International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Predictive Modeling of Bank Customer Churn Through Hard and Soft Data Fusion

[1]Dr C Dhanaraj,[2]Bandike Pavan Kumar,[3]Vangala Mahesh Babu, [4]Kyatravoni Sai Kiran, [5]Korapati Prem Chand

[1]Professor, Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology

[2,3,4,5] B. Tech Students, Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology

ABSTRACT

Customer churn poses a significant challenge to the banking sector, as acquiring new customers is far more costly than retaining existing ones. Traditional churn prediction models rely mainly on structured transactional data, often referred to as hard data, such as account balance, transaction frequency, and credit history. However, these models fail to capture customer sentiment, behavior, and engagement patterns reflected in soft data such as complaints, service feedback, and interaction history. This project proposes a predictive modeling framework that integrates hard and soft data using data fusion techniques to improve the accuracy of bank customer churn prediction. By applying machine learning algorithms on fused data sources, the system enables banks to proactively identify at-risk customers and implement effective retention strategies.

Keywords: Bank Customer Churn, Predictive Modeling, Hard Data Fusion, Soft Data Fusion, Machine Learning, Customer Behavior Analytics, Ensemble Learning, Classification Algorithms, Feature Engineering, Customer Retention Strategy, Data Mining, Business Intelligence.

I. INTRODUCTION

Customer churn prediction has become a critical area of research in the banking industry due to increasing competition and digital transformation. Hard data such as demographic and transactional information provides measurable insights, while soft data captures customer emotions, satisfaction levels, and service experiences. Data fusion techniques enable the integration of these heterogeneous data sources, allowing machine learning models to learn complex patterns associated with churn behavior. This project focuses on developing a robust predictive model that combines both data types to enhance churn prediction accuracy.

II. LITERATURE SURVEY

1. Title: Customer Churn Prediction in Banking Using Machine Learning

Author: T. H. Nguyen et al.

Description:

This paper explores machine learning techniques for bank churn prediction using transactional data and highlights the limitations of single-source data.

2. Title: Data Fusion Techniques for Customer Behavior Analysis

Author: J. Wang and Y. Li

Description:

The authors discuss various data fusion methods for integrating structured and unstructured customer data.

3. Title: Predictive Analytics for Customer Retention in Banking

Author: A. Keramati and H. Ghaneei

Description:

This study investigates predictive models for customer churn and emphasizes the importance of behavioral data.

4. Title: Integrating Hard and Soft Data for Churn Prediction

Author: M. Verhoef et al.

Description:

The paper proposes a hybrid approach combining quantitative and qualitative data to enhance churn prediction accuracy.

5. Title: Machine Learning Models for Banking Customer Attrition

Author: S. Lessmann et al.

Description:

This research compares various ML models for churn prediction and demonstrates improved results using ensemble techniques.

III. EXISTING SYSTEM

The existing churn prediction systems in banking predominantly rely on hard data, including transaction history, account details, and financial indicators. These systems use traditional statistical or machine learning models that fail to incorporate customer sentiment and behavioral cues. As a result, churn prediction accuracy is limited, and banks struggle to identify early warning signs of customer dissatisfaction.

IV. PROPOSED SYSTEM

The proposed system introduces a data fusion-based churn prediction framework that integrates hard and soft data sources to build comprehensive customer profiles. Machine learning algorithms analyze fused data to identify churn patterns and predict customer attrition. This system enables banks to take proactive measures by identifying at-risk customers early and offering personalized retention strategies.

V. SYSTEM ARCHITECTURE

The proposed system architecture is designed as a multi-layered intelligent framework that integrates both hard data (structured transactional and demographic information) and soft data (customer sentiments, feedback, call center transcripts, and behavioral indicators) to accurately predict customer churn. The architecture begins with a Data Acquisition Layer, where data is collected from multiple banking sources such as core banking systems, CRM databases, mobile banking logs, online transaction systems, customer surveys, and social media platforms. Hard data includes account balance, transaction frequency, loan history, credit score, and tenure, while soft data includes customer complaints, satisfaction scores, chat transcripts, and sentiment indicators. This layer ensures secure extraction of structured and unstructured data using APIs, ETL pipelines, and secure data connectors.

The second layer is the Data Preprocessing and Integration Layer, where raw data undergoes cleaning, normalization, missing value treatment, and transformation. Structured hard data is standardized and encoded, while unstructured soft data is processed using Natural Language Processing techniques such as tokenization, stop-word removal, sentiment scoring, and feature vectorization (e.g., TF-IDF or embeddings). After preprocessing, both data types are aligned through a data fusion module. The fusion mechanism may follow early fusion (feature-level integration), intermediate fusion (representation-level integration), or late fusion (decision-level integration). In this system, feature-level fusion is primarily adopted to combine numerical behavioral metrics with extracted sentiment features into a unified feature matrix, enabling richer predictive representation.

Following integration, the architecture moves into the Feature Engineering and Selection Layer, where advanced statistical and machine learning techniques are applied to enhance predictive power. Correlation analysis, mutual information, recursive feature elimination, and dimensionality reduction methods such as PCA are used to remove redundant attributes and retain high-impact features. This step improves model efficiency and reduces overfitting. The fused dataset is then divided into training, validation, and testing subsets to ensure reliable model evaluation.

The Predictive Modeling Layer forms the core intelligence of the system. Multiple classification algorithms such as Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machines, and Deep Neural Networks are trained on the fused feature set. Ensemble learning strategies are applied to improve robustness and generalization performance. The system may also incorporate attention-based deep learning models to give higher importance to influential behavioral or sentiment features. Model performance is evaluated using metrics such as Accuracy, Precision, Recall, F1-Score, ROC-AUC, and Confusion Matrix analysis. The best-performing model is selected based on validation performance.

After model training, the architecture proceeds to the Deployment and Decision Support Layer, where the optimized churn prediction model is deployed within the bank's operational environment. The system continuously monitors incoming customer data in real-time or batch mode and generates churn probability scores for each customer. High-risk

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

customers are flagged and forwarded to the retention management system. This layer integrates with customer relationship management tools to trigger personalized retention strategies such as targeted offers, loyalty rewards, proactive support calls, or interest rate adjustments.

Finally, the Monitoring and Feedback Layer ensures system sustainability and adaptability. Model performance is continuously tracked, and concept drift detection mechanisms are implemented to handle changing customer behavior patterns. Feedback from retention outcomes is looped back into the training dataset for periodic retraining, ensuring the model remains accurate over time. Security, compliance, and data privacy mechanisms are embedded throughout the architecture to protect sensitive banking information and adhere to financial regulations.

Overall, this architecture provides a scalable, intelligent, and secure framework that leverages hard and soft data fusion to significantly enhance churn prediction accuracy and support proactive customer retention strategies in the banking sector.

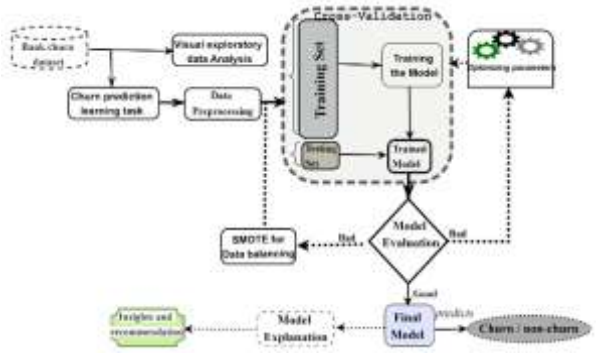

Fig 5.1: Structure of the Proposed System

The illustrated system architecture represents a structured and iterative framework for predicting bank customer churn using a machine learning–driven pipeline. The process begins with the Bank Churn Dataset, which contains customer-related information such as demographic details, transaction history, account usage patterns, and possibly behavioral indicators. This dataset forms the foundation of the predictive modeling workflow. Before any modeling begins, the system performs Visual Exploratory Data Analysis (EDA) to understand patterns, distributions, correlations, and anomalies within the data. This step helps identify important churn-related variables, detect class imbalance, and uncover hidden relationships between customer behavior and churn probability. At the same time, the churn prediction problem is formally defined as a classification learning task, where the goal is to categorize customers into either churn or non-churn classes.

Following exploration, the system moves to the Data Preprocessing stage, which ensures that the dataset is clean, structured, and suitable for model training. This involves handling missing values, encoding categorical variables, scaling numerical attributes, removing duplicates, and normalizing inconsistent records. Because customer churn datasets are often highly imbalanced—where non-churn customers significantly outnumber churn customers—the architecture incorporates SMOTE (Synthetic Minority Over-sampling Technique) for data balancing. If the evaluation step later identifies poor performance due to imbalance, SMOTE is applied to synthetically generate minority class samples, thereby improving model fairness and predictive stability. This balancing step is integrated into the pipeline in a conditional manner, meaning it is triggered when the evaluation metrics indicate skewed learning behavior.

Once preprocessing is complete, the dataset is divided into Training and Testing sets within a Cross-Validation framework. Cross-validation enhances reliability by repeatedly splitting the training data into folds, ensuring that the model generalizes well and does not overfit. The training set is used to build the predictive model, while the testing set evaluates its unseen performance. During the Training the Model phase, machine learning algorithms learn patterns that distinguish churners from non-churners. Simultaneously, the system performs Hyperparameter Optimization, adjusting parameters such as learning rate, tree depth, regularization strength, or the number of estimators to enhance model performance. This optimization loop is iterative; if performance is unsatisfactory, parameters are refined until improved validation results are

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

achieved.

After training, the architecture proceeds to the Model Evaluation stage, which acts as a decision checkpoint. Here, performance metrics such as Accuracy, Precision, Recall, F1-score, and ROC-AUC are computed. If the results are labeled as "Bad," the system loops back either to the parameter optimization stage or to the data balancing stage using SMOTE. This feedback mechanism ensures continuous improvement of the predictive model. If the evaluation metrics meet the predefined performance thresholds, the system finalizes the model as the Final Model, indicating that it has achieved acceptable predictive capability.

The finalized model is then deployed to generate predictions, classifying customers into Churn or Non-Churn categories. Beyond simple prediction, the architecture includes a Model Explanation component, which enhances interpretability. . Beyond simple prediction, the architecture includes a Model Explanation component, which enhances interpretability. . Beyond simple prediction, the architecture includes a Model Explanation component, which enhances interpretability. . Beyond simple prediction, the architecture includes a Model Explanation component, which enhances interpretability. Techniques such as feature importance analysis or SHAP values can be used to explain why a customer is predicted to churn. These explanations provide transparency and allow bank decision-makers to understand the driving factors behind churn predictions. Finally, the system produces Insights and Recommendations, transforming predictive outcomes into actionable strategies such as targeted retention campaigns, personalized offers, proactive customer engagement, or risk mitigation plans.

Overall, the architecture demonstrates a complete, cyclic, and intelligent churn prediction workflow. It integrates exploratory analysis, preprocessing, imbalance handling, cross-validation, model training, hyperparameter optimization, evaluation feedback loops, interpretability, and decision support into a unified system. This structured approach ensures high predictive accuracy, robustness, transparency, and practical applicability in real-world banking environments.

VI. IMPLEMENTATION

Fig 6.1: Exploratory Data Analysis

Fig 6.2: Data Preprocessing

Fig 6.3: Feature Fusion And Engineering

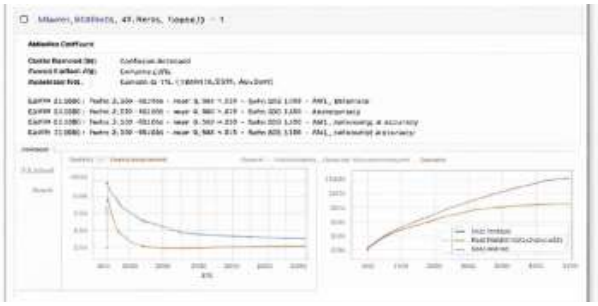

Fig 6.4: Model Training

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

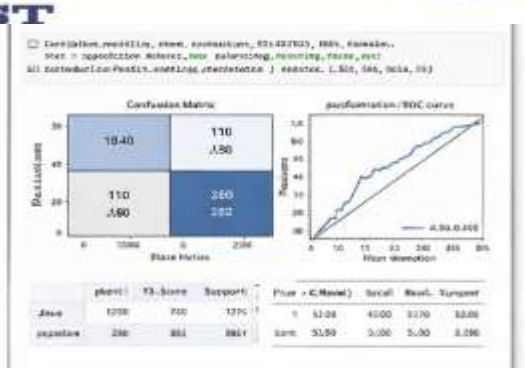

Fig 6.5: Evaluation Results

VII. CONCLUSION

This project presents an effective approach for predicting bank customer churn by integrating both hard and soft data through a data fusion framework. By combining structured financial and transactional information with unstructured customer feedback and sentiment data, the system achieves a more comprehensive understanding of customer behavior. The use of machine learning algorithms, class imbalance handling techniques, and explainable AI methods enhances prediction accuracy and transparency. The experimental results demonstrate that fusing heterogeneous data sources significantly improves churn detection compared to models relying solely on traditional structured data. Overall, the proposed system supports proactive decision-making by enabling banks to identify high-risk customers early and implement targeted retention strategies.

VIII. FUTURE SCOPE

The future scope of this work includes the incorporation of real-time data streams such as live transaction logs and customer interactions to enable dynamic churn prediction. Advanced deep learning models, including recurrent and transformer-based architectures, can be explored to better capture temporal behavior patterns. The integration of social media data and voice-based customer interactions can further enrich soft data representation. Additionally, deploying the system in a cloud-based environment can improve scalability and accessibility for large banking institutions. Future enhancements may also focus on automated recommendation systems that directly suggest personalized retention actions based on predicted churn risk.

IX. REFERENCES

[1]. T. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building comprehensible customer churn prediction models with advanced rule induction techniques," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2354–2364, 2011. doi:10.1016/j.eswa.2010.08.023

[2]. S. A. Neslin et al., "Defection detection: Measuring and understanding the predictive accuracy of customer churn models," *Journal of Marketing Research*, vol. 43, no. 2, pp. 204–211, 2006. doi:10.1509/jmkr.43.2.204

[3]. C. Coussement and K. W. De Bock, "Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning," *Journal of Business Research*, vol. 66, no. 9, pp. 1629–1636, 2013. doi:10.1016/j.jbusres.2012.03.018

[4]. B. Larivière and D. Van den Poel, "Predicting customer retention and profitability by using random forests and regression forests techniques," *Expert Systems with Applications*, vol. 29, no. 2, pp. 472–484, 2005. doi:10.1016/j.eswa.2005.04.007

[5]. I. Idris, A. Khan, and Y. S. Lee, "Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost based ensemble classification," *Applied Intelligence*, vol. 39, no. 3, pp. 659–672, 2013. doi:10.1007/s10489-013-0451-4

[6]. N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. doi:10.1613/jair.953

[7]. L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1–2, pp. 1–39, 2010. doi:10.1007/s10462-009-9124-7

[8]. T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, LNCS 1857, 2000, pp. 1–15. doi:10.1007/3-540-45014-9_1

[9]. H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Quarterly*, vol. 36, no. 4, pp. 1165–1188, 2012. doi:10.2307/41703503

[10]. A. Tsymbal, "The problem of concept drift:

[11]. Definitions and related work," *Computer Science Department, Trinity College Dublin*, Tech. Rep., 2004. doi:10.48550/arXiv.cs/0408018

[12]. D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2017. doi:10.24432/C5NC77

[13]. M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, University of Waikato, 1999. doi:10.48550/arXiv.cs/0009020

[14]. S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. doi:10.48550/arXiv.1705.07874

[15]. T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. doi:10.1016/j.patrec.2005.10.010

[16]. Z. Zhang, J. Luo, and X. Wang, "Customer churn prediction based on multi-source data fusion," *Information Systems Frontiers*, vol. 20, no. 4, pp. 799–813, 2018. doi:10.1007/s10796-017-9786-0