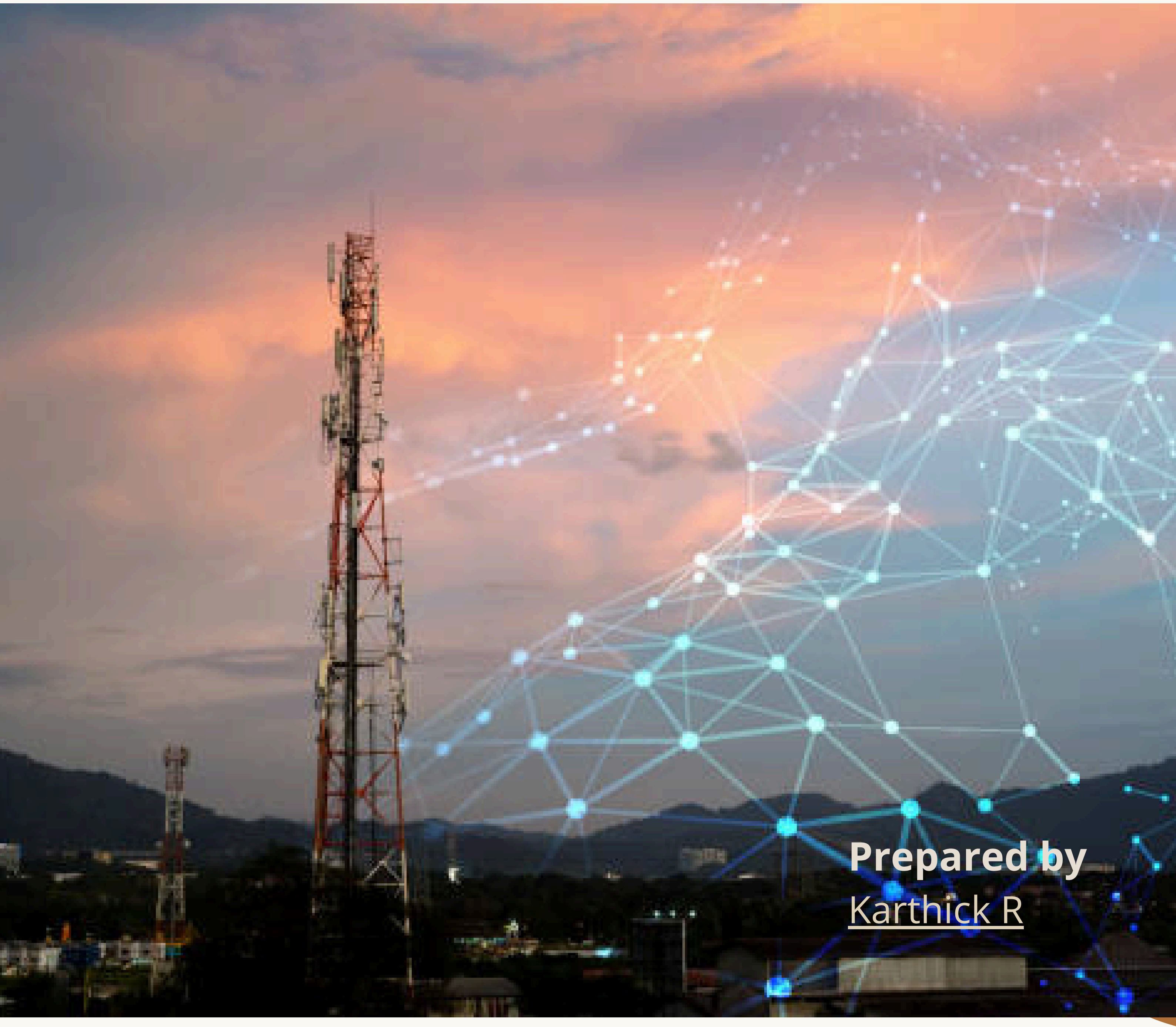
AlphaCom

CUSTOMER CHURN PREDICTION



Prepared by

Karthick R

TABLE OF CONTENT

Topics Page no

Objectives 6

Business Problem Statement 7

Data overview 10

Exploratory Data Analysis 17

Univariate analysis 17

Bivariate analysis 22

Data preprocessing 27

Model building 30

Hyperparameter Tuning 31

Feature Importance 44

Insights and Actionable business recommendations 46

LIST OF TABLES

Topics	Page no
1.Checking Distinct values	10
2. Top five rows of dataset	11
3. Data types of the column	12
4. Statistical summary of the dataset	15
5. Data on null values	16
6. Treating missing values	27
7. Shape after one hot encoding	29
8. Training (CV) and Validation score	30
9. Training performance	32
10. Validation performance	32
11. Training performance on Tuning	33
12. Validation performance on Tuning	33
13. Oversampled data	34
14. Training (CV) and Validation score	35
15. Undersampled Data	36
16. Training (CV) and Validation score	37

LIST OF TABLES

Topics Page no

17. Training performance on Tuning 38

18. Validation performance on Tuning 38

19. Training performance (Adaboosting) 39

20.Validation performance (Adaboosting) 40

21. Training performance(Log Reg) 41

22. Validation performance (Log Reg) 41

23. Training performance comparison 42

24. Validation performance comparison 42

25. Test Performance 43

26. Business Recommendation 48

LIST OF FIGURES

Topics **Page no**

1.Analysis on tenure 17

2. Analysis on Monthly Charges 18

3. Analysis on Total Charges 19

4. Analysis on Gender 20

5. Analysis on Churn 21

6. Analysis of Pairplot 22

7. Analysis between Churn and tenure 23

8. Churn and Total Charges 24

9. Internet Service and Churn 25

10. Correlation Matrix Analysis 26

11. Payment Method and Total Charges 28

12. Confusion Matrix Logistic Regression Model 43

13. Feature Importance 45

OBJECTIVES

The goal is to build a predictive model that accurately identifies customers who are likely to discontinue their service. In addition to prediction, the objective is to understand which factors most strongly contribute to churn. Achieving this will enable the business to take proactive action — by targeting at-risk customers with personalized retention strategies — thereby minimizing revenue loss, increasing customer lifetime value, and improving overall business stability and competitiveness.

BUSINESS PROBLEM STATEMENT

AlphaCom, a major telecommunications provider, is facing a rising rate of customer churn that is negatively affecting its revenue and market reputation. Existing retention strategies are ineffective because churn is driven by multiple interconnected factors—such as service usage patterns, billing behavior, contract terms, and customer demographics—making it difficult to anticipate who is likely to leave. The absence of a data-driven churn prediction mechanism forces the company to react only after customers defect, instead of intervening early.

To address this challenge, a predictive model is required to accurately identify customers at high risk of churn and to reveal the most influential drivers behind their decision to leave. These insights will enable AlphaCom to implement targeted retention strategies proactively, reduce financial losses, and enhance long-term customer loyalty.

NEED OF THE STUDY/PROJECT

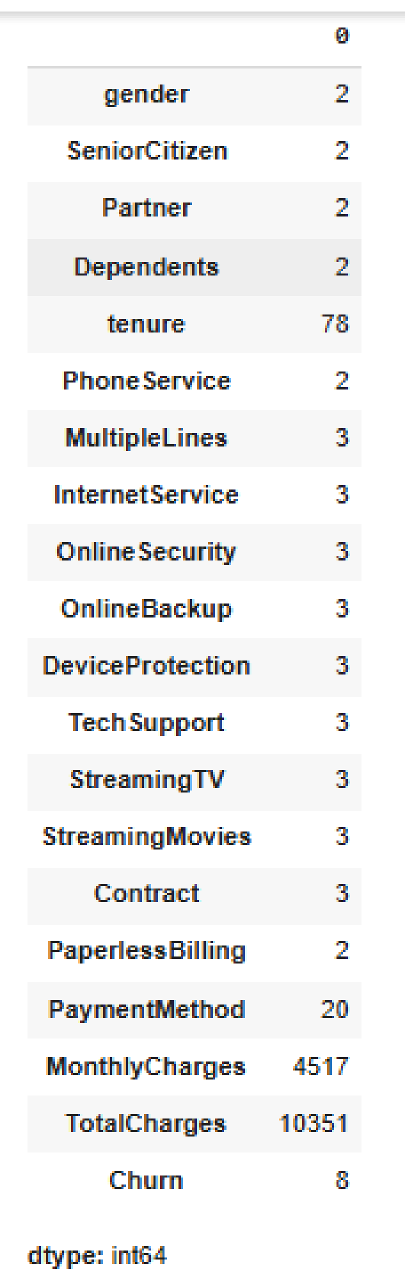
Customer churn poses a significant challenge for service-based industries like telecommunications, where revenue is heavily dependent on long-term customer retention rather than one-time acquisition. Despite having competitive offerings, AlphaCom is experiencing a steady rise in churn, leading to substantial financial loss, higher acquisition costs to replace lost customers, and potential damage to brand loyalty. Traditional intuition-based approaches lack the precision needed to identify at-risk customers early. Therefore, a data-driven predictive solution is essential to move from reactive to proactive retention planning. This study aims to fill that gap by applying analytical and machine learning techniques to forecast churn and uncover actionable drivers behind customer attrition.

UNDERSTANDING BUSINESS/SOCIAL OPPORTUNITY

From a business perspective, early identification of churners allows AlphaCom to intervene with targeted retention offers, optimize marketing budgets, preserve recurring revenue, and improve customer lifetime value. This not only strengthens the company's financial stability but also builds a competitive edge in a saturated market.

From a broader social and customer experience perspective, the insights gained from this study help the organization enhance service quality, personalize offerings, and improve customer satisfaction. Instead of blanket promotions or last-minute recovery attempts, the company can address root causes of dissatisfaction, creating a more transparent and responsive service ecosystem that benefits both the business and its customers.

DATA OVERVIEW



```
                    0
gender              2
SeniorCitizen       2
Partner             2
Dependents          2
tenure             78
PhoneService        2
MultipleLines       3
InternetService     3
OnlineSecurity      3
OnlineBackup        3
DeviceProtection    3
TechSupport         3
StreamingTV         3
StreamingMovies     3
Contract            3
PaperlessBilling    2
PaymentMethod      20
MonthlyCharges   4517
TotalCharges    10351
Churn               8

dtype: int64
```

- There are lot of unique values due to leading/trailing spaces and $ sign in numeric columns
- The Data need some preliminary correction before EDA.

Table 1: Checking Distinct values

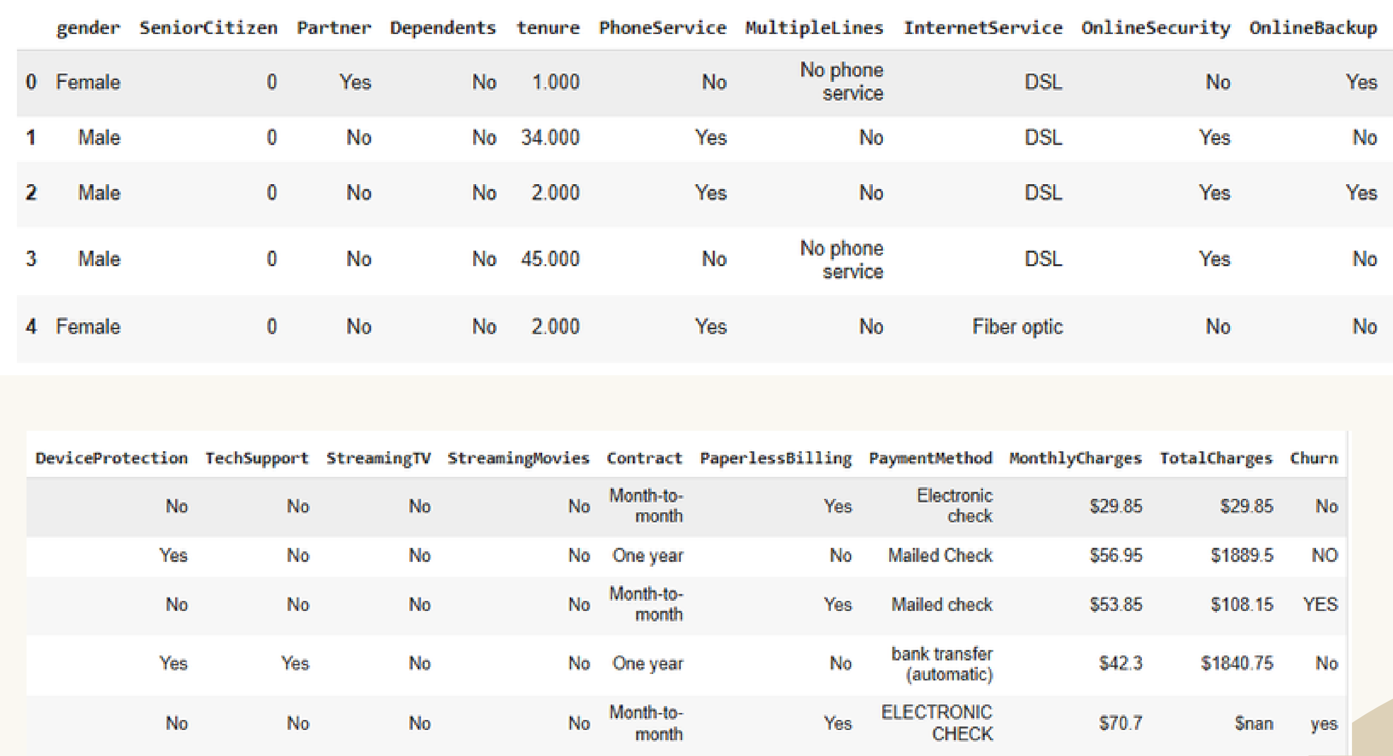
DATA CLEANING FOR EDA

- Removing symbols like $ from the dataset.
- Striping white space in between dataset.
- Standardize the values to "Yes" or "No"
- Removing leading/trailing spaces and make lowercase

DATA OVERVIEW

Total Rows: 12055
Total Columns: 20
Target Variable: Churn (Yes or no)

DISPLAYING THE FIRST 5 ROWS OF THE DATASET



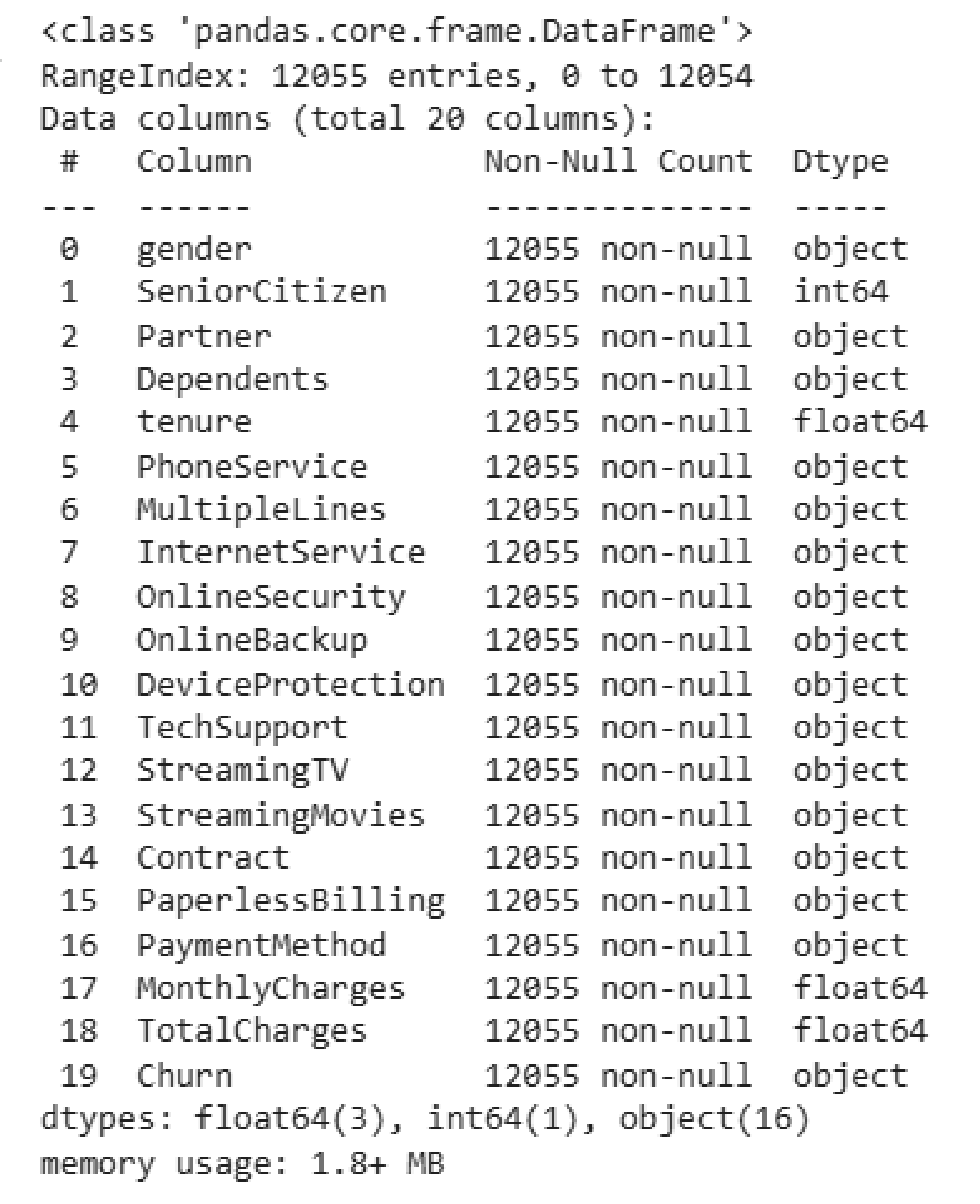
	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	Female	0	Yes	No	1.000	No	No phone service	DSL	No	Yes	No	No	No	No	Month-to-month	Yes	Electronic check	$29.85	$29.85	No
1	Male	0	No	No	34.000	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed Check	$56.95	$1889.5	NO
2	Male	0	No	No	2.000	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-month	Yes	Mailed check	$53.85	$108.15	YES
3	Male	0	No	No	45.000	No	No phone service	DSL	Yes	No	Yes	Yes	No	No	One year	No	bank transfer (automatic)	$42.3	$1840.75	No
4	Female	0	No	No	2.000	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-month	Yes	ELECTRONIC CHECK	$70.7	$nan	yes

Table 2: Top five rows of dataset

CHECKING THE SHAPE OF THE DATASET

- The dataset contains 12055 rows and 20 columns

CHECKING THE DATA TYPES OF THE COLUMNS FOR THE DATASET



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12055 entries, 0 to 12054
Data columns (total 20 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   gender            12055 non-null  object 
 1   SeniorCitizen     12055 non-null  int64  
 2   Partner           12055 non-null  object 
 3   Dependents        12055 non-null  object 
 4   tenure            12055 non-null  float64
 5   PhoneService      12055 non-null  object 
 6   MultipleLines     12055 non-null  object 
 7   InternetService   12055 non-null  object 
 8   OnlineSecurity    12055 non-null  object 
 9   OnlineBackup      12055 non-null  object 
 10  DeviceProtection  12055 non-null  object 
 11  TechSupport       12055 non-null  object 
 12  StreamingTV       12055 non-null  object 
 13  StreamingMovies   12055 non-null  object 
 14  Contract          12055 non-null  object 
 15  PaperlessBilling  12055 non-null  object 
 16  PaymentMethod     12055 non-null  object 
 17  MonthlyCharges    12055 non-null  float64
 18  TotalCharges      12055 non-null  float64
 19  Churn             12055 non-null  object 
dtypes: float64(3), int64(1), object(16)
memory usage: 1.8+ MB
```

Table 3: Data types of the column

Data Type Summary

- 16 categorical columns (object dtype)
- 3 numerical columns (tenure, MonthlyCharges, TotalCharges)
- 1 integer binary column (SeniorCitizen)

Dataset Description

- The dataset contains 12,055 customer records with 20 features related to a telecom company's customers.
- Each row represents a single customer, and the final column Churn indicates whether the customer has discontinued the service.

The features include:

Customer Demographics

- gender — Male/Female
- SeniorCitizen — Whether the customer is a senior citizen (0/1)
- Partner — Whether the customer lives with a partner
- Dependents — Whether the customer has dependents

Account & Subscription Information

- tenure — Number of months the customer has stayed with the company
- Contract — Type of contract (Month-to-month / One year / Two year)
- PaperlessBilling — Whether billing is paperless
- PaymentMethod — Mode of payment (Credit card, Bank transfer, etc.)

Service-related Features

- PhoneService — Whether the customer has phone service
- MultipleLines — Whether multiple lines are active
- InternetService — Type of internet (DSL / Fiber optic / None)
- OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport,
- StreamingTV, StreamingMovies — Add-on services subscribed (Yes/No/No internet)

Billing Information

- MonthlyCharges — Amount charged per month
- TotalCharges — Total amount billed to date

Target Variable

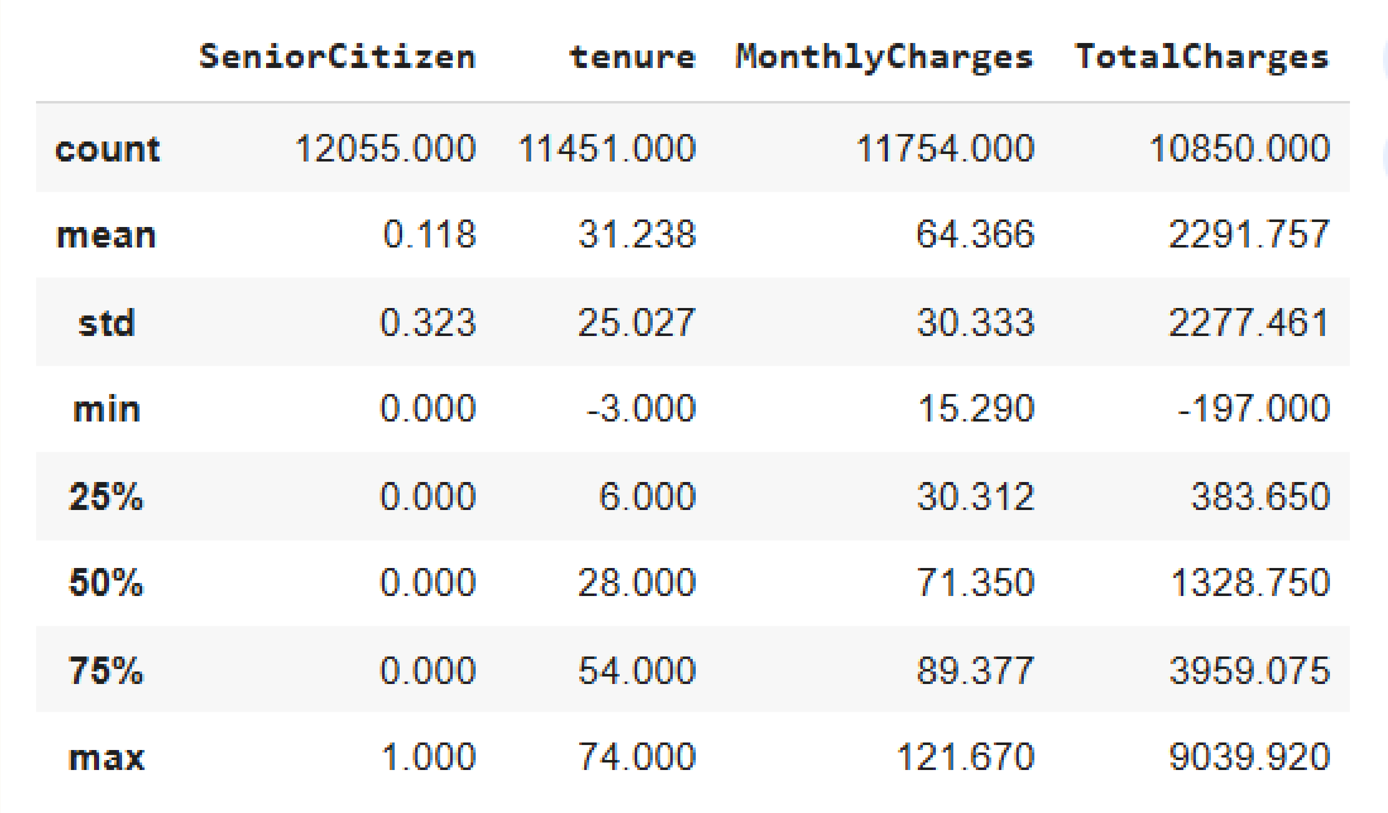
- Churn — Whether the customer left the company (Yes or No)

Purpose of the Dataset for the Project

This dataset supports:

1. **Predictive modeling** — to classify customers likely to churn
2. **Insight mining** — to identify what factors contribute most to churn
3. **Strategic planning** — enabling targeted retention interventions

Statistical summary of the dataset



	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
count	12055.000	11451.000	11754.000	10850.000
mean	0.118	31.238	64.366	2291.757
std	0.323	25.027	30.333	2277.461
min	0.000	-3.000	15.290	-197.000
25%	0.000	6.000	30.312	383.650
50%	0.000	28.000	71.350	1328.750
75%	0.000	54.000	89.377	3959.075
max	1.000	74.000	121.670	9039.920

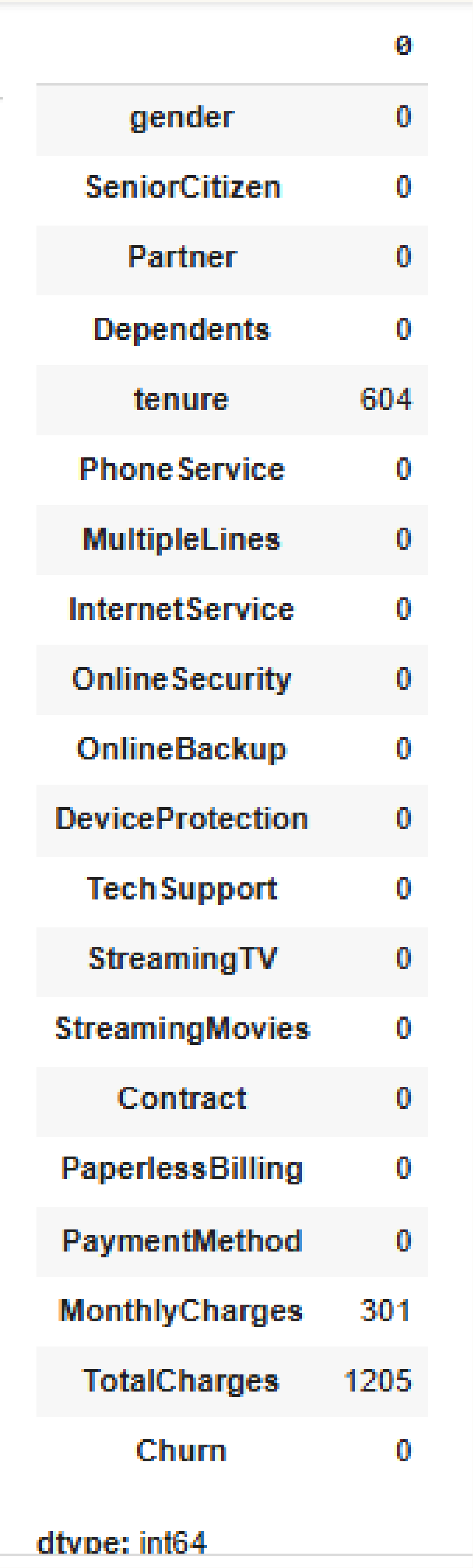
Table 4: Statistical summary of the dataset

- Average customer stays for ~31 months
- 25% of customers leave within first 6 months — early churn is a significant risk zone
- Average MonthlyCharges is ~₹64 (currency unit depends on dataset)
- Some customers pay over ₹120 per month, indicating premium plans
- Median TotalCharges (~₹1329) suggests many customers leave before staying long

Checking for Duplicate Values

- There are no duplicate values in the data.

Checking for Null Values



	0
gender	0
SeniorCitizen	0
Partner	0
Dependents	0
tenure	604
PhoneService	0
MultipleLines	0
InternetService	0
OnlineSecurity	0
OnlineBackup	0
DeviceProtection	0
TechSupport	0
StreamingTV	0
StreamingMovies	0
Contract	0
PaperlessBilling	0
PaymentMethod	0
MonthlyCharges	301
TotalCharges	1205
Churn	0

```
dtype: int64
```

Table 5: Data on null values

- There are some null values in the data on the column tenure, MonthlyCharges, TotalCharges .

EXPLORATORY DATA ANALYSIS

UNIVARIATE ANALYSIS

Analysis on tenure

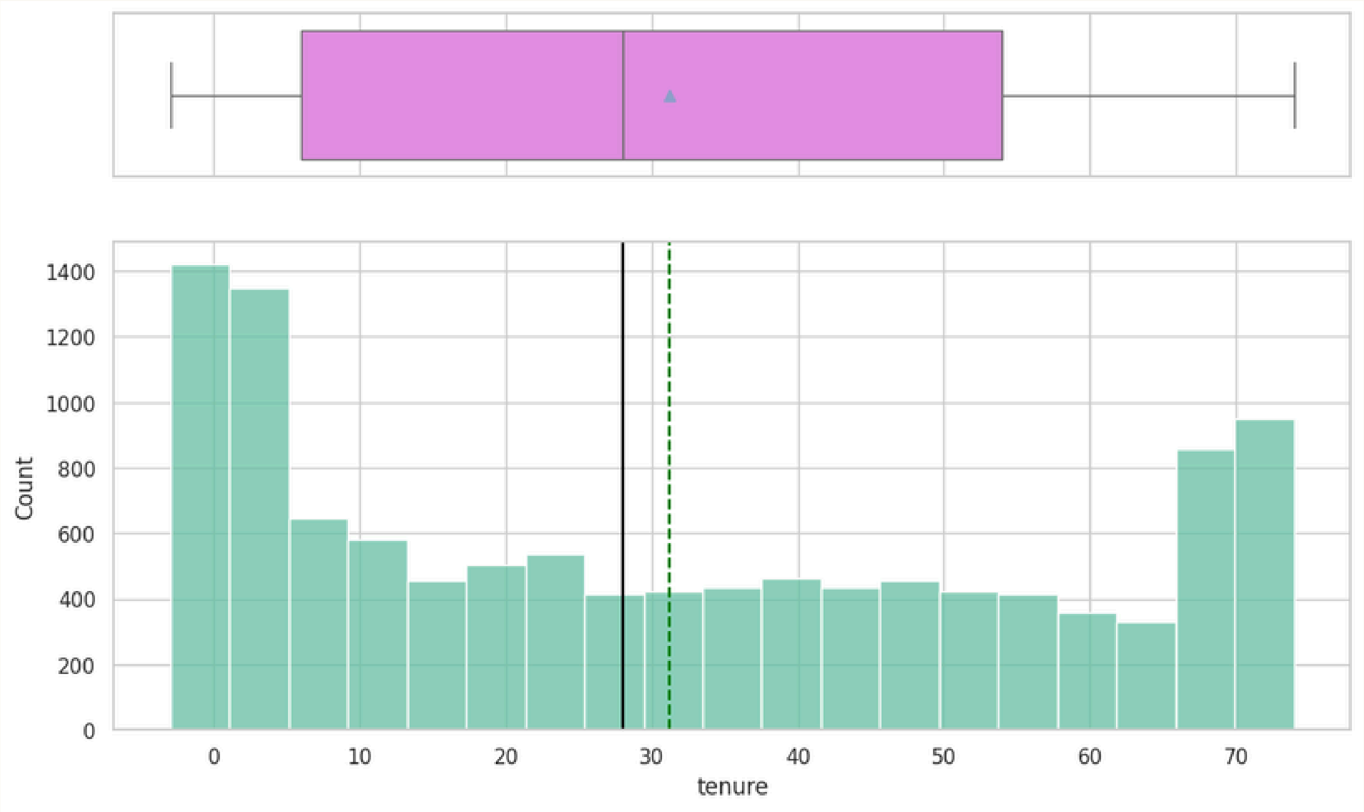


Figure 1: Analysis on tenure

Observations

- Tenure is heavily skewed toward new customers, with a large concentration in the 0–10 month range.
- There is another noticeable spike of long-term customers around 60–72 months, showing two distinct customer segments.
- The median (~28 months) and mean (~31 months) lie in the mid-range, indicating that although many are new, a sizable portion stays long enough to balance the average.

Analysis on Monthly Charges

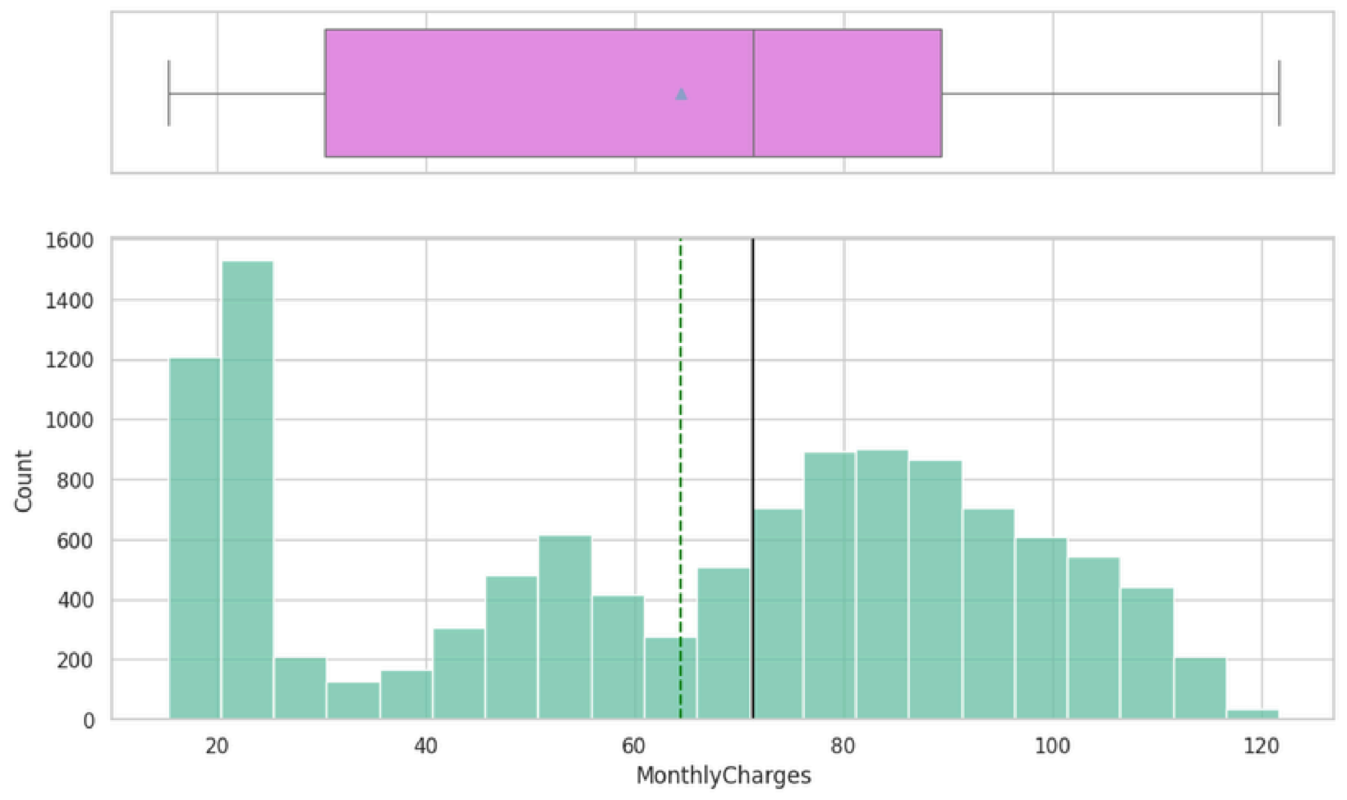


Figure 2: Analysis on Monthly Charges

Observations

- The distribution is clearly bimodal — one group paying around \$20–30 and another around \$70–100, indicating distinct plan segments (basic vs premium users).
- Both the mean and median lie in the higher range, showing the dataset is slightly right-shifted toward customers with higher monthly bills.
- There are no extreme outliers, but the spread is wide — suggesting billing amount varies significantly depending on subscribed services or contract type.

Analysis on Total Charges

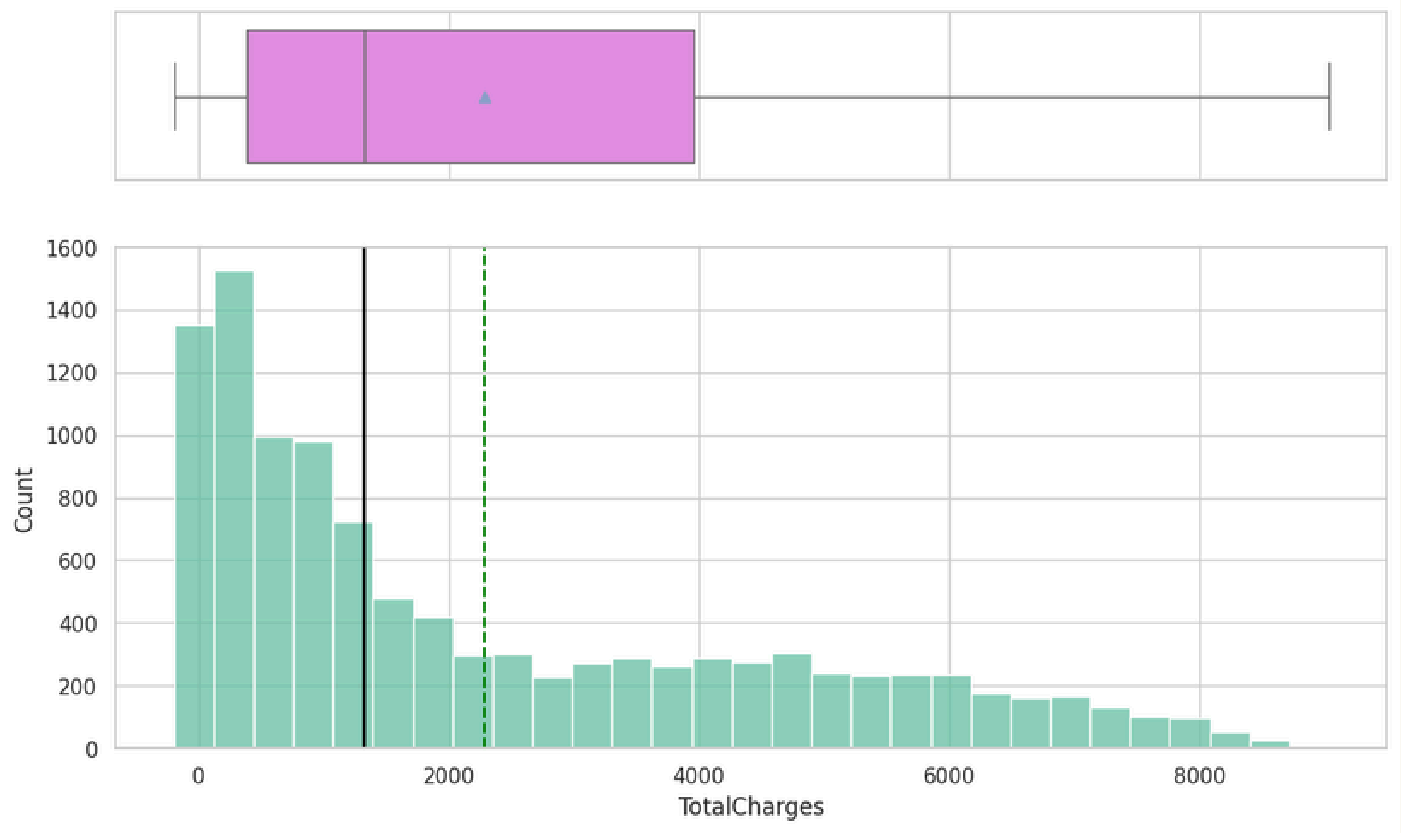


Figure 3: Analysis on Total Charges

Observations

- TotalCharges is heavily right-skewed, with most customers having paid relatively low cumulative amounts, while a small group of long-tenure users contributes to very high totals.
- The median is much lower than the mean, confirming that a few high-paying long-term customers pull the average upward.
- Occurrence of very small or near-zero TotalCharges likely corresponds to newly joined customers who have not been billed for long.

Analysis on Gender

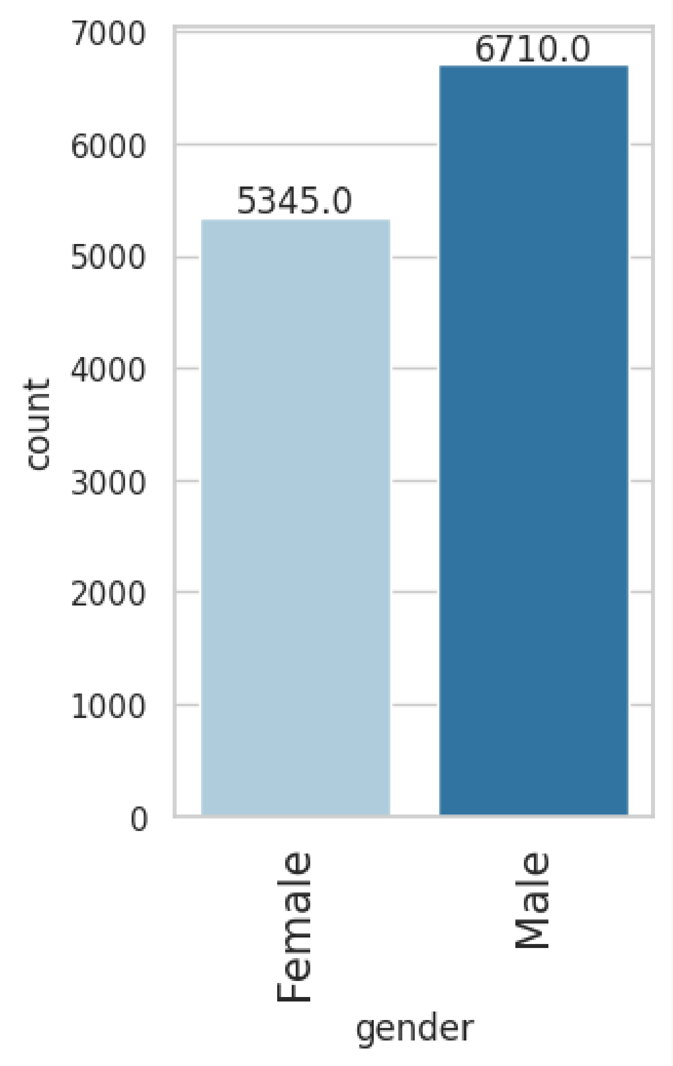


Figure 4: Analysis on Gender

Observations

- The dataset contains more male customers (≈6710) compared to female customers (≈5345), indicating a slight gender imbalance.
- Despite the difference in counts, both genders are well-represented, so gender can still be used reliably as a feature in modeling.

Analysis on Churn

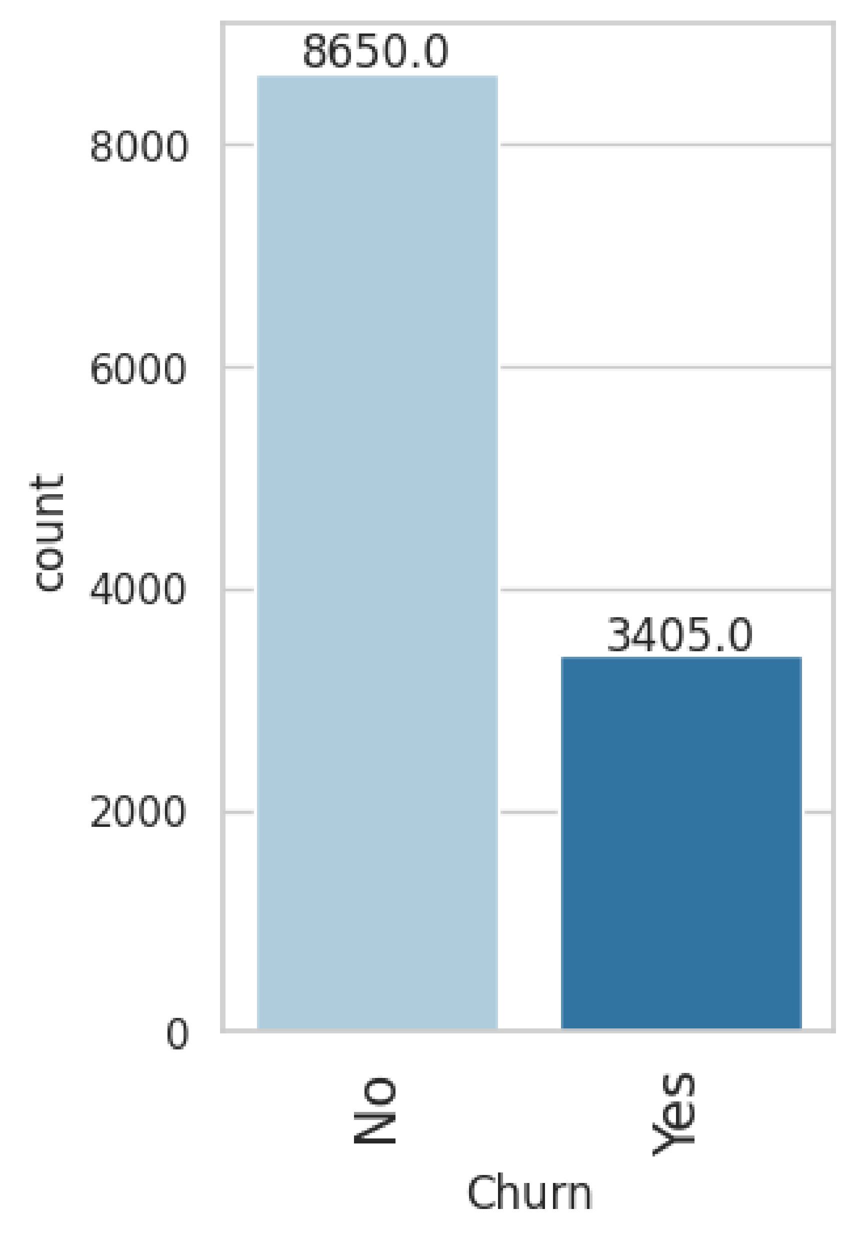


Figure 5: Analysis on Churn

Observations

- The dataset is imbalanced, with a much larger number of customers who did not churn (≈8650) compared to those who churned (≈3405).
- Roughly 28% of customers have churned, meaning churn is a significant issue but not the majority class.
- Because of this imbalance, evaluation metrics like recall, precision, F1, and ROC-AUC are more appropriate than plain accuracy when building models.

BIVARIATE ANALYSIS

Analysis of Pairplot

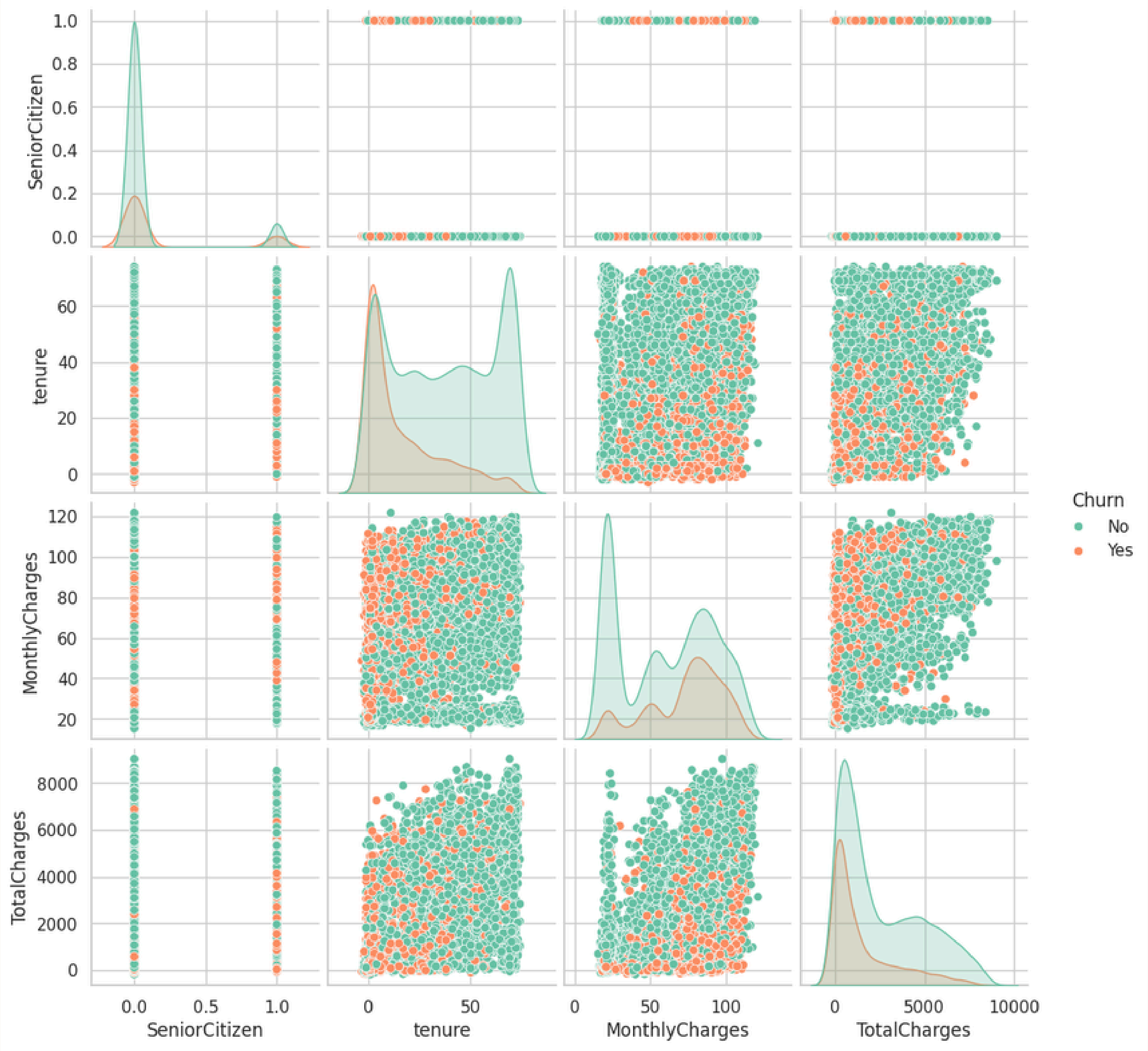


Figure 6: Analysis of Pairplot

Observations

- Customers who churned tend to have lower tenure, which means many leave within the early phase of their service period compared to retained customers who show higher concentrations at long tenure.
- Higher MonthlyCharges seems associated with higher churn density, suggesting cost-sensitive customers may be leaving more often.
- Senior citizens appear slightly more represented in churn cases, but the effect is weaker compared to tenure and charges; TotalCharges separation is mostly driven by tenure length.

Analysis between Churn and tenure

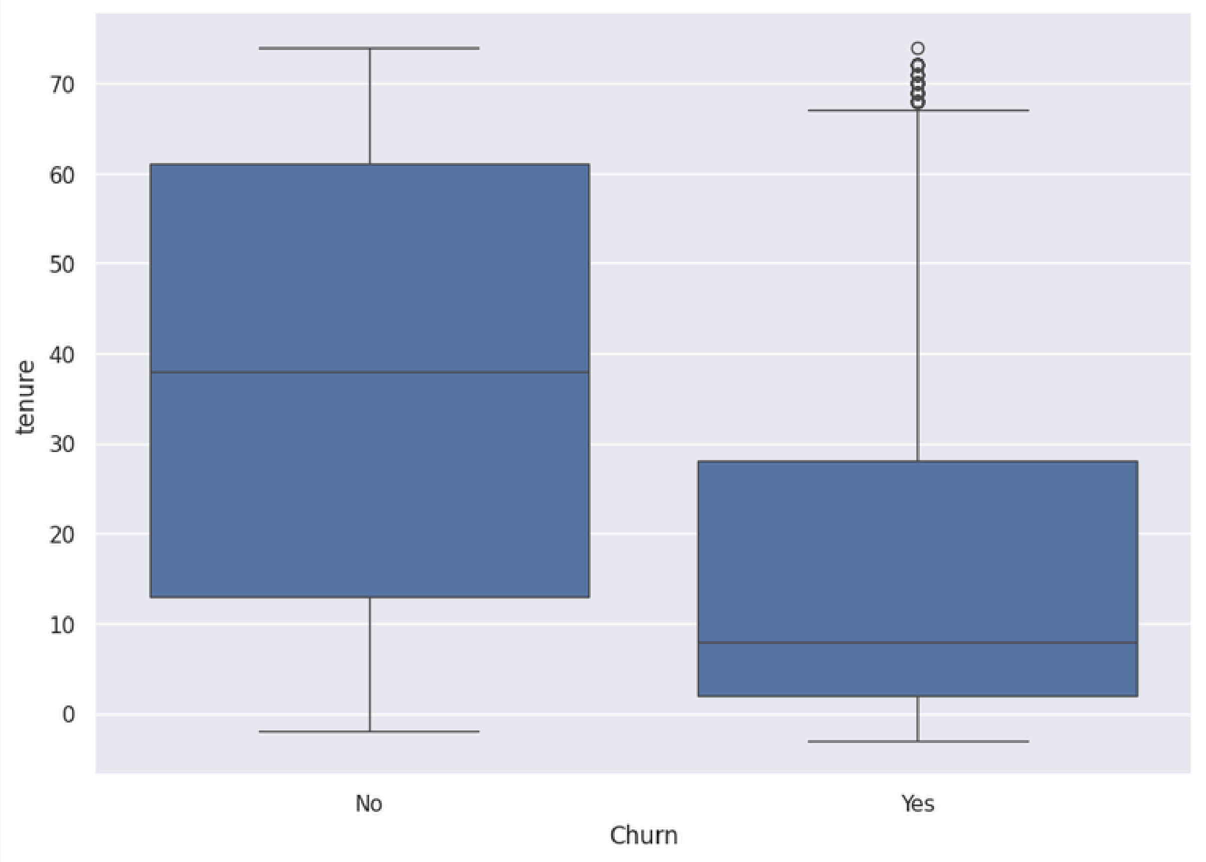


Figure 7: Analysis between Churn and tenure

Observations

- Customers who churned have significantly lower tenure — their median tenure is around a few months, indicating many leave early in the lifecycle.
- In contrast, non-churned customers have a much higher median and wider spread, meaning retention improves the longer customers stay.
- The presence of only a few long-tenure outliers in the churn group reinforces that late-stage churn is rare — churn is primarily an early-stage phenomenon.

Analysis between Churn and Total Charges

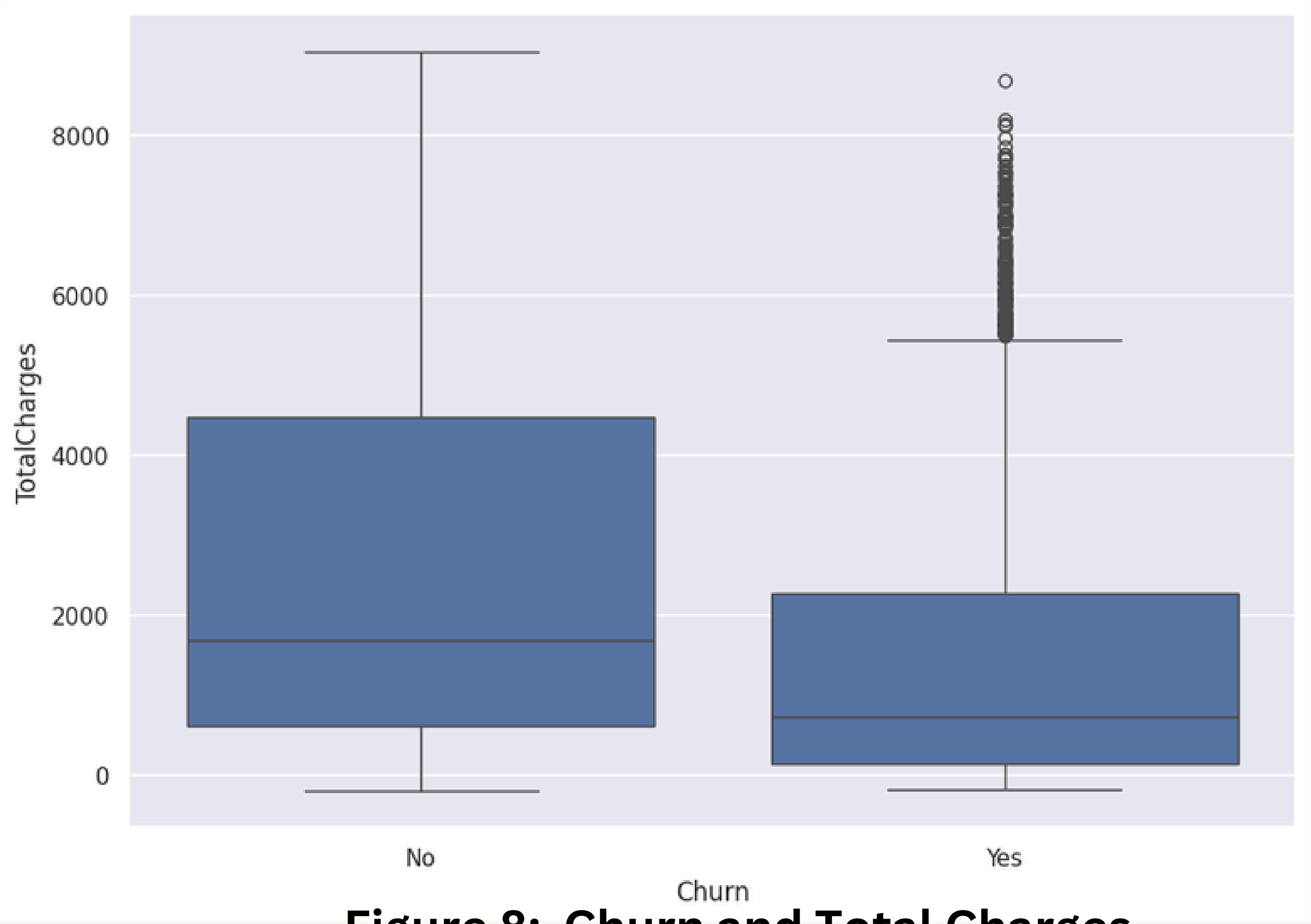


Figure 8: Churn and Total Charges

Observations

- The median TotalCharges is much lower for churned customers, meaning many of them leave before spending much with the company.
- Non-churned customers show higher totals and a wider spread, consistent with long-term retention and continued billing.
- A few churned customers have high TotalCharges (outliers), but these are rare — reinforcing that most churn happens early in the customer lifecycle.

Analysis between Internet Service and Churn

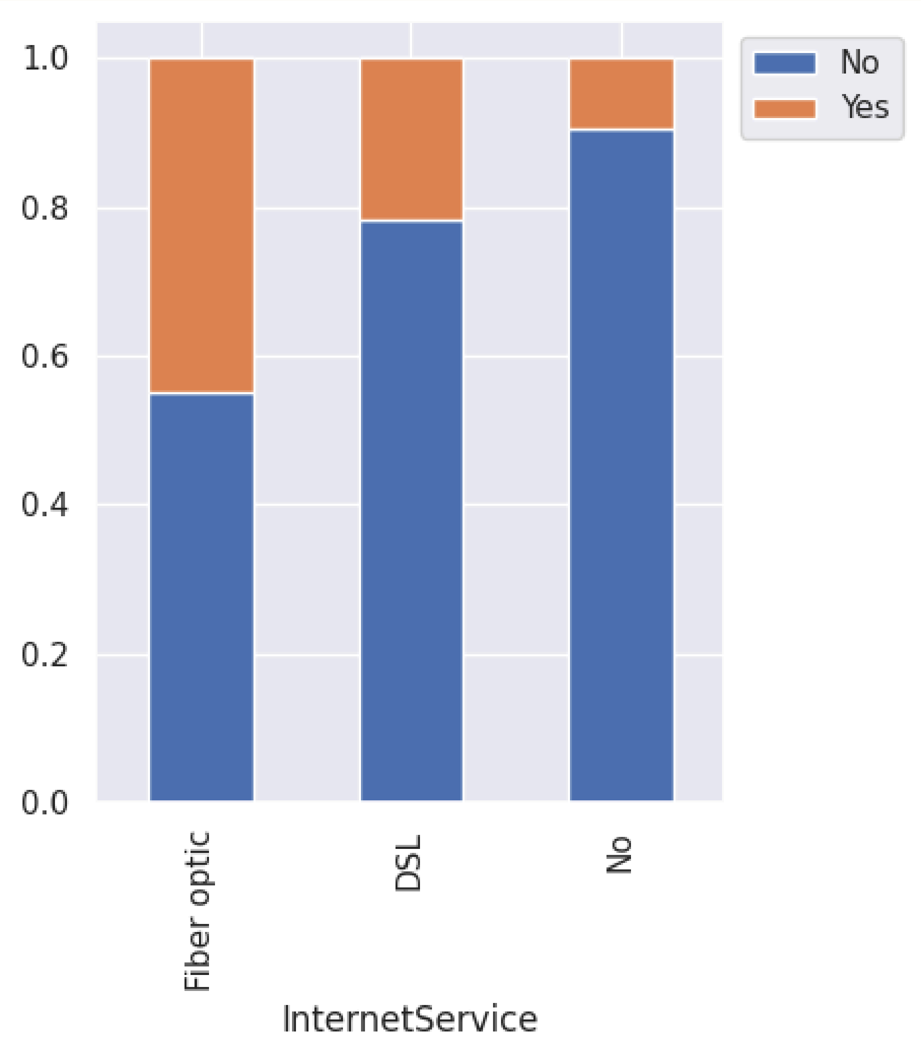


Figure 9: Internet Service and Churn

Observations

- Customers using Fiber optic service show the highest churn proportion, indicating dissatisfaction or higher pricing in this segment.
- DSL users churn less compared to Fiber, suggesting more stability or affordability in that plan type.
- Customers with no internet service have the lowest churn rate, implying churn is more strongly driven by internet-related service experiences than phone-only users.

Correlation Matrix Analysis

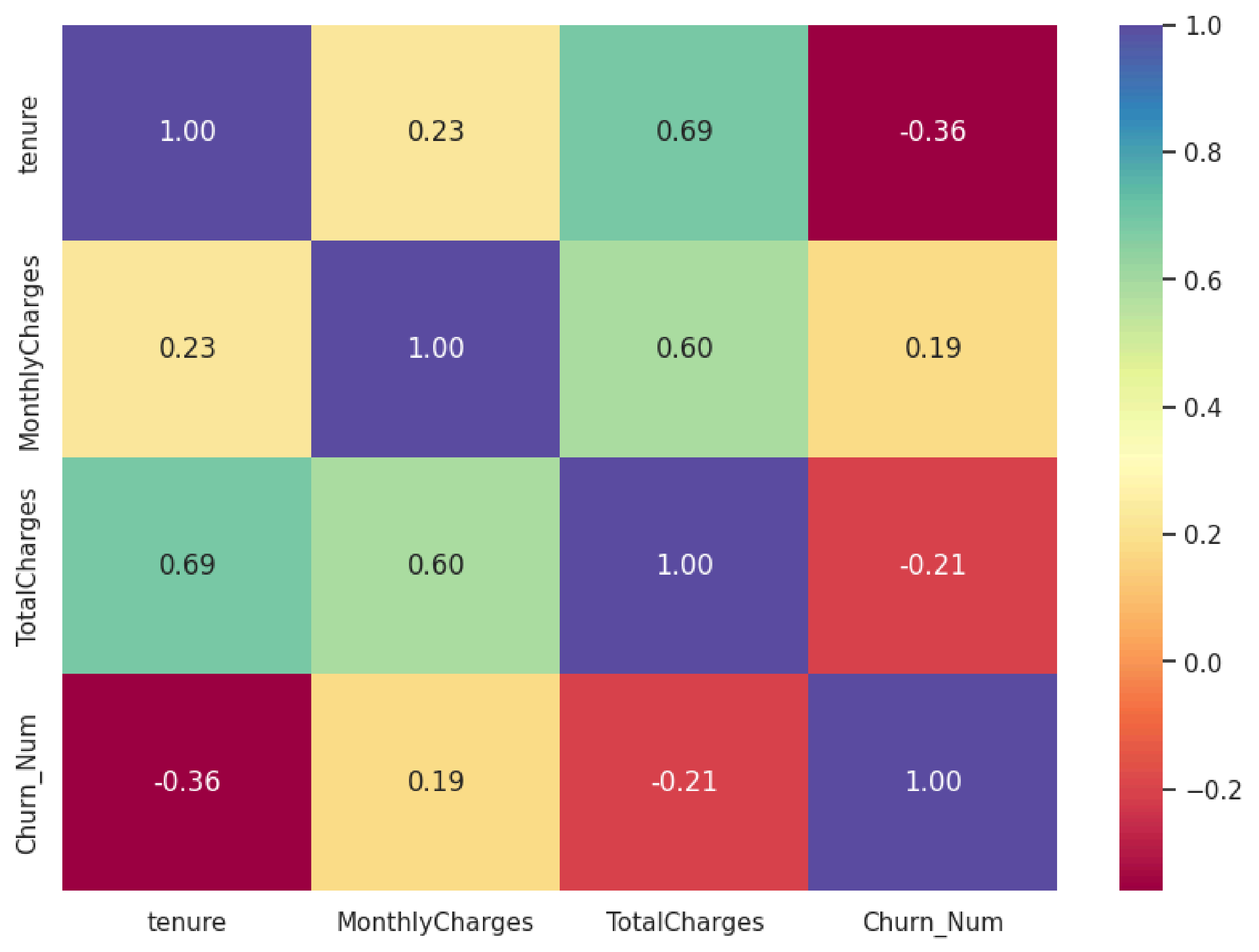


Figure 10: Correlation Matrix Analysis

Observations

- Tenure has the strongest relationship with churn (−0.36) — churn is more common among short-tenure customers.
- TotalCharges and MonthlyCharges are moderately correlated (0.60) and both relate to churn, but tenure is a more dominant driver of attrition.
- The weak positive correlation of MonthlyCharges with churn (0.19) suggests that higher billing may contribute to churn but not as strongly as tenure does.

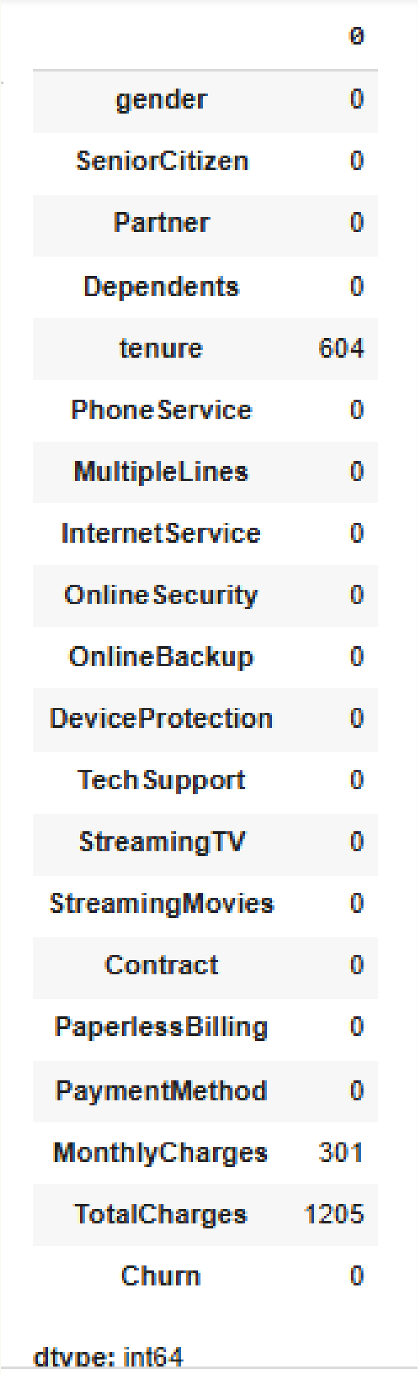
DATA PREPROCESSING

- Splitting the data first into training, validation and test to avoid data leakage.

- After splitting there are 7233 rows in train dataset, 2411 in validation and 2411 in test dataset with 19 clumns.

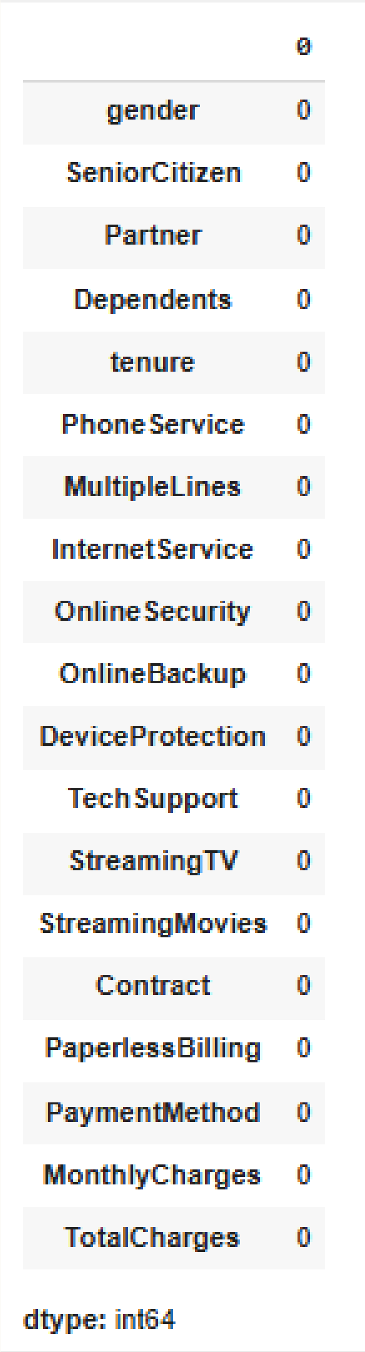
Treating for missing values

Before imputation



```
                      0
          gender      0
   SeniorCitizen      0
         Partner      0
      Dependents      0
          tenure    604
    PhoneService      0
   MultipleLines      0
 InternetService      0
  OnlineSecurity      0
    OnlineBackup      0
DeviceProtection      0
     TechSupport      0
     StreamingTV      0
 StreamingMovies      0
        Contract      0
PaperlessBilling      0
   PaymentMethod      0
  MonthlyCharges    301
    TotalCharges   1205
           Churn      0

dtype: int64
```

After imputation



```
                      0
          gender      0
   SeniorCitizen      0
         Partner      0
      Dependents      0
          tenure      0
    PhoneService      0
   MultipleLines      0
 InternetService      0
  OnlineSecurity      0
    OnlineBackup      0
DeviceProtection      0
     TechSupport      0
     StreamingTV      0
 StreamingMovies      0
        Contract      0
PaperlessBilling      0
   PaymentMethod      0
  MonthlyCharges      0
    TotalCharges      0

dtype: int64
```

Table 6: Treating missing values

- There are 604 missing values in tenure and 301 missing values in Monthly Charges and 1205 missing values in Total Charges.

- Using SimpleImputer to impute the missing values with the "most_frequent" value on the three columns("tenure", "MonthlyCharges","TotalCharges") with the missing values .

Checking for duplicate values

- There are two duplicate rows in the data. But not treating it because it cannot affect the prediction.

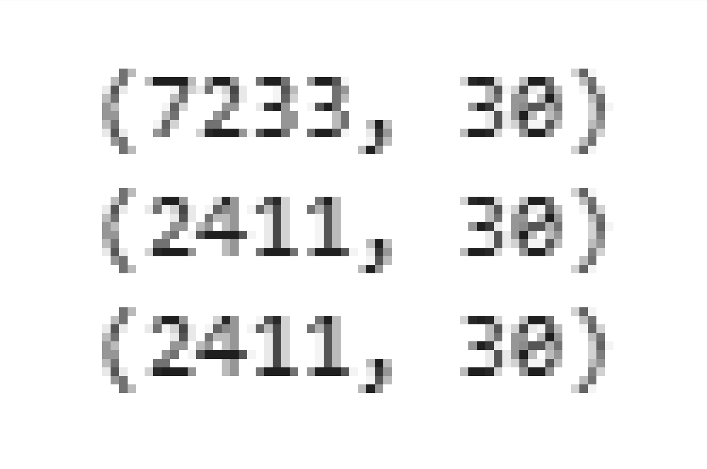
Feature engineering

- **Churn** is the target variable.
- 1. Features (X):
    - Removes the column **"Churn "** (the target variable) from **dataset**.
    - All remaining columns (e.g., "tenure", "MonthlyCharges","TotalCharges") become **input features** for the model.

- 2. Target Labels (y):
    - Extracts "Churn " (Yes/No) and converts it to binary (1/0):
    - **1** → "**Yes**"
    - **0** → "**No**"
    - This is done because machine learning models typically require numerical labels.

One-Hot Hncoding

- Converts categorical variables into dummy/indicator variables:



```
(7233, 30)
(2411, 30)
(2411, 30)
```

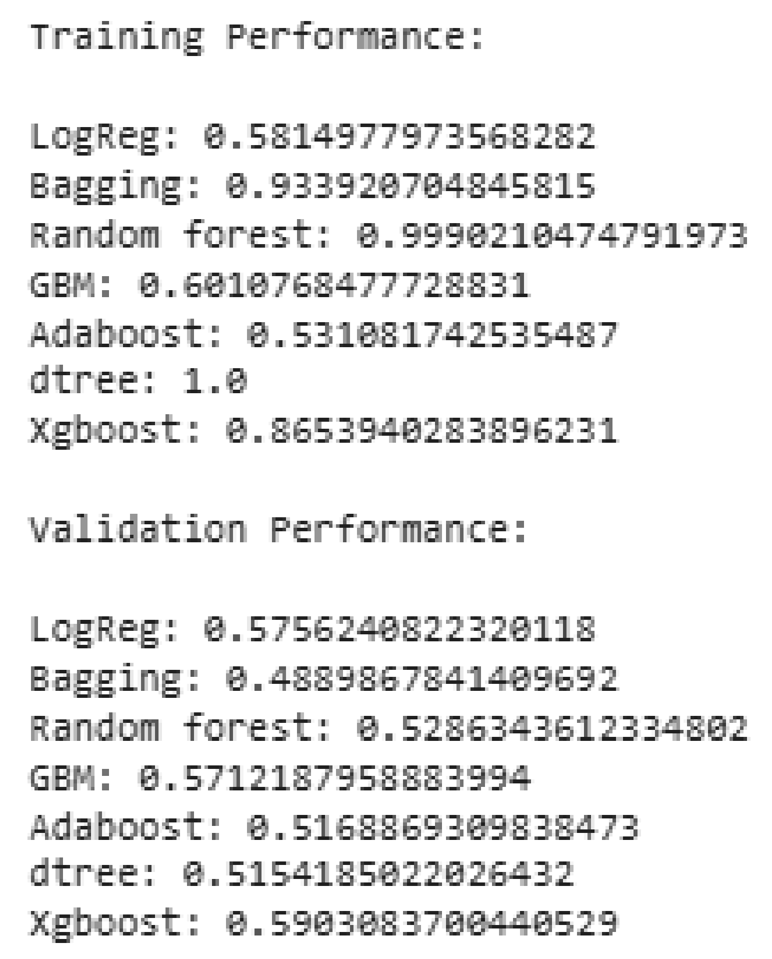
Table 7: Shape after one hot encoding

- After one hot encoding the training set have 7233 rows and 30 columns , validation set has 2411 rows and 30 columns and test set has 2411 rows and 30 columns.

MODEL BUILDING WITH ORGINAL DATA

- Using Logistic regression as a baseline model to predit the metrics

- **Purpose**: Initializes a list of 6 models:
- **Ensemble Methods: Logistic regression, Bagging, Random Forest, Gradient Boosting (GBM), AdaBoost, Decision tree, XGBoost.**
- **Baseline**: Logistic regression



```
Training Performance:

LogReg: 0.5814977973568282
Bagging: 0.933920704845815
Random forest: 0.9990210474791973
GBM: 0.6010768477728831
Adaboost: 0.531081742535487
dtree: 1.0
Xgboost: 0.8653940283896231

Validation Performance:

LogReg: 0.5756240822320118
Bagging: 0.4889867841409692
Random forest: 0.5286343612334802
GBM: 0.5712187958883994
Adaboost: 0.5168869309838473
dtree: 0.5154185022026432
Xgboost: 0.5903083700440529
```

Table 8: Training (CV) and Validation score

Overall Model Performance (Baseline Logistic Regression)

- **Overfitting is evident in Random Forest**, Decision Tree, and Bagging — training accuracy is near 1.0 but validation drops sharply.
- **Logistic Regression and AdaBoost** show the smallest train–validation gap, indicating better generalization.
- **GBM and XGBoost** perform decently but still need tuning (regularization or fewer estimators) to reduce overfitting.

Hyperparameter Tuning

Tuning AdaBoost model with Original data

Finding best parameters

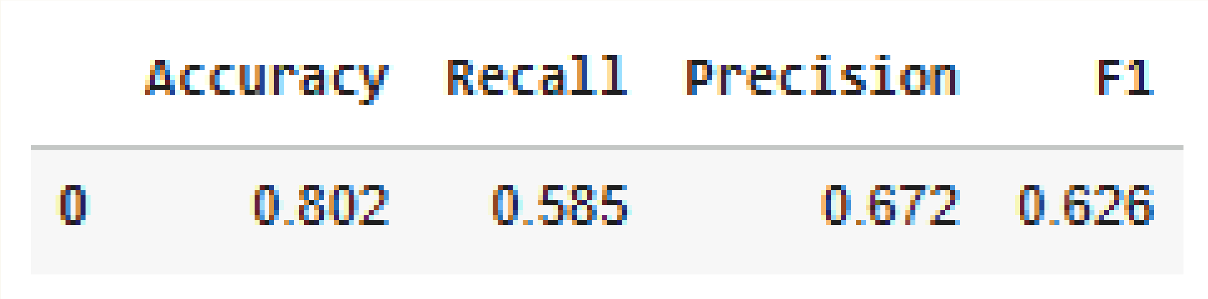
- Performing hyperparameter tuning for a Adaboost using GridSearchCV

- Using GridSearchCV Setup

Best parameters

- 'estimator': DecisionTreeClassifier(max_depth=3, random_state=1), 'learning_rate': 1.0, 'n_estimators': 100

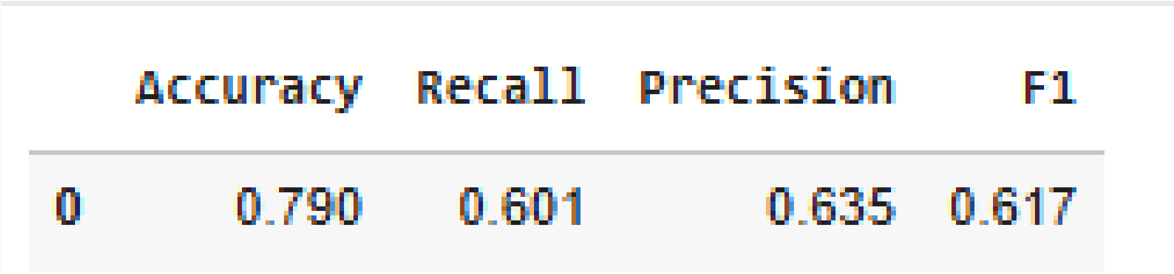
Checking model's performance on training set



	Accuracy	Recall	Precision	F1
0	0.802	0.585	0.672	0.626

Table 9: Training performance

Checking model's performance on Validation set



	Accuracy	Recall	Precision	F1
0	0.790	0.601	0.635	0.617

Table 10: Validation performance

Insights

1. Training and validation scores are very close, showing that the model is well-generalized with minimal overfitting.
2. Validation accuracy (0.79) and F1-score (0.617) indicate a balanced trade-off between precision and recall.
3. Slight improvement in validation recall (0.601 → 0.585) suggests the model captures positives better on unseen data.

Tuning Logistic Regression model with original data

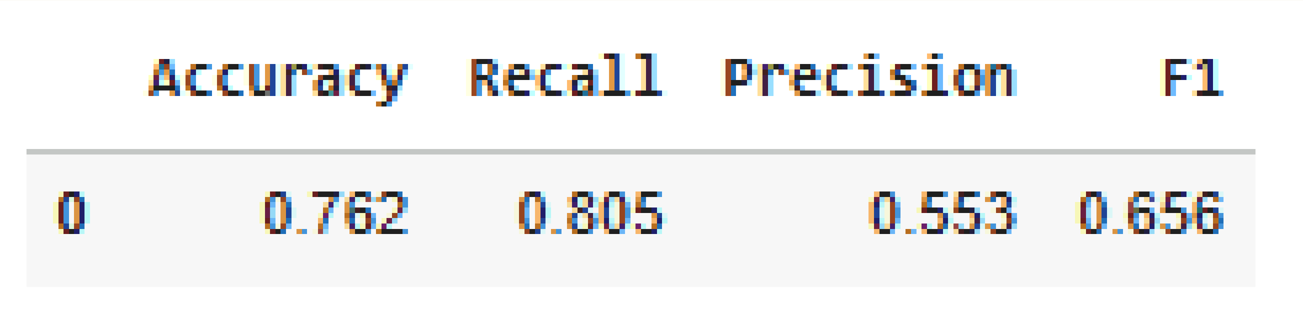
Finding best parameters

- Performing for a Logistic Regression using **RandomizedSearchCV**

Best parameters

- 'penalty': 'l1', 'class_weight': 'balanced', 'C': np.float64(0.3359818286283781)

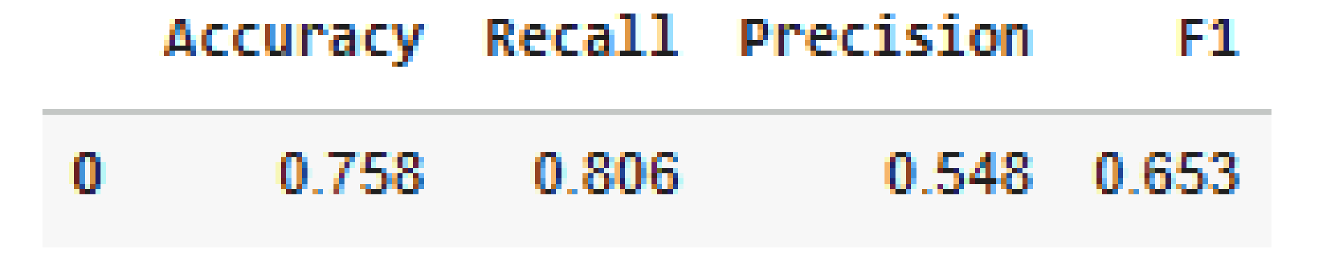
Checking model's performance on training set



	Accuracy	Recall	Precision	F1
0	0.762	0.805	0.553	0.656

Table 11: Training performance

Checking model's performance on Validation set



	Accuracy	Recall	Precision	F1
0	0.758	0.806	0.548	0.653

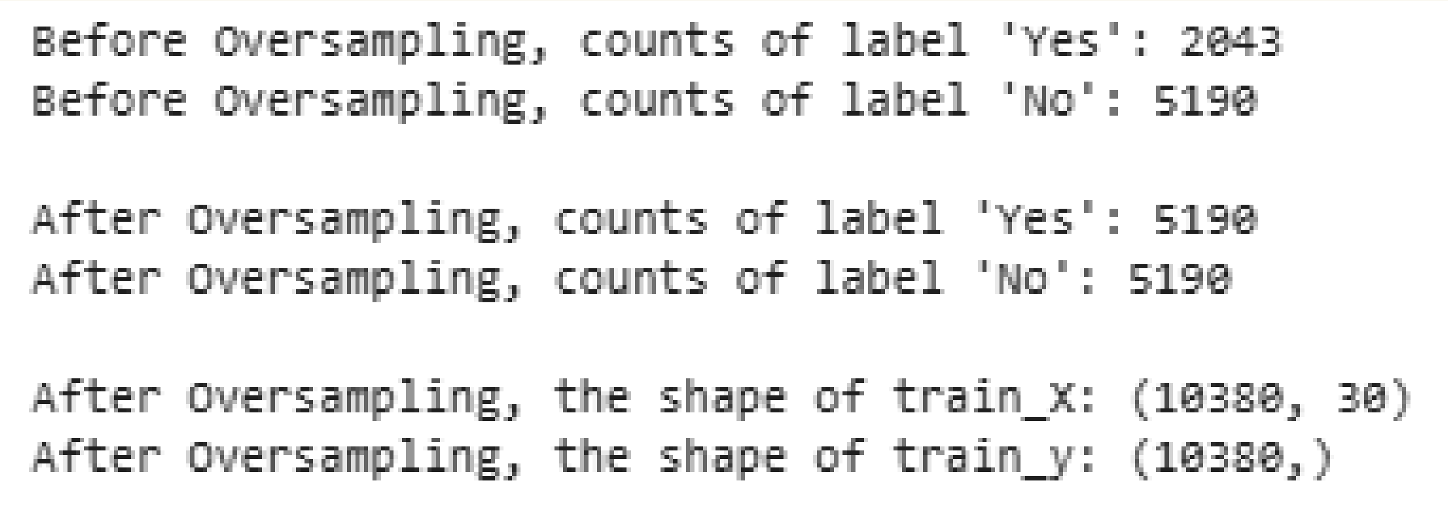
Table 12: Validation performance

Insights

1. Training and validation metrics are almost identical, showing the model is highly stable and not overfitting.
2. High recall (~0.80) indicates the model is excellent at capturing positive cases, though precision is moderate.
3. Balanced F1-scores (~0.65) across both sets confirm consistent performance and strong generalization ability.

Model Building - Oversampled Data

- Using **SMOTE** to resample the dataset.

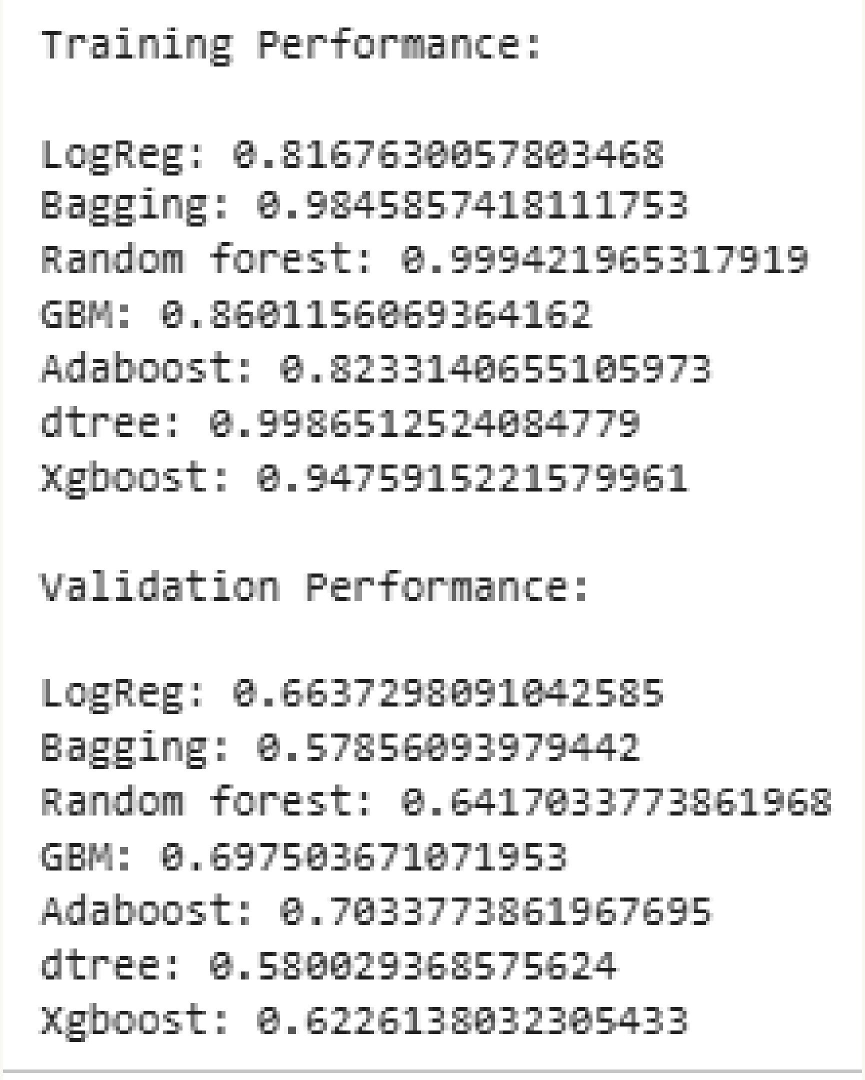


```
Before Oversampling, counts of label 'Yes': 2043
Before Oversampling, counts of label 'No': 5190

After Oversampling, counts of label 'Yes': 5190
After Oversampling, counts of label 'No': 5190

After Oversampling, the shape of train_X: (10380, 30)
After Oversampling, the shape of train_y: (10380,)
```

Table 13: Oversampled data

1. The dataset was imbalanced initially (‘Yes’: 2043 vs. ‘No’: 5190), which could bias the model toward the majority class.
2. After oversampling, both classes are equal (5190 each), ensuring balanced learning and fair prediction of both outcomes.
3. Final training data shape (10,380 × 30) confirms the synthetic samples were added correctly without altering feature count.

Training (CV) and Validation score



```
Training Performance:

LogReg: 0.8167630057803468
Bagging: 0.9845857418111753
Random forest: 0.999421965317919
GBM: 0.8601156069364162
Adaboost: 0.8233140655105973
dtree: 0.9986512524084779
Xgboost: 0.9475915221579961

Validation Performance:

LogReg: 0.6637298091042585
Bagging: 0.57856093979442
Random forest: 0.6417033773861968
GBM: 0.697503671071953
Adaboost: 0.7033773861967695
dtree: 0.580029368575624
Xgboost: 0.6226138032305433
```

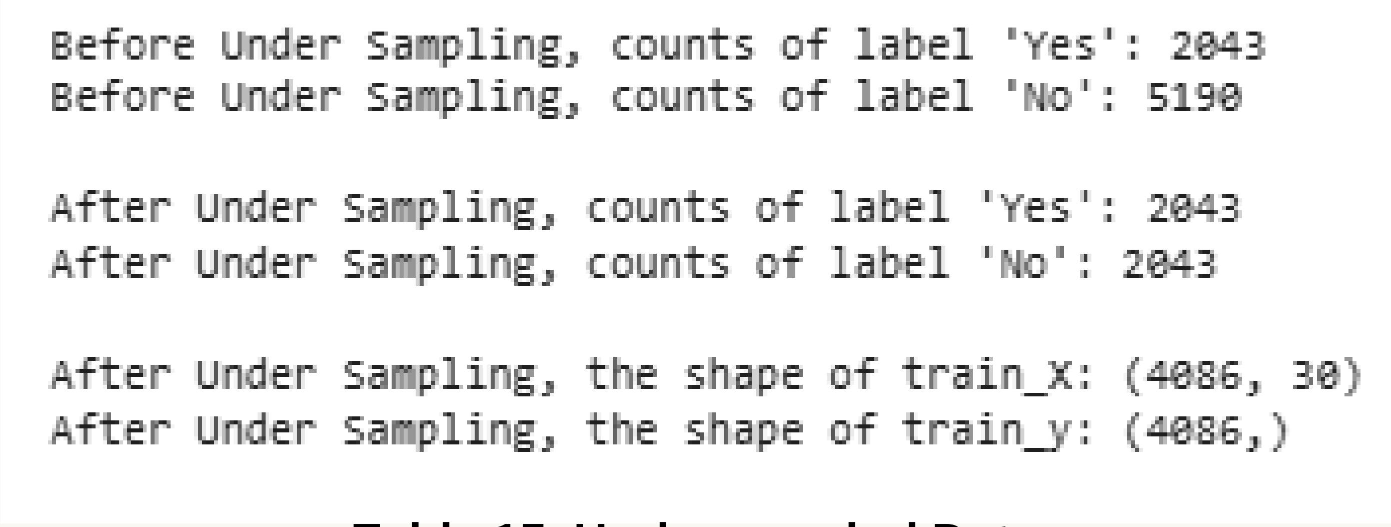
Table 14: Training (CV) and Validation score

Insights

1. Models like Random Forest and Decision Tree show extremely high training scores but large validation drops — clear overfitting.
2. Logistic Regression and AdaBoost maintain closer train–validation scores, indicating better generalization.
3. GBM achieves a good balance with strong validation performance (≈0.70), making it the most reliable model among the group.

Model Building - Under sampled Data

- Using **RandomUnderSampler** to resample the dataset.



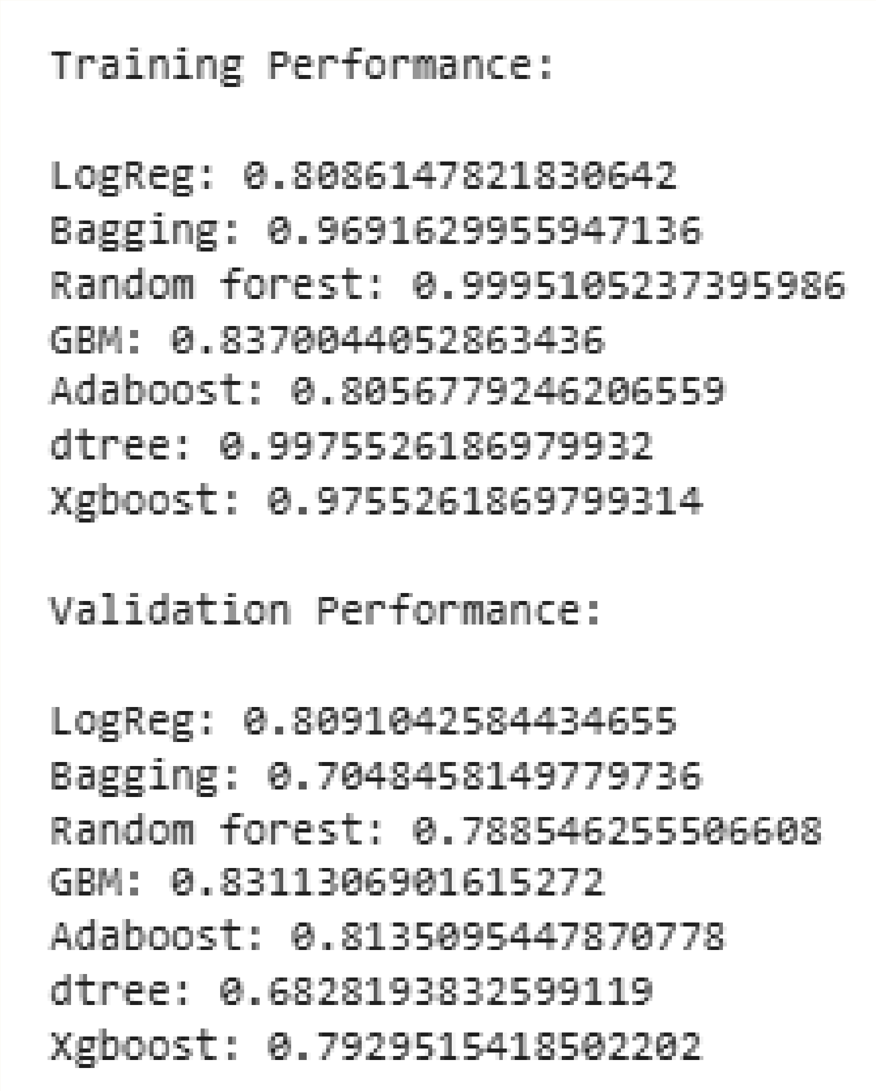
```
Before Under Sampling, counts of label 'Yes': 2043
Before Under Sampling, counts of label 'No': 5190

After Under Sampling, counts of label 'Yes': 2043
After Under Sampling, counts of label 'No': 2043

After Under Sampling, the shape of train_X: (4086, 30)
After Under Sampling, the shape of train_y: (4086,)
```

Table 15: Undersampled Data

Insights

- The dataset was imbalanced initially (‘Yes’: 2043 vs. ‘No’: 5190), favoring the majority ‘No’ class.
- After undersampling, both classes are equal (2043 each), creating a balanced dataset for fair model learning.
- Final training shape (4,086 × 30) confirms that the majority samples were reduced, helping prevent bias but reducing data volume.

Training (CV) and Validation score



```
Training Performance:

LogReg: 0.8086147821830642
Bagging: 0.9691629955947136
Random forest: 0.9995105237395986
GBM: 0.8370044052863436
Adaboost: 0.8056779246206559
dtree: 0.9975526186979932
Xgboost: 0.9755261869799314

Validation Performance:

LogReg: 0.8091042584434655
Bagging: 0.7048458149779736
Random forest: 0.7885462555066608
GBM: 0.8311306901615272
Adaboost: 0.8135095447870778
dtree: 0.6828193832599119
Xgboost: 0.7929515418502202
```

Table 16: Training (CV) and Validation score

Insights

1. Training and validation scores are well-aligned, showing models have generalized effectively after undersampling.
2. XGBoost and GBM deliver the best balance, with validation scores around 0.79–0.83, indicating strong and stable performance.
3. Random Forest and Decision Tree still show mild overfitting but improved significantly compared to the oversampled data.

Hyperparameter Tuning the Gradient Boosting Classifier with under sample data

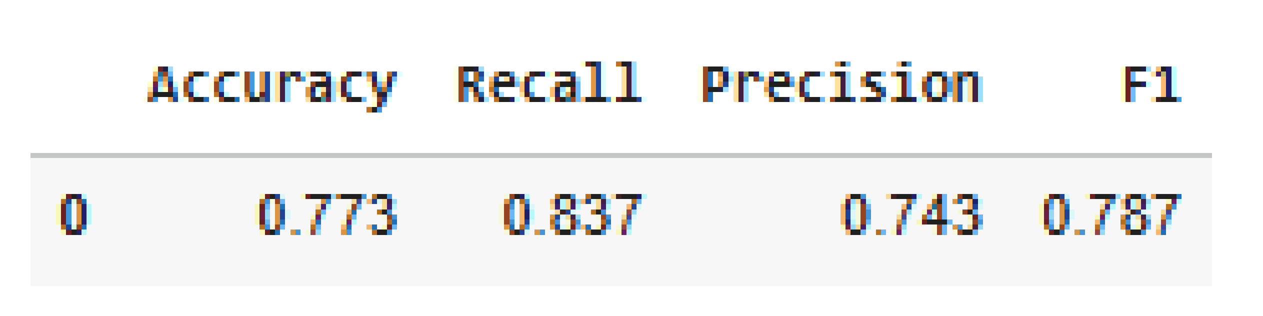
Finding best parameters

- Performing hyperparameter tuning for a Gradient Boosting Classifier using RandomizedSearchCV

Best parameters

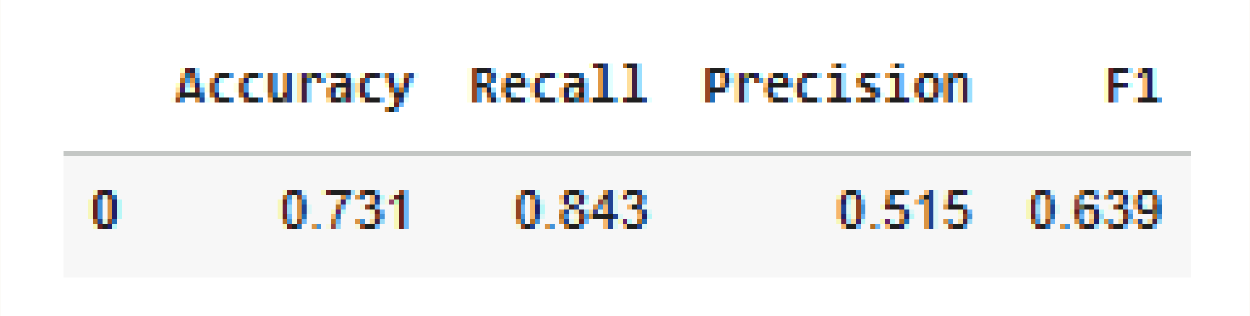
- 'subsample': 0.8, 'n_estimators': np.int64(150), 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': 2, 'learning_rate': 0.01

Checking model's performance on Training set and Validation Set



	Accuracy	Recall	Precision	F1
0	0.773	0.837	0.743	0.787

Table 17: Training performance on Tuning



	Accuracy	Recall	Precision	F1
0	0.731	0.843	0.515	0.639

Table 18: Validation performance on Tuning

Insights

- The model shows high recall (~0.84) on both sets, meaning it effectively captures most positive cases.
- Precision drops on validation (0.515), suggesting more false positives after tuning.
- Overall, good recall–precision trade-off but needs fine-tuning (e.g., threshold or regularization) to improve validation precision.

Hyperparameter Tuning the Ada Boosting Classifier with under sample data

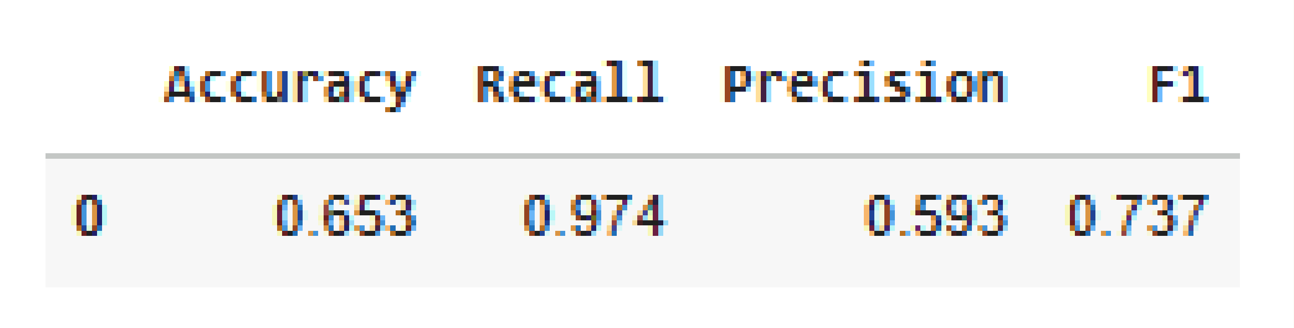
Finding best parameters

- Performing hyperparameter tuning for a Ada Boosting Classifier using **GridSearchCV**

Best parameters

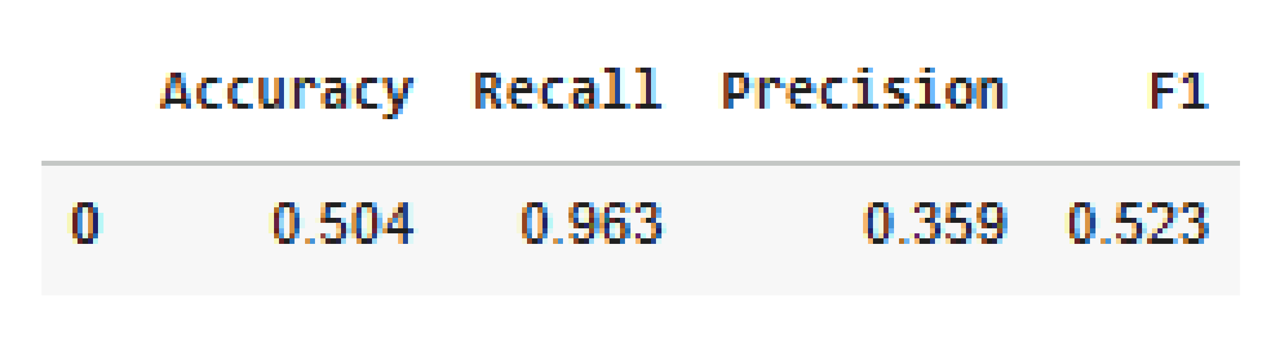
- 'estimator': DecisionTreeClassifier(max_depth=1, random_state=1), 'learning_rate': 0.01, 'n_estimators': 50

Checking model's performance on Training set and Validation Set



	Accuracy	Recall	Precision	F1
0	0.653	0.974	0.593	0.737

Table 19: Training performance(Adaboosting)



	Accuracy	Recall	Precision	F1
0	0.504	0.963	0.359	0.523

Table 20: Validation performance(Adaboosting)

Insights

- Recall is extremely high (~0.96) on both sets, meaning the model identifies nearly all positives — excellent sensitivity.
- However, precision is quite low (0.59 → 0.36), indicating many false positives after tuning.
- The model is overly recall-focused, so adjusting the decision threshold or using precision-weighted tuning could improve balance.

Hyperparameter Tuning the Logistic Regression with Undersampled data

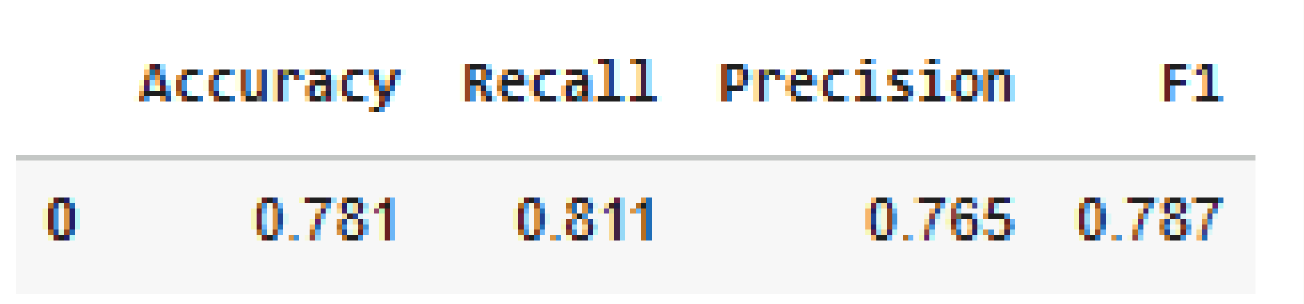
Finding best parameters

- Performing hyperparameter tuning for a Logistic Regression Classifier using RandomizedSearchCV

Best parameters

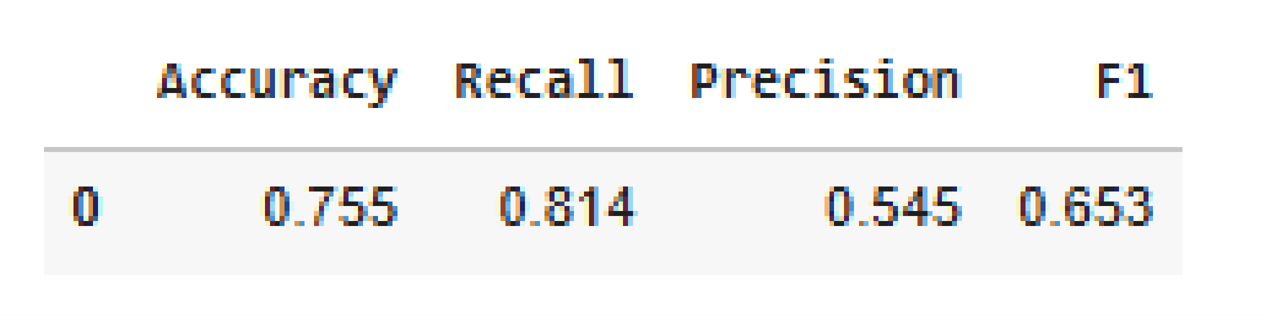
- 'penalty': 'l1', 'class_weight': None, 'C': np.float64(0.6951927961775606

Checking model's performance on Training set and Validation Set



	Accuracy	Recall	Precision	F1
0	0.781	0.811	0.765	0.787

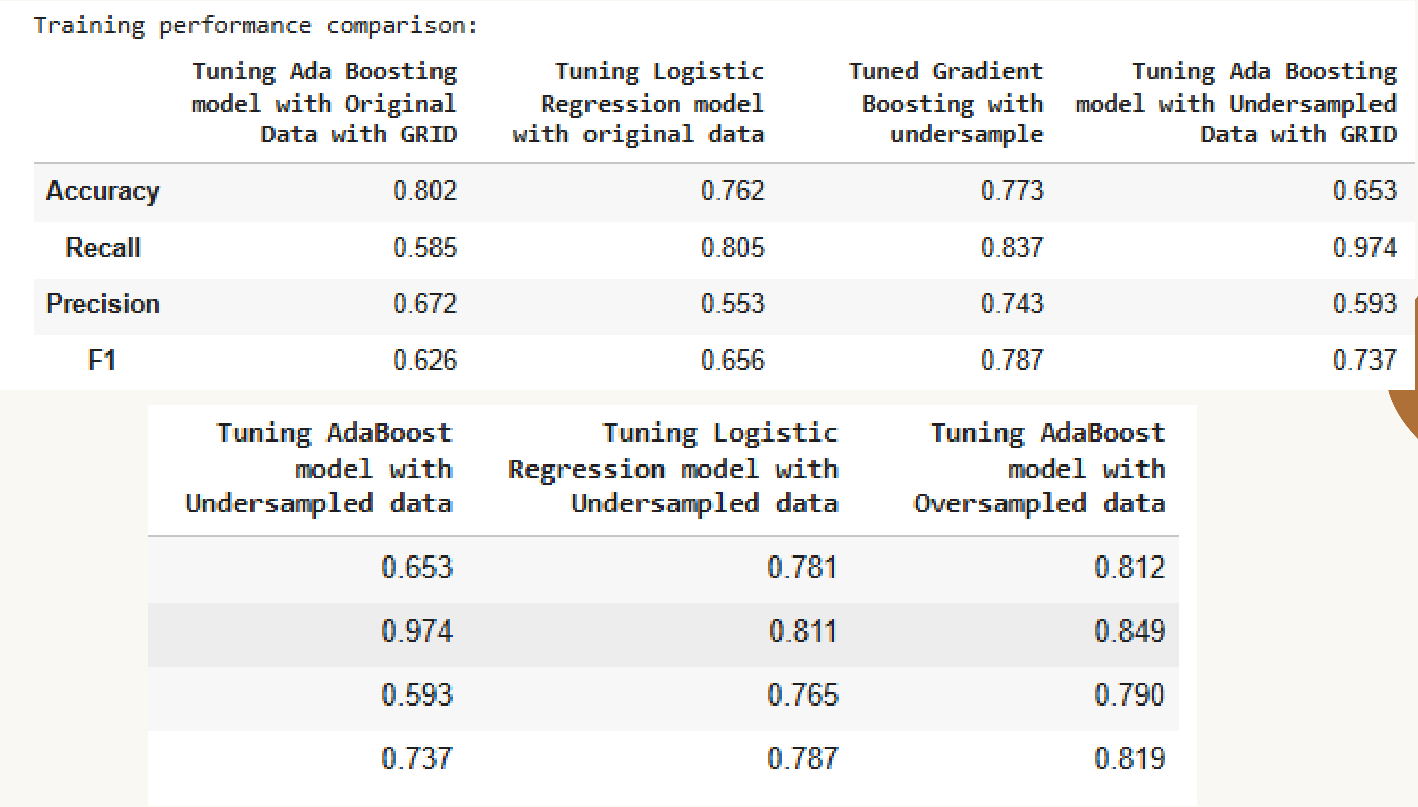
Table 21: Training performance(Logistic Reg)



	Accuracy	Recall	Precision	F1
0	0.755	0.814	0.545	0.653

Table 22: Validation performance (Log Reg)

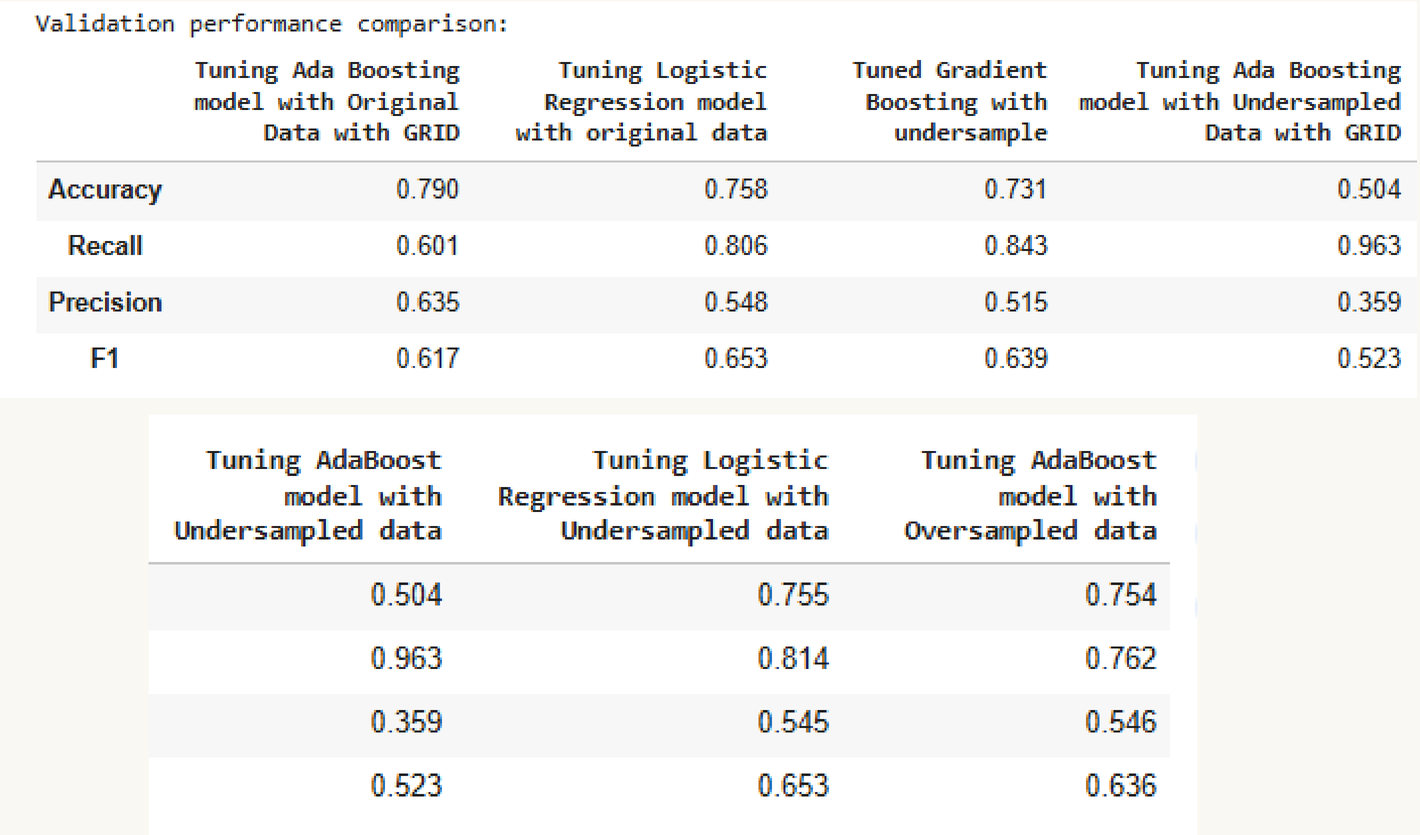
Model Performance Comparison



Training performance comparison:

	Tuning Ada Boosting model with Original Data with GRID	Tuning Logistic Regression model with original data	Tuned Gradient Boosting with undersample	Tuning Ada Boosting model with Undersampled Data with GRID	Tuning AdaBoost model with Undersampled data	Tuning Logistic Regression model with Undersampled data	Tuning AdaBoost model with Oversampled data
Accuracy	0.802	0.762	0.773	0.653	0.653	0.781	0.812
Recall	0.585	0.805	0.837	0.974	0.974	0.811	0.849
Precision	0.672	0.553	0.743	0.593	0.593	0.765	0.790
F1	0.626	0.656	0.787	0.737	0.737	0.787	0.819

Table 23: Training performance comparison



Validation performance comparison:

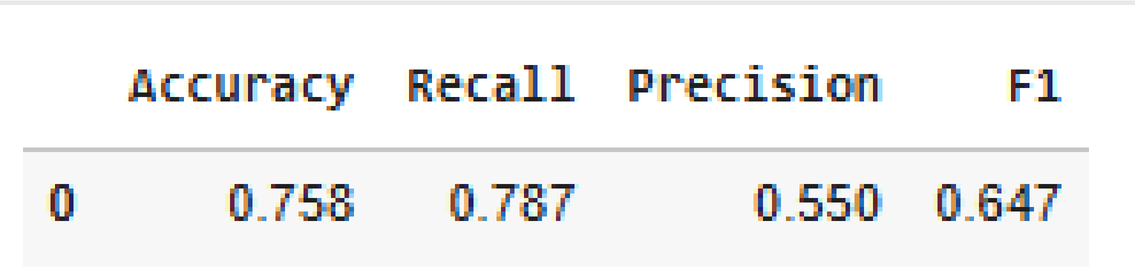
	Tuning Ada Boosting model with Original Data with GRID	Tuning Logistic Regression model with original data	Tuned Gradient Boosting with undersample	Tuning Ada Boosting model with Undersampled Data with GRID	Tuning AdaBoost model with Undersampled data	Tuning Logistic Regression model with Undersampled data	Tuning AdaBoost model with Oversampled data
Accuracy	0.790	0.758	0.731	0.504	0.504	0.755	0.754
Recall	0.601	0.806	0.843	0.963	0.963	0.814	0.762
Precision	0.635	0.548	0.515	0.359	0.359	0.545	0.546
F1	0.617	0.653	0.639	0.523	0.523	0.653	0.636

Table 24: Validation performance comparison

Choosing Best Model

- **Logistic regression with original data** gives the best score so far, so using the Logistic regression with original data on the test data.

Checking model's performance on Test set



	Accuracy	Recall	Precision	F1
0	0.758	0.787	0.550	0.647

Table 25: Test Performance

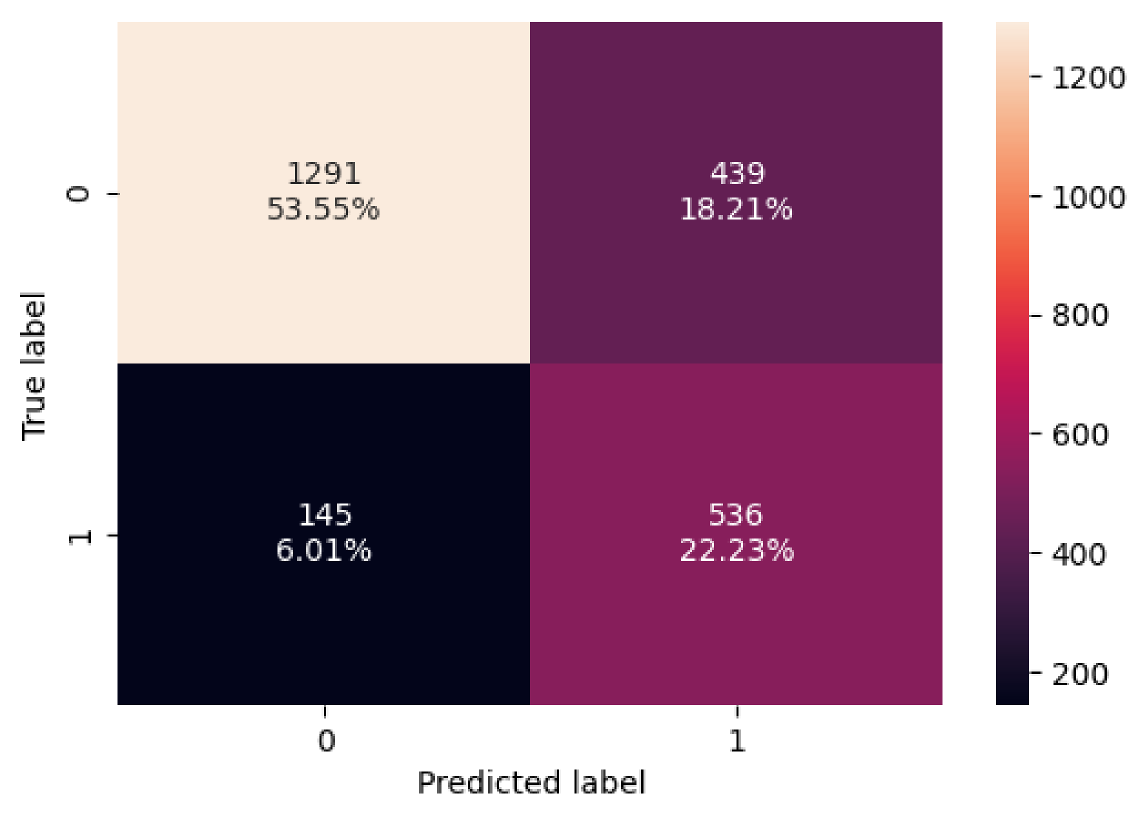


Figure 12: Confusion Matrix Logistic Regression Model

Test Performance Summary

- **Accuracy (0.758)** shows the model maintains strong generalization on unseen data — no major overfitting.

- **Recall (0.787)** indicates the model successfully captures most true positives, performing well on the target class.

- **Precision (0.550**) suggests some false positives remain, meaning the model favors recall slightly over precision.

- The **F1-score (0.647)** confirms a good overall balance between recall and precision, showing that the model generalizes effectively but could benefit from threshold or precision-oriented tuning for better reliability.

Insights

1. The model correctly classified 1291 true negatives (53.55%) and 536 true positives (22.23%), showing good recognition of both classes.
2. 145 false negatives (6.01%) are relatively low — the model is strong at identifying actual positives.
3. 439 false positives (18.21%) indicate some misclassification of negatives as positives, reducing precision.
4. Overall, the confusion matrix shows balanced performance with slightly higher recall strength than precision — useful for recall-sensitive tasks.

Feature Importance

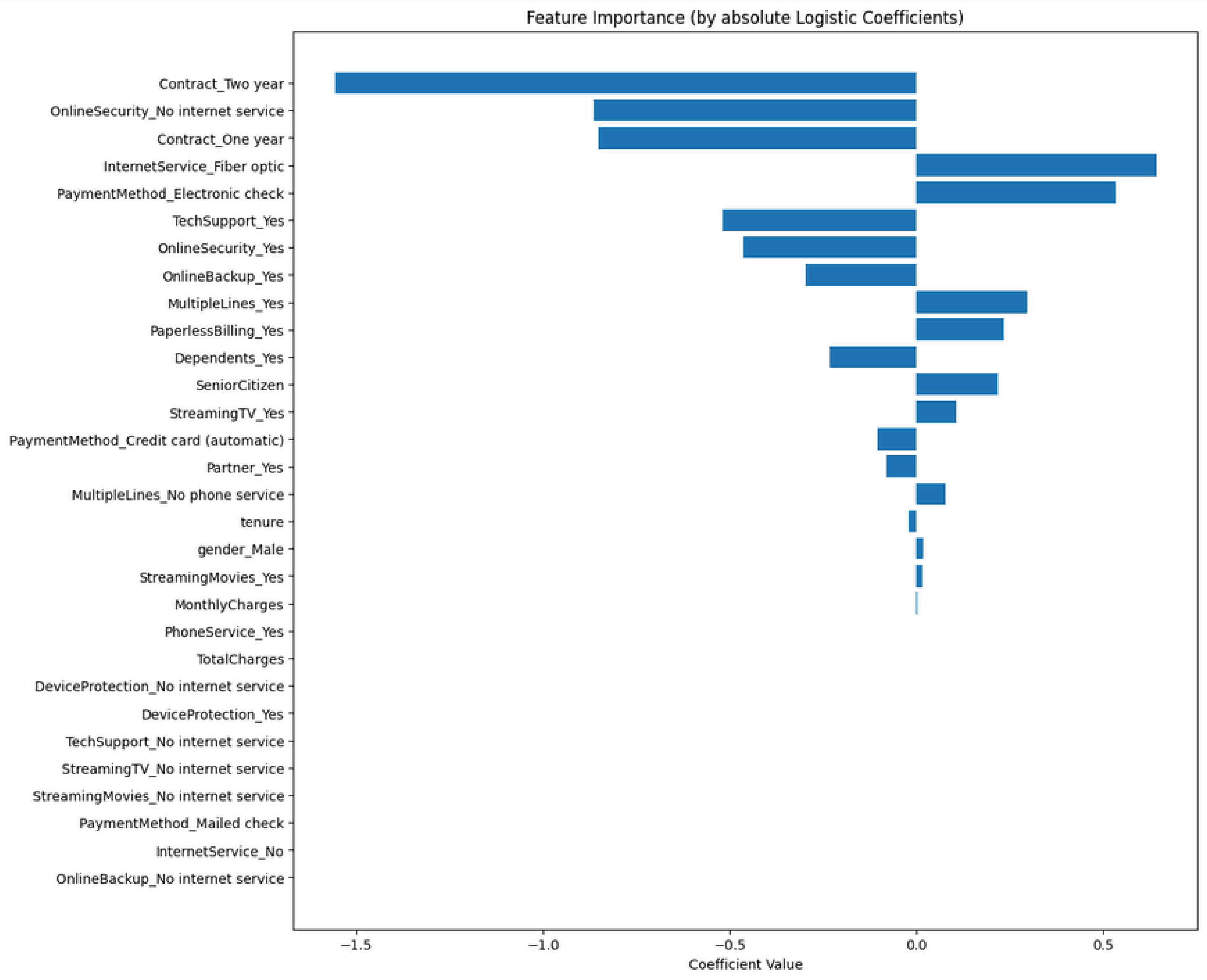


Figure 13: Feature Importance

Key Factors that REDUCE Churn (customers less likely to leave)

1. Contract type has the strongest impact

- "Contract_Two year" and "Contract_One year" have the largest negative coefficients, meaning longer contracts reduce the likelihood of churn.
- Customers with short-term or month-to-month contracts are more likely to leave.

2. Internet service and online security matter significantly

- “InternetService_Fiber optic” has a positive coefficient, indicating higher churn risk — likely due to higher cost or performance issues.
- Conversely, “OnlineSecurity_Yes” and “TechSupport_Yes” have negative coefficients, meaning customers with these services are less likely to churn.

3. Payment methods influence retention

- “PaymentMethod_Electronic check” strongly increases churn — these users are more likely to cancel.
- Credit card or automatic payments tend to decrease churn, reflecting greater stability.

4. Customer behavior and demographics

- Features like “PaperlessBilling_Yes” and “SeniorCitizen_Yes” have moderate positive coefficients, suggesting these groups churn slightly more.
- “Dependents_Yes” and “Partner_Yes” show negative coefficients — customers with families or partners are more loyal.

Overall Interpretation

- Customer loyalty improves with longer contracts, stable payment methods, and support/security services, while flexible contracts and electronic payments correlate with higher churn risk.

Insights from the analysis conducted and Actionable business recommendations

- **Contract Duration is the strongest churn driver.**
  - Customers with month-to-month contracts show significantly higher churn.
  - Longer-term contracts (one or two years) greatly reduce churn probability, as shown by strong negative coefficients.

- **Service-related features affect loyalty.**
  - Customers without Online Security or Tech Support are more likely to churn.
  - Those with Fiber Optic Internet churn more, possibly due to price sensitivity or service quality issues.

- **Payment method influences retention.**
  - Electronic check users have a high churn rate — these are likely less digitally engaged or short-term customers.
  - Credit card (auto-pay) users churn less, indicating greater convenience and commitment reduce cancellations.

- **Customer demographics and lifestyle matter.**
  - Senior citizens and paperless billing users show slightly higher churn.
  - Customers with dependents or partners are more stable, possibly due to shared service needs or household reliance.

- **Model performance is stable and generalizable.**
    - Accuracy ≈ 0.75 and F1 ≈ 0.65 on test data indicate balanced performance between detecting churners and minimizing false alarms.
    - Confusion matrix shows strong recall, meaning the model effectively identifies most churners.

Actionable Business Recommendations

Focus Area	Recommendation	Expected Impact
Contracts	Encourage **longer-term plans** with discounts or added benefits (e.g., 6-month or 12-month commitments).	Reduce churn by increasing customer stickiness.
Value-added Services	Bundle **Online Security** and **Tech Support** as free or discounted add-ons for at-risk customers.	Improve retention by increasing perceived value.
Payment Options	Promote **auto-pay via credit/debit card**; offer small incentives for switching from electronic checks.	Lower churn by simplifying payments.
Targeted Retention Campaigns	Use churn probabilities to **proactively reach at-risk customers** (e.g., short-term contract + electronic check users).	Prevent churn before cancellation occurs.

Focus Area	Recommendation	Expected Impact
Customer Segmentation	Create **separate engagement strategies** for senior citizens and younger users (e.g., phone-based support vs. digital self-service).	Improve satisfaction across customer groups.
Service Quality Review	Investigate **Fiber Optic service feedback** — pricing or network issues could be driving dissatisfaction.	Strengthen loyalty in a high-value customer segment.

Table 26: Business Recommendation

Strategic Summary

- The churn model reveals that contract type, service features, and payment methods are the primary levers of customer retention.
- By focusing on longer contracts, better service bundling, and convenient billing, the company can significantly lower churn and improve lifetime value.