iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83604

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

An Explainable AI Framework for Integrated Retail Analytics and Predictive Business Intelligence

Krovvidi Sri Harsha[1], Dr. B. Kranthi Kiran[2]
[1]*MTech Student, Computer Science and Engineering, JNTU Hyderabad, Telangana*
[2] *Professor of Computer Science and Engineering, JNTU Hyderabad, Telangana*

Abstract: ***In today's retail environment, organizations generate huge amounts of transactional data. However, many companies still use traditional reporting systems that are primarily concerned with historical analysis of sales. Such approaches provide limited insights into future customer behaviour and long term profitability. The paper introduces a Retail Analytics and Customer Intelligence Platform that uses a combination of machine learning techniques to integrate customer segmentation, Customer Lifetime Value (CLV) prediction and sales forecasting. The Online Retail dataset is pre-processed and transformed to generate RFM (Recency, Frequency, Monetary) features to understand customer purchasing behaviour. Machine learning algorithms such as Linear Regression, Random Forest and XG Boost are used to estimate customer lifetime value. XG Boost was the best model with an R2 score of 0.986, indicating good predictive power among these models. Customers are segmented into Low, Medium and High value segments to facilitate targeted marketing campaigns. There's also a sales forecasting module that predicts revenue trends to help with business planning. Developed an interactive dashboard with Streamlit for data visualization and decision making. The proposed system helps retailers to increase customer retention, optimize marketing efforts, and improve accuracy of revenue prediction.***

Keywords: Retail Analytics, Customer Lifetime Value (CLV), RFM Analysis, Machine Learning, Sales Forecasting, Customer Segmentation, XG Boost, Random Forest, Streamlit Dashboard, Predictive Analytics.

I. INTRODUCTION

Huge changes have come to the retail sector in the form of digital commerce and data-driven business models. Retailers amass huge amounts of transactional data from online sales platforms, billing systems, digital marketing campaigns and customer loyalty programs. But many organizations still work with old reporting systems that mostly supply descriptive statistics like total sales, monthly revenue and number of transactions, despite this richness of data. Such approaches are limited in terms of insights since they determine the past performance, but do not predict future customer behaviour or profitability in the long term.

In a competitive retail environment it is often more expensive to acquire new customers than it is to retain existing customers, therefore it is more important to retain the latter. Customer Lifetime Value (CLV) is one of the important metrics that has come about to help companies estimate the total revenue that a customer is expected to generate over the course of their relationship with the company. By predicting CLV accurately, organizations can identify valuable customers, allocate marketing resources effectively, and develop tailored engagement strategies. However, forecasting CLV is difficult due to variations in customer behaviour, sporadic purchasing habits, and complex connections in retail data.

To tackle these challenges, this project proposes a Retail Analytics and Customer Intelligence Platform that integrates data preprocessing, customer segmentation, predictive analytics and explainable artificial intelligence techniques. The system converts raw transactional data into meaningful analytical features by using Recency, Frequency, and Monetary (RFM) model, which is effective in capturing customer purchasing behaviour. The platform converts unstructured data into structured behavioral indicators to provide a reliable basis for machine learning analysis. Machine learning algorithms such as Linear Regression, Random Forest and XG Boost are used to predict Customer Lifetime Value and identify the most accurate predictive model. Among these techniques, XG Boost performed better with a high R2 score, which indicates good predictive ability. Customers are classified into Low, Medium, High value segments based on the predicted CLV values.

In conclusion, the proposed platform allows retailers to convert raw transactional data into actionable business intelligence to facilitate better decision making and profitability.

©IJRASET: All Rights are Reserved

II. LITERATURE REVIEW

Customer analytics and prediction of Customer Lifetime Value (CLV) are now important research fields in retail data mining and marketing analytics. The availability of large transactional datasets has helped the widespread use of machine learning techniques to analyse customer behaviour, segment customers and predict future revenue contributions.

The application of data mining techniques for customer analytics and retail sales forecast has been the subject of several research. Customer Lifetime Value is a crucial strategic statistic that aids businesses in identifying valuable clients and effectively allocating marketing resources, according to a study by Kumar and Reinartz [1]. Their research demonstrates how CLV-based analysis can enhance long-term profitability and client retention tactics.

In recent years, there has been a lot of interest in research on machine learning-based retail sales forecast. Using the BigMart dataset, a study by Gupta et al. examined the use of machine learning algorithms for retail sales prediction. Because ensemble learning techniques like Random Forest can capture non-linear correlations in transactional data, they perform better than traditional regression techniques, according to the authors' comparison of several regression models [2].

In marketing analytics, customer segmentation utilizing RFM (Recency, Frequency, Monetary) analysis has also been extensively researched. Hughes suggested the RFM framework as a useful method for determining valuable clients by looking at their purchase habits [3]. By classifying clients into high-value, medium-value, and low-value groups, this method assists companies in developing more focused marketing campaigns. Researchers have investigated the application of deep learning and advanced analytics in marketing applications in addition to conventional data mining techniques. Chaudhuri and Dayal talked about how using big datasets and sophisticated machine learning models, predictive analytics and data-driven decision systems might increase marketing effectiveness [4]. Their research emphasizes how crucial it is to combine business intelligence tools with predictive models in order to improve customer insights. Explainable artificial intelligence (XAI) approaches to enhance machine learning models' interpretability have also been the subject of recent research. A methodology called SHAP (Shapley Additive Explanations) was developed by Lundberg and Lee to consistently explain machine learning predictions [5]. This method enables businesses to comprehend how various characteristics affect model predictions, which is crucial for consumer analytics applications.

In a similar vein, Ribeiro et al. introduced the LIME (Local Interpretable Model-Agnostic Explanations) approach, which clarifies each complex machine learning model's prediction [6]. These methods assist decision-makers trust automated predictions made by AI systems and increase model transparency. Many current systems prioritize descriptive analysis over predictive analytics, despite advancements in machine learning-based customer analytics. For efficient retail analytics, integrated solutions that blend interactive visualization, predictive modeling, and customer segmentation are therefore required. In order to close this gap, the suggested approach combines sophisticated machine learning models like Random Forest and XG Boost with RFM-based feature engineering to precisely forecast Customer Lifetime Value.

III. METHODOLOGY

By using machine learning approaches to retail transactional data, the proposed system seeks to forecast Customer Lifetime Value (CLV). Data collection, preprocessing, model construction, and model evaluation are all part of the methodology's organized pipeline. These procedures guarantee that the unprocessed retail data is converted into insightful analysis and precise forecasts.

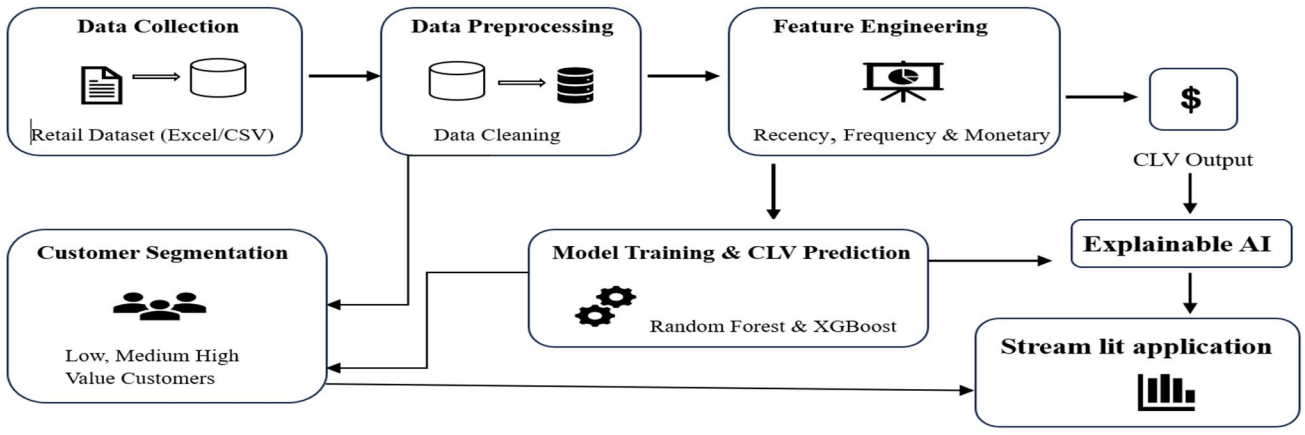


Fig.1. System Architecture

A. Data Acquisition

Gathering the dataset for analysis and model training is the first stage in the process. A retail transactional dataset with consumer purchase records was employed in this investigation. Customer ID, invoice number, product description, quantity, unit price, invoice date, and country are among the attributes included in the dataset. Customers past purchases throughout time are represented in this dataset. For additional processing and analysis, the data was imported into the Python environment after being saved in CSV format. Understanding buying trends and calculating the long-term worth of clients depend on historical transaction data.

TABLE I
DATASET ATTRIBUTES

Attribute	Description
Invoice	Transaction ID
Stock Code	Product Code
Description	Product Name
Quantity	Number of Products Purchased
Invoice Date	Transaction Date
Price	Product Price
Customer ID	Unique Customer Identifier
Country	Customer Country

Online retail transactional records with customer purchase history and product details make up the dataset used in this project. Invoice number, product description, quantity purchased, unit price, customer ID, and transaction date are among the transaction details contained in the dataset. These characteristics were used for segmentation, customer lifetime value prediction, and customer behaviour analysis.

TABLE II
SAMPLE DATASET

Invoice	Stock Code	Description	Quantity	Invoice Date	Price	Customer ID	Country	Total Price
489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085	United Kingdom	83.4
489434	79323	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085	United Kingdom	81
489434	79323	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085	United Kingdom	81
489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.1	13085	United Kingdom	100.8
489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085	United Kingdom	30

B. Data Preprocessing

Model performance may be impacted by missing values, duplicate records, and inconsistent data seen in raw retail datasets. In order to clean and get the dataset ready for analysis, data preparation is carried out. This phase involved handling missing values and eliminating invalid records, such as transactions with negative amounts or prices. In order to preserve data consistency, duplicate entries were also removed. To enable time-based analysis, the Invoice Date column was changed to datetime format. Quantity and Unit Price, which stand for the revenue from each transaction, were multiplied to create a new variable called Total Price. Furthermore, client purchase behaviour was represented by RFM characteristics (Recency, Frequency, and Monetary value)[7].

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

TABLE III
RFM FEATURES

Customer ID	Recency	Frequency	Monetary
12345	10	5	2500
12346	30	2	800

C. Model Development

Machine learning models were created to forecast Customer Lifetime Value following preprocessing and feature engineering. To construct and evaluate the models, the dataset was separated into training (70 %) and testing (30%) sets.
In this study, three regression algorithms were used:

- Linear Regression
- Random Forest Regressor
- XG Boost Regressor

While Random Forest and XG Boost were employed to identify intricate patterns in consumer purchase behaviour, Linear Regression served as a baseline model. The RFM features and other derived qualities were used as input variables to train these models.

D. Model Evaluation

Standard regression assessment measures were used to assess the created models' performance. These metrics assess how well projected values match the dataset's actual values.
The evaluation metrics used include:

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 Score (Coefficient of Determination)

IV. IMPLEMENTATION

Python-based machine learning and data analysis libraries were used to create the suggested Retail Analytics and Customer Intelligence Platform. Pandas and NumPy were first used to gather and process the retail transaction dataset. To enhance the dataset's quality, data preprocessing methods like resolving missing values, eliminating cancelled transactions, converting date variables, and producing total purchase values were used. Important customer behaviour measures, such as Recency, Frequency, Monetary value (RFM), Average Order Value (AOV), and Customer Lifetime Value (CLV), were then created through feature engineering.
Following preprocessing, machine learning models for CLV prediction and client segmentation were created. Customers were divided into various categories according to their shopping habits using K-Means clustering. The prepared dataset was used to train and evaluate different regression models, such as Random Forest, XG Boost, and Linear Regression, for CLV prediction. To determine the most accurate model, metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score were used to assess model performance [8],[9].
The performance of the implemented models was compared using the outcomes of these indicators. With the lowest prediction error and the greatest R2 score among the assessed models, XG Boost had the best performance, demonstrating its efficacy in predicting Customer Lifetime Value.

Table IV
Model Comparison

Model	RMSE	R^2 Score
Linear Regression	266595	-943.62
Random Forest	1814	0.956
XG Boost	1008	0.986

In order to visualize customer data, sales trends, model comparison findings, and forecast outputs, an interactive dashboard was developed using Streamlit. Through graphical representations, the dashboard enables users to forecast revenues, examine client groupings, and comprehend prediction explanations.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

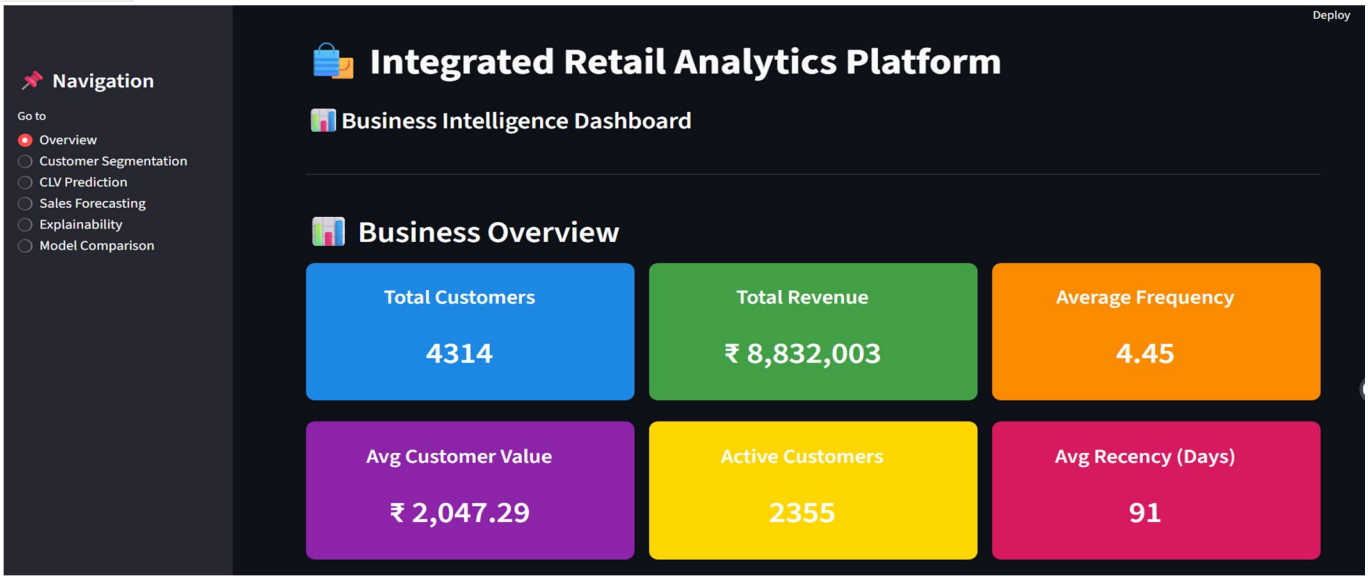


Fig. 2 Interactive Retail Analytics Dashboard

V. RESULTS & DISCUSSION

The outcomes of the suggested system show that machine learning methods are capable of accurately predicting Customer Lifetime Value and analysing consumer behaviour. Performance metrics including MSE, RMSE, and R2 Score were used to train and assess several models, including Random Forest, XG Boost, and Linear Regression. With the highest prediction accuracy and the lowest error values across these models, XG Boost performed the best. Clear insights into customer segmentation, sales trends, and business performance were also supplied by the created dashboard outputs and visualizations, which aided in improving comprehension of consumer decision-making and purchase habits.

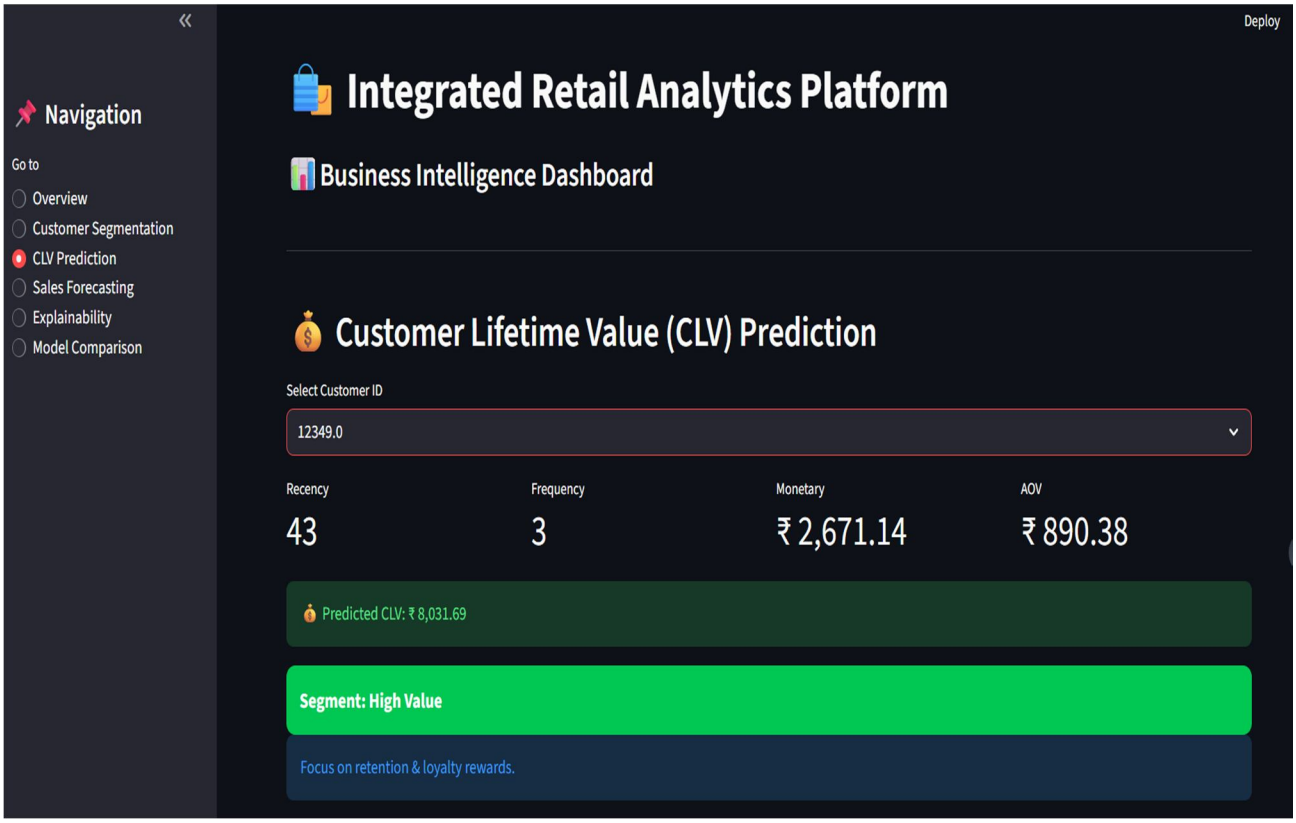


Fig. 3 Prediction of CLV for High Value Customers

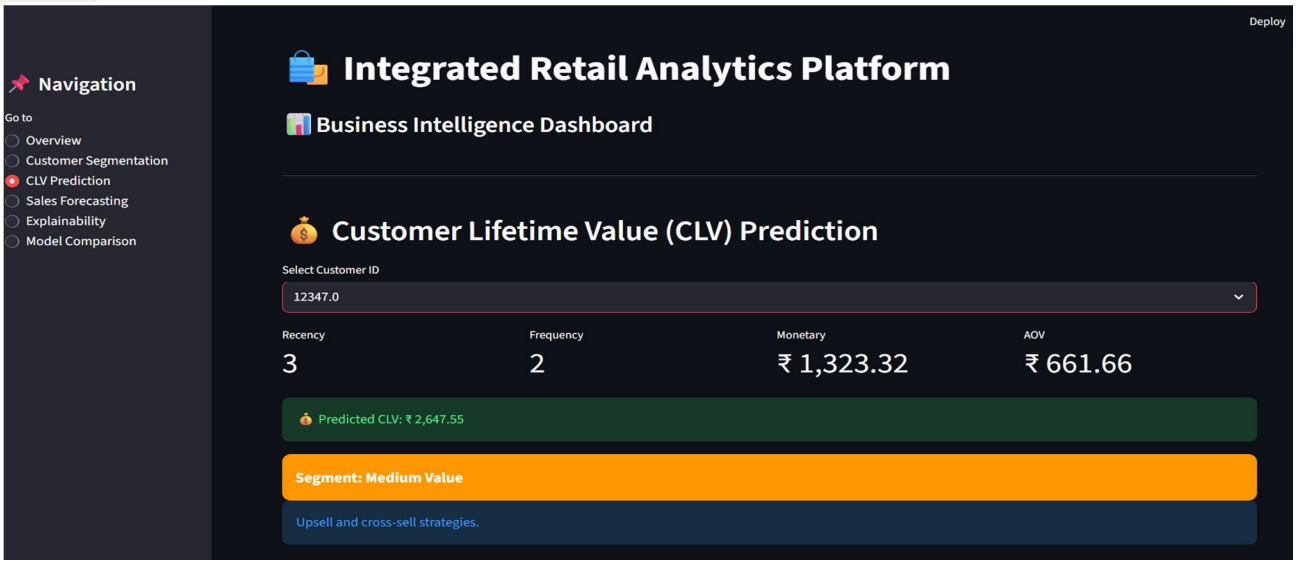


Fig. 4: Prediction of CLV for Medium Value Customers

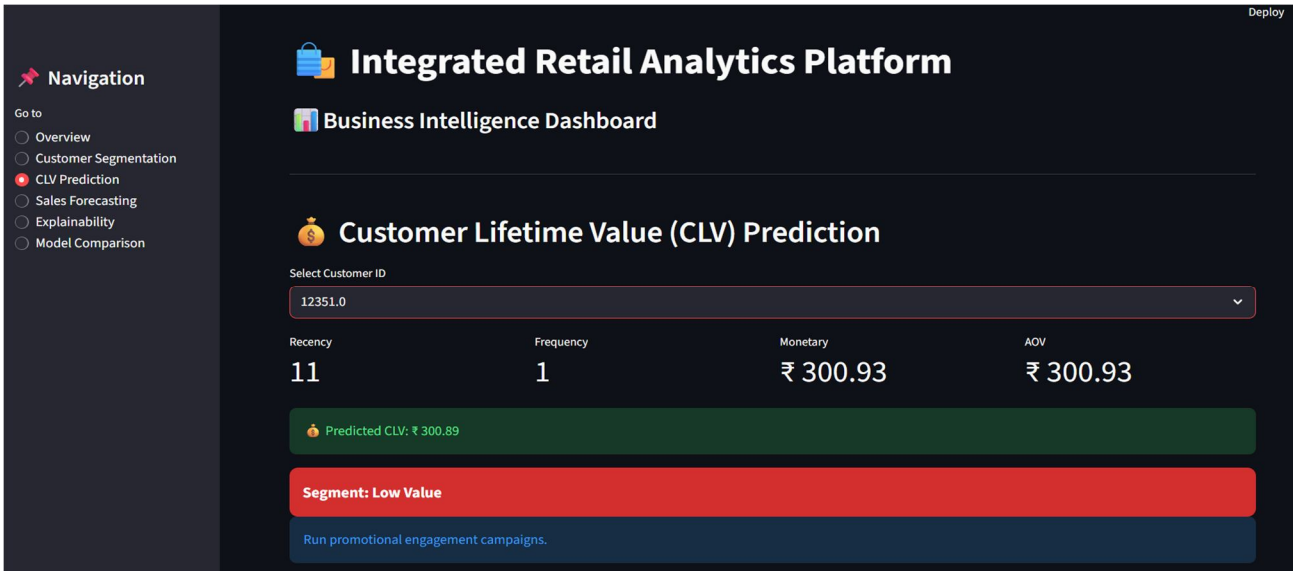


Fig. 5 Prediction of CLV for Low Value Customers

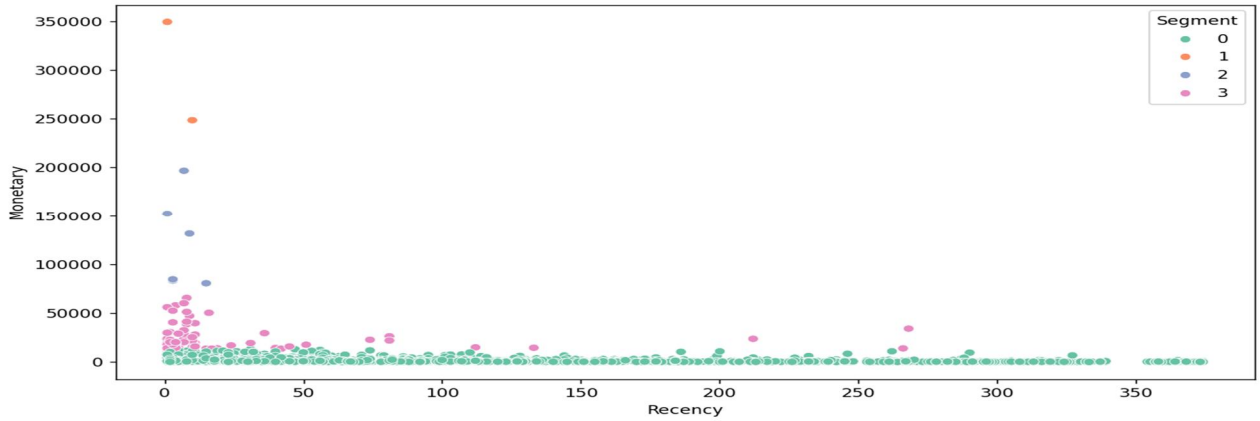


Fig. 6 Customer Segmentation Using RFM Features

	Recency	Frequency	Monetary	AOV	SHAP Explanation	LIME Explanation
0	165	11	372.86	33.89636	Customer shows high predicted CLV due to strong Frequency, AOV while lower values in Recency, Monetary slightly reduce it.	LIME analysis indicates that Frequency, 283.94, AOV, Recency significantly influence the prediction.
1	3	2	1323.32	661.66	Customer shows high predicted CLV due to strong Recency, Monetary, AOV while lower values in Frequency slightly reduce it.	LIME analysis indicates that Monetary, 1.00, AOV, Recency significantly influence the prediction.
2	74	1	222.16	222.16	Customer shows high predicted CLV due to strong AOV while lower values in Recency, Frequency, Monetary slightly reduce it.	LIME analysis indicates that Monetary, Frequency, 58.00, 171.86 significantly influence the prediction.
3	43	3	2671.14	890.38	Customer shows high predicted CLV due to strong Recency, Monetary, AOV while lower values in Frequency slightly reduce it.	LIME analysis indicates that Monetary, 2.00, AOV, 23.00 significantly influence the prediction.
4	11	1	300.93	300.93	Customer shows high predicted CLV due to strong Recency while lower values in Frequency, Monetary, AOV slightly reduce it.	LIME analysis indicates that Frequency, 283.94, Recency, 266.21 significantly influence the prediction.
5	11	2	343.8	171.9	Customer shows high predicted CLV due to strong Recency, Frequency while lower values in Monetary, AOV slightly reduce it.	LIME analysis indicates that 283.94, 171.86, 1.00, Recency significantly influence the prediction.
6	44	1	317.76	317.76	Customer shows while lower values in Recency, Frequency, Monetary, AOV slightly reduce it.	LIME analysis indicates that Frequency, 283.94, 23.00, 266.21 significantly influence the prediction.
7	203	1	488.21	488.21	Customer shows high predicted CLV due to strong AOV while lower values in Recency, Frequency, Monetary slightly reduce it.	LIME analysis indicates that Frequency, 283.94, AOV, Recency significantly influence the prediction.

Fig.7 SHAP & LIME Explanations

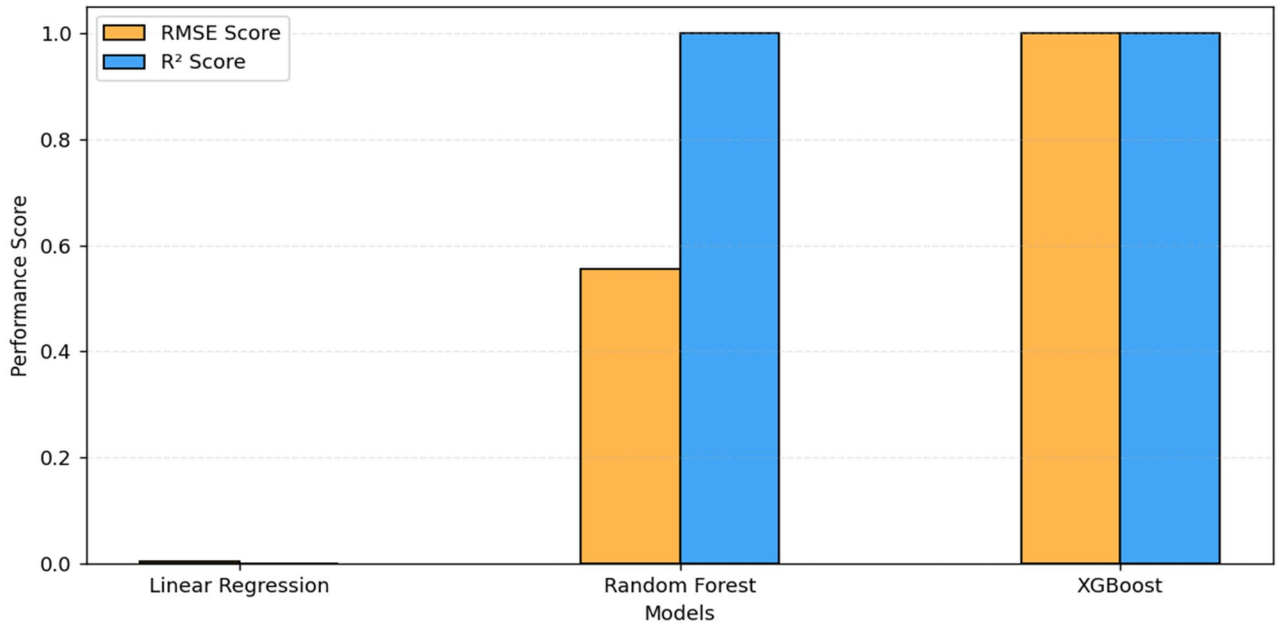


Fig.8 Model Comparison Graph

The above graph displays a near zero performance score for Linear Regression due to a very high RMSE value and a negative R2 score when compared to the other models. This indicates that the model was unable to capture complex non-linear relationships on the retail dataset. Random Forest and XG Boost performed much better on the prediction side, yielding much higher scores in the comparison graph.

VI. CONCLUSION

The proposed Retail Analytics and Customer Intelligence Platform successfully shows how machine learning can be used to analyse customer behaviour and improve business decision-making. The system processes retail transaction data, performs customer segmentation, predicts customer lifetime value, and analyses future sales trends. Different machine learning models were trained and compared using performance metrics such as RMSE and R^2 Score. Among all the models, XG Boost produced the best prediction results with higher accuracy and lower error values.

The project also includes an interactive Streamlit dashboard that helps users easily visualize customer insights, sales performance, and prediction results. Explainable AI techniques like SHAP and LIME were used to make the predictions more understandable and transparent. Overall, the system helps businesses identify valuable customers, improve customer retention strategies, and make better data-driven decisions for business growth.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

VII. FUTURE WORK

The proposed framework can be extended by incorporating interactive Power BI dashboards for real-time visualization and business decision support. Future research may focus on the development of Agentic AI-powered assistants to automate retail operations, including inventory monitoring, customer engagement, demand forecasting, and report generation. By integrating autonomous AI agents, we can significantly reduce manual efforts, improve operational efficiency, and accelerate decision-making processes.

VIII. ACKNOWLEDGMENT

I express my sincere gratitude to my guide, Dr. B. Kranti Kiran, Professor, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Hyderabad, for his valuable guidance, continuous support, and constructive suggestions throughout the development of this research work.

I also extend my heartfelt thanks to the faculty members of the Department of Computer Science and Engineering for their support and motivation during the course of this study.

REFERENCES

[1] V. Kumar and W. Reinartz, Customer Relationship Management: Concept, Strategy, and Tools, 3rd ed. Berlin, Germany: Springer, 2018. DOI: https://doi.org/10.1007/978-3-662-55381-7

[2] S. Gupta, H. Hanssens, B. Hardie, W. Kahn, V. Kumar, N. Lin, N. Ravishanker, and S. Sriram, "Modeling customer lifetime value," Journal of Service Research, vol. 9, no. 2, pp. 139–155, Nov. 2006.DOI: https://doi.org/10.1177/1094670506293810

[3] A. M. Hughes, Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable, Customer-Based Marketing Program, 4th ed. New York, NY, USA: McGraw-Hill, 2011.

[4] S. Chaudhuri and U. Dayal, "An overview of data warehousing and OLAP technology," ACM SIGMOD Record, vol. 26, no.1, pp. 65–74, Mar. 1997.DOI: https://doi.org/10.1145/248603.248616

[5] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.DOI: https://doi.org/10.48550/arXiv.1705.07874

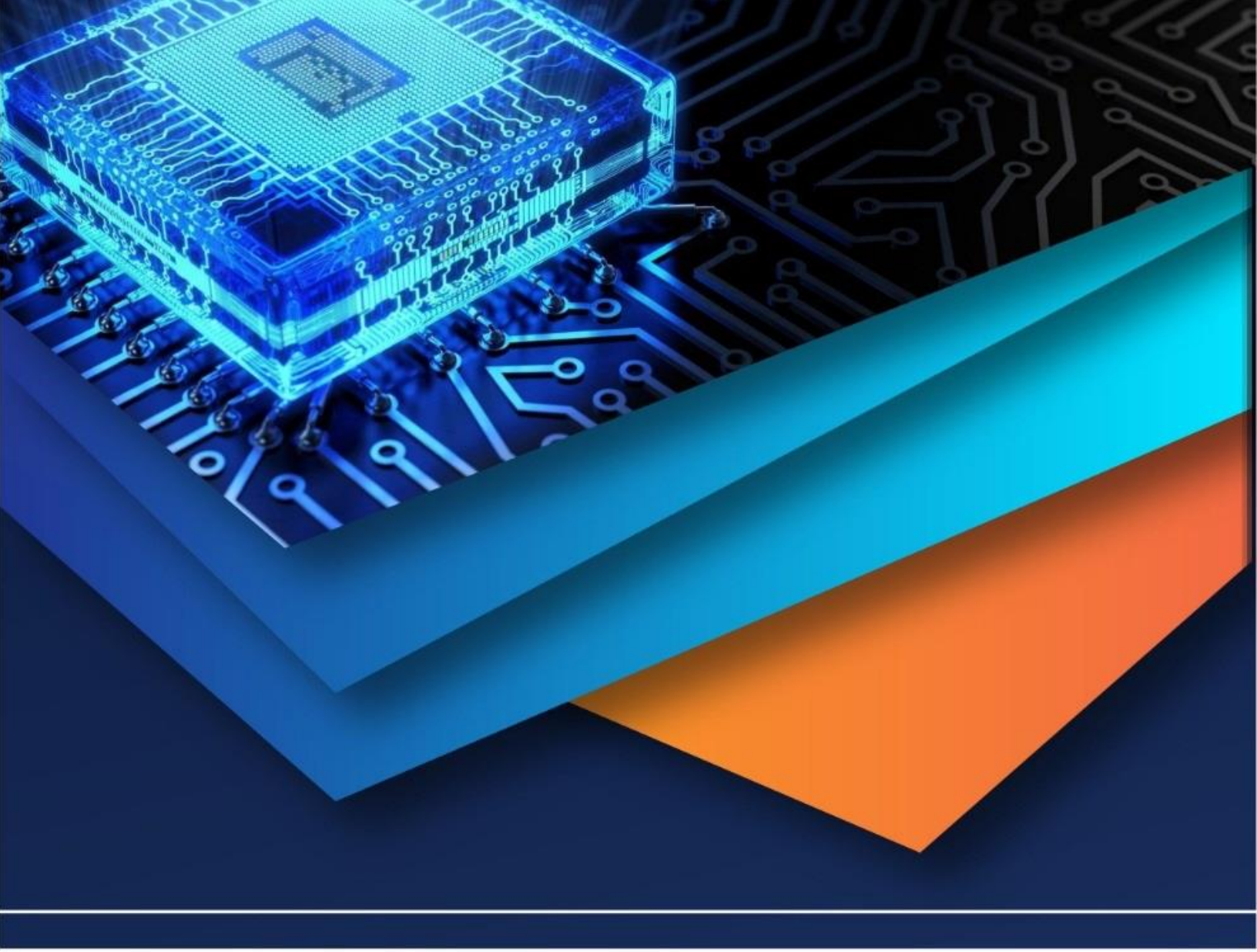
[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.DOI: https://doi.org/10.1145/2939672.2939778

[7] P. S. Fader, B. G. S. Hardie, and K. L. Lee, "RFM and CLV: Using Iso-Value Curves for Customer Base Analysis," Journal of Marketing Research, vol. 42, no. 4, pp. 415–430, Nov. 2005.DOI: https://doi.org/10.1509/jmkr.2005.42.4.415

[8] M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh, "Estimating Customer Future Value of Different Customer Segments Based on RFM Model in Retail Industry," Procedia Computer Science, vol. 3, pp. 1327–1332, 2011.DOI: https://doi.org/10.1016/j.procs.2011.01.050

[9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.
DOI: https://doi.org/10.1145/2939672.2939785

[10] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. DOI: https://doi.org/10.1023/A:1010933404324



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)