Customer Churn Prediction

BUSINESS REPORT

PGP DSA 18 JAN 2026

ANJU SAINI

INDEX

Graphs

1) **Understanding the Business Problem**

1.1) Defining the problem statement

Customer churn is a major problem for businesses because losing customers leads to loss of revenue. The company does not clearly know why customers stop using its services. This project aims to analyze customer data to identify the main factors that influence customer churn and help the business understand customer behavior better.

1.2) Need of the study/project

- Help to recover losses early
- changes on policies and create some new one so it may attract some new customer
- Targeted retention strategies can be created for high risk customers

1.3) Understanding business

AlphaCom is facing an increase in customer churn , where many customers are leaving the company. This is a serious issue because losing customers reduces revenue and harms the company position in the market.

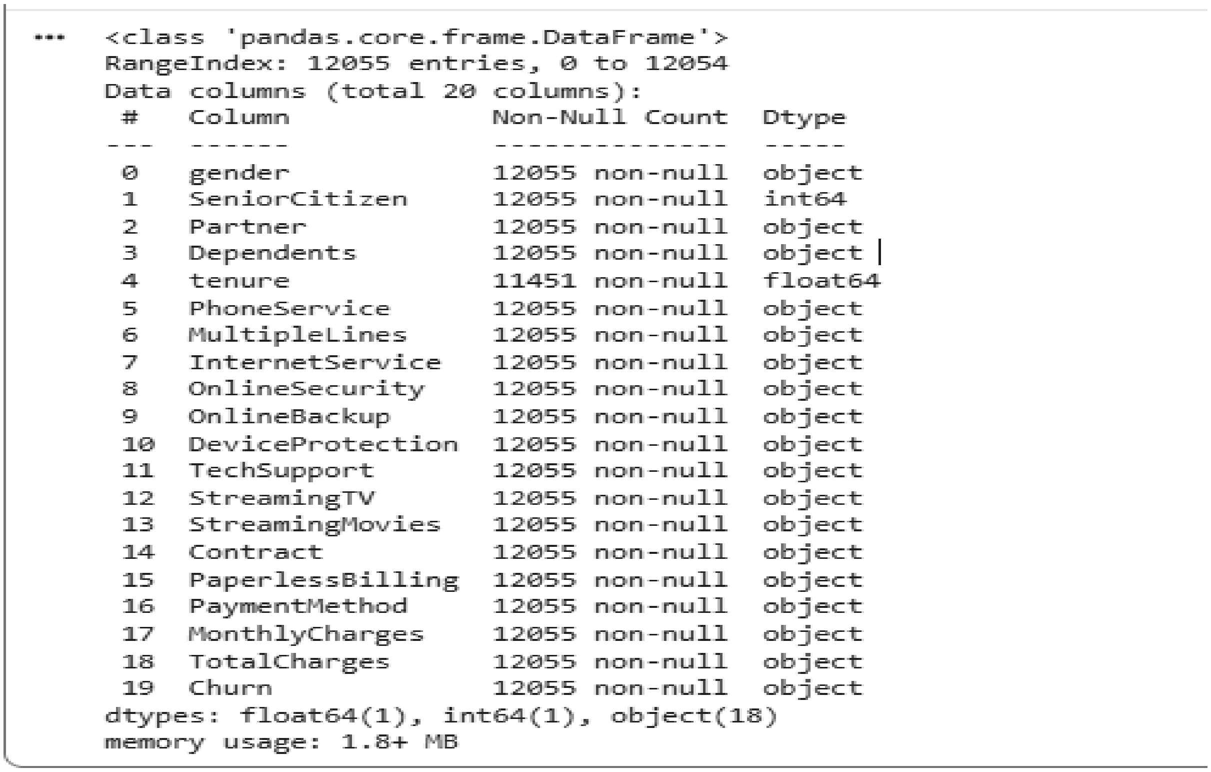
customer churn happens because of many reasons such as high price , contract type , not satisfied with company policy and services . So now companies really want to know what is the main reason behind this so they can stop churn.

1.4) Business / Social Opportunity

By understanding the reasons behind customer churn, the business can take steps to retain existing customers. This creates an opportunity to improve customer satisfaction, reduce revenue loss, and build long-term customer relationships. The insights from this analysis can help the company make better decisions and improve its services for customers.

2) Exploratory Data Analysis

2.1) Data overview

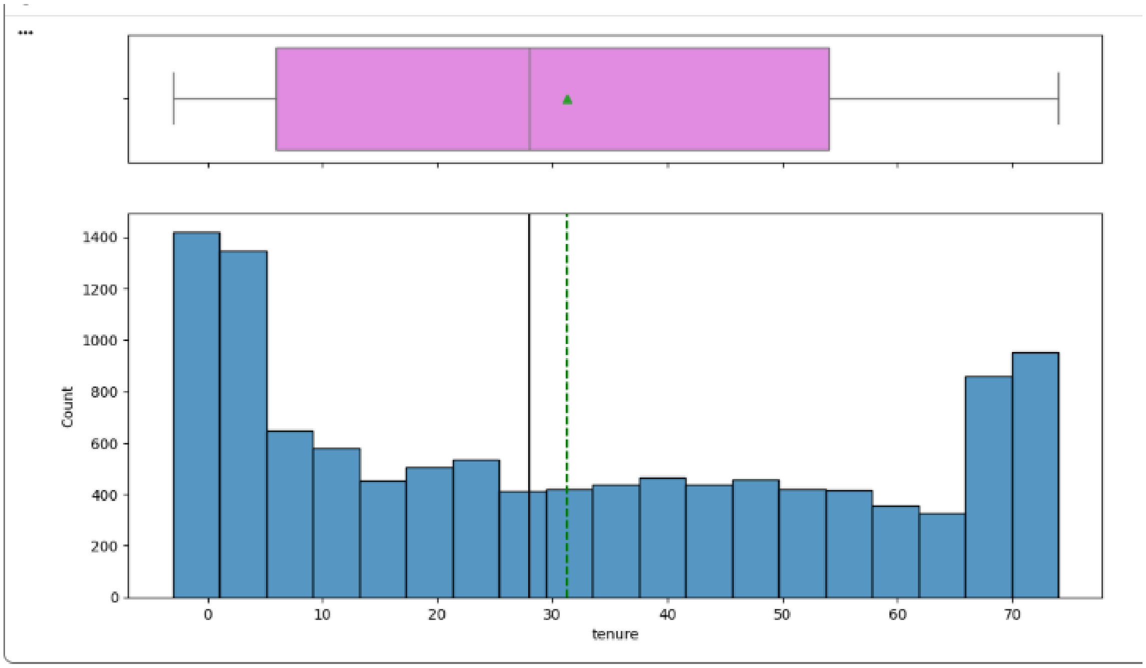


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12055 entries, 0 to 12054
Data columns (total 20 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   gender            12055 non-null  object
 1   SeniorCitizen     12055 non-null  int64
 2   Partner           12055 non-null  object
 3   Dependents        12055 non-null  object
 4   tenure            11451 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  object
 18  TotalCharges      12055 non-null  object
 19  Churn             12055 non-null  object
dtypes: float64(1), int64(1), object(18)
memory usage: 1.8+ MB
```

(1.1)

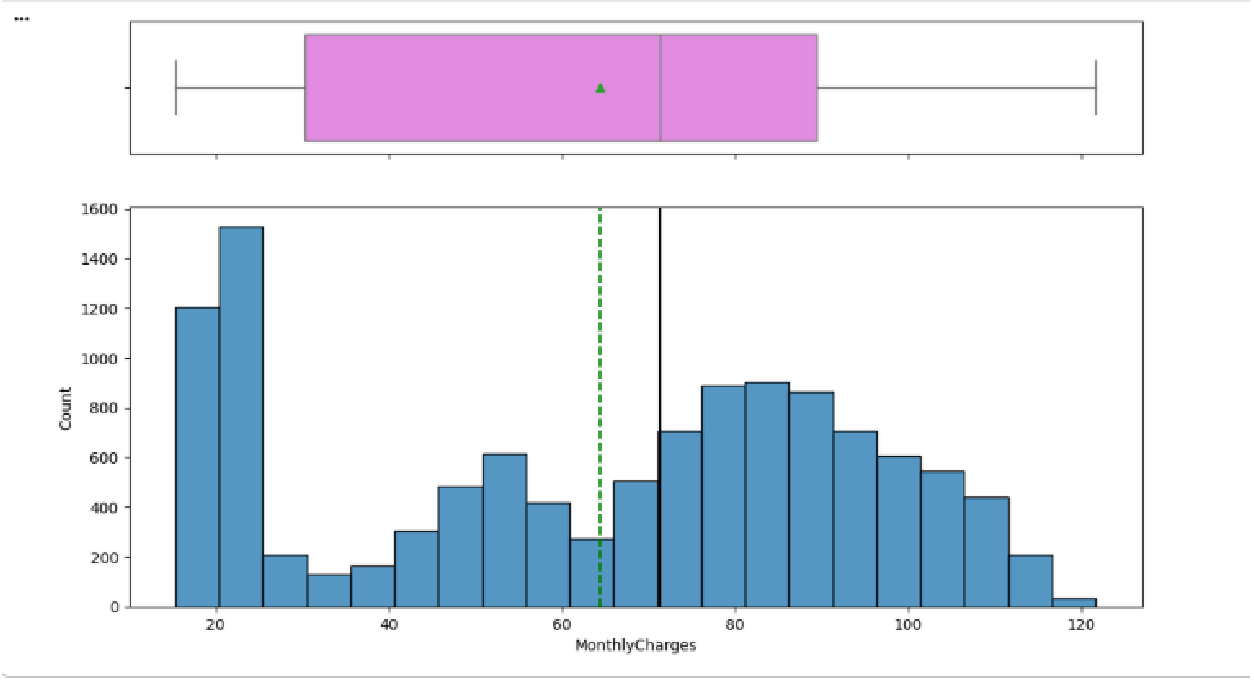
This dataset has 12055 rows and 20 columns.In that 18 is categorical type and 2 is numerical type .There is some missing value in tenure.

2.2) Univariate analysis



(1.2)

The boxplot shows no significant outliers for tenure. The median tenure is lower than the average tenure, which indicates that many customers leave early, while a smaller group of loyal customers stays for a much longer period.

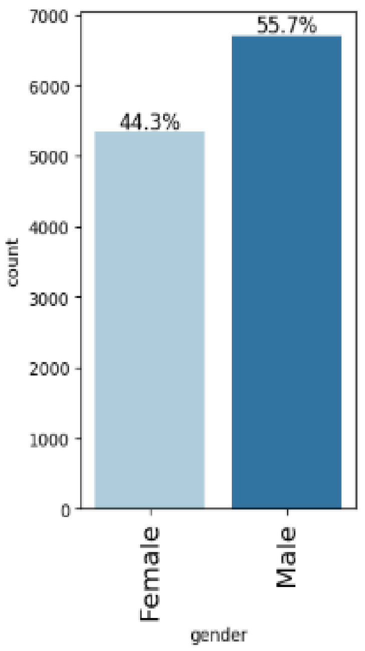


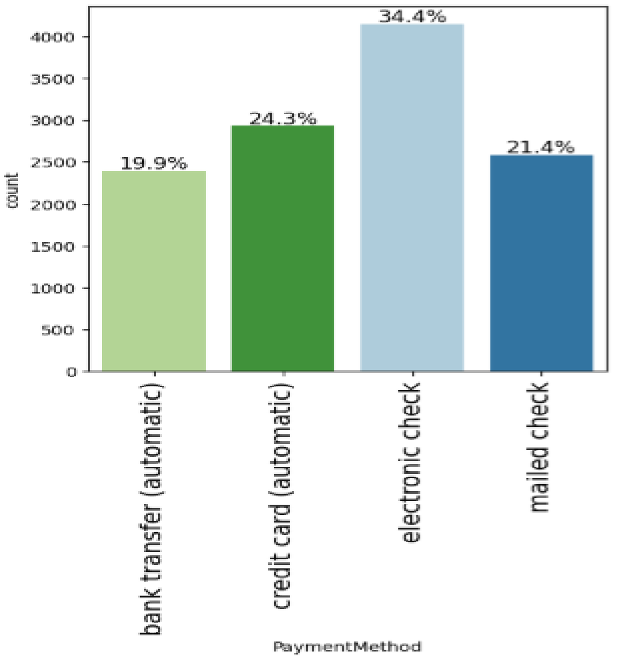
(1.3)

Boxplot shows there is no outlier

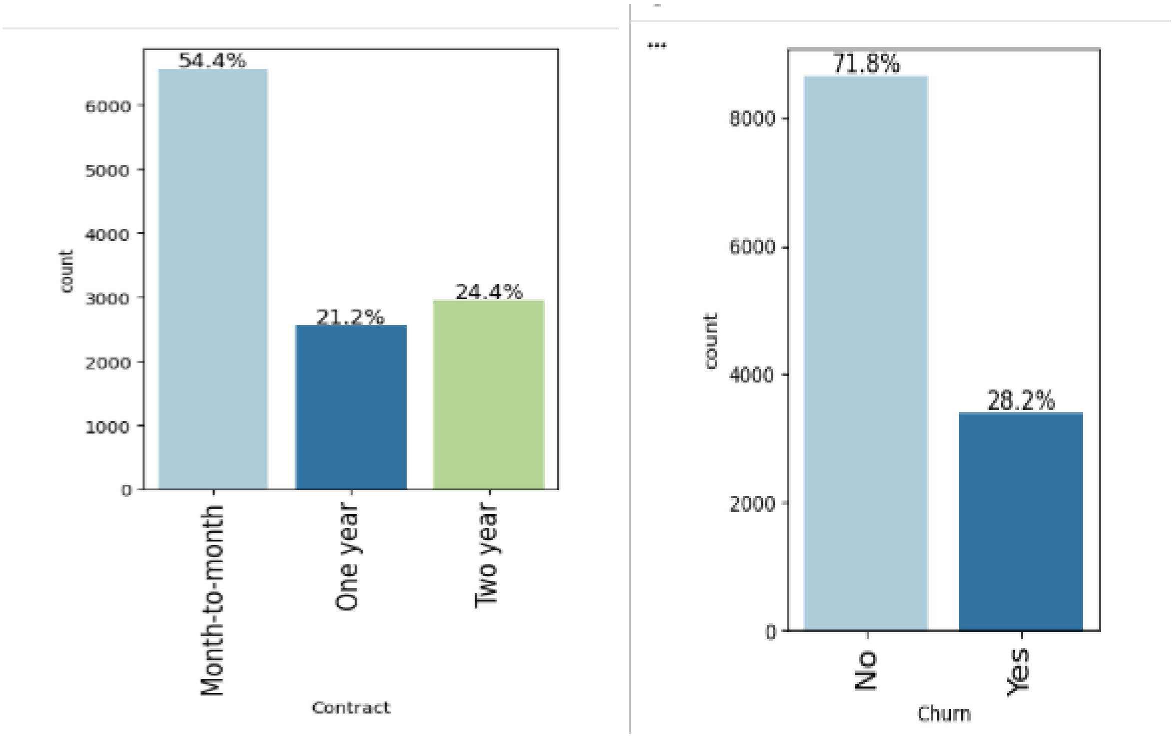
Monthly charges show a bimodal distribution, with one cluster around lower charges 10,30 and another around higher charges 100,120.

The median lies slightly above the lower cluster, while the mean is higher due to customers paying premium prices. TotalCharges is heavily right-skewed, with most customers having low cumulative charges.





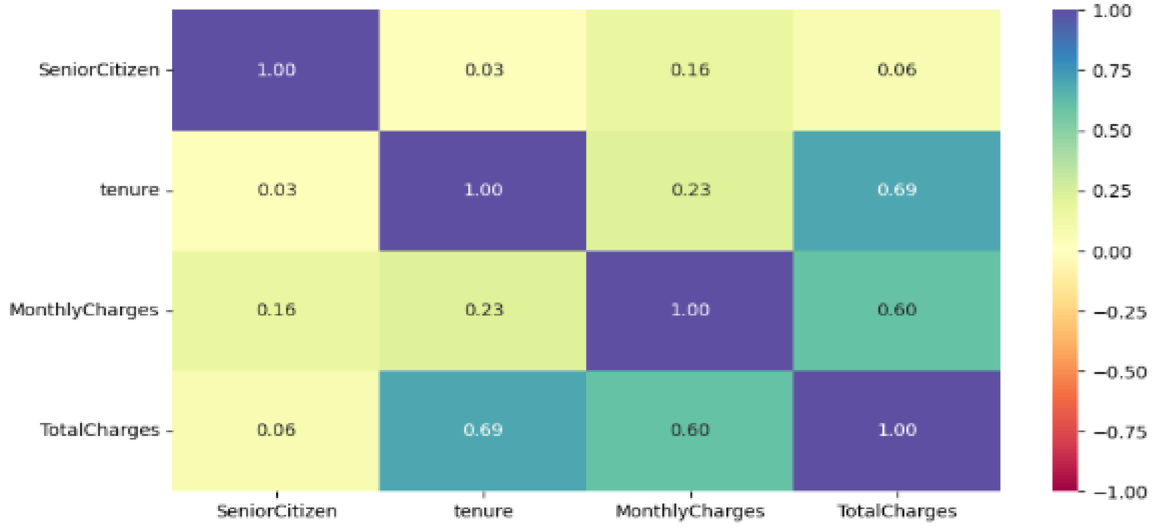
The customer base has slightly more males than females. About 34% pay via electronic check, and paperless billing usage is balanced. Half of the customers 50.5% don’t use online backup, and 23% have no internet service



(1.4)

Most customers 73% have no dependents, and 89% use phone services. Among internet users, fiber optic is the most common. The dataset is imbalanced, with 71% of customers not having churned.

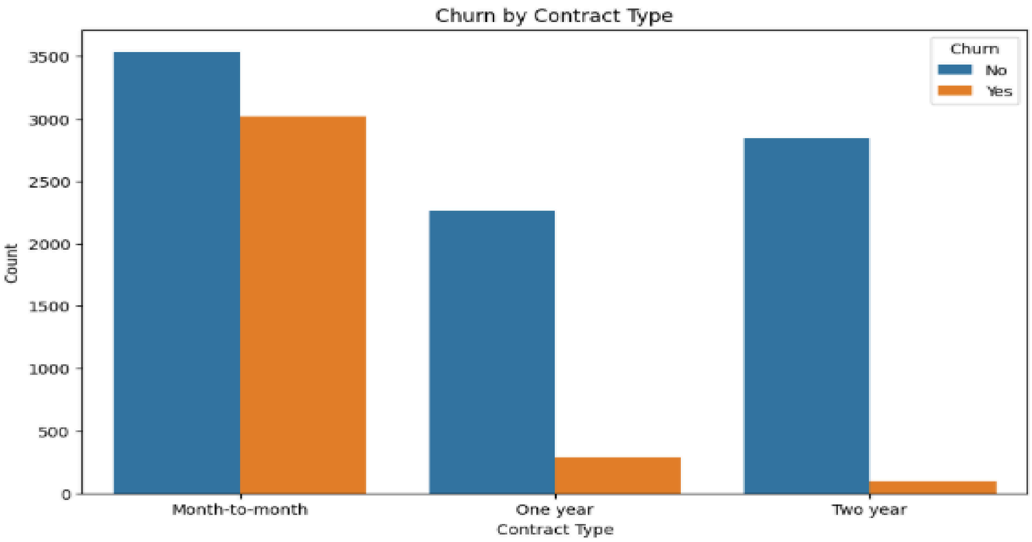
2.3) Bivariate analysis



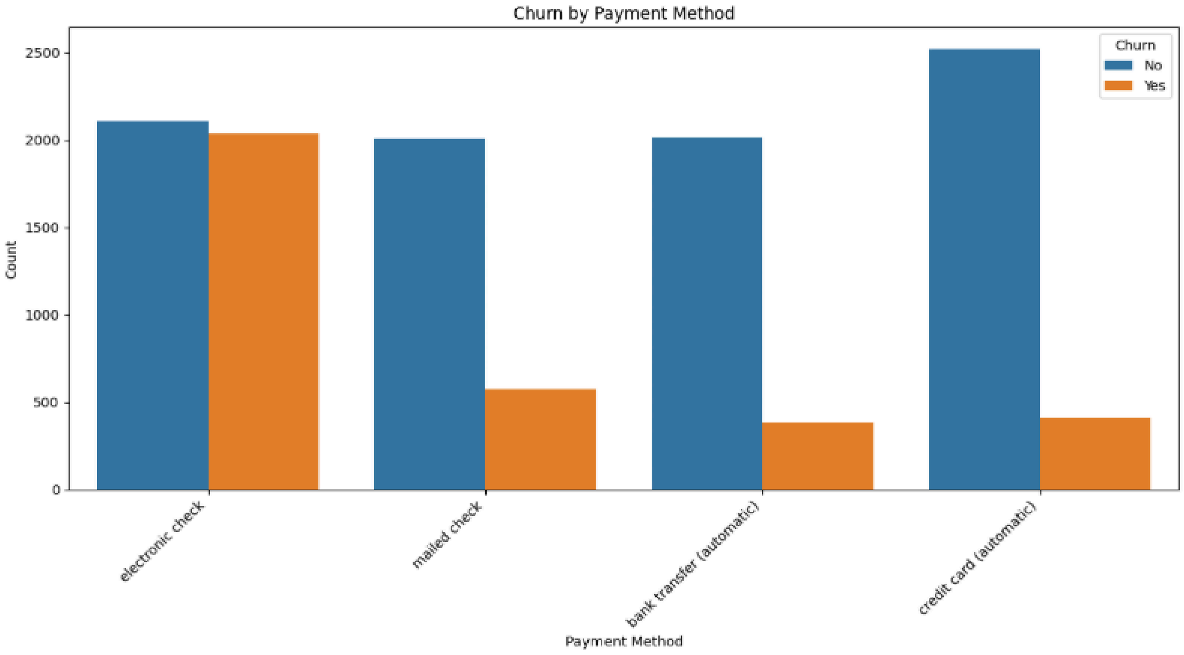
(1.5)

Customers with month to month contracts exhibit significantly higher churn compared to those with long term contracts. Higher monthly charges are associated with increased churn, particularly for customers with short tenure. Customers with short tenure are more likely to churn, indicating that early stage customers are at higher risk.

2.4) Multivariate analysis



(1.6)



(1.7)

When combining tenure, contract type, and charges, customers with short tenure, high monthly charges, and flexible contracts show the highest churn rates. Long term contracts and longer tenure appear to reduce churn even when monthly charges are relatively high**.**

2.5) Insights/observations based on EDA

Churn is more common among new customers.

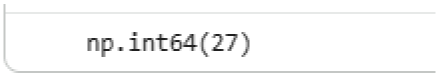
Month-to-month contracts are a strong indicator of churn risk.

High monthly charges increase churn risk, especially for customers with short tenure.

Retention strategies should focus on early customer engagement and encouraging long-term contracts.

3) Data Preprocessing

3.1) Duplicate value check



```
np.int64(27)
```

In this data set there are 27 duplicate value

3.1.1) Duplicate value treatment

TO Handle Duplicate value Simply Delete it

3.3) Data cleaning

1) There are **currency sign changes** in total charges after fix that use $ sign for each currency as mentioned in description

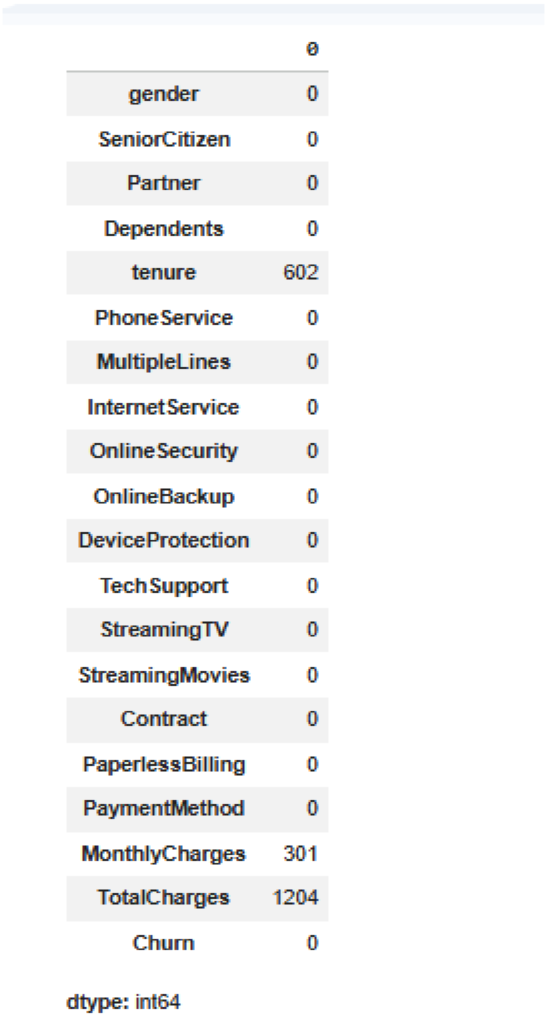
2) Remove currency sign from TotalCharges And MonthlyCharges .because logesic algorithm does not work with signs/Simbols .

3) TotalCharges And MonthlyCharges are **categorical types so change them into numerical types.**

3.4)Anomalous value check

There are some **negative value in Tenure and TotalCharges** so to treat this replace it with median, because there is no any outlier and we can fill these value with median .we can not drop these value because data set is small that -ve value contain many importent information

3.5)Missing value check



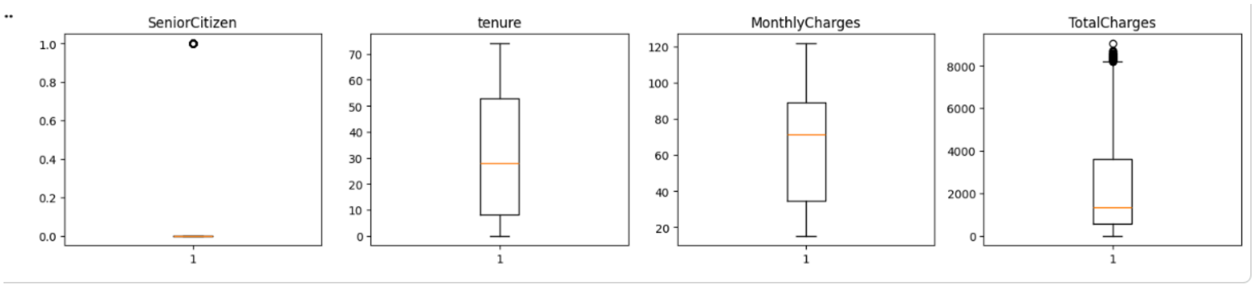
	0
gender	0
SeniorCitizen	0
Partner	0
Dependents	0
tenure	602
PhoneService	0
MultipleLines	0
InternetService	0
OnlineSecurity	0
OnlineBackup	0
DeviceProtection	0
TechSupport	0
StreamingTV	0
StreamingMovies	0
Contract	0
PaperlessBilling	0
PaymentMethod	0
MonthlyCharges	301
TotalCharges	1204
Churn	0

dtype: int64

(1.9)

In the given date set there are 602 null values in Tenure ,301 in Monthly charges and 1204 in Total charges .To fill Tenure we can't use Tenure × MonthlyCharges: because it describes into data description so we simply fill median for all these 3 values .

3.6) Outlier check



(2.0)

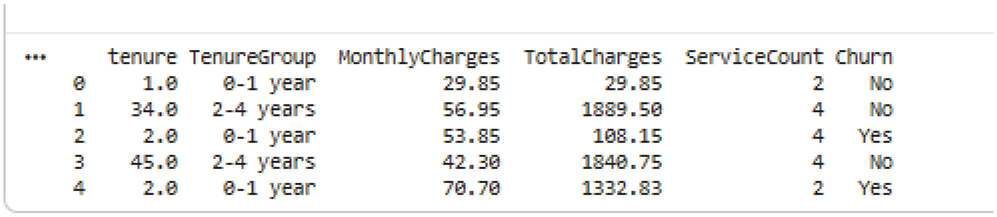
No outliers were found in tenure and MonthlyCharges, as their values are within a reasonable range. The SeniorCitizen variable contains only binary values 0 and 1, so outlier detection is not applicable. Outliers are observed in TotalCharges, but these values are genuine and represent long-term customers with higher service usage. Therefore, these outliers were not treated and were retained for further analysis.

3.7) Feature engineering

For Feature Engineering we create some new columns 1 create tenure in groups like 0-12 ,12-14 and so on for 1 ,2,3 groups . createing service_cols

For that we can count a user using how many services .

Replace 'Yes' with 1, and all 'No' variants and service types with 0 or 1 appropriately

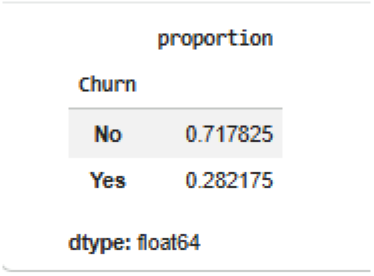


```
...    tenure TenureGroup  MonthlyCharges  TotalCharges  ServiceCount Churn
    0     1.0    0-1 year           29.85         29.85             2    No
    1    34.0   2-4 years           56.95       1889.50             4    No
    2     2.0    0-1 year           53.85        108.15             4   Yes
    3    45.0   2-4 years           42.30       1840.75             4    No
    4     2.0    0-1 year           70.70       1332.83             2   Yes
```

(2.1)

3.8) Class imbalance handling

There are class imbalance in churn because there is a huge mismatch between yes .There are 71 % no and 29 % yes



```
        proportion
Churn
  No      0.717825
  Yes     0.282175

dtype: float64
```

(2.2)

3.9) Feature scaling

For feature scaling convert churn yes and no into 0 and 1 and then shift churn to y and drop from x .

3.10) Data preparation for modeling

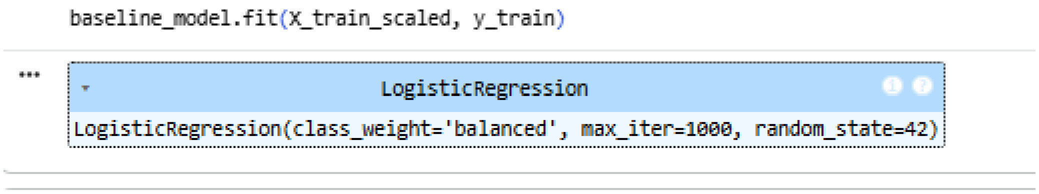
Add 1 hot Encoding and convert dummy values into 1 and 0.

3.11)Data leakage handling

To prevent data leakage split data int train tast and validate .

4)Model Building - Baseline Model- Decide metric of choice with rationale

4.1) Build a baseline (linear) model (For eg: Logistic Regression)

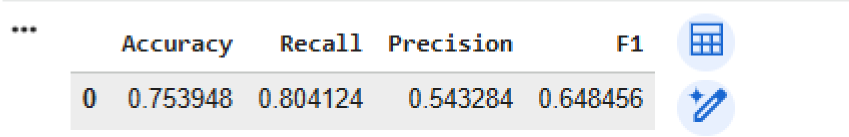


```
baseline_model.fit(X_train_scaled, y_train)
```

```
LogisticRegression
LogisticRegression(class_weight='balanced', max_iter=1000, random_state=42)
```

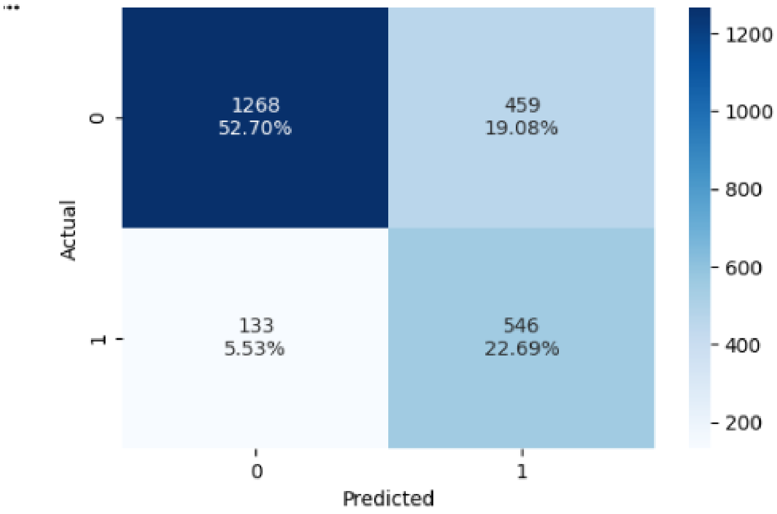
(2.3)

Validation Performance with Threshold = 0.5



	Accuracy	Recall	Precision	F1
0	0.753948	0.804124	0.543284	0.648456

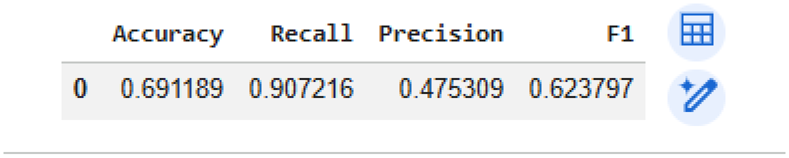
(2.4)



(2.5)

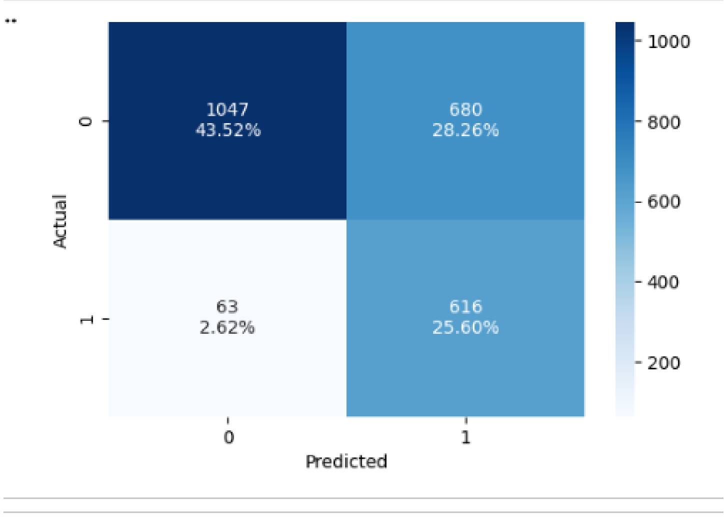
IT show Recall 0.80 with threshold 0.5

Validation Performance with Threshold = 0.35



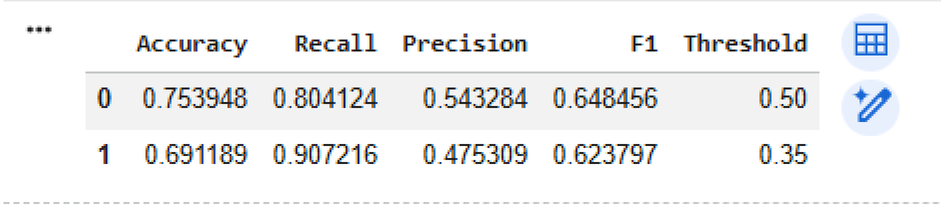
	Accuracy	Recall	Precision	F1
0	0.691189	0.907216	0.475309	0.623797

(2.6)



(2.7)

It show Recall 0.90 with threshold 0.5



	Accuracy	Recall	Precision	F1	Threshold
0	0.753948	0.804124	0.543284	0.648456	0.50
1	0.691189	0.907216	0.475309	0.623797	0.35

(2.8)

Recall 0.90 is good with threshold 0.35

4.2) Check and comment on the performance of the model on multiple metrics (including metric of choice)

The model was checked using accuracy, precision, recall, and F1-score. Since the data is imbalanced, recall was chosen as the main metric. With a threshold of 0.35, the model identifies most churned customers, even though accuracy and precision are slightly lower. This is acceptable because finding customers who may leave is more important than avoiding a few false alerts.

5) Insights

- Customer churn is higher among new customers with shorter tenure.
- Customers on month to month contracts are more likely to leave compared to those on long-term contracts.
- Higher monthly charges increase the chances of churn, especially for customers who have recently joined.
- Using a lower threshold helps the model identify more customers who are likely to churn.
- The model successfully highlights key factors such as tenure, contract type, and service usage that influence churn.

6) Recommendations

- Focus retention efforts on customers during their initial months with the company.
- Encourage customers to switch to long-term contracts through discounts or loyalty offers.
- Review pricing plans for customers with high monthly charges to reduce dissatisfaction.
- Use churn predictions to reach out to high risk customers early and offer personalized support.
- Continuously improve the model and explore advanced techniques for better churn prediction.