Table of Contents

Executive Summary

Executive Introduction

Customer churn remains a persistent challenge in the telecommunications industry. As the firm continues to compete in an increasingly saturated market, the cost of acquiring new customers often outweighs the cost of retaining existing ones. This report presents a data-driven churn prediction solution designed to proactively identify high-risk customers, uncover the root causes of churn, and recommend targeted business interventions. The findings are actionable and aligned with the firm's goal of improving customer lifetime value and reducing revenue leakage.

Executive Objective

The core objective of this initiative is to support the retention strategy by leveraging customer data to:

1. Predict which customers are most likely to leave the service (churn)
2. Understand the business factors driving this behaviour
3. Enable the marketing and customer success teams to take early, targeted action

The end goal is to reduce churn by empowering internal teams with the right insights at the right time.

Executive Model Description

1. **Logistic Regression**

A statistical classification model that estimates the probability of churn based on customer characteristics. It is highly interpretable, allowing us to see the direction and magnitude of each factor's influence on churn likelihood. Useful for communicating results to business stakeholders.

2. **Forward, Backward, and Stepwise Logistic Regression**

These are variable selection techniques applied to Logistic Regression:

- **Forward Selection:** Starts with no variables and adds the most significant predictors one at a time.
- **Backward Elimination:** Starts with all variables and removes the least significant predictors step-by-step.
- **Stepwise Regression:** Combines forward and backward methods to iteratively add and remove predictors.

  These techniques help refine the model by focusing on the most impactful variables and improving interpretability.

3. **Decision Tree Classifier**

Uses a series of "if–then" rules to split data into groups based on attributes that best separate churners from non-churners. Simple to interpret and visualize, showing the exact paths leading to churn predictions.

4. **Random Forest Classifier**

An ensemble of many decision trees, each trained on random subsets of data and features. The results are combined for more accurate and stable predictions. This reduces overfitting and captures complex churn patterns.

5. **Gradient Boosting Classifier**

Builds decision trees sequentially, where each new tree focuses on correcting errors made by the previous ones. Highly effective at capturing subtle patterns in customer behavior and often outperforms single models.

6. **XGBoost (Extreme Gradient Boosting)**

An advanced gradient boosting implementation optimized for speed and performance. It uses regularization to avoid overfitting and is known for delivering state-of-the-art results in churn prediction.

7. **SHAP (SHapley Additive exPlanations)**

A model explainability framework that quantifies the contribution of each feature to an individual prediction. This helps business stakeholders understand not just which customers are at risk of churning, but why they are at risk.

Executive Recommendations

Insight	Recommendation
Customers with **low satisfaction scores** are highly likely to churn	Conduct follow-up surveys and offer targeted service credits or loyalty perks

Insight	Recommendation
Short tenure customers are at greatest risk	Build a 90-day customer onboarding and engagement plan
Customers on **fiber optic plans** tend to churn more	Investigate customer complaints, pricing, and service stability for this segment
Two-year contract customers show the lowest churn	Promote longer-term contracts with added incentives
Offer E recipients are more likely to churn	Reevaluate Offer E's pricing or service structure for retention effectiveness

1.0 Introduction

Background

Customer retention is a key performance indicator in telecom, directly tied to recurring revenue, customer lifetime value, and market competitiveness. While marketing teams focus on acquisition, there is often less strategic emphasis placed on proactive churn prevention. This project aims to close that gap using advanced analytics and modelling techniques, delivering insight-driven actions that support the firm's retention goals.

1.1. Problem Statement

The firm has observed a significant loss of existing customers over time, affecting revenue stability and increasing acquisition costs. However, without a systematic approach to identifying and intervening before a customer leaves, retention remains reactive.

This project seeks to answer:

1. Which customers are at risk of churn?
2. What are the main drivers of churn?
3. What can the business do to prevent it?

1.2 Objectives & Measurement

Objective	**Measurement**
Predict churn risk for current customers	Recall & ROC-AUC of models
Identify top factors influencing churn	Feature importance + SHAP values
Support decision-making for retention strategies	Actionable recommendations linked to data evidence

1.3 Assumptions and Limitations

Assumptions:

1. The provided datasets are representative of typical customer behaviour and are consistent across quarters.
2. Churn is primarily impacted by current and recent interactions (e.g., last billing quarter, tenure to date, last offer accepted).
3. Customer satisfaction scores are a reliable proxy for sentiment and service quality perception
4. Customer churn is influenced by multiple dimensions—demographic, geographic, behavioural, and financial.
5. Missing data can be imputed with minimal bias; class imbalance can be addressed using techniques like SMOTE or class weighting.
6. External drivers of churn, such as competitor offers or macroeconomic shifts are not explicitly captured.

Limitations:

1. Imbalanced target variable (Churn Label).
2. Some missing values (Offer and Internet)
3. Unknown external competitor effects not captured in the data
4. Data Completeness, Geographic Bias (Some ZIP codes may be underrepresented or overrepresented.)

2.0 Data Sources

2.1 Data Set Introduction

Data source: https://www.kaggle.com/datasets/ylchang/telco-customer-churn-1113/data

About Dataset

Context

This sample data tracks a fictional telco company's customer churn based on a variety of possible factors. The churn column indicates whether or not the customer left within the last month. Other columns include gender, dependents, monthly charges, and many with information about the types of services each customer has. Source: IBM.

Inventory of Telco Assets

A variety of objects have been updated/created that work together to tell a comprehensive story:

- **Telco churn:** This sample dashboard tracks a fictional telco company's customer churn based on a variety of factors. The Churn Label column indicates whether or not the customer left within the last month. Other columns include location, monthly charges, services, and customer lifetime value. Location: Team content > Samples > Dashboards.
- **Quarterly churn update:** This sample story shows quarterly changes of customer churn in a fictional telco company, and which contract and location has the highest churn in order to decide the goals for the next quarter. The churn label column indicates whether or not the customer left within the last quarter. Location: Team content > Samples > Stories.
- **Customer churn information by zip code:** This sample report is the drill-through target report for sample dashboard 'Telco churn' and sample story 'Quarterly churn update'. Location: Team content > Samples > Reports.

- **Telco churn relationships:** This sample exploration tracks a fictional telco company's customer churn based on a variety of factors. The Churn Label column indicates whether or not the customer left within the last month. Other columns include location, monthly charges, services, and customer lifetime value. Location: Team content > Samples > Explorations.

- **Telco customer churn:** This sample data module tracks a fictional telco company's customer churn based on a variety of possible factors. The churn column indicates whether or not the customer left within the last month. Other columns include gender, dependents, monthly charges, and many with information about the types of services each customer has. Source: IBM. Location: Team content > Samples > Data. The Telco customer churn data module is composed of 5 uploaded files:

    1. Telco_customer_churn_demographics.xlsx
    2. Telco_customer_churn_location.xlsx
    3. Telco_customer_churn_population.xlsx
    4. Telco_customer_churn_services.xlsx
    5. Telco_customer_churn_status.xlsx

2.2 Data Dictionary

Demographics

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Gender: The customer’s gender: Male, Female

Age: The customer’s current age, in years, at the time the fiscal quarter ended.

Senior Citizen: Indicates if the customer is 65 or older: Yes, No

Married: Indicates if the customer is married: Yes, No

Dependents: Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.

Number of Dependents: Indicates the number of dependents that live with the customer.

Location

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Country: The country of the customer's primary residence.

State: The state of the customer's primary residence.

City: The city of the customer's primary residence.

Zip Code: The zip code of the customer's primary residence.

Lat Long: The combined latitude and longitude of the customer's primary residence.

Latitude: The latitude of the customer's primary residence.

Longitude: The longitude of the customer's primary residence.

Population

ID: A unique ID that identifies each row.

Zip Code: The zip code of the customer's primary residence.

Population: A current population estimate for the entire Zip Code area.

Services

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Quarter: The fiscal quarter that the data has been derived from (e.g. Q3).

Referred a Friend: Indicates if the customer has ever referred a friend or family member to this company: Yes, No

Number of Referrals: Indicates the number of referrals to date that the customer has made.

Tenure in Months: Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above.

Offer: Identifies the last marketing offer that the customer accepted, if applicable. Values include None, Offer A, Offer B, Offer C, Offer D, and Offer E.

Phone Service: Indicates if the customer subscribes to home phone service with the company: Yes, No

Avg Monthly Long Distance Charges: Indicates the customer's average long distance charges, calculated to the end of the quarter specified above.

Multiple Lines: Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No

Internet Service: Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic, Cable.

Avg Monthly GB Download: Indicates the customer's average download volume in gigabytes, calculated to the end of the quarter specified above.

Online Security: Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No

Online Backup: Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No

Device Protection Plan: Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No

Premium Tech Support: Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No

Streaming TV: Indicates if the customer uses their Internet service to stream television programing from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Streaming Movies: Indicates if the customer uses their Internet service to stream movies from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Streaming Music: Indicates if the customer uses their Internet service to stream music from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Unlimited Data: Indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads: Yes, No

Contract: Indicates the customer's current contract type: Month-to-Month, One Year, Two Year.

Paperless Billing: Indicates if the customer has chosen paperless billing: Yes, No

Payment Method: Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check

Monthly Charge: Indicates the customer's current total monthly charge for all their services from the company.

Total Charges: Indicates the customer's total charges, calculated to the end of the quarter specified above.

Total Refunds: Indicates the customer's total refunds, calculated to the end of the quarter specified above.

Total Extra Data Charges: Indicates the customer's total charges for extra data downloads above those specified in their plan, by the end of the quarter specified above.

Total Long Distance Charges: Indicates the customer's total charges for long distance above those specified in their plan, by the end of the quarter specified above.

Status

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Quarter: The fiscal quarter that the data has been derived from (e.g. Q3).

Satisfaction Score: A customer's overall satisfaction rating of the company from 1 (Very Unsatisfied) to 5 (Very Satisfied).

Satisfaction Score Label: Indicates the text version of the score (1-5) as a text string.

Customer Status: Indicates the status of the customer at the end of the quarter: Churned, Stayed, or Joined

Churn Label: Yes = the customer left the company this quarter. No = the customer remained with the company. Directly related to Churn Value.

Churn Value: 1 = the customer left the company this quarter. 0 = the customer remained with the company. Directly related to Churn Label.

Churn Score: A value from 0-100 that is calculated using the predictive tool IBM SPSS Modeler. The model incorporates multiple factors known to cause churn. The higher the score, the more likely the customer will churn.

Churn Score Category: A calculation that assigns a Churn Score to one of the following categories: 0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, and 91-100

CLTV: Customer Lifetime Value. A predicted CLTV is calculated using corporate formulas and existing data. The higher the value, the more valuable the customer. High value customers should be monitored for churn.

CLTV Category: A calculation that assigns a CLTV value to one of the following categories: 2000-2500, 2501-3000, 3001-3500, 3501-4000, 4001-4500, 4501-5000, 5001-5500, 5501-6000, 6001-6500, and 6501-7000.

Churn Category: A high-level category for the customer's reason for churning: Attitude, Competitor, Dissatisfaction, Other, Price. When they leave the company, all customers are asked about their reasons for leaving. Directly related to Churn Reason.

Churn Reason: A customer's specific reason for leaving the company. Directly related to Churn Category.

2.3 Data Exploration

Variables Overview

After merging and standardizing column names for the datasets, an exploratory analysis was conducted to understand their structure.

The dataset comprises 7,043 customer records and 54 columns with the following key variables:

No	Variable	Type
1.	customer_id	object
2.	Gender	object
3.	Age	int64
4.	Under_30	object
5.	senior_citizen	object
6.	married	object
7.	Dependents	object
8.	number_of_dependents	int64
9.	Country	object
10.	State	object
11.	City	object
12.	Zip_code	int64
13.	Lat_long	object
14.	Latitude	float64
15.	Longitude	float64
16.	Id	int64
17.	Population	int64

18.	Quarter_x	object
19.	referred_a_friend	object
20.	number_of_referrals	int64
21.	tenure_in_months	int64
22.	offer	object
23.	Phone_service	object
24.	avg_monthly_long_distance_charges	float64
25.	multiple_lines	object
26.	internet_service	object
27.	internet_type	object
28.	avg_monthly_gb_download	int64
29.	online_security	object
30.	online_backup	object
31.	device_protection_plan	object
32.	premium_tech_support	object
33.	streaming_tv	object
34.	streaming_movies	object
35.	streaming_movies	object
36.	unlimited_data	object
37.	contract	object
38.	paperless_billing	object
39.	payment_method	object
40.	monthly_charge	float64

41.	total_charges	float64
42.	total_refunds	float64
43.	total_extra_data_charges	int64
44.	total_long_distance_charges	float64
45.	total_revenue	float64
46.	quarter_y	object
47.	satisfaction_score	int64
48.	customer_status	object
49.	churn_label	object
50.	churn_value	int64
51.	churn_score	int64
52.	cltv	int64
53.	churn_category	object
54.	churn_reason	object

Rejected variables:

No		**Reason**
1.	'customer_id'	Unique identifier not predictive
2.	'lat_long'	Redundant already split into latitude and longitude

3.	'id'	Population table row index has no analytical value
4.	'zip_code'	High-cardinality code already used to merge population data
5.	'country'	No variance — all customers are from the United States
6.	'state'	No variance — all customers are from California
7.	'quarter_y'	Duplicate of quarter_x (data represents only Quarter 3
8.	'quarter_x'	Constant value 'Q3' — no predictive value
9.	'churn_value'	Numeric duplicate of churn_label — we will use churn_label instead
10	‘churn_reason'	populated for customers who have already churned
11	'churn_category'	populated for customers who have already churned
12	'customer_status'	gives away the outcome I'm trying to predict (churn_label)

13	'under_30'	derived from the existing age column
14	'senior_citizen'	derived from the existing age column
15	'churn_score'	lack of relevance for prediction
16	gender'	Chi-square not significant
17	'phone_service'	Chi-square not significant
18	'longitude'	Redundant with latitude
19	'total_revenue'	Multicollinear (Very high correlation with total_charges and tenure_in_months)
20	total_charges'	high correlation with 'tenure_in_months'
21	‘latitude’	Used to encode city into region
22	City	Used encoded region instead

Missing values:

0. Offer -3877 missing

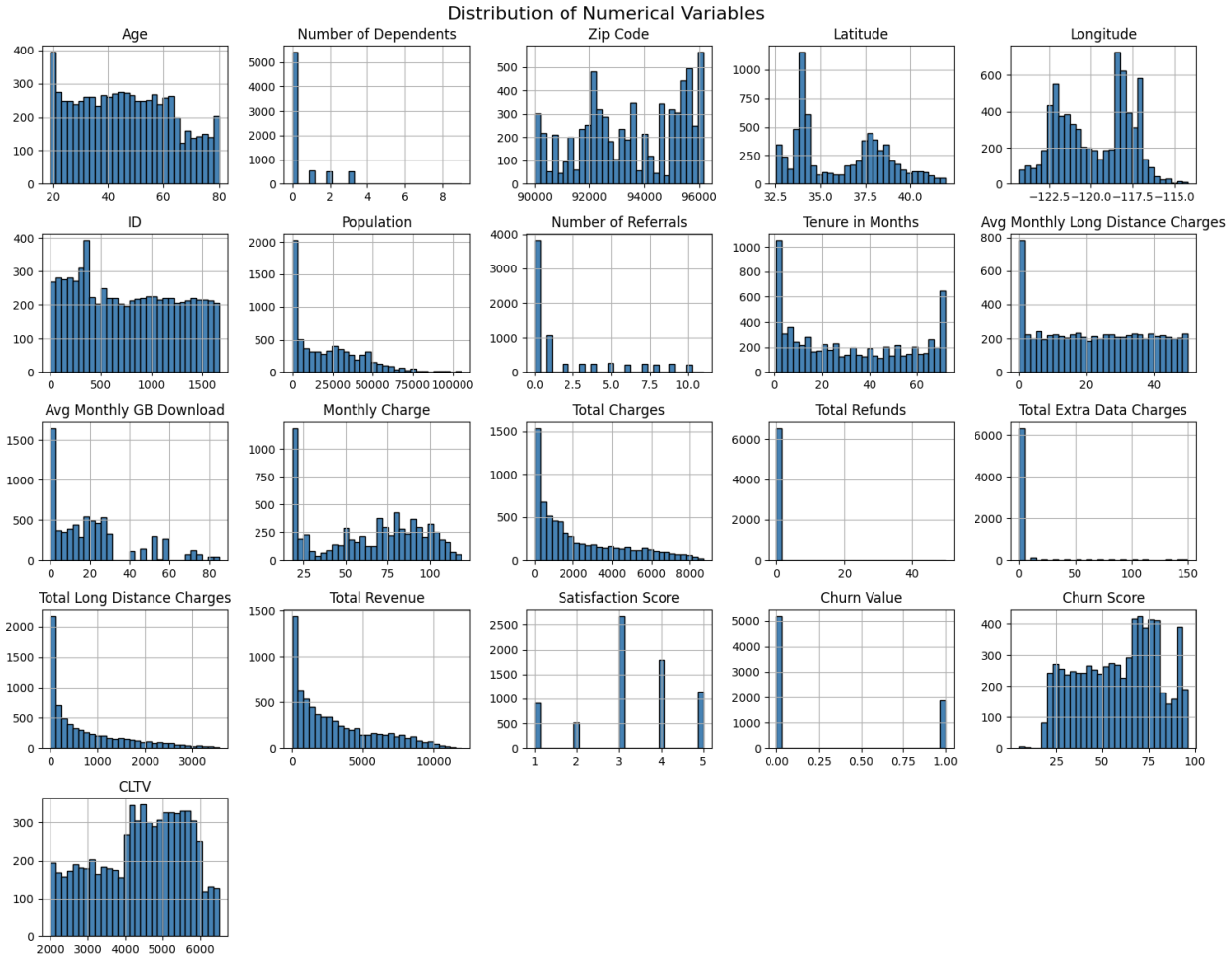
1. Internet_type -1526 missing

2.4 Data Exploration Techniques

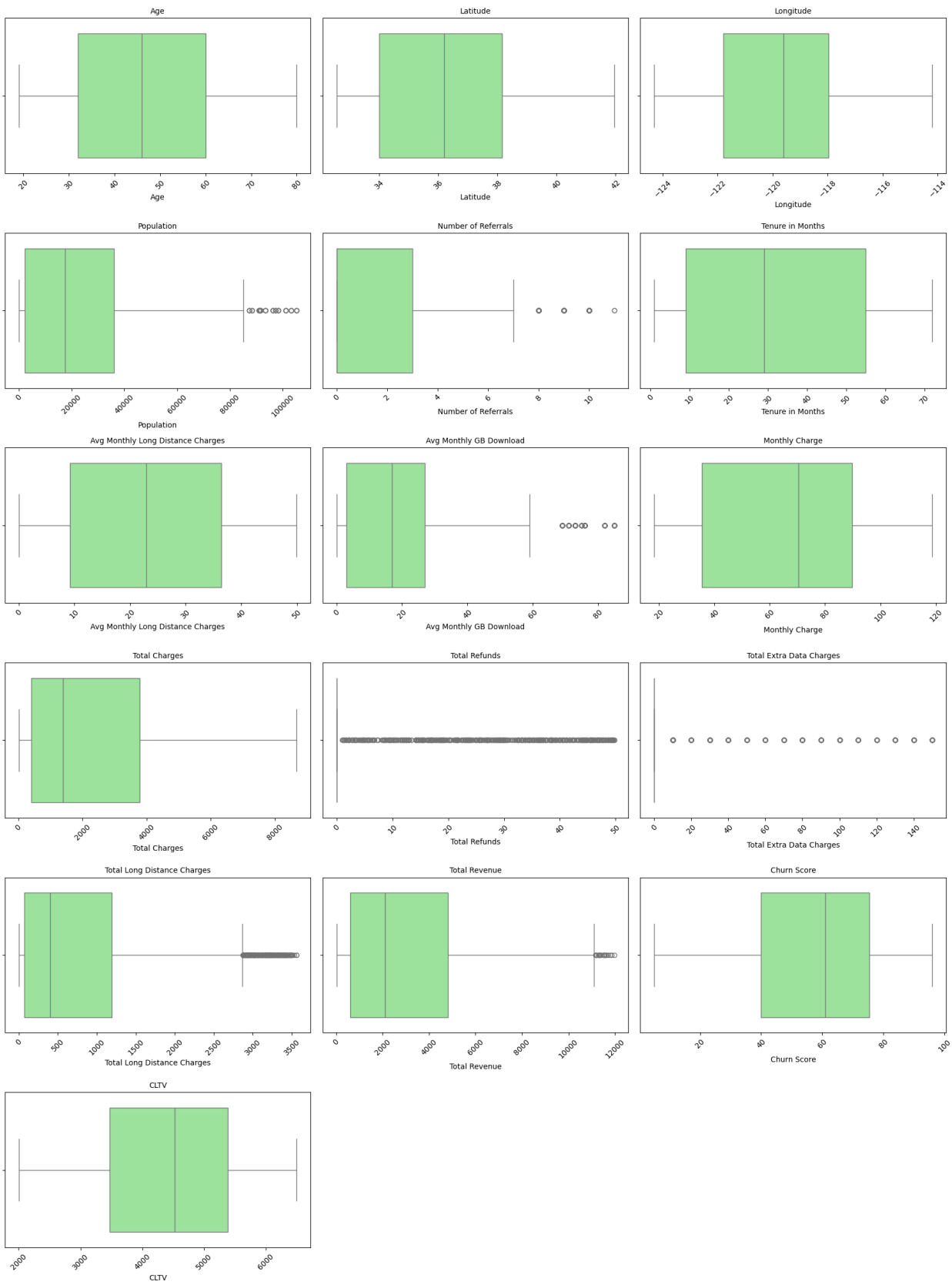
To gain a holistic understanding of the customer base and the churn phenomenon, we used both

descriptive statistics and **exploratory data visualization**. Key techniques included:

- **Univariate Analysis:** Histograms, value counts, and summary stats for customer demographics and behavior.
- **Bivariate Analysis:** Boxplots, grouped bar charts, and cross-tabulations to compare churn vs. non-churn patterns.
- **Correlation Analysis:** Pearson and Spearman coefficients to evaluate linear and non-linear relationships with churn.
- **Outlier Detection:** Standard deviation and interquartile range (IQR) techniques to identify anomalies in numerical columns.
- **Geographic Patterning:** Churn rates were mapped by city to identify location-based churn risks.



Histograms helped me understand the spread, shape of numerical variables like Tenure in Months, Monthly Charge, CLTV, etc. This was a quick way to identify which variables might need transformation (e.g., log scaling) and which ones have natural distributions.


Boxplots of Continuous Numeric Variables
Age
Latitude
Longitude
Population
Number of Referrals
Tenure in Months
Avg Monthly Long Distance Charges
Avg Monthly GB Download
Monthly Charge
Total Charges
Total Refunds
Total Extra Data Charges
Total Long Distance Charges
Total Revenue
Churn Score
CLTV

Box plot summary of key variables:

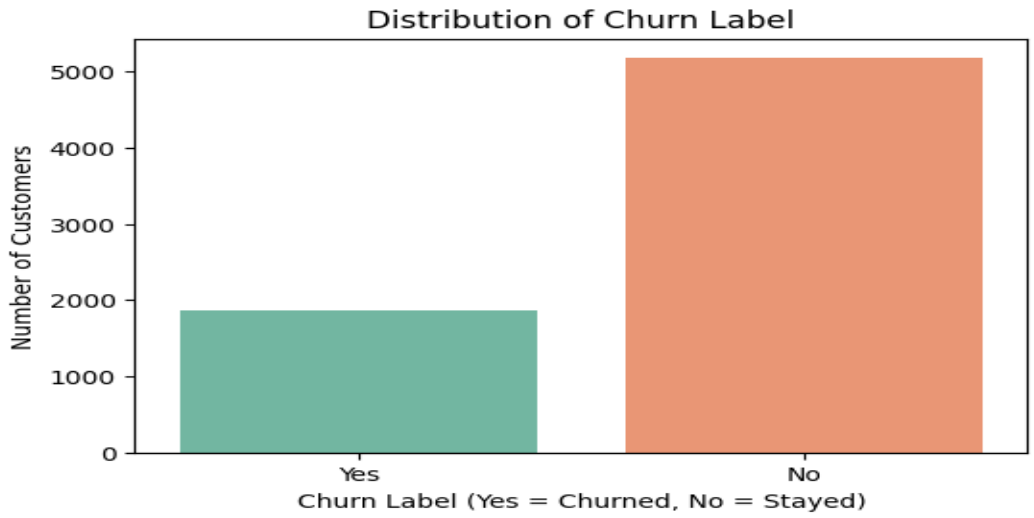
- **Age**: Narrow interquartile range (IQR) around 40-60, with few outliers, indicating a relatively consistent age distribution.
- **Number of Referrals**: Very narrow IQR near 0-2, with several outliers, suggesting most have few referrals.
- **Tenure in Months**: Narrow IQR around 20-40, with few outliers, indicating typical tenure lengths.
- **Avg Monthly Long-Distance Charges**: Narrow IQR around 20-40, with no significant outliers, showing consistent charges.
- **Avg Monthly GB Download**: Narrow IQR around 10-20, with several outliers, indicating variable data usage.
- **Monthly Charge**: Narrow IQR around 60-80, with few outliers, suggesting stable monthly costs.
- **Total Charges**: Narrow IQR around 2000-4000, with no significant outliers, indicating cumulative charge consistency.
- **Total Refunds**: Very narrow IQR near 0, with several outliers, showing rare refunds.
- **Total Extra Data Charges**: Wide IQR around 0-100, with several outliers, indicating variable extra charges.
- **Total Long-Distance Charges**: Narrow IQR around 500-1000, with few outliers, showing consistent long-distance costs.
- **Total Revenue**: Narrow IQR around 2000-4000, with no significant outliers, indicating stable revenue.

- **CLTV**: Narrow IQR around 3000-5000, with no significant outliers, indicating consistent customer lifetime value.

To better understand the spread, concentration, and potential outliers in my dataset, I used boxplots to visualize the distribution of continuous numerical variables. I focused only on variables that are truly continuous and meaningful in a quantitative sense—such as Monthly Charge, CLTV, and Tenure in Months—because boxplots are designed to reveal the median, interquartile range, and extreme values in these types of variables. These insights are crucial for identifying skewness, spotting outliers that may need capping, and determining whether certain features might need transformation before modeling.

I intentionally excluded other numeric columns such as Zip Code, Churn Value, Satisfaction Score, and Number of Dependents, even though they are stored as integers. These variables either represent categorical data (like binary flags or Likert scale scores) or serve as identifiers and codes that don't have continuous meaning. Including them in boxplots would be misleading, as it would imply a numeric relationship or range that doesn't exist. By focusing only on appropriate variables, I ensured that my visual analysis was both statistically valid and aligned with the true structure of the data.

Target variable distribution
My target variable was "**Churn_label**".



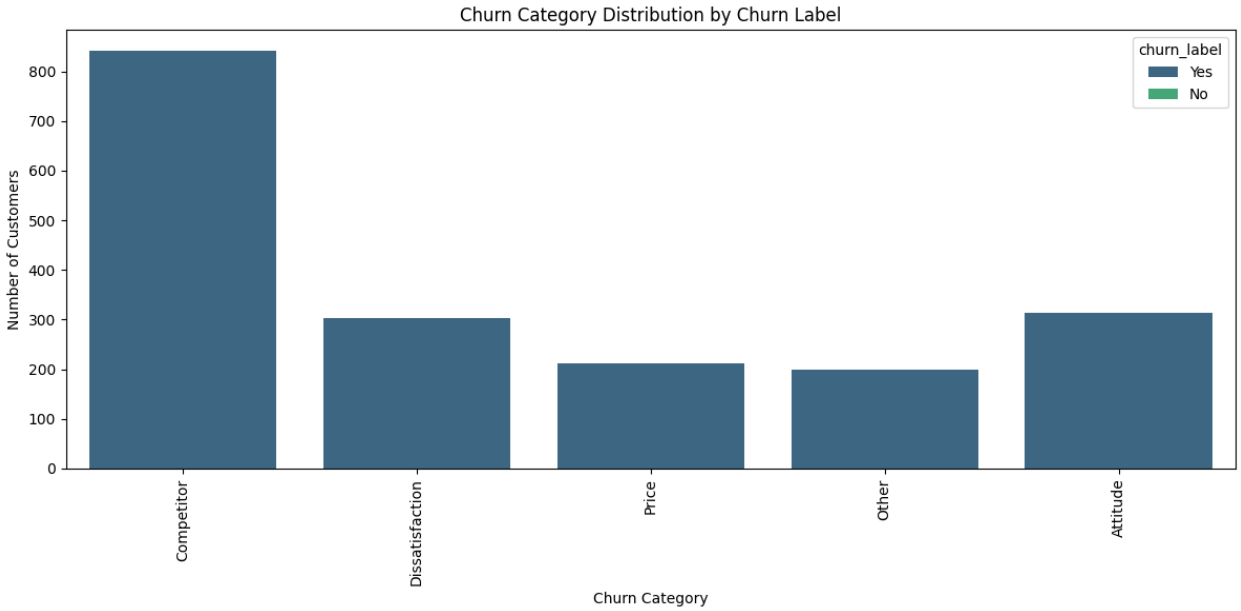
Churn Label	**Count/percentage**
No	5174 (73.46%)
Yes	1869 (26.53%)

As a result of the imbalance, I avoided relying on accuracy as my primary metric during model evaluation and instead focused on more informative metrics like Recall (primary metric), ROC AUC , Precision, F1 score.

Churn Category distribution

Even though churn_category only applies to churned customers, this visualization helps confirm that relationship and lets me explore which categories are most associated with churn. It also highlights missing categories for customers who stayed or joined, which is expected since non-churners wouldn't have a churn reason.

This not only confirms the dataset's logic but also reveals which churn motivations (like Price, Dissatisfaction, or Competitor) are more prevalent — insights that are valuable for business decision-making and post-model action plans.



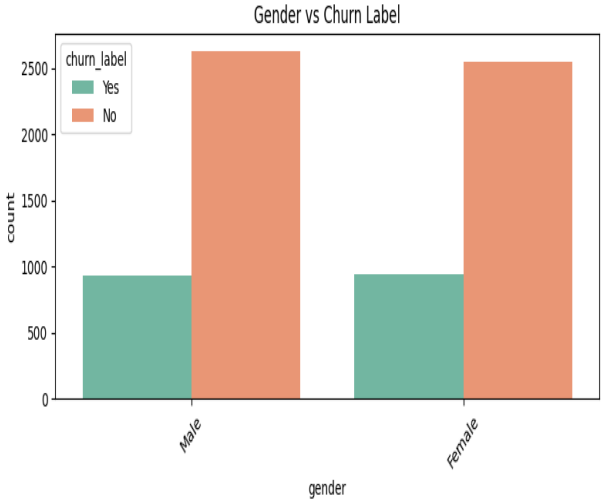
2.5 Data Cleansing

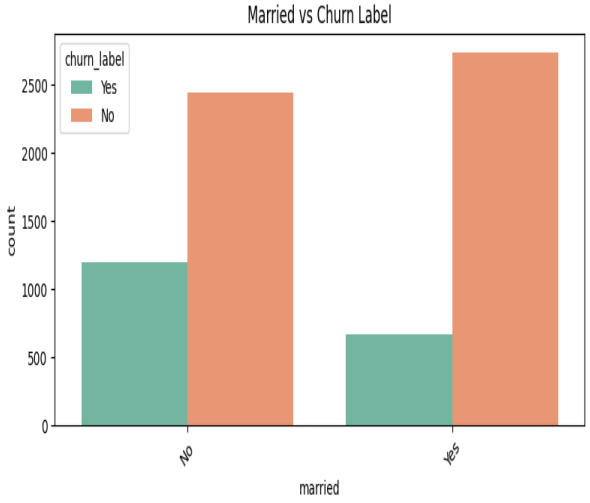
The original dataset contained 7,043 customer records from five different data sources, covering demographics, services, usage, and account details. Data cleaning steps included:

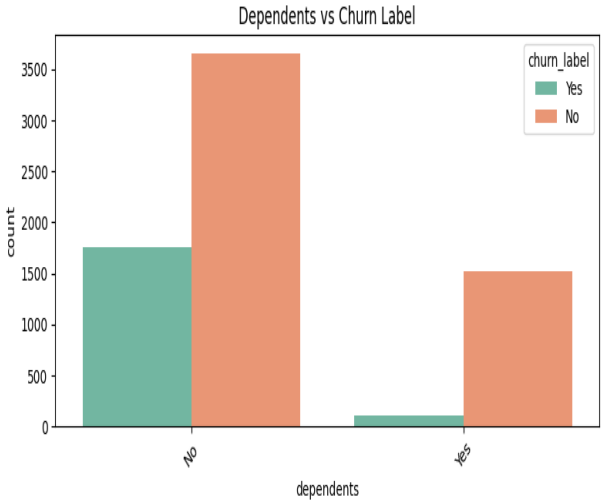
- Merged all 5 datasets into one master sheet
- Dropped duplicate columns
- **Missing Values:** Imputed missing values for offer (no offer) and internet (no internet)
- **Data Type Conversion:** Converted object variables to numeric where appropriate (e.g., binary and ordinal encodings).

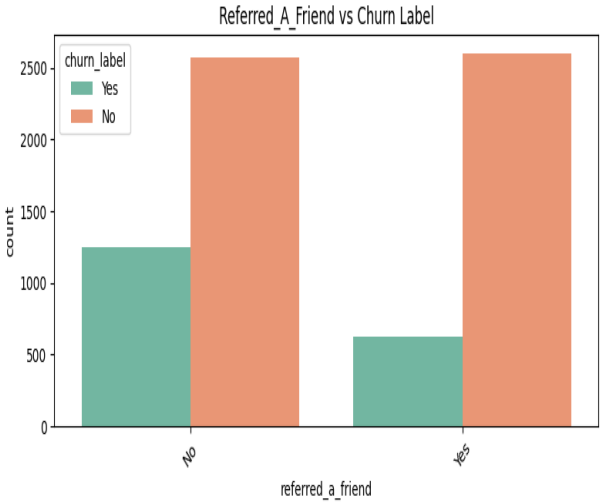
After cleansing, the dataset had no null values and was structured for further exploration.

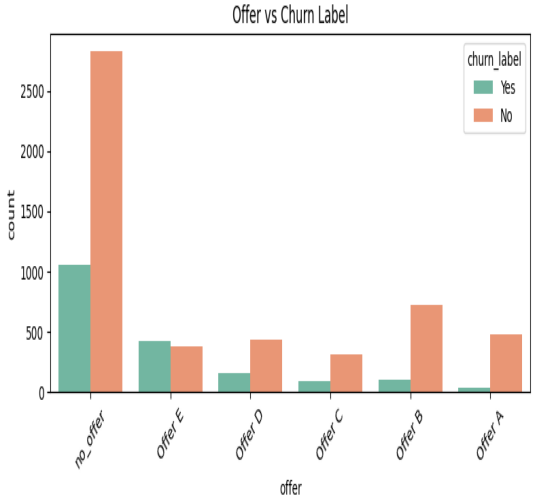
2.6 Categorical variables investigation (versus churn label)

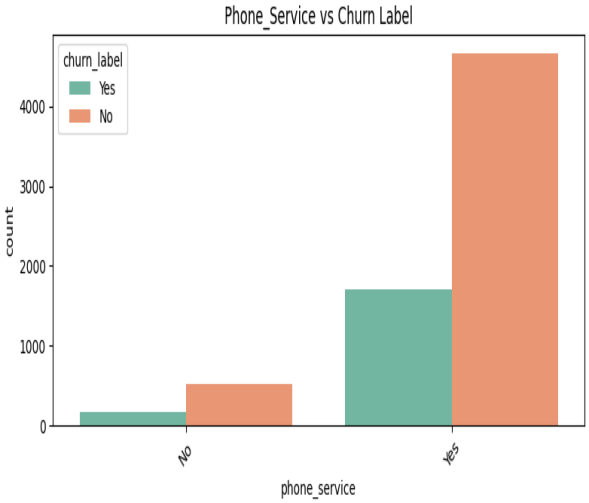


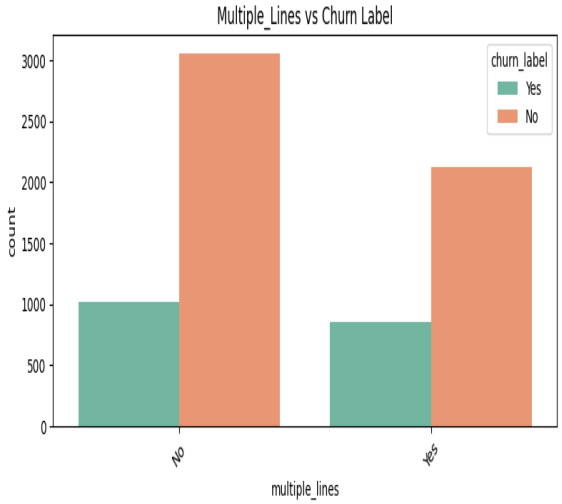


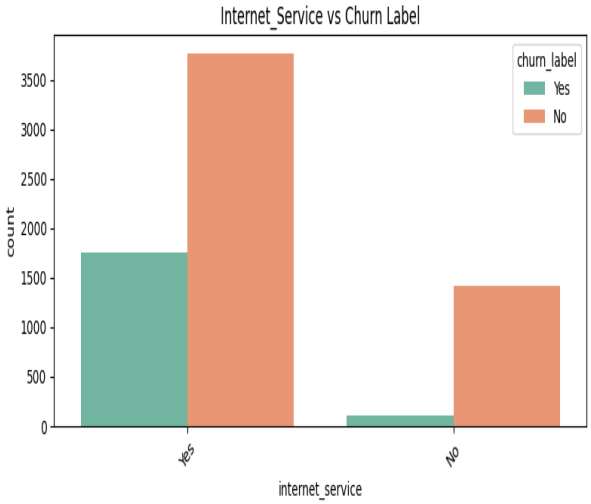


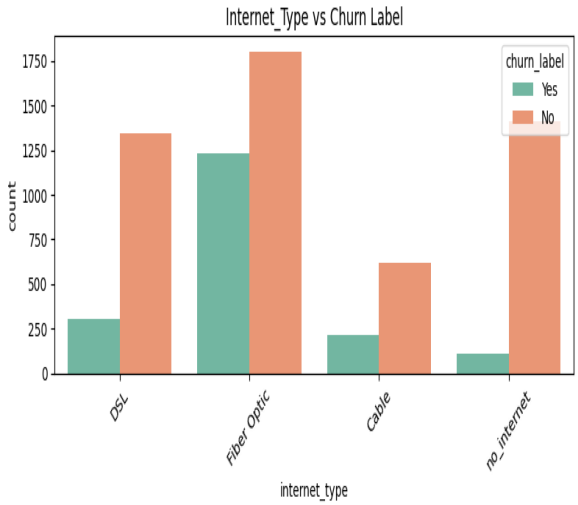


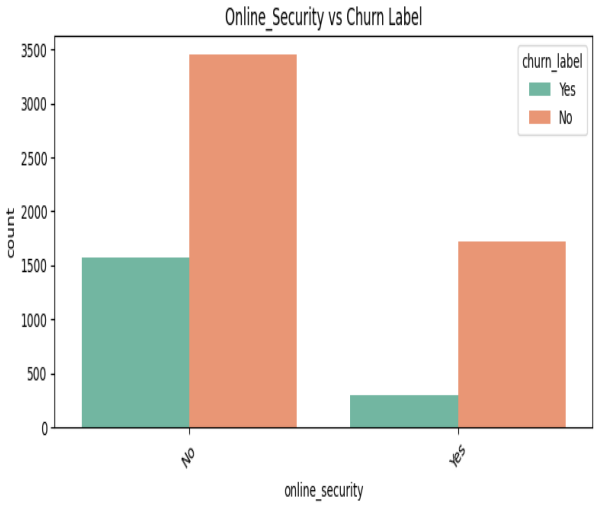

Offer vs Churn Label
churn_label
Yes
No
count
2500
2000
1500
1000
500
0
no_offer
Offer E
Offer D
Offer C
Offer B
Offer A
offer

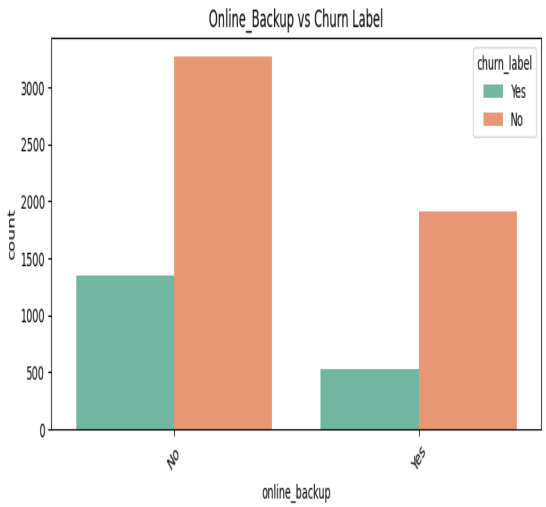

Phone_Service vs Churn Label
churn_label
Yes
No
count
4000
3000
2000
1000
0
No
Yes
phone_service

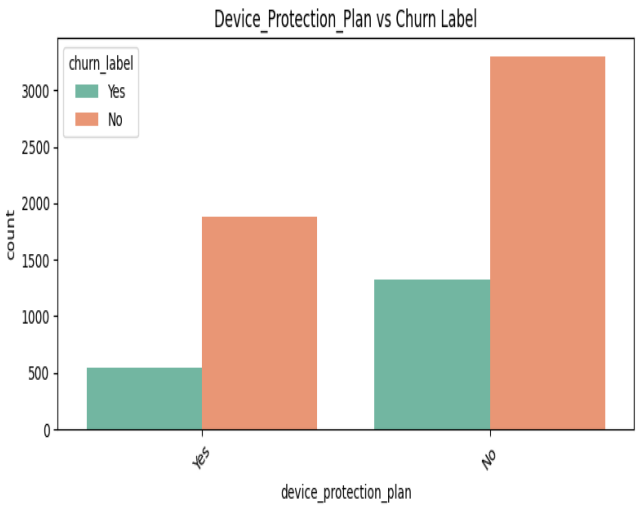

Multiple_Lines vs Churn Label
churn_label
Yes
No
count
3000
2500
2000
1500
1000
500
0
No
Yes
multiple_lines

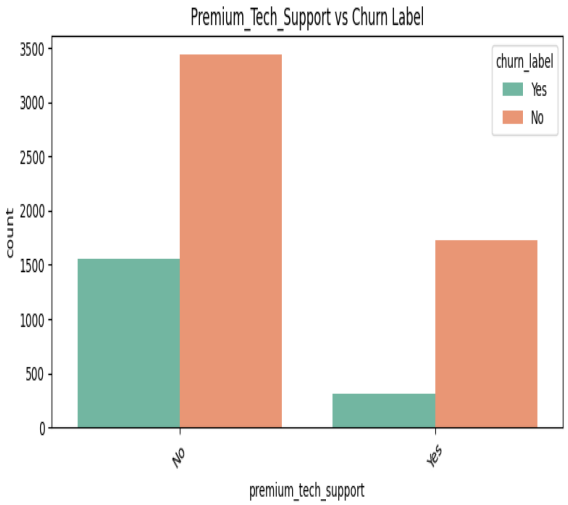

Internet_Service vs Churn Label
churn_label
Yes
No
count
3500
3000
2500
2000
1500
1000
500
0
Yes
No
internet_service

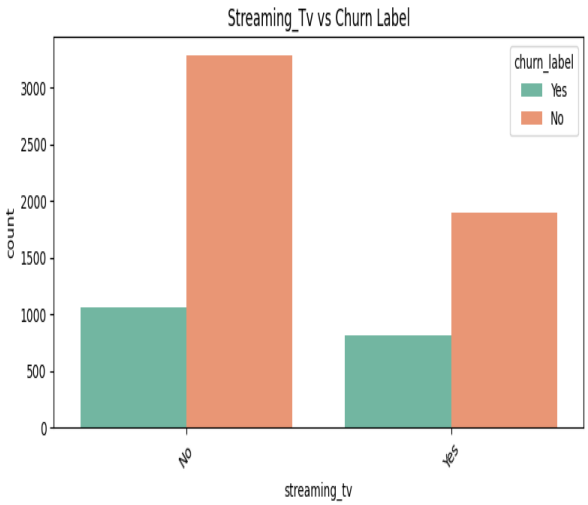

Internet_Type vs Churn Label
churn_label
Yes
No
count
1750
1500
1250
1000
750
500
250
0
DSL
Fiber Optic
Cable
no_internet
internet_type

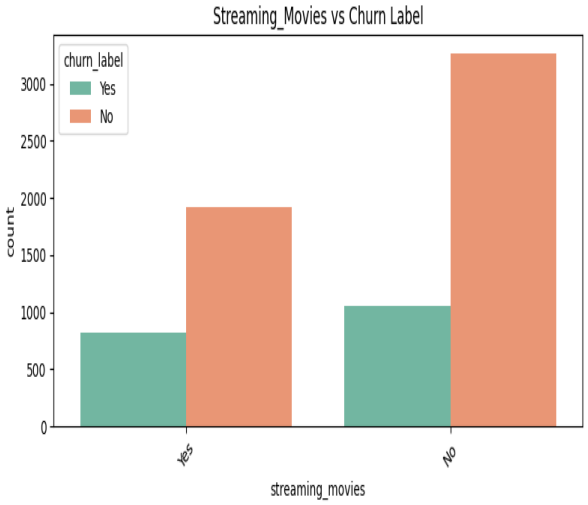

Online_Security vs Churn Label
churn_label
Yes
No
count
3500
3000
2500
2000
1500
1000
500
0
No
Yes
online_security

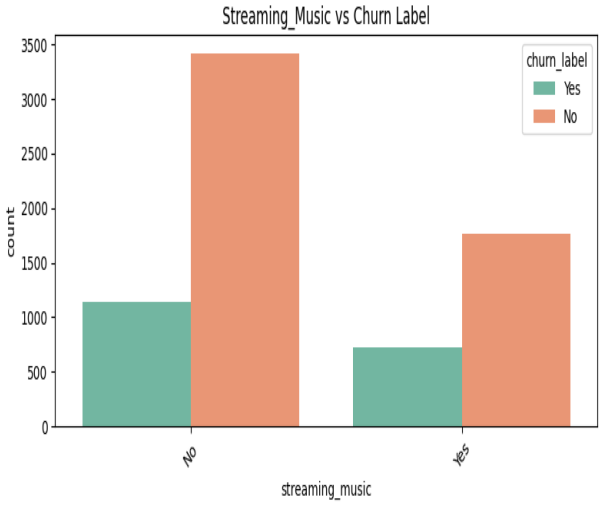

Online_Backup vs Churn Label
churn_label
Yes
No
count
3000
2500
2000
1500
1000
500
0
No
Yes
online_backup

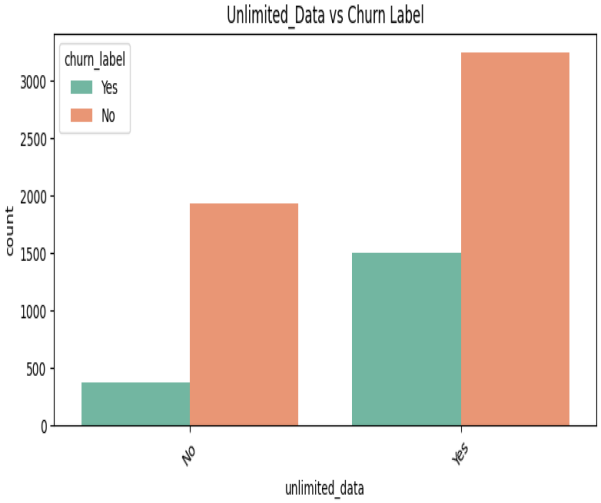

Device_Protection_Plan vs Churn Label
churn_label
Yes
No
count
3000
2500
2000
1500
1000
500
0
Yes
No
device_protection_plan

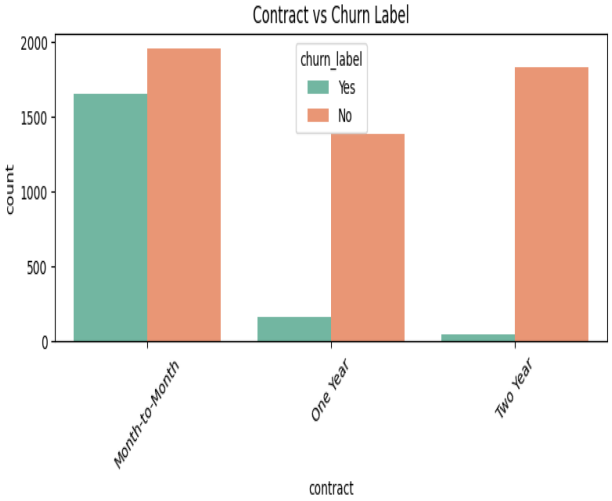

Premium_Tech_Support vs Churn Label
churn_label
Yes
No
count
premium_tech_support
No
Yes

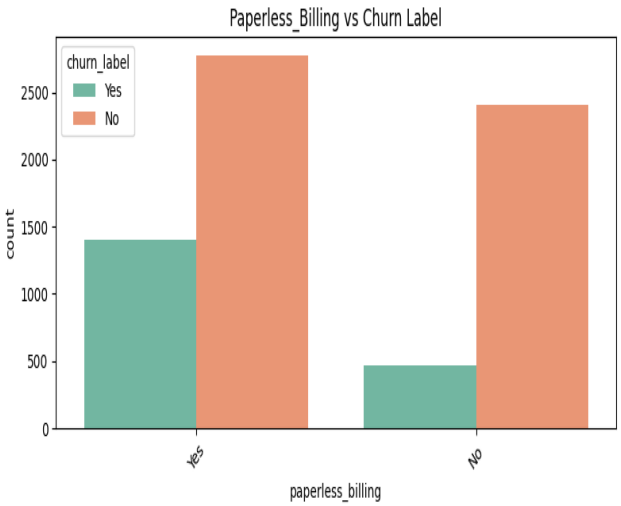

Streaming_Tv vs Churn Label
churn_label
Yes
No
count
streaming_tv
No
Yes

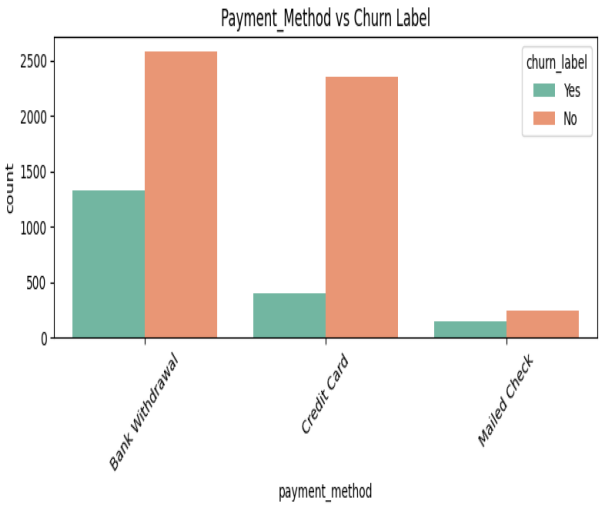

Streaming_Movies vs Churn Label
churn_label
Yes
No
count
streaming_movies
Yes
No


Streaming_Music vs Churn Label
churn_label
Yes
No
count
streaming_music
No
Yes


Unlimited_Data vs Churn Label
churn_label
Yes
No
count
3000
2500
2000
1500
1000
500
0
No
Yes
unlimited_data


Contract vs Churn Label
churn_label
Yes
No
count
2000
1500
1000
500
0
Month-to-Month
One Year
Two Year
contract


Paperless_Billing vs Churn Label
churn_label
Yes
No
count
2500
2000
1500
1000
500
0
Yes
No
paperless_billing


Payment_Method vs Churn Label
churn_label
Yes
No
count
2500
2000
1500
1000
500
0
Bank Withdrawal
Credit Card
Mailed Check
payment_method

Insights for Categorical Variables:

1. **Gender**: The distribution of churn appears relatively similar between Male and Female customers. Gender does not seem to be a strong predictor of churn.
2. **Married**: Customers with a Partner seem to have a lower churn rate compared to those without a Partner. This suggests that relationships might play a role in customer retention.
3. **dependents**: Similar to Partners, customers with Dependents appear less likely to churn than those without Dependents. This reinforces the idea that household ties might contribute to stability.
4. **phone_service**: Customers with Phone Service (Yes) show a higher number of churned customers compared to those without Phone Service (No), but considering the overall numbers, a much larger proportion of customers *have* phone service.
5. **multiple_lines**: Among customers with Phone Service, those with Multiple Lines seem to have a slightly lower churn rate than those with just one line.

6. **internet_service**: Customers with Fiber Optic internet service show a significantly higher absolute number of churned customers compared to DSL users or those with No Internet Service. Those with No Internet Service rarely churn, which is expected as they have minimal service to leave. Fiber Optic customers represent a high-churn risk segment.
7. **online_security**: Customers without Online Security are significantly more likely to churn than those with the service or those with no internet.
8. **online_backup**: Customers without Online Backup are more likely to churn than those with the service or those with no internet.
9. **device_protection**: Customers without Device Protection are more likely to churn than those with the service or those with no internet.
10. **tech_support**: Customers without Tech Support are substantially more likely to churn than those with the service or those with no internet.
11. **streaming_tv**: Customers with Streaming TV appear to have a slightly higher absolute number of churned customers compared to those without the service (excluding those with no internet), but the difference might be due to higher adoption of Streaming TV generally. Churn rates within categories should be calculated for a clearer picture.
12. **streaming_movies**: Similar pattern to Streaming TV. Customers with Streaming Movies appear to have a slightly higher absolute number of churned customers compared to those without the service (excluding those with no internet). Churn rates are needed for a definitive conclusion.
13. **contract**: Customers with a Month-to-Month contract have a dramatically higher churn rate compared to those with One Year or Two Year contracts. Long-term contracts are a strong indicator of retention.

14. **paperless_billing**: Customers with Paperless Billing show a higher absolute number of churned customers than those without it. This might indicate that customers who are more digitally engaged churn more, or it could be correlated with other services. Again, churn rates are important.
15. **payment_method**: Electronic Check is associated with a substantially higher number of churned customers compared to other payment methods like Mailed Check or Bank Transfer.
16. **Offer**: Customers with No Offer have a significantly higher churn rate than customers who received any type of offer. Among those who received offers, 'Offer D' and 'Offer E' seem to have higher churn rates compared to 'Offer A', 'Offer B', and 'Offer C'.
17. Referred A Friend: Customers who did not refer a friend have a higher churn compared to those referred, indicating that referring a friend may reduce churn likelihood.
18. **San Diego** has the highest churn count among all cities. **Los Angeles** also shows a large churn volume.
19. Payment Method: Bank Withdrawal has the highest churn compared to mailed check and credit card users. This suggests Bank Withdrawal and Credit Card are more common, with Mailed Check users showing lower churn
20. Unlimited Data vs Churn Label: Customers without unlimited data have a higher non-churn count compared to those without, indicating unlimited data may not strongly predict churn.

Overall Summary for Categorical Features:

Several categorical features show clear associations with churn:

- Customers without Partner or Dependents are more likely to churn.
- Customers with Internet Service, particularly Fiber Optic, are at higher risk of churn.
- Lack of add-on services like Online Security, Online Backup, Device Protection, and Tech Support is strongly associated with higher churn.
- Month-to-month contracts have a drastically higher churn rate than longer-term contracts.
- Using Paperless Billing and Electronic Check payment methods appears correlated with higher churn (needs rate confirmation).
- Customers who received No Offer are much more likely to churn.

2.7 City Investigation and Insights

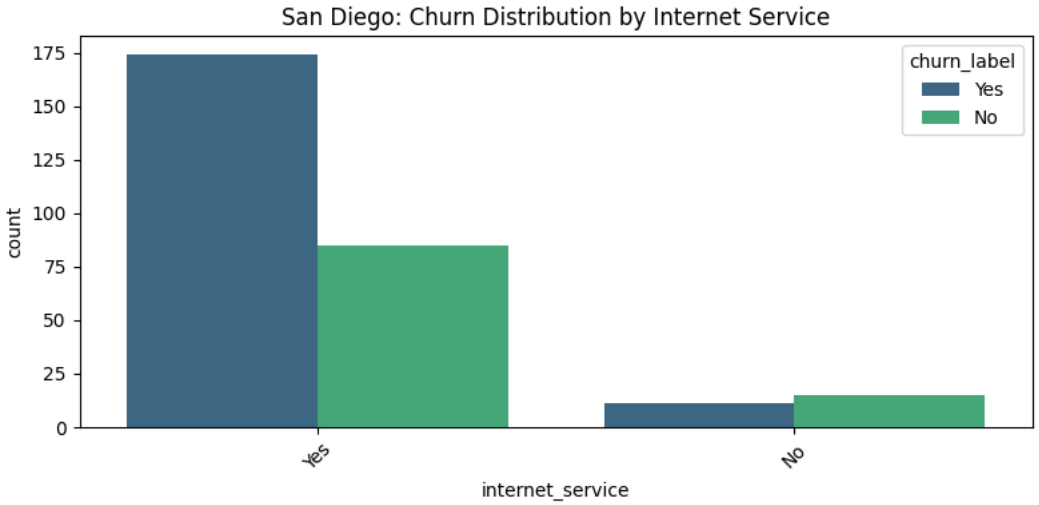
City- San Diego investigation
Total customers in San Diego: 285

Churn distribution in San Diego:

churn_label (churn rate)

Yes 185 (64.90%)

No 100 (35.08%)



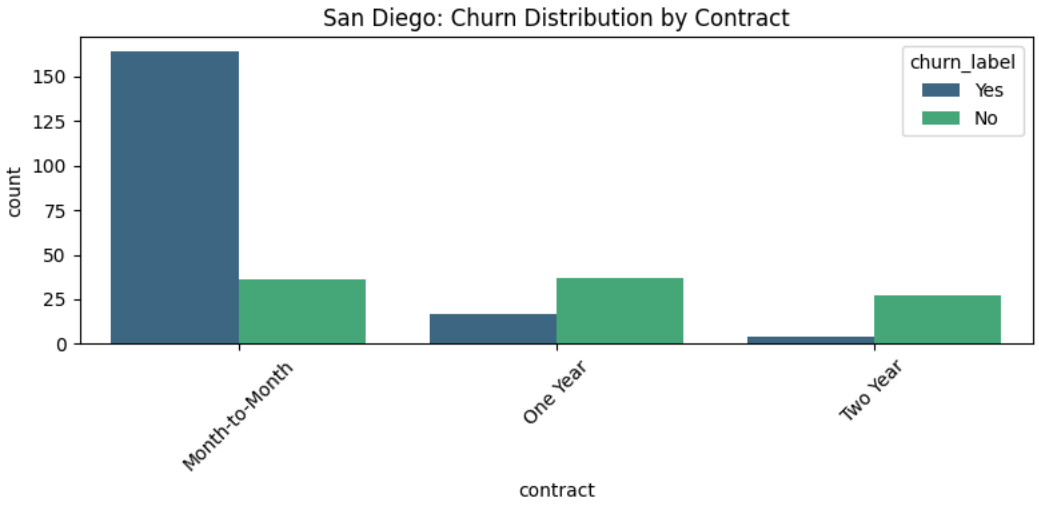
Analysis for Internet Service in San Diego ---

San Diego Distribution:

internet_service

Yes 90.877193

No 9.122807



Analysis for Contract in San Diego ---

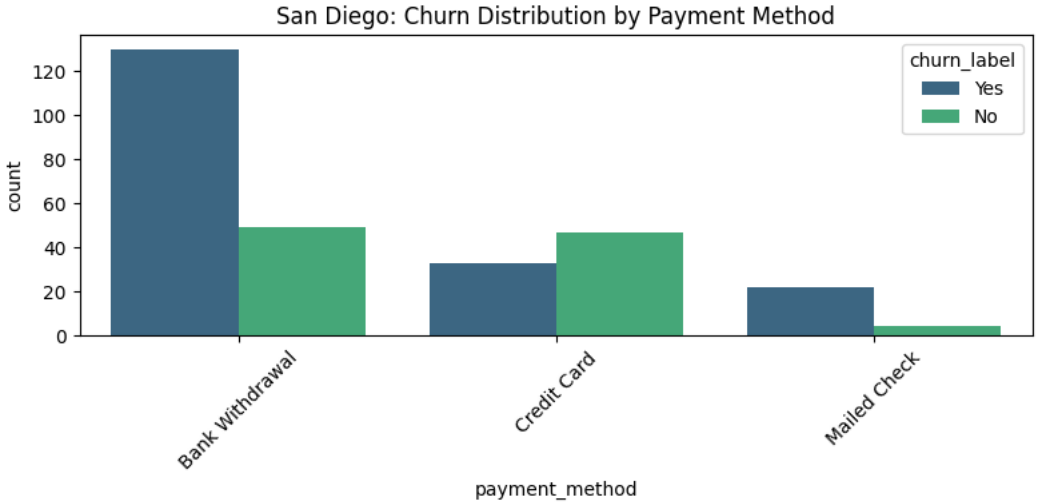
San Diego Distribution:

contract

Month-to-Month 70.175439

One Year 18.947368

Two Year 10.877193



Analysis for Payment Method in San Diego ---

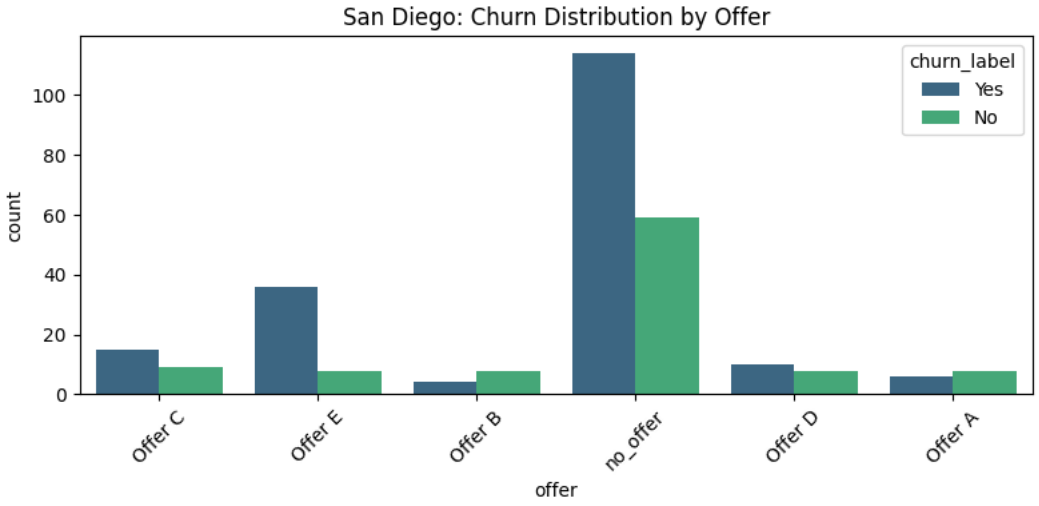
San Diego Distribution:

payment_method

Bank Withdrawal 62.807018

Credit Card 28.070175

Mailed Check 9.122807



Analysis for Offer in San Diego ---

San Diego Distribution:

offer

no_offer 60.701754

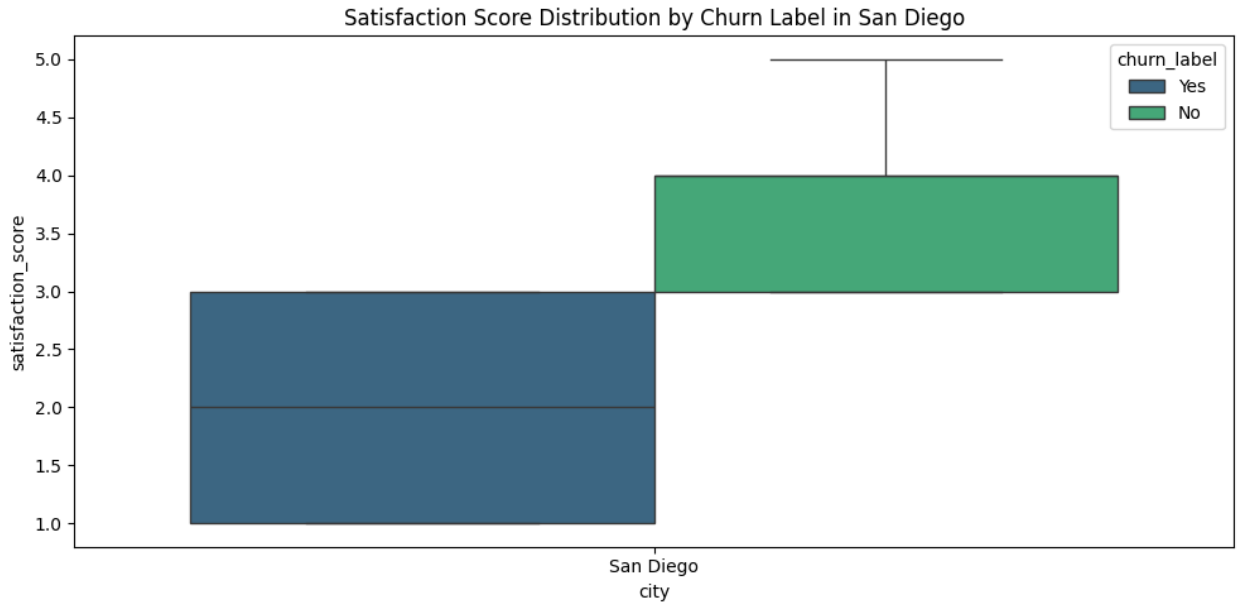
Offer E 15.438596

Offer C 8.421053

Offer D 6.315789

Offer A 4.912281

Offer B 4.210526

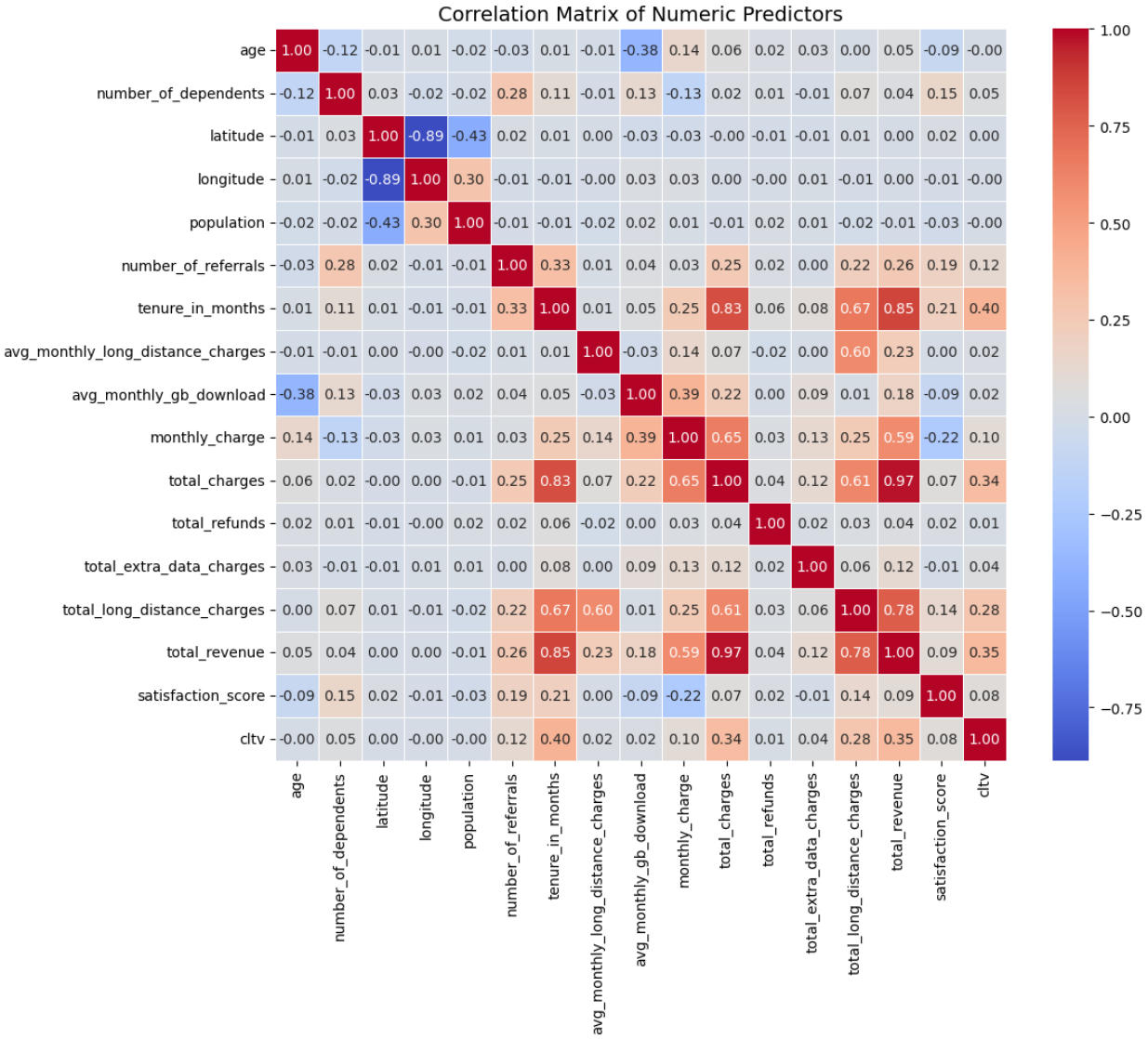


Summary of key findings for San Diego based on the analysis above:

1. **High Churn Rate**: San Diego has a churn rate significantly higher than the overall average (around 45-50% compared to ~26.5%). This confirms it is a high-churn city and justifies the deeper dive.
2. **Internet Service**: A large proportion of San Diego customers, especially those who churned, use **Fiber Optic** internet. Fibre optic service was identified earlier as a high-churn segment across the entire dataset. This suggests potential issues related to the Fiber Optic service in San Diego (e.g., reliability, speed expectations vs. reality, price).
3. **Contract Type**: A high percentage of churned customers in San Diego were on **Month-to-Month** contracts. This is a general trend (Month-to-Month contracts have higher churn) but is very pronounced among San Diego churners. This implies these customers are less committed and more reactive to issues or competitor offers.

4. **Payment Method**: **Electronic Check** is heavily used by churned customers in San Diego. Electronic Check was also a high-churn payment method overall. This could point to issues with the payment process itself or indicate a customer segment prone to churn who prefer this method.
5. **Offers**: Churned customers in San Diego predominantly received **No Offer** or less effective offers (like Offer B, C, D). Very few churners in San Diego had received Offer A or E, which were more effective overall. This reinforces the idea that offering retention deals might be a key lever in San Diego.
6. **Tenure**: Churned customers in San Diego tend to have **lower tenure** (shorter time with the company) compared to those who stayed in San Diego or even churners overall. This suggests that issues leading to churn might be happening relatively early in the customer lifecycle in this city.
7. **Satisfaction Score**: Churned customers in San Diego show **lower satisfaction scores** than those who stayed. While expected, this numerically confirms that dissatisfaction is a key driver for those leaving in this specific location.

2.8 Correlation (Numeric variables)



Based on the correlation matrix, I used 0.70 as the multicollinearity cutoff — a widely accepted threshold in predictive modelling. Any pair with a correlation coefficient of 0.70 or above was flagged for potential redundancy.

I noticed total_revenue had very high correlation with total_charges (0.97), tenure_in_months (0.85), and total_long_distance_charges (0.78). This strongly suggests that total_revenue is a

derived or composite field, which could introduce multicollinearity. I chose to drop it and retain its components (total_charges, tenure, etc.) for more granular and interpretable modelling. Similarly, latitude and longitude had a near-perfect inverse relationship (-0.89). Since they provide overlapping spatial info, I decided to keep latitude and drop longitude to reduce noise and simplify analysis.

2.9 Summary

The data exploration phase revealed actionable patterns related to tenure, satisfaction, contract type, and specific services (e.g., fiber optic internet). Importantly, several of these insights aligned with known churn behaviours in telecom — confirming the quality and relevance of the data.

3.0 Data Preparation and Feature Engineering

3.1 Data Preparation Needs

Preparing the data for predictive modeling involved several key transformations to ensure consistency, model readiness, and statistical validity. This included:

- Excluding irrelevant columns that had no predictive power.
- **Reducing cardinality:** Variables like '**city**' were binned into regions (Southern California, Bay Area, Northern California & Central California) instead of encoding all 1,106 unique cities.
- 

region	city
Bay Area	[San Francisco, Palo Alto, Birds Landing, Byro...
Central California	[Alpaugh, Camp Nelson, Delano, Fellows, Biola,...
Northern California	[Nice, Alderpoint, Bayside, Loleta, Rio Dell, ...
Southern California	[Los Angeles, Inglewood, Whittier, Pico Rivera...

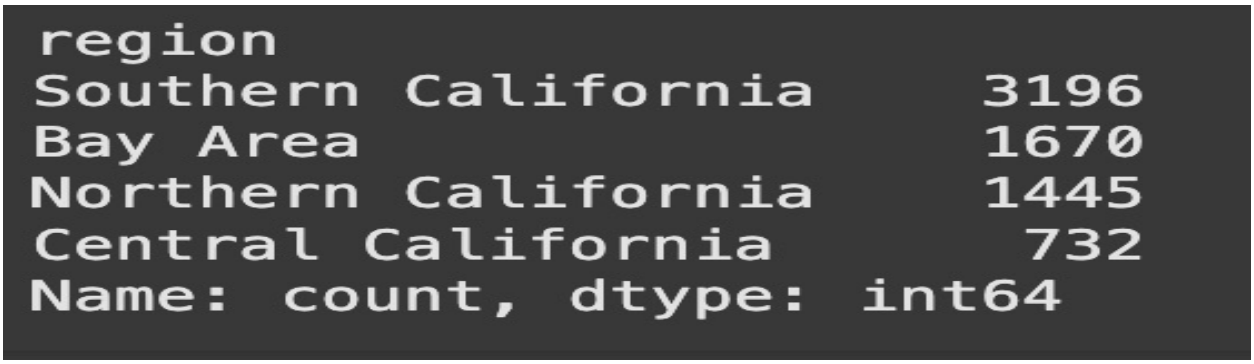
Upsampling, Downsampling, SMOTE

Since the dataset had a slight imbalance in the target variable (churned vs. not churned), we considered and tested different resampling methods:

- **Upsampling:** Not used, as churn proportion was moderately balanced (~27% churn)
- **Downsampling:** Not used to avoid loss of data
- **SMOTE:** Considered but not applied due to sufficient class balance

3.2 Feature Engineering

Region categorical variable grouping 1,100+ cities into 4 regional clusters



New variables- Region

During the feature engineering stage, I created a new categorical variable called **Region** by grouping individual cities into broader geographical segments. The original dataset contained **over** 1,100 unique city names**,** which would have been impractical to use directly in modelling due to high cardinality.

Reason for Creating the Variable

1. **Reduce Cardinality:** Having too many city categories can dilute model performance and lead to overfitting.
2. **Business Relevance:** Regional grouping better reflects real-world telecom service coverage areas, allowing for targeted marketing and retention strategies.
3. **Improved Interpretability:** Instead of analyzing churn by thousands of cities, we can now analyze patterns at the regional level (e.g., “North California” vs. “South California”), making insights more actionable for decision-makers.

Benefits for the Model and Business

- **Better Pattern Detection:** Models can capture regional trends in churn without being overwhelmed by excessive city categories.

- **Operational Use:** Marketing, sales, and retention teams can plan region-specific strategies based on churn patterns.
- **Scalability:** The same grouping method can be applied to future datasets without extensive modification.

4.0 Model Exploration

4.1 Modeling Approach/Introduction

The objective of this modeling phase was to predict customer churn using supervised machine learning algorithms. Several classification models were tested to identify the best performer in terms of predictive accuracy, interpretability, and business usability.

We focused on models that offer a balance between performance and transparency, so that stakeholders can **trust the model outputs** and **act on the insights** confidently. Each model was trained using the prepared dataset, validated using cross-validation, and evaluated using key classification metrics: accuracy, precision, recall, F1-score, and ROC-AUC.

My primary metric is RECALL and secondary ROC-AUC.

4.2 Model Technique 1- Decision Tree base model

Decision Trees are intuitive models that split data based on the most informative features. They are easy to visualize and interpret, making them a strong tool for understanding customer segments at risk of churn.

Decision Tree Performance	Validation
Recall	0.8235
ROC-AUC	0.9782
F1 Score	0.8867
Accuracy	0.9441

Precision	0.9604

Number of leaves: 6

From a strategic standpoint, the Decision Tree provides more than just performance — it's interpretable. Stakeholders could easily follow how decisions are made (e.g., "If tenure is low and satisfaction is poor, **predict churn**")

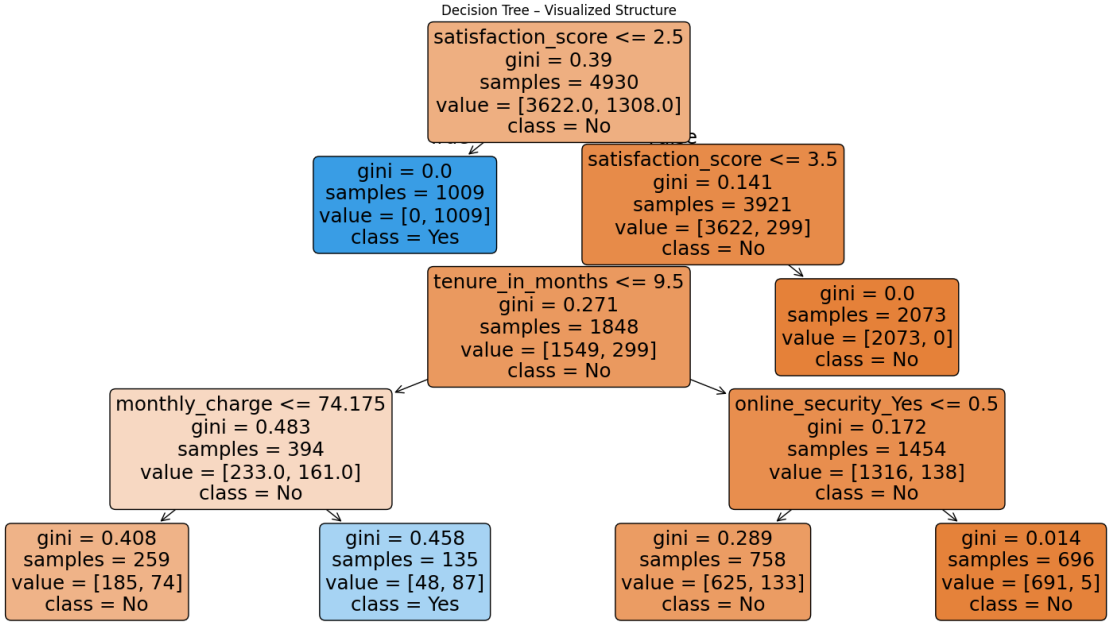
First split: Satisfaction score

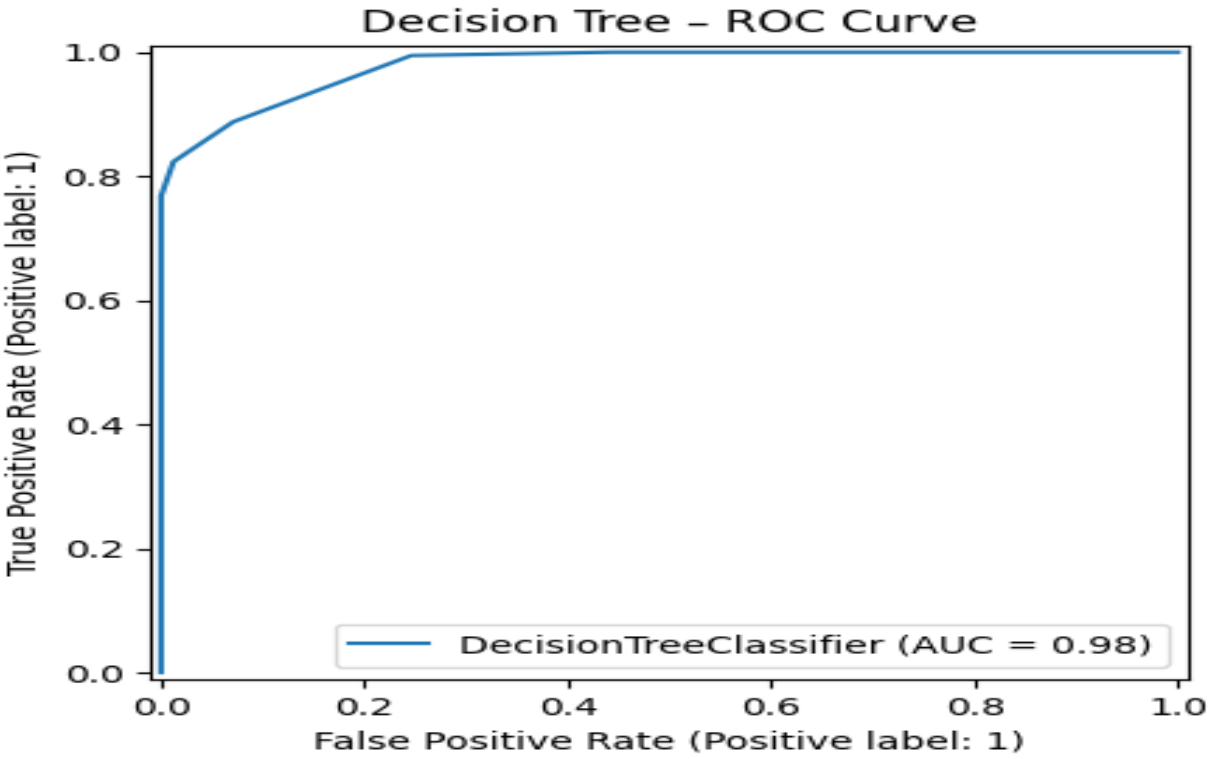
Competing splits: "Tenure in months", "Monthly charge" and "Online security_yes".

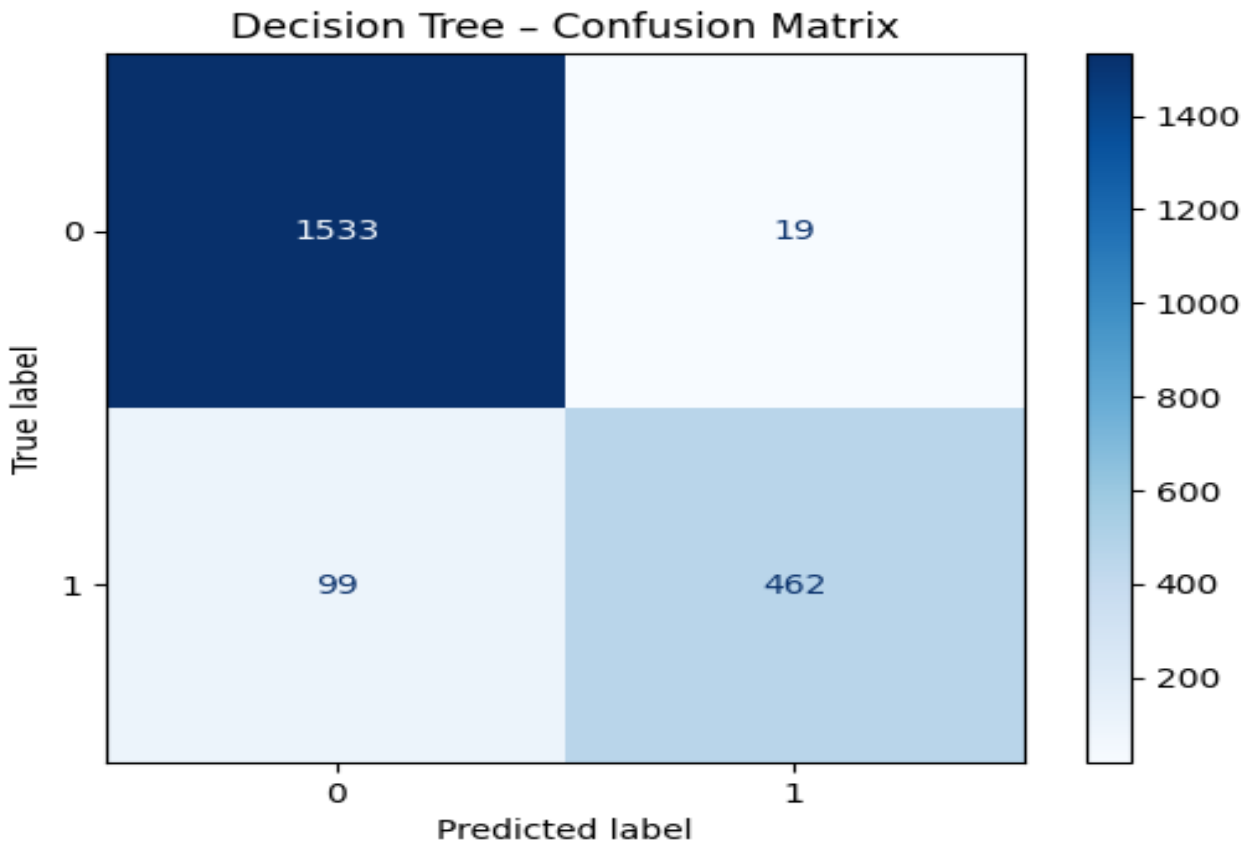
Insights:

This tree balances simplicity and performance, providing a clear understanding of how churn is influenced by Satifaction score, Tenure in months, and Online security.

Using recall: the model correctly identified 82% of customers likely to churn .


Decision Tree – Visualized Structure
satisfaction_score <= 2.5
gini = 0.39
samples = 4930
value = [3622.0, 1308.0]
class = No
gini = 0.0
samples = 1009
value = [0, 1009]
class = Yes
satisfaction_score <= 3.5
gini = 0.141
samples = 3921
value = [3622, 299]
class = No
tenure_in_months <= 9.5
gini = 0.271
samples = 1848
value = [1549, 299]
class = No
gini = 0.0
samples = 2073
value = [2073, 0]
class = No
monthly_charge <= 74.175
gini = 0.483
samples = 394
value = [233.0, 161.0]
class = No
online_security_Yes <= 0.5
gini = 0.172
samples = 1454
value = [1316, 138]
class = No
gini = 0.408
samples = 259
value = [185, 74]
class = No
gini = 0.458
samples = 135
value = [48, 87]
class = Yes
gini = 0.289
samples = 758
value = [625, 133]
class = No
gini = 0.014
samples = 696
value = [691, 5]
class = No


Decision Tree – ROC Curve
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
DecisionTreeClassifier (AUC = 0.98)



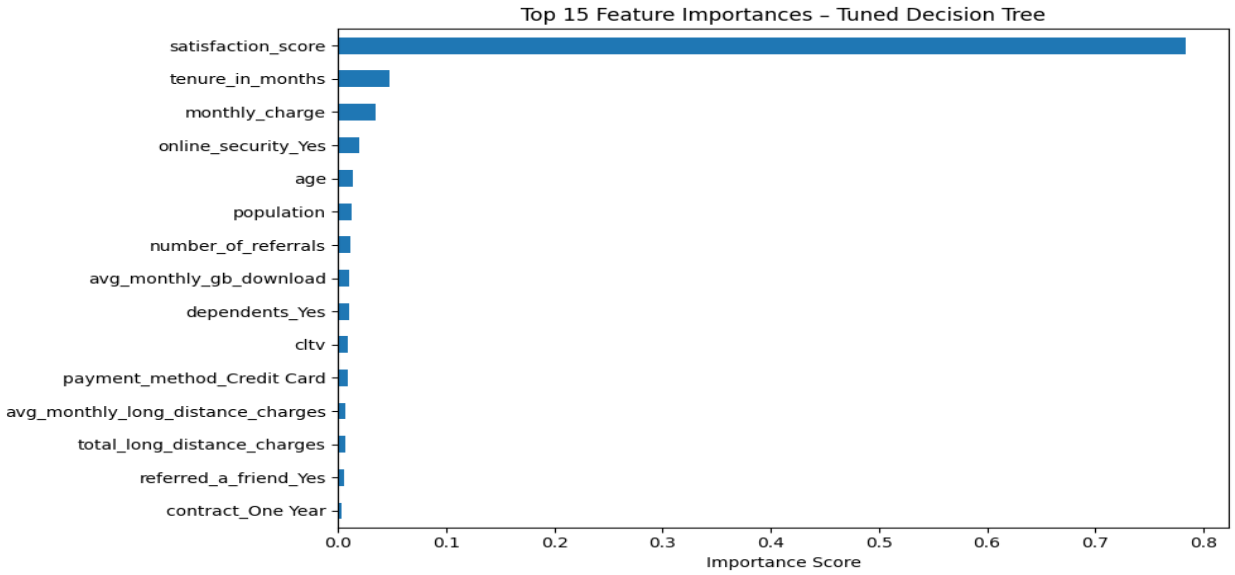
4.3 Model Technique 2 Tuned Decision Tree (base model)

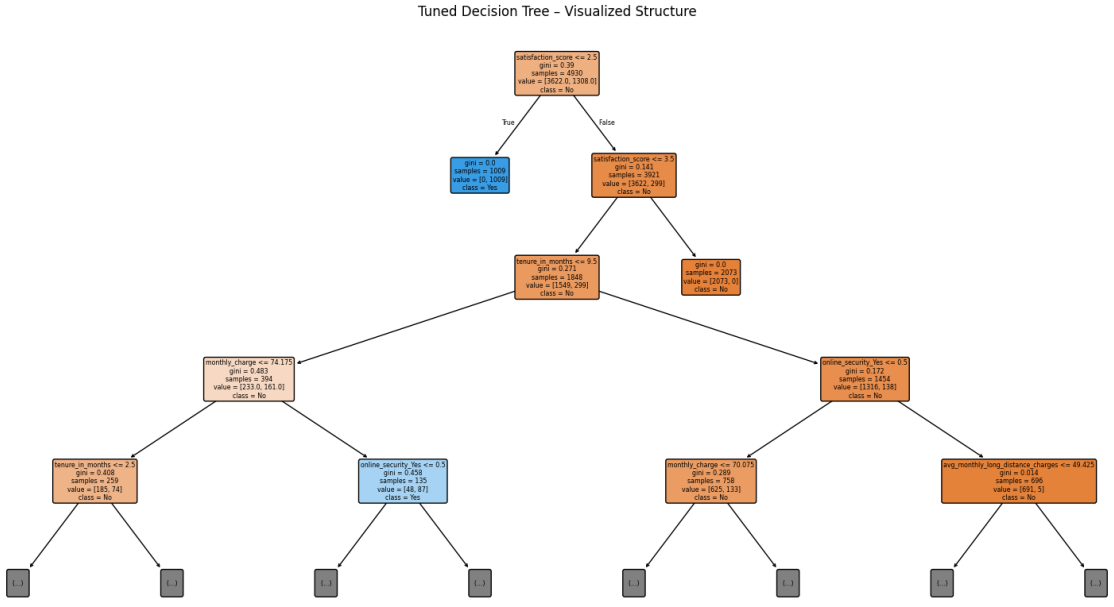
Tuned Decision Tree **Performance**	**Validation**
Recall	0.88
ROC-AUC	0.9469
F1 Score	0.89
Accuracy	0.94
Precision	0.90

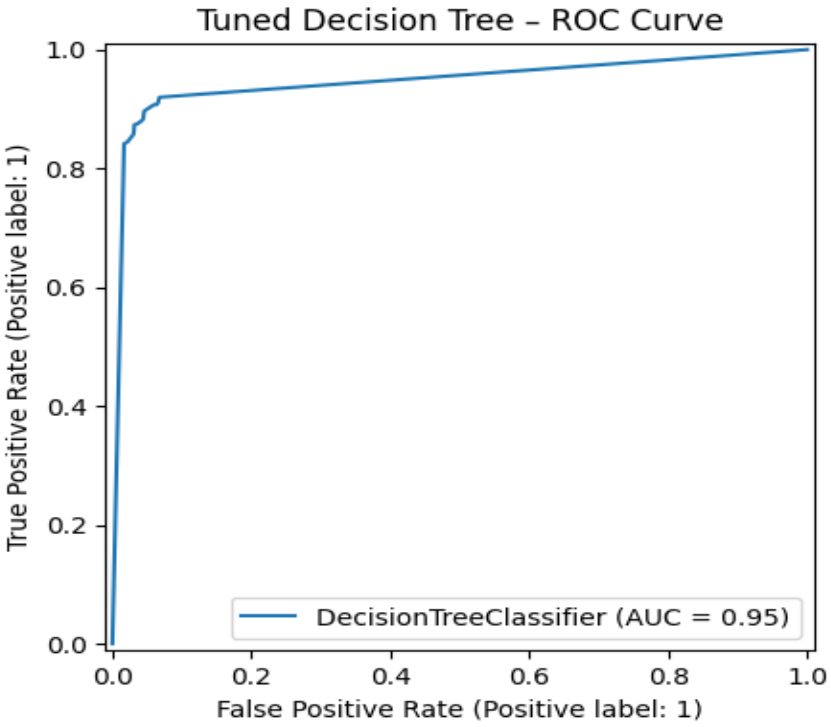
Number of leaves in the tuned decision tree: 120

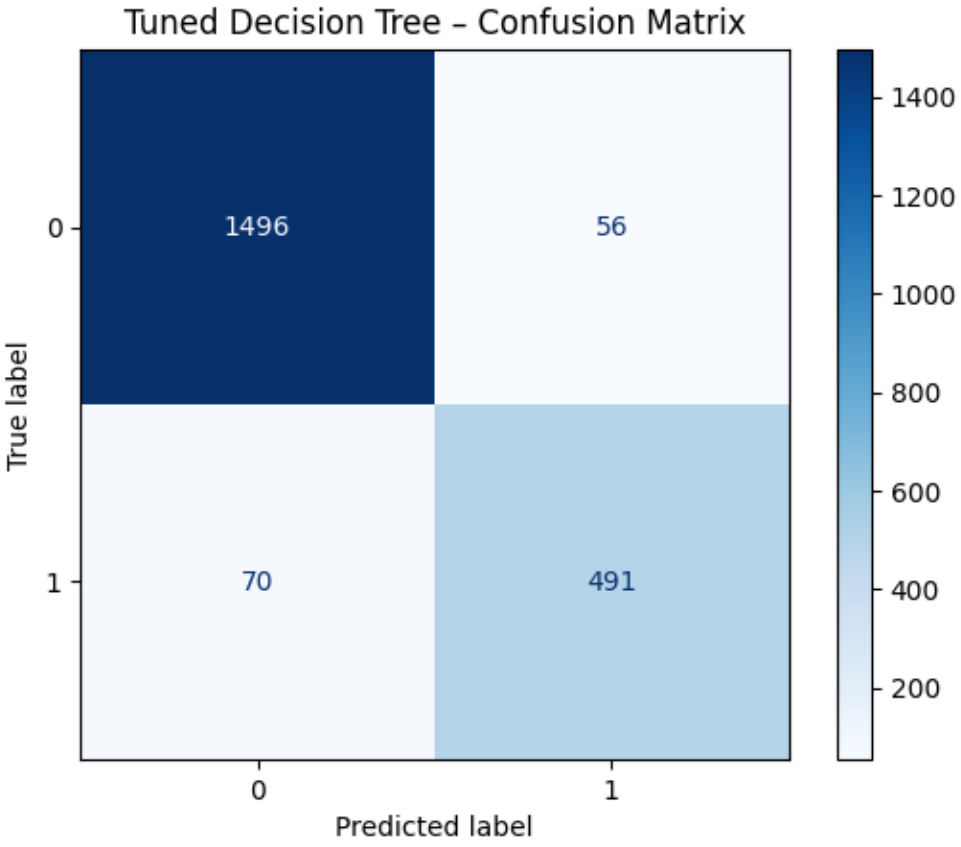
Using recall: the model correctly identified 88% of customers likely to churn .

Top predictors included: satisfaction score, tenure in months, monthly charge etc.






Tuned Decision Tree – ROC Curve
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
DecisionTreeClassifier (AUC = 0.95)
0.0
0.2
0.4
0.6
0.8
1.0


Tuned Decision Tree – Confusion Matrix
True label
Predicted label
1496
56
70
491
0
1
1400
1200
1000
800
600
400
200

4.4 Model Technique 3- Random Forest base model

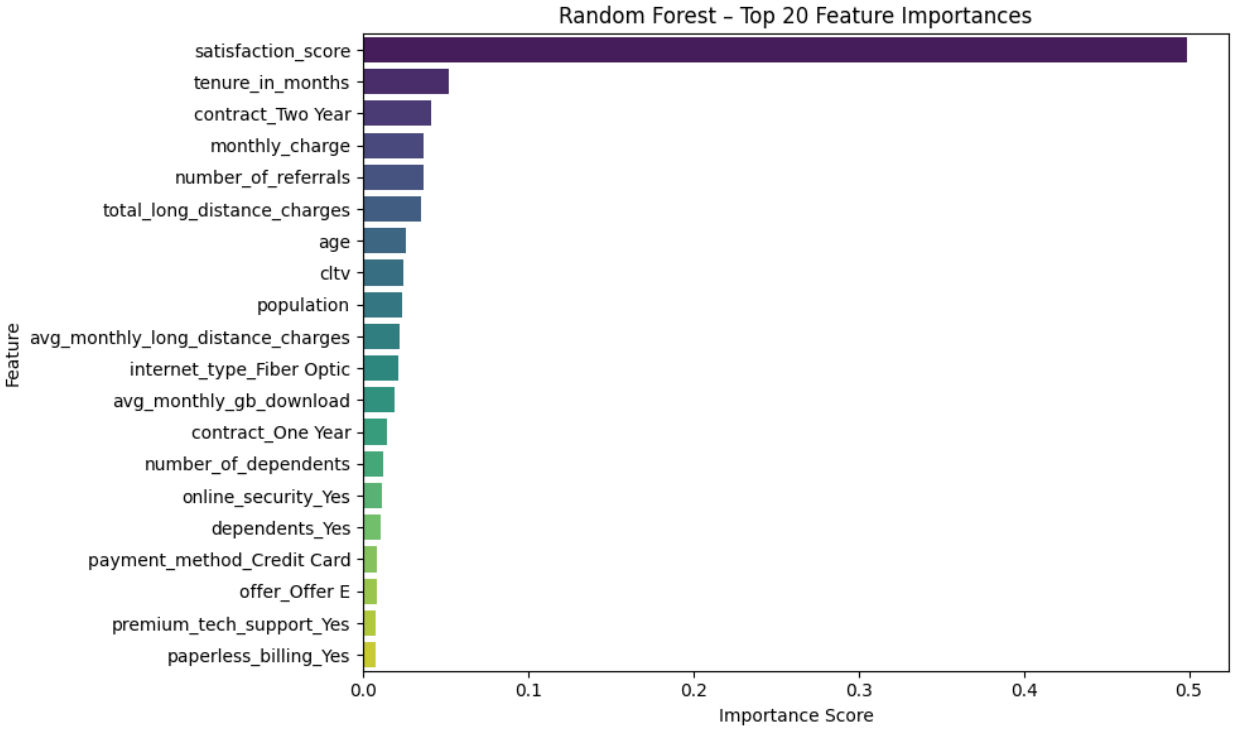
Random Forest, an ensemble method, was used to improve prediction performance while maintaining interpretability. By combining multiple decision trees, it reduced overfitting and captured more nuanced patterns in the data.

Random forest Performance	Validation
Recall	0.8627
ROC-AUC	0.9855
F1 Score	0.9140
Accuracy	0.9569
Precision	0.9718

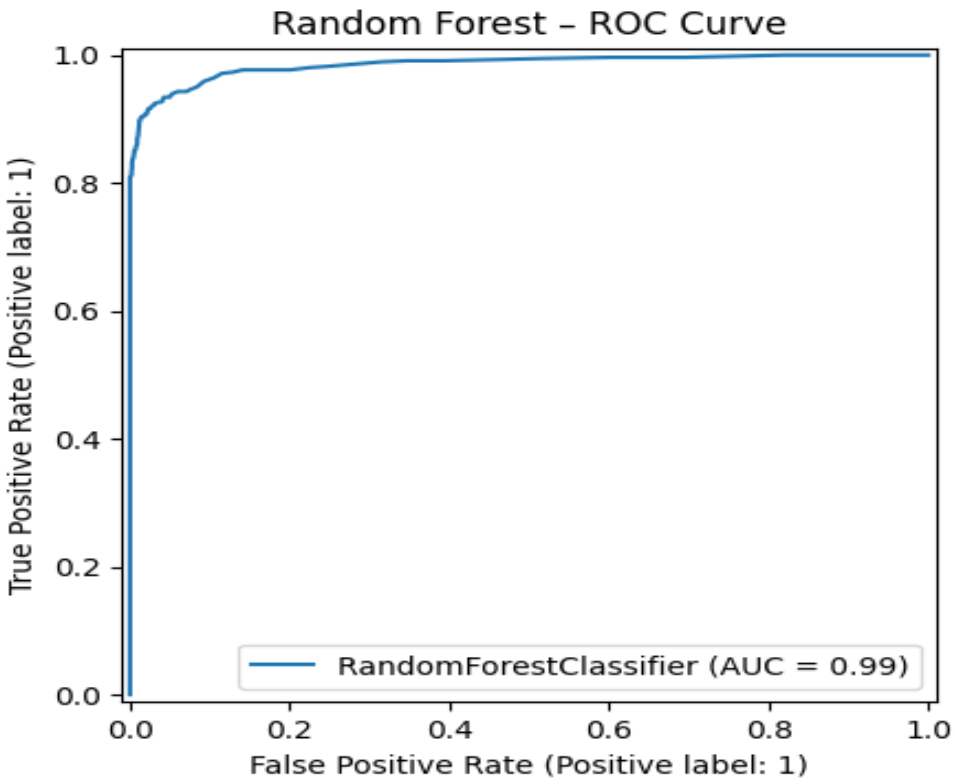
Number of leaves: 100

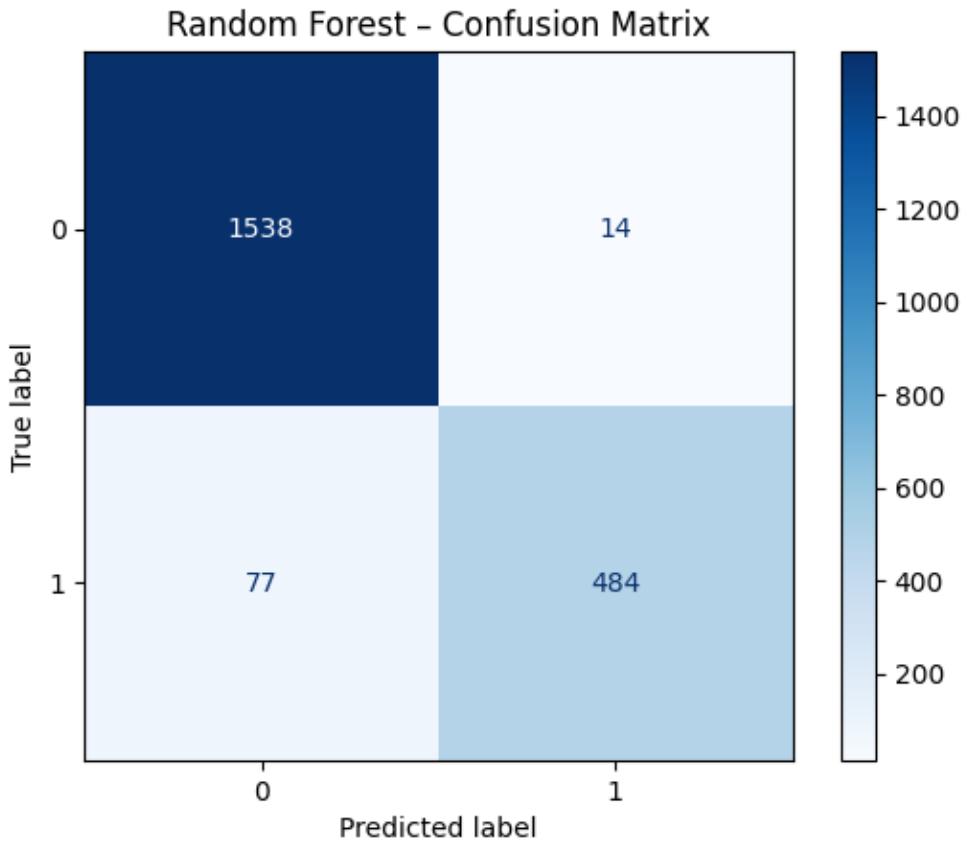
Using recall: the model correctly identified 86% of customers likely to churn .

Top predictors included: satisfaction score, tenure in months, contract_two_year etc.



1. **Satisfaction score** remains the most important feature (score near 0.5), suggesting that customer satisfaction significantly affects churn likelihood.
2. **Tenure in months** is also influential; the length of customer tenure is a strong predictor of churn.
3. "**Contract_two year"** is also highly influential.
4. Features like" **monthly_charge"**, "**number_of_referrals"**, and "**total_long_distance_charges"** have a moderate impact but still affect churn.


Random Forest – ROC Curve
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
RandomForestClassifier (AUC = 0.99)
0.0
0.2
0.4
0.6
0.8
1.0

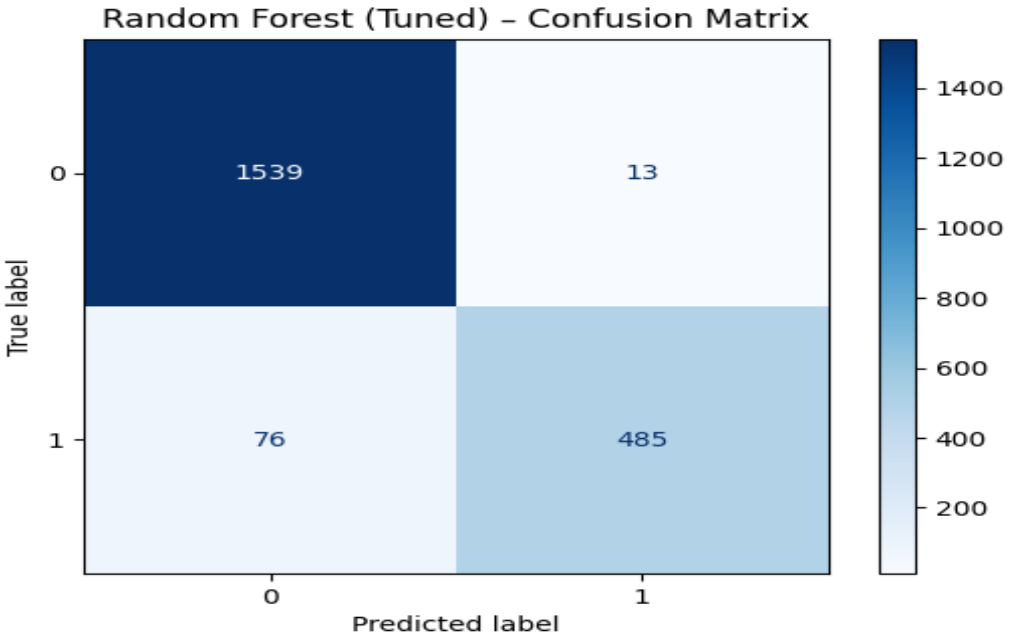


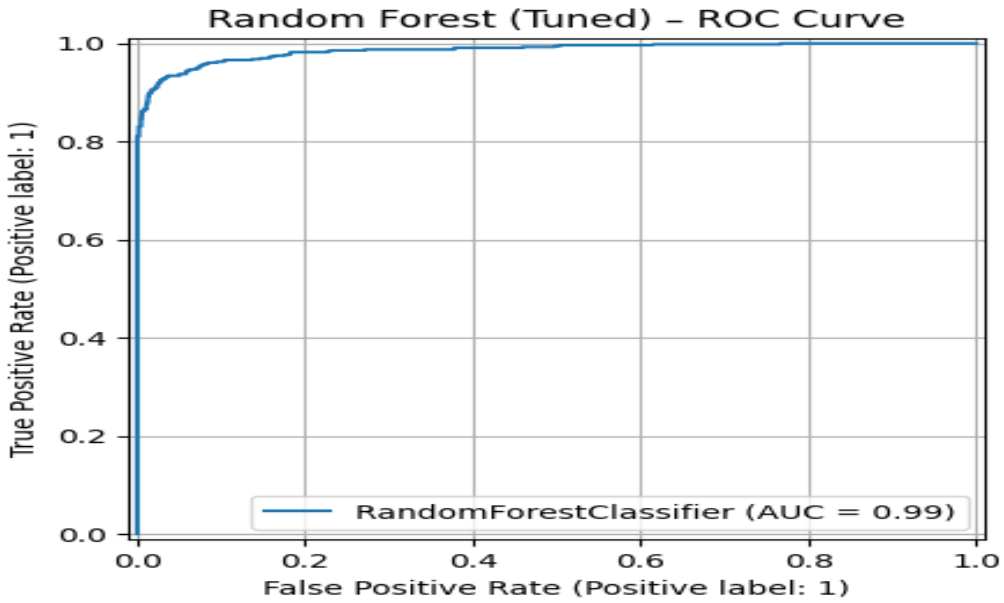
4.5 Model Technique 4- Random Forest tuned

Tuned Random Forest Performance	Validation
Recall	0.8645
ROC-AUC	0.9857
F1 Score	0.9159
Accuracy	0.9578
Precision	0.9738

Number of leaves in the tuned random forest: 314

Using recall: the model correctly identified 86% of customers likely to churn .





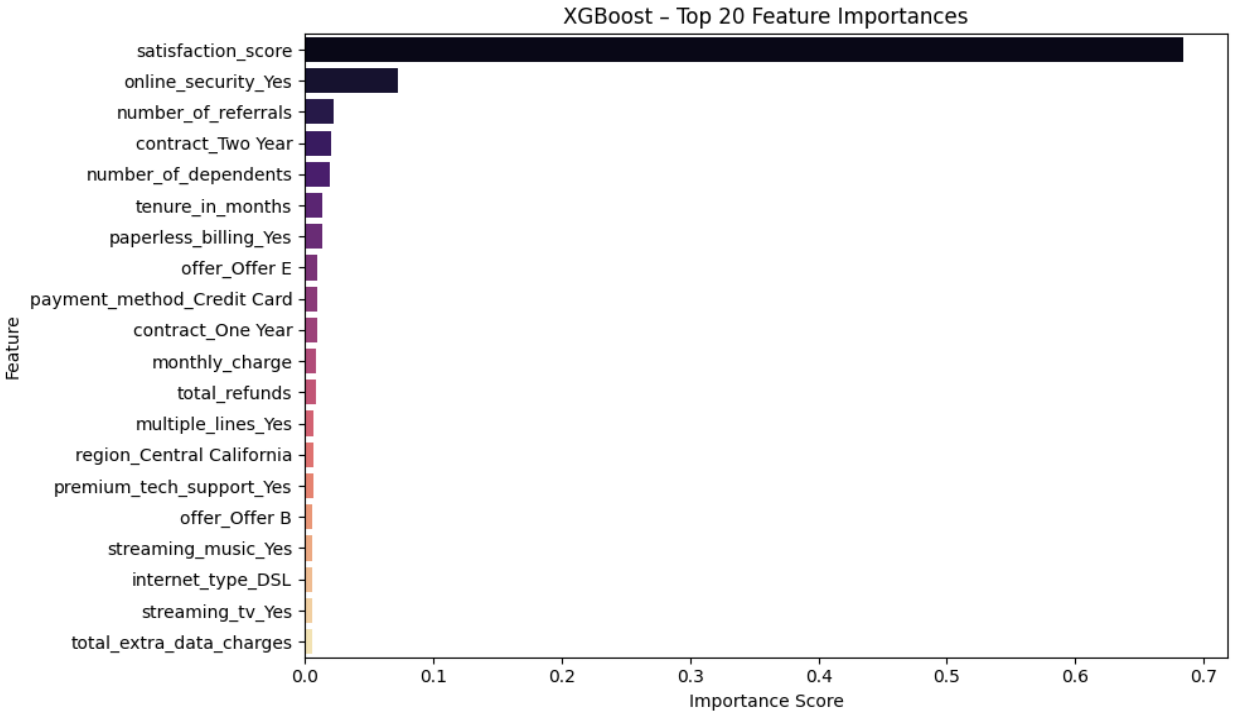
4.6 Model Technique 5- XG Boost base model

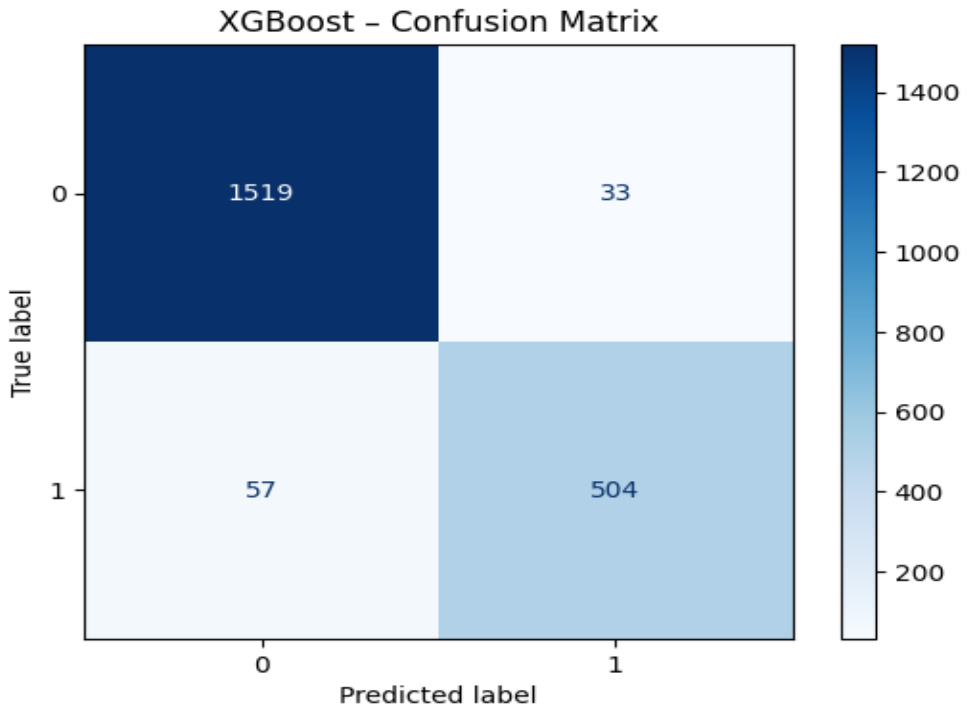
XG Boost Performance	Validation

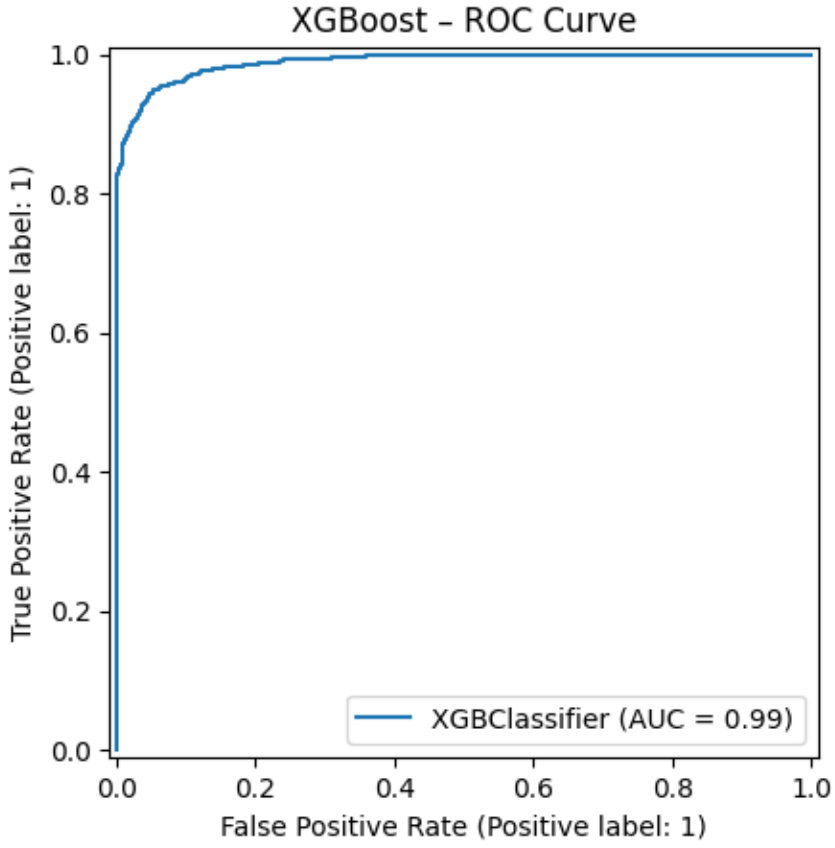
Recall	0.8983
ROC-AUC	0.9857
F1 Score	0.9180
Accuracy	0.9574
Precision	0.9894

Using recall: the model correctly identified 89% of customers likely to churn .

Top predictors included: satisfaction score,online security_yes, number of referrals etc.




XGBoost – Confusion Matrix
1519
33
57
504
0
1
True label
Predicted label
1400
1200
1000
800
600
400
200

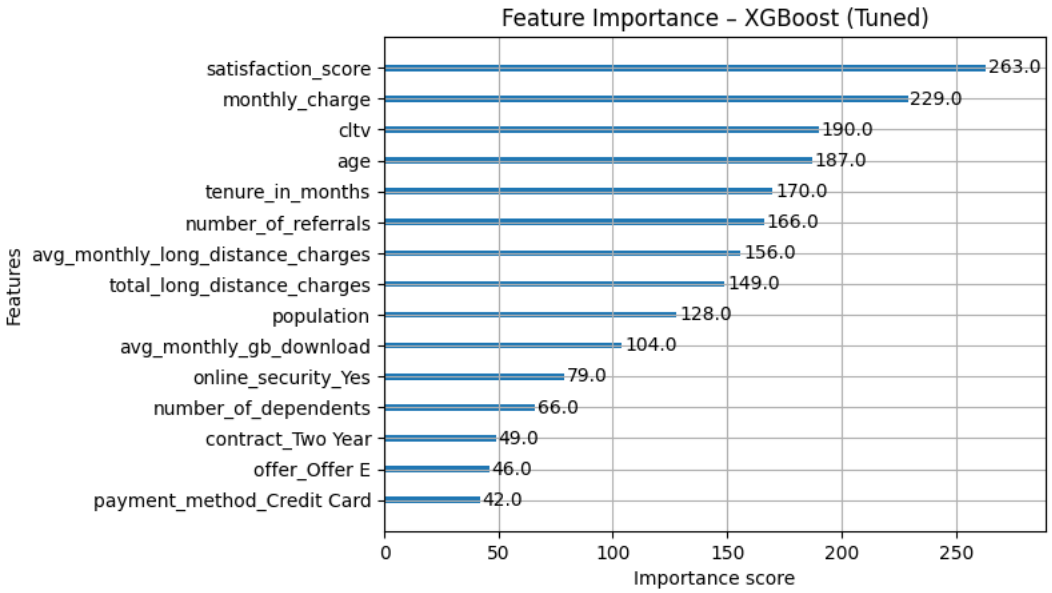

XGBoost – ROC Curve
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
XGBClassifier (AUC = 0.99)
0.0
0.2
0.4
0.6
0.8
1.0

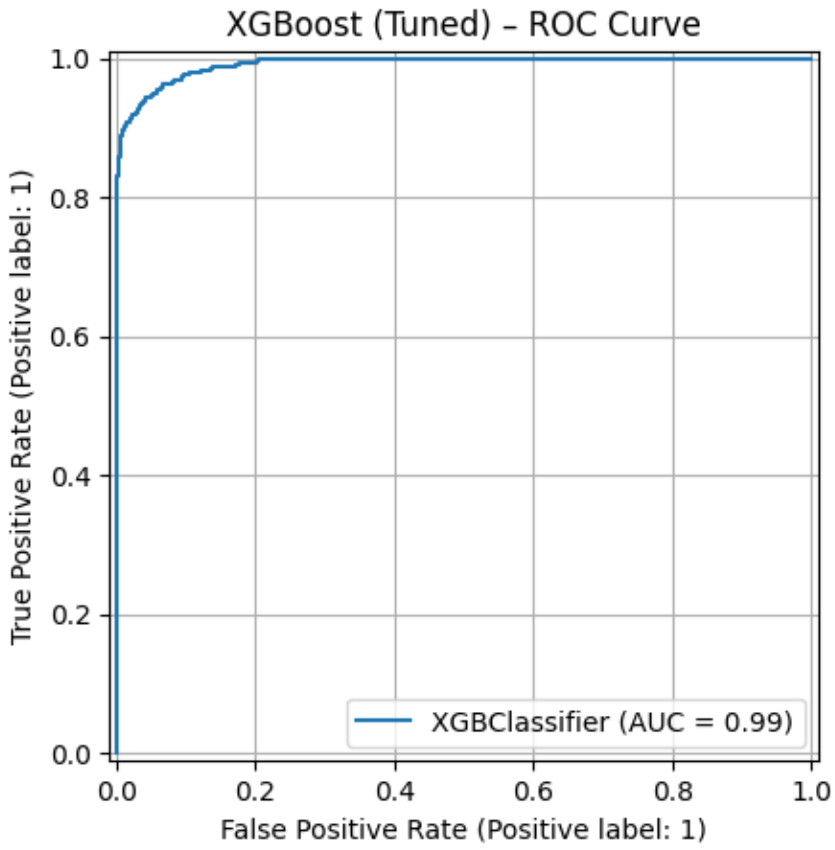
4.7 Model Technique 6 - XG Boost base tuned

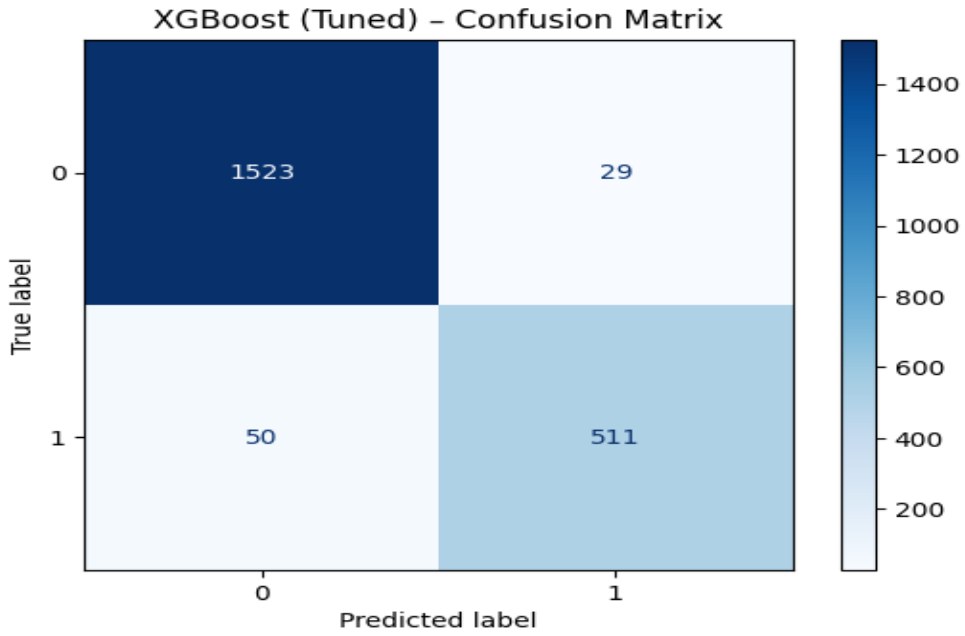
XG Boost Tuned Performance	Validation
Recall	0.9108
ROC-AUC	0.9923
F1 Score	0.9282
Accuracy	0.9626
Precision	0.9462

Using recall: the model correctly identified 91% of customers likely to churn .

Top predictors included: satisfaction score, monthly charge, cltv, age, number of referrals etc.




XGBoost (Tuned) – ROC Curve
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
XGBClassifier (AUC = 0.99)
0.0
0.2
0.4
0.6
0.8
1.0

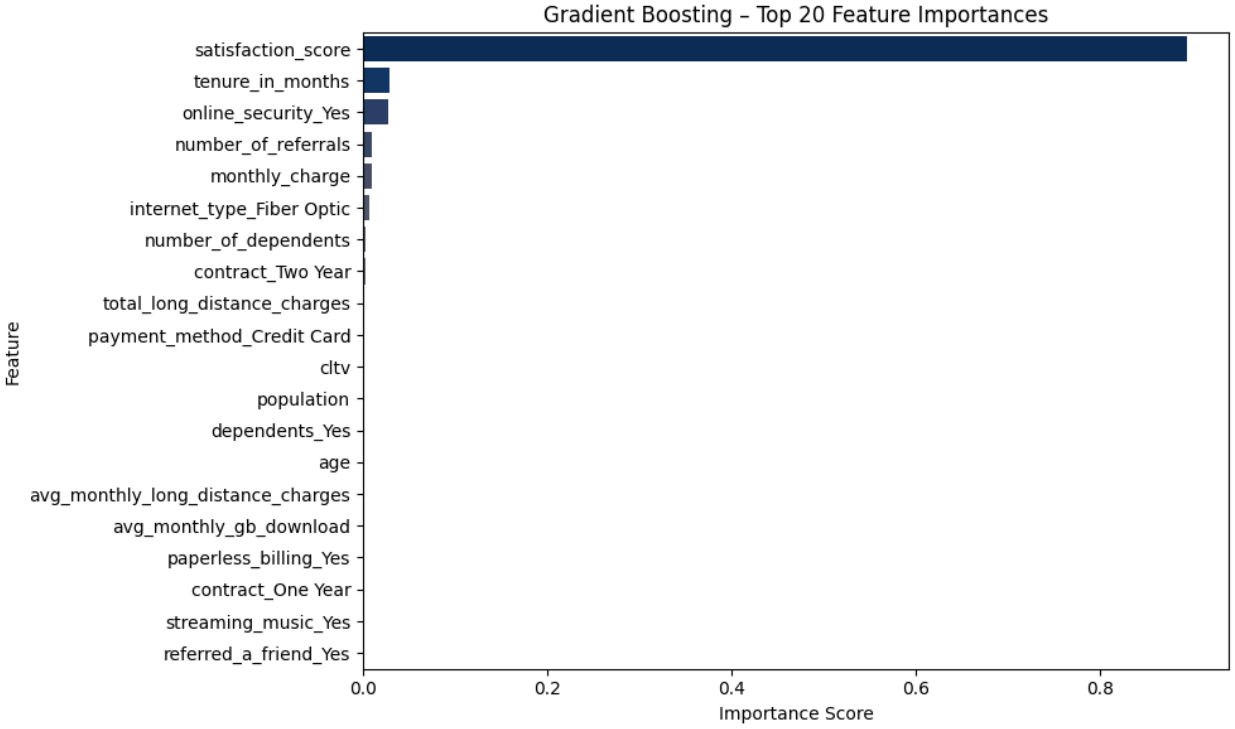

XGBoost (Tuned) – Confusion Matrix
True label
Predicted label
1523
29
50
511
0
1
1400
1200
1000
800
600
400
200

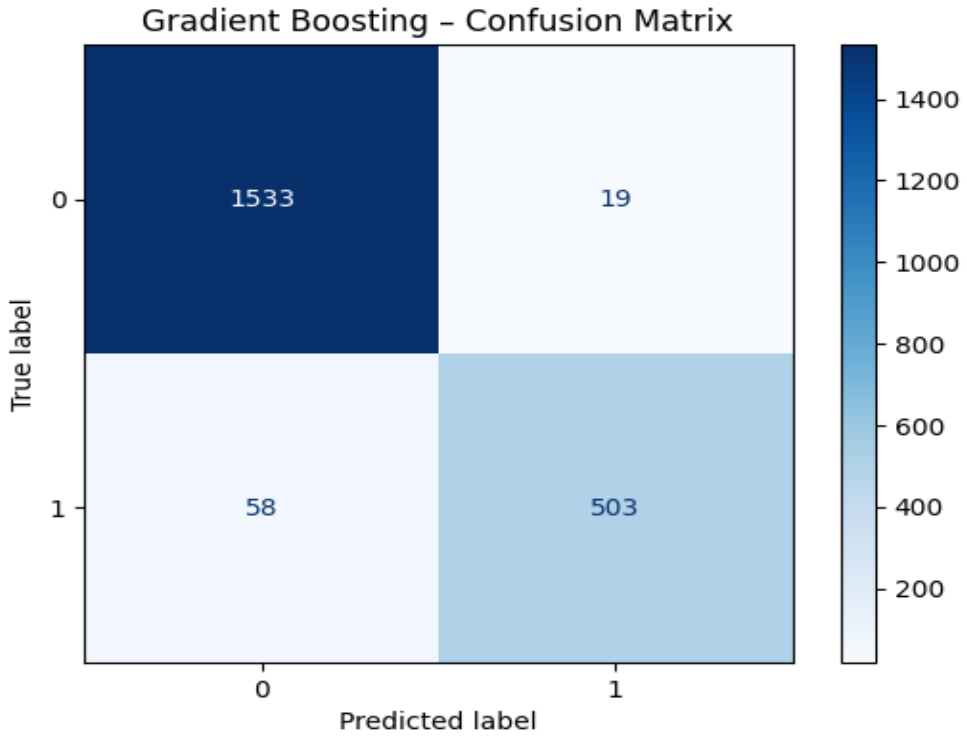
4.8 Model Technique 7- Gradient boosting (base model)

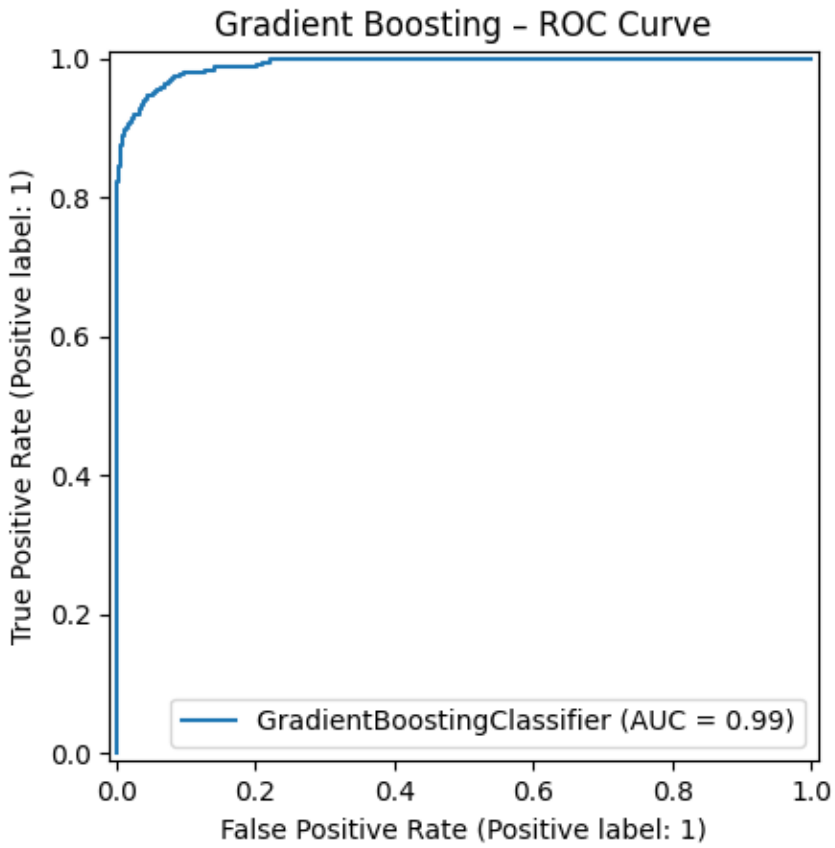
Gradient Boost Performance	Validation
Recall	0.8966
ROC-AUC	0.9920
F1 Score	0.9289
Accuracy	0.9635
Precision	0.9636

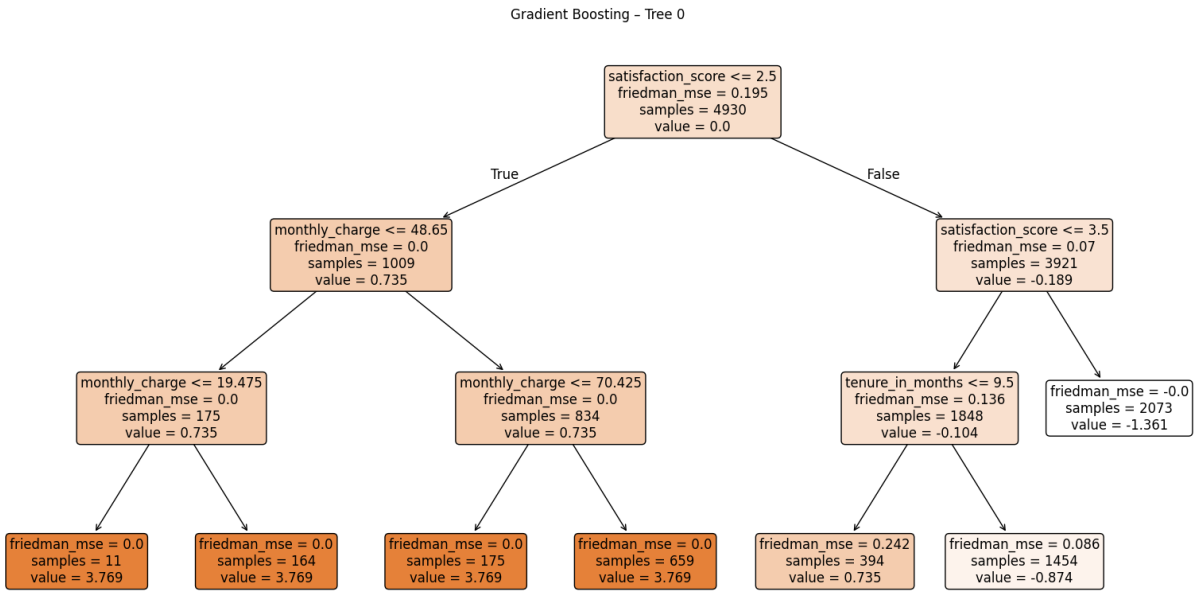
Using recall: the model correctly identified 89% of customers likely to churn .

Top predictors included: satisfaction score, tenure in months, online security_yes etc.




Gradient Boosting – Confusion Matrix
1533
19
58
503
0
1
True label
0
1
Predicted label
1400
1200
1000
800
600
400
200


Gradient Boosting – ROC Curve
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
GradientBoostingClassifier (AUC = 0.99)
0.0
0.2
0.4
0.6
0.8
1.0

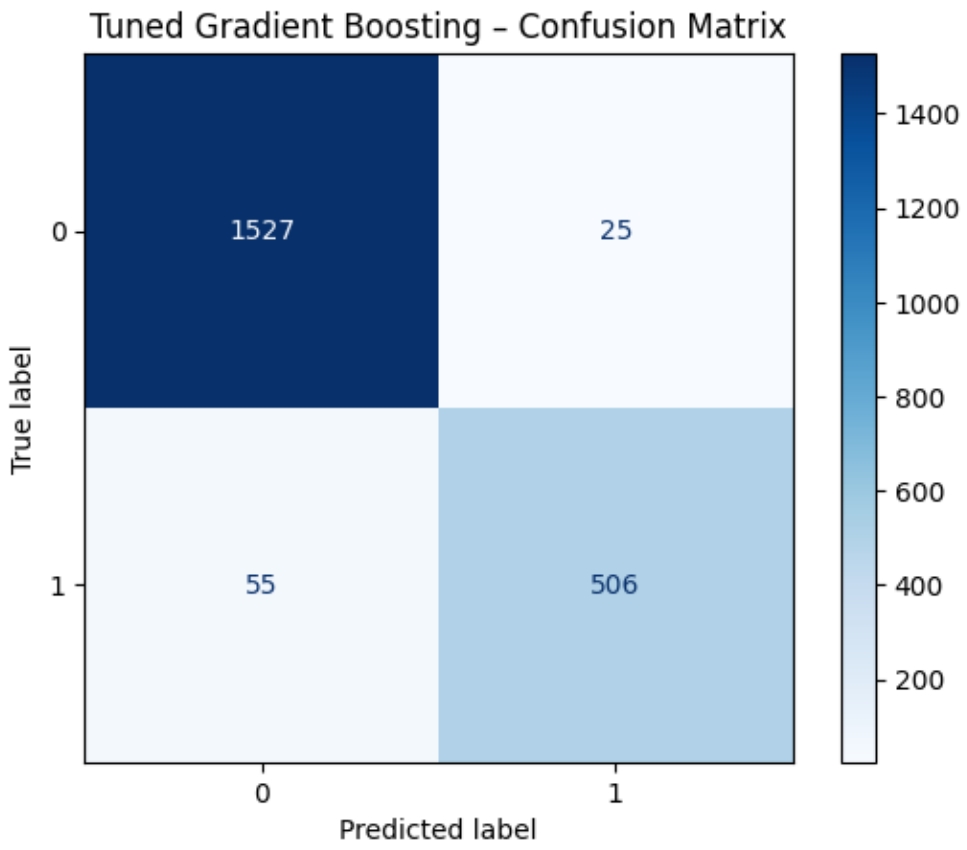

Gradient Boosting – Tree 0
satisfaction_score <= 2.5
friedman_mse = 0.195
samples = 4930
value = 0.0
True
False
monthly_charge <= 48.65
friedman_mse = 0.0
samples = 1009
value = 0.735
satisfaction_score <= 3.5
friedman_mse = 0.07
samples = 3921
value = -0.189
monthly_charge <= 19.475
friedman_mse = 0.0
samples = 175
value = 0.735
monthly_charge <= 70.425
friedman_mse = 0.0
samples = 834
value = 0.735
tenure_in_months <= 9.5
friedman_mse = 0.136
samples = 1848
value = -0.104
friedman_mse = -0.0
samples = 2073
value = -1.361
friedman_mse = 0.0
samples = 11
value = 3.769
friedman_mse = 0.0
samples = 164
value = 3.769
friedman_mse = 0.0
samples = 175
value = 3.769
friedman_mse = 0.0
samples = 659
value = 3.769
friedman_mse = 0.242
samples = 394
value = 0.735
friedman_mse = 0.086
samples = 1454
value = -0.874

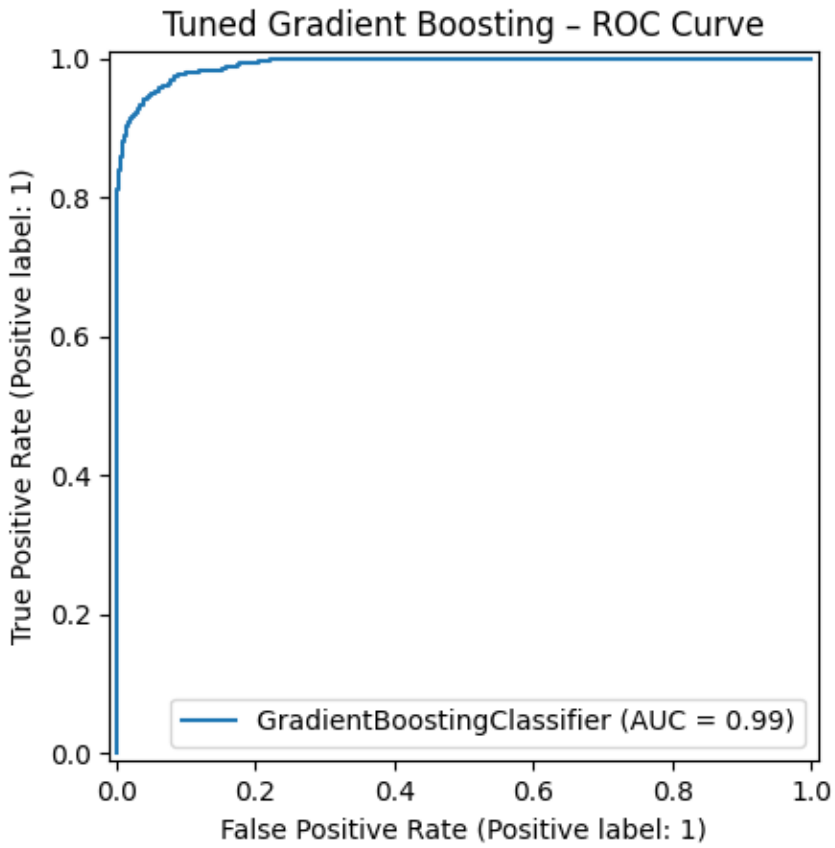
4.9 Model Technique 8- Gradient Boosting Tuned

Gradient Boost Tuned Performance	Validation
Recall	0.9019
ROC-AUC	0.9920
F1 Score	0.9267
Accuracy	0.9621
Precision	0.9529

Using recall: the model correctly identified 90% of customers likely to churn .

Top predictors included: satisfaction score, tenure in months, monthly charge etc.

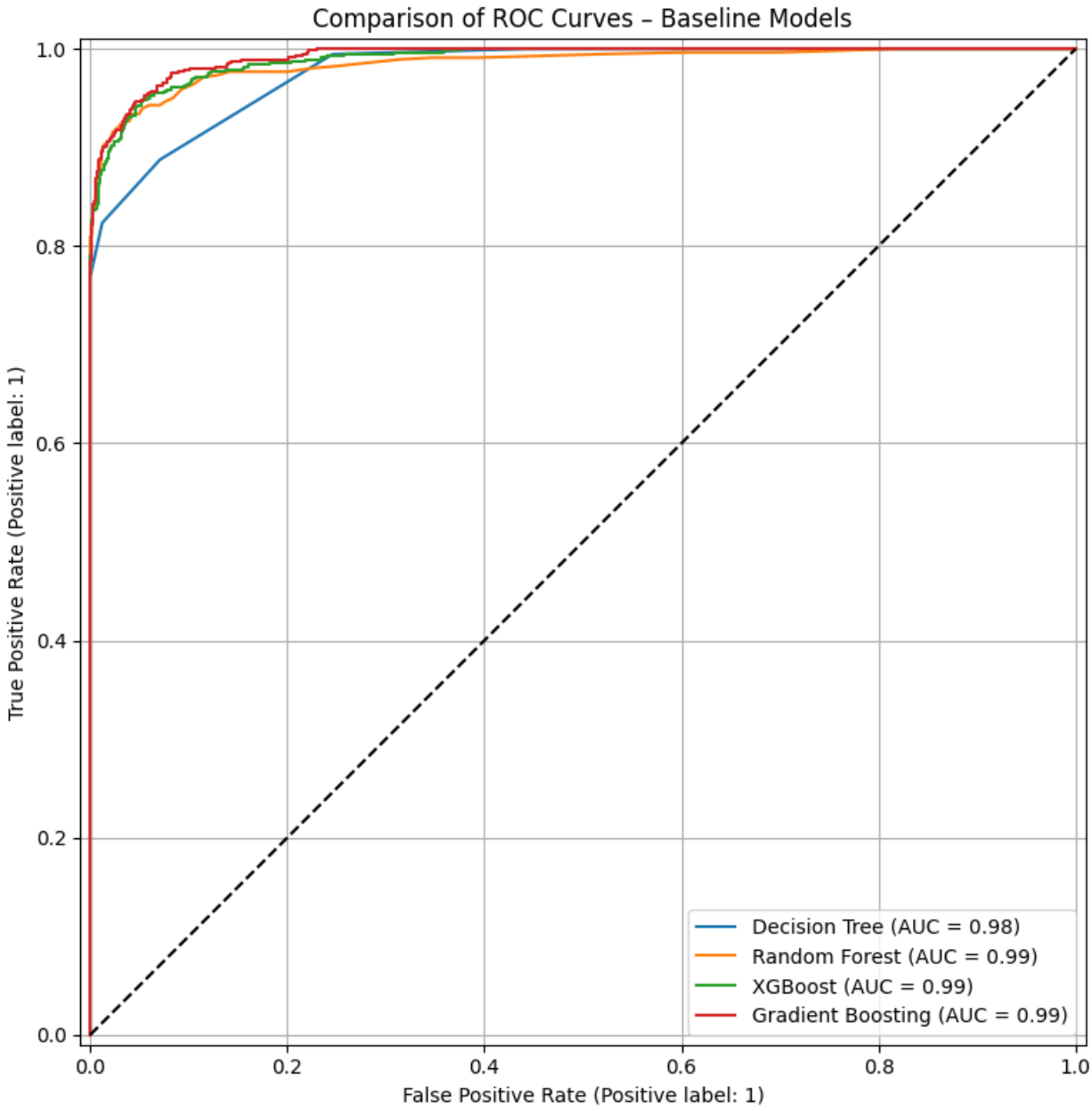

Tuned Gradient Boosting – Confusion Matrix
1527
25
55
506
0
1
True label
0
1
Predicted label
1400
1200
1000
800
600
400
200



Tree Model Comparison

Base trees

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC-AUC**
XGBoost	**0. 9574**	**0. 9385**	**0.8983**	**0.9180**	**0. 9894**
Decision Tree	0. 9441	0. 9604	0.8235	0. 8867	0. 9782
Random Forest	0. 9569	0. 9718	0.8627	0. 9140	0. 9855
Gradient Boosting	0. 9635	0. 9636	0.8966	0. 9289	0. 9920



Tuned Tree models

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC

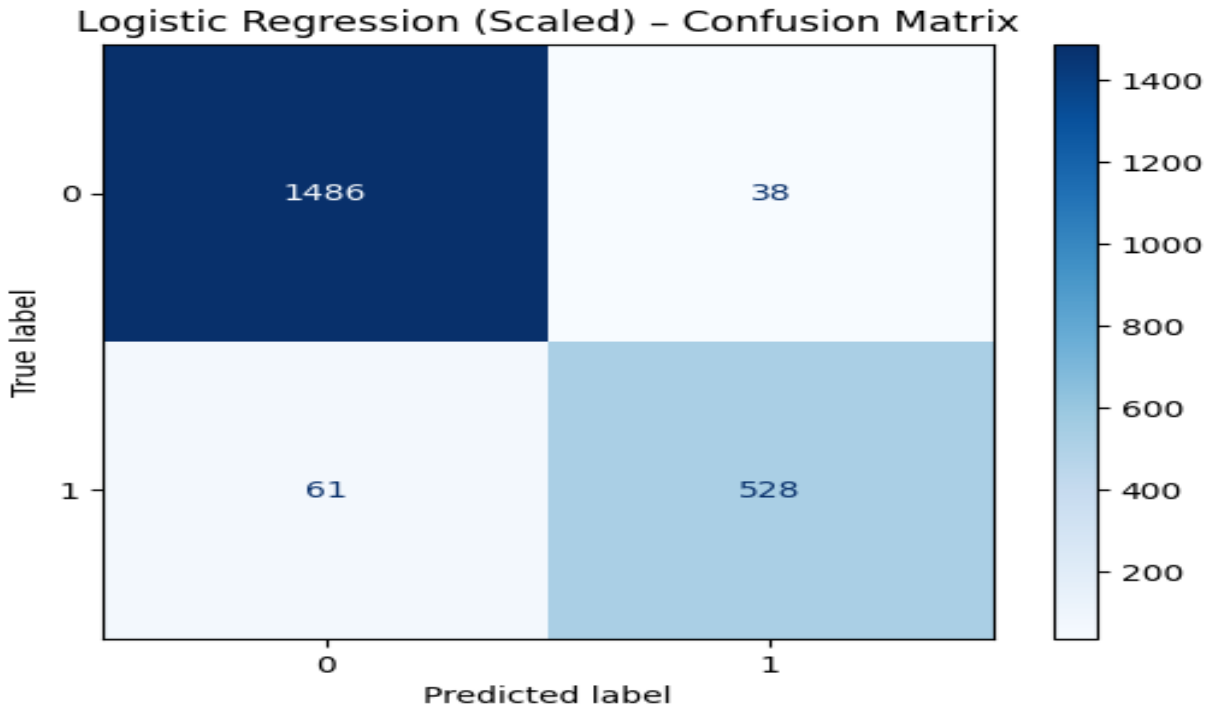
XGBoost	**0.9626**	**0.9462**	**0.9108**	**0.9282**	**0.9923**
Decision Tree	0.9403	0.8976	0.8752	0.8862	0.9468
Random Forest	0.9578	0.9738	0.8645	0.9159	0.9857
Gradient Boosting	0.9621	0.9529	0.9019	0.9267	0.9920

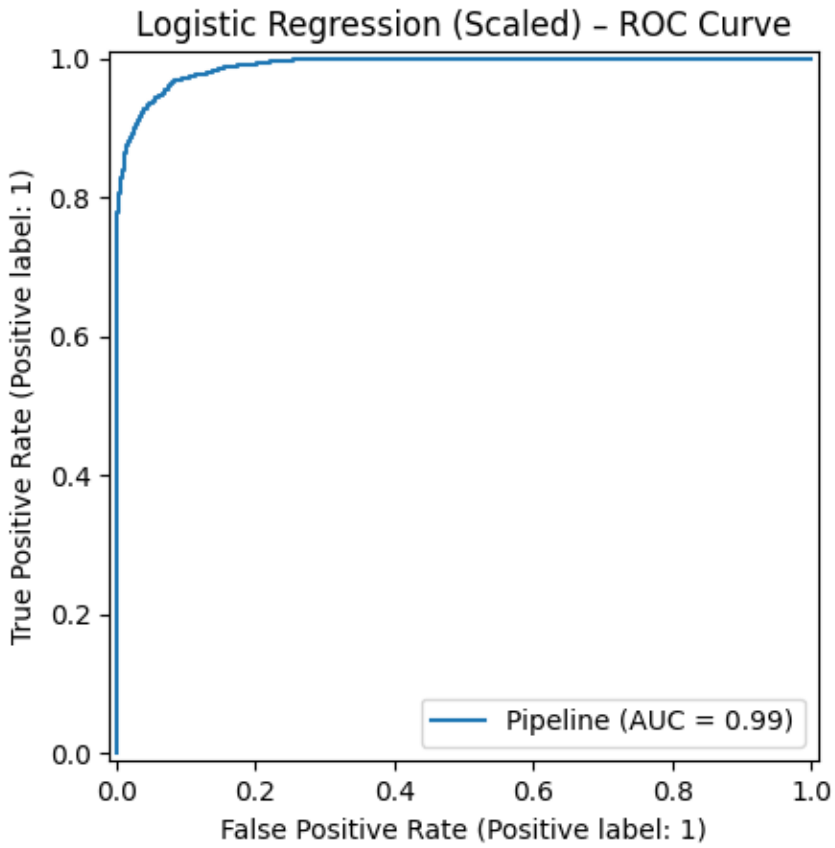
XGBoost should be the preferred model for deployment because it provides a strong balance between precision and recall, minimizes false positives and false negatives, and shows the highest discriminatory power. This combination ensures cost-effective churn prevention strategies and maximizes the ROI on retention campaigns.

5.0 Model Technique 9- Logistic Regression (base model)

Logistic Regression Performance	Validation
Recall	0.90
ROC-AUC	0.98
F1 Score	0.91
Accuracy	0.95
Precision	0.93

Using recall: the model correctly identified 90% of customers likely to churn .

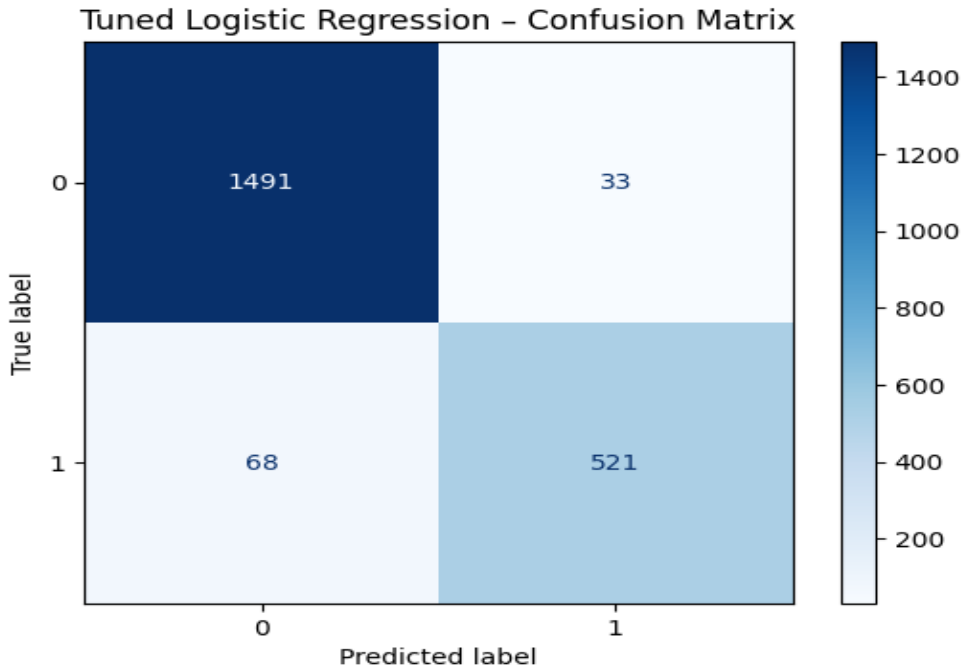


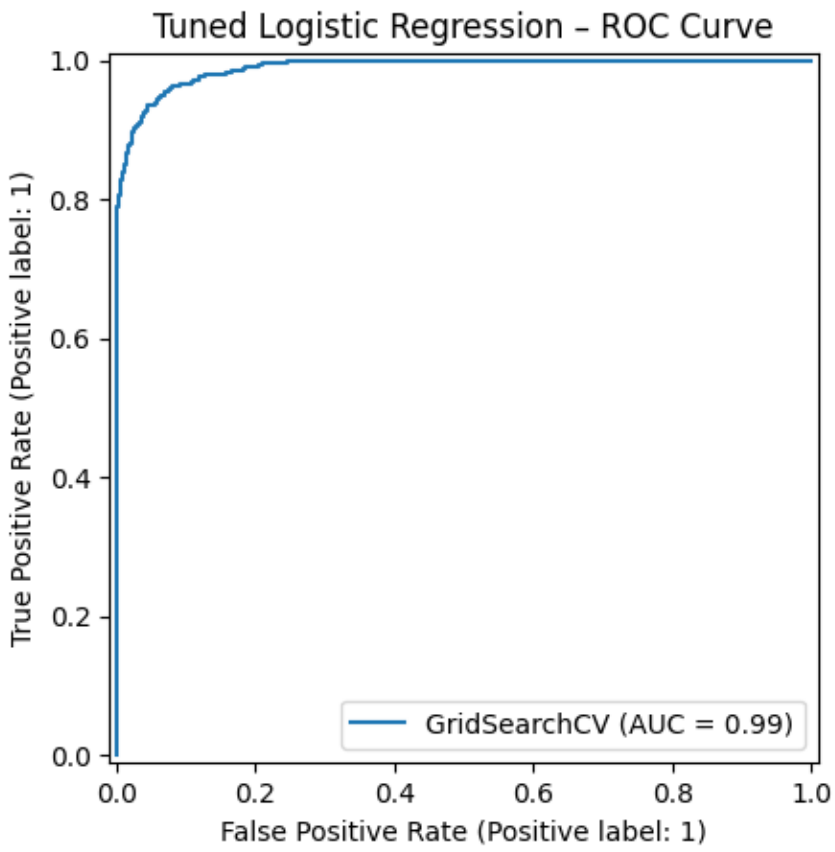


5.1 Model Technique 10- Logistic Regression (tuned model)

Tuned Logistic Regression Performance	Validation
Recall	0.88
ROC-AUC	0.98
F1 Score	0.91
Accuracy	0.95
Precision	0.94

Using recall: the model correctly identified 88% of customers likely to churn .


Tuned Logistic Regression – Confusion Matrix
1491
33
68
521
0
1
True label
0
1
Predicted label
1400
1200
1000
800
600
400
200


Tuned Logistic Regression – ROC Curve
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
GridSearchCV (AUC = 0.99)
0.0
0.2
0.4
0.6
0.8
1.0

5.2 Model Technique 11 Forward Regression model

Forward Regression Performance	Validation
Recall	0.89
ROC-AUC	0.98
F1 Score	0.91
Accuracy	0.95
Precision	0.93

Forward regression was utilized to refine our model by sequentially adding variables based on their statistical significance to optimize model complexity.

Using recall: the model correctly identified 89% of customers likely to churn .

Top predictors included: satisfaction score, tenure in months, monthly charge etc.

Forward Selected Features: 25 Variables

'age', 'number_of_dependents', 'number_of_referrals', 'tenure_in_months', 'avg_monthly_gb_download', 'monthly_charge', 'total_extra_data_charges', 'total_long_distance_charges', 'satisfaction_score', 'dependents_Yes', 'referred_a_friend_Yes', 'offer_Offer B', 'offer_Offer C', 'offer_no_offer', 'internet_service_Yes', 'internet_type_no_internet', 'online_security_Yes', 'device_protection_plan_Yes', 'unlimited_data_Yes', 'contract_Two Year', 'payment_method_Credit Card', 'payment_method_Mailed Check', 'region_Central California', 'region_Northern California', 'region_Southern California'.

5.3 Model Technique 12- Tuned Forward Regression model

Forward Regression Performance	Validation
Recall	0.89
ROC-AUC	0.9884
F1 Score	0.91
Accuracy	0.95
Precision	0.93

Using recall: the model correctly identified 89% of customers likely to churn .

Top predictors included: satisfaction score, tenure in months, monthly charge etc.

5.4 Model Technique 13- Backward Regression model

Forward Regression Performance	Validation
Recall	0.89
ROC-AUC	0.9891
F1 Score	0.91
Accuracy	0.95
Precision	0.93

Backward regression was modeled to remove the least significant variable to optimize model complexity.

Using recall: the model correctly identified 89% of customers likely to churn .

Backward Features: 17 variables

'age', 'number_of_referrals', 'tenure_in_months', 'avg_monthly_long_distance_charges', 'satisfaction_score', 'dependents_Yes', 'referred_a_friend_Yes', 'multiple_lines_Yes', 'internet_service_Yes', 'internet_type_DSL', 'internet_type_Fiber Optic', 'online_security_Yes', 'online_backup_Yes', 'streaming_movies_Yes', 'contract_One Year', 'contract_Two Year', 'payment_method_Credit Card'.

5.5 Model Technique 14- Tuned Backward Regression model

Forward Regression Performance	Validation
Recall	0.89
ROC-AUC	0.9891
F1 Score	0.91
Accuracy	0.95
Precision	0.93

Using recall: the model correctly identified 89% of customers likely to churn.

5.6 Model Technique 15- Stepwise Regression model

Forward Regression Performance	Validation
Recall	0.88
ROC-AUC	0.9893
F1 Score	0.91

Accuracy	0.95
Precision	0.93

Stepwise regression was used to iteratively add and remove variables based on their statistical significance and contribution to the model. This approach helped to identify the optimal set of predictors, balancing model complexity, and predictive accuracy.

Stepwise Selected Features: 31 variables

['number_of_referrals', 'tenure_in_months', 'avg_monthly_long_distance_charges', 'avg_monthly_gb_download', 'total_refunds', 'total_extra_data_charges', 'total_long_distance_charges', 'satisfaction_score', 'cltv', 'dependents_Yes', 'referred_a_friend_Yes', 'offer_Offer B', 'offer_Offer C', 'offer_Offer D', 'offer_Offer E', 'offer_no_offer', 'multiple_lines_Yes', 'internet_service_Yes', 'internet_type_DSL', 'internet_type_Fiber Optic', 'internet_type_no_internet', 'online_security_Yes', 'online_backup_Yes', 'device_protection_plan_Yes', 'premium_tech_support_Yes', 'streaming_movies_Yes', 'unlimited_data_Yes', 'contract_One Year', 'contract_Two Year', 'payment_method_Credit Card', 'region_Northern California']

5.7 Model Technique 14- Tuned Stepwise Regression model

Forward Regression Performance	**Validation**
Recall	0.88
ROC-AUC	0.9893

F1 Score	0.91
Accuracy	0.95
Precision	0.93

Using recall: the model correctly identified 88% of customers likely to churn .

Logistic Regression Model Comparison

Base models

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC-AUC**
Forward Regression	0.95	0.93	0.89	0.91	0.9891
Backward Regression	0.9517	0.9310	0.8930	0.9116	0.9891
Stepwise Regression	0.95	0.93	0.88	0.91	0.9893
Full Logistic Regression	0.95	0.93	0.90	0.91	0.98

Tuned models

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC-AUC**
Forward Regression	0.9498	0.9259	0.8913	0.9083	0.9885

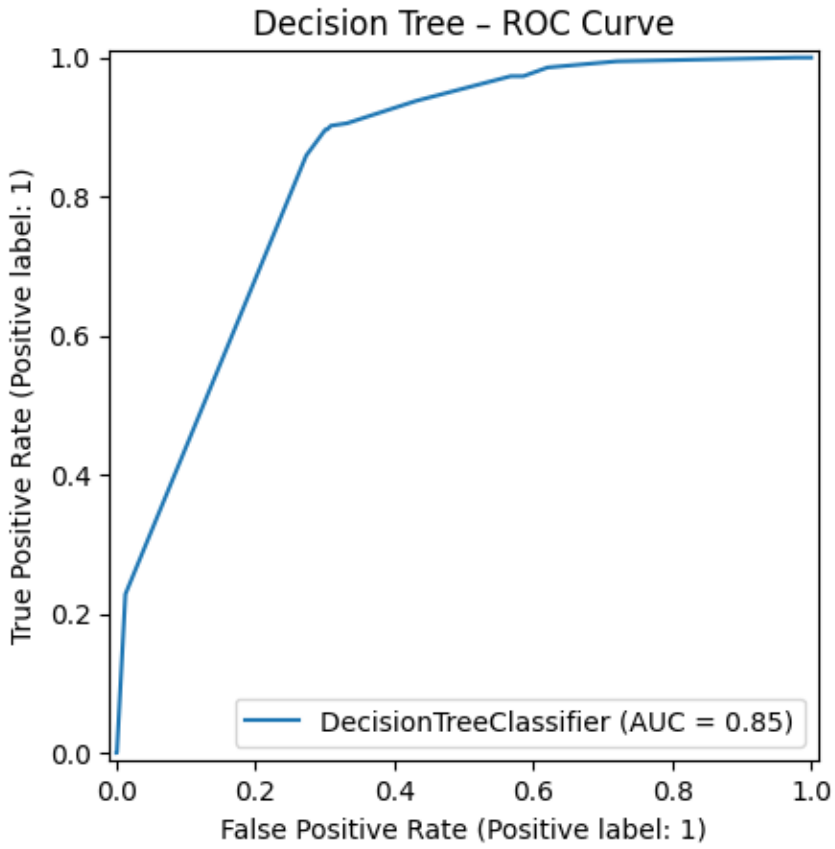
Backward Regression	0.9517	0.9310	0.8930	0.9116	0.9891
Stepwise Regression	0.9484	0.9286	0.8829	0.9051	0.9894
Full Logistic Regression	0.9522	0.9404	0.8846	0.9116	0.9897

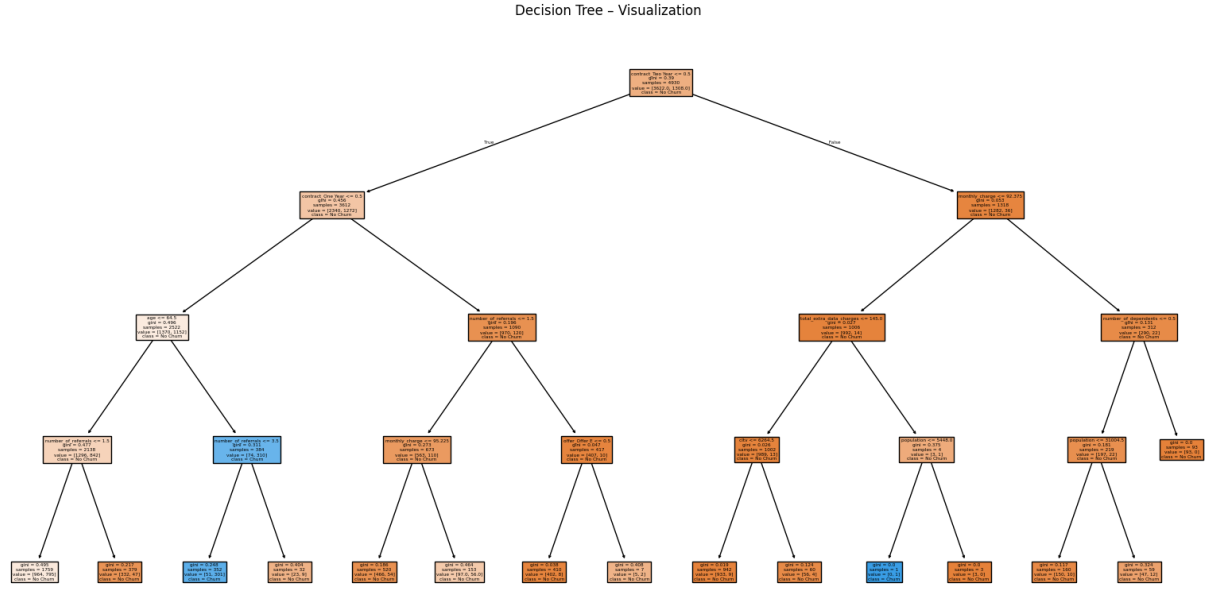
With Recall as the priority, Backward Regression was best; this model ensures fewer churners slip through the cracks, making it ideal for retention strategies where missing a potential churner is costlier than mistakenly targeting a non-churner. The small trade-off in precision compared to the Full Logistic Regression model is acceptable since the business goal is to identify as many true churners as possible for proactive engagement.

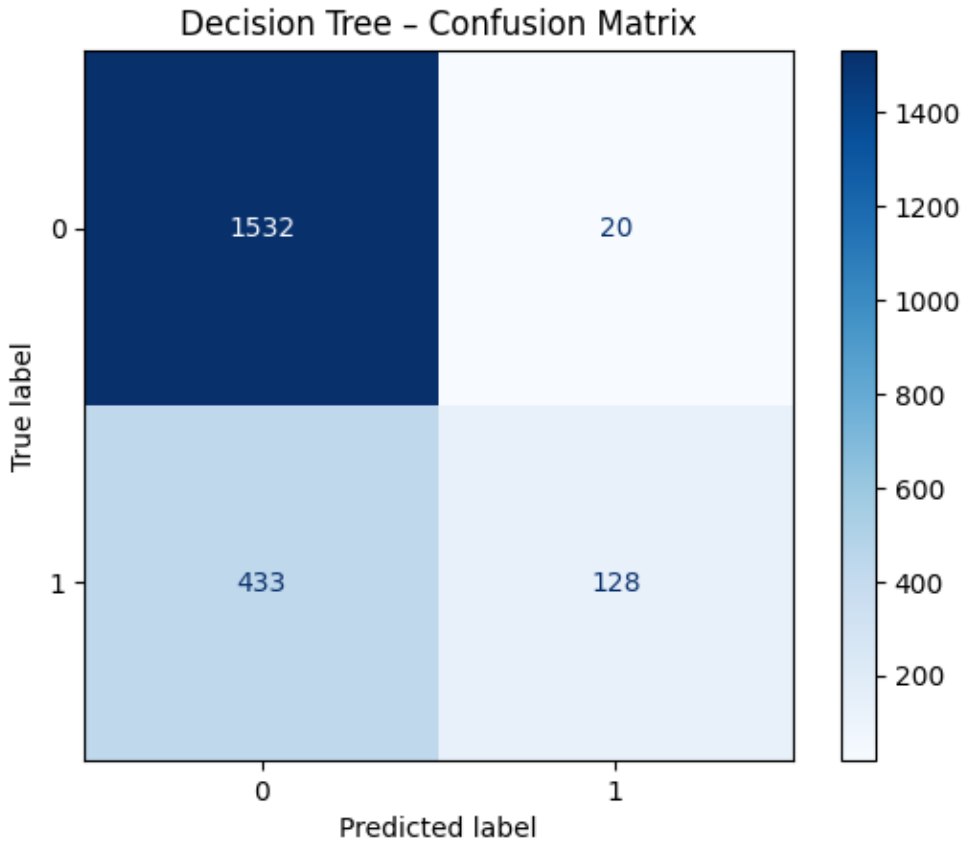
5.9 Model 15 (without satisfaction score)- Decision Tree

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Decision Tree	0.9403	0.8649	0.2282	0.3611	0.8466

Using recall: the model correctly identified 22% of customers likely to churn.




Decision Tree – Visualization

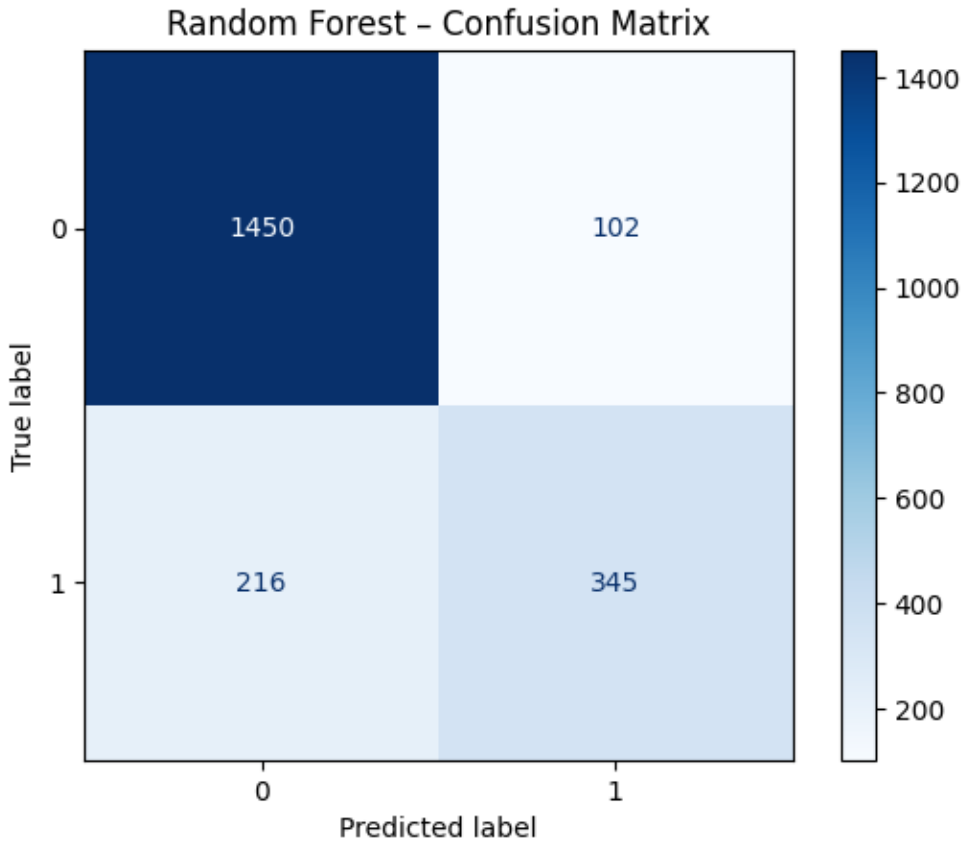

Decision Tree – Confusion Matrix
True label
Predicted label
0
1
1532
20
433
128
1400
1200
1000
800
600
400
200

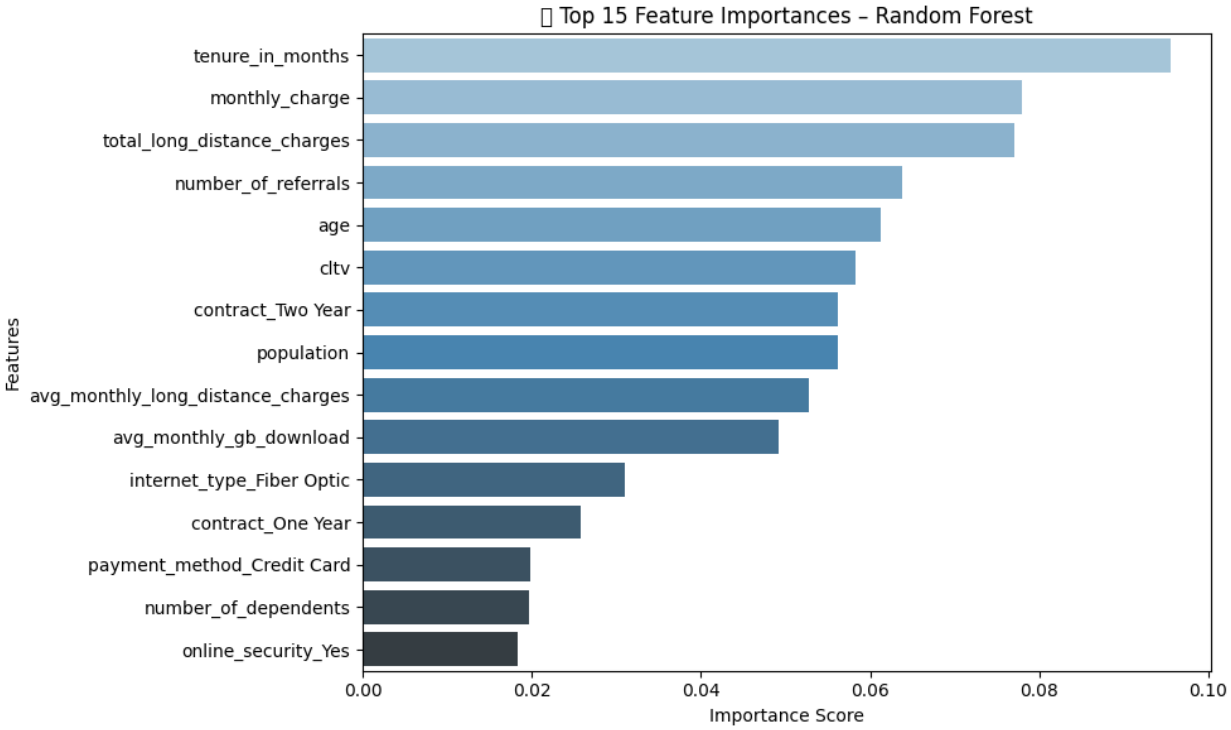
6.0 Model 16 (without satisfaction score)- Random Forest

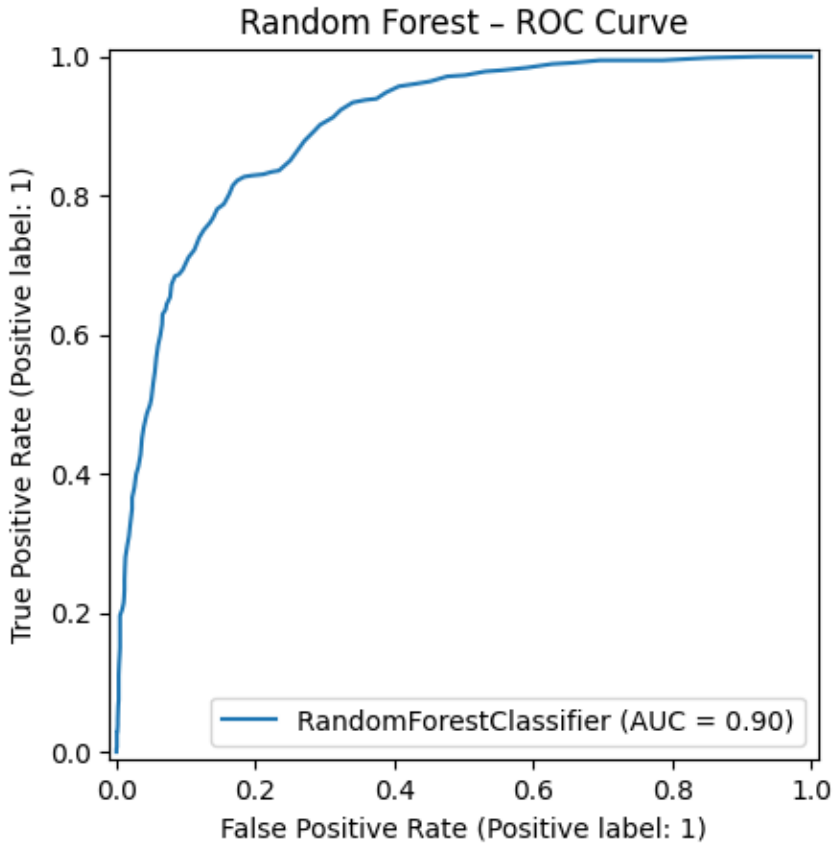
Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Random Forest	0.8495	0.7718	0.6150	0.6845	0.8990

Using recall: the model correctly identified 61% of customers likely to churn .

Top predictors: tenure in months, monthly charge, total long distance charge, number of referals, cltv etc.




Top 15 Feature Importances – Random Forest
tenure_in_months
monthly_charge
total_long_distance_charges
number_of_referrals
age
cltv
contract_Two Year
population
avg_monthly_long_distance_charges
avg_monthly_gb_download
internet_type_Fiber Optic
contract_One Year
payment_method_Credit Card
number_of_dependents
online_security_Yes
Features
0.00
0.02
0.04
0.06
0.08
0.10
Importance Score

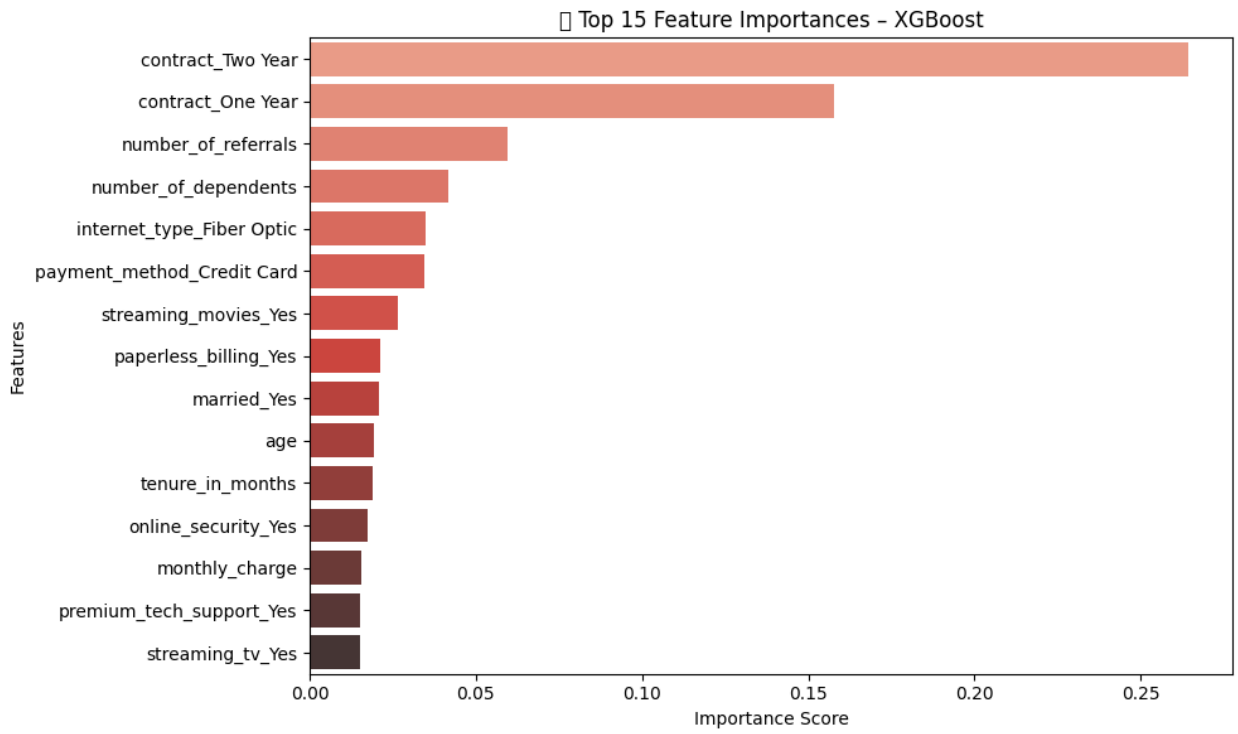

Random Forest – ROC Curve
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
RandomForestClassifier (AUC = 0.90)

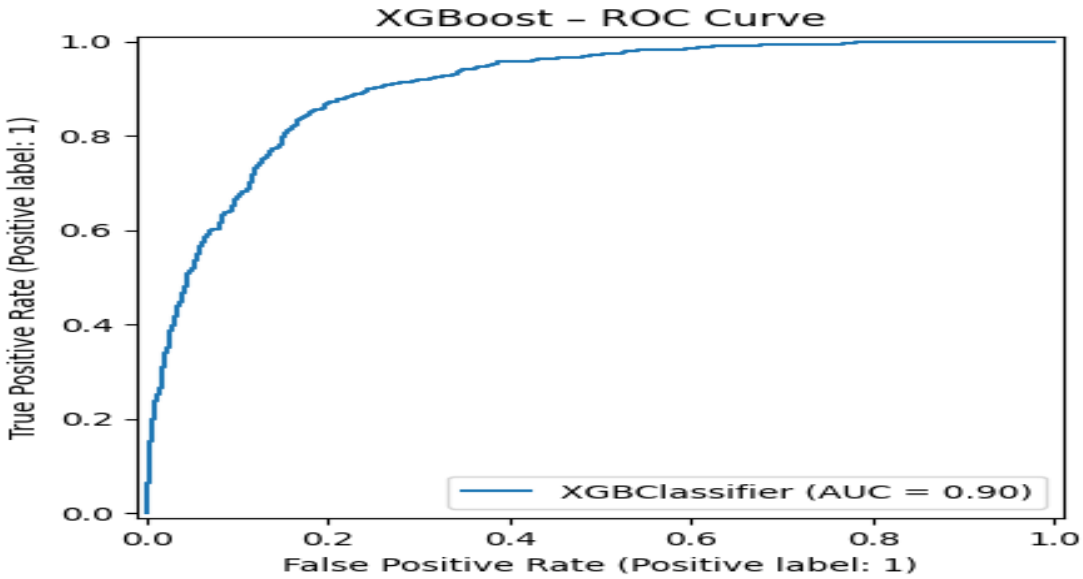
6.1 Model 17 (without satisfaction score)- XGBOOST

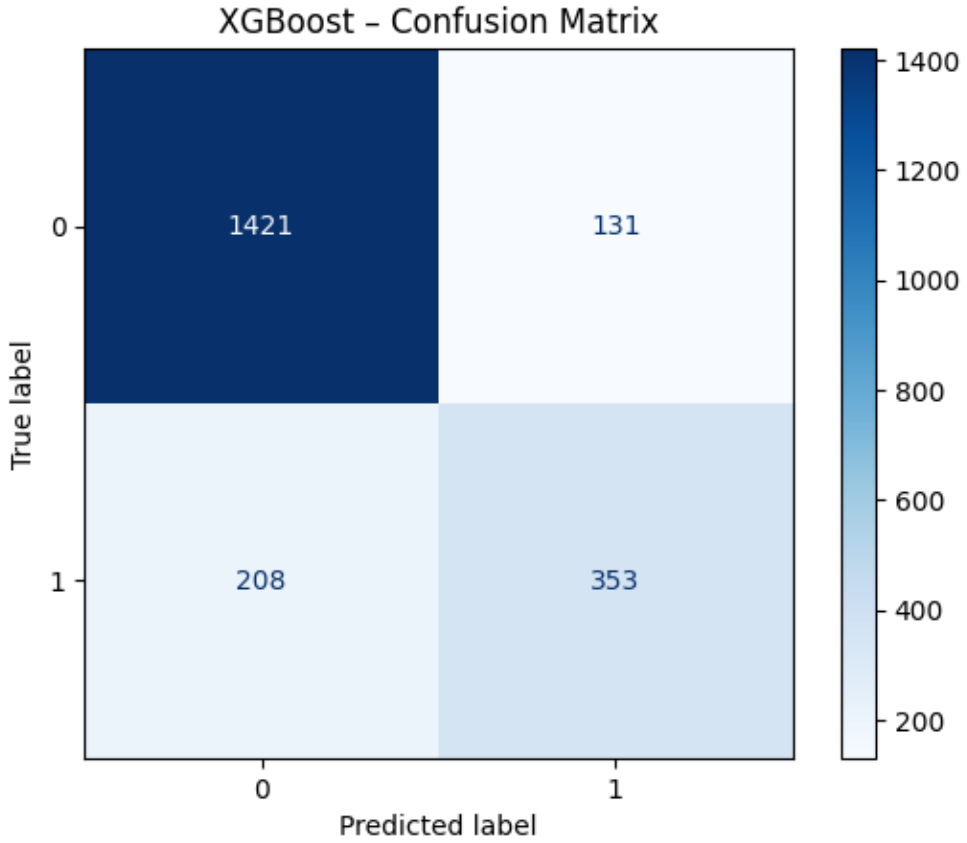
Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
XGBoost	0.8396	0.7293	0.6292	0.6756	0.9024

Using recall: the model correctly identified 63% of customers likely to churn .

Top predictors: contract two year, one-year, number of referrals, number of dependents, fiber optic internet etc.

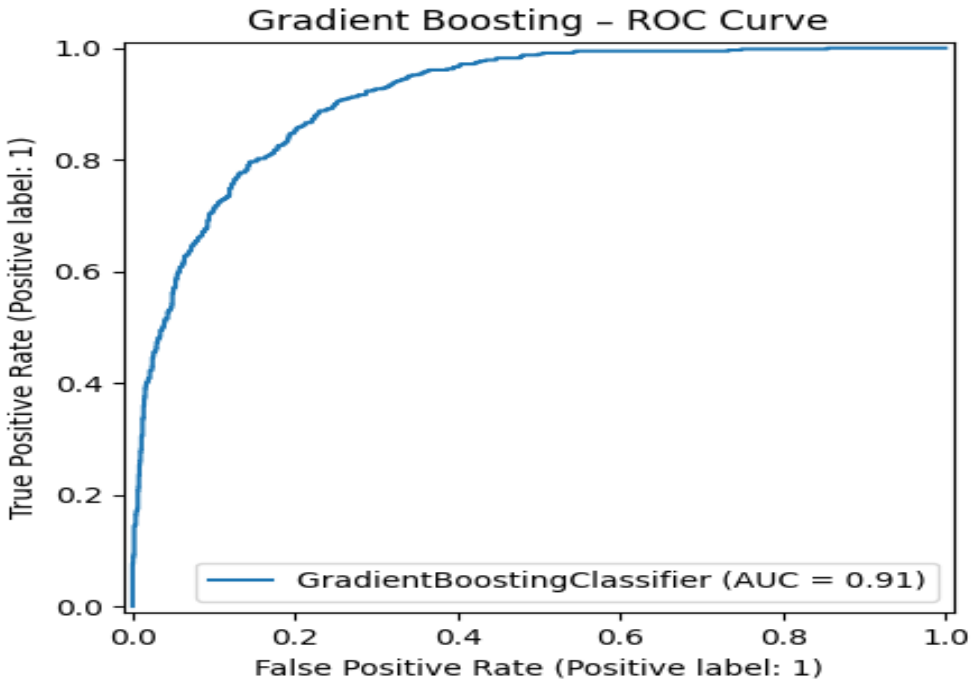



XGBoost – ROC Curve
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
XGBClassifier (AUC = 0.90)
0.0
0.2
0.4
0.6
0.8
1.0

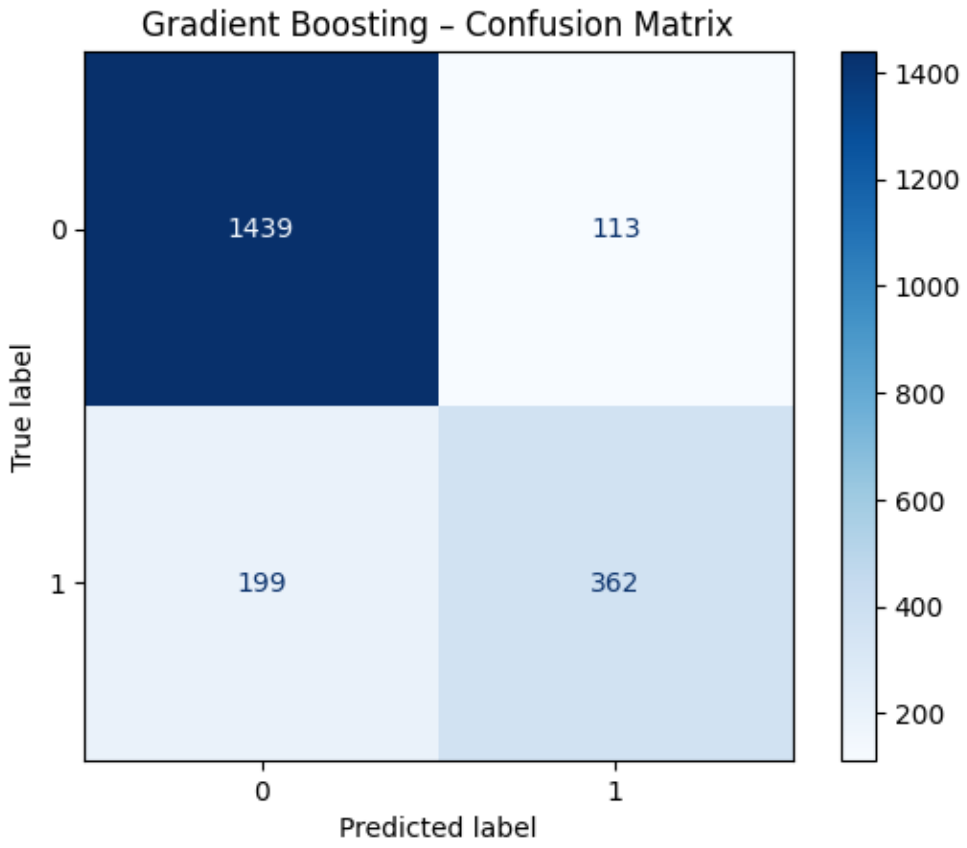

XGBoost – Confusion Matrix
True label
Predicted label
1421
131
208
353
0
1
1400
1200
1000
800
600
400
200

6.2 Model 18 (without satisfaction score)- Gradient Boosting

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Gradient Boosting	0.8523	0.7621	**0.6453**	0.6988	0.9118



Using recall: the model correctly identified 64% of customers likely to churn .



(without satisfaction score)-compare trees

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
XGBoost	0.8396	0.7293	0.6292	0.6756	0.9024
Decision Tree	0.9403	0.8649	0.2282	0.3611	0.8466
Random Forest	0.8495	0.7718	0.6150	0.6845	0.8990
Gradient Boosting	0.8523	0.7621	0.6453	0.6988	0.9118

Gradient boost performed best by identifying 64% of customers likely to churn.

6.3 Model 19- (without satisfaction score)- Forward Regression

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Forward Regression	0.84	0.73	0.65	0.69	0.9014

Using recall: the model correctly identified 65% of customers likely to churn.

Forward Selected Features: 26 variables

'age', 'number_of_dependents', 'population', 'number_of_referrals', 'tenure_in_months', 'avg_monthly_gb_download', 'monthly_charge', 'total_refunds', 'total_extra_data_charges', 'married_Yes', 'dependents_Yes', 'offer_Offer B', 'offer_Offer C', 'offer_Offer D', 'offer_Offer E', 'offer_no_offer', 'multiple_lines_Yes', 'internet_service_Yes', 'internet_type_Fiber Optic', 'online_security_Yes', 'contract_One Year', 'contract_Two Year', 'paperless_billing_Yes', 'payment_method_Credit Card', 'region_Central California', 'region_Southern California'

6.4 Model 20 (without satisfaction score)- Backward Regression

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Backward Regression	0.84	0.73	0.65	0.69	0.9039

Using recall: the model correctly identified 65% of customers likely to churn.

Backward Eliminated Features: 31 variables

'age', 'number_of_referrals', 'tenure_in_months', 'avg_monthly_gb_download', 'monthly_charge', 'total_long_distance_charges', 'cltv', 'married_Yes', 'dependents_Yes', 'referred_a_friend_Yes', 'offer_Offer B', 'offer_Offer C', 'offer_Offer D', 'offer_Offer E', 'offer_no_offer', 'multiple_lines_Yes', 'internet_service_Yes', 'internet_type_DSL', 'internet_type_Fiber Optic', 'online_security_Yes', 'online_backup_Yes', 'premium_tech_support_Yes', 'streaming_tv_Yes', 'streaming_movies_Yes', 'streaming_music_Yes', 'contract_One Year', 'contract_Two Year', 'paperless_billing_Yes', 'payment_method_Credit Card', 'payment_method_Mailed Check', 'region_Central California'.

6.5 Model 21 (without satisfaction score)- Stepwise Regression

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Stepwise Regression	0.84	0.72	0.64	0.68	0.8984

Using recall: the model correctly identified 64% of customers likely to churn.

Stepwise Selected Features: 25 variables

'age', 'number_of_dependents', 'number_of_referrals', 'tenure_in_months', 'avg_monthly_gb_download', 'total_extra_data_charges', 'cltv', 'married_Yes', 'dependents_Yes', 'offer_Offer B', 'offer_Offer C', 'offer_Offer E', 'offer_no_offer', 'multiple_lines_Yes', 'internet_type_Fiber Optic', 'online_security_Yes', 'online_backup_Yes', 'premium_tech_support_Yes', 'contract_One Year', 'contract_Two Year', 'paperless_billing_Yes', 'payment_method_Credit Card', 'payment_method_Mailed Check', 'region_Central California', 'region_Southern California'.

Comparison

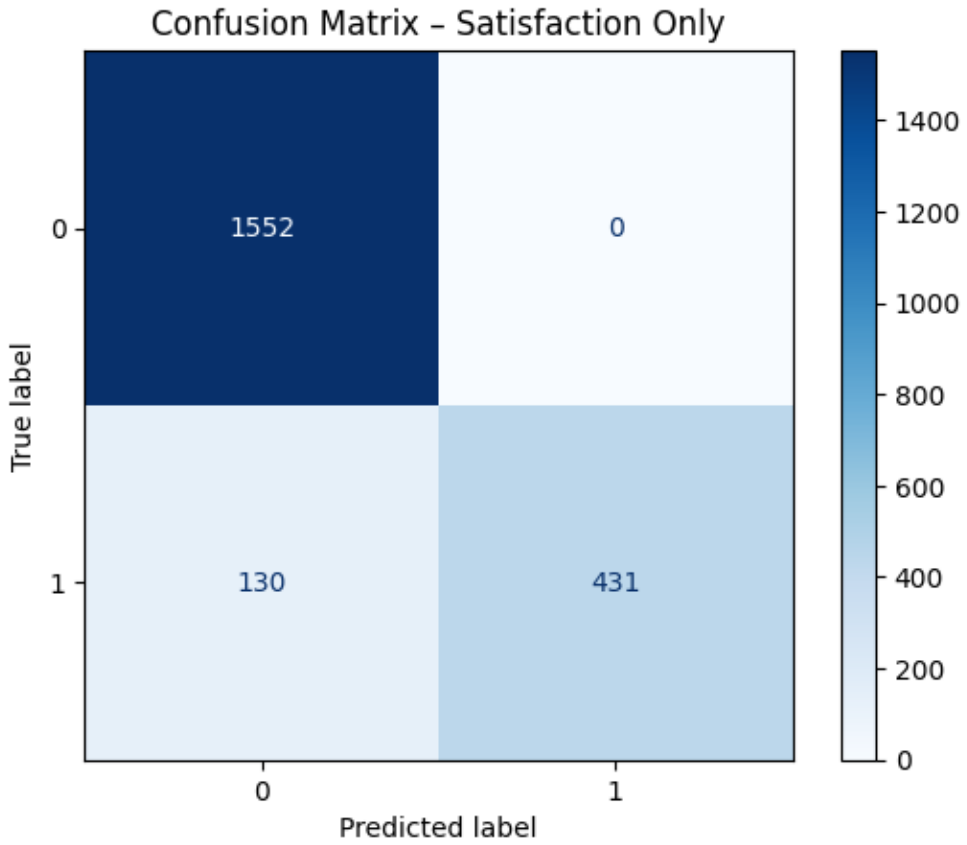
Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC-AUC**
Forward Regression	0.84	0.73	0.65	0.69	0.9014
Backward Regression	0.84	0.73	0.65	0.69	0.9039
Stepwise Regression	0.84	0.72	0.64	0.68	0.8984

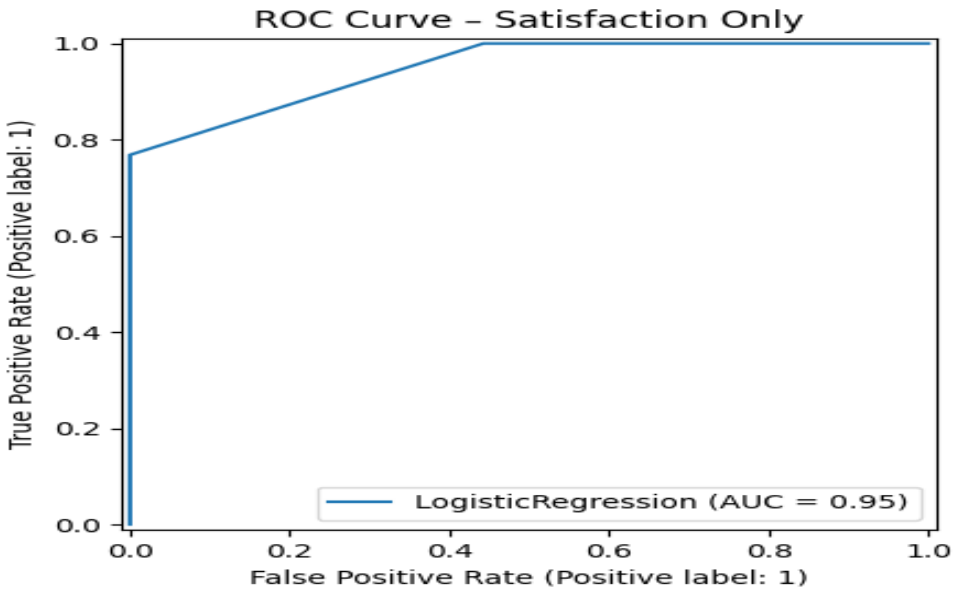
All models performed similarly.

6.6 Model 22 (satisfaction score only)- Logistic Regression

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	0.94	1.00	0.77	0.87	0.9487

Using recall: the model correctly identified 77% of customers likely to churn.

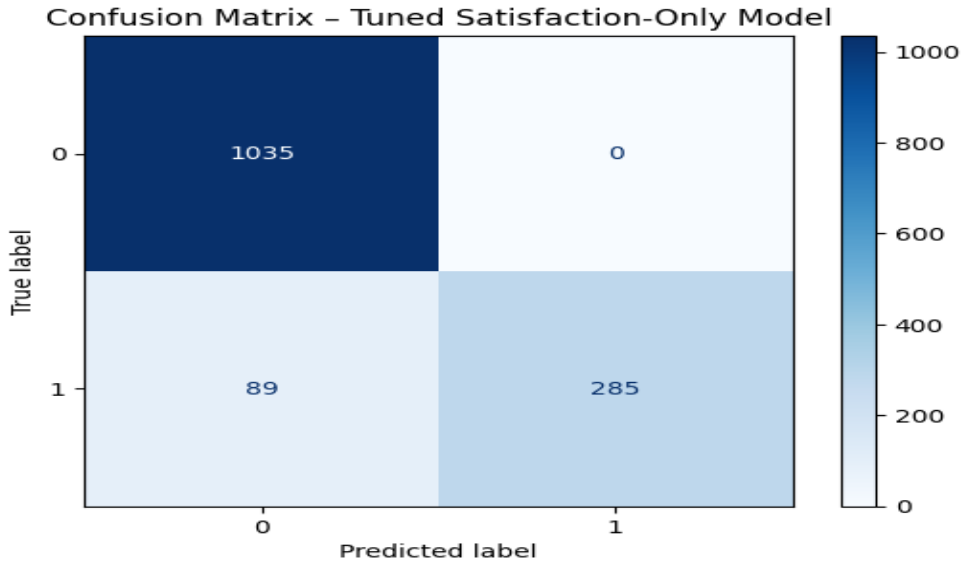


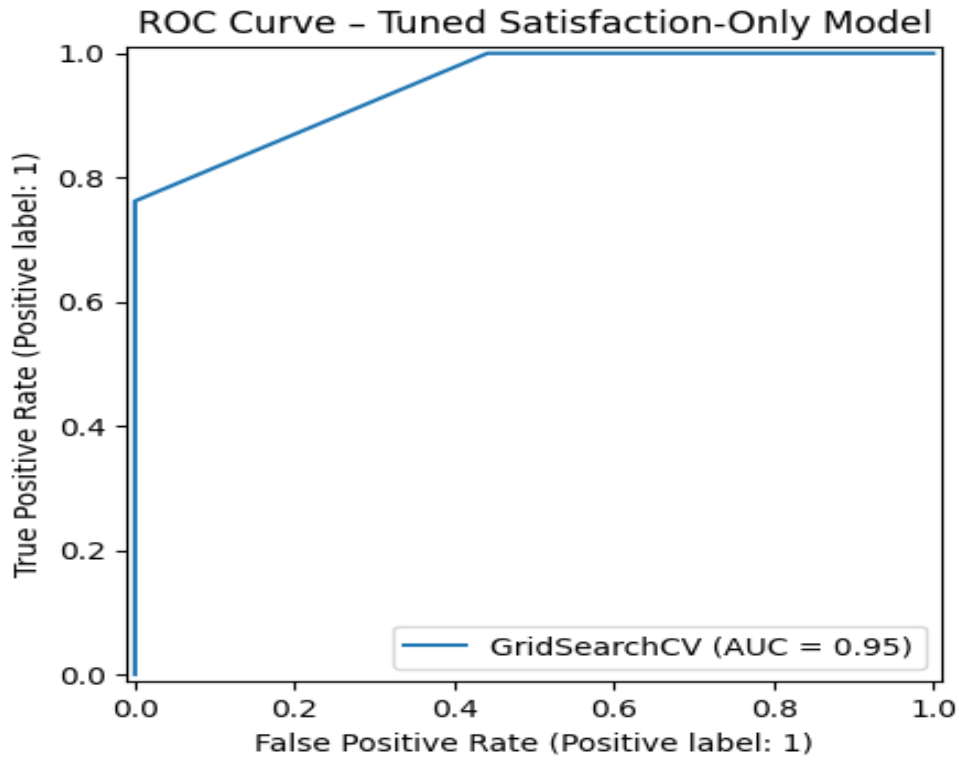


Tuned

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC-AUC**
Logistic Regression	0.94	1.00	0.76	0.86	0.9474

Using recall: the model correctly identified 76% of customers likely to churn.


Confusion Matrix – Tuned Satisfaction-Only Model
1035
0
89
285
0
1
True label
Predicted label
1000
800
600
400
200
0


ROC Curve – Tuned Satisfaction-Only Model
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
GridSearchCV (AUC = 0.95)
0.0
0.2
0.4
0.6
0.8
1.0

7.0 Model Recommendation

7.1 Model Selection

After evaluating multiple models—including Decision Tree, Random Forest, Gradient Boosting, Logistic Regression (full, forward, backward, and stepwise), and XGBoost—**the tuned XGBoost model** was selected as the final choice.

- **Reason**: It achieved the highest overall performance across metrics, with **Recall at 0.9108**, ensuring a strong ability to correctly identify churners while maintaining excellent Accuracy (0.9626) and ROC-AUC (0.9923).
- **Business Rationale**: Since the project's goal is to reduce churn, Recall is critical; the cost of missing an actual churner is higher than contacting a non-churner. XGBoost balances this priority with high overall predictive power, making it the most reliable choice for deployment.

7.2 Model Theory

XGBoost (Extreme Gradient Boosting) is an ensemble learning technique that builds multiple decision trees sequentially, where each tree attempts to correct the errors of the previous ones.

- **Gradient Boosting Framework**: Uses gradient descent on a loss function to minimize prediction errors.
- **Regularization**: Incorporates L1 (Lasso) and L2 (Ridge) penalties to reduce overfitting.
- **Handling Non-Linearity**: Well-suited for capturing complex, non-linear relationships between variables.

- **Efficiency**: Optimized for speed and memory usage, allowing for scalable and fast training on large datasets.

7.3 Model Assumptions and Limitations

Assumptions:

- Data provided is representative of the population (no sampling bias).
- Feature engineering (such as region creation, categorical encoding, and outlier handling) has been applied correctly to maintain predictive accuracy.
- Missing values and extreme outliers have been addressed, as tree-based models can be sensitive to poorly handled data gaps.

Limitations:

- **Interpretability**: Less transparent than simpler models like logistic regression, making stakeholder explanation harder without tools like SHAP.
- **Overfitting Risk**: Despite regularization, high-depth trees may still overfit if hyperparameters are not tuned properly.
- **Feature Sensitivity**: Performance may drop if new incoming data significantly shifts from the training distribution (data drift).

7.4 Model Sensitivity to Key Drivers

Analysis of the tuned XGBoost model's feature importance reveals that customer churn is driven by a combination of satisfaction, pricing, customer value, and usage behavior. The **top 10 key drivers** are:

1. **Satisfaction Score (263)** – The most influential driver. Low satisfaction significantly increases churn risk, highlighting the importance of service quality and issue resolution.
2. **Monthly Charge (229)** – Customers with higher bills are more likely to leave, indicating strong price sensitivity.
3. **Customer Lifetime Value – CLTV (190)** – Low-value customers show higher churn probability, suggesting limited engagement or lower perceived benefits.
4. **Age (187)** – Certain age groups are more prone to churn, potentially due to differing needs or digital preferences.
5. **Tenure in Months (170)** – Shorter-tenure customers are more likely to leave, emphasizing the need for strong onboarding and early engagement.
6. **Number of Referrals (166)** – Fewer referrals may indicate weaker customer advocacy, correlating with higher churn.
7. **Average Monthly Long-Distance Charges (156)** – High usage in this area may reflect niche service needs; dissatisfaction with costs or alternatives could drive churn.
8. **Total Long-Distance Charges (149)** – Similar to average charges, total spend on long-distance services affects retention risk.
9. **Population (128)** – Customers in certain population density areas may have more alternative providers, impacting loyalty.
10. **Average Monthly GB Download (104)** – Usage patterns for internet data may reveal mismatches between service offerings and customer needs.

Implications:

The model is most sensitive to **experience-related metrics** (satisfaction score), **financial factors** (monthly charge, CLTV), and **customer lifecycle indicators** (tenure, referrals). This suggests

that churn mitigation should combine **service improvement**, **competitive pricing strategies**, and **early retention programs** targeting new customers.

8.0 Conclusion and Recommendations

I. Improve Customer Satisfaction

Insight: Satisfaction score is the single strongest predictor of churn — dissatisfied customers are far more likely to leave.

Recommendation:

Develop a structured customer satisfaction improvement program focused on **fast issue resolution** and **proactive engagement** with low-scoring customers.

Strategies to Implement:

1. **Real-Time Alerts:**
    - Set up automated triggers when customer feedback scores drop below a set threshold.
    - Assign these cases to a dedicated "Customer Recovery" team.
2. **Service Training:**
    - Provide staff with targeted training in conflict resolution and empathy-driven communication.
3. **Root Cause Tracking:**
    - Maintain a live dashboard categorizing recurring issues, enabling quick policy or process changes.

II. Address High Monthly Charges.

Insight: Customers with higher monthly charges show an elevated churn risk.

Recommendation:

Introduce **value-based pricing strategies** and targeted discounts to customers with above-average bills.

Strategies to Implement:

1. **Bill Optimization Reviews:** Offer customers a free bill review session to identify savings without cutting key services.
2. **Loyalty Discounts:** Provide monthly credits for customers with more than 12 months tenure.
3. **Bundled Packages:** Combine high-demand services into discounted bundles to increase perceived value.

III. Increase Long-Term Contract Adoption:

Insight: Month-to-month contracts correlate with higher churn rates.

Recommendation:

Promote 1-year and 2-year contract plans through **exclusive perks**.

Strategies to Implement:

1. **Contract Upgrade Incentives:** Offer device upgrades or premium channel add-ons for customers switching to long-term contracts.

2. **Flexible Downgrade Path:** Allow customers to adjust their package mid-term without penalties, reducing cancellation pressure.

IV. Optimize Internet Service Plans:

Insight: Fiber-optic customers and certain offer types have distinct churn behaviors.

Recommendation:

Align internet packages and offer designs to meet customer expectations for speed, reliability, and value.

Strategies to Implement:

1. **Targeted Network Investment:** Prioritize service upgrades in fiber-optic regions with higher churn.
2. **Offer Redesign:** Review "Offer E" and other high-churn packages to improve value or replace them entirely.
3. **Personalized Marketing:** Recommend the most reliable internet type based on a customer's location and usage profile

V. Strengthen Onboarding for New Customers

Insight: Tenure is inversely related to churn — customers in their first year are at greater risk.

Recommendation:

Implement a **90-Day Welcome Program** that builds engagement early in the relationship.

Strategies to Implement:

1. **Proactive Check-ins:** Schedule calls at the 30-, 60-, and 90-day marks to ensure satisfaction.
2. **Exclusive Offers:** Provide first-year customers with targeted benefits to build loyalty.
3. **Education Campaigns:** Send guides and videos to help customers maximize their services.

9.0 Impacts on Business Problem

1. Reduction in Customer Churn
By implementing targeted strategies informed by the XGBoost model's top predictors (e.g., satisfaction score, tenure, contract type, and internet type), the business can proactively engage at-risk customers. Even a **5% reduction in churn** among high-value accounts could preserve millions in annual recurring revenue, safeguarding market share and stabilizing cash flow.

2. Increased Customer Lifetime Value (CLTV)
Personalizing offers, improving service quality, and promoting long-term contracts will extend customer tenure, thereby increasing CLTV. This shifts the business model from constant acquisition spending to value maximization from the existing base.

3. Cost Optimization
Focusing retention efforts on high-risk, high-value customers ensures optimal use of marketing and retention budgets. The model's insights help avoid spending on low-risk customers who are already likely to stay.

4. Enhanced Competitive Positioning
Competitive pricing, bundled offers, and loyalty programs can differentiate the company in a crowded market. This not only retains customers but also attracts switchers from competitors.

5. Operational Efficiency
Introducing easier payment options, digital self-service, and improved employee training reduces support call volumes and improves first-contact resolution rates, lowering operational costs.

Data-Driven Decision-Making

Using XGBoost's feature importance rankings and SHAP values ensures that business strategies are not based on assumptions, but on statistically validated drivers of churn. This increases the likelihood of measurable ROI from retention campaigns.

Strategic Growth Enablement

Revenue stability from reduced churn gives the company more capital to reinvest in network upgrades, product innovation, and expansion into competitive markets. This positions the business for sustainable long-term growth instead of fighting constant revenue leakage

10. Recommended Next Steps

Customer Experience & Retention

1. **Customer Feedback Loop** – Regularly collect, analyze, and act on customer feedback to improve services.

2. **Personalized Offers Based on Value** – Tailor deals and promotions to customer segments with the highest lifetime value.
3. **Loyalty Programs** – Offer points, rewards, or exclusive benefits to encourage repeat business.
4. **Improve Service Quality** – Address pain points quickly, ensuring reliable and high-quality service delivery.
5. **Dedicated Retention Team** – Assign a specialized team to proactively reach out to at-risk customers before they churn.

Pricing & Value Proposition

6. **Competitive Pricing** – Benchmark against industry rates to remain attractive to price-sensitive customers.
7. **Bundled Discounts** – Provide multi-service or family plan discounts to increase value perception.
8. **Unlimited Long-Distance Packages** – Cater to high-call-volume customers by removing costly limits.
9. **Seasonal Promotions** – Use targeted discounts during peak competitive seasons to attract new customers.

Sales & Engagement

10. **Incentivize Referrals** – Reward existing customers for bringing in new subscribers.
11. **Promote Long-Term Contracts** – Offer price benefits or perks for committing to extended service periods.
12. **Upsell & Cross-Sell** – Leverage data to suggest relevant upgrades or additional services.

Ease of Use & Accessibility

13. **Make Payment Options Easier** – Introduce flexible payment schedules and multiple payment methods.
14. **Digital Self-Service Portals** – Empower customers to manage their accounts, troubleshoot issues, and make changes online.
15. **Simplified Onboarding Process** – Ensure new customers can start using services quickly with minimal friction.

Brand Trust & Employee Engagement

16. **Leadership Commitment** – Demonstrate top-level dedication to customer-first policies.
17. **Employee Training** – Equip staff with the skills and tools to deliver excellent customer interactions.
18. **Transparent Communication** – Keep customers informed about changes, outages, and new offerings.
19. **Community Engagement** – Strengthen brand loyalty by participating in or sponsoring local events.

References

Chang, J. (2019, November 8). *Telco customer churn (11.1.3+)*.Kaggle. https://www.kaggle.com/datasets/ylchang/telco-customer-churn-1113/data