Enhanced Over the Top Subscribers Retention Prediction using Random Forest with Hyperparameter Tuning

V Karthick, Associate Professor
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai - 602105.
vkarthick86@gmail.com

Suryaa KS
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai - 602105.
210701273@rajalakshmi.edu.in

Syed Javith R
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai - 602105.
210701278@rajalakshmi.edu.in

ABSTRACT

With rapid development of Over-the-top(OTT) platforms in the entertainment industry, predicting the user's churn behavior has become a crucial one which helps in gaining the profit and analyzing their trends. The trends in the user interaction with the product will help the company to gain insights of their areas of improvement for future purpose. Thus playing a crucial role in business and customer management. We suggest a Machine Learning based churn prediction model which has been trained specifically on the OTT platform data like user preferences, demographic information and user viewing habits. With the use of historical data on the user subscription and interaction pattern, our Machine Learning model tries to incorporate the features that may determine the churn of the user. Retaining the users has been a significant problem for the companies.The dataset utilized for this model takes many variables into account. In this proposed methodology, After evaluating a variety of machine learning algorithms for efficiency and accuracy, the final accuracy of 85% was achieved using Random Forest algorithm with hyperparameter tuning.

Keywords : Over-the-top(OTT),Subscribers,Classification, Tuning,Interaction,Business.

INTRODUCTION

- In the ever evolving landscape of Entertainment industry, Over-the-top(OTT) platforms have witnessed a great importance in popularity and significance. Despite this, it also has seen some challenges in understanding the behavioral patterns of the subscribers,who are the primary consumers of these platforms. Thus gaining the insights and developing the application according to those insights will be helpful in driving profit out of the consumer behavior.
- The companies can strategically plan their areas of improvement ensuring better management of customers. Moreover the companies can also manage their businesses in an efficient manner to tackle the upcoming challenges on the way. The usage of Over-The-Top (OTT) platforms has influenced the way people consume entertainment, offering convenience and adaptability. However, in this fast paced growing platform, retaining subscribers and reducing retention rates are inevitable for the consistent success of OTT platforms.
- This project aims to develop a machine learning solution for subscriber retention prediction customized specifically for OTT platforms. By exploring a wide range of user data, including preferences, demographic information, and viewing habits, the project predicts whether a customer is likely to retain their subscription.
- In addition to that, targeted notifications are sent to at-risk customers, offering personalized incentives or recommendations to retain their subscription. This application improves user retention rates, improves customer satisfaction, and enhances sustainable growth for OTT platforms in the ever growing digital entertainment landscape.

LITERATURE REVIEWS

PAPER	AUTHORS	DATE PUBLISHED	REVIEW
Archimedes Optimization Algorithm-Based Feature Selection with Hybrid Deep-Learning-Based Churn Prediction in Telecom Industries	H. A. Mengash, N.Al Ruwais, F. Kouki, C. Singla, E. S. Abdel Hameed, and A. Mahmud	Dec, 2023	Uses deep-learning based churn prediction model that utilizes Archimedes Optimization Algorithm for feature selection.They used objective function to compute the effectiveness of the feature subsets by archimedes principle
Research on customer churn prediction and model interpretability analysis	K. Peng, Y. Peng, and W. Li	Dec , 2023	Uses interpretability analysis for predicting the churn of the customer.They incorporated various sampling methods including SMOTEEN,SMOTE.They used XGB model for better optimizing the F1 and AUC values to more than 90%.
Artificial Intelligence Based Customer Churn Prediction Model for Business Markets	J. Faritha Banu, S. Neelakandan, B. T. Geetha, V. Selvalakshmi, A. Umadevi, and E. O. Martinson	Sep , 2022	Used the CCP model to distinguish the churners and non churners.They used FRC and QPSO to optimize feature assortment and validation of the datasets.
Bayesian Networks for Churn Prediction in the Mobile Telecommunications Industry	B. Brandusoiu	2020	The research on Bayesian Network[7]for retention prediction in telecommunication speaks on the volume of influence on Bayesian networks in preventing the customer to stay within their subscriptions.

PAPER	AUTHORS	DATE PUBLISHED	REVIEW
Hyperparameter Tuning with Python: Boost your machine learning model's performance via hyperparameter tuning	L. Owen	2022	Explored the ways to control the key features of the model resulting in boosting the performance of the ML model.It explores HS and optimization methods to boost the accuracy of a model.
Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient	T. Agrawal	2020	Explores the step by step guide for hyperparameter optimization.It addresses the brute force approaches including the problem of time and memory constraints.
Early warning of telecom enterprise customer churn based on ensemble learning,	Y. Zhou, W. Chen, X. Sun, and D. Yang	Oct , 2023	Used ensemble learning to predict the exit of the customer from the subscriptions.They have used BPNN and RF as the ensemble models to analyze customer data.They provided insights which helps telecom companies to improve their retention strategies for the customers.
An efficient churn prediction model using gradient boosting machine and metaheuristic optimization	I. AlShourbaji, N. Helian, Y. Sun, A. G. Hussien, L. Abualigah, and B. Elnaim	Sep , 2023	Used GB and meta optimization techniques to predict the exit of the customer.They use a modified PSO method along with Artificial ecosystem optimization to perform hypertuning
New CRM System for Telecoms Customer Churn Analysis Based on Ensemble Learning and RFM	T.Xu	2023	Telecom customer exit analysis used ensemble learning and RF to predict the exit of the customer.Staking models such as XGB, LR, DT are used for soft voting.

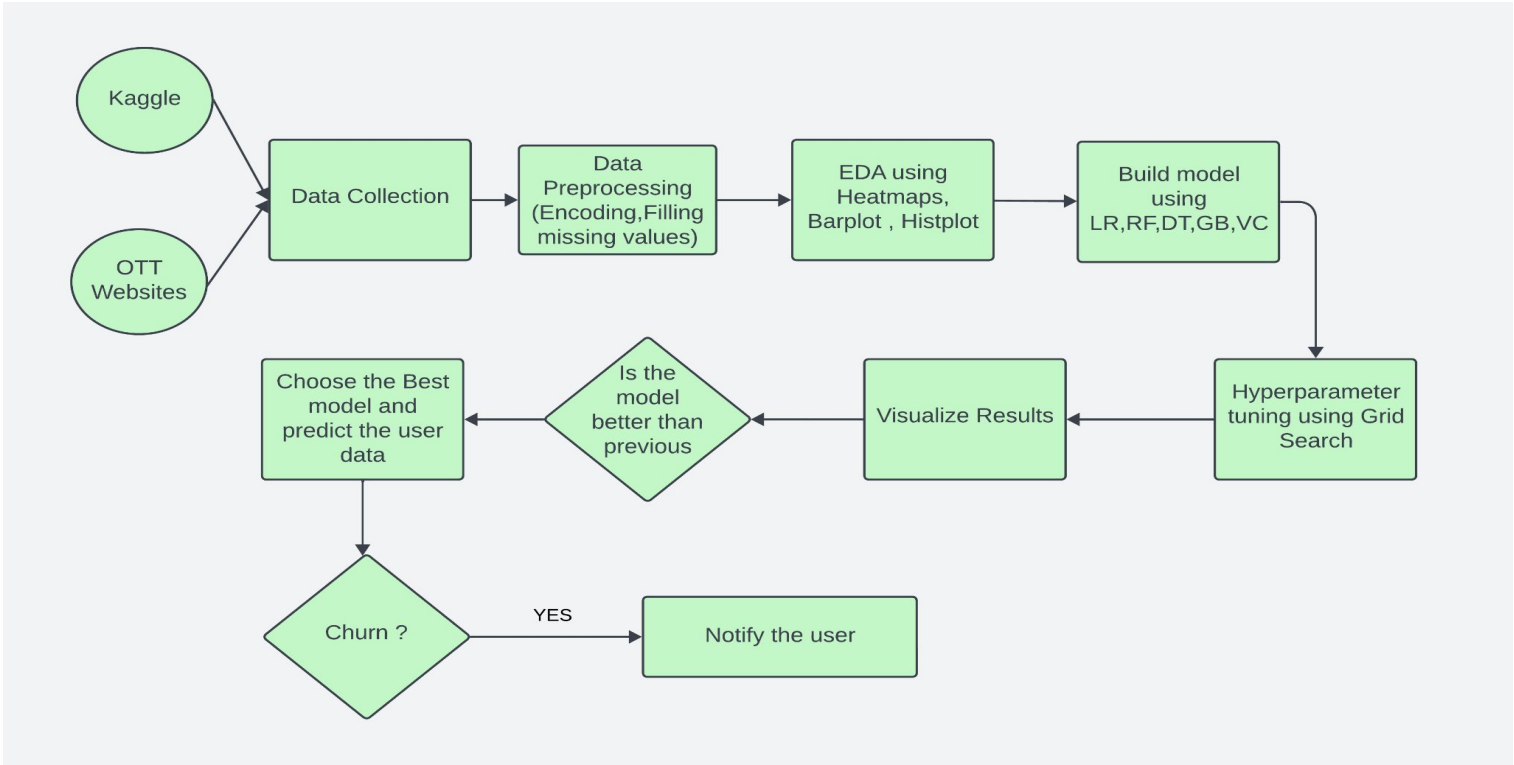
RESEARCH GAP

While existing machine learning-based churn prediction systems for OTT platforms exist, they may have limitations. These limitations could include

- Using a narrow range of user data (excluding user preferences)
- Being static and not adapting to change
- Failing to provide actionable strategies for user retention.

Our proposed system, which incorporates continuous learning and explainability, has the potential to address these shortcomings and offer a more comprehensive approach to churn prediction and user retention in the competitive OTT market.

PROPOSED METHODOLOGY



PROPOSED METHODOLOGY

A. Data Preprocessing : The Data collected won't be in a state that can be used for training purposes hence, the data should undergo the step of preprocessing in which common problems are eradicated such as missing values , improper spelling in data or incorrectness in data etc.

B. EDA : The acquired data is analyzed for its relation within the data. Any outliers or deviation of data can be inferred at this point and also this helps to gain the significance of each data column.Heatmaps are extensively used to know the correlation between various attributes.

C. Build Models : The next step involves building various machine learning models in supervised classification algorithms such as LR , DT , Random Forest , KNN and also several ensemble algorithms such as A-Boosting ,G-Boosting , V-Classifier are used to build and keep track of the models' performance.

D. Hyperparameter tuning : Once the basic models are built, the models are then tuned based on their Hyperparameters such as max_depth, iteration count to improve the performance of the existing models. The tuning of the models will help it to find the best parameters for training.

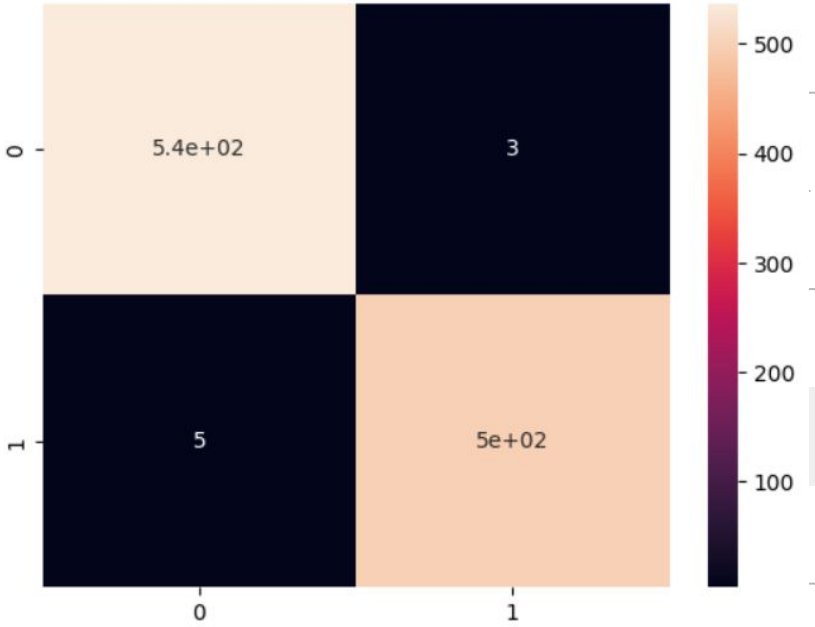
E. Visualizing Results : The results of various parameters and also the accuracies along the time are analyzed in this step to get insights of the working of various models. The Confusion matrix plays an important role in analyzing the performance of models. With the help of an intuitive bar graph, we can infer the nobel model for the project.

F. Choosing Best Model : As we have a track of the model performance we can choose the best model among the trained models and can utilize it for the further development of the project which leads to an iterative development process.

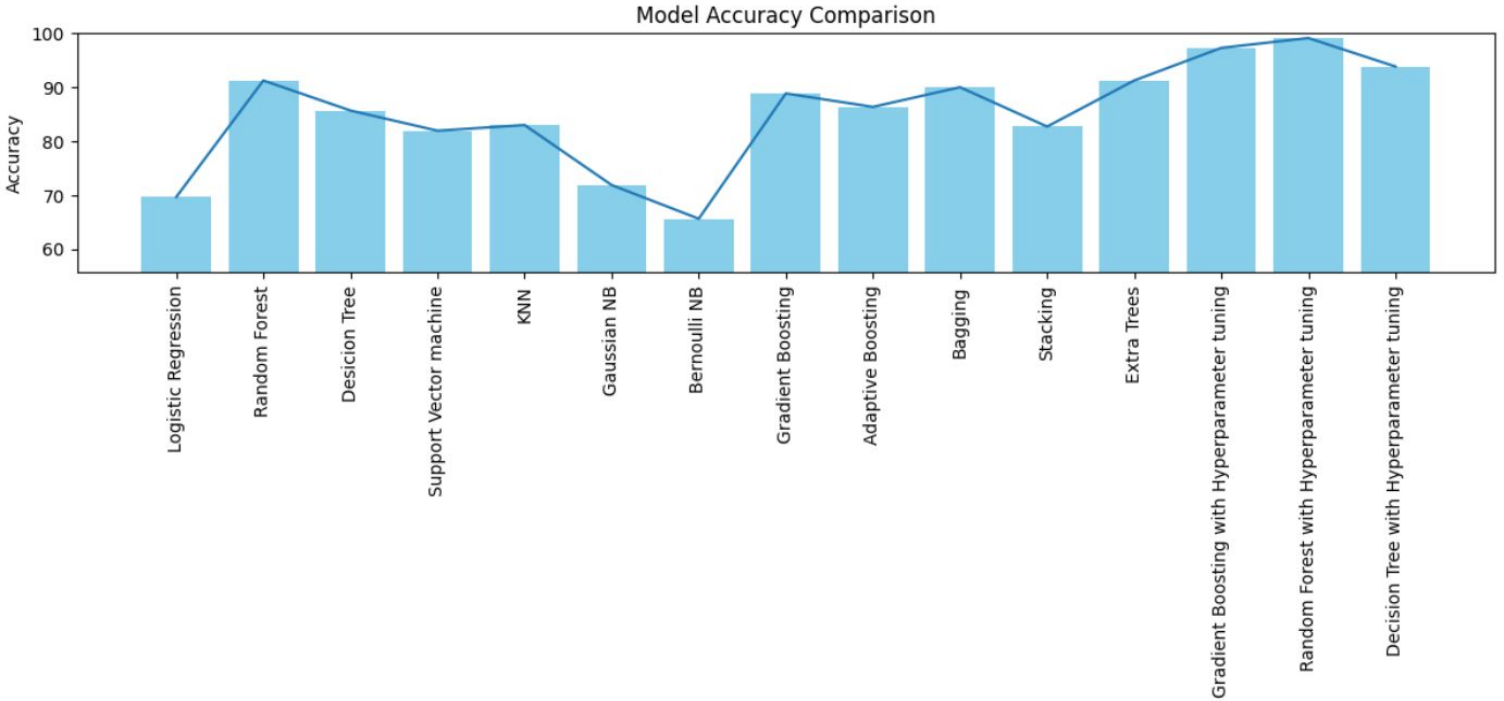
RESULT AND DISCUSSION

The classification report for our customer churn prediction model provides a detailed analysis of the model's performance. Key metrics include precision, recall, and F1-score for both the 'retain' and 'churn' classes.

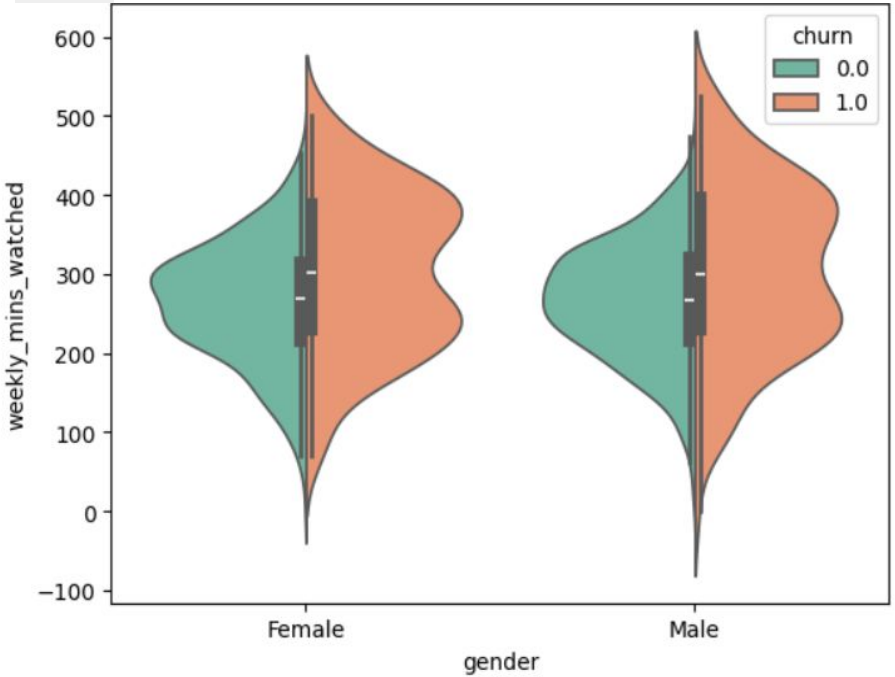
Precision indicates the accuracy of the positive predictions, recall measures the model's ability to identify all actual positives, and the F1-score provides a balance between precision and recall. High values in these metrics reflect the model's effectiveness in correctly predicting customer behavior



RESULT AND DISCUSSION



RESULT AND DISCUSSION



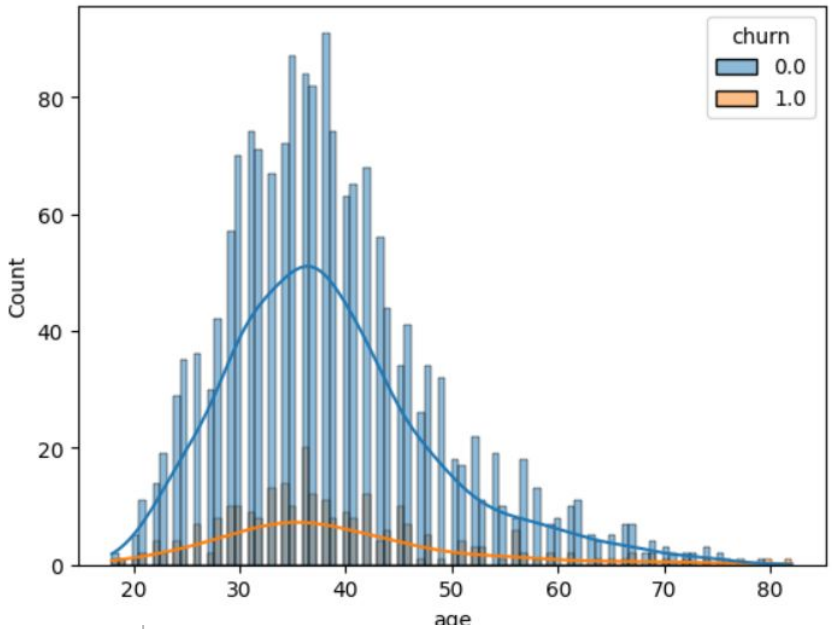
The violin graph illustrates the distribution of age categories and gender among users who are predicted to churn from their subscription plan. By visualizing this data,it was able to identify specific age groups and gender profiles that are more likely not to retain their subscription.

The wider sections of the violin plots indicate a higher density of users within those age and gender segments. This insight allows for targeted retention strategies aimed at these high-risk groups, enhancing the effectiveness of interventions designed to reduce churn and improve customer retention.

RESULT AND DISCUSSION

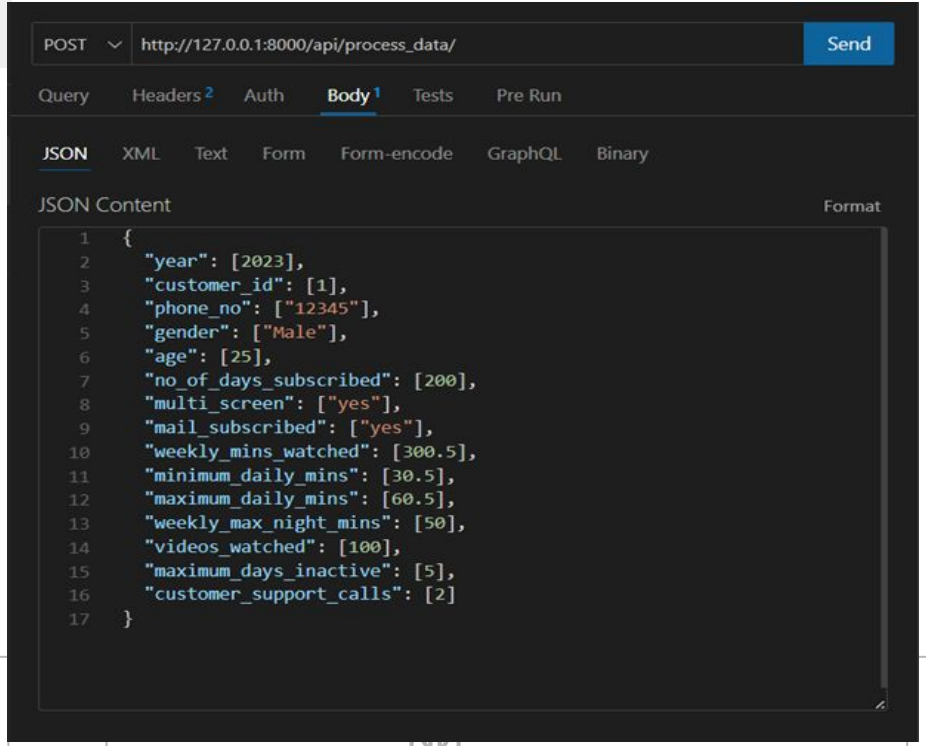
The histogram displays the age distribution of users who are predicted to churn from their subscription plan. Each bar represents the frequency of users within specific age ranges, highlighting which age groups are more prone to not retaining their subscription.

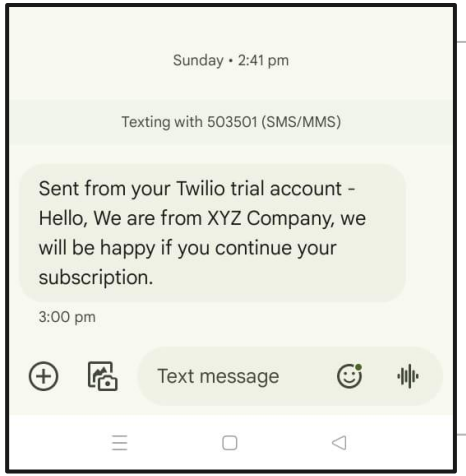
This visual analysis helps identify critical age segments that require focused retention efforts.By understanding the age demographics most associated with churn, businesses can tailor their engagement strategies to address the needs and preferences of these high-risk age groups



RESULT AND DISCUSSION

The API for sending SMS notifications to churned customers was tested using Postman. These tests ensured the accuracy and reliability of the message delivery system, confirming that notifications are sent promptly to at-risk users.





COMPARATIVE ANALYSIS

MODEL	ACCURACY	F1-SCORE
Logistic Regression	76.17	0.82
BernoulliNB	76.5	0.83
Adaptive Boosting	77.33	0.84
Extra Trees	79	0.84
Bagging	80.17	0.84
Gradient Boosting with Hyperparameter tuning	82.5	0.87

CONCLUSION & FUTURE ENHANCEMENT

Thus the system represents a significant advancement in exploring the ML model techniques to overcome the challenges faced in the digital industries.It achieved in identifying the customer's intention to stay or leave and engaged at-risk customers without letting them from their subscription plan.

Future enhancements of the system include incorporating advanced techniques such as DL and NLP to extract deeper insights from customer interactions and feedback. Additionally, integrating real-time data sources and implementing dynamic pricing or personalized content recommendations based on exit risk profiles could further improve the project's effectiveness in retaining customers. Moreover, implementing RL algorithms to continuously optimize user behaviors and market dynamics could offer a more adaptive and proactive approach.

REFERENCES

[1] Kalorth and Nithin, *Exploring the Impact of OTT Media on Global Societies*. IGI Global, 2024.

[2] L. Harte, *OTT Business Opportunities: Streaming TV, Advertising, TV Apps, Social TV, and TCommerce*. Discovernet, 2020.

[3] G. Sheetrit, *Over The Top SEO (OTT) is a digital marketing & Professional SEO Agency: Professional SEO OTT*. BookRix, 2023.

[4] H. A. Mengash, N. Al Ruwais, F. Kouki, C. Singla, E. S. Abdel Hameed, and A. Mahmud, "Archimedes Optimization Algorithm-Based Feature Selection with Hybrid Deep-Learning-Based Churn Prediction in Telecom Industries," *Biomimetics*, vol. 9, no. 1, Dec. 2023, doi: 10.3390/biomimetics9010001.

[5] K. Peng, Y. Peng, and W. Li, "Research on customer churn prediction and model interpretability analysis," *PLoS One*, vol. 18, no. 12, p. e0289724, Dec. 2023.

[6] J. Faritha Banu, S. Neelakandan, B. T. Geetha, V. Selvalakshmi, A. Umadevi, and E. O. Martinson, "Artificial Intelligence Based Customer Churn Prediction Model for Business Markets," *Comput. Intell. Neurosci.*, vol. 2022, p. 1703696, Sep. 2022.

[7] I. B. Brandusoiu, *BAYESIAN NETWORKS FOR CHURN PREDICTION IN THE MOBILE TELECOMMUNICATIONS INDUSTRY*. GAER Publishing House, 2020.

[8] L. Owen, *Hyperparameter Tuning with Python: Boost your machine learning model's performance via hyperparameter tuning*. Packt Publishing Ltd, 2022.

[9] T. Agrawal, *Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient*. Apress, 2020.

REFERENCES

[10] Y. Zhou, W. Chen, X. Sun, and D. Yang, "Early warning of telecom enterprise customer churn based on ensemble learning," *PLoS One*, vol. 18, no. 10, p. e0292466, Oct. 2023.

[11] I. AlShourbaji, N. Helian, Y. Sun, A. G. Hussien, L. Abualigah, and B. Elnaim, "An efficient churn prediction model using gradient boosting machine and metaheuristic optimization," *Sci. Rep.*, vol. 13, no. 1, p. 14441, Sep. 2023.

[12] T. Xu, *A New CRM System for Telecoms Customer Churn Analysis Based on Ensemble Learning and RFM*. Overseas Chinese Press, 2023.

[13] E. Crincoli *et al.*, "Deep learning for automatic prediction of early activation of treatment naïve non-exudative MNVs in AMD," *Retina*, Mar. 2024, doi: 10.1097/IAE.0000000000004106.

[14] G. Fatima, S. Khan, F. Aadil, D. H. Kim, G. Atteia, and M. Alabdulhafith, "An autonomous mixed data oversampling method for AIOT-based churn recognition and personalized recommendations using behavioral segmentation," *PeerJ Comput Sci*, vol. 10, p. e1756, Jan. 2024.

[15] M. Zheng, *Spatially Explicit Hyperparameter Optimization for Neural Networks*. Springer Nature, 2021.

[16] I. Vasilev, D. Slater, G. Spacagna, P. Roelants, and V. Zocca, *Python Deep Learning: Exploring deep learning techniques and neural network architectures with PyTorch, Keras, and TensorFlow*. Packt Publishing Ltd, 2019.

[17] S. Vajjala, B. Majumder, A. Gupta, and H. Surana, *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. O'Reilly Media, 2020.

[18] S. Liu *et al.*, "Interaction Pattern Disentangling for Multi-Agent Reinforcement Learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, May 2024, doi: 10.1109/TPAMI.2024.3399936.

THANK YOU