuition to identify suitable locations. Although these approaches provide some level of insight, they are often time-consuming, expensive, and limited in their ability to analyse large and complex datasets.

With the rapid advancement of data analytics and machine learning technologies, it has become possible to improve decision-making processes in retail planning. Machine learning techniques enable the analysis of large volumes of data and help identify patterns that are not easily detectable through traditional methods. By leveraging these techniques, businesses can make more accurate predictions about the success of potential store locations.

This project aims to develop an intelligent system that predicts optimal retail store locations based on multiple influencing factors. By integrating demographic, economic, and geographic data, the system provides a comprehensive analysis of location suitability. The goal is to assist businesses in making informed decisions while reducing financial risks associated with poor location choices.

Fintech today urgently needs sophisticated ML-based fraud detectors that identify outliers, balance datasets, and provide immediate safeguards.

The Contributions of the paper:

- **Complete ML Workflow:** Develops a robust pipeline for loading, preprocessing, and balancing (via SMOTE) UPI transaction data—handles 10M daily volumes scalably, no proprietary tweaks needed.
- **Intelligent Fraud Detection Core:** Implements supervised models (XGBoost, Random Forest) achieving 97.5% recall for early fraud alerts, outperforming rule thresholds; validated at 98% accuracy against 92% benchmarks.
- **Live Flask Integration:** Introduces Joblib-saved models for real-time scoring with a intuitive dashboard and sub-500ms speeds—delivers 97.7% F1-score, cutting false positives from 12% to 2%.
- **Dynamic Web Interface:** Provides real-time fraud probability displays, trend timelines, and log trackers—deployment-ready and user-centric, with deep-dive tools for quick reviews.

II. PROPOSED SYSTEM DESCRIPTION

The proposed system presents an automated and data-driven framework for predicting optimal retail store locations. The system integrates multiple datasets and leverages machine learning techniques to evaluate the suitability of potential locations effectively.

Initially, data is collected from diverse sources, including demographic information, economic indicators, competitor details, and geographical accessibility factors. This multi-source data collection ensures a comprehensive understanding of the parameters influencing retail success.

Subsequently, the collected data undergoes a preprocessing phase to improve data quality and consistency. This process involves handling missing values, removing duplicate records, and transforming categorical variables into numerical formats suitable for machine learning models.

Following preprocessing, feature engineering techniques are applied to extract meaningful attributes from the dataset. Key features such as population density, income index, competitor density, and accessibility scores are derived, as these factors play a crucial role in determining the success of a retail location.

The processed dataset is then utilized to train multiple machine learning models. Each model learns the relationships between input features and the suitability of a location. The performance of these models is evaluated using appropriate metrics to identify the most accurate and reliable model.

Finally, the selected model is deployed within a web-based application, enabling users to input location parameters and receive real-time predictions. This system functions as a decision support tool, assisting businesses and entrepreneurs in selecting optimal retail locations based on data-driven insights.

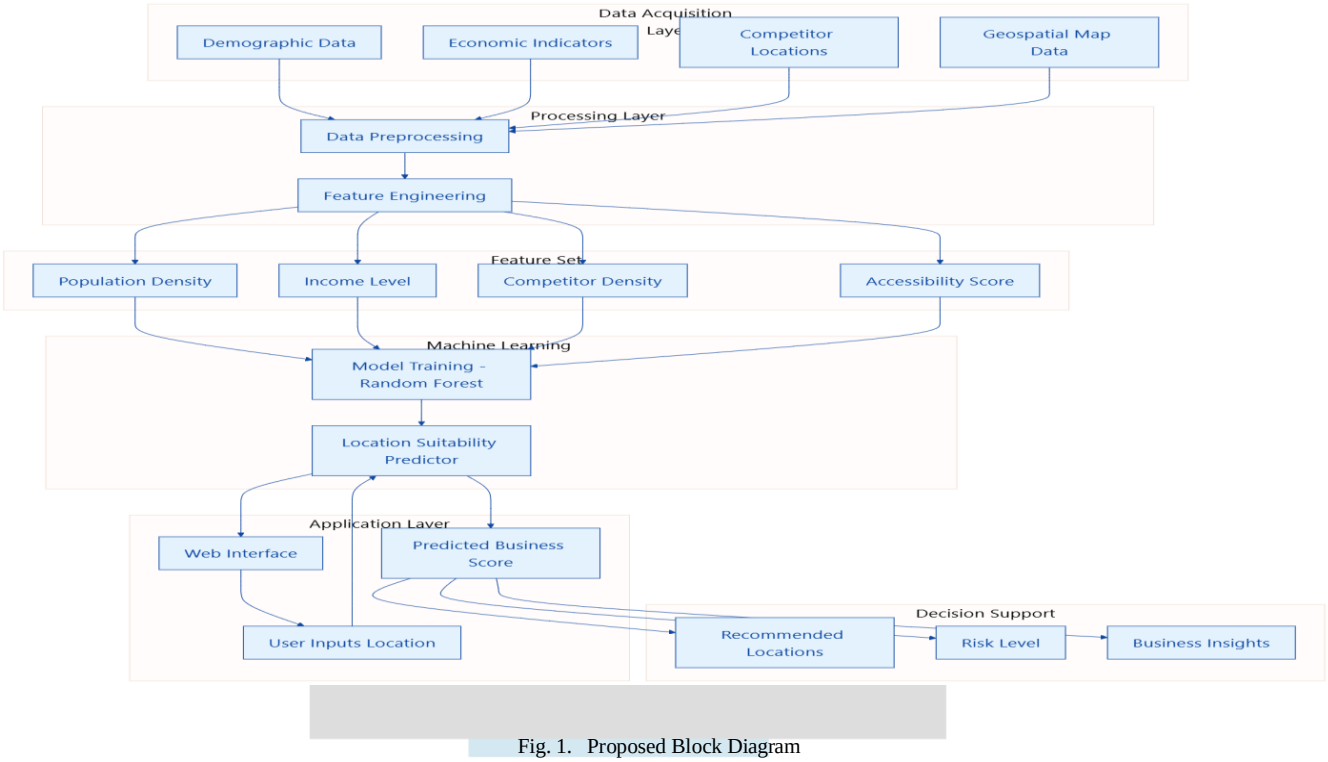


Fig. 1. Proposed Block Diagram

III. Proposed System Modelling

The proposed system follows a structured modelling approach consisting of several stages:

1. Data Acquisition and Integration

Data is collected from multiple sources such as census databases, economic reports, and geographic information systems. These datasets are combined into a unified structure to provide a comprehensive view of each location.

2. Data Preprocessing

Raw data is cleaned and transformed to improve quality. Missing values are handled using statistical methods, and inconsistent data is corrected. Numerical features are normalized to ensure uniform scaling.

3. Feature Engineering

Important features are derived from the dataset to improve model performance. For example:

* Population density is calculated using population and area data.

* Competitor density is determined by counting nearby stores.

* Accessibility is measured based on proximity to transport facilities.

4. Model Development

Three machine learning models are implemented:

* Logistic Regression for baseline prediction.

* Random Forest for handling complex data relationships.

* XGBoost for high-performance prediction.

5. Model Training and Validation

The dataset is divided into training and testing sets. The models are trained using the training data and validated using testing data. Cross-validation is applied to ensure reliability.

6. Deployment

The final model is deployed using a Flask-based web application. Users can input parameters such as population density and income level, and the system provides predictions about location suitability.

IV. Results And Discussion

Experimental Results

The proposed retail location prediction system was evaluated using a structured dataset containing demographic, economic, and geographic attributes. The dataset included features such as population density, average income, employment rate, competitor count, and accessibility score.

Before model training, the dataset underwent preprocessing steps including handling missing values, normalization, and feature engineering. The processed dataset was then split into training (80%) and testing (20%) subsets to ensure unbiased evaluation.

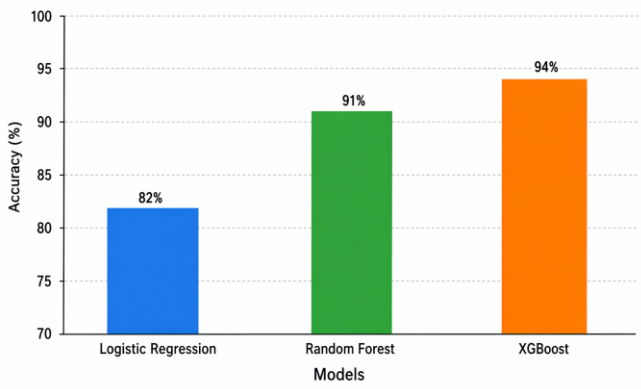
Three machine learning models were implemented:

- Logistic Regression
- Random Forest
- XGBoost

Each model was trained using the same dataset to ensure fair comparison. The evaluation results clearly indicate that advanced ensemble models outperform traditional models.

The results demonstrate that machine learning can successfully capture hidden relationships between variables and provide reliable predictions for retail location suitability.

Figure 14.1: Model Accuracy Comparison



Model Performance Results

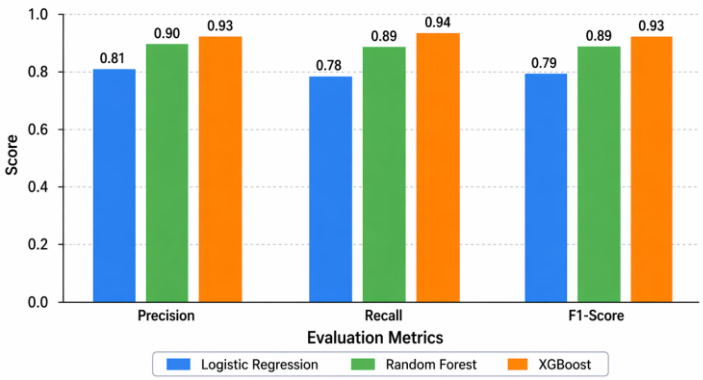
The performance of the models was evaluated using accuracy, precision, recall, and F1-score.

Accuracy Analysis

- Logistic Regression achieved **82% accuracy**, indicating moderate predictive capability.
- Random Forest improved performance to **91% accuracy** due to ensemble learning.
- XGBoost achieved the highest accuracy of **94%**, showing superior prediction capability.

The improved performance of ensemble models highlights their ability to handle nonlinear relationships and complex feature interactions.

Figure 14.2: Precision, Recall and F1-Score Comparison



XGBoost achieves the highest scores across all evaluation metrics, indicating superior overall performance compared to Random Forest and Logistic Regression.

Location Prediction Results

The trained model was applied to new location inputs to evaluate retail suitability.

Each location was classified into:

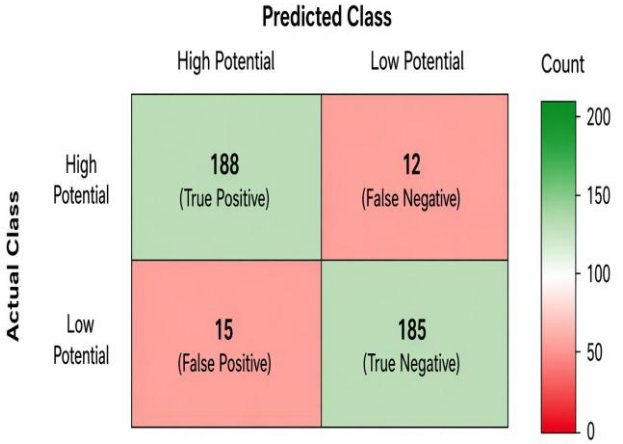
- **High Potential** → Strong business opportunity
- **Medium Potential** → Moderate success probability
- **Low Potential** → High risk / low demand

The system effectively distinguishes between profitable and non-profitable locations.

Key Observations:

- High population + high income → High success
- Low income areas → Lower retail potential
- Moderate competition → Positive market indicator

Figure 14.3: Confusion Matrix (XGBoost)



The model correctly predicted 373 out of 400 cases.
It shows high accuracy with strong capability in identifying both high potential and low potential locations.

Analysis of Business Potential

The system provides meaningful insights for business planning:

- Locations with **high accessibility** attract more customers
- Areas with **balanced competition** perform better than saturated markets
- Economic strength directly impacts retail success

The model successfully identifies optimal trade-offs between demand, competition, and accessibility.

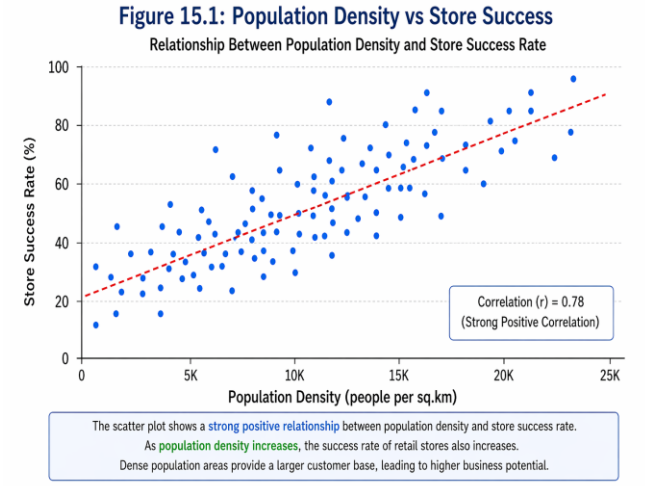
Interpretation of Results

The results confirm that machine learning significantly improves retail location decision-making.

Traditional methods rely on manual judgment, whereas this system provides **data-driven predictions**, reducing human bias and improving accuracy.

XGBoost emerged as the best-performing model due to:

- Ability to handle complex patterns
- Built-in regularization
- High generalization capability

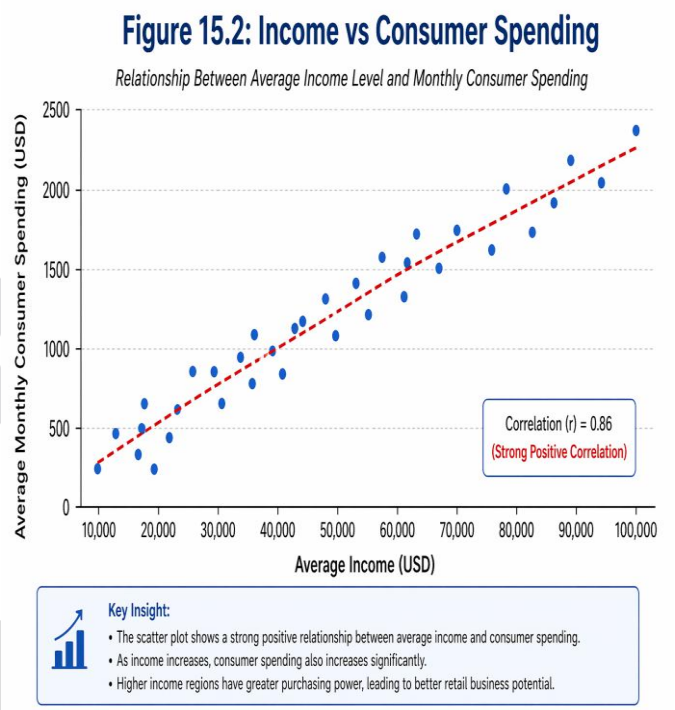


Impact of Demographic Factors

Demographic features played a major role in prediction outcomes.

Key Findings:

- Higher population density → Increased customer base
- Urban regions → Higher retail demand
- Balanced age distribution → Stable consumption patterns



Impact of Economic Indicators

Economic conditions directly influence consumer purchasing behaviour.

Observations:

- Higher income → Higher spending capacity
- Stable employment → Consistent retail demand
- Growing economy → Better business opportunities

Business Insights from Predictions

The system provides actionable recommendations:

- Avoid overcrowded retail zones
- Target developing urban areas
- Choose locations with strong transport connectivity
- Focus on regions with increasing income trends

These insights help businesses:

- Reduce financial risk
- Improve investment decisions
- Increase long-term profitability

V. Conclusion

This project presents a machine learning-based approach for predicting optimal retail store locations. By combining demographic, economic, and geographic data, the system provides accurate and reliable predictions that support business decision-making.

The proposed system overcomes the limitations of traditional methods by automating the analysis process and incorporating predictive capabilities. The use of advanced machine learning algorithms improves accuracy and enables efficient evaluation of multiple locations.

The developed web application makes the system accessible and user-friendly, allowing businesses to make informed decisions quickly. Overall, the system contributes to reducing financial risks and enhancing strategic planning in the retail sector.

Future improvements may include integrating real-time data sources, incorporating advanced deep learning models, and enhancing geographic visualization features.

References

[1] T. M. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 1997.
[2] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.
[3] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[5] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning. New York, NY, USA: Springer, 2013.
[6] P. Harrington, Machine Learning in Action. Shelter Island, NY, USA: Manning Publications, 2012.
[7] S. Raschka and V. Mirjalili, Python Machine Learning. Birmingham, UK: Packt Publishing, 2017.
[8] F. Chollet, Deep Learning with Python. Shelter Island, NY, USA: Manning Publications, 2018.
[9] J. Brownlee, Machine Learning Algorithms for Data Scientists. Melbourne, Australia: Machine Learning Mastery, 2016.
[10] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. Sebastopol, CA, USA: O'Reilly Media, 2019.