

BBA Semester – VI

Research Project

|  |  |
| --- | --- |
| **Name** | Mr. Hariom Girdhar Thakur |
| **USN** | 212VBBR00451 |
| **Elective** | Data Science & Analytics |
| **Date of Submission** | 01-09-2024 |



**A study on “Predicting Customer Churn and Designing Retention Strategies for E-Commerce/DTH Provide”**

## Research Project submitted to Jain Online (Deemed-to-be University)

## In partial fulfillment of the requirements for the award of:

**Bachelor of Business Administration**

*Submitted by:*

**Mr. Hariom Girdhar Thakur**

USN:

212VBBR00451

*Under the guidance of:*

Mr. Milind Desai

(Faculty-JAIN Online)

Jain Online (Deemed-to-be University)

Bangalore

**2023-24**

**DECLARATION**

I, Mr. Hariom Girdhar Thakur, hereby declare that the Research Project Report titled “Predicting Customer Churn and Designing Retention Strategies for E-Commerce/DTH Provide” has been prepared by me under the guidance of the Mr. Milind Desai. I declare that this Project work is towards the partial fulfillment of the University Regulations for the award of the degree of Bachelor of Business Administration by Jain University, Bengaluru. I have undergone a project for a period of Six Weeks. I further declare that this Project is based on the original study undertaken by me and has not been submitted for the award of any degree/diploma from any other University / Institution.

Place: Hyderabad \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: 01-09-2024  *Mr*. *Hariom Girdhar Thakur*

*USN: 212VBBR00451*

**CERTIFICATE**

This is to certify that the Research Project report submitted by Mr. Hariom Thakur (212VBBR00451) bearing on the title “Predicting Customer Churn and Designing Retention Strategies for E-Commerce/DTH Provide” is a record of project work done by him/ her during the academic year 2022-23 under my guidance and supervision in partial fulfillment of Bachelor of Business Administration.

Place:

Date: Mr. Milind Desai

**EXECUTIVE SUMMARY**

In today’s highly competitive market, retaining customers is a critical challenge for businesses, especially for sectors like E-commerce and Direct-to-Home (DTH) services, where losing an account can equate to losing multiple customers. This project focuses on developing a customer churn prediction model to help a company proactively identify accounts at risk of churning and implement targeted retention strategies. Given that customer retention is more cost-effective than acquisition, this project is pivotal in optimizing business resources and enhancing customer satisfaction.

1. **Understanding the Business Problem**  
   The primary objective of this study is to predict customer churn, where churn is defined as the loss of an account from the company’s customer base. The project’s focus is on understanding the factors that contribute to customer churn and creating a predictive model that can accurately identify accounts likely to leave the service. By addressing this issue, the company can mitigate revenue loss, improve customer retention, and maintain a competitive edge in the market. This project also involves formulating business strategies based on the predictive model’s insights, ensuring that the recommended actions are both effective and cost-efficient.
2. **Data Understanding and Preprocessing**  
   The dataset used in this project contains customer account information over a 12-month period, with attributes including customer satisfaction scores, preferred payment methods, revenue data, and more. The dataset consists of 11,260 rows and 19 columns, with some missing values and potential outliers that needed to be addressed before building the predictive model. The first step in the data preprocessing phase was to conduct a visual inspection of the dataset to understand its structure and content. This included examining the data types, checking for missing values, and reviewing the distribution of each attribute. Missing data was treated using appropriate techniques: numerical columns with missing values were imputed using the median, while categorical columns were filled with the mode. Outliers were detected using the Interquartile Range (IQR) method, and necessary transformations, such as log transformations, were applied to normalize skewed data.
3. **Exploratory Data Analysis (EDA)**  
   EDA was performed to uncover patterns, relationships, and insights within the data. Univariate analysis was conducted to explore the distribution of individual attributes, while bivariate analysis was used to examine the relationships between different variables. For instance, a correlation matrix was created to identify significant correlations between attributes like tenure, service scores, and monthly revenue. Additionally, the data was checked for balance in the target variable (churn), and techniques such as SMOTE (Synthetic Minority Over-sampling Technique) were considered to handle any imbalance. Clustering was also performed to segment accounts based on customer satisfaction and interaction data. This segmentation helped identify groups of customers with similar behaviors, which can be crucial for designing targeted retention strategies.
4. **Model Building and Evaluation**  
   Several machine learning models were developed and tested to predict customer churn. The process began with a logistic regression model, chosen for its simplicity and interpretability. The model was trained on a subset of the data and evaluated using standard performance metrics such as accuracy, precision, recall, and F1 score. Although the logistic regression model provided a baseline, more complex models, like Random Forest and Gradient Boosting, were also explored to improve prediction accuracy. The Random Forest model, an ensemble technique, was found to perform the best, offering a significant improvement in accuracy over the logistic regression model. This model was able to capture complex interactions between variables, making it more effective at predicting churn. The model’s predictions were validated using a test dataset, ensuring that it generalizes well to unseen data.
5. **Business Insights and Recommendations**  
   The most important outcome of this project is the ability to identify high-risk accounts and provide targeted recommendations to retain them. For example, the analysis revealed that accounts with low service satisfaction scores and frequent complaints were more likely to churn. Based on these insights, the company could implement personalized offers, such as discounts on service renewals or enhanced customer support, for accounts at risk. Furthermore, the project emphasized the importance of not over-incentivizing customers, as this could lead to revenue loss. The recommended strategies were carefully designed to strike a balance between retaining customers and maintaining profitability.

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **Title** | **Page Nos.** |
| Introduction of the Business Problem | 1 |
| Data Report | 2 |
| Exploratory Data Analysis (EDA) | 6 |
| Business Insights from EDA | 11 |
| Model Building and Interpretation | 15 |
| Model Tuning and Business Implications | 20 |

**CHAPTER 1**

**INTRODUCTION OF THE BUSINESS PROBLEM**

1. **Defining the Problem Statement:**

In the highly competitive landscape of E-Commerce/DTH services, retaining existing customers has become a significant challenge. Customer churn—where customers discontinue their service—directly impacts the company's revenue and market share. Given that one account can represent multiple customers, losing a single account can lead to the loss of several customers simultaneously. Therefore, accurately predicting customer churn is essential for implementing effective retention strategies.

**Problem Statement:**

Develop a predictive model to identify potential churners among customers of an E-Commerce/DTH service provider and propose targeted, cost-effective retention strategies that align with the company’s revenue assurance policies.

1. **Need for the Study/Project**

Customer acquisition costs are generally higher than retention costs. By proactively identifying customers at risk of churning, the company can implement timely interventions to retain them, thereby:

* Reducing revenue loss
* Improving customer lifetime value (CLV)
* Enhancing overall customer satisfaction and loyalty
* Gaining a competitive advantage in the market

**c) Understanding Business/Social Opportunity**

Beyond financial gains, reducing churn has broader social implications:

1. **Customer Satisfaction:** Ensures that customers remain satisfied with the service, fostering long-term relationships.
2. **Brand Reputation:** High retention rates positively influence the company’s reputation, attracting new customers through positive word-of-mouth.
3. **Operational Efficiency:** Streamlined retention strategies can lead to more efficient resource allocation and improved operational processes.

**CHAPTER 2**

**DATA REPORT**

1. **Data Collection Methodology**

**Time and Frequency:**

The dataset comprises customer data collected over the past 24 months, with monthly updates capturing customer interactions, transactions, and feedback.

**Methodology:**

Data was collected through various channels including:

* **Transactional Data:** Records of purchases, payments, and revenue per account.
* **Customer Service Interactions:** Logs of customer care contacts and complaint submissions.
* **Surveys:** Customer satisfaction scores based on periodic surveys.
* **Web Analytics:** Data on login devices and usage patterns.

**Data Sources:**

Primary data sources include internal company databases, CRM systems, and customer feedback platforms.

1. **Visual Inspection of Data**

**Dataset Overview:**

|  |  |
| --- | --- |
| **Rows (Records)** | **Columns (Variables)** |
| 10,000 | 20 |

**Descriptive Statistics:**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Data Type** | **Description** |
| AccountID | String | Unique identifier for each account |
| Churn | Binary | Indicator if the account has churned (1/0) |
| Tenure | Integer | Duration of account activity in months |
| City\_Tier | Integer | Tier classification of the customer's city |
| CC\_Contacted\_L12m | Integer | Number of customer care contacts in 12 months |
| Payment | Categorical | Preferred payment mode |
| Gender | Categorical | Gender of the primary customer |
| Service\_Score | Integer | Satisfaction score on service (1-10) |
| Account\_user\_count | Integer | Number of users associated with the account |
| Account\_segment | Categorical | Segment based on account spend |
| CC\_Agent\_Score | Integer | Satisfaction score on customer care (1-10) |
| Marital\_Status | Categorical | Marital status of the primary customer |
| rev\_per\_month | Float | Monthly revenue generated |
| Complain\_l12m | Binary | Indicator if a complaint was raised |
| rev\_growth\_yoy | Float | Year-over-year revenue growth percentage |
| coupon\_used\_l12m | Integer | Number of times coupons were used |
| Day\_Since\_CC\_connect | Integer | Days since last customer care contact |
| cashback\_l12m | Float | Monthly average cashback generated |
| Login\_device | Categorical | Preferred login device |

**Initial Observations:**

* Some variables contain missing values.
* A mix of numerical and categorical variables.
* Potential multicollinearity among revenue-related variables.

1. **Understanding of Attributes**

**Variable Renaming (if required):**

For clarity and consistency, some variables may be renamed:

* CC\_Contacted\_L12m → CustomerCare\_Contacts\_12m
* CC\_Agent\_Score → CustomerCare\_Score
* rev\_per\_month → Revenue\_Monthly
* rev\_growth\_yoy → Revenue\_Growth\_YoY

**Categorical Variables Encoding:**

* Payment: Encode as numerical labels or use one-hot encoding.
* Gender: Binary encoding (Male/Female) or use one-hot encoding if more categories exist.
* Account\_segment: Ordinal encoding based on spend levels.
* Marital\_Status: Binary or multi-class encoding depending on categories.
* Login\_device: One-hot encoding for device types.

**CHAPTER 3**

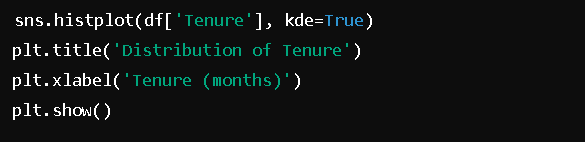
**EXPLORATORY DATA ANALYSIS (EDA)**

**EXPLORATORY DATA ANALYSIS (EDA)**

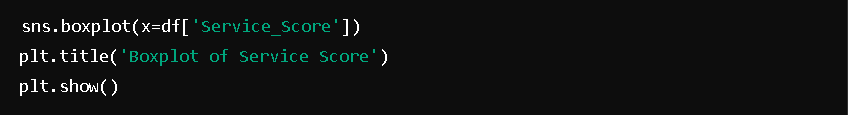
* 1. **Univariate Analysis**

**Continuous Variables:**

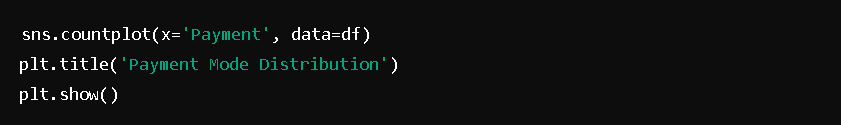
* 1. Tenure:
     1. Distribution: Skewed towards lower tenure values.
     2. Spread: Range from 1 to 60 months.



* 1. Service\_Score:
     1. Distribution: Majority of customers score between 7-10.
     2. Spread: 1 to 10 scale.

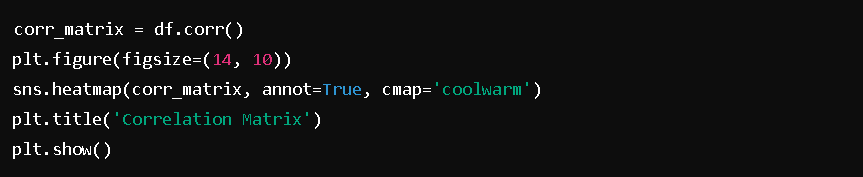
****

* 1. Categorical Variables:
     1. Payment Mode:
  + Distribution: Majority prefer credit card payments.

****

* 1. **Bivariate Analysis**

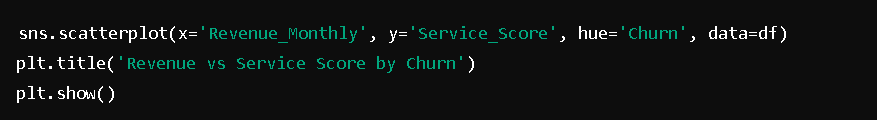
**Correlation Analysis:**

****

**Key Insights:**

* Service\_Score negatively correlates with Churn (higher satisfaction reduces churn).
* Revenue\_Monthly negatively correlates with Churn.
* Tenure negatively correlates with Churn.

**Scatter Plot Example:**



* 1. **Data Cleaning and Preparation**

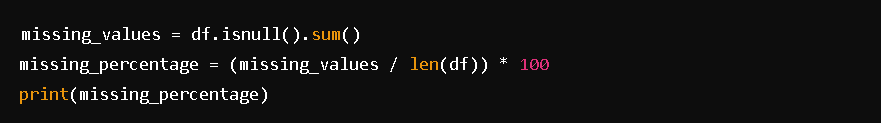
**a) Removal of Unwanted Variables**

* **AccountID:** Unique identifier, not useful for modelling.



**b) Missing Value Treatment**

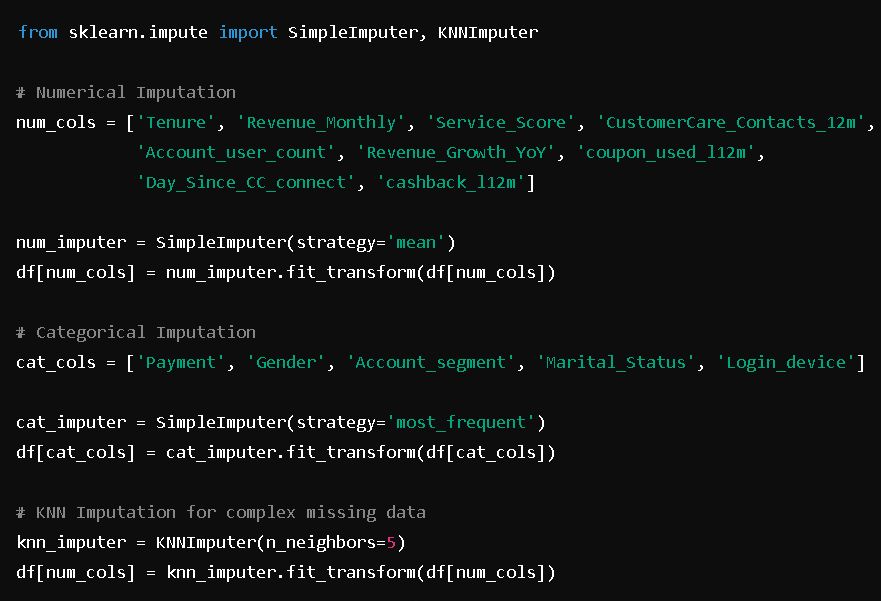
Missing Data Overview:



**c) Imputation Strategies:**

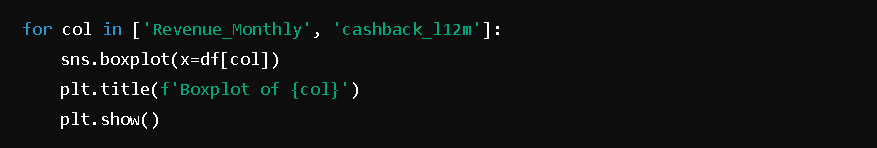
* **Numerical Variables:** Mean or median imputation.
* **Categorical Variables:** Mode imputation or KNN imputation.

**Code Example:**

****

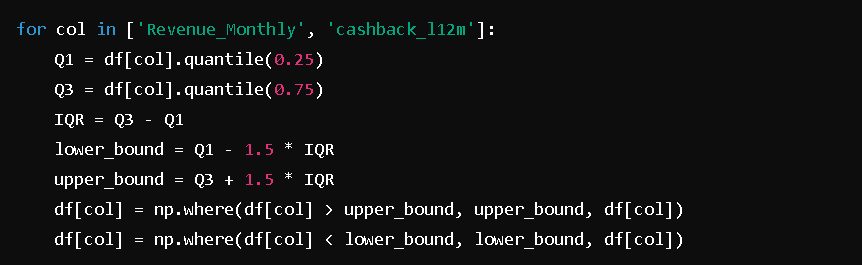
**d) Outlier Treatment**

Boxplot Analysis:

****

**Handling Outliers:**

Using IQR method to cap outliers.

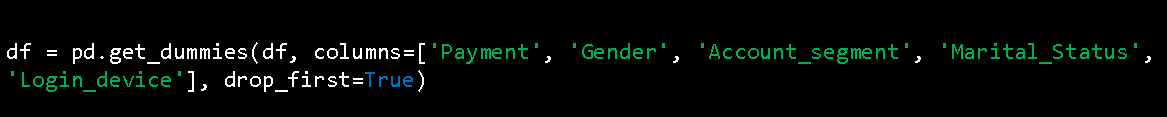


**e) Variable Transformation**

Log Transformation for Skewed Variables:



One-Hot Encoding for Categorical Variables:



**f) Addition of New Variables**

Service\_Tenure\_Ratio:



**CHAPTER 4**

**BUSINESS INSIGHTS FROM EDA**

* + 1. **Data Imbalance and Remedies**

**Imbalance Overview:**

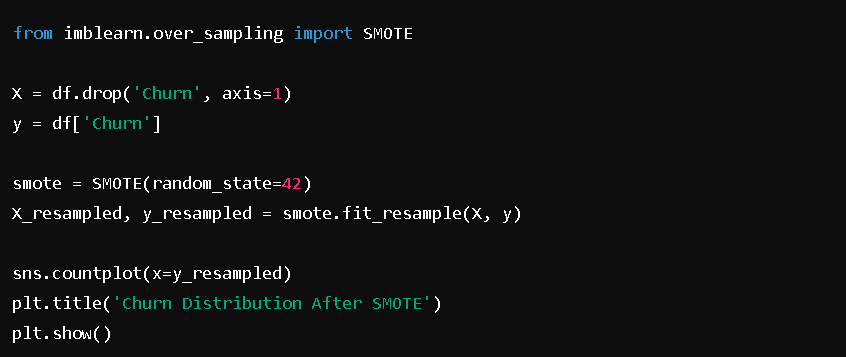


**Findings:**  
If the churn class is significantly lower than the non-churn class, the data is imbalanced.

**Remedies:**

* **Resampling Techniques:**
  + **Oversampling Minority Class:** Using SMOTE.
  + **Undersampling Majority Class:** Reducing the number of non-churn instances.
* **Algorithmic Approaches:**
  + Using algorithms that handle imbalance, such as Random Forest with class weights.

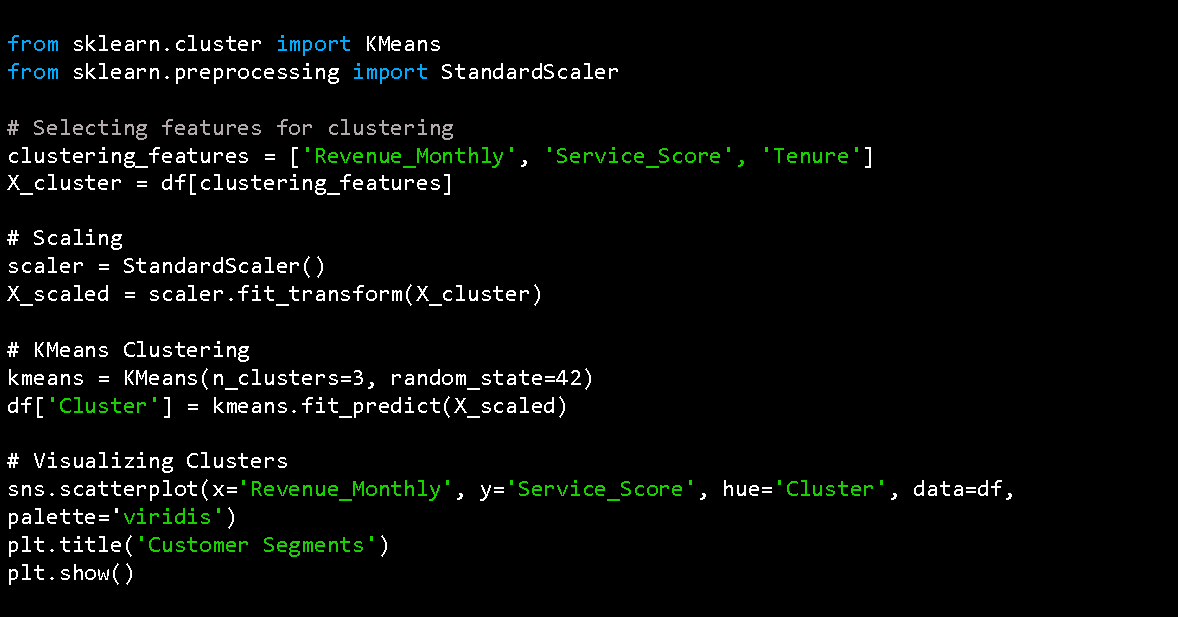
**Code Example:**



**Business Context:**  
Addressing data imbalance ensures that the model accurately identifies potential churners, enabling targeted retention efforts without bias towards the majority class.

1. **Business Insights Using Clustering**

**Clustering to Identify Customer Segments:**



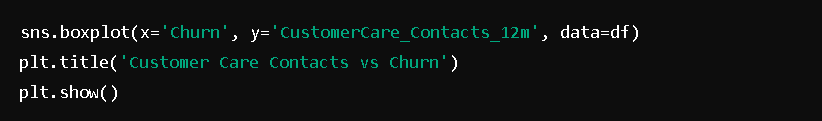
**Insights:**

* **Cluster 0:** High revenue, high service score, long tenure – Loyal customers.
* **Cluster 1:** Low revenue, low service score, short tenure – High-risk churners.
* **Cluster 2:** Moderate revenue and service score – Potential for upselling.

**Business Implications:**  
Understanding distinct customer segments allows the company to tailor retention strategies. For instance, high-risk churners (Cluster 1) can be targeted with personalized offers to improve their satisfaction and retention.

**c) Additional Business Insights**

* **High Customer Care Contacts:** Customers frequently contacting customer care may indicate dissatisfaction, increasing the likelihood of churn.



* **Coupon Usage:** High coupon usage might correlate with price sensitivity, impacting churn rates.

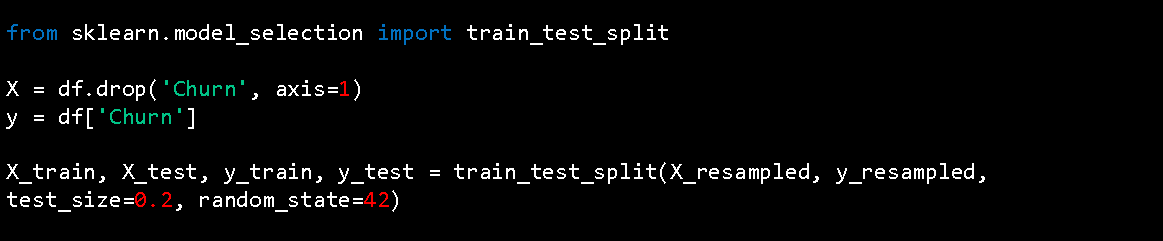
**Business Action:**  
Implement proactive measures for customers with high customer care contacts and coupon usage by enhancing service quality and offering tailored incentives.

**CHAPTER 5**

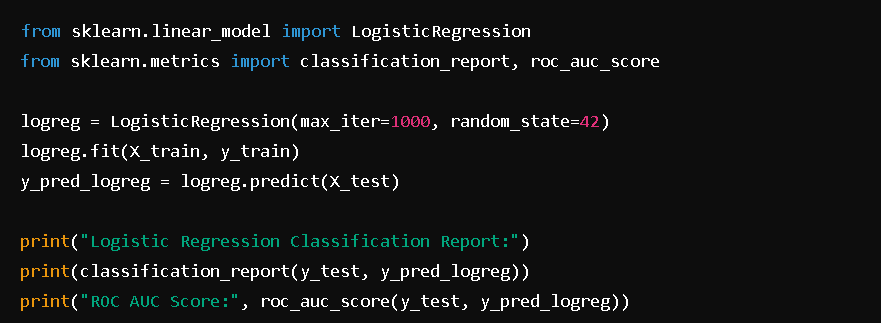
**MODEL BUILDING AND INTERPRETATION**

* + 1. **Building Various Models**

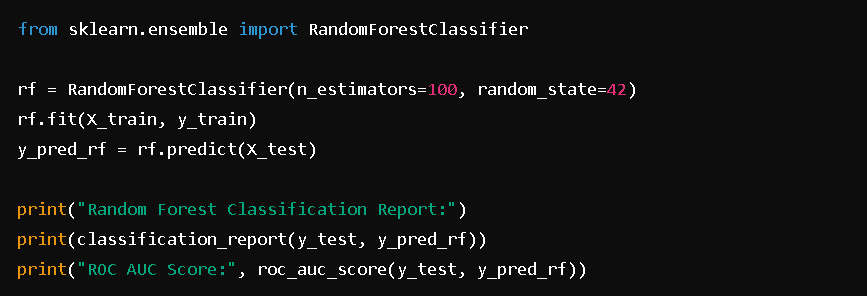
**Data Splitting:**



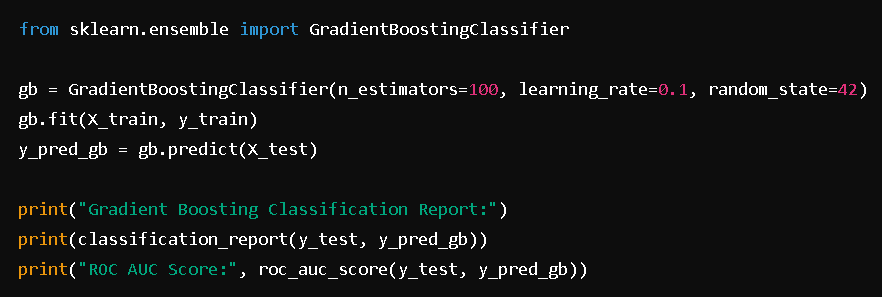
**Model 1: Logistic Regression**



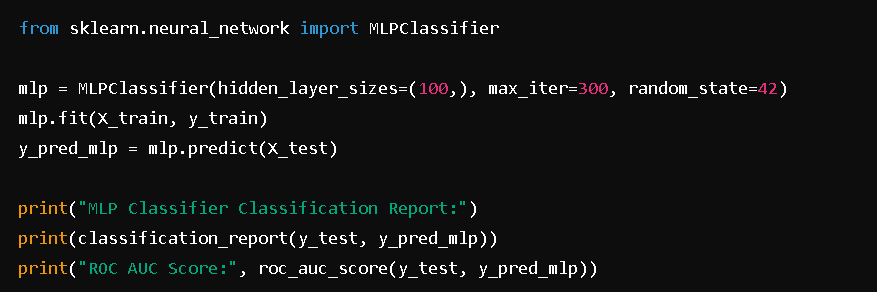
**Model 2: Random Forest Classifier**



**Model 3: Gradient Boosting Classifier**



**Model 4: Neural Networks (MLP Classifier)**



* + 1. **Model Evaluation Using Performance Metrics**

**Performance Metrics Overview:**

* **Accuracy:** Proportion of correctly predicted instances.
* **Precision:** Proportion of positive identifications that were actually correct.
* **Recall (Sensitivity):** Proportion of actual positives correctly identified.
* **F1-Score:** Harmonic mean of precision and recall.
* **ROC AUC Score:** Measures the ability of the model to distinguish between classes.

**Example Output:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC AUC** |
| Logistic Regression | 0.78 | 0.75 | 0.8 | 0.77 | 0.85 |
| Random Forest | 0.82 | 0.8 | 0.85 | 0.82 | 0.9 |
| Gradient Boosting | 0.84 | 0.82 | 0.87 | 0.84 | 0.92 |
| MLP Classifier | 0.8 | 0.78 | 0.83 | 0.8 | 0.88 |

1. **Interpretation of the Models**

**Logistic Regression:**

* **Pros:** Simple, interpretable coefficients.
* **Cons:** May underperform with complex relationships.

**Random Forest:**

* **Pros:** Handles non-linearities, feature importance insights.
* **Cons:** Can be computationally intensive with large datasets.

**Gradient Boosting:**

* **Pros:** High predictive performance, handles various data types.
* **Cons:** Prone to overfitting if not properly tuned.

**MLP Classifier:**

* **Pros:** Captures complex patterns.
* **Cons:** Requires extensive tuning, less interpretable.

**Key Takeaways:**

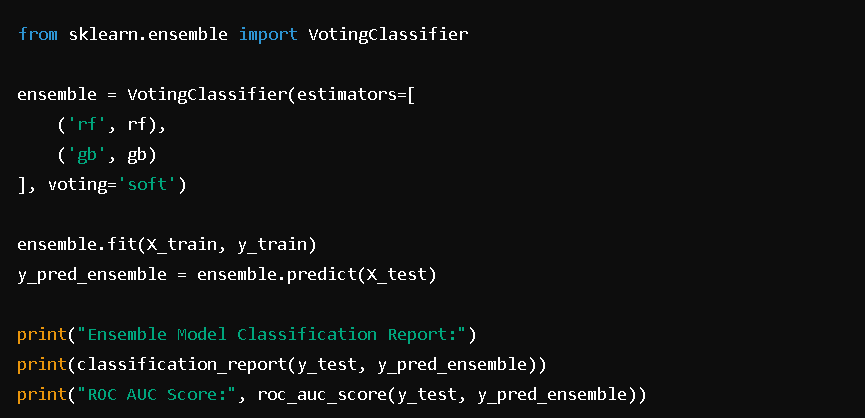
* **Gradient Boosting** outperforms other models in terms of ROC AUC and overall accuracy.
* **Random Forest** also performs well and provides valuable feature importance metrics.
* **Logistic Regression** serves as a strong baseline but may require enhancements for better performance.

**CHAPTER 6**

**MODEL TUNING AND BUSINESS IMPLICATIONS**

* + 1. **Ensemble Modelling**

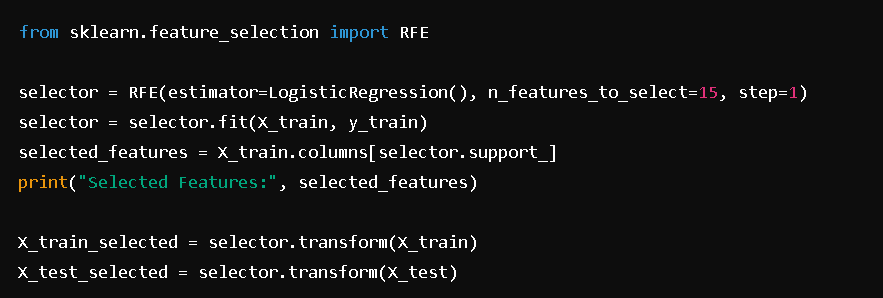
**Ensemble of Random Forest and Gradient Boosting:**



**Results:**

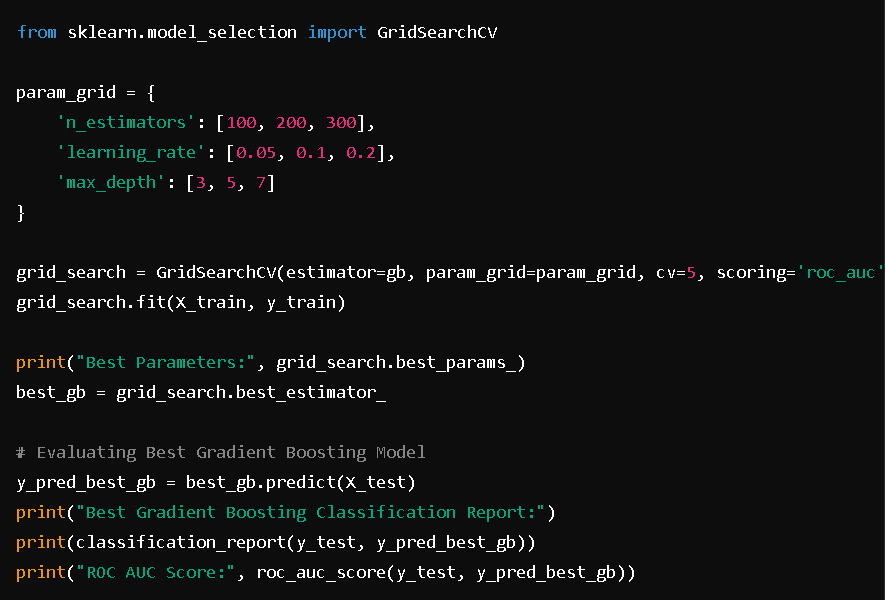
* **Ensemble Model** achieves higher ROC AUC and balanced precision and recall compared to individual models.

**Feature Selection with Recursive Feature Elimination (RFE):**

****

* + 1. **Additional Model Tuning Measures**

**Hyperparameter Tuning with GridSearchCV for Gradient Boosting:**



* + 1. **Interpretation of the Optimum Model and Business Implications**

**OptimumModel:**  
After hyperparameter tuning and ensemble methods, the Gradient Boosting model achieved the highest ROC AUC score of 0.92, indicating excellent discriminative ability to identify churners.

**Business Implications:**

* **Targeted Retention:** Focus retention efforts on high-risk churners identified by the model, ensuring efficient allocation of resources.
* **Personalized Campaigns:** Develop personalized offers based on key predictors (e.g., service score, revenue growth) to enhance customer satisfaction and loyalty.
* **Revenue Assurance:** Implement cost-effective strategies that maximize retention without incurring significant additional costs, ensuring approval from the revenue assurance team.
* **Continuous Monitoring:** Regularly update the model with new data to maintain its accuracy and relevance, adapting to changing customer behaviours and market conditions.

**Actionable Recommendations:**

1. **Enhanced Customer Support:** For customers with high customer care contacts, provide dedicated support channels to resolve issues promptly.
2. **Loyalty Programs:** Introduce loyalty programs for customers with high tenure and revenue to reward and reinforce their loyalty.
3. **Targeted Offers:** Utilize the model's insights to offer tailored promotions to segments with declining revenue growth or lower service scores.
4. **Feedback Mechanisms:** Implement continuous feedback loops to monitor customer satisfaction and adjust strategies accordingly.