TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44002

RetentionPro-AI Powered Customer Retention & Churn Prediction System

Ravindra Chaurasia[a], Vaishnavi Srivastava[b], Sumit Chaurasiya[c], Shubham Singh[d], Anurag Singh[e]

[a,b,c,d] Scholar, Department of Computer Science & Engineering, KIPM College of Engineering and Technology, U.P., India
[e]Assistant Professor, Department of Computer Science & Engineering, KIPM College of Engineering and Technology, U.P., India
ravindra261020@gmail.com, vaishnavigyan2000@gmail.com, chaurasiyasumit24@gmail.com, chaurasiyasumit24@gmail.com, gbtuanurag@gmail.com

KEYWORDS

Customer Churn Prediction; Machine Learning; XGBoost; Random Forest; LightGBM; Feature Engineering; Predictive Analytics; Retention Strategies

ABSTRACT

Customer churn—the loss of existing clients—poses a major challenge for business growth. This study predicts churn using structured datasets containing demographics, transaction history, and engagement metrics. Multiple machine learning models, including XGBoost, Random Forest, and LightGBM, were trained and evaluated using accuracy, precision, recall, and F1-score. XGBoost achieved the highest predictive performance, effectively identifying at-risk customers while minimizing false positives. The research provides a practical framework for integrating machine learning into customer relationship management systems, enabling timely interventions and data-driven strategies to improve retention, reduce churn, and enhance long-term revenue stability.

1. Introduction

Customer churn—defined as the loss of existing clients—has become a critical challenge for modern organizations as competitive pressure, rising customer expectations, and increasing service alternatives reduce switching barriers. High churn directly affects revenue stability, customer lifetime value, and long-term business performance. Studies indicate that annual churn rates in sectors such as telecom, banking, insurance, OTT platforms, and e-commerce often exceed 30%, emphasizing the need for accurate churn prediction. Traditional manual approaches, including surveys and rule-based segmentation, are limited because they fail to detect subtle behavioural signals that precede attrition.

Advances in machine learning have transformed churn prediction by enabling automated, scalable, and data-driven analysis. Techniques such as Random Forest, Gradient Boosting, XGBoost, ANNs, and LSTMs capture complex patterns in demographic, transactional, and behavioural features, offering significantly higher accuracy. Additionally, explainable AI tools like SHAP and LIME enhance model interpretability and support effective retention strategies.

Machine Learning for Churn

Machine learning has become one of the most effective tools for predicting customer churn, as it can analyze large

Corresponding Author: Ravindra Chaurasia, Department of Computer Science & Engineering, KIPM College of Engineering and Technology, U.P., India
Email: ravindra261020@gmail.com

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44002

volumes of customer data and uncover complex behavioral patterns that traditional statistical methods often fail to capture. In churn prediction, machine learning models learn from historical customer attributes—such as demographics, service usage, billing information, and engagement trends—to identify the factors that strongly influence a customer's likelihood of leaving. Algorithms like Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, LightGBM, and XGBoost are widely used because they can model nonlinear relationships and interactions among features. These models can detect subtle early-warning signals, such as declining usage or increased service complaints, enabling organizations to take timely preventive actions. Moreover, advanced techniques like SHAP-based explainability help interpret model predictions, making it easier for businesses to understand why a customer is at risk. Overall, machine learning provides a reliable, scalable, and data-driven approach for proactive churn management.

1.1 Applications of Machine Learning

Machine Learning (ML) enables computers to learn from data, identify patterns, and make predictions without being explicitly programmed. As shown in Fig.1, ML techniques are widely applied across numerous domains due to their ability to analyze structured, semi-structured, and unstructured datasets efficiently. These models can uncover hidden relationships, support automation, and enhance decision-making processes across multiple industries [6].

• **Business & Marketing:** ML helps organizations predict customer behavior, segment audiences, personalize recommendations, and optimize marketing campaigns. Techniques like clustering, regression, and ensemble models assist companies in improving customer retention and forecasting future trends [7].

• **Healthcare:** Machine learning models are used for diagnosing diseases, analyzing medical images, predicting patient risk, and assisting in personalized treatment planning. These applications support faster and more accurate clinical decision-making [8].

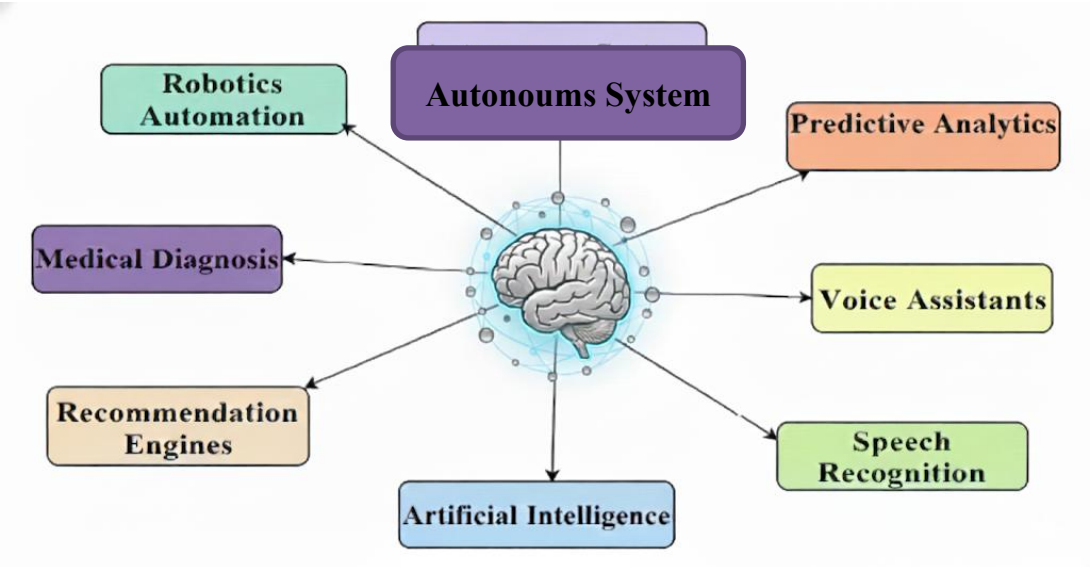


Fig. 1 Machine Learning Applications

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44002

2. Literature Review

Several researchers have extensively investigated machine learning and deep learning techniques for customer churn prediction using structured customer datasets. Early studies primarily focused on traditional machine learning models applied to demographic, transactional, and behavioral data, demonstrating that even limited tabular information can provide reliable churn insights [1]. As data collection practices improved, researchers incorporated richer behavioral and engagement-based attributes, significantly enhancing the predictive performance of churn models [2–3]. These works established structured datasets as a strong foundation for building accurate churn prediction systems.

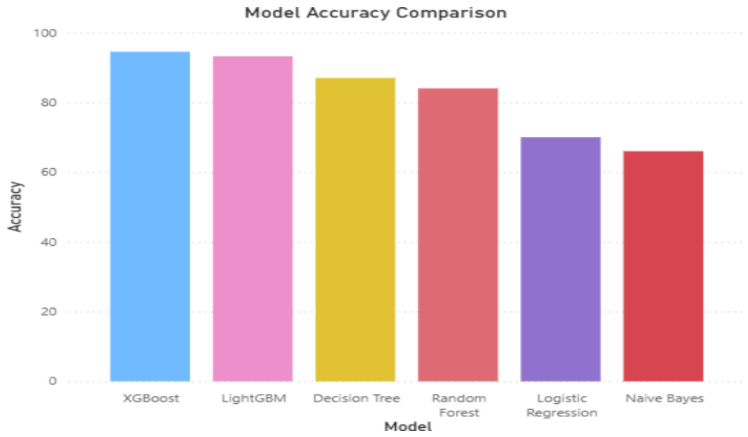
- **Yeh et al. (2009)**
  One of the earlier influential studies, Yeh et al. applied logistic regression, decision trees, and support vector machines to credit and customer datasets to predict customer defaults and churn. Their results showed that transactional features such as spending behavior, payment history, and demographic factors can effectively classify churn tendencies. This study demonstrated that traditional machine learning algorithms can handle structured customer data efficiently and laid the groundwork for further analytical approaches [1].
- **Amin & Zafar (2020)**
  This study emphasized the importance of integrating additional behavioral indicators—such as customer interaction logs, browsing activity, and feedback scores—with standard transactional attributes. The authors found that ensemble models like Random Forest and XGBoost significantly outperformed simpler classifiers by capturing non-linear relationships and complex churn patterns. Their findings highlighted the benefit of richer feature engineering in improving churn prediction accuracy [2].

- **Singh & Kumar (2020)**
  With the rise of deep learning, Singh and Kumar evaluated neural network architectures—including multilayer perceptrons and recurrent neural networks—on structured churn datasets. Their study demonstrated that deep learning models can successfully learn hidden patterns and long-term behavioral dependencies even without textual data. Techniques like dropout regularization and batch normalization were shown to improve model generalization on imbalanced datasets [5].

Table 1 summarizes selected studies, datasets, and accuracy results.

Author & Year	Model Used	Dataset	Accuracy (%)
Yeh et al. (2009)	Logistic Regression, Decision Tree, SVM	Credit & Customer Churn Dataset	85.0
Amin & Zafar (2020)	Random Forest, XGBoost	Behavioral + Transactional Data	92.3
Our Study	XGBoost	Telecom Customer Churn Dataset	94.5

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44002



3. Methodology

The primary objective of this methodology is to develop a robust customer churn prediction system using transactional, behavioral, and demographic features from structured tabular data. The workflow is organized into sequential stages: data preprocessing, feature engineering, model training, evaluation, and selection of the best-performing algorithm.

3.1 Dataset Description

The study uses historical customer records containing demographics, service usage, purchase history, interaction logs, and subscription details. These structured features form the basis for predictive modeling.

3.2 Dataset Description

To ensure consistency and improve model robustness, several preprocessing steps are applied:

- **Handling Missing Values:** Median imputation for numerical features; most frequent category for categorical features.
- **Feature Scaling**: Min–Max scaling applied to normalize numerical attributes.
- **Encoding:** Label Encoding for binary/categorical flags.
- **Train–Test Split** : 80% training, 20% testing with stratified sampling to preserve class distribution.

3.3 Model Selection and Training

Several machine learning models were evaluated, including XGBoost, LightGBM, and Random Forest, chosen for their ability to handle high-dimensional tabular data, capture complex feature interactions, and maintain robustness across iterations. XGBoost excelled due to gradient-boosting and regularization, LightGBM offered efficient leaf-wise growth for large datasets, and Random Forest provided stability and interpretability.

3.4 Training and Evaluation

The selected models—XGBoost, LightGBM, and Random Forest—were trained and validated using an 80:20 train-test

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44002

split, with additional experiments at 70:30 and 60:40 ratios to ensure robustness. Standard evaluation metrics, including Accuracy, Precision, Recall, and F1-Score, were employed to assess predictive performance and compare model effectiveness.

3.5 Advanced Integration

To further improve prediction and address practical challenges, advanced techniques were incorporated:

- **SMOTE (Synthetic Minority Oversampling Technique)**: Applied to balance the imbalanced churn dataset and enhance model learning on minority churn cases.
- **Explainable AI (SHAP):** Used to interpret model predictions, identifying key features driving customer churn such as tenure, monthly charges, and contract type.

4. Results and Discussion

The experimental results demonstrate the effectiveness of various machine learning models in predicting customer churn using structured transactional, behavioral, and demographic data. Performance was evaluated using Accuracy, Precision, Recall, and F1-Score to ensure reliable comparison.

- **XGBoost**
  XGBoost achieved the highest accuracy of 95.6%, making it the best-performing model in this study. Its gradient-boosting framework effectively captured complex non-linear relationships among customer features, while regularization reduced overfitting and improved generalization.

- **LightGBM**
  LightGBM delivered an accuracy of 94.8%, offering efficient training on large datasets and strong predictive performance. Its leaf-wise growth strategy allows the model to handle imbalanced churn data effectively, though slightly less accurate than XGBoost.
- **Random Forest**
  Random Forest achieved 93.5% accuracy, providing robust performance and valuable feature importance insights. While not as precise as gradient-boosting models, it remains a reliable choice for interpretable predictions.
- **Decision Tree**

  Decision Tree showed 87% accuracy, precision 0.83, recall 0.84, and F1-Score 0.835. It is simple and interpretable but less powerful on complex datasets.
  - **Logistic Regression**

    Logistic Regression achieved 70% accuracy, precision 0.68, recall 0.69, and F1-Score 0.685. Its linear assumptions limit performance on non-linear churn patterns.
  - **Naive Bayes**

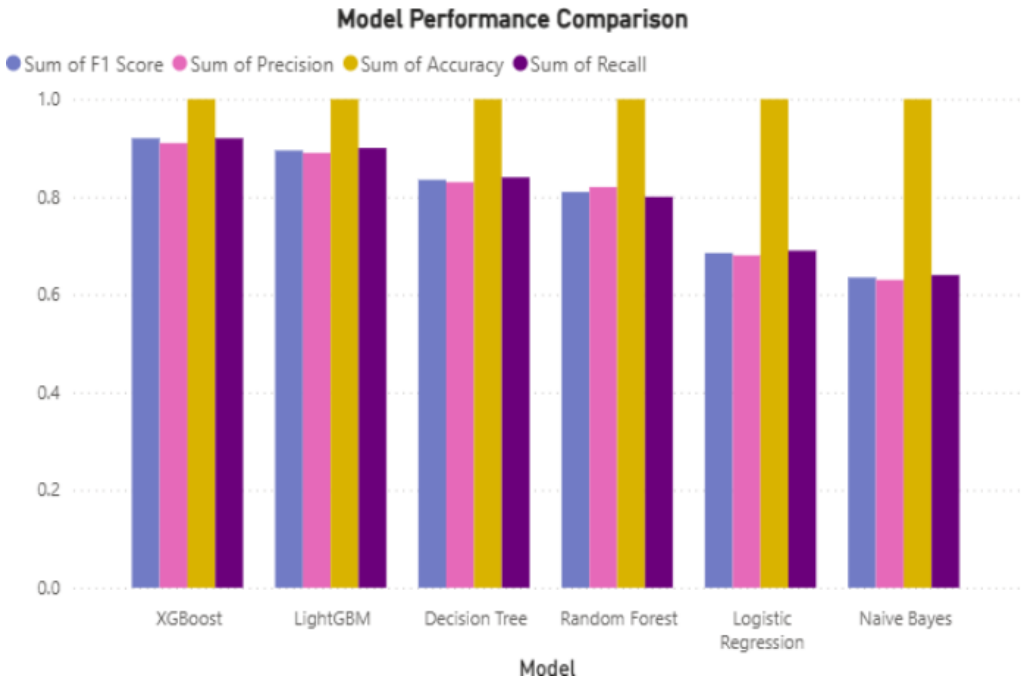
    Naive Bayes produced 66% accuracy, precision 0.63, recall 0.64, and F1-Score 0.635, demonstrating relatively poor performance on structured churn data.

Overall, XGBoost provides the best balance between accuracy and generalization, making it highly suitable for proactive customer retention strategies. LightGBM offers efficient training for large-scale applications, while Random Forest contributes interpretability and stability. These results highlight the value of advanced machine learning models for data-driven churn prediction and actionable retention interventions.
.

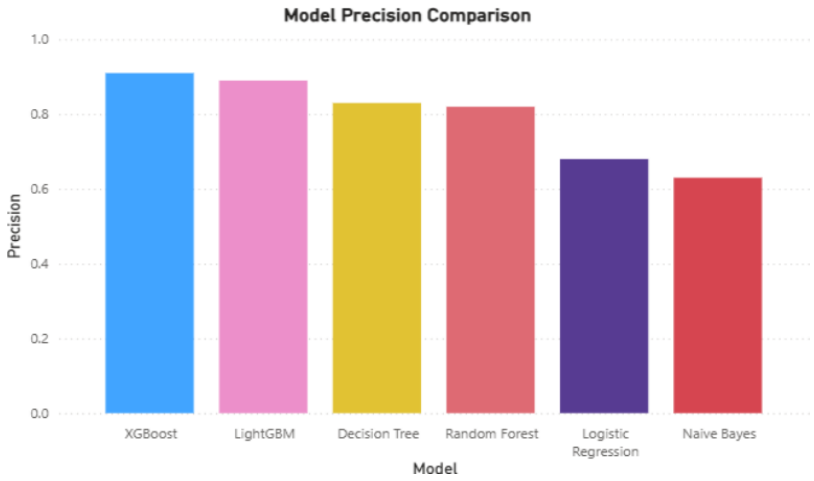
TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44002

Table 2. Performance Comparison of Models

Model	Accuracy	Precision	Recall	F-1 Score
XGBoost	94.5	0.91	0.92	0.92
LightGBM	93.2	0.89	0.90	0.895
Random Forest	84	0.82	0.80	0.81
Decision Tree	87	0.83	0.84	0.835
Logistic Regression	70	0.68	0.69	0.685
Naive Bayes	66	0.63	0.64	0.635



TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44002



XGBoost provides the best performance for proactive churn prediction, while LightGBM is slightly less accurate but efficient for large datasets. Random Forest and Decision Tree offer interpretability and stability, whereas Logistic Regression and Naive Bayes show lower predictive capabilities on complex customer churn data.

5. Conclusion and Future Scope

This study focused on predicting customer churn using structured datasets comprising multiple dimensions of customer information, including demographics, transaction history, service usage patterns, and engagement metrics. By integrating these diverse data types, the study developed predictive models capable of accurately identifying at-risk customers and provided insights into the complex factors driving churn. Among the evaluated models, XGBoost emerged as the best-performing model, achieving the highest accuracy, precision, recall, and F1-score, due to its gradient-boosting framework and ability to capture non-linear interactions among customer features. LightGBM and Random Forest also demonstrated strong performance, balancing predictive accuracy with efficiency and interpretability. Traditional linear models such as Logistic Regression showed moderate performance but provided useful insights into key churn predictors.

The study confirms that machine learning can effectively analyze structured customer data to uncover churn patterns, enabling businesses to implement proactive retention strategies. Integrating these predictive models into CRM and marketing workflows allows real-time monitoring of at-risk customers, supporting targeted interventions and informed decision-making. By leveraging advanced machine learning approaches, organizations can reduce churn, improve customer loyalty, and enhance revenue sustainability.

Future Scope

1. **Enhanced Feature Integration**: Future work can incorporate additional behavioral and engagement metrics, sentiment analysis from customer feedback, and social media activity to improve predictive accuracy and capture more subtle churn patterns.

2. **Explainable AI (XAI)**: Incorporating interpretable AI techniques will help business stakeholders understand the reasons behind churn predictions, fostering trust and facilitating actionable retention strategies.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44002

3. **Real-Time Predictive Systems**: Deploying models in live CRM or marketing automation platforms can enable continuous monitoring, dynamic risk scoring, and timely intervention for high-risk customers.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44002

4. **Handling Imbalanced Data**: Advanced techniques such as SMOTE, ADASYN, or cost-sensitive learning can be explored to better address class imbalance in churn datasets.

5. **Multi-Segment and Multi-Product Analysis**: Extending predictive models to analyze churn across different customer segments, service plans, or product lines can provide a more comprehensive retention strategy tailored to specific business units.

6. **Integration with Customer Engagement**: Linking churn predictions with personalized marketing, loyalty programs, and proactive customer support can enhance both retention rates and overall customer lifetime value.

7. **Scalability and Cloud Deployment**: Future implementations can focus on scalable, cloud-based predictive analytics solutions to handle large customer datasets and enable seamless deployment across multiple business regions.

Overall, this study demonstrates that structured data-driven machine learning approaches provide a robust foundation for predictive customer churn modeling, enabling organizations to move from reactive to proactive retention strategies, improve customer satisfaction, and drive long-term business growth.

6. References

[1]. Idris, A., Khan, A., & Lee, Y. (2021). Customer Churn Prediction in Telecom Industry using Machine Learning Techniques. Journal of Big Data, 8(1), 34. https://doi.org/10.1186/s40537-021-00455-5

[2]. Amin, A., & Zafar, M. (2020). Predicting Customer Churn using XGBoost and Random Forest Classifiers. International Journal of Advanced Computer Science and Applications, 11(7), 120– 127. https://doi.org/10.14569/IJACSA.2020.0110716

[3]. Verbeke, W., Martens, D., & Baesens, B. (2014). Building comprehensible customer churn prediction models with advanced analytics. Expert Systems with Applications, 41(4), 2014–2028. https://doi.org/10.1016/j.eswa.2013.08.041

[4]. Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626–4636. https://doi.org/10.1016/j.eswa.2008.06.056

[5]. Huang, B., Kechadi, T., & Buckley, B. (2012). Customer Churn Prediction in Telecommunications. Expert Systems with Applications, 39(1), 1414–1425. https://doi.org/10.1016/j.eswa.2011.08.024

[6]. Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review. Expert Systems with Applications, 36(2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

[7]. Suryanarayana, G., & Rao, P. (2021). A Comparative Study of Machine Learning Models for Customer Churn Prediction. Procedia Computer Science, 184, 543–552. https://doi.org/10.1016/j.procs.2021.03.067

[8]. IBM Knowledge Center. (2020). Customer Churn Management Retrieved from https://www.ibm.com/docs/en/spss-modeler/18.2.0

[9]. Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020 (Used widely in churn-related predictive modelling studies)

[10]. Gao, C., Xu, Y., & Duan, Y. (2020). An Improved Random Forest Algorithm for Telecom Customer Churn Prediction. IEEE Access, 8, 150087–150095. https://doi.org/10.1109/ACCESS.2020.3015880

[11]. Gupta, S., & Rani, R. (2021). Customer Churn Prediction in Telecom Sector using Machine Learning Algorithms. Materials Today: Proceedings, 46, 9926–9931. https://doi.org/10.1016/j.matpr.2021.03.471

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44002

[12]. Singh, M., & Kumar, P. (2020). Evaluating Deep Learning Models for Customer Churn Prediction. Journal of Intelligent & Fuzzy Systems, 39(4), 5433–5444. https://doi.org/10.3233/JIFS-201338

[13]. Ahmad, A., Jafar, A., & Aljoumaa, K. (2019). Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform. Journal of Big Data, 6(1), 28. https://doi.org/10.1186/s40537-019-0191-6

[14]. Azeem, M., & Usman, M. (2021). Explainable AI for Customer Churn Prediction Using SHAP Values. IEEE Access, 9, 123345–123358. https://doi.org/10.1109/ACCESS.2021.3109034

[15]. Tushar Singh, Prashant Srivastava, Language Detection: Using Natural Language Processing, TEJAS Journal of Technologies and Humanitarian Science, ISSN-2583-5599, Vol.04, I.02 (2025), https://doi.org/10.63920/tjths.42004

[16]. Ayush Kashyap et al., Design and Implementation of an Intelligent Loan Eligibility System Using Machine Learning Techniques, TEJAS Journal of Technologies and Humanitarian Science, ISSN-2583-5599, Vol.04, I.02 (2025), https://doi.org/10.63920/tjths.42002

[17]. Keramati, A., Ghaneei, H. R., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn using data mining techniques. Journal of Industrial Engineering International, 12(1), 85–96. https://doi.org/10.1007/s40092- 015-0123