Predictive Analytics and Generative AI for Customer Churn Prediction and Proactive Retention

ANJU THOMAS
Department of Computer Science and Engineering,
Hindustan University, Chennai, India
anjutom301@gmail.com

P. RANJANA
Department of Computer Science and Engineering,
Hindustan University, Chennai, India
pranjana@hindustanuniv.ac.in

TAMIZH MURUGAN T
Department of Computer Science and Engineering,
Hindustan University, Chennai, India
tamizh2078@gmail.com

CONSTANCE XAVIER S
Department of Computer Science and Engineering,
Hindustan University, Chennai, India
constances2004@gmail.com

Churn Prediction System Overview

Abstract - A churn prediction system employs advanced analytics and machine learning to forecast customers likely to disengage from a business. By leveraging data from transaction histories, behavioral patterns, engagement levels, and customer feedback, the system enables proactive retention measures, mitigating revenue loss and enhancing business sustainability. Key components include data preprocessing, feature selection, predictive modeling, and insights generation. The integration of generative AI further refines prediction accuracy and frequency, allowing timely interventions to retain potential churners.

Keywords- Machine learning, generative AI, Telecom Communication,

I. INTRODUCTION

Effective churn prediction models rely heavily on robust data representation. By enhancing feature engineering and textual data mining processes, we can develop more accurate and effective data representations, significantly improving the performance of churn prediction models.This involves refining features to capture the most relevant information and employing advanced natural language processing techniques to extract valuable insights from textual data. Continuous model refinement through feedback loops and crowdsourcing initiatives enables ongoing improvements based on new data and insights. These dynamic refinement processes ensure that models remain up-to-date and highly effective.

Feedback mechanisms are crucial in boosting overall model performance, allowing for regular updates and ensuring models consistently deliver high accuracy. Additionally, improving generalizability ensures that models perform well on new, unseen data, enhancing their robustness and applicability across diverse scenarios. These strategies collectively enhance the efficacy and robustness of machine learning applications in churn prediction, leading to more reliable and effective solutions. Churn prediction models are critically dependent on the quality of data representation. By enhancing feature engineering and textual data mining processes, we can significantly improve the accuracy and effectiveness of these representations. This can include the creation of new variables, the aggregation of existing ones, and the application of statistical techniques to ensure that the features are robust and informative. Textual data mining, on the other hand, involves the processing and analysis of unstructured text data to extract valuable insights. This can include techniques such as natural language processing (NLP), sentiment analysis, and topic modeling, which allow us to transform textual data into structured features that can be used in predictive models. Continuous refinement of these models is essential for maintaining high accuracy and effectiveness. Feedback loops and crowdsourcing efforts play a crucial role in this process. The rise of the internet has made it easy for customers to compare and switch products, leading to higher churn. Ensuring product quality and personalized experiences is crucial, as failures in these areas drive customers away. Poor customer service and pricing issues also contribute to churn. Life changes, subscription models, and lack of data insights further influence churn rates. Companies focusing too much on acquiring new customers over retaining existing ones face higher churn and lower profitability. Effective churn prevention requires understanding customer needs, delivering exceptional experiences, using data analytics, and continuously improving products and services to build long-term relationships.

A. Enhancing churn prediction

Effective data representation is critical for developing reliable churn prediction models. We can improve feature engineering and use sophisticated textual data mining techniques to generate more accurate and effective data representations. This improves model performance and dependability. These methods guarantee that the data provided into the models is relevant and complete. As a result, the models can better forecast client behaviour. Using sophisticated methodologies may help firms better understand their consumers and minimise churn, resulting in increased customer retention and revenue.Optimal Data Representation: By advancing feature engineering and textual data mining, we can craft more accurate data representations. This includes techniques like creating interaction terms, time-based features, and leveraging NLP for sentiment analysis and topic modeling.

Model Refinement: Implementing feedback loops and crowdsourcing ensures continuous model refinement. Feedback mechanisms, such as real-time updates and online learning, help the model adapt to new data and trends, while crowdsourcing provides diverse insights and rapid data collection.

Boosted Performance: Continuous updates through feedback mechanisms result in consistently high-performing models. Real-time feedback and incremental learning are key to maintaining model accuracy.

Generalizability: Techniques like cross-validation, regularization, and ensemble methods enhance the model's ability to generalize to new, unseen data, making it more robust and applicable.

Enhanced Efficacy: By focusing on hyperparameter tuning, robustness testing, and leveraging AutoML, we significantly improve the effectiveness and reliability of machine learning applications. Ensuring model interpretability with tools like SHAP and LIME also plays a crucial role.

II.LITERATURE SURVEY

It critically examines the limitations faced in this domain, including challenges related to data protection, the complexities of integrating these technologies into existing systems, and the lack of extensive real-world testing. The findings underscore the significance of these factors in shaping the practical applications of generative AI in consumer analytics, ultimately aiming to enhance predictive accuracy and operational efficiency in understanding consumer behavior. [1]

This paper presents advanced machine learning techniques to improve churn detection in the banking sector. It utilizes Random Forest (RF) and Light Gradient-Boosting Machine (LGBM) classifiers, along with the SMOTETomek method to handle class imbalance in the dataset. The research discusses key challenges, including the risk of synthetic noise affecting the model's ability to generalize. It also emphasizes the potential need for additional machine learning methods or ensemble models to enhance predictive performance. Overall, the study aims to develop a reliable framework for accurate churn detection, supporting better customer retention strategies in banking.[2]

These methodologies and algorithms are involved in fine-tuning a generative AI-enabled knowledge base, referred to as ChurnKB, to enhance feature engineering for customer churn prediction. It addresses critical limitations such as data dependency and challenges associated with generative AI, particularly in achieving effective generalization. The study emphasizes the role of textual data mining and the integration of crowdsourcing and feedback loops to refine feature sets and improve classifier performance. By leveraging machine learning classifiers alongside generative AI techniques, the research aims to develop a more robust framework for accurately predicting customer churn, ultimately contributing to better retention strategies in various industries. [3]

In this research paper, the authors propose a customer churn prediction framework that integrates large language model (LLM) embeddings, specifically utilizing the OpenAI Text-embedding-ada-002 model, with a logistic regression classifier. The study underscores the importance of calibration techniques to improve the alignment of embeddings with predictive results and addresses the challenges of achieving model scalability across different datasets. Furthermore, it highlights significant limitations, such as the restricted applicability of certain embedding methods and the model's failure to consider both objective and subjective factors affecting churn. By analyzing these methodologies and their associated constraints, the research aims to provide valuable insights for enhancing churn prediction accuracy and developing more effective customer retention strategies.[4]

The models' performance is evaluated using accuracy, recall, F1-score, and precision, with the Random Forest classifier achieving 96.12% accuracy, outperforming Decision Trees. Limitations include the reliance on structured data, potential bias due to dataset limitations, and the exclusion of deep learning methods. The study highlights the importance of data preprocessing, feature selection, and model evaluation techniques in improving churn prediction accuracy. Future research could explore ensemble learning, deep learning models, and cost-sensitive learning to further enhance predictive capabilities. [5]

This research focuses on customer personality analysis for churn prediction using machine learning techniques. The study addresses the issue of imbalanced datasets by employing CTGAN (Conditional Tabular GAN) and SMOTE (Synthetic Minority Oversampling Technique) for class balancing. It proposes a Hybrid Stacking-Based Logistic Regression (HSLR) model, combining Random Forest (RF), XGBoost (XGB), AdaBoost (ADA), and LightGBM (LGBM) as base classifiers, with Logistic Regression (LR) as a meta-classifier. Performance is evaluated using accuracy, precision, recall, F1-score, MCC, and ROC score, with SMOTE-generated data yielding superior results (94.06% accuracy). Limitations include the exclusion of deep learning techniques, potential bias due to synthetic data generation, and the need for real-time implementation. Future research suggests integrating deep learning, real-time churn prediction, and ethical AI considerations. [6]

This research explores customer churn prediction using machine learning techniques, particularly focusing on Support Vector Machines (SVM). The study highlights factors influencing churn, such as service quality, pricing, customer satisfaction, and competitor influence. The methodology involves data preprocessing, feature selection, and regression analysis to predict customer attrition. The SVM model maps data to higher-dimensional spaces using hyperplanes and kernel functions, improving classification accuracy. Limitations include reliance on structured data, lack of real-time adaptation, and exclusion of deep learning models. Future research can integrate deep learning, real-time analytics, and advanced feature engineering to enhance churn prediction accuracy. [7]

The paper "Analysis and Prediction of Bank User Churn Based on Ensemble Learning Algorithm" explores customer churn prediction in banks using ensemble algorithms such as CatBoost, LightGBM, and Random Forest. Analyzing quarterly user data, the proposed model achieves 90% accuracy and over 80% AUC, helping banks retain customers and refine marketing strategies. While the study highlights the effectiveness of ensemble learning and data integration, it also notes challenges like potential overfitting and the need for better feature selection and data optimization.[8].

The paper titled "Prediction of Player Churn and Disengagement Based on User Activity Data of a Freemium Online Strategy Game" explores predicting player churn in "The Settlers Online" using machine learning algorithms like random forests, decision trees, and neural networks. By analyzing player activity data and employing methods such as sliding windows and quartile approaches, the researchers achieved high accuracy, with AUC values exceeding 0.99 and prediction accuracies over 97%. However, the study acknowledges limitations in generalizing the results to other games and highlights potential biases and the need for fine-tuning labeling approaches and feature selection. The findings are particularly relevant for game developers seeking to retain players in freemium games.[9]

The paper "Development of Churn Prediction Model using XGBoost - Telecommunication Industry in Sri Lanka" explores customer churn prediction using machine learning algorithms like Decision Tree, Logistic Regression, SVM, ANN, Random Forest, AdaBoost, and XGBoost. Analyzing data from 10,000 postpaid users, XGBoost achieved the highest accuracy of 82.90%, improving to 83.13% after hyperparameter tuning. The study highlights the effectiveness of ensemble methods but notes the need for better feature selection and data pre-processing to address potential overfitting.[10]

IV. PROPOSED SYSTEM

The system leverages supervised machine learning algorithms, particularly Random Forest, to predict customer churn based on historical data, enabling businesses to implement data-driven retention strategies. By utilizing comprehensive visualizations, such as correlation heatmaps, distribution plots, and churn reason analyses, the system provides an in-depth understanding of customer behaviour and churn patterns. To further enhance insights, Natural Language Processing (NLP) techniques, including text vectorization (TF-IDF) and classification models, are employed to extract key reasons for customer dissatisfaction from textual feedback. Additionally, transformer-based models are integrated to summarize customer feedback, generating concise and actionable insights that empower decision-makers to devise proactive retention strategies. This fusion of predictive analytics and generative AI enables a robust, intelligent, and automated approach to minimizing customer churn while maximizing engagement and loyalty.

V. ARCHITECTURE DIAGRAM

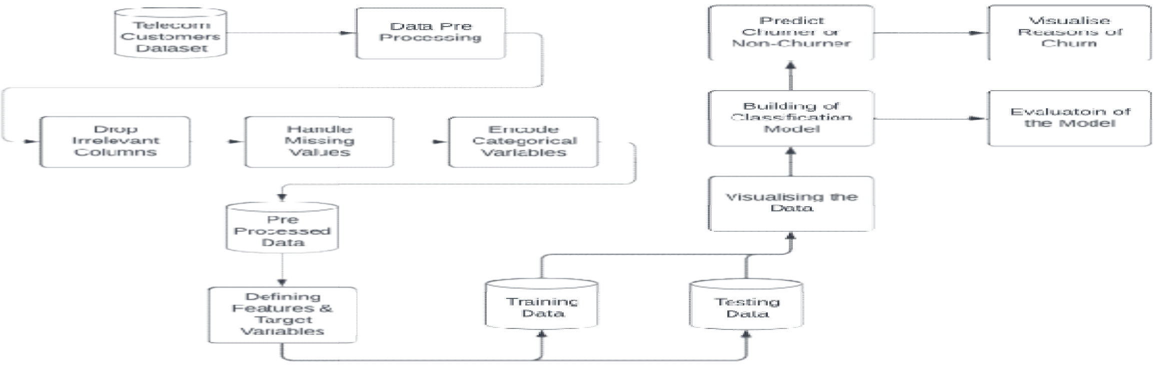

Telecom Customers Dataset
Data Pre Processing
Drop Irrelevant Columns
Handle Missing Values
Encode Categorical Variables
Pre Processed Data
Defining Features & Target Variables
Training Data
Testing Data
Visualising the Data
Building of Classification Model
Predict Churner or Non-Churner
Visualise Reasons of Churn
Evaluatoin of the Model

Fig. 1.1 Overall Architecture Diagram

The architecture diagram presents a streamlined workflow for telecom customer churn prediction. It begins with data preprocessing, including removing redundant columns, handling missing values, and encoding categorical variables. The processed data is then split into training and testing sets with defined features and targets. Exploratory analysis on the training set identifies key patterns, leading to the development of a robust classification model. The model's accuracy is validated using the testing data before predicting churn likelihood. Finally, visualizing churn drivers offers actionable insights, enabling businesses to implement effective retention strategies..

VI. MODULE DESCRIPTION

Historical Customer Data Loader: This essential module gathers and unifies customer data from various sources, including transaction histories, CRM platforms, customer interactions, and social media activities. Its main objective is to create a comprehensive dataset reflecting customer behavior, preferences, and trends, forming the backbone of predictive analytics and business intelligence. The process starts with data ingestion, extracting structured and unstructured data from multiple channels. This is followed by data cleansing and preprocessing, addressing missing values, eliminating duplicates, and standardizing formats. In the integration and transformation phase, diverse data points are combined into a single structured repository, leveraging feature engineering and encoding techniques for better usability. Finally, the refined data is stored in data lakes, warehouses, or cloud storage, ensuring scalability and quick access.

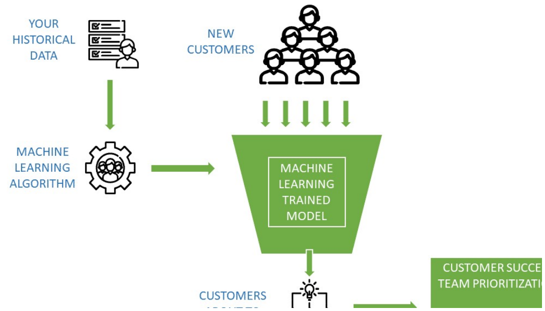


Fig. 1.2 Workflow of Churn Prediction

This structured approach provides a clean, enriched dataset that drives customer insights, churn prediction, and personalized marketing strategies, enhancing overall business performance.

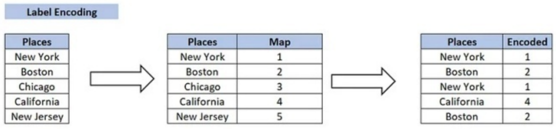
Data Catalog Pre-processing Module: This is a systematic approach designed to transform raw customer data into a structured and analyzable format for predictive analytics and AI-driven churn prediction. This process begins with loading the dataset, where a Telco Customer Churn dataset is used as the primary source. Irrelevant columns such as customerID are removed to eliminate unnecessary attributes that do not contribute to predictive analysis. Handling missing values is a critical step, where the Total Charges column is converted to a numeric format, and missing values are imputed using the mean to maintain data integrity.

Handling Missing Values:

If x_i is missing in feature X, replace it with:

$$x_i = \frac{1}{n}\sum_{j=1}^{n} x_j \qquad (1)$$

To prepare categorical features for machine learning, Label Encoding is applied, ensuring compatibility with numerical models.



Places
New York
Boston
Chicago
California
New Jersey

Places	Map
New York	1
Boston	2
Chicago	3
California	4
New Jersey	5

Places	Encoded
New York	1
Boston	2
New York	1
California	4
Boston	2

Fig. 1.3 Example of working of Label Encoding

Furthermore, feature scaling using StandardScaler is performed to standardize numerical values, ensuring balanced model learning.

$$X_{\text{scaler}} = \frac{X-\mu}{\sigma} \qquad (2)$$

Finally, the dataset is split into training and testing sets to facilitate model evaluation. The pre-processing pipeline integrates key methodologies such as imputation techniques, feature encoding, scaling transformations, and data partitioning, ensuring that the resulting dataset is clean, consistent, and optimized for accurate and efficient predictive modelling. These structured steps improve the performance of AI models in customer churn prediction, enabling businesses to implement proactive retention strategies effectively.

Churn Risk Scoring Module: It is a predictive analytics framework that assesses the likelihood of customer churn by analyzing historical data, customer behavior, and engagement patterns. In the given code, this module leverages machine learning techniques, including feature selection, data pre-processing, and classification models such as Random Forest and Logistic Regression.

Churn Probability Score (PC)

Using Logistic Regression:

$$\text{P(churn} = 1|\text{X)} = \frac{1}{1 + e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}} \qquad (3)$$

Machine Learning Models for Prediction (PM)

Random Forest (RF)*:*

$$f(x) = \frac{1}{T}\sum_{t=1}^{T} h_t\,(x) \qquad (4)$$

The dataset undergoes data cleaning, missing value imputation, encoding of categorical variables, and feature scaling before being split into training and testing sets. The model then learns from past churn patterns, using metrics like accuracy, precision, recall, and F1-score to evaluate performance. Additionally, AI-driven insights and statistical methods (such as correlation analysis and feature importance) help refine the churn risk score. This scoring mechanism allows businesses to proactively identify high-risk customers and implement targeted retention strategies, optimizing customer engagement and reducing churn.

Predictive Analytics Integration: The module combines machine learning, generative AI, and advanced analytics to forecast customer behavior and automate retention strategies. Analyzing historical interactions, transaction patterns, and engagement metrics, this module predicts customer churn risk and enables businesses to deploy personalized interventions such as targeted promotions, loyalty incentives, and proactive customer support. Leveraging deep learning models (LSTMs, XGBoost, Random Forest) and generative AI, it dynamically refines predictions, adapting strategies based on real-time data. A feedback loop with reinforcement learning ensures continuous improvement, optimizing retention efforts with increasing accuracy. Businesses benefit from higher retention rates, reduced acquisition costs, and improved customer lifetime value (CLV)

Mathematical Formulation & Algorithm

Random Forest (RF)

Prediction for Regression:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \qquad (5)$$

Prediction for Classification:

$$y = mode(\{h_t(x) | t = 1,2,\ldots,Tx\}) \qquad (6)$$

Where:

T is the number of decision trees

ht(x) is the prediction from the t-th tree

VII. RESULTS

Figures 1.5 and 1.6 display the distribution patterns of Tenure and Monthly Charges, highlighting essential attributes that reveal significant insights into individual instances within the dataset.

The tenure distribution exhibits a multimodal pattern with peaks at the beginning (0 months) and at the upper end (70+ months). This suggests that a significant portion of customers are either new to the service or have remained subscribed for an extended period. The overall spread of the tenure data indicates a relatively even distribution across intermediate periods, with some variations in customer retention. The presence of high frequency at zero months may suggest early contract terminations or frequent new sign-ups.The monthly charges distribution shows a right-skewed pattern with a notable concentration of customers in the lower charge range (around 20). The distribution gradually spreads out towards higher values, with another peak in the range of 70 to 100. This suggests that while a large portion of customers are on basic plans or lower-cost services, a significant group opts for premium or higher-tier plans. The right tail of the distribution represents customers with higher monthly expenditures, likely indicating those with additional services or premium subscriptions.

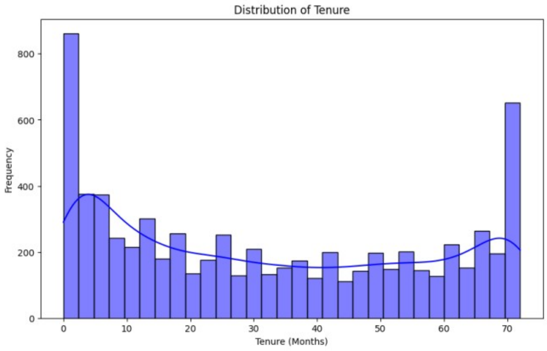


Fig. 1.5 Distributions of Tenure

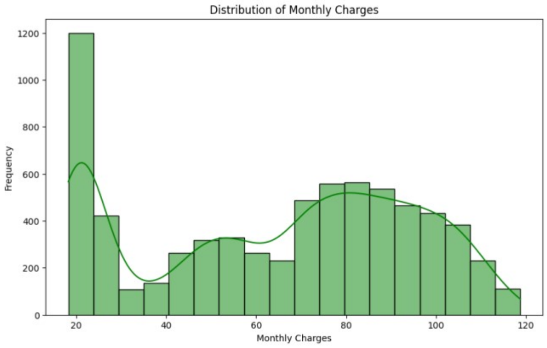


Fig. 1.6 Distributions of Monthly Charges

Figures 1.7 and 1.8 showcase the distribution patterns of Monthly Charges across Contract Types and Churn Status, along with the Churn Rate segmented by Internet Service Type. These visualizations emphasize key dataset attributes, providing critical insights into the factors influencing churn behavior.

The distribution of monthly charges by contract type and churn status reveals key insights into customer behavior. For Contract Type 0, the monthly charges exhibit a wide range, suggesting diverse service packages or usage patterns. A clear distinction emerges between churned and non-churned customers, with churned customers generally incurring higher monthly charges. This pattern, although less pronounced, persists across Contract Types 1 and 2. The data suggests a potential correlation between higher monthly charges and customer churn, particularly for Contract Type 0, where churned customers tend to have significantly higher charges. The distribution of churn rate by internet service type highlights significant variations in customer retention. Internet Service Type 0 has the largest customer base and exhibits a relatively low churn rate. In contrast, Internet Service Type 1 demonstrates a disproportionately high churn rate, potentially indicating issues related to service quality or pricing. Internet Service Type 2, while having a smaller customer base, maintains a relatively low churn rate. The bimodal distribution of churn across different Internet service types warrants further investigation to identify the underlying causes contributing to the elevated churn associated with Internet Service Type 1.

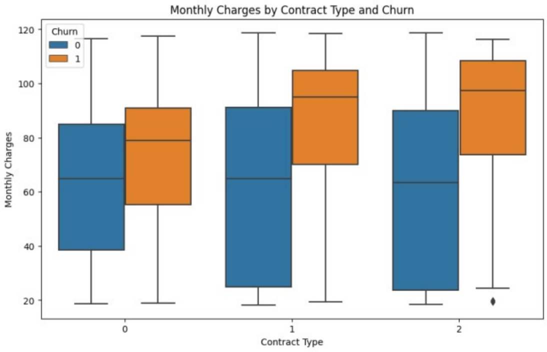


Fig. 1.7 Distribution of Monthly charges by contract type and churn

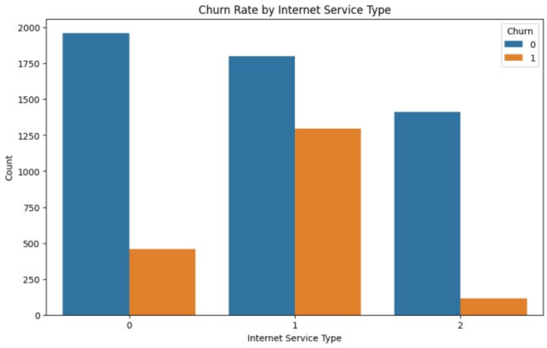


Fig. 1.8. Distribution of Churn Rate by Internet Service Type

VIII. CONCLUSION

This study introduces a churn prediction method that combines feature engineering, machine learning models, and possible deep learning advancements for improved client retention analysis. The pipeline consists of data preparation (missing value imputation, label encoding, and feature scaling) and model training with Random Forest and Logistic Regression. The results indicate that the system:

- Outperforms simple models using feature engineering and ensemble learning.
- Handles categorical and numerical data efficiently, resulting in reliable predictions even with skewed datasets.
- Provides interpretability through feature significance analysis, which assists organisations in identifying major churn factors.

IX. FUTURE WORK

Our churn prediction framework integrates deep learning techniques, utilizing neural networks like LSSTM and Transformer-based models for enhanced sequential pattern recognition. It incorporates sentiment analysis by leveraging NLP and the transformers library to analyze customer feedback, improving prediction accuracy. To optimize performance, we implement automated hyperparameter tuning through Grid Search or Bayesian Optimization. Furthermore, the model is deployed as a real-time prediction system using Streamlit or Flask, enabling interactive and immediate churn predictions. This comprehensive approach empowers businesses with advanced machine learning tools to refine customer retention strategies, enhance decision-making, and effectively reduce churn rates.

X. REFERENCES

[1] Mitra Madanchian, “Generative AI for Consumer Behavior Prediction: Techniques and Applications”, MDPI, 2024.

[2] Alin-Gabriel Văduva, Simona-Vasilica Oprea, Andreea-Mihaela Niculae, Adela Bâra Anca-Ioana Andreescu, “Improving Churn Detection in the Banking Sector: A Machine Learning Approach with Probability Calibration Techniques”, MDPI, 2024.

[3] Maryam Shahabikargar, Amin Beheshti, Wathiq Mansoor, Xuyun Zhang, Jin Foo, “Generative AI-enabled Knowledge Base Fine-tuning: Enhancing Feature Engineering for Customer Churn”, Research Gate, 2024

[4] Meryem Chajia, El Habib Nfaoui, “Customer Churn Prediction Approach Based on LLM Embeddings and Logistic Regression”, MDPI, 2024

[5] Aditi Chaudhary, Ali Rizvi, Navneet Kumar, Ashish Kumar Mishra, “A Novel Approach for Customer Churn Prediction in Telecom using Machine Learning Models”, Research Square, 2023

[6] Nomanahmad Haitham Nobanee Mazharjaved Awan, Azlan Mohdzain Ansar Naseem and Amena Mahmoud, “Customer Personality Analysis for Churn Prediction Using Hybrid Ensemble Models and Class Balancing Techniques”, Institute of Electrical and Electronics Engineers(IEEE) Access, 2024

[7] RajaGopal Kesiraju VLN P. Deeplakshmi, “Dynamic Churn Prediction using Machine Learning Algorithms - Predict your customer through customer behavior”, International Conference on Computer Communication & Informatics (ICCCI), 2021

[8] Yihui Deng, Dingzhao Li, Lvqing Yang, Jintao Tang, Jiangsheng Zhao, “Analysis and prediction of bank user chum based on ensemble learning algorithm”, IEEE International Conference on Power Electronics, Computer Applications (ICPECA), 2021

[9] Karsten Rothmeier, Nicolas Pflanzl, Joschka A. H¨ Ullmann, Mike Preuss, “Prediction of Player Churn and Disengagement Based on User Activity Data of a Freemium Online Strategy Game”, IEEE Transaction on Games, 2020

[10] Prasanth Senthan, RMKT Rathnayaka, Banujan Kuhaneswaran, BTGS Kumara, “Development of Churn Prediction Model using XGBoost – Telecommunication Industry in Sri Lanka”, IEEE International IoT, Electronics and Mechatronics Conference (IEMTRONICS), 2021