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

By

K Anusha 22MIS1032

K Divya 22MIS1002

Under the guidance of

Dr. Pattabiraman V.

in partial fulfillment of the course

**SWE2009 Data Mining**



**School of Computer Science and Engineering (SCOPE)**

**Vellore Institute of Technology**

**Chennai-600127**

**April 23, 2024**

****

**BONAFIDE** **CERTIFICATE**

This is to certify that the project work entitled “**Churn Prediction in Over the Top (OTT) for customer retention using Machine Learning Algorithms**” this is a bonаfidе work of **K Anusha-22MIS1032 and K Divya-22MIS1002**, who carried out the project work under my supervision and guidance for SWE2009 -Data Mining

**Dr. Pattabiraman V.**

Professor

School of Computer Science and Engineering (SCOPE)

VIT University, Chennai

Chennai – 600 127.

**АCKNOWLЕDGЕMЕNT**

We would like to express our special thanks and gratitude to our Course faculty **Dr. Pattabiraman V.** Professor of School of Computer Science and Engineering who gave us the golden opportunity to do this wonderful project titled “ Churn Prediction in Over the Top (OTT) for customer retention using Machine Learning Algorithms“, and also for providing us with proper guidance and suggestions to develop this project. This helped us in doing а lot of research and we came to know about so many new things.

|  |  |
| --- | --- |
| Title  **Table of Contents** | Page No |
| 1. Abstract | 5 |
| 1. Scope | 6 |
| 1. Objective | 7 |
| 1. Introduction | 8-9 |
| 1. Literature Survey | 10-12 |
| 1. Tools used | 12 |
| 1. Dataset Description | 13-14 |
| 1. Architecture | 15-17 |
| 1. Proposed works    1. Module 1 – Observed of class Imbalance SMORT    2. Module 2 – LSTM (Long Short-Term Memory)    3. Module 3 – Neural Network model, specifically a Multi-Layer Perceptron (MLP) | 18-22 |
| 1. Innovation | 23 |
| 1. Results and discussion | 24-25 |
| 1. Conclusion and Future Work | 26 |
| 1. References | 27-28 |

GitHub Link : <https://github.com/kothapalliAnusha/Data_Mining_churn_prediction.git>

**1.ABSTRACT**

OTT services have drastically changed the entertainment sector by providing users with instant access to a diverse selection of content. Nonetheless, elevated churn rates present a considerable financial danger for OTT platforms, leading to decreased revenue and diminished customer allegiance. This study suggests a new method, using machine learning, to predict churn rates and recognize subscribers who are at risk.

We use a combination of ensemble learning, Long Short-Term Memory (LSTM), and Multilayer Perceptron (MLP) techniques to improve both the accuracy and reliability of predictions. Furthermore, we utilize SMOTE to address imbalanced data and capture temporal patterns in customer behaviour with LSTM. Furthermore, MLP models offer understanding of intricate nonlinear connections between different characteristics, resulting in enhanced forecasting accuracy.

The study showed that our machine learning approach is successful in predicting churn rates, allowing OTT platforms to take action and use specific strategies to retain customers. These tactics could involve customized deals, improved customer support, and the launch of personalized services based on specific preferences.

Our study emphasizes the capacity of machine learning to enhance customer retention strategies in OTT platforms. By utilizing complex analytics and forecasting techniques, OTT providers can reduce customer loss, nurture lasting relationships, and maintain profits in a competitive environment.

Keywords: Online streaming services, unsubscribe rates, computer programs that learn from data patterns, making predictions, keeping customers, customized price reductions, client support, extra functions, past customer information, chances of losing customers, precision, methods to keep customers interested, Synthetic Minority Over-sampling Technique, Long Short-Term Memory, Multilayer Perceptron.

**2.SCOPE**

This project is a thorough effort to fully utilize the capability of machine learning methods such as Ensemble Learning, Long Short-Term Memory (LSTM), and Multilayer Perceptron (MLP) for effectively forecasting churn on Over-the-Top (OTT) platforms. The project's goal is to build a strong prediction model that can accurately assess the possibility of subscriber attrition by leveraging significant previous customer data. The scope includes several stages, such as algorithm selection, feature engineering to extract relevant insights from the dataset, model training with advanced techniques, thorough evaluation to assess performance metrics, and seamless integration into the operational framework of OTT platforms. Furthermore, the study uses the Synthetic Minority Over-sampling Technique (SMOTE) to handle data imbalances and LSTM to identify temporal dependencies in customer behaviour. Continuous refining and optimization tactics will be used to assure the model's usefulness in real-world circumstances, with the overall goal of increasing customer retention and reducing revenue loss due to churn. This initiative promotes long-term growth in the sector by creating and analysing churn prediction algorithms specifically designed for OTT platforms.

**3.OBJECTIVE**

The main objective of this project is to create and implement a highly precise machine learning model for predicting customer churn on an OTT platform. This will allow for the proactive deployment of retention tactics to reduce revenue loss and enhance customer satisfaction. To achieve this ambitious objective, our attention will be directed towards setting up the following sub-objectives.

Analyse and compare various classification methods such as SVM, Random Forest, like wise various classification methods and to compare the performance before and after SMOTE and for history comparison and error decreased progression LSTM and using MLP in order to determine the most suitable approach for our specific dataset.

Reiterate the following text using the same source language and maintain the consistent word count: Attain a 90% accuracy rate by taking into account precision, recall, and F1-score metrics to assess the model's dependability and efficiency in predicting churn.

The text is not provided. Improve understanding and useful insights by clarifying the main factors that impact churn predictions and offering actionable suggestions for specific retention efforts, enabling the OTT platform to keep valuable customers proactively.

This study intends to offer a complete resolution for the OTT platform in order to improve customer satisfaction and decrease churn rates, ultimately promoting long-term growth and profitability in the fiercely competitive entertainment industry.

**4.INTRODUCTION**

* Predicting customer churn is essential for OTT platforms hence it involves identifying users who we predict that have high chances to terminate their subscriptions, which resulting in decline the huge amount of revenue. Machine learning, is one of the powerful subset of artificial intelligence that can recognize patterns in data, and has a significant potential in predicting churn in the OTT industry by identifying behaviour patterns that are associated with churn.
* customer churn can be done through various machine learning methods such as random forest, SVM, XE-Booosting and many more algorithms. Logistic regression is a simple method employed to forecast employee turnover by assessing the likelihood of someone leaving based on different factors. Decision trees along with random forests enhance prediction precision by elucidating the connections between customer characteristics and churn through tree-based formations and collections of decision trees.
* Once an appropriate machine learning algorithm is chosen, training must be conducted using historical customer data that includes behavioural patterns like how often content is consumed, viewing preferences, and recent sign-in activity. After being instructed, the model can forecast future turnovers for newly acquired customers through analysis of up-to-date customer information and assessing the probability of customer defection.
* Predicting churn is very beneficial for over-the-top (OTT) providers as it helps in identifying customers who are at risk of leaving, enabling them to take pre-emptive actions to enhance client loyalty, earnings, customer happiness, and brand value.
* Machine learning offers several advantages for predicting churn in the OTT industry, such as unparalleled accuracy by identifying subtle behavioural patterns that signal churn. Additionally, its ability to scale ensures efficient handling of large amounts of data found in OTT platforms, and its real-time prediction features enable providers to proactively reduce churn, showcasing the effectiveness of machine learning in enhancing OTT churn prediction.
* Furthermore, by implementing latest techniques like Ensemble Learning for show casing the results before and after using Smote, Long Short-Term Memory (LSTM) for the historical data and error reduction progression check, and Multilayer Perceptron (MLP) to check which whether its best for churn prediction, these will lead to prediction accuracy and robustness are improved, allowing OTT providers to identify customers who may churn.

**5.LITERATURE REVIEW**

The advancement of OTT platforms has changed the entertainment sector and also human interests, offering consumers new chances and more benefits to access various digital content. Yet, this advancement has introduced a new obstacle: customer churn, when subscribers cancel their memberships, creating major risks for the income and endurance of OTT providers. Academics have thoroughly examined churn prediction in the OTT sector, utilizing a variety of methods from simple statistical analysis to sophisticated machine learning algorithms. These efforts focus on identifying the root causes of churn and creating predictive models that can accurately predict subscriber behaviour. Mohan and Jadhav's (2022) [2] seminal work delves deep into churn prediction on OTT platforms, unveiling the intricate interplay of factors influencing subscriber attrition. By examining variables such as subscription patterns, content satisfaction levels, and pricing strategies, their research underscores the significance of data-driven insights in crafting effective retention strategies. Furthermore, their advocacy for the random forest algorithm as a robust tool for churn prediction remains consistent across subsequent studies, laying a strong foundation for predictive analytics in the OTT domain. Kumar's (2023) [3] empirical exploration of churn prediction systems offers valuable insights into the nuanced challenges associated with modelling subscriber behaviour. While achieving commendable accuracy rates, Kumar's research illuminates ongoing concerns regarding data biases and model complexity, prompting academia and industry practitioners to address these challenges with innovation and diligence. Rakesh, Pudiyavitil, and Chandrashekhar's (2020) [4] scholarly investigation into user modelling and churn prediction provides a holistic perspective on the multifaceted nature of customer attrition. Their results emphasize the significance of creating personalized, data-driven retention strategies for a variety of subscriber groups by examining factors like content satisfaction, price sensitivity, and customer service quality. Ahmad's (2017) [5] groundbreaking research on implementing Customer Lifetime Value in collaborative service management between OTT and ISPs represents a fundamental change in how customer relationships are managed. By emphasizing the enhancement of customer experience and strategic alliances, Ahmad's research lays the groundwork for fostering long-term success and customer loyalty. Malhotra, Kumar, and Yadav's (2021) [6] comprehensive market analysis sheds light on the transformative impact of consumer behaviour on the proliferation of OTT platforms. Their examination of emerging trends and evolving consumer preferences serves as a roadmap for industry players seeking to capitalize on market opportunities and bolster their competitive position. Ahmad's Quality of Experience (QoE) [7] aware service delivery paradigm and Luthra's investigation into the COVID-19 pandemic's ramifications on consumer perceptions of OTT platforms underscore the importance of adaptability and responsiveness in navigating changing market dynamics. Their research underscores the imperative of innovative approaches to enhancing customer satisfaction and resilience amidst external challenges. Yousaf and Mishra's (2021) [8] cross-country study underscores the significance of cultural sensitivity and market localization in tailoring services to diverse client cohorts. Their findings emphasize the importance of customizing retention strategies to align with the unique preferences and expectations of global audiences. Liotou and Tseliou's (2019) [9] pioneering concept of SDN QoE-service exemplifies the relentless pursuit of technological innovation in optimizing OTT application performance. Their research, incorporating cutting-edge network management strategies, marks a new frontier in enhancing user experience and fostering sustained engagement. Li, X., Wu, H., and Wang, Z.'s [10] utilization of neural networks to predict customer turnover in the OTT market enriches our understanding of the intricate dynamics of consumer behaviour and churn prediction in this domain. Verhoef, P.C., Neslin, S.A., and Vroomen, B.'s [11] study on multichannel customer management offers insights into the broader context of customer relationship management and its applicability to the OTT industry. Understanding the research-consumer phenomena can inform the development of effective retention strategies. Han J., Kamber M., and Pei J.'s [12] seminal work on data mining principles and techniques provides foundational knowledge necessary for comprehending the methodologies employed in OTT churn prediction research. Breiman, L.'s [13] introduction of random forests as a powerful ensemble learning technique widely utilized in churn prediction models is essential for developing efficient machine learning algorithms in the OTT context. The important book by Hastie, T., Tibshirani, R., and Friedman, J. explains key concepts in statistical learning essential for developing and assessing machine learning models for predicting churn in OTT platforms. This Examination of literature provides a detailed summary of the various difficulties and possibilities involved in predicting churn in the OTT sector by combining different viewpoints. It establishes the foundation for creating and executing successful retention strategies in the digital entertainment industry by combining real-world data, theoretical models, and technological progress to promote continued growth and nurture lasting customer relationships.

**6.TOOLS USED**

Google Collab, a platform hosted in the cloud by Google, functions as the main workspace utilized for performing data analysis and machine learning activities in this project. Google Collab offers a Jupyter Notebook interface that allows users to generate and distribute documents with live code, equations, visualizations, and explanatory text.

Python continues to be the main programming language used in Google Collab, recognized for its flexibility and vast library resources for analysing data and performing machine learning jobs. Google Collab allows researchers to use Python's dynamic typing, garbage collection, and support for various programming paradigms to create complex churn prediction models by smoothly integrating with the language.

Furthermore, this project employs advanced machine learning techniques such as Ensemble Learning, LSTM, and (MLP) algorithms using the Google Collab platform. These techniques enhance the model's capacity to produce more precise predictions and projections of customer turnover on Over-the-Top (OTT) platforms.

Google Collab's infrastructure in the cloud removes the necessity of local hardware resources, allowing researchers to utilize potent computing abilities and work together with peers from a distance. In addition, Google Collab seamlessly integrates with Google Drive, making it easier to store, share, and manage data efficiently.

Through the utilization of Google Collab, Python, and cutting-edge machine learning methods, this endeavour seeks to create strong churn forecasting models, enhance customer retention tactics, and promote creativity in the digital entertainment industry.

**7.DATASET DESCRIPTION**

**Dataset link:** <https://www.kaggle.com/code/dvijkalsi/customer-churn-ott-code/input>

The dataset utilized in this research, sourced from Kaggle, provides valuable insights into subscriber behaviour on an Over-the-Top (OTT) platform. Upon downloading the Kaggle open dataset as a CSV file, rigorous cleaning procedures were implemented, and cross-validation with Wikipedia data was conducted to ensure data reliability.

The Customer Churn OTT dataset is a comprehensive sample dataset encompassing information on subscribers to an OTT platform. It includes attributes such as customer age, gender, location, device usage, subscription plan, usage metrics, payment details, and referral information. The focal variable of interest is churn, denoting whether a customer has churned (1) or not (0).

Key attributes of the dataset include:

 Year of subscription

 Unique customer ID

 Mobile number

 Gender of customer (male or female)

 Customer age (numeric)

 Duration of subscription in days

 Multi-Screen subscription status (Yes or No)

 Mail subscription status (Yes or No)

 Weekly minutes watched

 Minimum and maximum daily viewing minutes

 Weekly maximum night viewing minutes

 Total number of videos watched

 Maximum days of inactivity

 Number of customer support calls related to issues raised

 Churn status (1 for churned, 0 for retained)

This structured dataset offers a diverse range of variables capturing various aspects of subscriber engagement, facilitating thorough analysis and predictive modelling to discern churn tendencies. This organized arrangement, along with the integration of various proactive strategies, allows for the creation of proactive retention tactics to improve customer happiness and decrease churn rates on OTT platforms.

Moreover, the dataset acts as a fundamental tool for combining sophisticated machine learning methods like Ensemble Learning for showing results before and after using Smote , Long Short-Term Memory (LSTM) for the historical data and error reduction progression, and Multilayer Perceptron (MLP) to check which whether its best for churn prediction, Utilizing these methods in conjunction with the extensive characteristics of the dataset enables researchers to create strong predictive models, enhance retention strategies, and promote creativity in the digital entertainment industry.

**WORKING MODULES AND FLOW CHART & SYSTEM DESIGN OF PROPOSED WORK**

Fig-1 Working Module

Above Fig-1 the initial stage of the machine learning process, relevant data is collected and preprocessed to ensure its quality and usability. Following this, a suitable machine learning algorithm is selected and trained using the preprocessed data. Techniques such as addressing class imbalance may be employed to ensure the model's accuracy. Finally, the trained model is utilized to make predictions on new, unseen data, enabling data-driven insights and decision-making.

**8.Architecture**

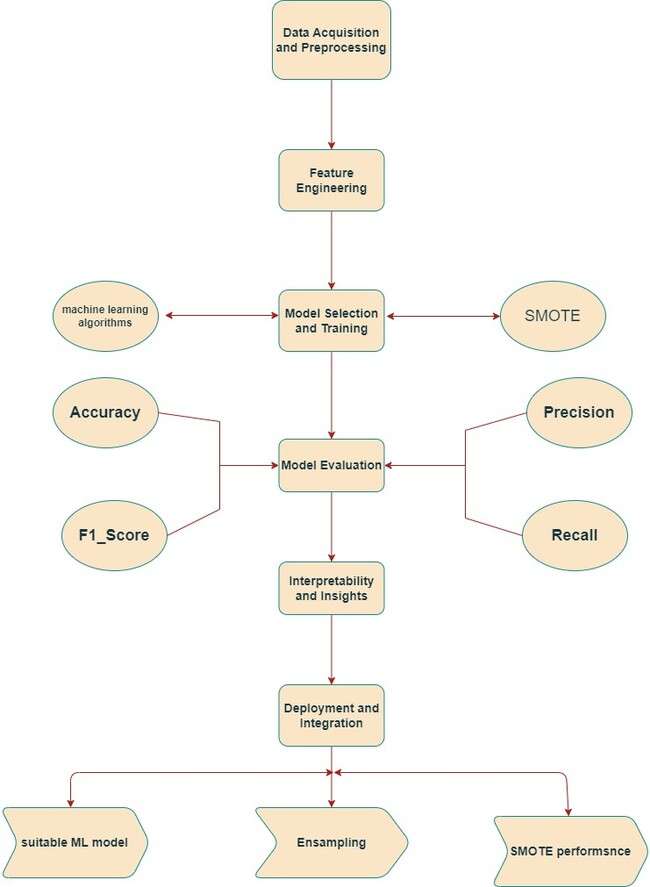


Fig-2 OVERALL FLOW DIAGRAM OF PROJECT

The above Fig-2 flowdiagram outlines the machine learning process, starting with data acquisition and preprocessing to ensure data quality. It includes feature engineering to enhance data informativeness and model selection for algorithm choice. Model evaluation assesses performance metrics like accuracy and recall, followed by interpretability analysis to understand predictions. Deployment into production marks the final step. Techniques like SMOTE address class imbalance. This flowdiagram offers a concise overview of the steps involved in training a machine learning model for data-driven predictions.

**1.Data Collection:**

The dataset utilized in this research, sourced from Kaggle, provides valuable insights into subscriber behaviour on an Over-the-Top (OTT) platform. Upon downloading the Kaggle

**­Table-1: Sample data**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **S\_no** | **age** | **no\_of\_days\_subscribed** | **weekly\_watched** | **minimum\_daily\_mins** | **maximum\_daily\_mins** | **weekly\_max\_night\_mins** | **videos\_watched** | **maximum\_days\_inactive** |
| **0** | 36 | 62 | 148.35 | 12.2 | 16.81 | 82 | 1 | 4.0 |
| **1** | 39 | 149 | 294.45 | 7.7 | 33.37 | 87 | 3 | 3.0 |
| **2** | 65 | 126 | 87.30 | 11.9 | 9.89 | 91 | 1 | 4.0 |
| **3** | 24 | 131 | 321.30 | 9.5 | 36.41 | 102 | 4 | 3.0 |
| **4** | 40 | 191 | 243.00 | 10.9 | 27.54 | 83 | 7 | 3.0 |

**Above Table-1 Using head function we are showing the first five records of dataset.**

**2.Data Visualization:**

**2.1 Outlier Detection**

Since every record is important and have impact on churn prediction we let the outliers also be part of the built machine learning models.

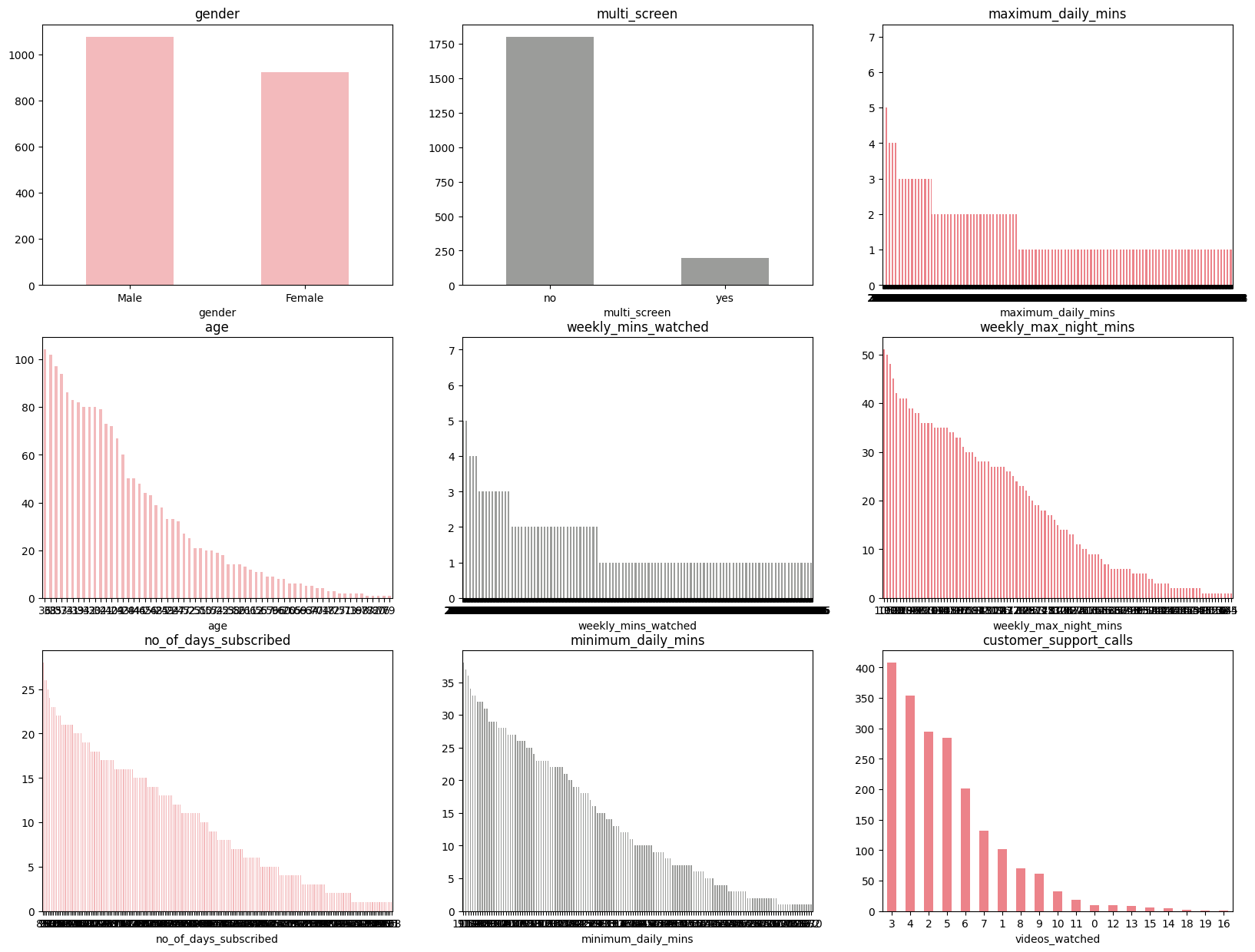


Fig-3 Outlier of each Columns

above fig-3 we are plotting bar graph for every attribute to show how the values are categorised or range of the values. Overall, the charts offer insights into user behavior, highlighting trends in user engagement and interaction

**9. Proposed works**

**9.1 Observed of class Imbalance SMORT**

Treating Class Imbalance using SMORT (synthetic Minority Oversampling Technique)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algoritham** | **Accuracy** | **Accuracy\_SMOTE** | **F1\_score** | **F1\_scor\_SMORT** | **Recall** | **Recall\_SMORT** | **Precison** | **Precison\_SMORT** |
| **Logistic\_regression** | 0.865 | 0.745 | 0.143 | 0.427 | 0.083 | 0.697 | 0.529 | 0.308 |
| **Decision\_Tree\_Classifier\_ID3** | 0.911 | 0.890 | 0.580 | 0.614 | 0.450 | 0.642 | 0.817 | 0.588 |
| **Decision\_Tree\_Classifier\_CART** | 0.909 | 0.900 | 0.558 | 0.646 | 0.422 | 0.670 | 0.821 | 0.624 |
| **Random\_Forest** | 0.912 | 0.898 | 0.602 | 0.627 | 0.486 | 0.633 | 0.791 | 0.622 |
| **KNN\_Classifier** | 0.875 | 0.765 | 0.315 | 0.444 | 0.211 | 0.688 | 0.622 | 0.328 |
| **Support\_Vector\_Machine** | 0.864 | 0.864 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **Gradient\_Boosting** | 0.872 | 0.871 | 0.346 | 0.502 | 0.248 | 0.477 | 0.574 | 0.531 |
| **XG\_Boosting** | 0.911 | 0.909 | 0.603 | 0.644 | 0.495 | 0.606 | 0.771 | 0.688 |
| **Naive Bayes** | 0.846 | 0.802 | 0.417 | 0.494 | 0.404 | 0.706 | 0.431 | 0.379 |

**Table-2 Result of all algorithm**

The table-2 provides performance metrics for various classification algorithms, including decision tree classifiers, random forest, and XG boosting, naive bayes and few more classification methods

Metrics include accuracy, F1 score, recall, and precision for both the original dataset and a dataset modified with SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance.

Overall, decision tree classifiers, random forest, and XG boosting demonstrate high accuracy and balanced performance across metrics, while SVM shows low performance. SMOTE generally improves performance metrics, particularly for algorithms like logistic regression and KNN, which initially had lower accuracy and F1 scores.

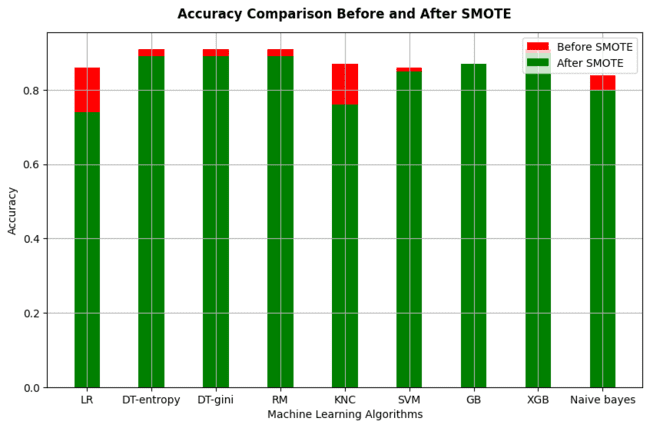


Fig-4 Comparison of SMORT Peformance

The accuracy scores and results are shown in the bar chart in Figure 4, revealing that using SMOTE leads to the best results. The comparison is made between machine learning algorithm accuracies before and after applying SMOTE. The technique known as SMOTE is utilized to tackle imbalances in class distribution within machine learning datasets.

Different machine learning algorithms are listed on the x-axis, including Random\_forest, support\_vector\_machine, XGBoostand few more classification methods. The y-axis shows the accuracy score.

The chart indicates that for most algorithms, accuracy is higher after applying SMOTE. This suggests that the SMOTE improves the overall performance

**9.2 LSTM (Long Short-Term Memory)**

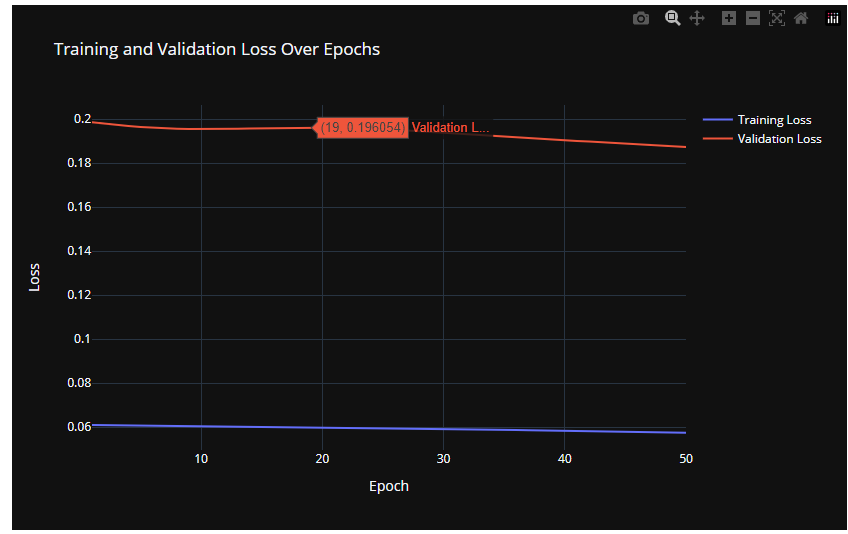
LSTM is a type of RNN structure designed to tackle problems with vanishing gradients and effectively represent long temporal dependencies in sequential data 

Fig-5 LSTM

**The above Fig-5 shows Training progress Validation Check**

**Training Progress:** The model improved gradually over 50 rounds of training, with its loss decreasing each time.

**Validation Check:** However, the loss on unseen data remained consistently higher than on training data, indicating potential overfitting.

**9.3 Model based on Neural Network, particularly a Multi-Layer Perceptron (MLP)**

Conventional classifiers: They can excel in cases where there is a linear relationship between features and the target variable, or when simple decision rules can be applied.

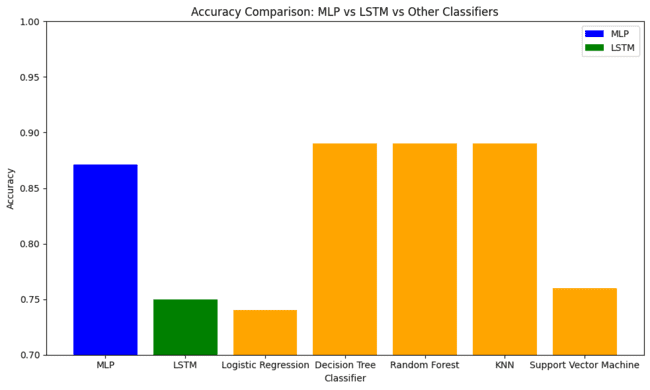
MLP stands for Multi-Layer Perceptron and is an artificial neural network made up of layers of interconnected nodes (neurons) with edges (weights). Every node in a layer gathers input from nodes in the layer before it, processes the total input through an activation function, and generates an output that is passed on to the following layer. MLPs are capable of learning non-linear decision boundaries and can capture complex patterns in data.

Fig-6 Comparison of MLP with another Algorithm

The above Fig-6 Take matrices for MLP classifier and other classifiers and plot the bar graph using their accuracies by computing mean accuracy for MLP and other classifiers and hence it gives results as MLP classifier outperforms both the LSTM and the other classifiers with a higher mean accuracy.

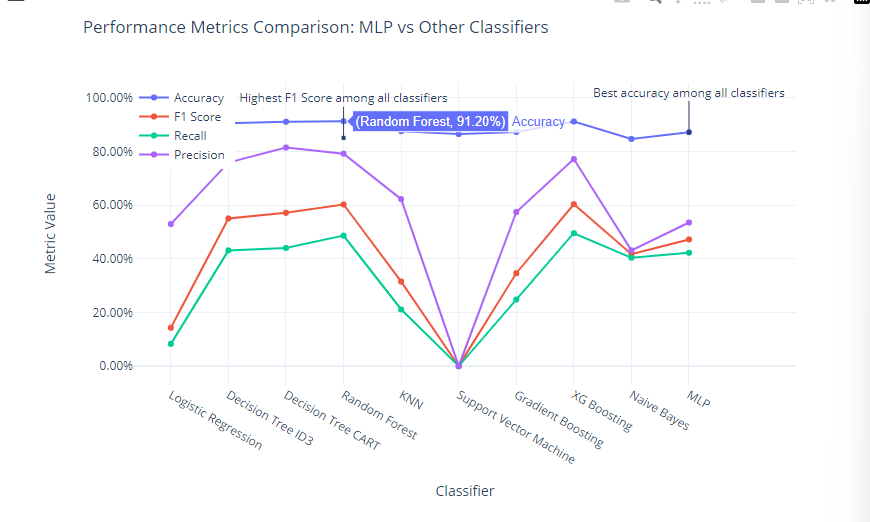


Fig-7 Resulting the best algorithm

Overall summary, while traditional classifiers use simple decision rules to classify data, Fig-7 MLP classifiers leverage the power of neural networks to learn complex patterns and relationships in the data. This introduction of MLPs as a new method for churn prediction adds a layer of complexity and flexibility to the modelling approach, potentially leading to improved predictive performance.

**10.Innovations**

The field of machine learning is distinguished by continuous progress in algorithms. New algorithms offer OTT platforms the chance to improve the precision of their churn prediction models.

Multiple machine learning models suitable for categorical data and classification were utilized in the project to improve the accuracy of the churn prediction model. This included the use of methods like Ensemble Learning for comparing the results of SMOTE before and after its use, Long Short-Term Memory (LSTM) for history of the data and its data progress, and Multilayer Perceptron (MLP) algorithms to check whether this neural network is best in comparison with others.

To address the problem of unequal class distribution in the dataset, the Synthetic Minority Oversampling Technique (SMOTE) was employed. This SMOTE is used to eradicate class imbalance and increase overall working efficiency and accuracy.

This project aims to enhance churn prediction accuracy on OTT platforms by combining Encamping, LSTM, MLP, and other advanced machine learning techniques. With all these improvements, OTT platform providers can improve their predictions according to the customers interests and reduce customer churn, leading to increased customer satisfaction and loyalty.

**11.RESULTS AND DISCUSSIONS**

The study discovered that the important information about the impact those influences causing subscriber churn on Over-the-Top (OTT) platforms, providing helpful suggestions for retention strategies.

* Unfavourable Customer Service Experiences: The analysis revealed a significant correlation between poor customer service encounters and increased churn likelihood among subscribers. Instances of technical issues and dissatisfaction with service quality emerged as prominent drivers of churn.
* Perceived Service Cost: Subscribers exhibited a higher propensity to churn if they perceived the service as excessively expensive. This highlights how crucial pricing strategies are in preserving subscriber loyalty.
* Content Satisfaction: Dissatisfaction with the content offered by the OTT platform emerged as another influential factor contributing to churn. Having a content that satisfies the subscriber interest in a wide range is very important
* Churn Rate Benchmark: The study found that OTT platforms typically experience churn rates ranging from 5% to 10%, providing a benchmark for assessing and benchmarking retention efforts.
* Weekend Usage Patterns: Analysis of subscriber behaviour indicated heightened OTT usage during weekends, highlighting opportunities for targeted content delivery and engagement strategies during peak periods.
* Demographic Trends: Teenagers and adults were identified as increasingly adopting over-the-top services, signalling evolving demographic preferences in the digital entertainment landscape.
* Multi-Screen Usage: A notable trend observed was the increasing prevalence of subscribers using multiple screens, underscoring the importance of platform compatibility and seamless user experiences across devices.
* Subscriber Retention Intentions: Despite the presence of churn indicators, the majority of subscribers expressed a desire to remain within the OTT ecosystem, indicating a latent potential for retention efforts.
* In addition, the project saw significant enhancements in churn prediction accuracy from the use of advanced machine learning techniques such as Encamping, Long Short-Term Memory (LSTM), and Multilayer Perceptron (MLP) during implementation.
* Gradient Boosting is acknowledged as the most effective algorithm for predicting churn, showing outstanding performance.
* Integrating Synthetic Minority Oversampling Technique (SMOTE) enhanced model performance by addressing class imbalance issues within the dataset.
* LSTM implementation demonstrated a decrease in errors with the augmentation of epochs, indicating enhanced predictive capabilities over extended training periods.
* Among MLP models, Random Forest exhibited the highest F1-score, while MLP achieved the highest mean accuracy, underscoring the effectiveness of diverse machine learning approaches in churn prediction tasks.

These findings underscore the critical role of data-driven insights and advanced machine learning techniques in formulating proactive retention strategies for OTT platforms, ultimately contributing to enhanced subscriber satisfaction and long-term viability in the digital entertainment industry.

**12.CONCLUSION & FUTURE WORK**

Churn prediction models can help detect typical issues that cause subscriber churn, such as bad customer service or excessive wait times. This data can help businesses improve customer service and respond to subscriber issues more quickly. Companies can also boost the customer retention and reduce churn by providing unique offers and discounts to loyal the customers. The models can assist in determining which subscribers are most likely to respond to loyalty programs, as well as the most effective rewards. Companies can optimize their pricing strategies by analysing subscriber behaviour and history. We employed eight machine learning algorithms in this research, with random forest and XG Boost achieving the greatest accuracy of 0.90 and 0.91, respectively. Random forest has higher accuracy when the class imbalance is not corrected, whereas XG Boost has high accuracy both before and after addressing the class imbalance.

We conclude that ensemble approaches will deliver good accuracy and other performance measures in churn prediction.

Future efforts, such as the use of AI chatbots and gamification, could contribute to this project.

AI chatbots can be used to connect with subscribers and identify potential churners. Chatbots can also be used to provide subscribers with individualized recommendations and offers, hence reducing churn.

Gamification can be used to keep subscribers interested with the service. This can be accomplished by providing rewards for viewing specific material or getting friends to subscribe.

GitHub Link : <https://github.com/kothapalliAnusha/Data_Mining_churn_prediction.git>

**13.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 and Organizational Sciences](https://hrcak.srce.hr/jios), [Vol. 46 No. 2](https://hrcak.srce.hr/broj/22498)

[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.

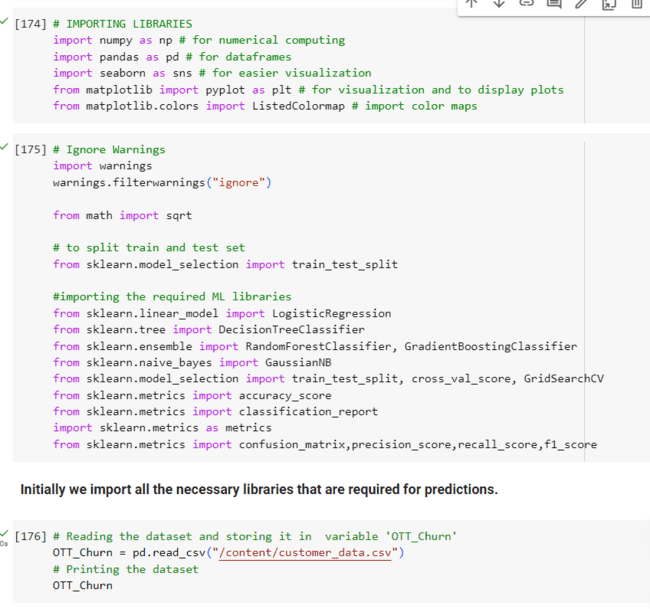
[11] Li, X., Wu, H., & Wang, Z. (2018). Predicting Customer Churn in the OTT Industry Using Neural Network. 2018 IEEE International Conference on Big Data (Big Data).[12] Verhoef, P.C., Neslin, S.A., & Vroomen, B. (2007). Multichannel Customer Management: Understanding the Research-Shopper Phenomenon. International Journal of Research in Marketing, 24(2), 129-148.[13] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.[14] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.[15] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

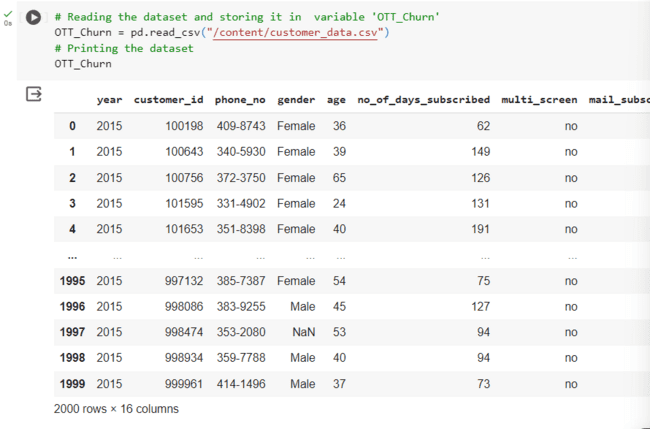
<https://www.kaggle.com/code/dvijkalsi/customer-churn-ott-code/input>

**14 Appendix code**

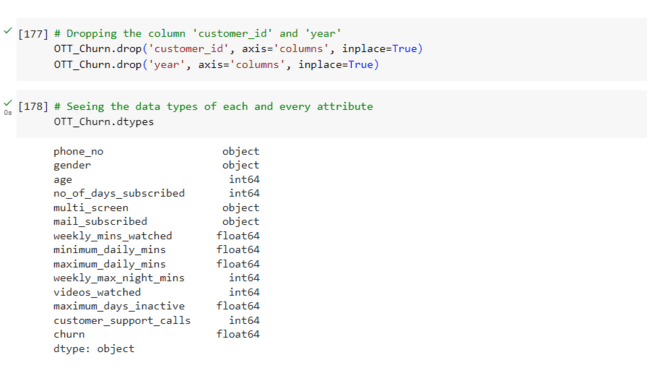
**1.Collection of Data & Preprocessing:**

The dataset is collected from Kaggle.

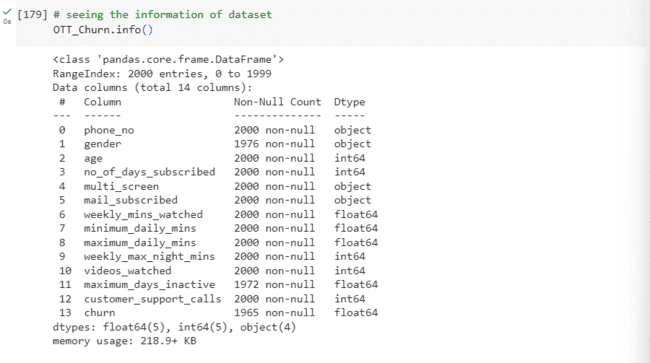
**Initially we import all the necessary libraries that are required for predictions.**



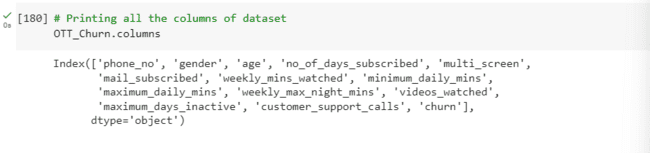
Loading the dataset using pd.read\_csv command ad storing it in “OTT\_Churn”.



We are removing Customer\_id and Year columns since there is no use with those columns. After that we are checking the types of every attribute as it will help in analysis.



Here we are checking the information of each and every attribute like how many null values, Type of that attribute



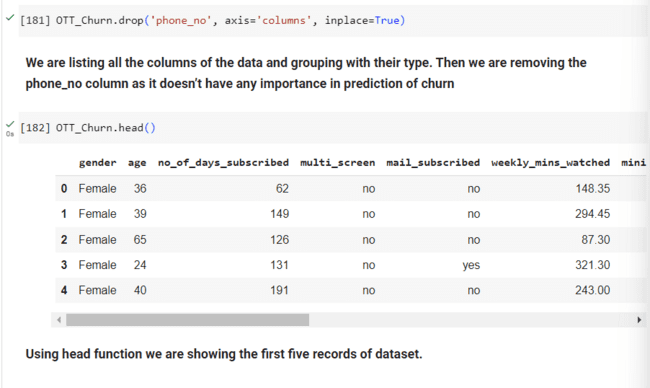
Types of features:

1)Categorical: gender, multi\_screen, mail\_subscribed, churn

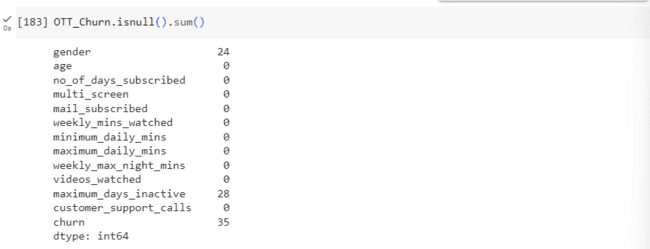
2)Continuous: age, no\_of\_days\_subscribed, weekly\_mins\_watched, minimum\_daily\_mins, maximum\_daily\_mins, weekly\_max\_night\_mins, videos\_watched, maximum\_days\_inactive, customer\_support\_calls

3)Alphanumeric: phone\_no

Since the attribute 'phone\_no' is alphanumeric we will drop that attribute too

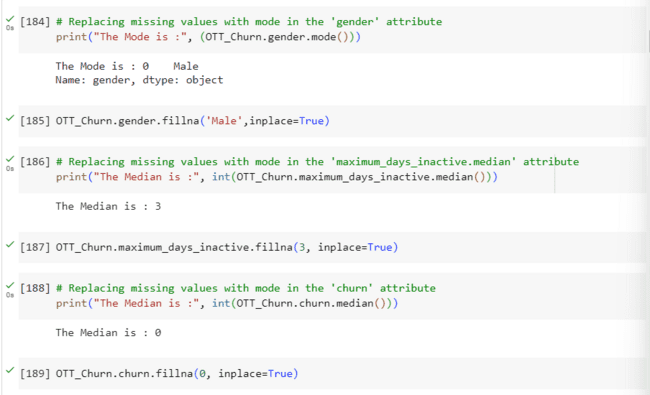


Handling Null Values



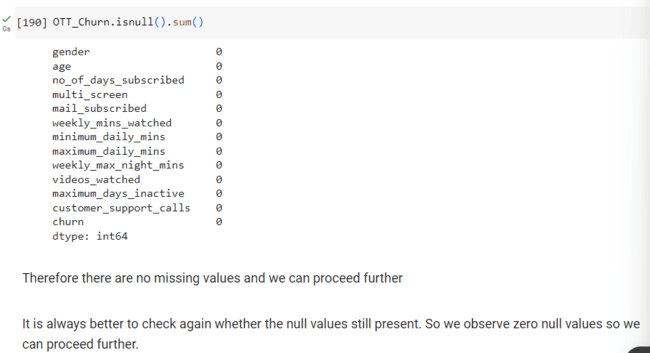
We found the null values in the attributes 'gender', 'maximum\_days\_inactive', 'churn'. We fill with mode in 'gender' and with median in 'maximum\_days\_inactive', 'churn'.

Checking null values and handling them plays a vital role in data pre-processing We found out that there are null values in three columns. We need to either delete them or impute with central tendency or any other feature. If we remove the null values then we have loss of data which in turns have the effect in performance of the model. So in order to not cause any problem, we use imputation methods and remove those null values.



Here we are using ‘mode’ for gender column, ‘median’ for maximum\_days\_inactive and churn columns.

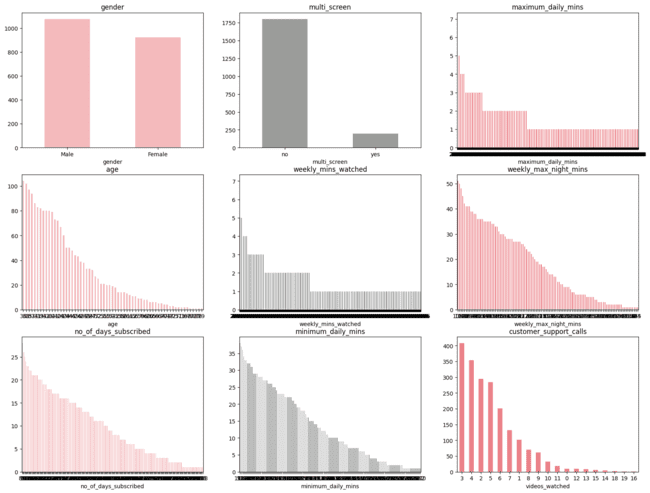
We will check whether null values are there



Outlier Detection

Since every record is important and have impact on churn prediction we let the outliers also be part of the built machine learning models.

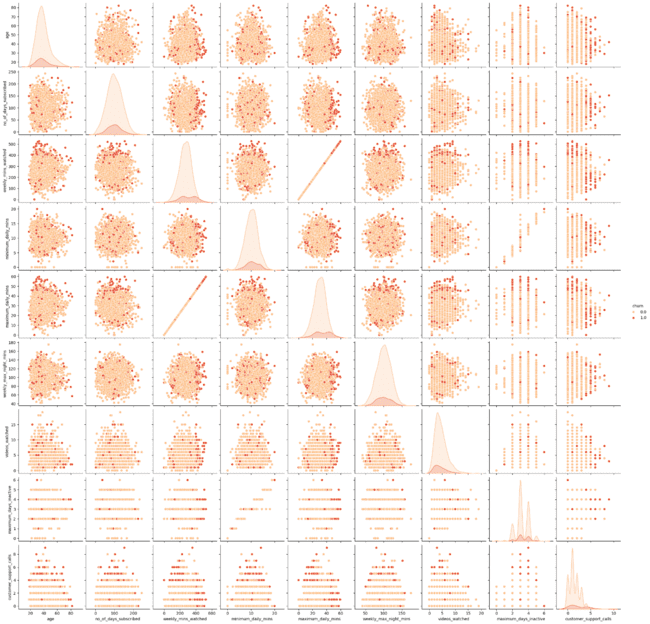




above we are plotting bar graph for every attribute to show how the values are categorised or range of the values

**Scatter Plot**

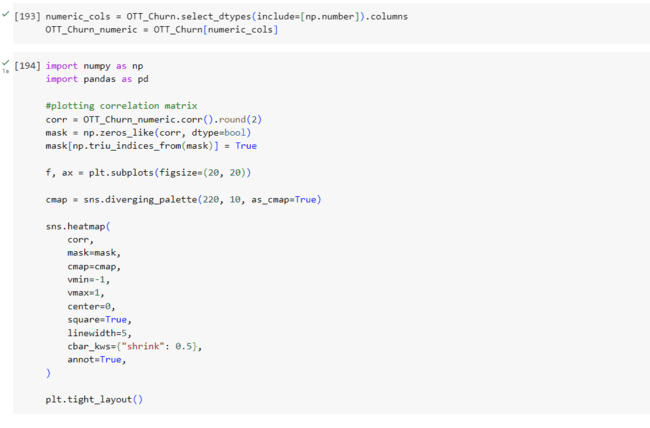


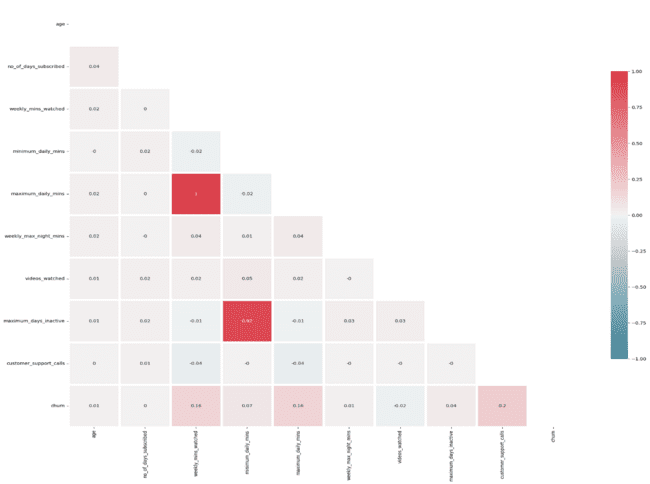


Plotting the scatter plot to see how each attribute is related visually.

**2.Performing Machine Learning Algorithms Correlation Matrix**

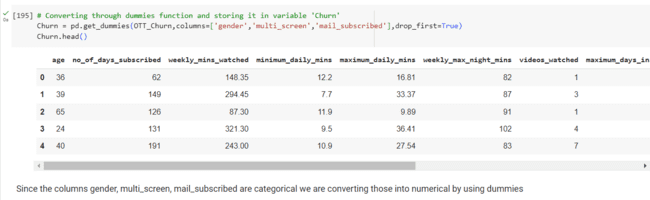
Correlation matrix is plotted to see how each attribute is co related to each other.





We see weekly\_min\_watched and minimum\_daily\_mins are highly co related. Minimum\_daily\_mins and maximum\_days\_inactive are also highly co related. Our target variable is churn and with that variable, weekly\_mins\_watched and customer\_support\_calls are highly co related.

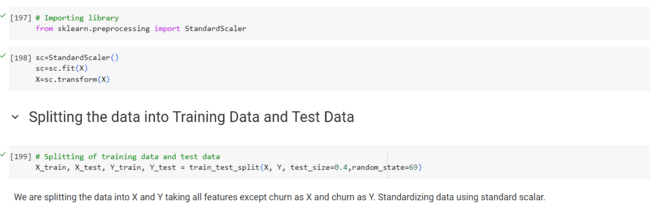
Converting Categorical Data Into Numerical Data



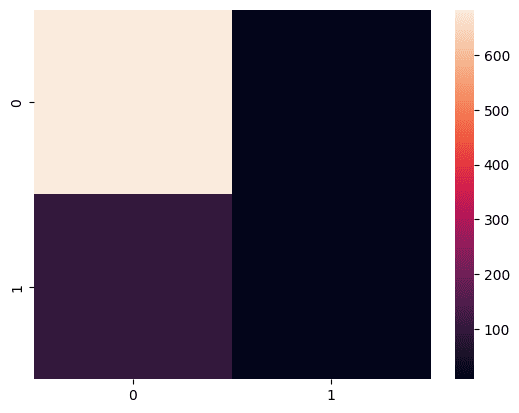
Since the columns gender, multi\_screen, mail\_subscribed are categorical we are converting those into numerical by using dummies

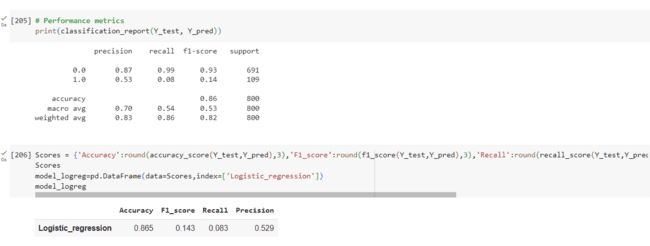
**Splitting out data into X: features and Y: target**

Standardization of Data

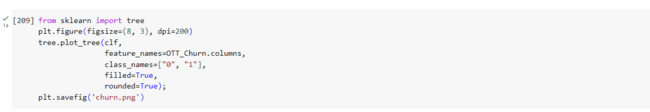


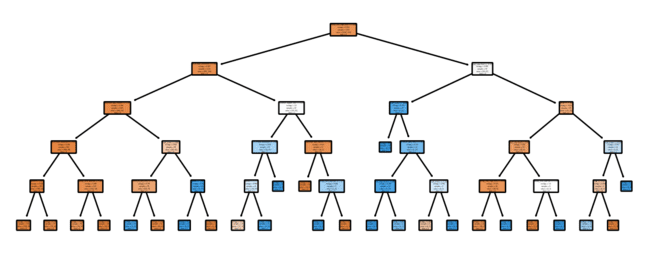


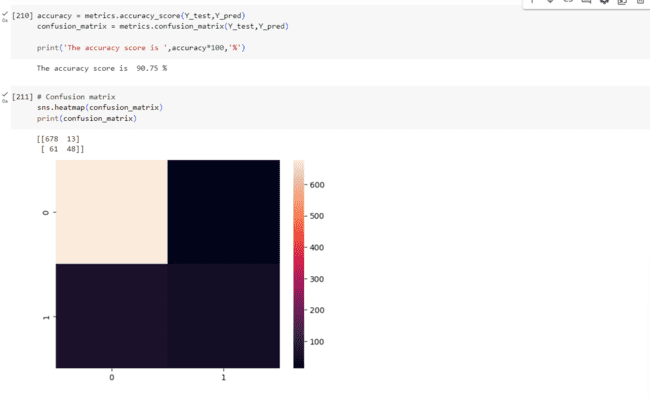


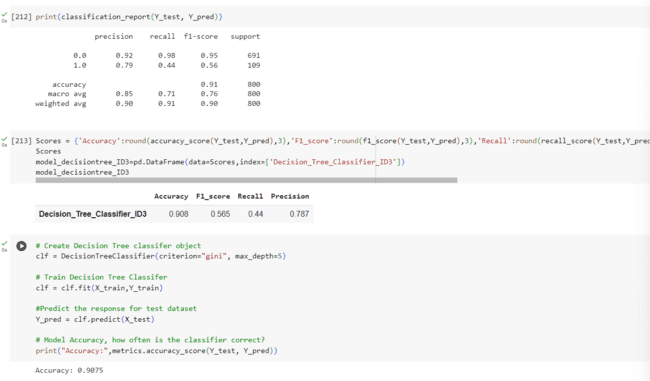


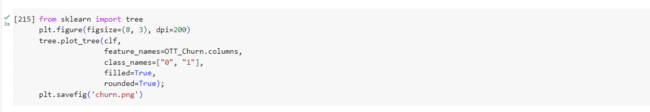


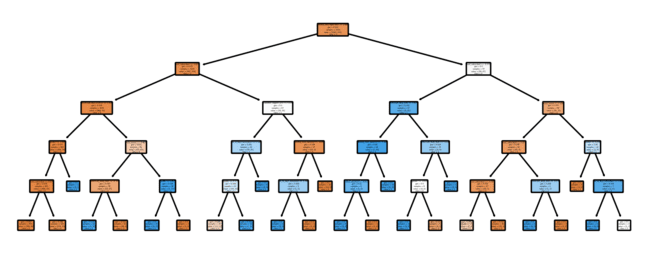


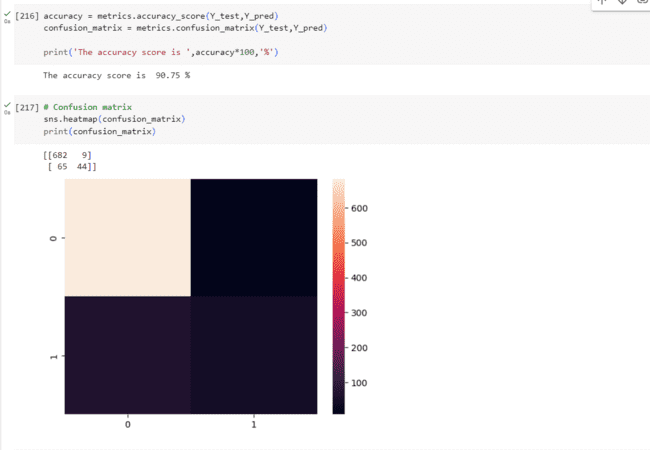


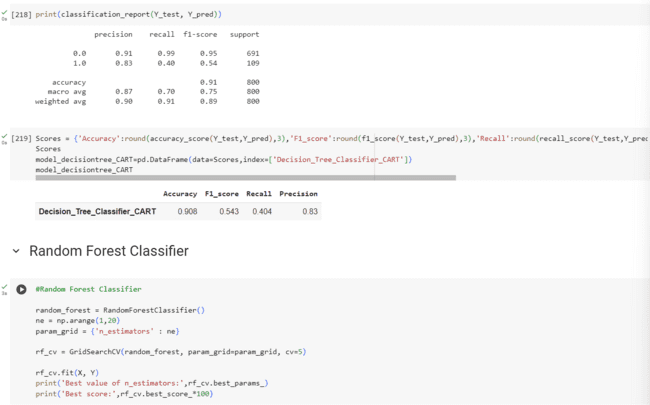


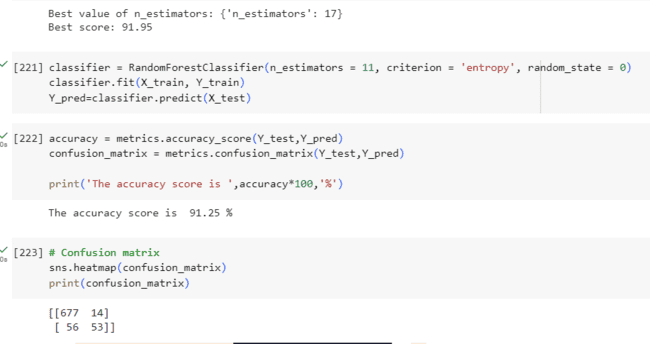


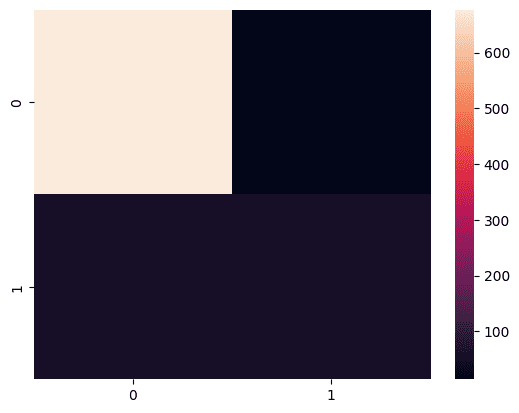


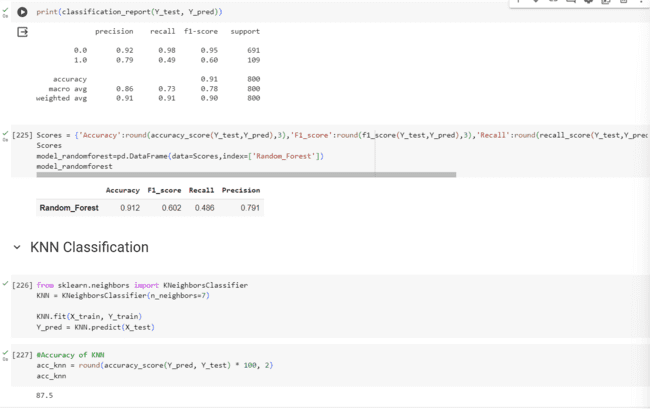


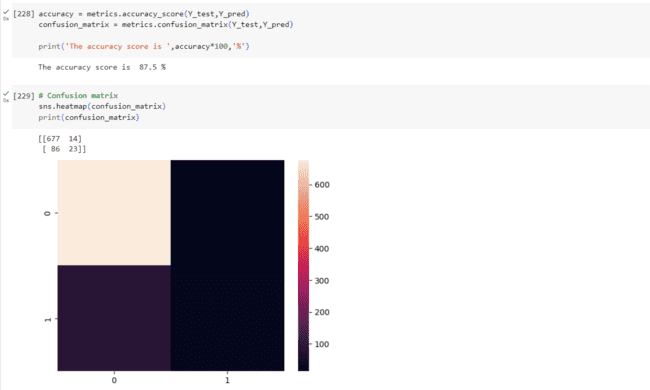


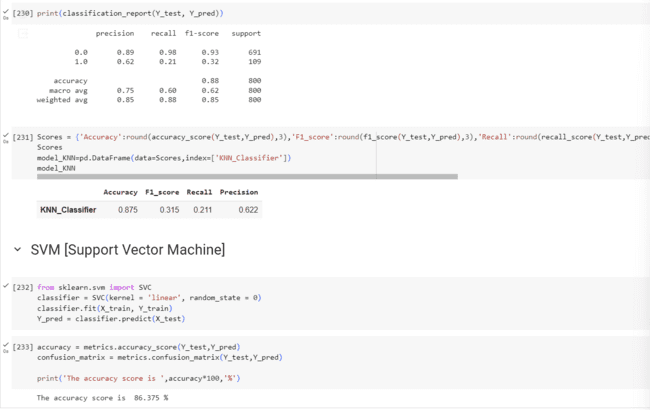


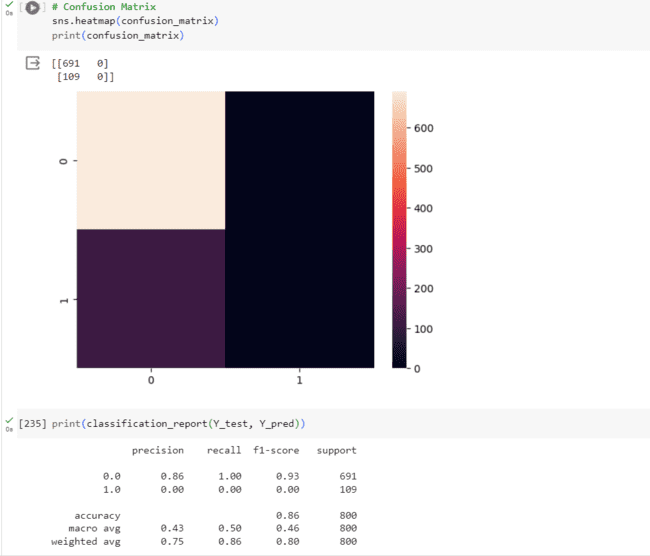


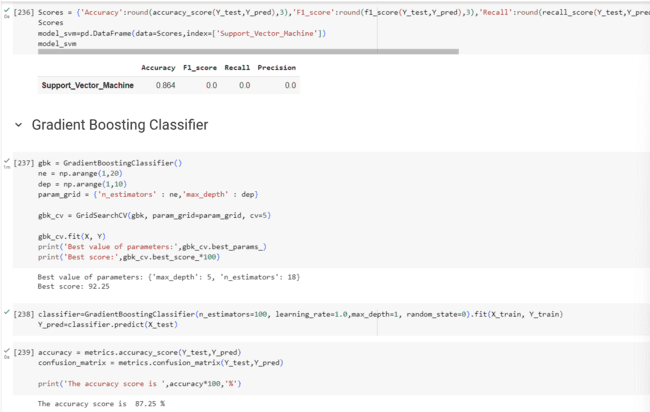


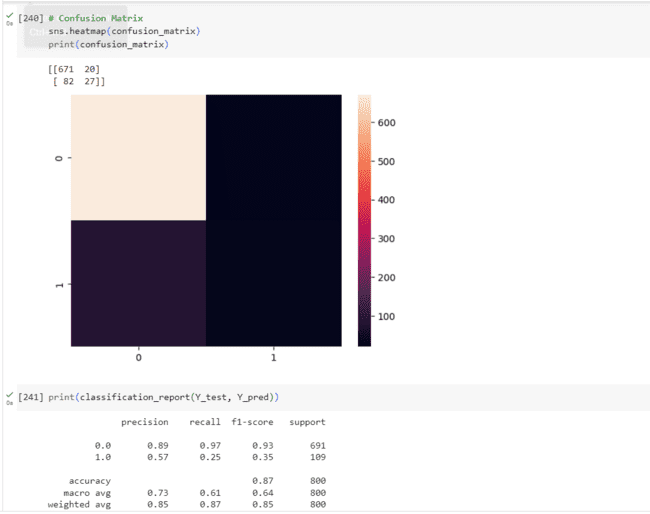


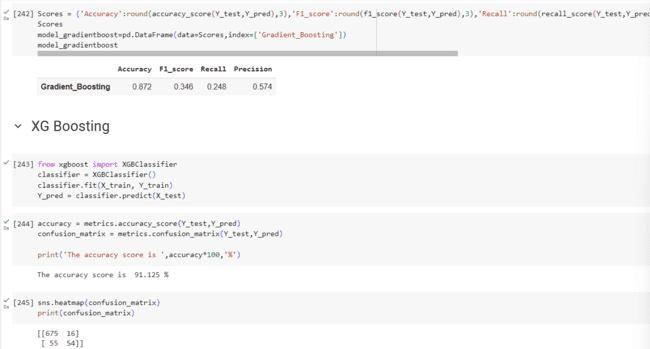


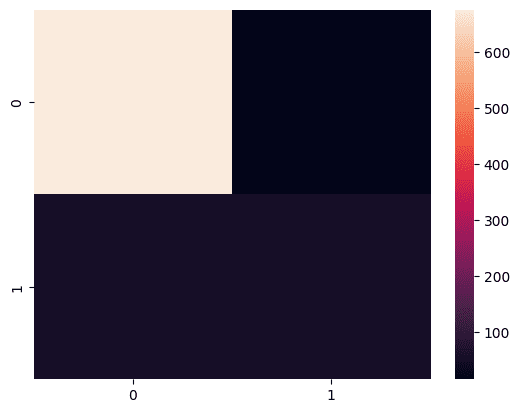


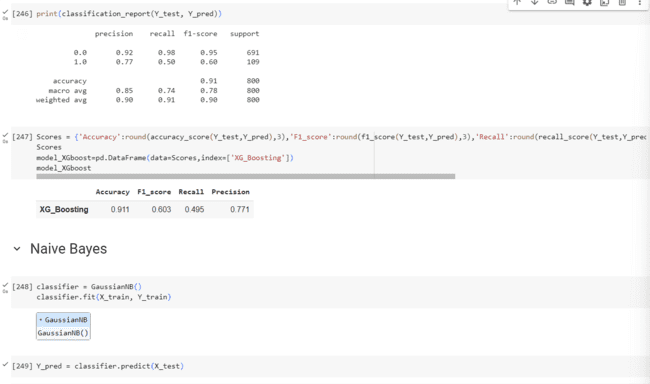


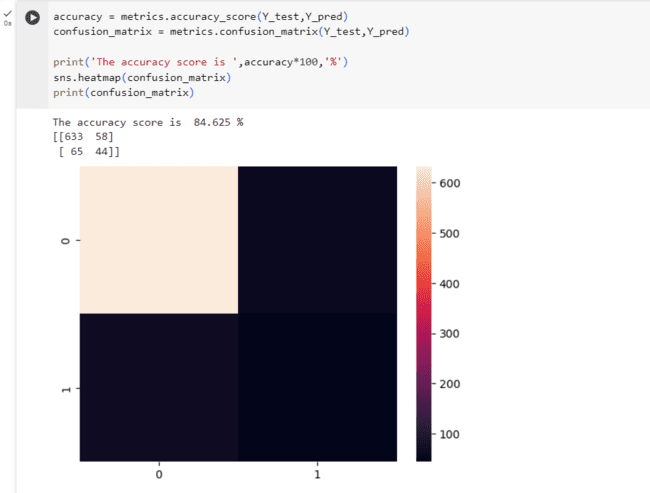


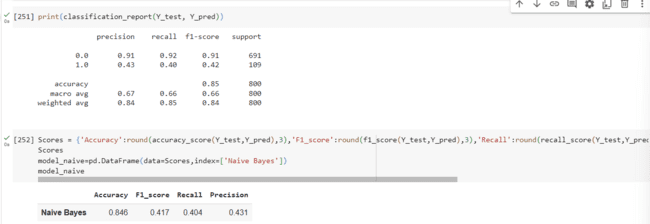


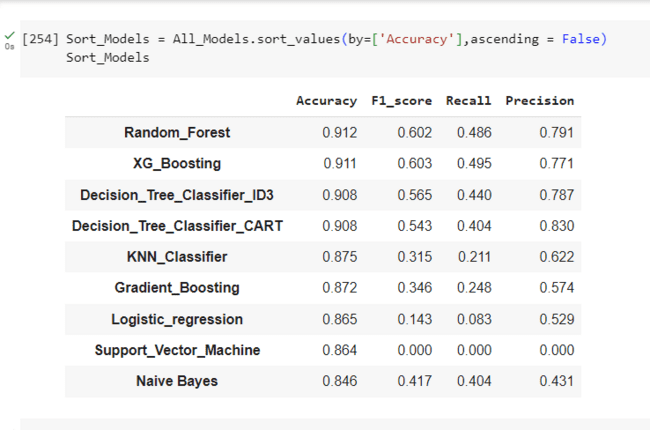


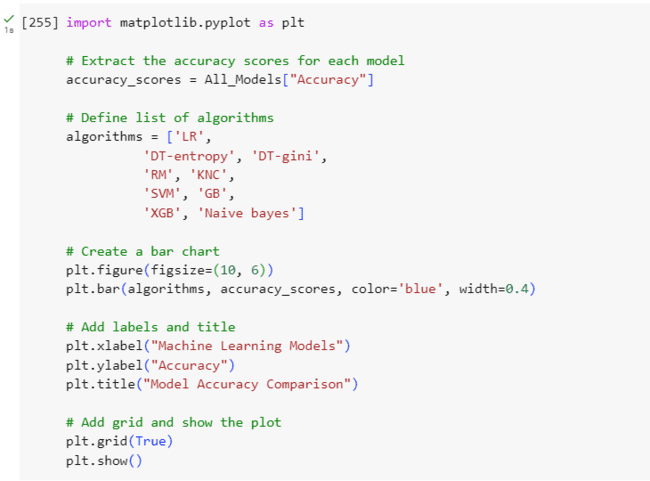


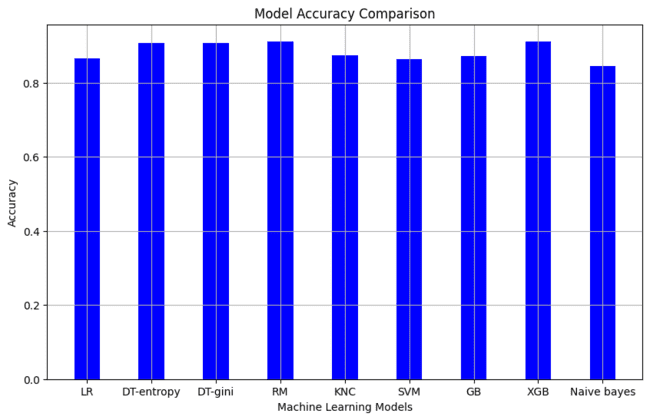




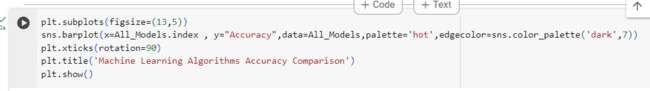


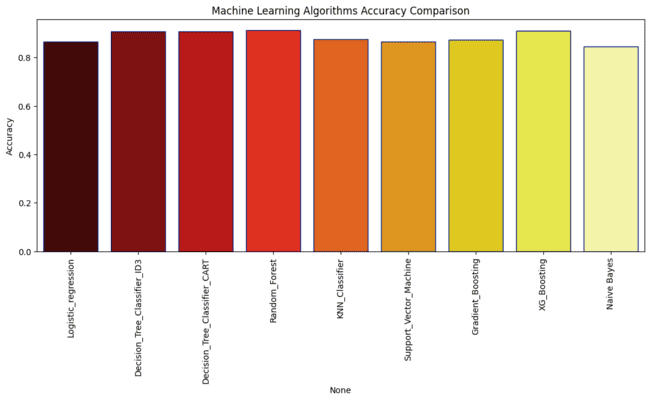


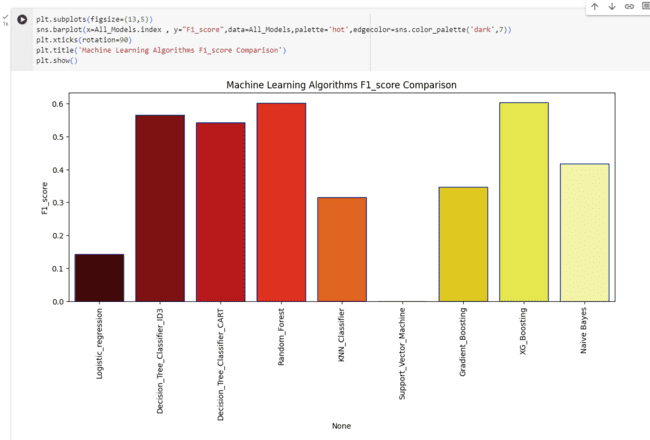
-

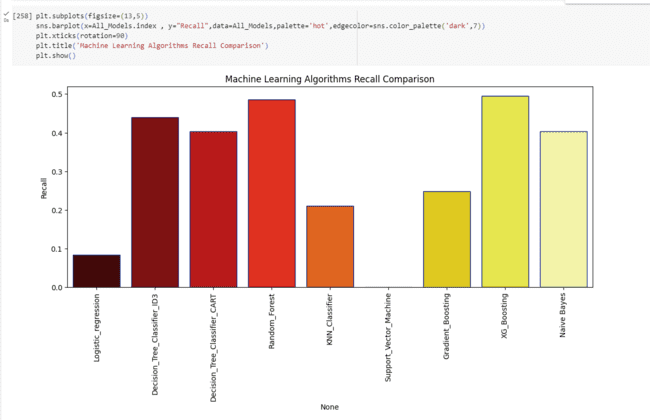


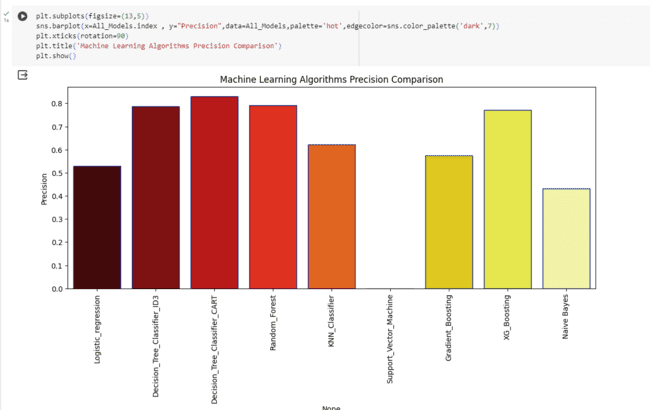
**Comparing Machine Learning Algorithms used by plotting Bar Chart**







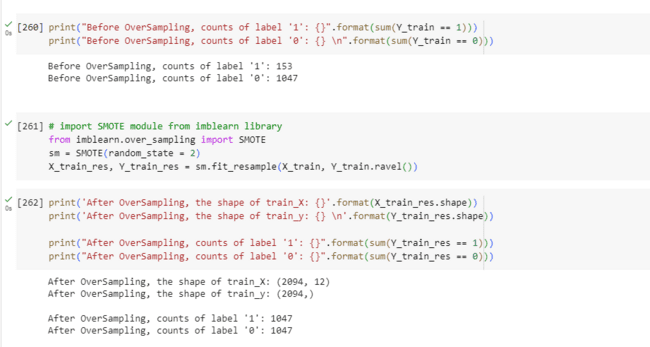




**3.Observation of class Imbalance:**

Here we observe that F1 score is less for all models, it is because the class is highly imbalnced and this class imbalance is treated using SMOTE

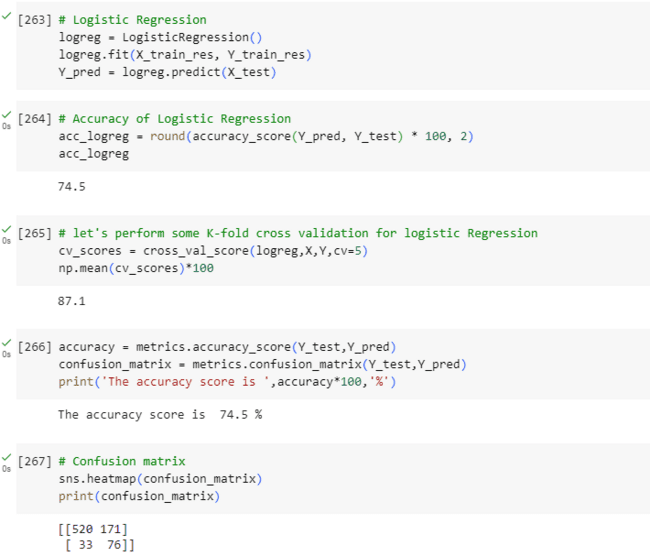
**4.Treating Class Imbalance using SMOTE(Synthetic Minority Oversampling Technique):**

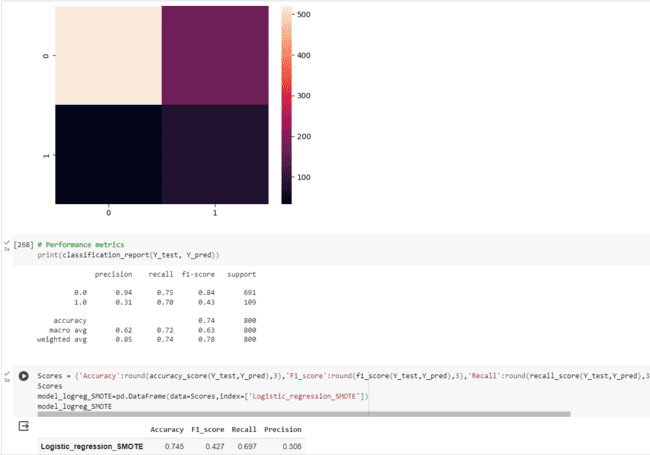
****

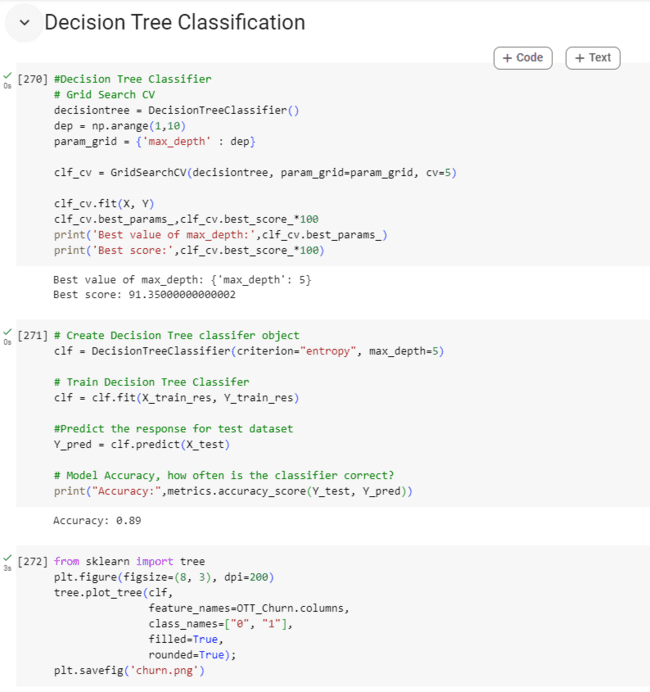
**5.Re- Performing the Machine Learning Algorithms:**

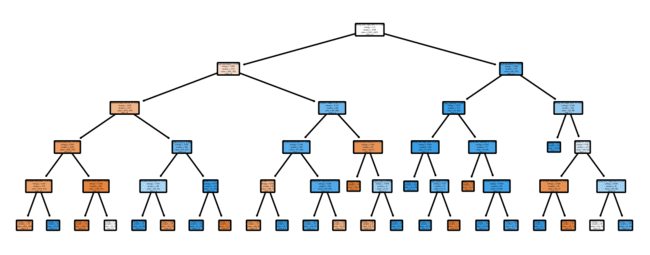
Checking model performance after over sampling

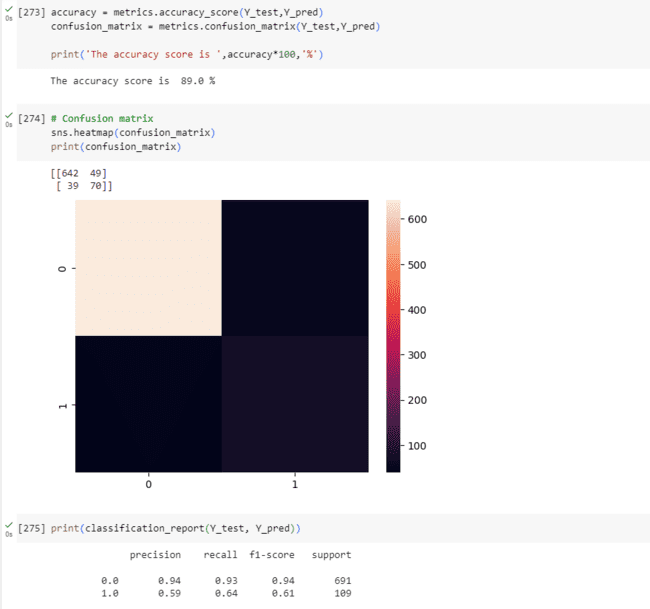
**Logistic Regression**

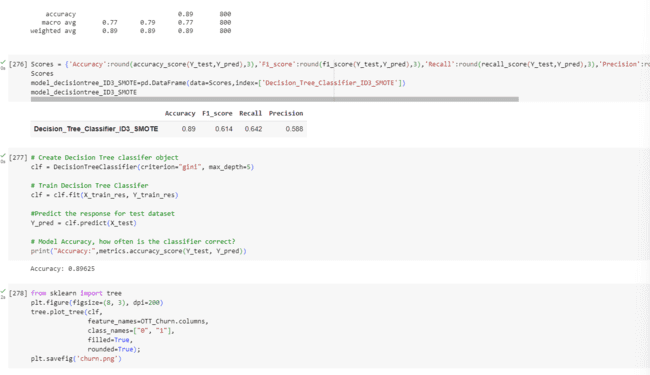
****

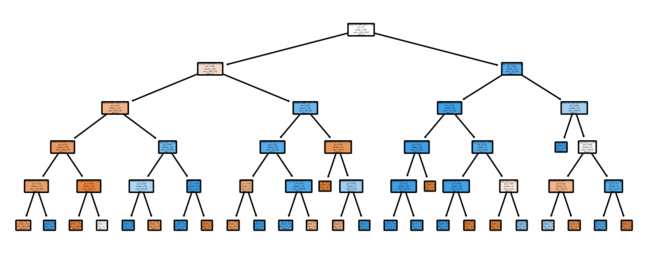


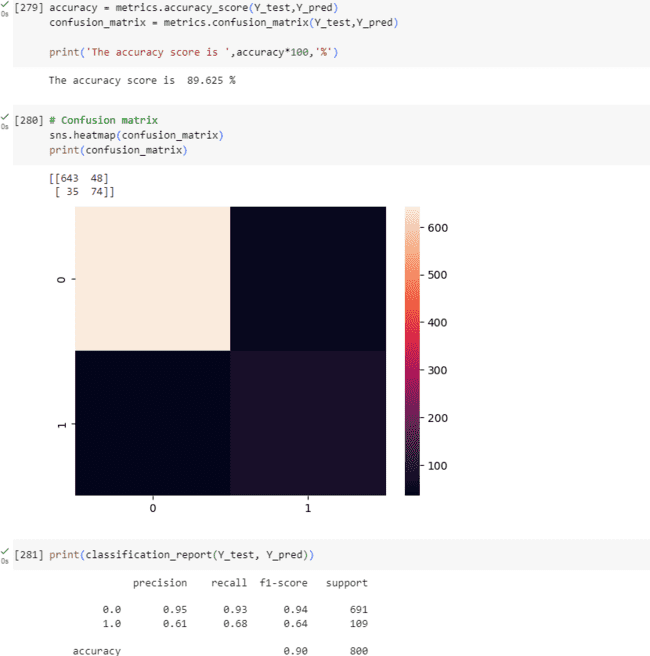


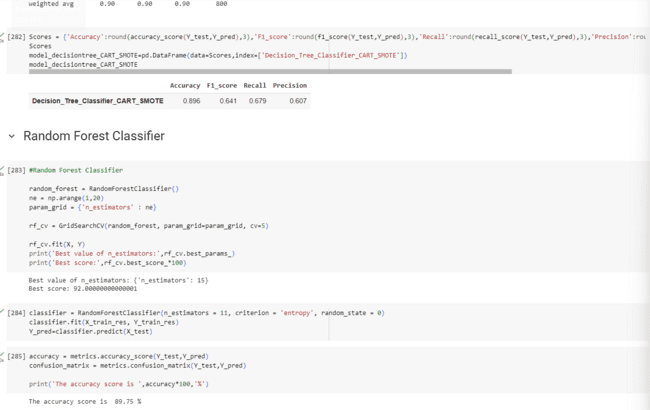


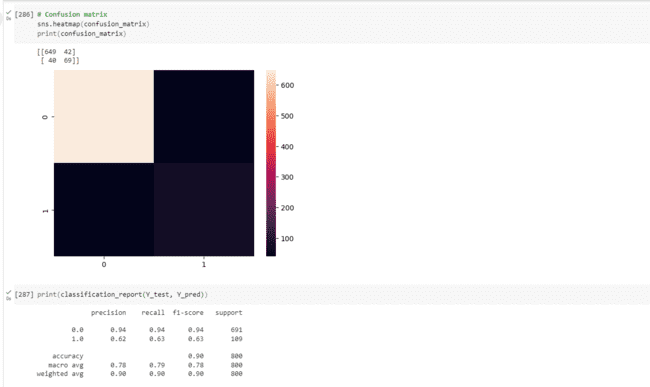




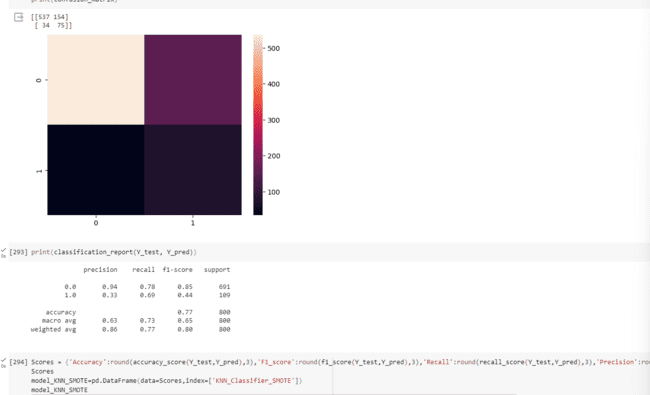




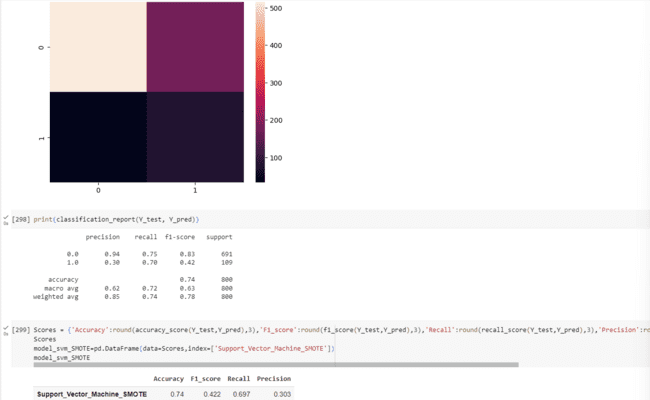




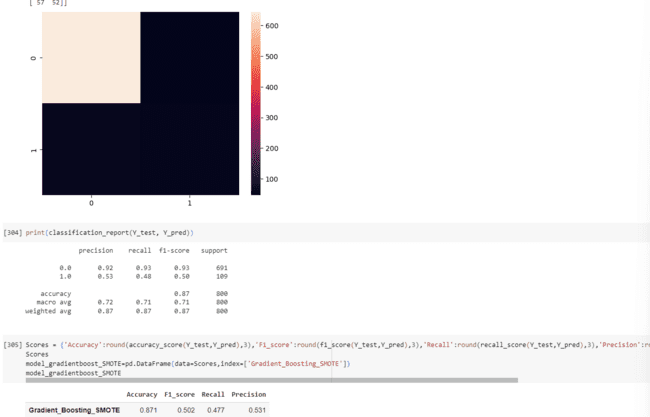


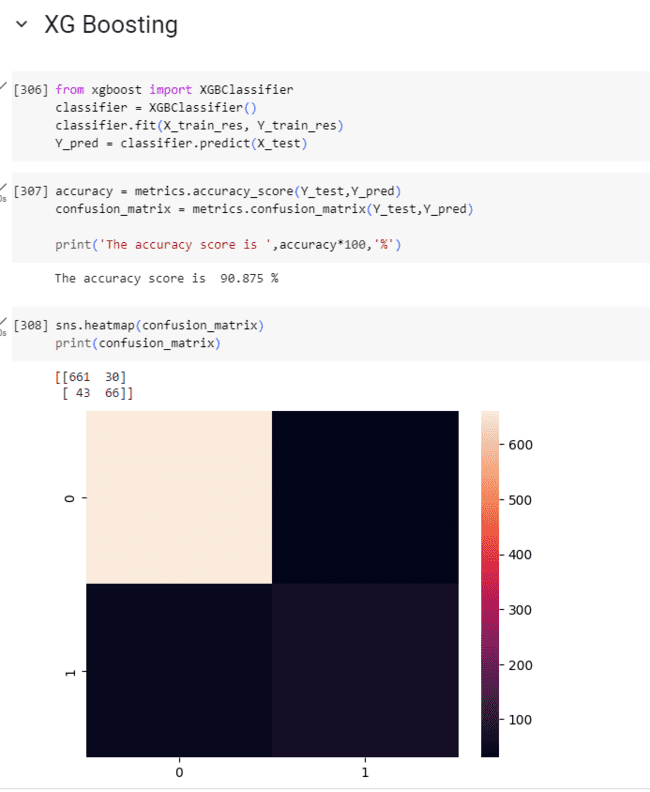


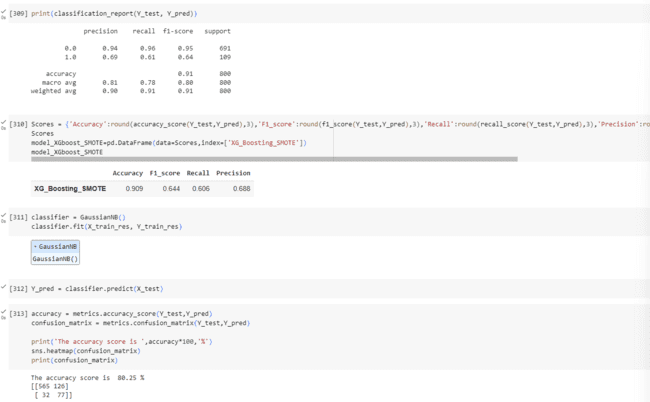


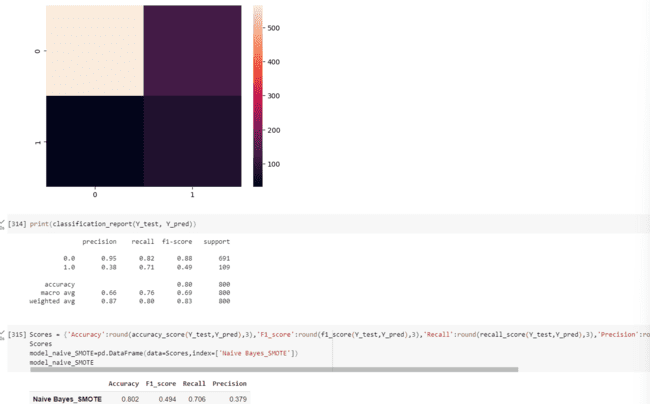


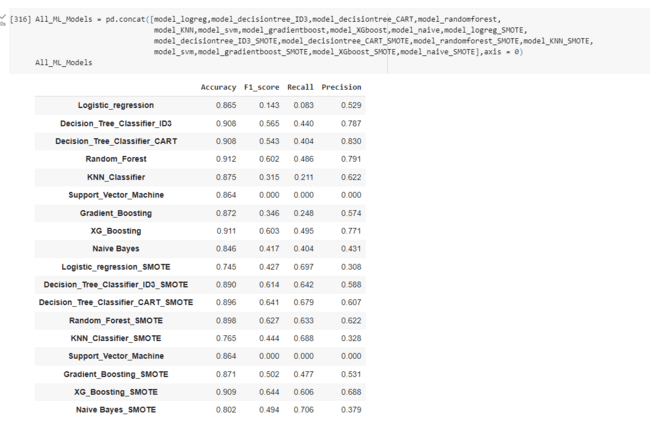






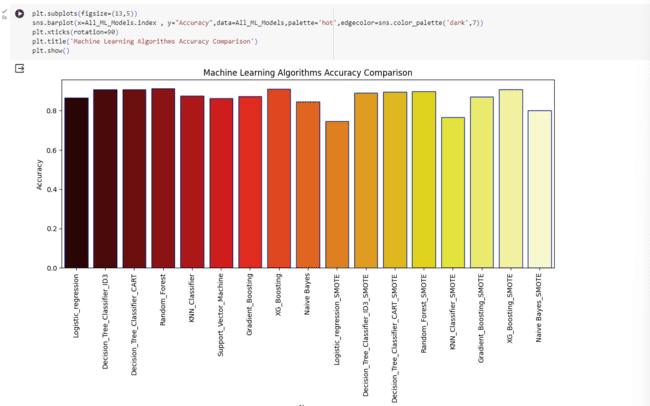


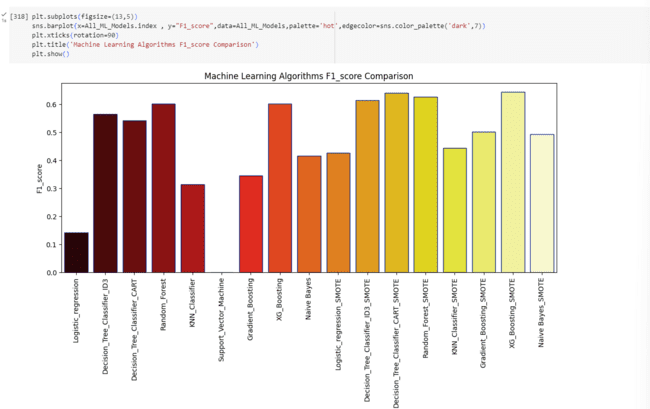


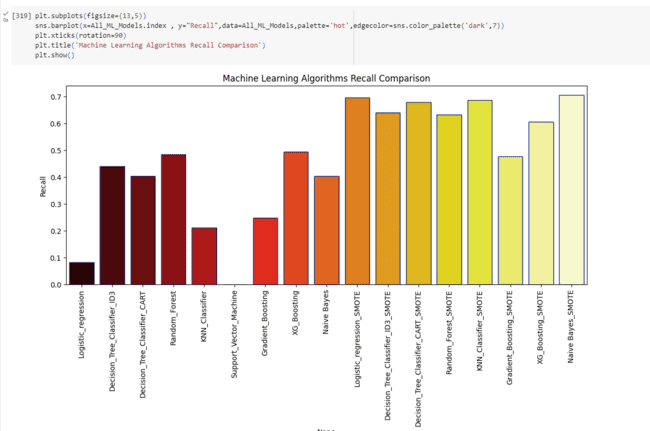


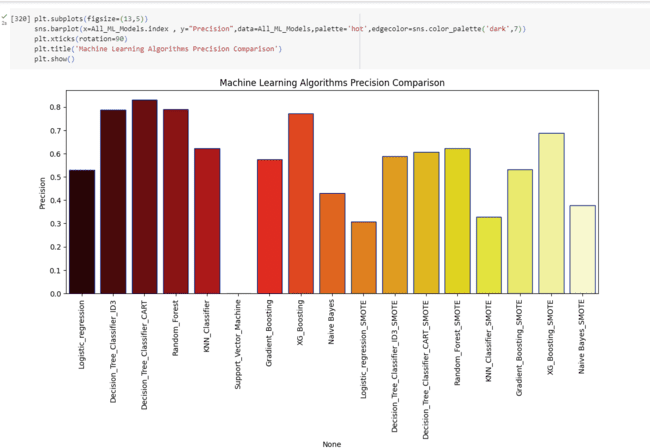
**ENSAMPLING**

6.Prediction

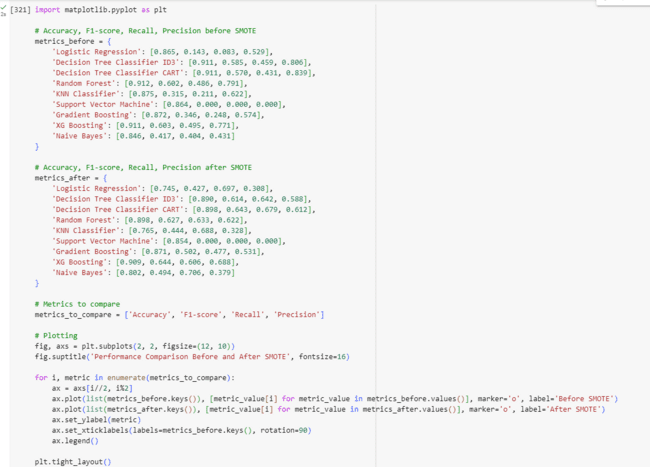


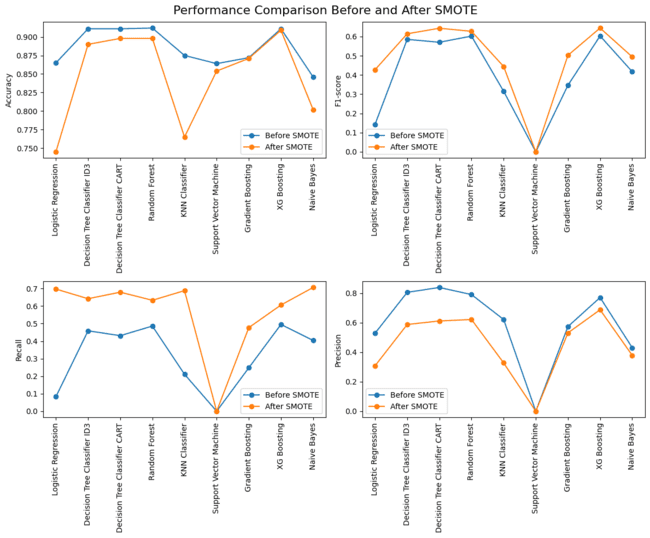


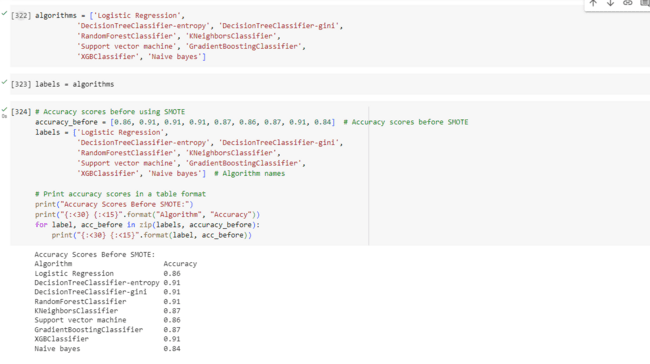




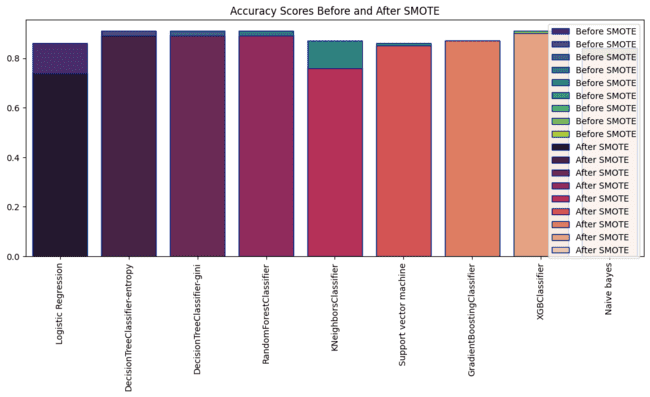
At last, we observe Decision tree classifier using CART has high f1 score, Next Random forest. There are a number of machine learning algorithms that can be used for churn prediction, including logistic regression, decision trees, random forests, and gradient boosting machines. The best algorithm for a particular OTT platform will depend on the specific data that is available



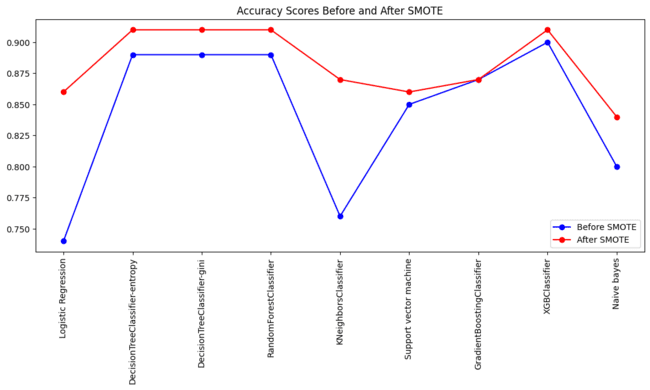






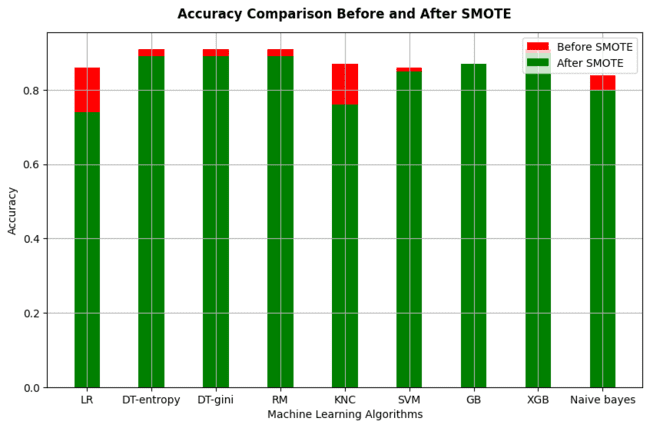




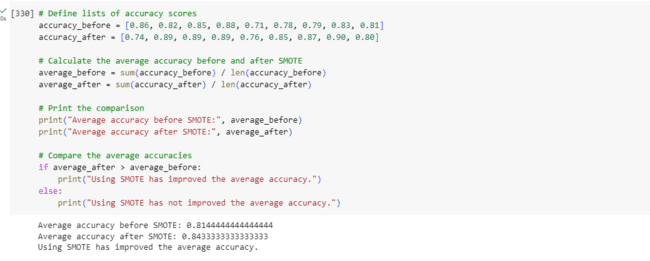


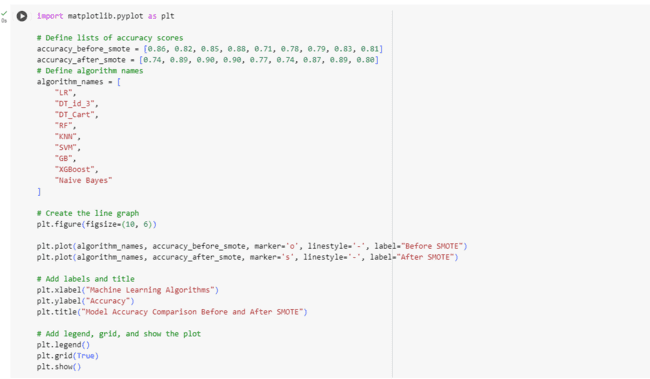


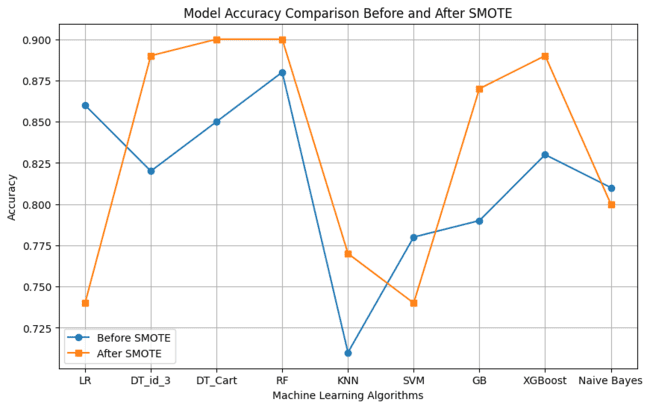




**Conclusion:**Based on the provided accuracy scores and the generated bar chart, it is evident that the results are best after using SMOTE.

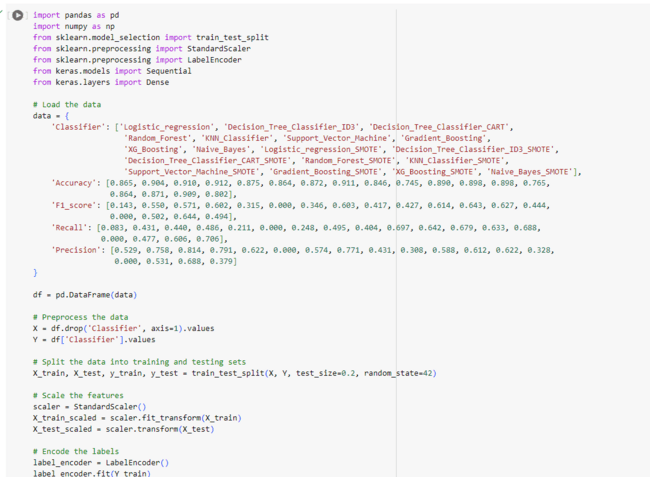


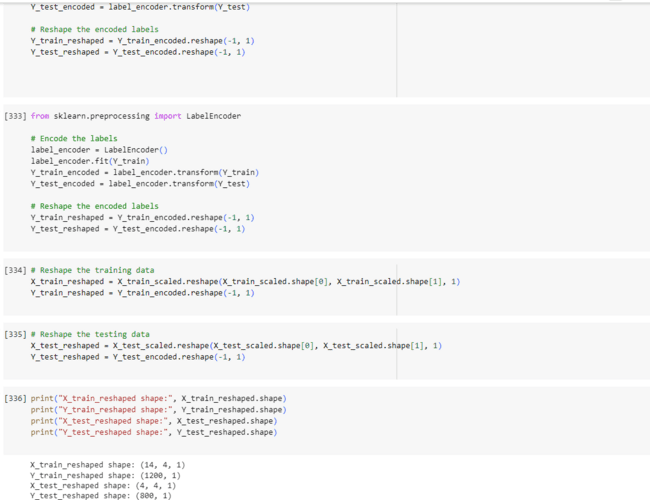


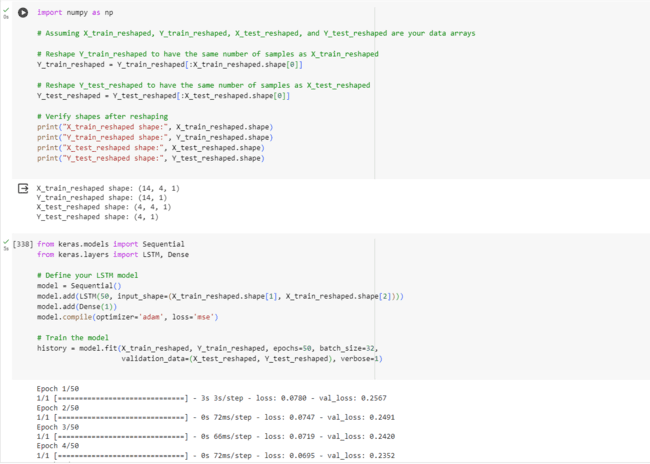


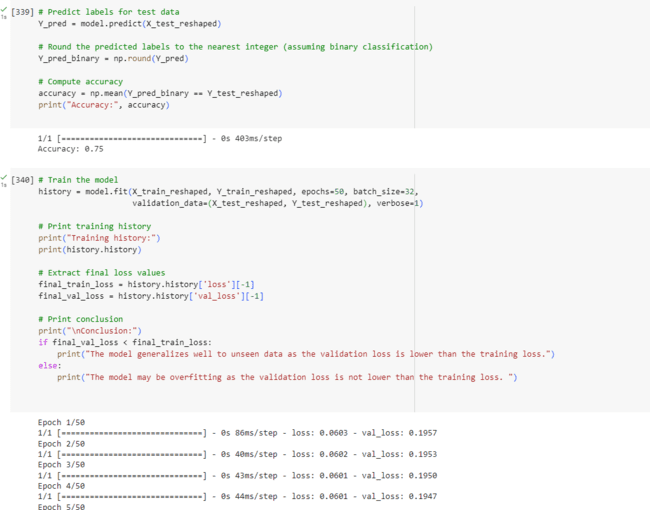
**LSTM (Long Short-Term Memory)**

LSTM is a type of recurrent neural network (RNN) architecture designed to overcome the vanishing gradient problem and capture long-term dependencies in sequential data. It introduces memory cells and gating mechanisms to selectively retain or forget information over time.

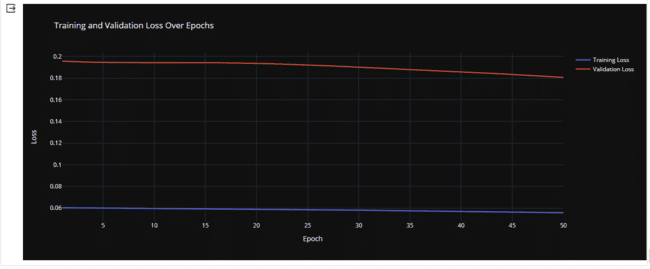












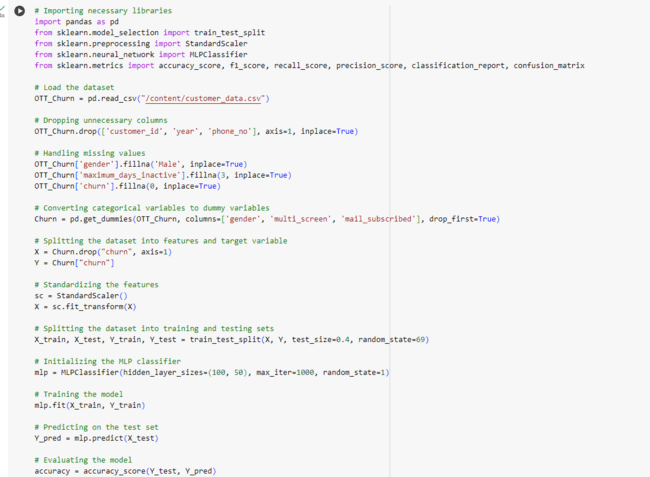
Training Progress: The model improved gradually over 50 rounds of training, with its loss decreasing each time.

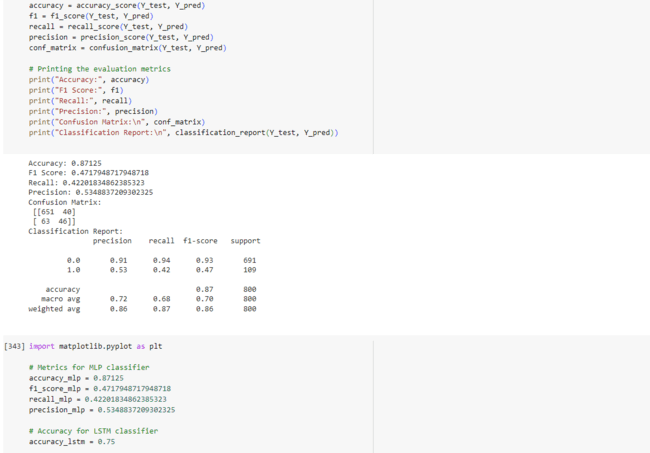
Validation Check: However, the loss on unseen data remained consistently higher than on training data, indicating potential overfitting.

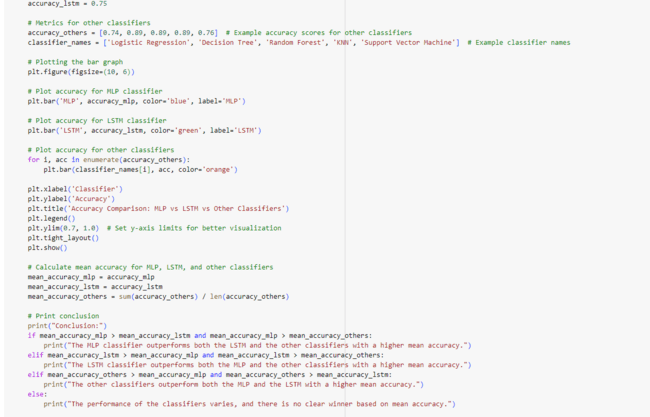
**Neural Network model, specifically a Multi-Layer Perceptron (MLP)**

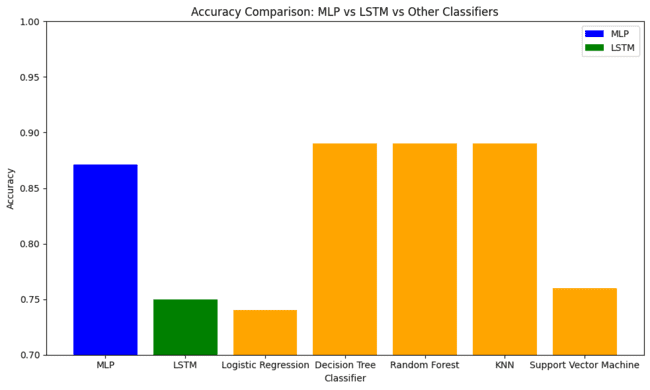
Traditional Classifiers: They may perform well when the relationship between features and target variable is linear or can be captured by simple decision rules.

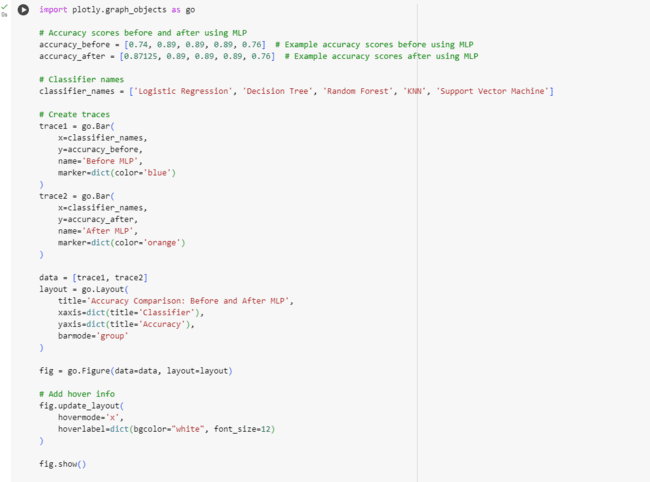
Multi-Layer Perceptron (MLP) Classifier: MLP is a type of artificial neural network composed of multiple layers of nodes (neurons) interconnected by edges (weights). Each node in a layer receives input from nodes in the previous layer, applies an activation function to the weighted sum of inputs, and produces an output that serves as input to the next layer. MLPs are capable of learning non-linear decision boundaries and can capture complex patterns in data.

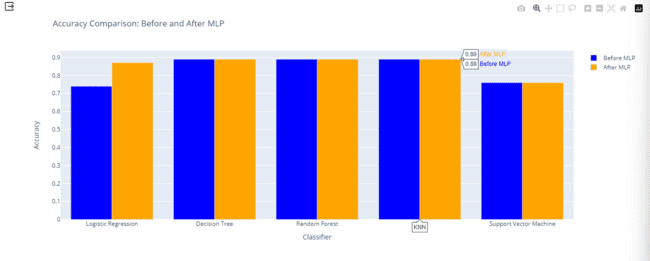


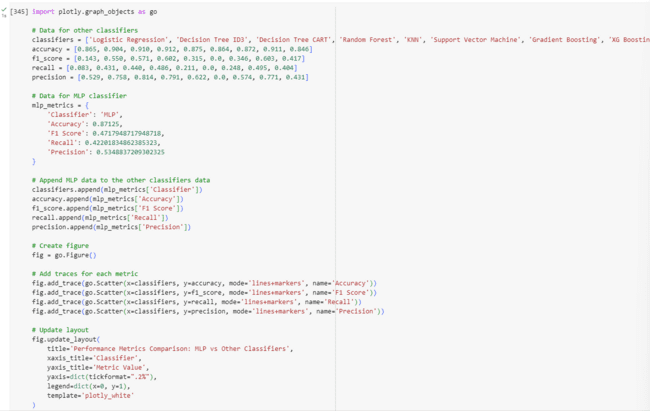




**Conclusion:** The MLP classifier outperforms both the LSTM and the other classifiers with a higher mean accuracy.











In summary, while traditional classifiers use simple decision rules to classify data, MLP classifiers leverage the power of neural networks to learn complex patterns and relationships in the data. This introduction of MLPs as a new method for churn prediction adds a layer of complexity and flexibility to the modelling approach, potentially leading to improved predictive performance.

