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 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 over the top platform data like user preferences, demographic information and user viewing habits. With the use of historical data on the user subscription and interaction pattern using almost 12 distinct features for the model, our ML model tries to incorporate these features that may determine the exit of the user. Retaining the users has been a significant problem for the companies.Approximately 5-7% percentage of the total users gets exited from companies subscription plan.The dataset utilized for this model takes these 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 RF algorithm with HP tuning

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

I . INTRODUCTION

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]. Moreover the companies can also manage their businesses in an efficient manner to tackle the upcoming challenges on the way. The usage of Enhanced Over-The-Top (EOTT) 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 platforms[5]. This project aims to develop a machine learning solution for subscriber retention prediction customized specifically for EOTT 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 platforms in the ever growing digital entertainment landscape.

Hanan Abdullah Mengash, Nuha Alruwais[6] 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.It allows to select the best performing subset as the optimal solution.

Yan peng [7]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%.

Faritha Banu along with Neelakandan S [8]in their research work 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.

The research on Bayesian Network[9]for retention prediction in telecommunication speaks on the volume of influence on Bayesian networks in preventing the customer to stay within their subscriptions.

Owen L [2] in his study on Hyperparameter tuning 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.

Research paper from Agrawal T[10] explores the step by step guide for hyperparameter optimization.It addresses the brute force approaches including the problem of time and memory constraints.

Zhou Chen and Sun X[11] 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.

AlShourbaji I,Helian N[12] 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.They used novel base learner to perform multiple boosting stages.

Xu T in his research work [13] for 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.They produce much better accuracy than other algorithms.

While these models demonstrate good results,they exhibit various limitations in terms of scalability and generalizability across datasets.Their reliance on specific algorithms may limit their flexibility to explore.Their interpretability of the models and the practical implications may be straightforward which may cause trouble in decision making.

II. MATERIALS AND METHODS

For the project the hardware requirements include a PC with minimum of 16GB RAM, 256 GB SSD, and a quad-core processor. Besides these, A GPU is recommended for better performance. Also ensure a Good network connectivity throughout the training.

The Software requirements that need to be installed for the better development of the project include Jupyter Notebook , Python ,Visual Studio Code , Python Package manager, External libraries (such as Numpy, Pandas, Matplotlib, Seaborn, and ScikitLearn), Microsoft Excel and Git, Django, Postman , Twilio.

The Dataset for the project is collected from kaggle named “Churn Modeling for OTT Platform” with a size of 50KB. It took almost 300 iterations to complete the project.

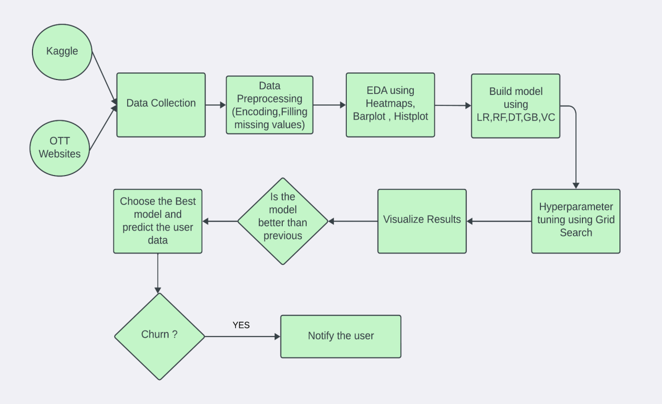
III. EXISTING ALGORITHM

The Existing system for OTT platforms utilizes machine learning algorithms to analyze the user 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 were 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 LR[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.

IV. 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.


Kaggle
OTT Websites
Data Collection
Data Preprocessing (Encoding,Filling missing values)
EDA using Heatmaps, Barplot , Histplot
Build model using LR,RF,DT,GB,VC
Hyperparameter tuning using Grid Search
Visualize Results
Is the model better than previous
Choose the Best model and predict the user data
Churn ?
YES
Notify the user

V. METHODOLOGY

A. Data Collection :

The first phase of the project involves data from trusted sources such as kaggle and Google dataset search in desired format such as CSV, JSON. The data set collected should have desired data columns and be able to provide better results and the size should be sufficient enough.The Dataset can also be collected from other trusted sources such as real-time systems.

B. 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.

C. 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.

D. 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.

E. 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.

F. Visualizing Results :

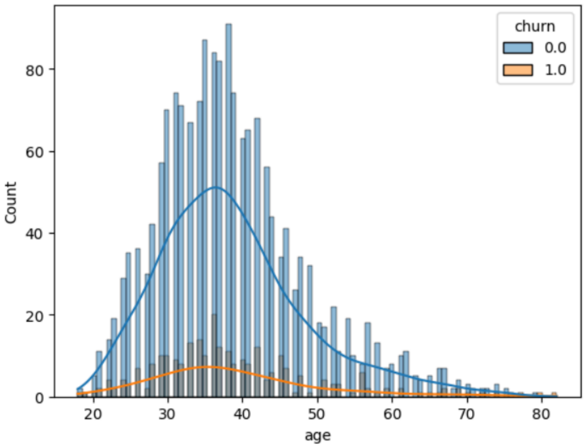
The results of various parameters and also the accuracies along the time are analyzed in this step to get insights of the working of various models. The Confusion matrix plays an important role in analyzing the performance of models. With the help of an intuitive bar graph, we can infer the nobel model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

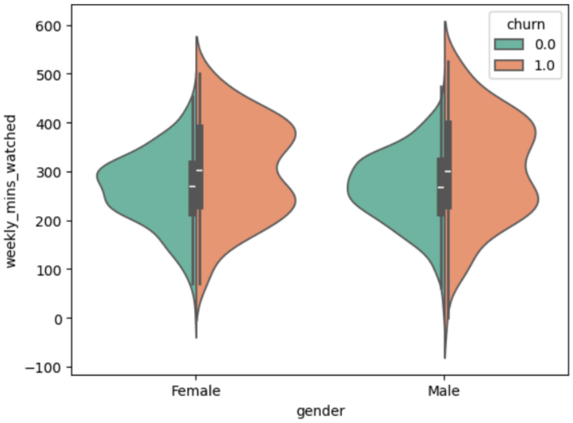
G. Choosing Best Model :

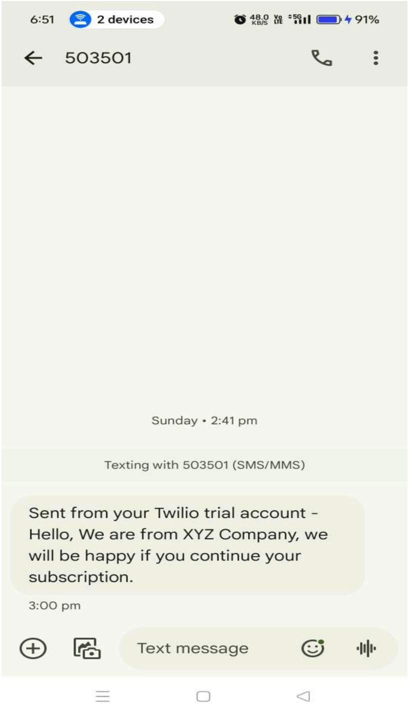
As we have a track of the model performance we can choose the best model among the trained models and can utilize it for the further development of the project which leads to an iterative development process. Among those models, the Enhanced Random Forest algorithm with tuning is concluded as the best model as its performance was at the top. report and accuracy score form the basis of the evaluation of the model.

VI RESULTS

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%.







Sunday • 2:41 pm

Texting with 503501 (SMS/MMS)

Sent from your Twilio trial account - Hello, We are from XYZ Company, we will be happy if you continue your subscription.

3:00 pm

VII DISCUSSION

In our project, the crucial importance of customer interactions in predicting the exit patterns have been clearly emphasized using various machine learning algorithms. The study has shown that the customer behaviors play a vital role in profit generation and trend analysis. Handling the user data for the sake of providing assistance to the OTT platforms resulted in improved performance in the area of improvements. The Project has analyzed various existing methodologies and algorithms for predicting the behavior, followed by the training and analyzing, we have come up with the Random Forest algorithm with hyperparameter tuning as a better model with an estimated efficiency of 85%.

As the model has been trained with different algorithms, it possesses a wide scope and flexibility with the change in data and trends. The model can be furnished and improved in its performance, thus exhibiting an adaptability behavior which is crucial for a software throughout its lifetime. These features will also give the model an upperhand in encompassing new trends in the field of the Over-the-top platforms.

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 , Streamlit 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.

VIII 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.

IX 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: 10.1109/TPAMI.2024.3399936.