CHURN PREDICTION IN OVER-THE-TOP(OTT) FOR CUSTOMER RETENSION USING MACHINE LEARNING ALGORITHMS

Malepati Amulya
SCOPE
Vellore Institute of Technology
Chennai, India
Malepati.amulya2021@vitstudent.ac.in

Giri Praneetha
SCOPE
Vellore Institute of Technology
Chennai, India
Giri.praneetha2021@vitstudent.ac.in

Praveen Joe I R
SCOPE
Vellore Institute of Technology
Chennai, India
Praveen.joe@vit.ac.in

Abstract

In the view of content providers like Over-The-Top (OTT), the ability to predict the amount of churn is a key part of the organization. With these predictions, the company can make better strategies in order to reduce the churn rate. This paper presents a comprehensive study on Churn Prediction in Over-The-Top (OTT) using various Machine Learning Algorithms. These include Logistic Regression, Decision Tree using both Entropy and Gini as parameters, Random Forest, K-Nearest Neighbour classification, Naive Bayes, Support Vector Machine (SVM), XG Boost, Gradient Boost algorithms. In which the class imbalance is found and treated using Synthetic Minority Over-sampling Technique (SMOTE) and re-performed the machine learning algorithms, in which the accuracy all algorithms are greater than 74% and better F1-Score, These findings can be useful to the companies with real time data and to find the reasons behind customer attrition and increase their customer life value and customer satisfaction.

Key Words

Over-The-Top, Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbour classification, Naive Bayes, Support Vector Machine (SVM), XG Boost, Gradient Boost, Over Sampling, SMOTE, Churn Prediction.

INTRODUCTION

Churn is defined as the how many number of customers are decided to leave the particular company. Churn Prediction is the process of identifying consumers who pose a danger of cancelling their subscriptions or closing their accounts altogether. It also detects the customers those who are in risk of rejecting the subscription. Over-The-Top (OTT) provides content like movies, web series, etc. customers take subscription in order to get entertained but due to some reasons they drop the subscriptions in the middle and leave the platform, these platforms will predict on what reasons customers are leaving. In OTT platforms churn prediction is a vital concern. Churn prediction is basically use machine learning algorithms to detect the subscribed people leaving the platform. Churn prediction considers both the "why" and the "who". Companies may learn a lot about the factors behind customer attrition by examining the data used to forecast churn. Various causes may contribute to this, such as competitive products, unsatisfactory customer service, absence of desired features, or cost.

Machine learning is very important for analysing the customer data for future prediction over churn in OTT. First, we collect the data on customers. Train the data by using machine learning algorithms like Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbour (KNN), Naive Bayes, Support Vector Machine (SVM), XG Boost, Gradient Boosting. After completion of training, we need to evaluate the model to know the accuracy in predicting churn. Once the model is done then it used to predict the churn and reasons for the same.

Developing a model that can precisely anticipate whether a customer will stick with this platform or not is the aim of utilizing

machine learning to forecast subscriptions for Over-The-Top (OTT) services. Strategy for client segmentation can be used with churn prediction. Businesses are able to construct more individualized customer experiences that meet a range of wants and preferences by segmenting their client base according to churn risk and other pertinent characteristics. OTT firms need this information to better understand and manage their marketing and retention campaigns. To train the model, pertinent data such as watching preferences and consumer demographics is gathered and examined. In this procedure, significant characteristics are chosen, the data is cleaned, and a machine learning model is trained.

In the following sections, we see each and every methodology in detail, how the models are performed, how the performance metrics is calculated, finding the class imbalance and performing over sampling using SMOTE, drawing recommendations and conclusions.

RELATED WORKS

[1] uses machine learning techniques like Hierarchical logistic regression, decision tree, random forest, Ada Boost. Factors like multiple subscription, switching frequency, content satisfaction, price satisfaction have higher impact on customer churn. These are found with most effective algorithm among all those Random Forest, this approach has higher accuracy compared to others.

Retension strategies in order to reduce churn in OTT platforms are clearly discussed in [2]. Finding the most significant attributes for churn of a particular individual, the OTT platform can take necessary steps in order to reduce churn, Content satisfaction shows more effect on churn, so by taking the videos, movies, that are highly satisfied by the viewers churn can be reduced. The highly satisfied content can be collected by viewers or customers feedback, review of a particular movie or video, etc. Showing the related content to the viewers is another strategy of OTT. This can be done by having the data of one viewer, what kind of movies they are continuously watching, what genre they are interested. Author used logistic regression, multi-layer perceptron, random forest, decision trees, and gradient boosting machines and also bought the accuracy of 80%. However they faced the problems with the data, the model built was complex, model drift.

Churn of one organization depends on the competitors. With the increase of technology, OTT platforms are increasing day-by-day, this can become a big hurdle for one platform. So, these should build strategies by keeping competitors in mind. These are explained in detailed in [3]. The availability of competing services, other platform price, have effect in churn.

Other than machine learning algorithms, Comprehensive Understanding was done in [4]. They stated that there will be increase of paid subscribers by 16.1% by the year 2028 i.e, the subscriber market will increase from USD104.2 billion to USD293 billion.

Over-The-Top (OTT) providers and Internet Services Providers (ISPs) joint service management approach based on Customer Lifetime Value (CLV) and benefits of joint services management are discussed in [5]. They also stated that this can improve the customer experiences, increase customer loyalty which are key factors in reducing churn. Over-The-Top (OTT) providers and Internet Services Providers (ISPs) joint service management approach based on Quality of Experience (QoE) and benefits of joint services management [7]. This is a measure of satisfaction got by particular viewer with respect to the service they received. Regression analysis is performed between QoE and Churn in order to get a relationship among those.

Over-The-Top (OTT) Churn is more effected by content provided by particular OTT platforms and also the price charged for that, plans and subscription options provided. This is analysed by performing content analysis and economic analysis [6]. Found that customers are increasing because they feel that OTT platforms are for providing entertainment, treat as stress busters. This was found based on various factors like variety of content, affordability of OTT subscriptions.

[8] The author stated there are increase of OTT subscribers during Covid-19. Almost 7.5 million subscribers have been increased from the year 2019 to 2020. Factors like Covid-19 pandemic, increasing availability of high-speed

internet, growth popularity of streaming devices showed significant effect on this particular growth.

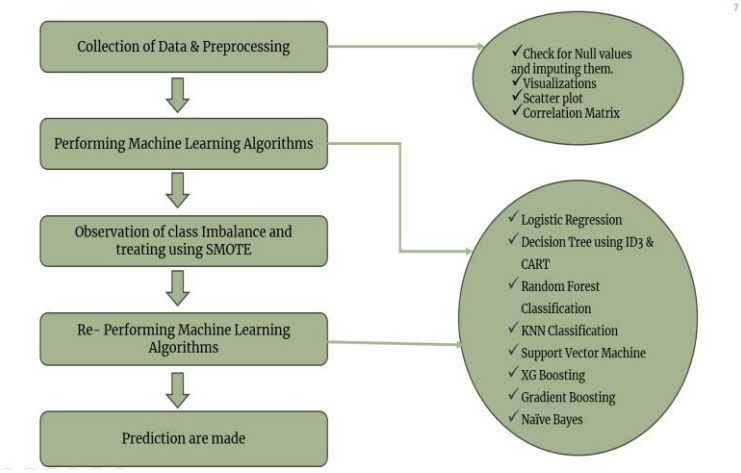
METHODOLOGY

About the dataset

The dataset contains 16 attributes along with the target variable 'churn' that is binary which states 0-no, 1-yes. The 15 independent variables are

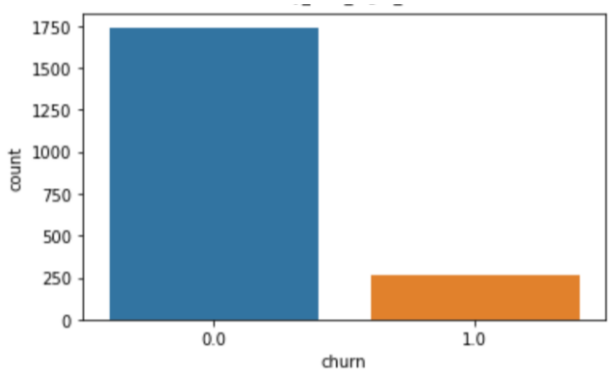
Year, customer_id, phone_no, Gender, Age, no_of_days_subcribed, multi_screen,

mail_subscrided, weekly_mins_watched, minimum_daily_mins, maximum_daily_mins, weekly_max_night_mins, videos_watched, maximum_days_inactive, customer_support_calls. The dataset contains 2000 entries.



(i) Data Preprocessing

The initial step of the project includes preprocessing steps like removing unnecessary attributes, handling null values, outlier detection, some visualizations, creation of dummies. Here we removed customer_id, year, phone_no attributes. While treating with null values, we found the attributes gender, maximum_days_inactive, churn have null values. Using imputation methods, we treated these, gender is filled with mode, maximum_days_inactive is filled with median, churn is filled with mode. No outliers are detected hence proceed further. When correlation matrix is plotted, found out that maximum_days_inactive is highly correlated minimum_daily_min. Attributes like gender, multi_screen, mail_subscribed are converted categorical to binary through dummies.



(ii) Model Selection

Splitting of the dataset into training and testing is done, 40% of the data is provided for testing and 60% for training and random state is taken as 78 and built several machine learning algorithms. Models like Logistic Regression, Decision Tree using entropy and gini, Random Forest, K-Nearest Neighbour Classification (KNN), Support Vector Machine (SVM), Gradient Boosting, XG Boosting, Naive Bayes are trained, tested, and validated.

(iii) Performance Evaluation

Performance metrics, Confusion metrics is plotted for each model and Accuracy, Precision, Recall, F1-Score are retained, from that the efficiency of the model is been said. Random Forest has higher accuracy followed by XG Boost and then K-Nearest Neighbour. However, we got F1-Score less than 0.5 for most of the models.

Accuracy = TP + TN / TP + FP +TN + FN

Precision = TP / TP + FP

Recall = TP / TP + FN

F1-Score = 2 * Precision * Recall / Precision + Recall

Out[80]:

	Accuracy	F1_score	Recall	Precision
Random_Forest	0.912	0.602	0.486	0.791
Decision_Tree_Classifier_CART	0.910	0.566	0.431	0.825
Decision_Tree_Classifier_ID3	0.909	0.563	0.431	0.810
XG_Boosting	0.908	0.593	0.495	0.740
KNN_Classifier	0.875	0.315	0.211	0.622
Gradient_Boosting	0.872	0.346	0.248	0.574
Logistic_regression	0.865	0.143	0.083	0.529
Support_Vector_Machine	0.864	0.000	0.000	0.000
Naive Bayes	0.846	0.417	0.404	0.431

(iv) Treating with Class Imbalance

Less F1-Score is occurred due to class imbalance, this is treated with Synthetic Minority Over-sampling Technique (SMOTE), for oversampling the data, after performing SMOTE, all the above machine learning models are performed again for the new data and do the evaluation. After this we found out that XG Boost has high accuracy followed by Random Forest and Decision Tree with entropy.

RESULTS & DISCUSSION

All models are compared with accuracy, precision, recall and F1-score to understand the performance of various algorithms for churn prediction. In all the proposed models, XG Boost has high accuracy of 91% followed by Random Forest with accuracy of 89% and Decision Tree with entropy with accuracy of 89%. Logistic Regression has less accuracy of 74% and Support Vector Machine (SVM) has accuracy of 86% but precision, recall, F1-score are 0%, this states SVM does not support churn Prediction in OTT. All the performance metrics is as mentioned in the following table.

	Accuracy	F1_score	Recall	Precision
XG_Boosting_SMOTE	0.910	0.633	0.569	0.713
Decision_Tree_Classifier_CART_SMOTE	0.898	0.643	0.679	0.612
Random_Forest_SMOTE	0.898	0.627	0.633	0.622
Decision_Tree_Classifier_ID3_SMOTE	0.890	0.614	0.642	0.588
Gradient_Boosting_SMOTE	0.870	0.500	0.477	0.525
Support_Vector_Machine	0.864	0.000	0.000	0.000
Naive Bayes_SMOTE	0.802	0.494	0.706	0.379
KNN_Classifier_SMOTE	0.765	0.444	0.688	0.328
Logistic_regression_SMOTE	0.745	0.427	0.697	0.308

CONCLUSION

Churn prediction models are useful in identifying typical problems, including excessive wait times or terrible customer service, that result in subscriber churns. Businesses may utilize this data to promptly handle subscriber problems and enhance their customer service. Businesses may lower attrition and boost subscriber retention by rewarding devoted customers with special offers and incentives. Which subscribers are most likely to respond to loyalty programs and what kinds of rewards work best may be determined with the use of these models. Through the analysis of subscriber behavior and subscription history, businesses may enhance their pricing methods.

In this project, eight machine learning algorithms had been used and we got the highest accuracy to be 0.90 and 0.91 in random forest and XG Boost. Random forest has higher accuracy when class imbalance is not treated where as XG Boost has high accuracy before and after treating the class imbalance.

We draw the conclusion that ensemble approaches for churn prediction will yield excellent accuracy as well as additional performance measures. Future developments, such as the use of AI chatbots and gamification, may contribute to this effort.

Chatbots with artificial intelligence (AI) can engage with subscribers and detect those who are likely to leave. In order to reduce customer attrition, chatbots may also be utilized to send subscribers tailored offers and suggestions.
It is possible to utilize gamification to motivate users to stick around on the service. One way to achieve this is by providing incentives for viewing particular material or urging others to sign up.

REFERENCES

[1] Mohan, M., & Jadhav, A. (2022). Predicting customer churn on OTT platforms: Customers with subscription of multiple service providers. Journal of the Association for Information Science and Technology, 73(1), 1-15.

[2] Senthil Kumar, Needhi Devan, "Ott Subscriber Churn Prediction Using Machine

Learning" (2023). Electronic Theses, Projects, and Dissertations. 1660.

[3] Manish Mohan, Anil Jadhav (2022). Predicting Customer Churn on OTT Platforms: Customers with Subscription of Multiple Service Providers. Journal of Information & Organizational Sciences.

[4] Sistla Srivalli Leela Praveena, Dr. Vinay Negi-Over-The-Top (OTT) 2021-Video Market:Rise of Paid Subscription Viewers Study

[5] A. Ahmad, A. Floris and L. Atzori, "OTT-ISP joint service management: A Customer Lifetime Value based approach," 2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), Lisbon, Portugal, 2017, pp. 1017-1022, doi: 10.23919/INM.2017.7987431

[6] Priya Malhotra, Akshay Kumar (2021),"Market Research and Analytics on Rise of OTT Platforms: A study of Consumer Behaviour" International Journal of Advances in Engineering and Management (IJAEM) Volume 3, Issue 7 July 2021, pp: 4005-4012

[7] A. Ahmad, A. Floris and L. Atzori, "QoE-aware service delivery: A joint-venture approach for content and network providers," 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, Portugal, 2016, pp. 1-6, doi: 10.1109/QoMEX.2016.7498972.

[8] Sachika Luthra- The Impact of Covid-19 on Consumer Perception Towards Subscription Based OTT Platforms.

[9] Anish Yousaf, Abhishek Mishra 2021- A cross-country analysis of the determinants of customer recommendation intentions for over-the-top (OTT) platforms.

[10] E. Liotou, G. Tseliou, K. Samdanis, D. Tsolkas, F. Adelantado and C. Verikoukis, "An SDN QoE-service for dynamically enhancing the performance of OTT applications," 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), Pilos, Greece, 2015, pp. 1-2, doi: 10.1109/QoMEX.2015.7148106.