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

MINI PROJECT REPORT

Submitted by

SURYAA KS **210701273**

SYED JAVITH R **210701278**

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

ANNA UNIVERSITY: CHENNAI 600 025

APRIL 2024

RAJALAKSHMI ENGINEERING COLLEGE,

CHENNAI

BONAFIDE CERTIFICATE

Certified that this Report titled "**Enhanced Over the Top Subscribers Retention Prediction using Random Forest with Hyperparameter Tuning**" is the bonafide work of **"SURYAA KS (210701273), SYED JAVITH R (210701278)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Karthick. V

Associate Professor,
Department of Computer Science and Engineering,

Rajalakshmi Engineering College,
Chennai – 602015

Submitted to Mini Project Viva-Voce Examination held on _____

Internal Examiner **External Examiner**

ABSTRACT

With rapid development of Over-the-top (OTT) platforms in the entertainment industry, predicting the user's 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 ML based exit of customer 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 ML model tries to incorporate the features that may determine the exit 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 the Random Forest algorithm with Hyper Parameter tuning

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavor to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time.

We express our sincere thanks to **Dr. P. KUMAR, Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, **Rahul Chiranjeevi V** Professor, Department of Computer Science and Engineering. Rajalakshmi Engineering College for his valuable guidance throughout the course of the project.

SURYAA KS -210701273

SYED JAVITH-210701278

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

OTT	Over-the-top
EDA	Exploratory Data Analysis
CR	Classification Report
BPNN	Backpropagation Neural Network
NLP	Natural Language Processing
PSO	Particle Swarm Optimization

CHAPTER 1
INTRODUCTION

1.1 GENERAL

In the ever-evolving landscape of the Entertainment industry, Enhanced Over-the-top (EOTT) platforms have witnessed a great importance in popularity and significance[3]. 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[4].

1.2 OBJECTIVE

Our main objective for this project lies in providing valuable insights to the owners of OTT platforms to get aware of their areas of improvement and provide a business solution that will be betterment for the economic enhancement in long run. By adopting the strategies for the customer retention will be crucial for any OTT company, leading to an increase in profit and reputation.

1.3 EXISTING SYSTEM

The Existing system for OTT platforms utilizes machine learning algorithms to analyze the users' behavior and predict the likelihood of cancellation of the subscription. It uses GA-XGB to achieve optimal F1 and AUC values[7].These models provide adequate insights for decision making in the banking and telecommunication sector. There was also existing system that relies on ensemble learning which utilizes soft voting for predicting the exit of the customer[14].These systems conducted various evaluation measures

to produce a better accuracy for the newly entered data.

However, the system relies on a simple algorithm such as Logistic Regression[15] which may sometimes lack to detect the comprehensive and complex features from the given dataset and possesses a lack of intricate understanding. The existing system also lacks in preemptively predicting the detection of the customer from OTT platform. It doesn't have the capability to handle huge amounts of data.

1.4 PROPOSED SYSTEM

The proposed system analyzes the user preferences, demographic information and viewing habits and predicts the exit of the customer from their subscription plan. The system aims to identify the most effective model for the retention of the customer. It compares the accuracy of various algorithms and chooses Random forest algorithm for optimal results. After the identification of the best model, it is deployed into the system to find the likelihood of a customer retaining in their subscription plan. The proposed system also uses hyper parameter tuning to control the complexity of the model and to achieve better accuracy. It finds the optimal combination of hyperparameters[16] that produces the best performance of the Machine Learning model on the data which is new for training. This system also avoids overfitting due to the hyper parameter tuning of the algorithms employed. The system also notifies the customer who are not likely to get retained by a SMS using Twilio and Django. Thus our project aims to help all the existing digital over the top platforms in understanding the pros and cons and customers expectation from their platforms. It notifies priorly about the signs of customer likely to leave their platform and subscription and helps them to take necessary actions to improve the customer retention.

CHAPTER 2
LITERATURE SURVEY

Hanan Abdullah Mengash and Nuha AL Ruwais [6] developed a deep-learning-based churn prediction model using the Archimedes Optimization Algorithm for feature selection. This algorithm assesses the effectiveness of feature subsets through an objective function inspired by Archimedes' principle, enabling the selection of the best-performing subset. Their approach enhances the model's accuracy by focusing on the most relevant features, reducing computational complexity, and improving performance. The use of deep learning in conjunction with a sophisticated optimization algorithm allows for handling large datasets and complex patterns. This method provides a robust framework for predicting customer churn, though it may require significant computational resources. Overall, their model offers a promising solution for telecom companies seeking to improve customer retention by accurately identifying potential churners.

Yan Peng's research integrates interpretability analysis [7] for customer churn prediction, employing various sampling methods like SMOTEEN and SMOTE to balance the dataset. They used the XGBoost (XGB) model, optimizing the F1 and AUC values to exceed 90%, ensuring high predictive accuracy. The incorporation of interpretability analysis aids in understanding the underlying factors influencing churn, making the model's predictions more transparent. This approach helps businesses not only predict churn but also develop strategies to mitigate it by addressing the identified factors. The use of advanced sampling methods addresses class imbalance, enhancing model performance. However, the complexity of the XGB model and the interpretability methods might require extensive computational resources. Overall, the study provides a comprehensive framework for accurate and interpretable churn prediction.

Faritha Banu and Neelakandan S utilized the Customer Churn Prediction (CCP) model [8]to distinguish churners from non-churners effectively. They employed Fuzzy Rough Set (FRC) and Quantum Particle Swarm Optimization (QPSO) to optimize feature selection and validate datasets. This approach improves the efficiency of feature selection, enhancing the model's accuracy and robustness. The use of FRC and QPSO allows the model to handle complex customer data, providing a more accurate prediction

of churn. Their methodology ensures that the most relevant features are selected, reducing noise and improving prediction reliability. However, the complexity of their optimization techniques may require significant computational power. Overall, their research offers a sophisticated approach to customer churn prediction, benefiting telecom companies in retaining customers.

The study on Bayesian Network for retention prediction in telecommunications [9] underscores the significant role of Bayesian Networks (BN) in preventing customer churn. BN provides a probabilistic graphical model that captures dependencies among various factors influencing churn, allowing for accurate predictions. This method helps telecom companies identify and address the key drivers of churn, improving customer retention strategies. The probabilistic nature of BN offers insights into the likelihood of different churn scenarios, aiding in strategic decision-making. However, the complexity of BN and the computational resources required may pose challenges. Despite these challenges, the study demonstrates the effectiveness of BN in providing a comprehensive understanding of customer behavior. Overall, the research highlights the potential of Bayesian Networks in enhancing customer retention efforts.

Owen L's study on hyperparameter tuning [10] explores methods to control key features of machine learning models, significantly boosting their performance. By examining techniques such as Harmony Search (HS) and other optimization methods, the study provides insights into improving model accuracy and efficiency. This research offers practical guidelines for selecting and tuning hyperparameters, contributing to more robust and effective predictive models. The study emphasizes the importance of systematic hyperparameter tuning in achieving optimal model performance. By exploring various optimization techniques, it provides a comprehensive framework for enhancing model accuracy and reliability. However, the complexity of these methods may require substantial computational resources. Overall, the study offers valuable insights into improving machine learning models through effective hyperparameter tuning.

Agrawal T's research presents a step-by-step guide for hyperparameter optimization[11], addressing the challenges of time and memory constraints associated with brute-force approaches. The study evaluates various optimization techniques, offering strategies to streamline the hyperparameter tuning process. This comprehensive guide aids researchers and practitioners in efficiently optimizing their models, balancing performance and computational resource requirements. By addressing the limitations of brute-force methods, the study provides a more efficient approach to hyperparameter tuning. The proposed strategies help in reducing the time and memory required for optimization, making the process more practical. However, the effectiveness of these techniques may vary depending on the specific model and dataset. Overall, the research offers practical solutions for enhancing model performance through efficient hyperparameter optimization.

Zhou Chen and Sun X utilized ensemble learning [12] to predict customer churn, combining Back Propagation Neural Network (BPNN) and Random Forest (RF) models. Their ensemble approach leverages the strengths of both models to analyze customer data, providing valuable insights that help telecom companies enhance their retention strategies. By integrating BPNN and RF, the model achieves higher predictive accuracy and robustness. The ensemble method allows for a comprehensive analysis of customer behavior, improving the reliability of churn predictions. However, the complexity of implementing and managing ensemble models may require significant computational resources. Despite these challenges, the research demonstrates the effectiveness of ensemble learning in customer churn prediction. Overall, their study provides a robust framework for improving customer retention through accurate churn prediction.

AlShourbaji I and Helian N employed Gradient Boosting (GB) and meta-optimization[13]techniques to predict customer churn, incorporating a modified Particle Swarm Optimization (PSO) method and Artificial Ecosystem Optimization for hyperparameter tuning. Their approach involves multiple boosting stages with a novel base learner, resulting in a highly optimized and accurate churn prediction model. The use of GB and meta-optimization techniques enhances model performance, providing precise churn predictions. The novel base learner and multiple boosting stages ensure the model's robustness and reliability. However, the complexity of their optimization methods may present implementation challenges and require substantial computational

power. Despite these challenges, their research offers a sophisticated approach to churn prediction. Overall, the study provides a comprehensive framework for improving customer retention through advanced optimization techniques.

Xu T's research on telecom customer exit analysis used ensemble learning[14], particularly Random Forest (RF), to predict churn. By employing stacking models such as XGBoost (XGB), Logistic Regression (LR), and Decision Trees (DT) for soft voting, they achieved superior accuracy compared to other algorithms. The ensemble approach leverages the strengths of multiple models, enhancing predictive accuracy and robustness. This method ensures a comprehensive analysis of customer behavior, providing reliable churn predictions. However, the complexity of stacking models and the computational resources required may pose challenges. Despite these challenges, the research demonstrates the effectiveness of ensemble learning in churn prediction. Overall, Xu T's study offers a robust framework for improving customer retention through accurate and reliable churn predictions.

CHAPTER 3
SYSTEM DESIGN

3.1 DEVELOPMENT ENVIRONMENT

3.1.1 HARDWARE SPECIFICATIONS

This project uses minimal hardware but in order to run the project efficiently without any lack of user experience, the following specifications are recommended

Table 3.1.1 Hardware Specifications

PROCESSOR	Intel Core i5
RAM	4GB or above (DDR4 RAM)
GPU	Intel Integrated Graphics
HARD DISK	6GB
PROCESSOR FREQUENCY	1.5 GHz or above

3.1.2 SOFTWARE SPECIFICATIONS

The software specifications in order to execute the project has been listed down in the below table. The requirements in terms of the software that needs to be pre-installed and the languages needed to develop the project has been listed out below.

Table 3.1.2 Software Specifications

BACK END	Python, Django
LIBRARIES	Numpy, Pandas, Matplotlib, Seaborn, Scikit Learn
SOFTWARES USED	Visual Studio, Jupyter Notebook, Microsoft Excel

3.2 SYSTEM DESIGN

3.2.1 ARCHITECTURE DIAGRAM

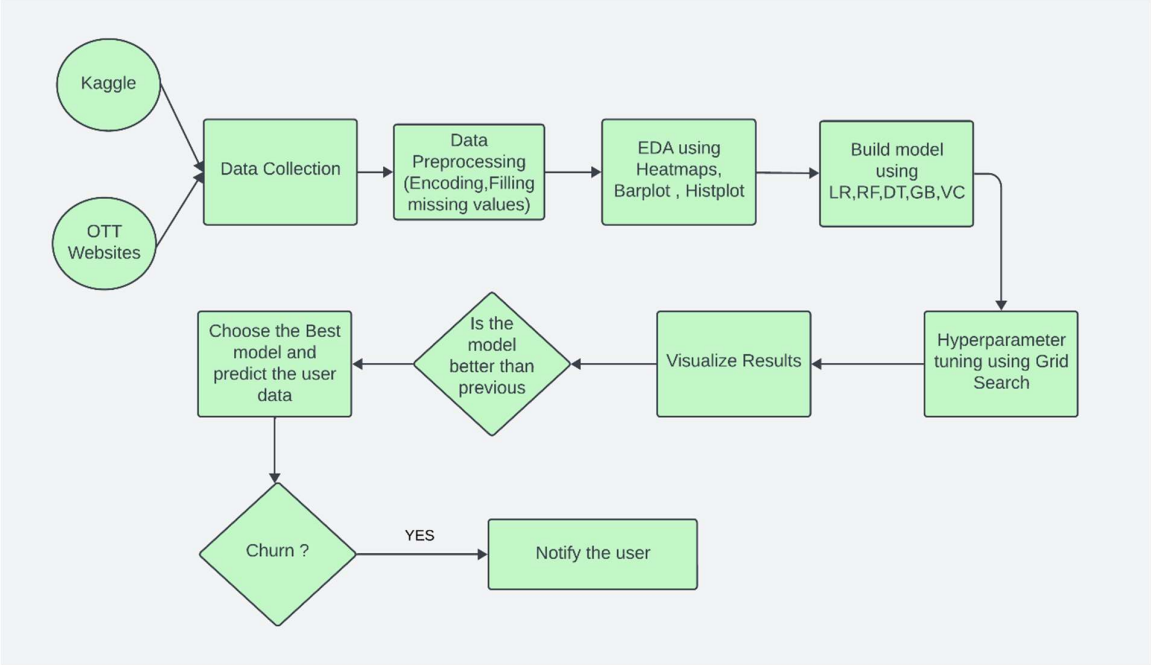

Fig 3.2.1 Architecture Diagram

CHAPTER 4
PROJECT DESCRIPTION

4.1 MODULE DESCRIPTION

4.1.1 DATA COLLECTION:

The first phase of the project involves data from trusted sources such as kaggle and Google dataset search. The data set collected should have desired data columns and be able provide better results and the size should be sufficient enough.

4.1.2 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. Various python libraries specialized for data analysis can be utilized for this purpose such as Numpy, Pandas. This step is crucial for the project as these may cause inefficiency if they are fed directly to the model.

4.1.3 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. Libraries of these are visualization tools commonly used in the project. Through EDA, we concluded that several attributes of users such as phone number, user id etc. are redundant and thus they are dropped. Heatmaps are extensively used to know the correlation between various attributes

4.1.4 BUILD MODELS:

The next step in the project 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. The library will be helpful in building those models.

4.1.5 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.

4.1.6 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 CR and accuracy scores of the models play a vital role in analyzing them.

4.1.7 CHOOSING BEST MODEL:

As we have a track of the model's 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. Among those models, the Enhanced Random Forest algorithm with tuning is concluded as the best model as its performance was at the top. report and accuracy score form the basis of the evaluation of the model.

CHAPTER 5
IMPLEMENTATION AND RESULTS

5.1 IMPLEMENTATION

5.1.1 Random Forest:

Random Forest is a versatile and widely-used ensemble learning method for classification, regression, and other tasks. It works by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. The key idea behind Random Forest is to reduce the overfitting that individual decision trees often exhibit by averaging multiple trees, each trained on a random subset of the data and features. This randomness introduces diversity among the trees, leading to more robust and generalizable predictions.

The strength of Random Forest lies in its ability to handle large datasets with higher dimensionality, including datasets with many features and missing values. It provides an estimate of feature importance, which can be useful for understanding the influence of different variables in the predictive model. Additionally, Random Forests are relatively easy to tune and parallelize, making them efficient in practice. Despite its complexity, the method often performs well out-of-the-box and is less sensitive to hyperparameter settings compared to other machine learning algorithms, making it a popular choice for a wide range of applications.

5.1.2 Hyperparameter Tuning:

Hyperparameter tuning is the process of optimizing the parameters that govern the behavior of a machine learning model but are not learned from the data. These parameters, known as hyperparameters, can significantly affect the performance of a model. Examples include the learning rate in neural networks, the number of trees in a Random Forest, and the regularization strength in logistic regression. Effective hyperparameter tuning can lead to

improved model accuracy, robustness, and generalization to new data. The tuning process typically involves searching over a predefined space of hyperparameters and evaluating the model's performance using techniques such as cross-validation.

There are several methods for hyperparameter tuning, including grid search, random search, and more advanced techniques like Bayesian optimization and genetic algorithms. Grid search exhaustively searches through a specified subset of hyperparameters, evaluating every possible combination, which can be computationally expensive but thorough. Random search, on the other hand, samples hyperparameters randomly, often leading to better results in less time. Bayesian optimization builds a probabilistic model of the objective function and uses it to select the most promising hyperparameters to evaluate, balancing exploration and exploitation efficiently. These methods help identify the best hyperparameters, ultimately enhancing the predictive performance and reliability of machine learning models.

5.2 OUTPUT SCREENSHOTS

After analyzing various user criteria, the system successfully predicted whether a user is likely to retain their subscription for the platform or to churn from the subscription plan. Additionally, the system was able to identify the age category and gender of users who are more likely not to retain their plan. Furthermore, the system notifies users who are likely to churn. These results were achieved with an accuracy of 85%.

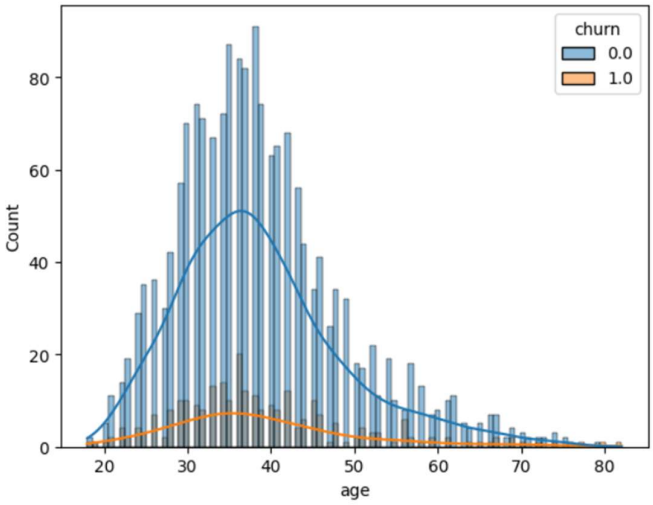

Fig 5.2.1 Distribution plot (age vs churn)

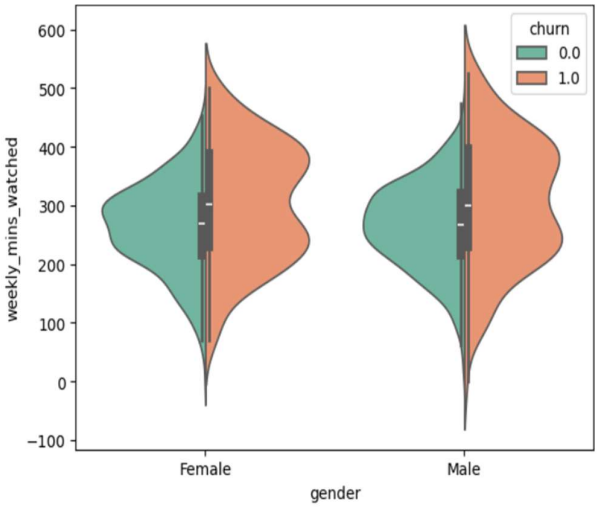

Fig 5.2.2 Weekly minutes watched

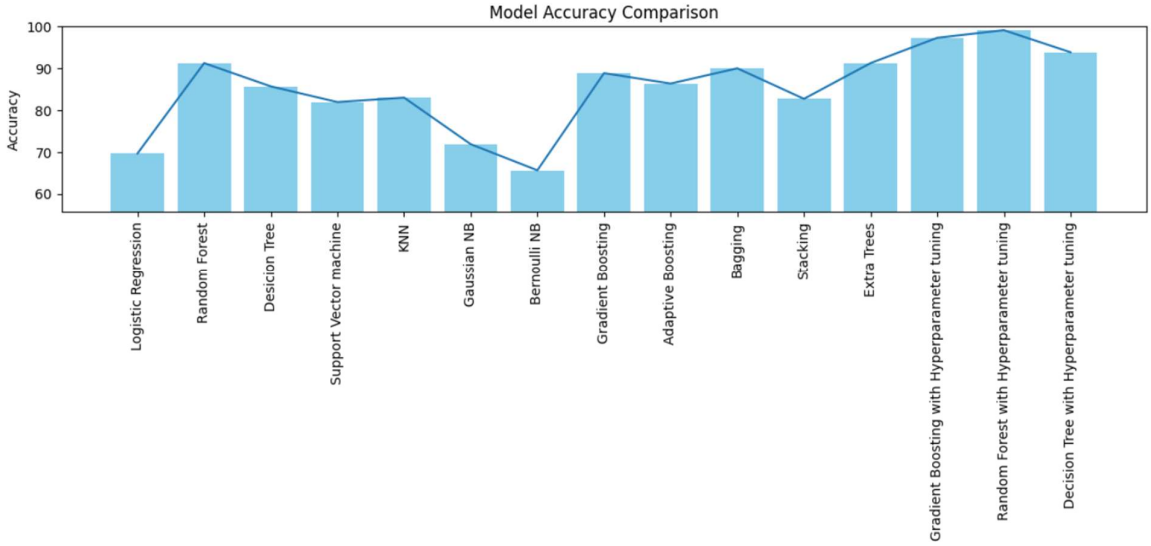

Fig 5.2.3 Models vs Accuracy

CHAPTER 6
CONCLUSION AND FUTURE ENHANCEMENTS

6.1 CONCLUSION

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 [17]and NLP[18] 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[19] algorithms to continuously optimize user behaviors and market dynamics could offer a more adaptive and proactive approach.

6.2 FUTURE ENHANCEMENTS

The Future enhancement encompasses the model to be integrated with a real time application where the user can provide data and the model will predict and behave accordingly. The use of backend web technologies such as Django, flask, Stream lit etc. can facilitate this real time application along with a user-friendly interface development. As the model gets used to the real system, it will be exposed to a recent, large and diverse dataset which further enhances the studies of the user behavioral pattern and interactions.

REFERENCES

[1] W. Sullivan, *Decision Tree and Random Forest: Machine Learning and Algorithms: The Future Is Here!* Createspace Independent Publishing Platform, 2018.

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

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

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

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

[6] H. A. Mengash, N. Al Ruwais, F. Kouki, C. Singla, E. S. Abd Elhameed, 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.

[7] 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.

[8] 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.

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

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

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

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