International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.066 ISSN: 2581-4621

Artificial Intelligence Tool For Churn Prediction Model and Customer Segmentation

A.Chandra Sekhar[1], J. Vignesh[2], C. Nabi Harshad[3], G. Mohammad Shaheed[4], D. Sreekanth[5], P. Vishnu Vardhan Reddy[6], A. Neela Vardhan[7]
chandra.ambarapu@gmail.com[1], jabadevignesh@gmail.com[2], nharshad.jhc@gmail.com[3], mohammadshaheed020@gmail.com[4], sreekanth.d2003@gmail.com[5], peramvishnuvardhanreddy551@gmail.com[6], amminenineelavardhan123@gmail.com[7] .
DEPARTMENT OF COMPUTER SCIENCE ENGINEERING (Artificial Intelligence) GATES INSTITUTE OF TECHNOLOGY, Gooty.

ABSTRACT
The telecommunications sector has experienced remarkable expansion over the past few decades, driven by rising competition, rapid technological advancements, and ever-changing customer expectations. To remain competitive in this evolving environment, telecom operators must prioritize customer retention and the delivery of personalized services. Two of the most pressing challenges in this context are churn prediction and customer segmentation. Churn prediction involves identifying subscribers who are at high risk of discontinuing their service and migrating to a competitor, allowing providers to develop timely retention strategies. Meanwhile, customer segmentation focuses on categorizing users into distinct groups based on shared traits and behavioral patterns, which enables more precise marketing campaigns, personalized service offerings, and optimized pricing models. Traditionally, telecom providers depended on conventional statistical approaches for churn prediction, which were often manual, limited in scalability, and insufficient for real-time decision-making. The emergence of Artificial Intelligence (AI) and machine learning has revolutionized this landscape, equipping telecom companies with advanced tools to enhance churn forecasting and refine customer segmentation. These technologies have empowered companies to leverage large-scale customer data for more accurate predictions and strategic segmentation, enabling proactive engagement with at-risk customers and fostering long-term loyalty. Consequently, this research focuses on the integration of AI-driven solutions for churn prediction and customer segmentation, which has become vital for telecom firms aiming to minimize customer attrition, strengthen competitive advantage, and deliver an exceptional customer experience in an industry shaped by constant innovation and fierce market rivalry.

INTRODUCTION
The development of an AI-driven tool for churn prediction and customer segmentation using unsupervised learning is primarily motivated by the growing emphasis on customer-centric strategies in modern business environments. In today's fiercely competitive markets, organizations across industries have come to recognize that customer retention and personalized engagement are fundamental to maintaining long-term profitability and sustainable growth. Customer churn the loss of existing clients represents a considerable risk, as replacing them often demands significantly more resources than retaining them. Addressing this challenge requires advanced AI-powered solutions capable of processing and analyzing extensive customer data to uncover hidden patterns and behavioral trends that conventional methods or human analysts might overlook. Such tools offer businesses actionable insights, allowing them to proactively mitigate churn by understanding its root causes and implementing targeted retention strategies. Furthermore, these systems facilitate precise customer segmentation, enabling companies to design customized marketing campaigns and product offerings that cater to the unique needs of distinct customer groups. The rapid advancements in artificial intelligence and machine learning have enhanced the scalability, speed,

DOI:10.46647/ijetms.2025.v09i02.066 ISSN: 2581-4621

and accuracy of data analysis, making the adoption of these technologies both practical and necessary for businesses striving to stay competitive. Ultimately, the motivation behind developing AI-based churn prediction and segmentation tools stems from the need to leverage data-driven intelligence to strengthen customer relationships, minimize attrition, and drive sustained business growth in an era where success is increasingly defined by a company's ability to deliver personalized, data-informed customer experiences.

LITERATURE SURVEY

Customer churn the rate at which customers discontinue their association with a company remains one of the biggest challenges in competitive industries, especially for telecom providers. Identifying customers likely to leave and understanding their reasons for doing so is vital for any organization aiming to maintain long-term profitability. Thanks to recent advancements in data analytics and Customer Relationship Management (CRM) systems, businesses have shifted toward customer-focused strategies over traditional product-centric approaches. This shift has opened the door to new, data-driven marketing opportunities, where minimizing churn and retaining loyal customers have become key elements in maximizing shareholder value and sustaining growth.

In the telecom sector, where customer loyalty directly impacts survival, the loss of subscribers is often unpredictable and costly. As acquiring new customers is significantly more expensive than retaining existing ones, telecom companies prioritize churn prevention. Studies indicate that the cost of retaining a customer is five to ten times lower than the cost of acquisition, and a mere 5% reduction in churn can potentially increase profits by up to 85%. This highlights the urgent need for accurate churn prediction systems as part of strategic business planning. Machine learning and data mining technologies have become essential tools in this space, empowering companies to detect patterns of customer attrition and respond proactively.

Beyond churn prediction, AI-based solutions offer a clear advantage in crafting personalized marketing campaigns and strengthening customer relationships. With the growth of social media marketing and the widespread availability of customer data, businesses now rely on AI tools to automate insights and enable sales and support teams to deliver more effective and personalized service. Deep learning techniques, combined with customer behavior analysis, can also help businesses predict what resonates with their audience, optimize marketing efforts, and enhance conversion rates. Furthermore, AI-driven systems can reduce manual workload, streamline processes, and uncover opportunities for smarter customer engagement across platforms.

While prior research has addressed churn prediction across multiple sectors including employee churn and subscription-based services the telecom industry remains a primary focus, given its high financial losses linked to customer departures. Different studies have approached this problem from two angles: maximizing predictive accuracy and maximizing profitability. However, few have attempted to balance both objectives, which is essential for real-world application developing churn models that are not only accurate but also cost-effective and scalable without the need for complex or resource-intensive methods.

To tackle this challenge, the present study investigates various machine learning strategies including traditional classifiers, ensemble techniques, and deep learning methods. These models were tested using two publicly available datasets: one from the Indian and Southeast Asian telecom sector, and another from the American market. The goal is to develop churn prediction systems that balance accuracy and business value. Researchers like Shirazi and Mohammadi have explored how blending structured data with unstructured sources, such as web traffic and call logs, can improve churn detection. Other work, such as Zdravevski's, has shown the benefits of using cloud-based ETL pipelines to handle large datasets and detect over 98% of churn cases, providing actionable insights for sales and retention teams.

Unstructured data such as phone conversations have also proven valuable in churn forecasting. Colleagues demonstrated how machine learning models trained on call log data can effectively

DOI:10.46647/ijetms.2025.v09i02.066 ISSN: 2581-4621

predict churn and offer business intelligence by analyzing both customer sentiment and behavioral trends. Similarly, profit-aware churn prediction models have emerged, combining accuracy with business value. One example is the Expected Maximum Profit measure (EMPC), which integrates financial priorities into predictive models. Studies have shown that embedding EMPC into decision trees and logistic models can lead to higher profits and better customer retention outcomes, as highlighted in works by Stripling and others, where evolutionary algorithms were combined with churn prediction to optimize both accuracy and return on investment.

RELATED WORK

In the field of customer churn prediction and segmentation, various machine learning techniques have been explored to enhance predictive accuracy and business insights. One widely adopted approach is **Gradient Boosted Trees (GBT)**, an ensemble-based learning technique designed to address both classification and regression tasks by combining the predictions of multiple decision trees. The model builds trees in a sequential manner, with each new tree trained to minimize the errors left by its predecessors, ultimately refining prediction accuracy through iterative learning. This step-by-step correction process allows GBT to capture complex data relationships and deliver reliable outcomes, making it an important reference point in data-driven customer retention strategies.

Another influential method in the literature is the **XGBoost Regressor**, an advanced extension of gradient boosting that introduces regularization and optimized learning rates to further strengthen model robustness and accuracy. The XGBoost framework starts with a basic decision tree and progressively builds new trees to reduce the residual errors identified in previous iterations. By implementing shrinkage techniques and applying penalty terms to control model complexity, XGBoost reduces the likelihood of overfitting while improving the generalization of predictions. This algorithm has become especially popular for handling large-scale prediction problems, including customer churn analysis, due to its efficiency and high performance on structured datasets.

Despite the strengths of these models, several challenges have been identified in the literature. Both GBT and XGBoost require precise **hyperparameter tuning** to achieve optimal performance, which can be resource-intensive and time-consuming. These models are also sensitive to **outliers** and demand significant computational resources, which limits their deployment in real-time and low-power systems. Moreover, the **interpretability** of ensemble models is often reduced as the complexity of the tree ensemble grows, making it difficult to trace specific decision paths. Additionally, XGBoost lacks **native support for categorical variables** and requires manual preprocessing, and its ability to handle **time-dependent data** is limited compared to specialized temporal models like recurrent neural networks (RNNs). These limitations highlight the need for alternative AI-driven approaches that can overcome the drawbacks of existing techniques and deliver more scalable, interpretable, and adaptable solutions for churn prediction and customer segmentation.

PROPOSED SYSTEM

The proposed system is a comprehensive, data-driven framework designed to help telecom operators proactively identify customers at risk of leaving and to uncover meaningful customer segments for targeted retention efforts. We begin by ingesting raw subscriber data into a distributed Spark environment, where a series of preprocessing steps such as cleaning missing entries, encoding categorical fields, and normalizing numerical features ensures that the dataset is analysis-ready. With the data prepared, we explore key variables and their relationships to churn through visual analyses, which inform both feature engineering and model selection.

For the churn-prediction component, we implement two robust ensemble classifiers Random Forest and Gradient-Boosted Trees tuning each via cross-validated grid searches and assessing their ability to discriminate between churners and non-churners using ROC-AUC metrics. Recognizing

International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.066 ISSN: 2581-4621

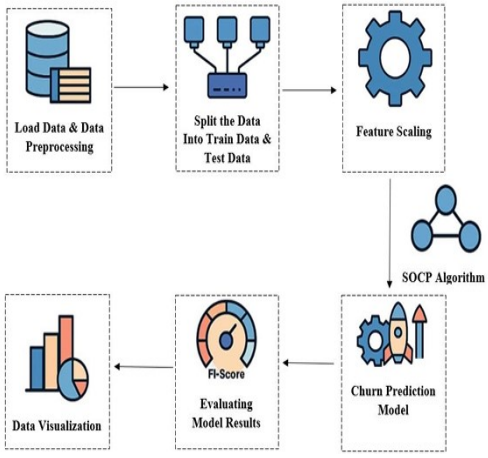
that revenue at risk is as critical as churn likelihood, we extend our pipeline to include regression models: an XGBoost regressor fine-tuned for optimal error rates, alongside a small feed-forward neural network built in Keras. These models quantify potential revenue loss, enabling the business to prioritize high-value accounts for retention campaigns.

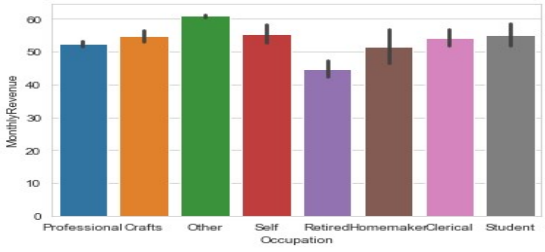
Beyond prediction, our system segments the customer base using K-Means clustering. By evaluating cluster quality with silhouette scores and profiling each segment's characteristic behaviors such as usage patterns, billing metrics, and tenure we generate actionable groupings that marketing and customer-success teams can use to tailor offers, communications, and incentives. In this way, the proposed system not only flags at-risk individuals but also uncovers distinct customer archetypes, empowering telecom providers to deploy personalized, cost-effective strategies that drive loyalty and reduce churn.

ADVANTAGES OF PROPOSED SYSTEM

- Utilizes Apache Spark to process large-scale telecom data in a distributed, scalable manner
- Applies K-Means clustering, which remains efficient even in high-dimensional feature spaces
- Employs Random Forest and Gradient-Boosted Trees with hyperparameter tuning to achieve high predictive accuracy
- Conducts extensive data cleaning, missing-value handling, and categorical encoding to ensure data quality and speed up computation
- Uses Seaborn and Matplotlib visualizations to uncover key drivers of churn and guide retention strategies
- Identifies high-risk customers early, enabling targeted campaigns that reduce revenue loss
- Focuses retention efforts on the most profitable, high-propensity customers to minimize overall costs
- Produces clear, reproducible cluster profiles that aid stakeholder understanding and decision-making
- Supports both classification (churn prediction) and regression (revenue-loss forecasting) within a unified pipeline
- Integrates multiple machine-learning libraries (scikit-learn, XGBoost, Keras) to tackle diverse analytical challenges

ARCHITECTURE

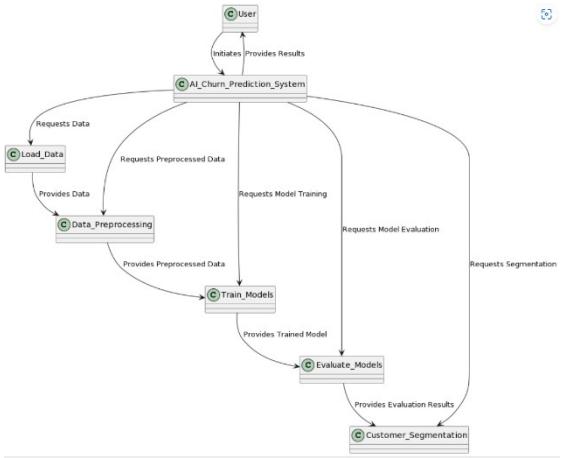




International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.066 ISSN: 2581-4621

DATA FLOW DIAGRAM

- User initiates the AI Churn Prediction System and receives the final results.



- The system requests and loads raw data via the Load Data component.
- Loaded data is sent to Data Preprocessing for cleaning, encoding, and transformation.
- Pre-processed data is passed to Train Models to build and return trained predictive models.
- The system invokes Evaluate Models to assess model performance and generate evaluation metrics.
- Customer Segmentation applies K- Means clustering on pre-processed data to produce segment labels.
- The AI Churn Prediction System aggregates model outputs, evaluation results, and customer segments for the user.

RESULTS

Figure 1 displays a bar plot showing how monthly revenue varies across different occupation categories. It provides insights into the revenue distribution among different professional groups.

Figure 2 presents a categorical distribution of the "Months Inservice" variable, categorized into 12 groups. It illustrates how the duration of service is distributed among the customers.

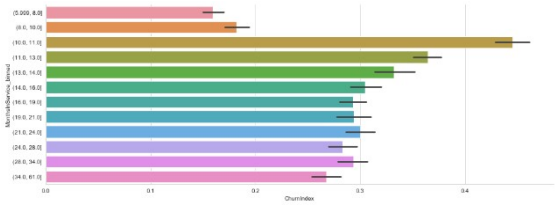


Figure 3 shows a series of bar plots representing the relationship between categorical variables (HandsetRefurbished, HandsetWebCapable,and MadeCallToRetentionTeam) and the ChurnIndex. It helps understand how these variables influence the likelihood of customer churn.

International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.066 ISSN: 2581-4621

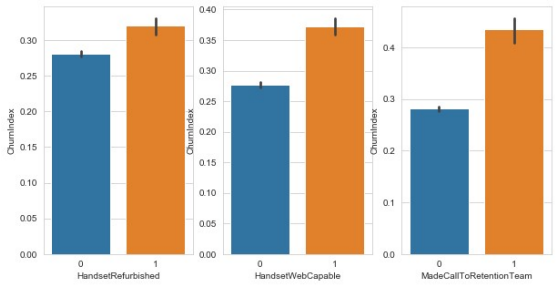
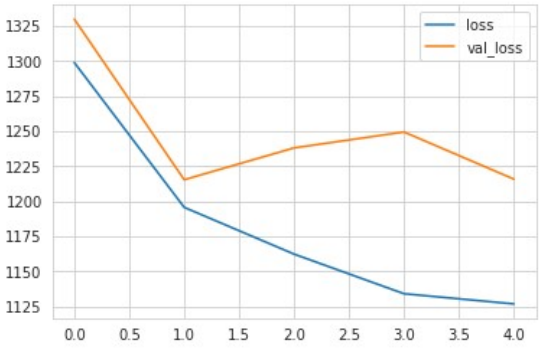
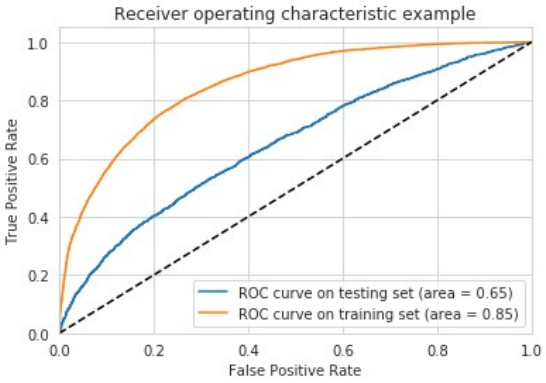


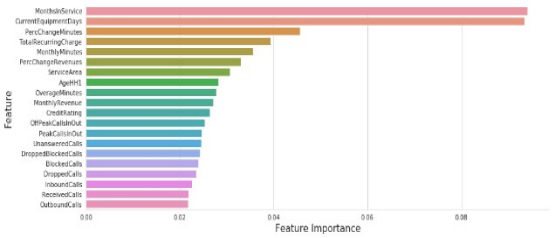
Figure 4 displays the Receiver Operating Characteristic (ROC) curve for a Random Forest Classifier. It provides a visual representation of the model's performance in terms of true positive

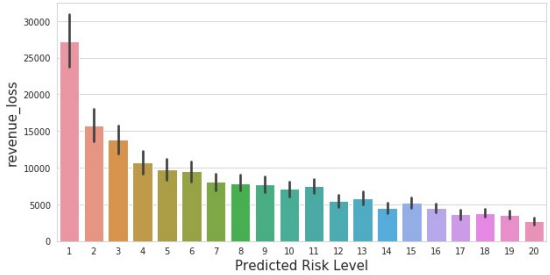




rate against false positive rate.

Figure 5 shows a bar plot or similar visualization indicating the percentage importance of different features in the dataset. It helps identify which features contribute the most to the predictive power of the model.





International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.066 ISSN: 2581-4621

Figure 6 A textual summary of the architecture and parameters of a Feedforward Neural Network model. It provides a concise overview of the model's structure and settings.

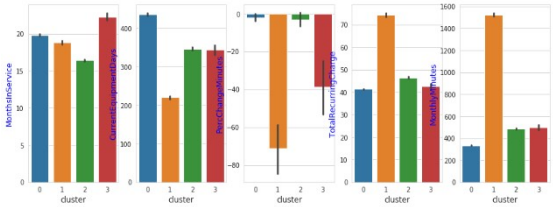
Figure 7 illustrates the training and validation loss over epochs during the model training process. It helps assess how well the model is learning from the training data and if it is overfitting or underfitting.

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 67)	3819
dropout_1 (Dropout)	(None, 67)	0
dense_2 (Dense)	(None, 56)	3808
dropout_2 (Dropout)	(None, 56)	0
dense_3 (Dense)	(None, 44)	2508
dense_4 (Dense)	(None, 1)	45

Total params: 10,180
Trainable params: 10,180
Non-trainable params: 0

Figure 8 displays a bar plot or similar visualization indicating the percentage of clients classified in the highest risk group. It provides insights into the proportion of high-risk clients in the dataset.

Figure 9 contains multiple subplots, each showing the distribution of a specific feature across different clusters. It helps understand how different features vary among the identified clusters.



CONCLUSION

This research has effectively addressed the complex challenge of predicting customer churn and estimating potential revenue loss within the telecom sector. By implementing powerful machine learning models such as Random Forest and Gradient Boosted Trees for churn prediction, the system was carefully refined through hyperparameter tuning and rigorous cross-validation to ensure strong and reliable performance. The accuracy and trustworthiness of these models were further validated using ROC curve analysis. Alongside this, advanced models like XGBoost Regressor and Keras -based Neural Networks were employed to predict potential revenue losses. Through detailed parameter tuning with grid search and performance evaluation using the Mean Absolute Error (MAE), the models were optimized for precision. To enhance understanding, visualizations were created to highlight how different customer segments contribute to revenue loss, offering meaningful insights for more informed decision-making.

FUTURE WORKS AND EXTENSIONS

Looking ahead, this project offers a wide range of possibilities for growth and refinement. One promising direction is the integration of more advanced machine learning techniques, including ensemble models and deep learning frameworks, to boost the accuracy of both churn and revenue loss predictions. Continuous exploration and innovation in feature engineering could also uncover

DOI:10.46647/ijetms.2025.v09i02.066 ISSN: 2581-4621

new patterns or enhance existing features, further improving model performance. Tackling class imbalance, which is a common hurdle in churn prediction, could be addressed through methods like oversampling or advanced resampling techniques such as SMOTE.

Beyond prediction, future work could focus on transforming these insights into real-world customer retention strategies, such as offering personalized deals or services to customers identified as high-risk. Implementing real-time prediction systems with tools like Apache Kafka or Apache Storm could enable businesses to respond instantly to potential churn events. Additionally, developing interactive dashboards using platforms like Tableau or Power BI would give business teams the ability to visualize insights and model outputs dynamically, supporting smarter and quicker decision-making.

To maintain the relevance and accuracy of the models over time, it will be essential to set up ongoing monitoring systems. If shifts in customer behavior or data trends are detected, the models can be retrained or adjusted accordingly. Collaboration with domain experts and business teams will also play a key role in ensuring that the solutions stay aligned with real-world business goals. By pursuing these future enhancements, the project has the potential to grow into a powerful, intelligent tool that not only predicts outcomes but also guides telecom providers in making impactful business decisions.

REFERENCES

[1]. Mishra, A.; Reddy, U.S. A comparative study of customer churn prediction in telecom industry using ensemble-based classifiers. In Proceedings of the 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, India, 23–24 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 721–725.

[2]. Shirazi, F.; Mohammadi, M. A big data analytics model for customer churn prediction in the retiree segment. Int. J. Inf. Manag. 2019, 48, 238–253.

[3]. Bhattacharyya, J.; Dash, M.K. Investigation of customer churn insights and intelligence from social media: A netnographic research. Online Inf. Rev. 2020, 45, 174–206.

[4]. Ahmad, A.K.; Jafar, A.; Aljoumaa, K. Customer churn prediction in telecom using machine learning in big data platform. J. Big Data 2019, 6, 28. [Green Version]

[5]. Coussement, K.; Lessmann, S.; Verstraeten, G. A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decis. Support Syst. 2017, 95, 27–36.

[6]. Ly, T.V.; Son, D.V.T. Churn prediction in telecommunication industry using kernel Support Vector Machines. PLoS ONE 2022, 17, e0267935.

[7]. De Caigny, A.; Coussement, K.; De Bock, K.W. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. Eur. J. Oper. Res. 2018, 269, 760–772.

[8]. El-Gohary, H.; Trueman, M.; Fukukawa, K. Understanding the factors affecting the adoption of E-Marketing by small business enterprises. In E-Commerce Adoption and Small Business in the Global Marketplace; Thomas, B., Simmons, G., Eds.; IGI Global: Hershey, PA, USA, 2009; pp. 237–258.

[9]. El-Gohary, H. E-Marketing: Towards a conceptualization of a new marketing philosophy e book chapter. In E-Business Issues, Challenges and Opportunities for SMEs: Driving Competitiveness; IGI Global: Hershey, PA, USA, 2010.

[10]. Jain, N.; Tomar, A.; Jana, P.K. A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning. J. Intell. Inf. Syst. 2021, 56, 279–302.

[11]. Indian and Southeast Asian Telecom Industry Dataset Which Is. Available online: https://www.kaggle.com/datasets/priyankanavgire/telecom-churn (accessed on 22 March 2021).

[12]. American Telecom Market Dataset. Available online: https://www.kaggle.com/datasets/mnassrib/telecom-churn-datasets (accessed on 18 February 2020).

DOI:10.46647/ijetms.2025.v09i02.066 ISSN: 2581-4621

[13]. Zdravevski, E.; Lameski, P.; Apanowicz, C.; Ślęzak, D. From Big Data to business analytics: The case study of churn prediction. Appl. Soft Comput. 2020, 90, 106164.
[14]. Vo, N.N.; Liu, S.; Li, X.; Xu, G. Leveraging unstructured call log data for customer churn prediction. Knowl.-Based Syst. 2021, 212, 106586.