Customer Churn Prediction

BUSINESS REPORT

PGP DSA 1 FAB 2026

ANJU SAINI

INDEX

Graphs

1) **Understanding the Business Problem**

1.1) Defining the Problem Statement

Customer churn is a major challenge for businesses because when customers leave the company loses both current revenue and future growth opportunities. In highly competitive industries such as telecommunications, customers can easily switch to competitors due to similar pricing and service offerings. AlphaCom currently lacks a clear understanding of why customers stop using its services. This project aims to analyze customer data to identify the key factors influencing churn and to help the business predict customers who are at high risk of leaving.

Industry studies indicate that acquiring a new customer costs nearly 5 to 6 times more than retaining an existing one. Even a 5% improvement in customer retention can increase profits by 25% to 95%. The churn rate observed in this dataset is comparable to typical telecom industry churn rates, which are generally high due to intense competition and low switching costs. This highlights a clear business risk and the need for proactive churn management. Identifying high risk churn customers early allows the company to apply targeted retention strategies, reducing revenue loss and improving long term profitability.

1.2) Need of the Study / Project

The need for this project arises from the significant financial impact of customer churn. Retaining existing customers is more cost effective than acquiring new ones, as acquisition costs are substantially higher. In addition, the success rate of selling new products or services to existing customers is approximately 60 to 70%, compared to only 5 to 20% for new customers. Research also shows that U.S. companies lose approximately $136.8 billion per year due to avoidable customer churn.

By identifying the key drivers of churn, AlphaCom can focus on retaining valuable customers, reducing avoidable losses, improving profitability, and strengthening its market position through cost effective retention strategies.

1.3) Understanding the Business Context

AlphaCom is experiencing customer churn that directly affects revenue, profitability, and brand reputation. In the telecom industry, churn is commonly driven by factors such as pricing sensitivity, contract flexibility, service quality, and competitive offerings. The churn rate reflected in the dataset aligns with general industry behavior, indicating that this is a realistic and relevant business problem. Customers may leave due to high charges, unfavorable contract terms, dissatisfaction with services, or better alternatives offered by competitors. Understanding these patterns is critical for making informed, data driven business decisions.

1.4) Business / Social Opportunity

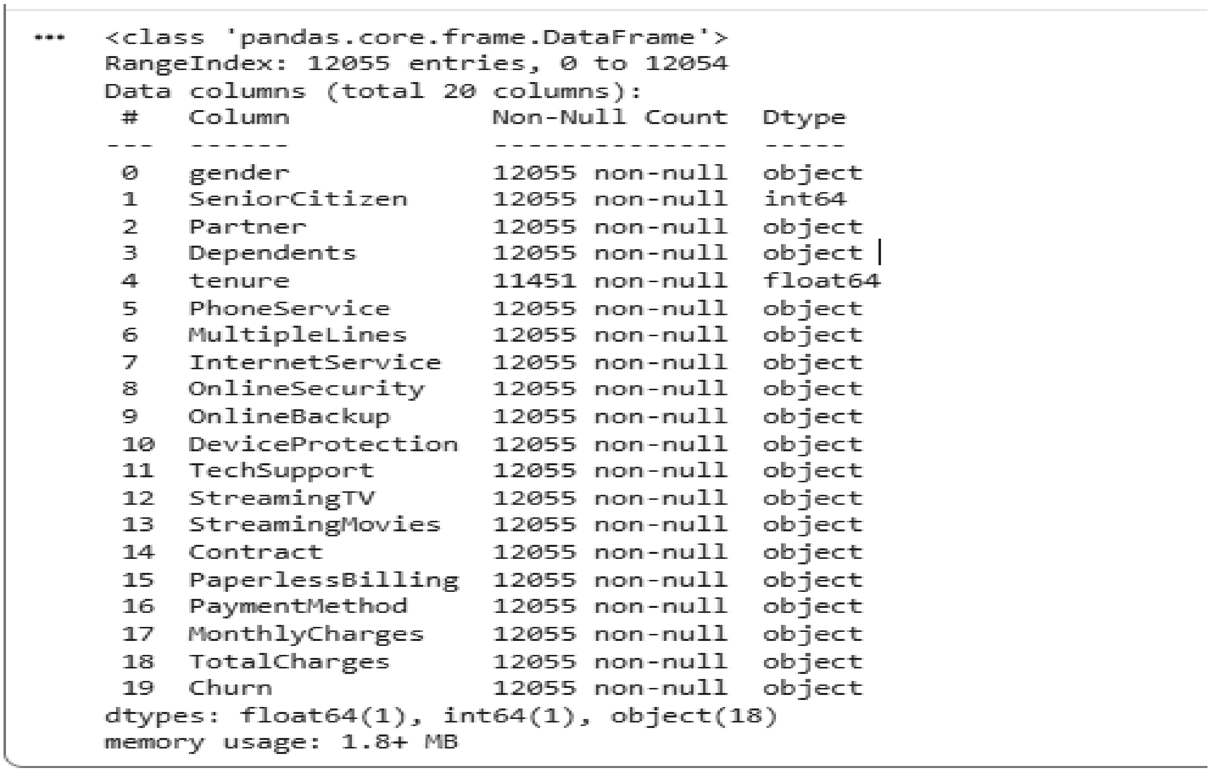
By understanding the reasons behind customer churn, AlphaCom can design targeted retention strategies for customers who are most likely to leave. This creates an opportunity to reduce revenue loss, increase customer lifetime value, and improve overall customer satisfaction. From a business perspective, lower churn contributes to higher profitability and a stronger brand reputation. From a broader perspective, improved customer engagement and service quality support sustainable business growth and long term customer relationships.

1.5) Objective

The objective of this project is to identify customers who are at high risk of churn so that AlphaCom can take proactive retention actions. The scope of the study is limited to customer demographic information, service usage patterns, billing details, and contract related variables available in the dataset. Given the competitive nature of the telecom industry and the high costs associated with customer churn, reducing churn is critical for protecting revenue, improving customer lifetime value, and maintaining a strong market position.

2) Exploratory Data Analysis

2.1) Data overview

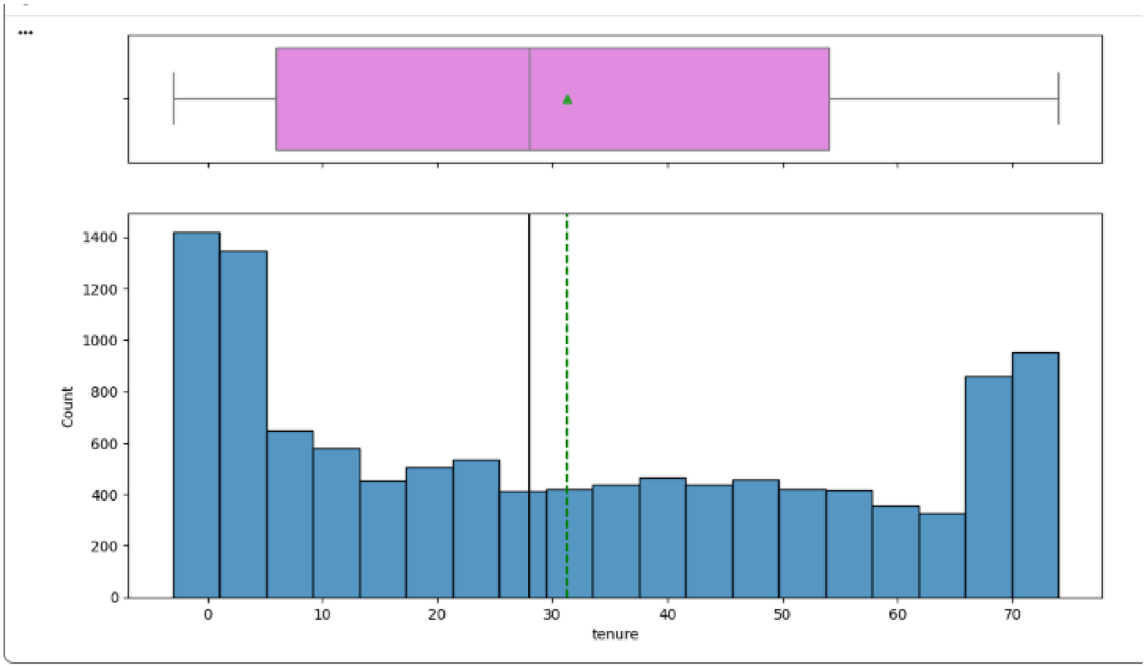


```
...  <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 12055 entries, 0 to 12054
     Data columns (total 20 columns):
      #   Column            Non-Null Count  Dtype
     ---  ------            --------------  -----
      0   gender            12055 non-null  object
      1   SeniorCitizen     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)

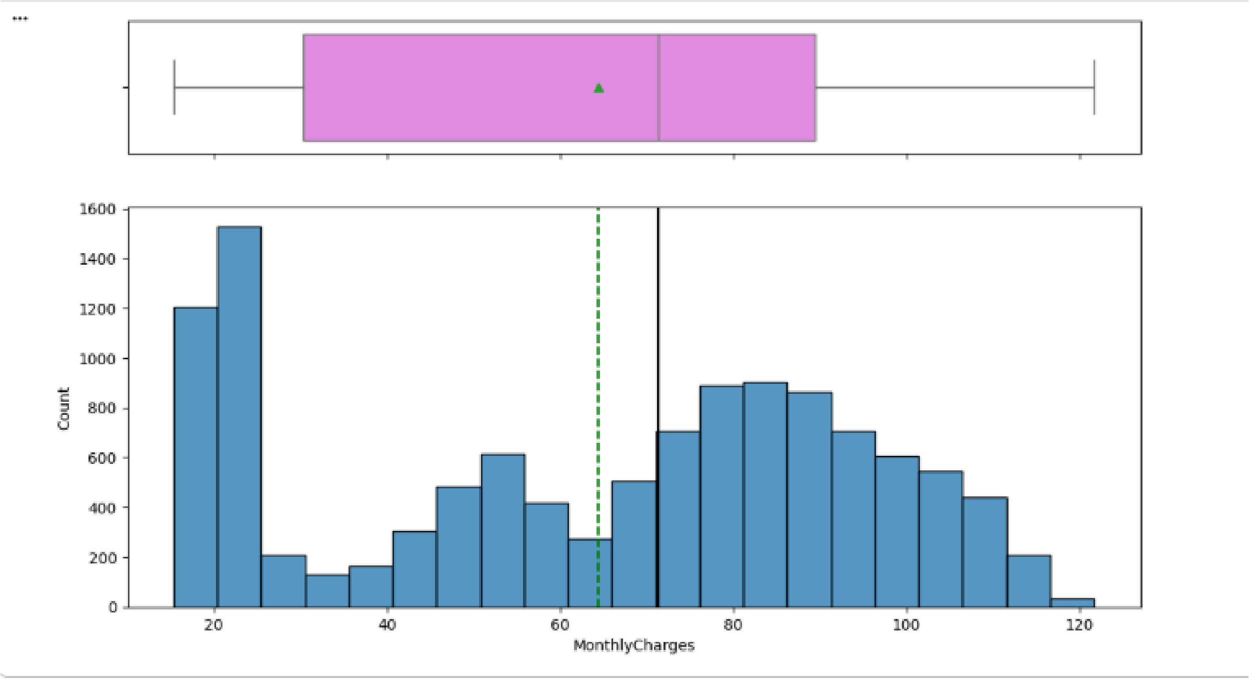
The target variable is Churn, which shows whether a customer has left the company Yes or is still using the service No. The dataset includes 12,055 customer records and 20 input features. These features include numerical variables such as tenure and charges, as well as categorical variables like contract type and payment method.

2.2) Univariate analysis



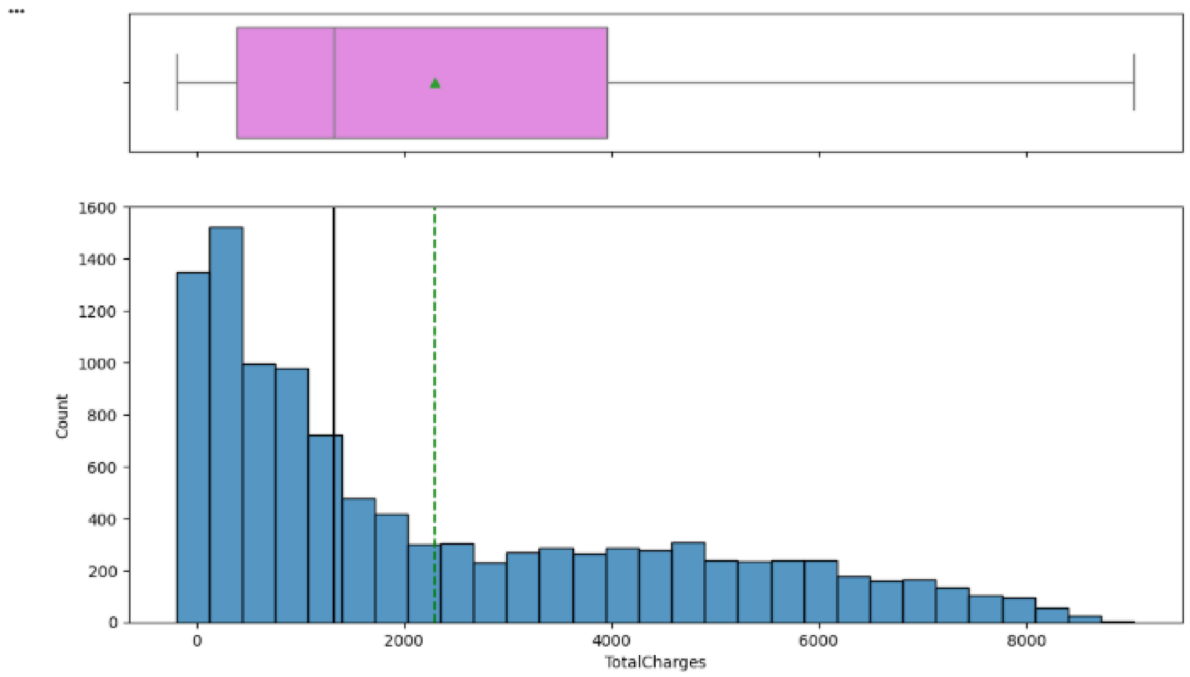
(1.2)

Many customers leave early, while a smaller group stays for a long period. Since churn typically occurs at lower tenure values, tenure provides clear separation for the target variable and is expected to be a strong predictor of churn.



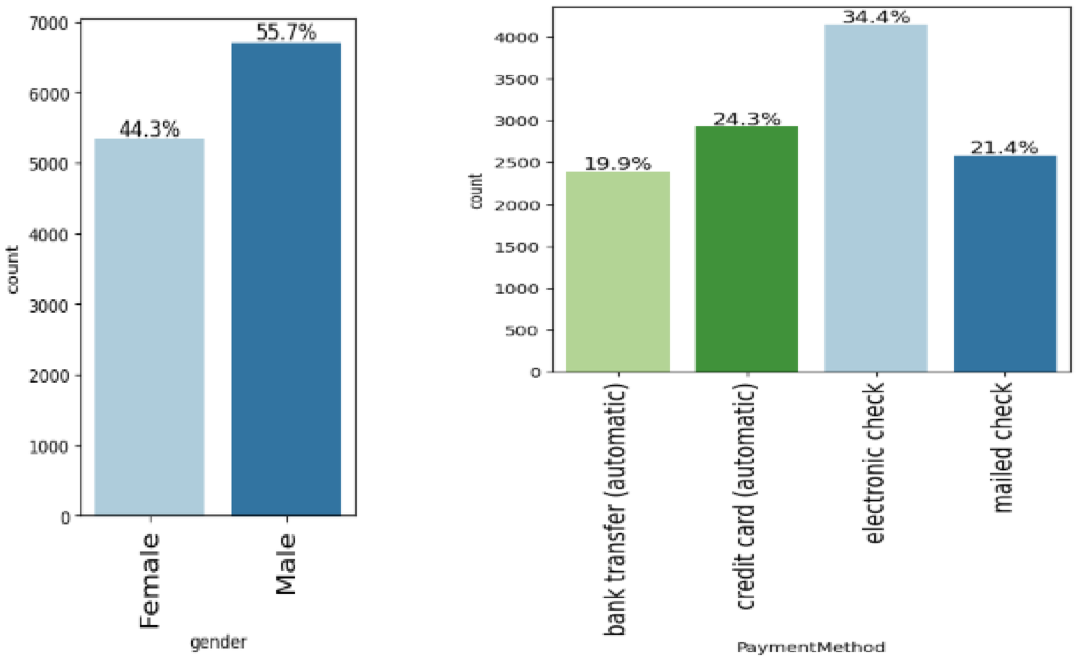
(1.3)

sSmaller group of customers pays significantly higher charges. Customers with higher monthly charges are more likely to churn, making this variable relevant for churn prediction.



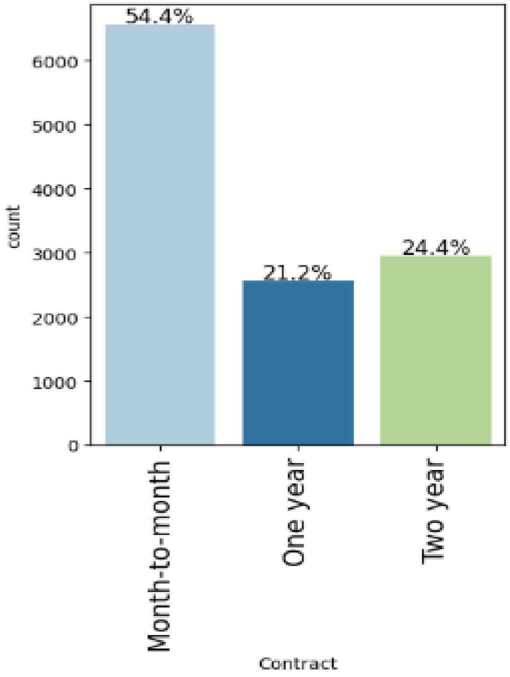
(1.4)

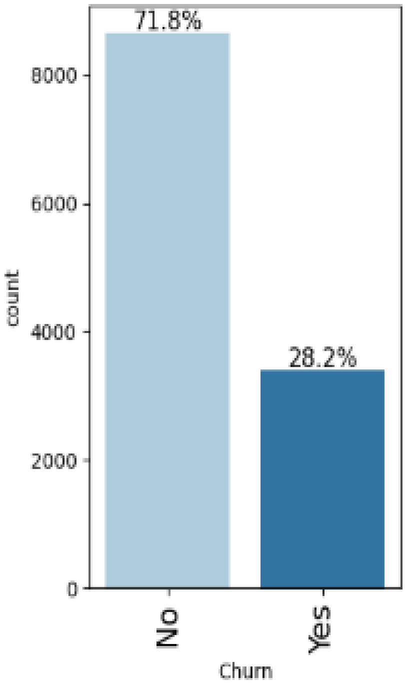
Customers with higher TotalCharges generally represent long tenure and high value customers, who are less likely to churn. This variable indirectly captures customer loyalty.



(1.5)

Gender distribution is nearly balanced, suggesting gender alone may not strongly influence churn.Customers using electronic check payment methods may have higher churn risk due to lack of automatic payment commitment.



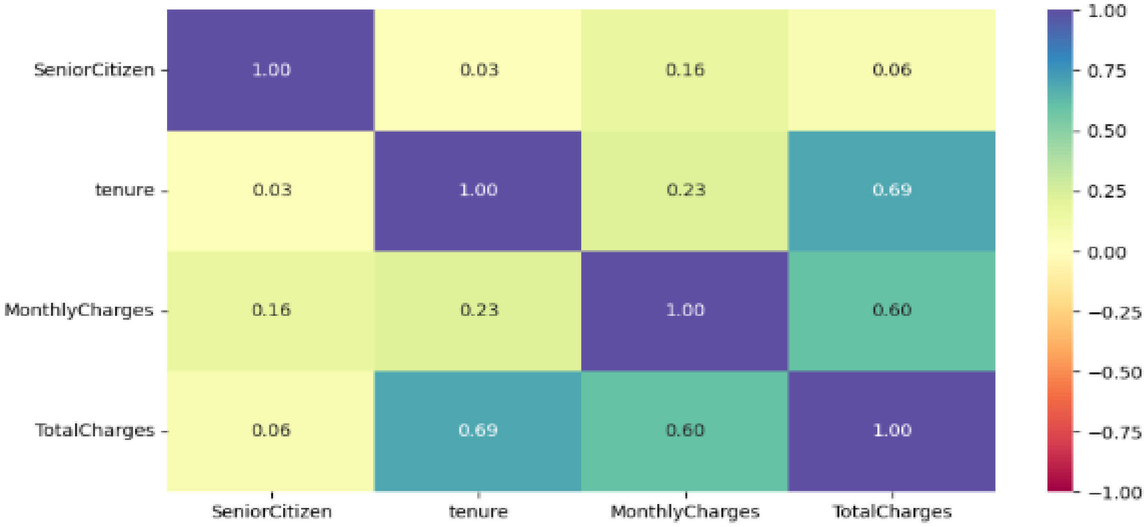


(1.6)

Contract type shows dominance of month to month contracts, which are typically associated with higher churn compared to long term contracts.

The churn variable shows a clear class imbalance, with approximately 71% non churned customers and 29% churned customers, indicating the need for imbalance handling during modeling.

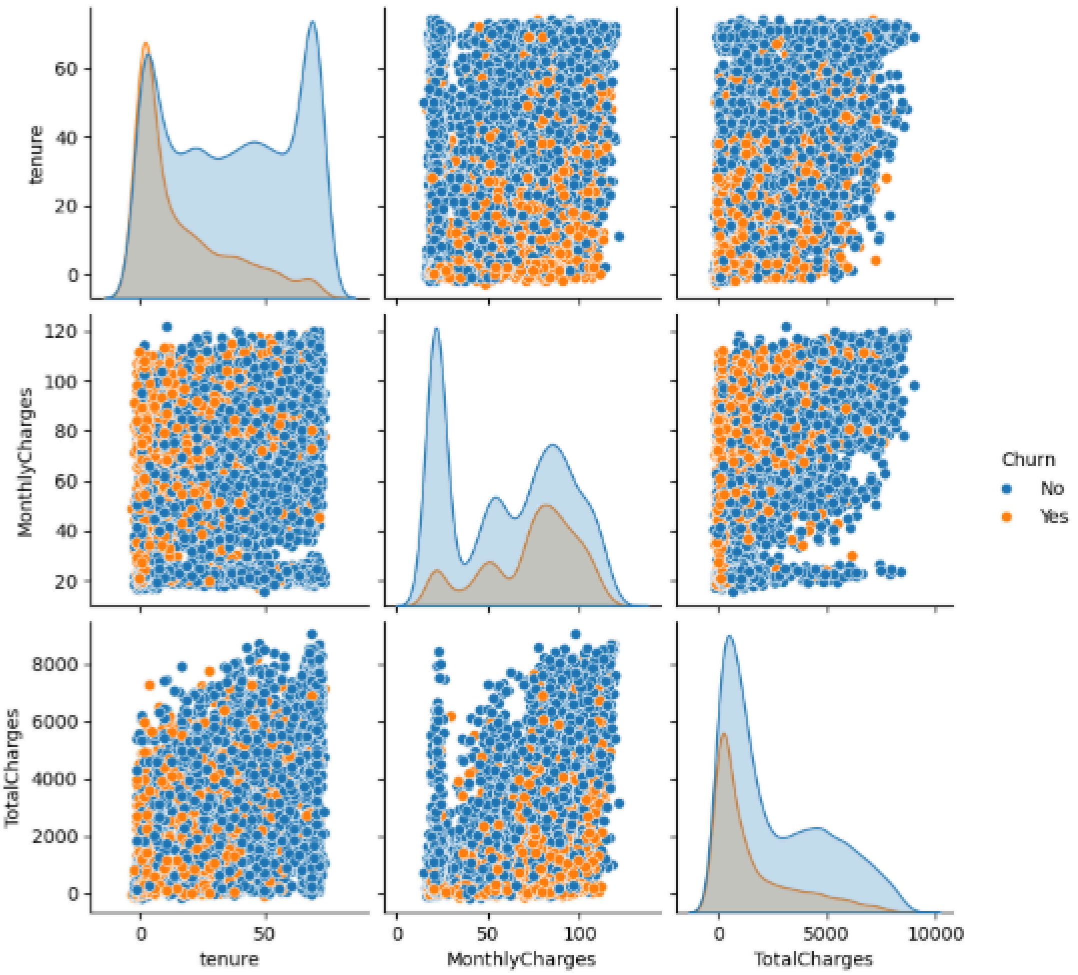
2.3) Bivariate analysis



(1.7)

The correlation heatmap shows a strong positive correlation between tenure and TotalCharges, which is expected because customers who stay longer accumulate higher total charges over time. MonthlyCharges show a weaker correlation with tenure, indicating that pricing and customer duration are relatively independent. No extremely high correlations are observed among numerical variables, suggesting that strong Relationship is not a concern for model building.

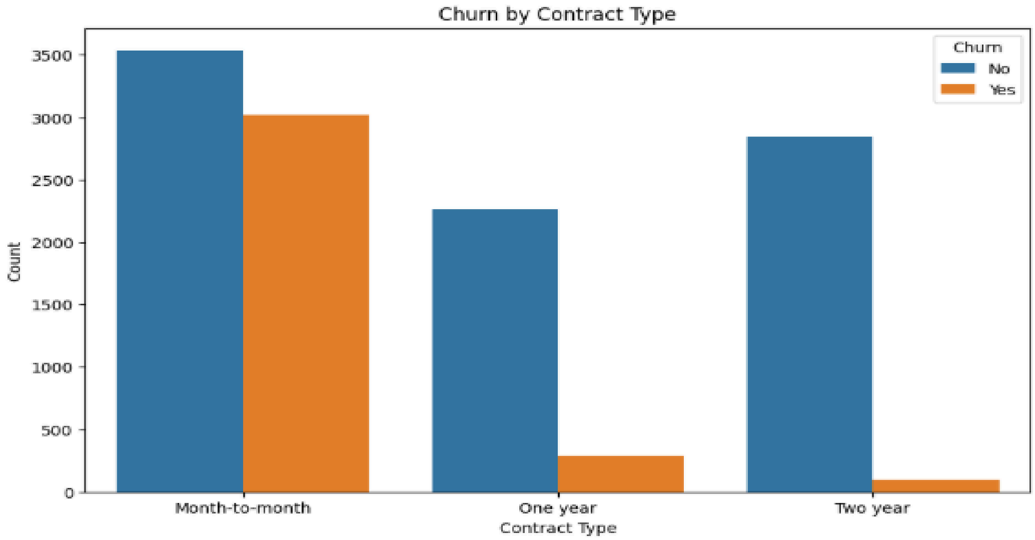
2.4) Relationship between variables



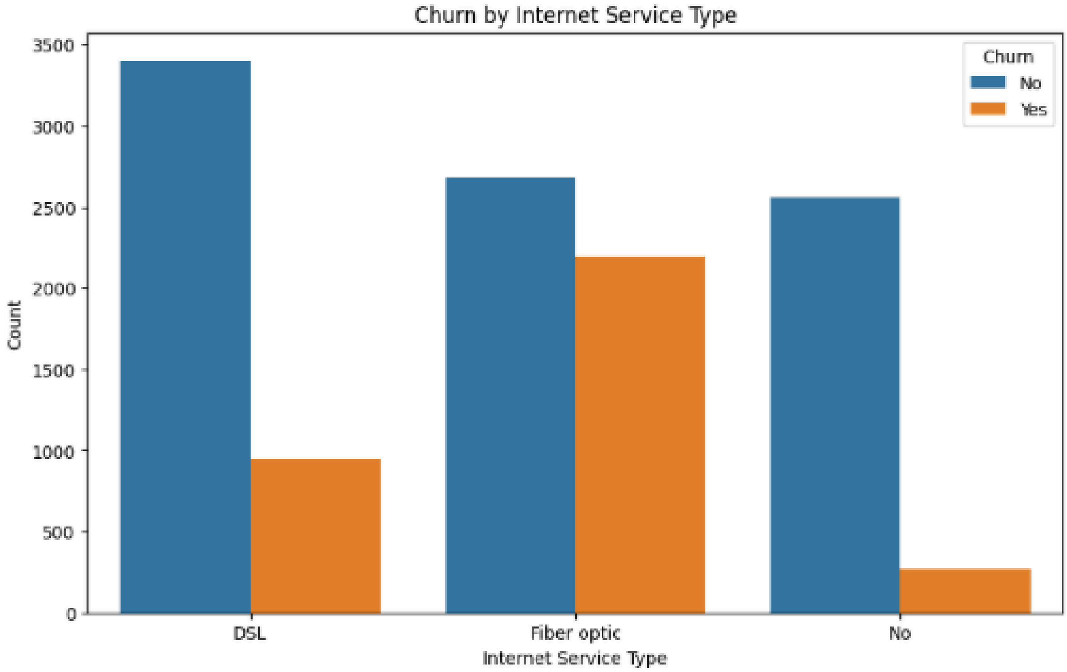
(1.8)

This suggests a potential Strong Relationship between tenure and TotalCharges. The pairplot highlights that customers who churn are mostly concentrated at low tenure values, while customers with higher tenure are more likely to stay. TotalCharges increase steadily with tenure, explaining why non churned customers have higher total spending. MonthlyCharges partially separate churned and non churned customers, with higher charges increasing churn risk mainly for short tenure customers. Overall, tenure emerges as the most influential numerical variable affecting churn.

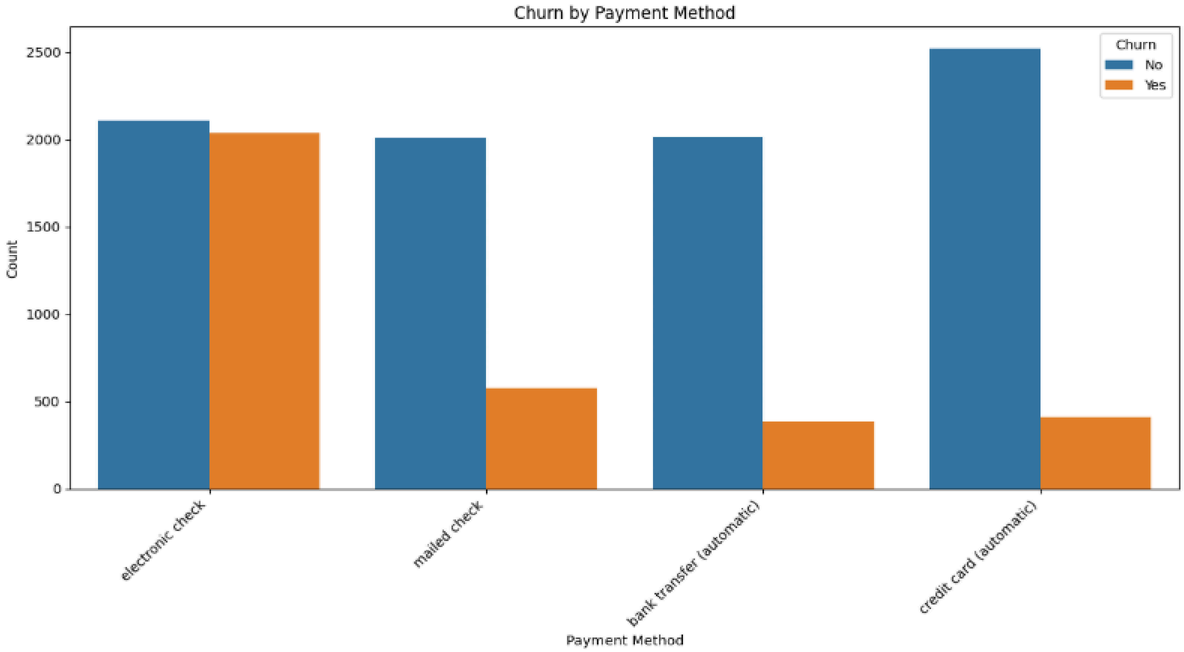
2.5) Multivariate analysis



(1.9)



(1.10)



(1.11)

Tenure is the most important factor influencing churn. Customers with shorter tenure are much more likely to churn, while long tenure customers tend to remain loyal. This shows that churn risk is highest during the early stages of the customer lifecycle.

MonthlyCharges impact churn, especially for newer customers. Customers with higher monthly charges show a higher likelihood of churn, indicating pricing sensitivity among short tenure customers.

TotalCharges reflect customer loyalty and lifetime value. Higher TotalCharges are associated with longer tenure and lower churn, while customers with low TotalCharges often churn early.

2.6) Key Observations and Insights from EDA

Tenure is the most important factor influencing churn. Customers with shorter tenure are much more likely to churn, while long tenure customers tend to remain loyal. This shows that churn risk is highest during the early stages of the customer lifecycle.

MonthlyCharges impact churn, especially for newer customers. Customers with higher monthly charges show a higher likelihood of churn, indicating pricing sensitivity among short tenure customers.

TotalCharges reflect customer loyalty and lifetime value. Higher TotalCharges are associated with longer tenure and lower churn, while customers with low TotalCharges often churn early.

The payment method shows meaningful churn patterns. Customers using electronic check payments appear more likely to churn compared to those using automatic payment methods, suggesting lower engagement.

Churn data is imbalanced. A larger proportion of customers have not churned, indicating the need for class imbalance handling during model building.

Multivariate analysis confirms tenure as the strongest predictor. The pairplot and correlation heatmap show a strong relationship between tenure and TotalCharges, with churned customers concentrated at lower tenure values.

3) Data Preprocessing

3.1) Duplicate value check

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

The dataset was checked for duplicate records, and 27 duplicate rows were identified.

3.1.1) Duplicate value treatment

Duplicate records do not add any new information and may bias the analysis, they were removed using the in place option to ensure data consistency.

3.2) Data cleaning

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

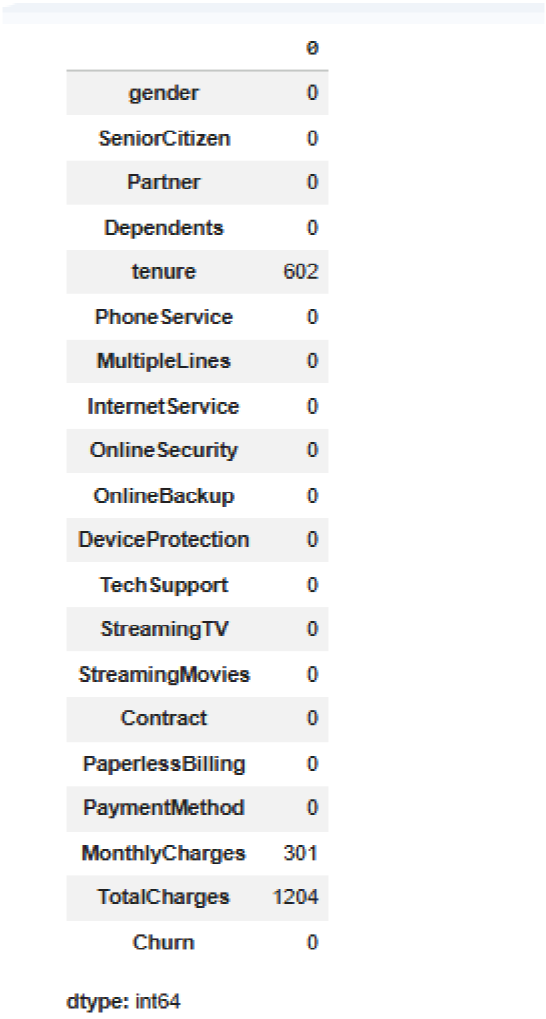
2) Remove currency sign from TotalCharges And MonthlyCharges .because logesic algorithm does not work with signs/Simbols .

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

3.3)Anomalous value check

Anomalous values such as negative tenure and negative TotalCharges were identified, which are not logically possible. These values were treated by replacing them with the median of the respective feature to maintain data consistency without distorting the overall distribution.

3.4)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

(1.12)

There are multiple missing values in data set tenure 602 ,Monthly charges 301 and Total charges 1204 Missing Values .

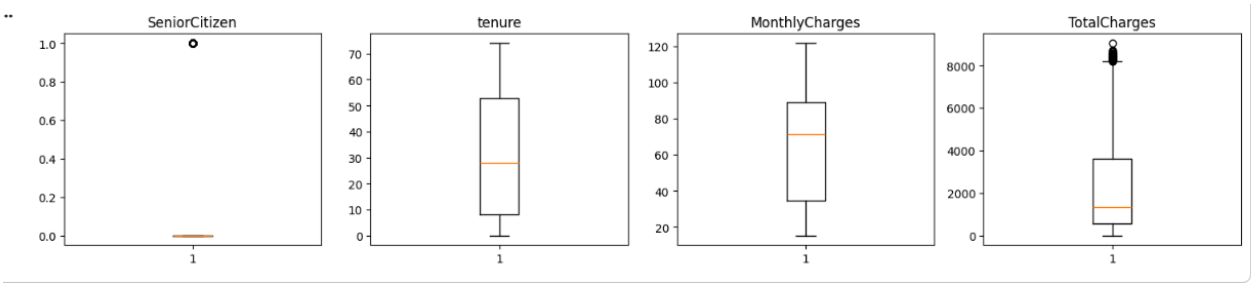
Missing values were found in multiple numerical variables: 602 in tenure, 1204 in

TotalCharges, and 301 in MonthlyCharges. These missing values were treated using median imputation, as the data was skewed and the median is more robust than the mean.

Although TotalCharges is related to tenure and MonthlyCharges, it was not recalculated using a formula, as the dataset description indicates that these values are not always mathematically equal. Therefore, median imputation was considered the safest approach.

Missing values were identified in numerical variables. Since the data was skewed, missing values were imputed using the median, which is more robust to outliers compared to the mean.

3.5) Outlier check

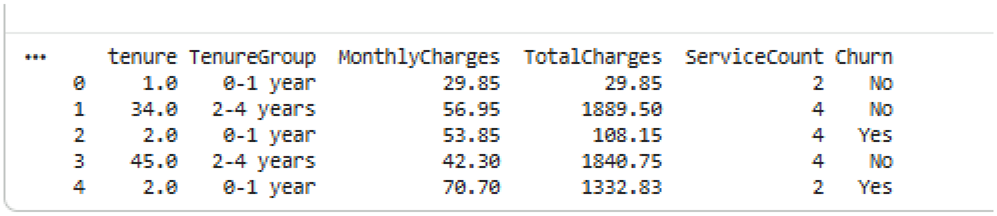


(1.13)

Outliers were examined using boxplots. No significant outliers were found in tenure and MonthlyCharges. Although TotalCharges showed extreme values, these were retained as they represent long term customers with higher cumulative spending and are therefore considered genuine.

3.6) Feature engineering

Categorical variables were transformed using one hot encoding to make them suitable for machine learning models. This step allowed the model to interpret categorical information numerically without introducing unintended order.



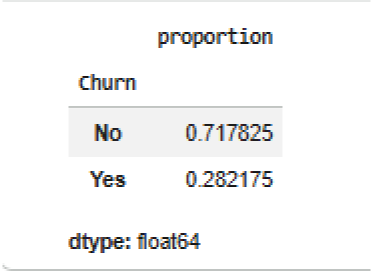
```
...    tenure TenureGroup  MonthlyCharges  TotalCharges  ServiceCount Churn
    0     1.0    0-1 year           29.85         29.85             2    No
    1    34.0   2-4 years           56.95       1889.50             4    No
    2     2.0    0-1 year           53.85        108.15             4   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
```

(1.14)

A new feature called ServiceCount is created to show how many services a customer is using.Service related columns such as online security, backup, device protection, tech support, streaming services, phone service, and internet service are considered. Each service is converted into numbers, where Yes is treated as 1 and all types of No are treated as 0.These values are then added together to calculate ServiceCount, which helps indicate customer engagement and potential loyalty.

3.7) Class imbalance handling

Although the class imbalance is moderate, stratified sampling was used during train test splitting to preserve class distribution. Resampling techniques were deferred to the modeling stage, where recall focused optimization will be evaluated.



```
        proportion
Churn
  No      0.717825
  Yes     0.282175

dtype: float64
```

(1.15)

3.8) Feature scaling

Feature scaling was applied to numerical variables using standardization. The scaler was fitted only on the training data and then applied to validation and test sets to prevent data leakage.

3.9) Data preparation for modeling

Add 1 hot Encoding and convert dummy values/categorical types value into 1 and 0.

3.10)Data leakage handling

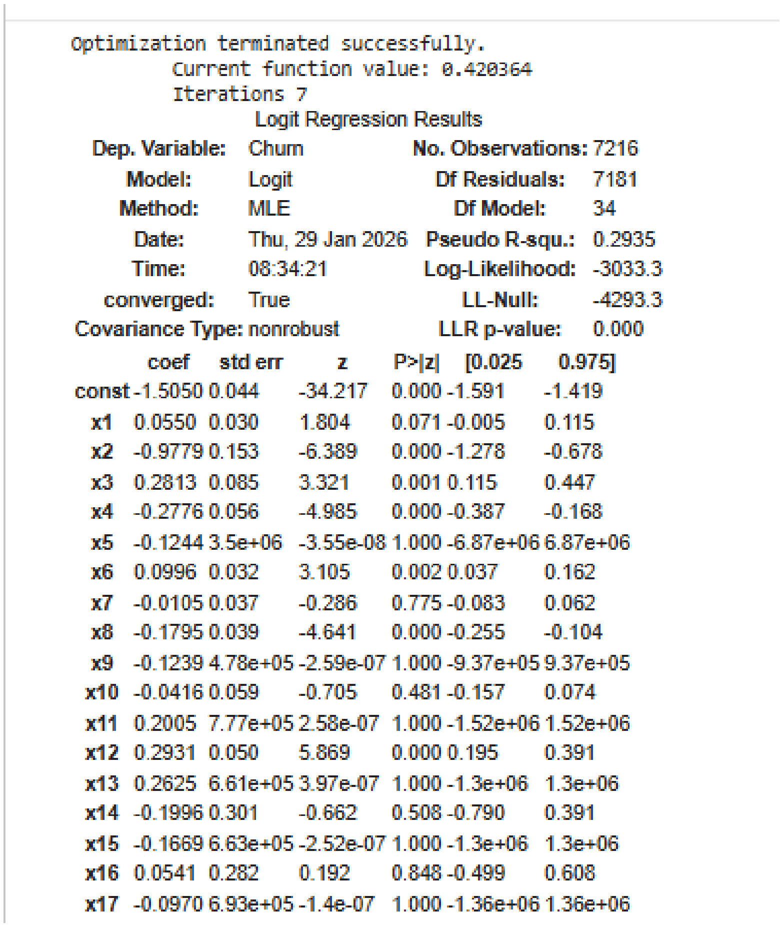
To prevent data leakage, split data into train tast and validate .

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

4.1) Model Selection Rationale

Since the target variable Churn is binary, classification models are appropriate. Logistic Regression was selected as the baseline linear model because it is interpretable, handles binary outcomes well, and allows statistical evaluation of feature significance using coefficients and p values. Non linear models such as Decision Trees can capture complex relationships but lack interpretability, making Logistic Regression suitable for initial analysis and business insights.

4.2) Metric Selection Rationale

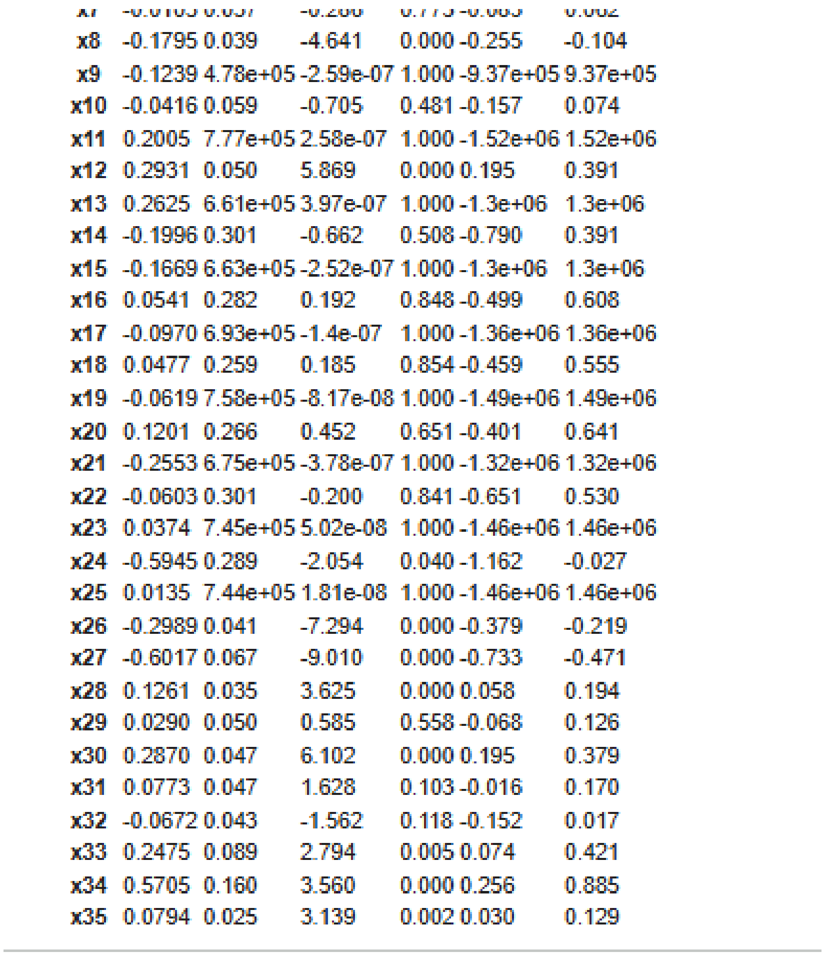


```
Optimization terminated successfully.
         Current function value: 0.420364
         Iterations 7
```

Logit Regression Results

Dep. Variable:	Churn	**No. Observations:**	7216
Model:	Logit	**Df Residuals:**	7181
Method:	MLE	**Df Model:**	34
Date:	Thu, 29 Jan 2026	**Pseudo R-squ.:**	0.2935
Time:	08:34:21	**Log-Likelihood:**	-3033.3
converged:	True	**LL-Null:**	-4293.3
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-1.5050	0.044	-34.217	0.000	-1.591	-1.419
x1	0.0550	0.030	1.804	0.071	-0.005	0.115
x2	-0.9779	0.153	-6.389	0.000	-1.278	-0.678
x3	0.2813	0.085	3.321	0.001	0.115	0.447
x4	-0.2776	0.056	-4.985	0.000	-0.387	-0.168
x5	-0.1244	3.5e+06	-3.55e-08	1.000	-6.87e+06	6.87e+06
x6	0.0996	0.032	3.105	0.002	0.037	0.162
x7	-0.0105	0.037	-0.286	0.775	-0.083	0.062
x8	-0.1795	0.039	-4.641	0.000	-0.255	-0.104
x9	-0.1239	4.78e+05	-2.59e-07	1.000	-9.37e+05	9.37e+05
x10	-0.0416	0.059	-0.705	0.481	-0.157	0.074
x11	0.2005	7.77e+05	2.58e-07	1.000	-1.52e+06	1.52e+06
x12	0.2931	0.050	5.869	0.000	0.195	0.391
x13	0.2625	6.61e+05	3.97e-07	1.000	-1.3e+06	1.3e+06
x14	-0.1996	0.301	-0.662	0.508	-0.790	0.391
x15	-0.1669	6.63e+05	-2.52e-07	1.000	-1.3e+06	1.3e+06
x16	0.0541	0.282	0.192	0.848	-0.499	0.608
x17	-0.0970	6.93e+05	-1.4e-07	1.000	-1.36e+06	1.36e+06

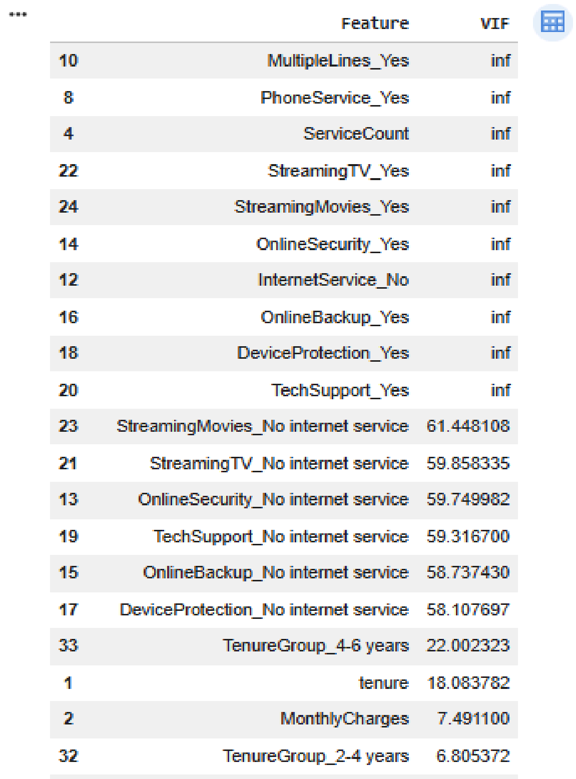


x7	-0.0105	0.037	-0.286	0.775	-0.083	0.062
x8	-0.1795	0.039	-4.641	0.000	-0.255	-0.104
x9	-0.1239	4.78e+05	-2.59e-07	1.000	-9.37e+05	9.37e+05
x10	-0.0416	0.059	-0.705	0.481	-0.157	0.074
x11	0.2005	7.77e+05	2.58e-07	1.000	-1.52e+06	1.52e+06
x12	0.2931	0.050	5.869	0.000	0.195	0.391
x13	0.2625	6.61e+05	3.97e-07	1.000	-1.3e+06	1.3e+06
x14	-0.1996	0.301	-0.662	0.508	-0.790	0.391
x15	-0.1669	6.63e+05	-2.52e-07	1.000	-1.3e+06	1.3e+06
x16	0.0541	0.282	0.192	0.848	-0.499	0.608
x17	-0.0970	6.93e+05	-1.4e-07	1.000	-1.36e+06	1.36e+06
x18	0.0477	0.259	0.185	0.854	-0.459	0.555
x19	-0.0619	7.58e+05	-8.17e-08	1.000	-1.49e+06	1.49e+06
x20	0.1201	0.266	0.452	0.651	-0.401	0.641
x21	-0.2553	6.75e+05	-3.78e-07	1.000	-1.32e+06	1.32e+06
x22	-0.0603	0.301	-0.200	0.841	-0.651	0.530
x23	0.0374	7.45e+05	5.02e-08	1.000	-1.46e+06	1.46e+06
x24	-0.5945	0.289	-2.054	0.040	-1.162	-0.027
x25	0.0135	7.44e+05	1.81e-08	1.000	-1.46e+06	1.46e+06
x26	-0.2989	0.041	-7.294	0.000	-0.379	-0.219
x27	-0.6017	0.067	-9.010	0.000	-0.733	-0.471
x28	0.1261	0.035	3.625	0.000	0.058	0.194
x29	0.0290	0.050	0.585	0.558	-0.068	0.126
x30	0.2870	0.047	6.102	0.000	0.195	0.379
x31	0.0773	0.047	1.628	0.103	-0.016	0.170
x32	-0.0672	0.043	-1.562	0.118	-0.152	0.017
x33	0.2475	0.089	2.794	0.005	0.074	0.421
x34	0.5705	0.160	3.560	0.000	0.256	0.885
x35	0.0794	0.025	3.139	0.002	0.030	0.129

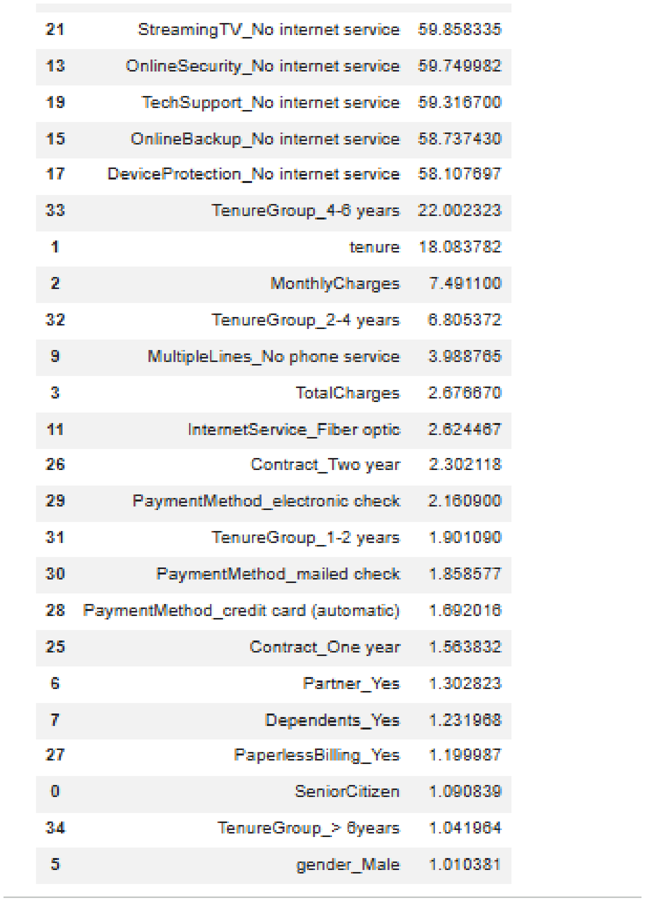
(1.16)

Accuracy alone is not sufficient for churn prediction due to class imbalance. Recall was chosen as the primary evaluation metric because the business objective is to minimize false negatives, customers who churn but are incorrectly predicted as non churn. Precision, F1 score, and ROC AUC were also monitored to ensure balanced model performance**.**

4.3)VIF / Strong Relationship Check



	Feature	VIF
10	MultipleLines_Yes	inf
8	PhoneService_Yes	inf
4	ServiceCount	inf
22	StreamingTV_Yes	inf
24	StreamingMovies_Yes	inf
14	OnlineSecurity_Yes	inf
12	InternetService_No	inf
16	OnlineBackup_Yes	inf
18	DeviceProtection_Yes	inf
20	TechSupport_Yes	inf
23	StreamingMovies_No internet service	61.448108
21	StreamingTV_No internet service	59.858335
13	OnlineSecurity_No internet service	59.749982
19	TechSupport_No internet service	59.316700
15	OnlineBackup_No internet service	58.737430
17	DeviceProtection_No internet service	58.107697
33	TenureGroup_4-6 years	22.002323
1	tenure	18.083782
2	MonthlyCharges	7.491100
32	TenureGroup_2-4 years	6.805372



21	StreamingTV_No internet service	59.858335
13	OnlineSecurity_No internet service	59.749982
19	TechSupport_No internet service	59.316700
15	OnlineBackup_No internet service	58.737430
17	DeviceProtection_No internet service	58.107697
33	TenureGroup_4-6 years	22.002323
1	tenure	18.083782
2	MonthlyCharges	7.491100
32	TenureGroup_2-4 years	6.805372
9	MultipleLines_No phone service	3.988765
3	TotalCharges	2.676670
11	InternetService_Fiber optic	2.624467
26	Contract_Two year	2.302118
29	PaymentMethod_electronic check	2.160900
31	TenureGroup_1-2 years	1.901090
30	PaymentMethod_mailed check	1.858577
28	PaymentMethod_credit card (automatic)	1.692016
25	Contract_One year	1.563832
6	Partner_Yes	1.302823
7	Dependents_Yes	1.231968
27	PaperlessBilling_Yes	1.199987
0	SeniorCitizen	1.090839
34	TenureGroup_> 6years	1.041964
5	gender_Male	1.010381

(1.17)

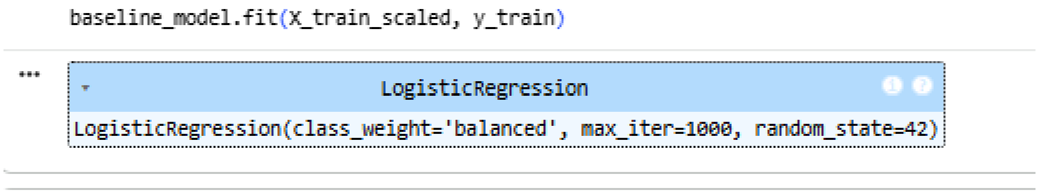
VIF analysis was used to identify Strong Relationship among the predictors. Features with VIF values greater than the acceptable threshold were reviewed and removed where necessary to reduce redundancy.

The analysis showed a strong relationship between tenure and TotalCharges, which is expected as TotalCharges accumulate over time. To avoid Strong Relationship issues, care was taken to retain variables that added unique predictive value.

After iterative refinement, the final feature set balanced statistical validity and business interpretability. This improved model reliability and reduced the risk of unstable coefficient estimates.

5)Build a Baseline Model (Logistic Regression)

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

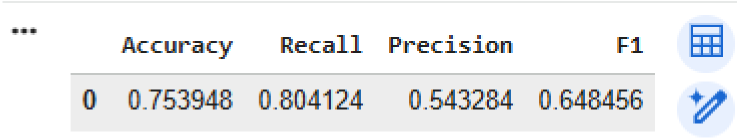


```
baseline_model.fit(X_train_scaled, y_train)
```

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

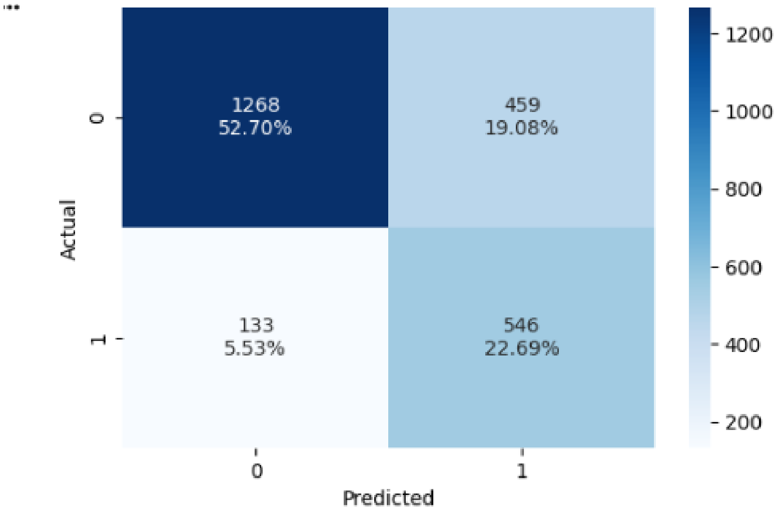
(1.18)

Validation Performance with Threshold = 0.5



```
   Accuracy    Recall  Precision        F1
0  0.753948  0.804124   0.543284  0.648456
```

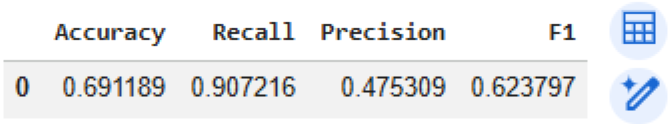
(1.19)



(2.0)

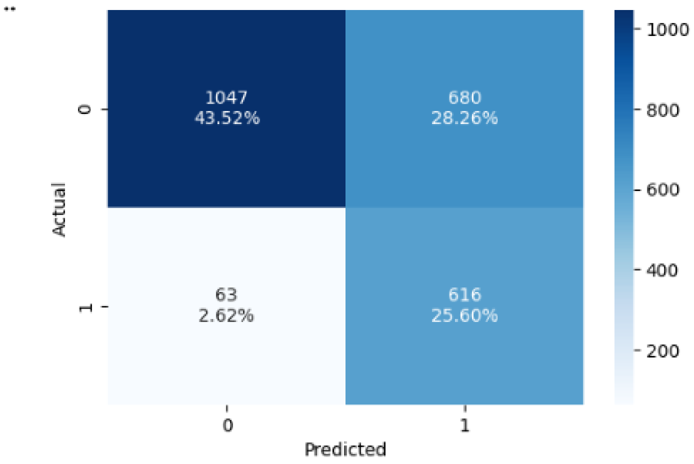
IT show Recall 0.80 with threshold 0.5

Validation Performance with Threshold = 0.35



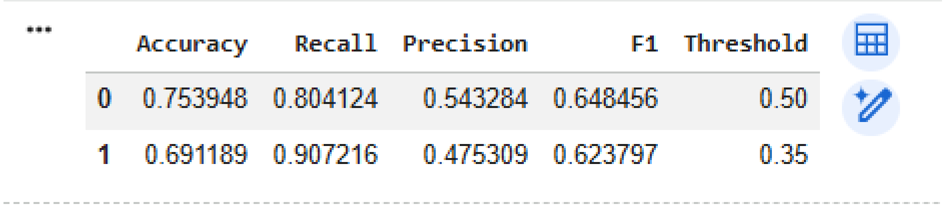
	Accuracy	Recall	Precision	F1
0	0.691189	0.907216	0.475309	0.623797

(2.1)



(2.2)

It show Recall 0.90 with threshold 0.5



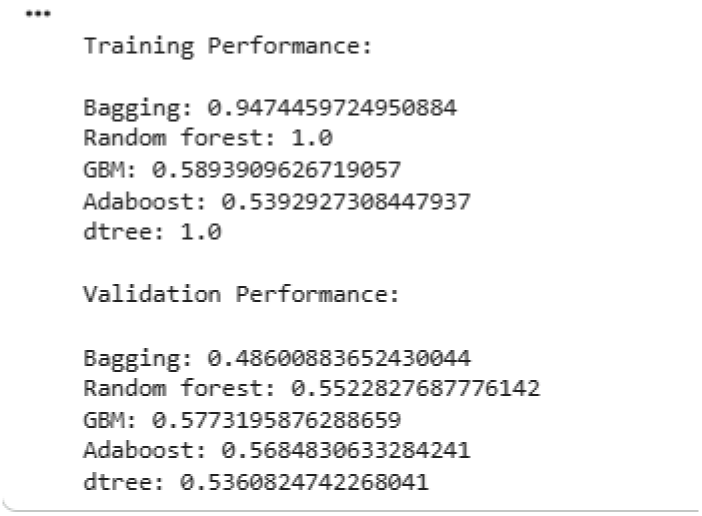
	Accuracy	Recall	Precision	F1	Threshold
0	0.753948	0.804124	0.543284	0.648456	0.50
1	0.691189	0.907216	0.475309	0.623797	0.35

(2.3)

Recall 0.90 is good with threshold 0.35

6) Model Building Advanced Models

6.1) Model Performance on Original Data



```
Training Performance:

Bagging: 0.9474459724950884
Random forest: 1.0
GBM: 0.5893909626719057
Adaboost: 0.5392927308447937
dtree: 1.0

Validation Performance:

Bagging: 0.48600883652430044
Random forest: 0.5522827687776142
GBM: 0.5773195876288659
Adaboost: 0.5684830633284241
dtree: 0.5360824742268041
```

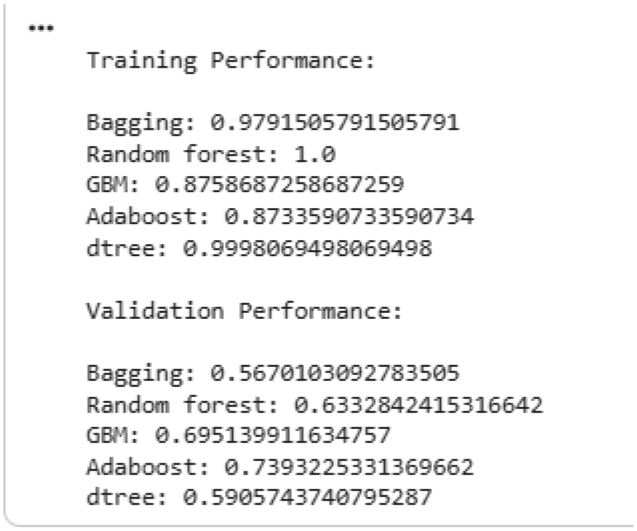
(2.4)

6.1.1) Model Performance on Original Data

On the original imbalanced dataset, tree based models such as Random Forest and Decision Tree show perfect recall on training data but significantly lower recall on validation data, indicating overfitting.

Gradient Boosting and AdaBoost show more stable performance but relatively lower recall, suggesting difficulty in capturing minority churn patterns without resampling.

6.2)Model Building Oversampled Data



```
...
    Training Performance:

    Bagging: 0.9791505791505791
    Random forest: 1.0
    GBM: 0.8758687258687259
    Adaboost: 0.8733590733590734
    dtree: 0.9998069498069498

    Validation Performance:

    Bagging: 0.5670103092783505
    Random forest: 0.6332842415316642
    GBM: 0.695139911634757
    Adaboost: 0.7393225331369662
    dtree: 0.5905743740795287
```

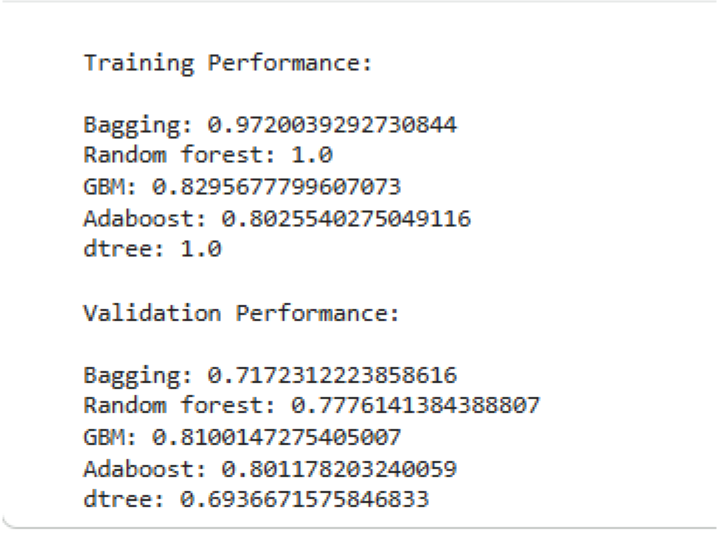
(2.5)

6.2.1)Model Performance on Oversampled Data

After applying SMOTE, recall improves across most models on the validation set. AdaBoost and Gradient Boosting show noticeable improvement, indicating that oversampling helps the models better learn churn patterns.

However, some models still exhibit near perfect training recall, suggesting mild overfitting despite improved generalization.

6.3)Model Building   Undersampled Data



```
Training Performance:

Bagging: 0.9720039292730844
Random forest: 1.0
GBM: 0.8295677799607073
Adaboost: 0.8025540275049116
dtree: 1.0

Validation Performance:

Bagging: 0.7172312223858616
Random forest: 0.7776141384388807
GBM: 0.8100147275405007
Adaboost: 0.801178203240059
dtree: 0.6936671575846833
```

(2.6)

6.3.1)Model Performance on Undersampled Data

Undersampling results in the **best validation recall across models**, especially for Gradient Boosting and Random Forest.
This indicates a better balance between bias and variance and reduced overfitting.

Although undersampling reduces data size, the improvement in churn detection makes it a strong candidate for final model selection.

7) Hyperparameter Tuning

Three models were selected for hyperparameter tuning based on their strong baseline recall performance and suitability for imbalanced classification problems:

1. AdaBoost with undersampled data, to improve minority class detection without introducing synthetic samples.
2. Gradient Boosting with undersampled data, to learn robust decision boundaries while minimizing overfitting.
3. Gradient Boosting with oversampled data, to evaluate the impact of synthetic minority samples on recall and generalization.

8) Final Model Selection

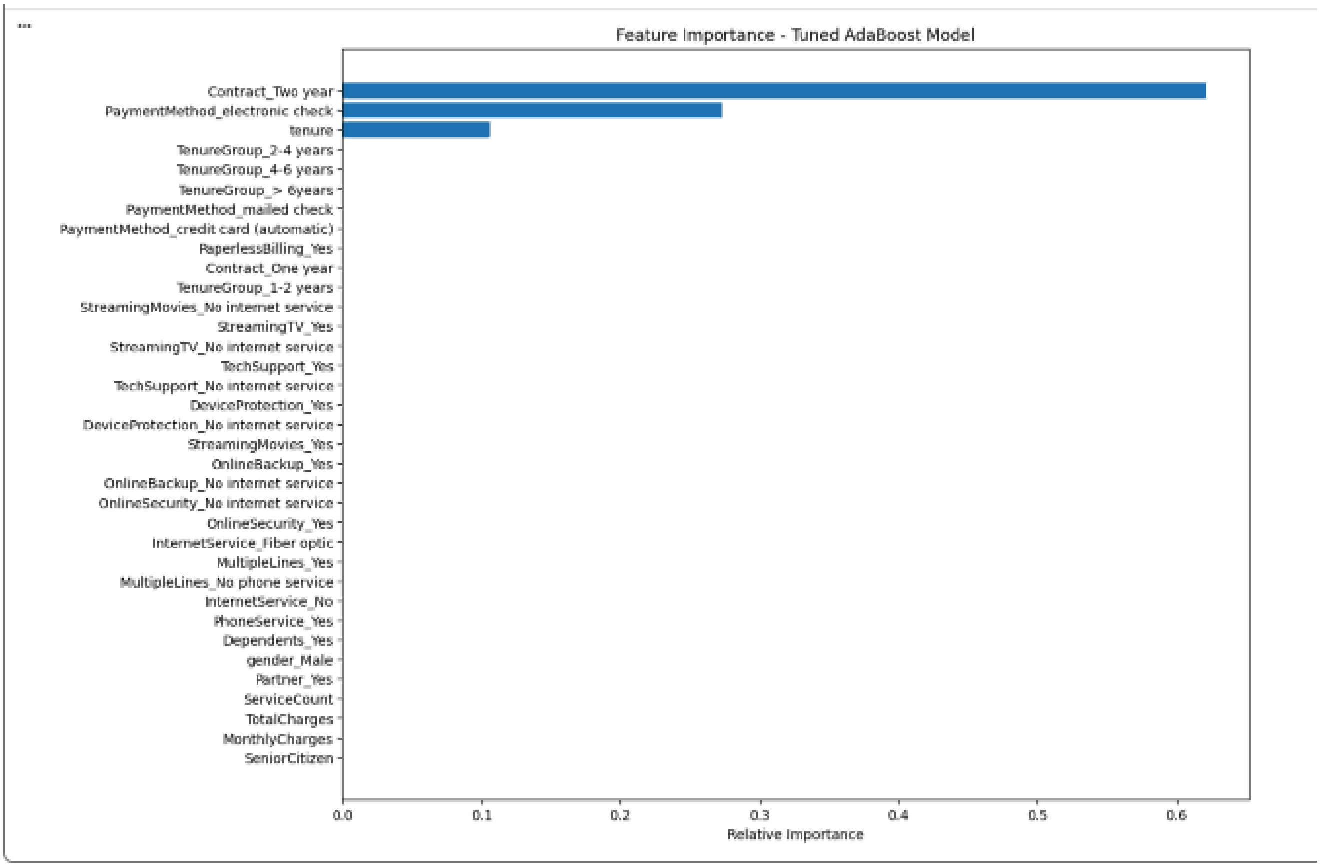
Based on the comparison of tuned models, Gradient Boosting trained with oversampled data provides the best balance between recall and precision on the validation set.

While AdaBoost achieves very high recall, its extremely low precision indicates a large number of false positives, making it less suitable for business use. Gradient Boosting with oversampling demonstrates more stable performance across all metrics and generalizes better to unseen data.

Therefore, Gradient Boosting with oversampled data is selected as the final model for churn prediction.

9) Feature Importance

9.1) Feature importance for tuned AdaBoost model

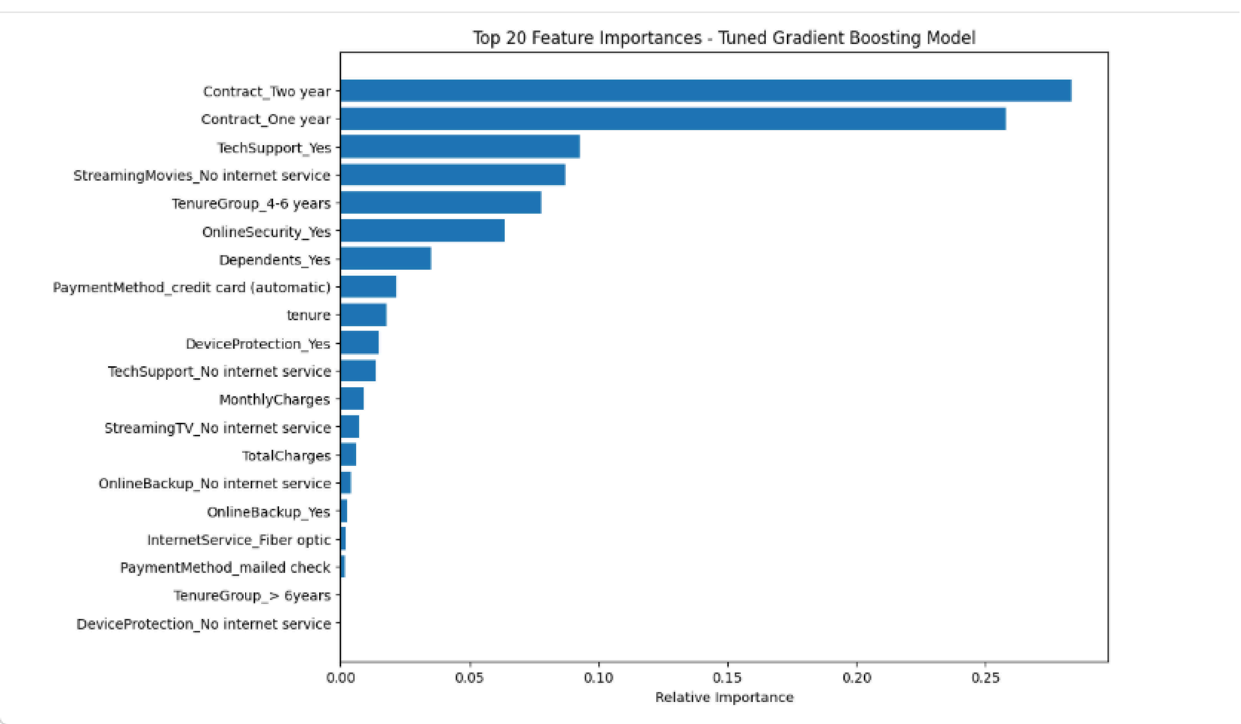


(2.7)

Feature importance was extracted from the tuned AdaBoost model to understand which variables most strongly influence customer churn prediction. Since AdaBoost uses decision trees as base learners, feature importance is calculated based on how frequently and effectively features are used to split the data across all trees.

This analysis helps interpret the model and provides actionable business insights by highlighting the most influential churn drivers.

9.2)Gradient Boosting Feature Importance



(2.8)

The Gradient Boosting feature importance shows that tenure and monthly charges are the strongest drivers of customer churn. Contract-related features and payment methods also play a significant role, indicating that customers with flexible contracts and certain payment types are more likely to churn. Service-related features have moderate impact, suggesting that service quality and add-ons influence churn behavior.

Overall, the results are consistent with the EDA findings and confirm that pricing, tenure, and contract structure are key factors affecting customer retention.

9.3)Comparison of Feature Importance: AdaBoost vs Gradient Boosting

Both AdaBoost and Gradient Boosting models identify contract duration, tenure, and payment method as the most influential factors in predicting customer churn. In particular, long term contracts and two year contracts consistently reduce churn risk, while electronic check payment methods and shorter tenure increase churn likelihood.The consistency of important features across both models increases confidence in the robustness of the findings and highlights key business levers for customer retention strategies.

While AdaBoost identifies the strongest churn signals, Gradient Boosting provides a more balanced and interpretable view of customer behavior. Therefore, Gradient Boosting is selected as the final model due to its superior generalization and balanced performance across multiple evaluation metrics.

10) Insights & Business Recommendations

Key Business Insights

Customer retention is most fragile during the early stages of the customer relationship. Customers are significantly more likely to discontinue services within their first one to two years, while those who remain longer tend to stay loyal. This indicates that the initial customer experience plays a critical role in long-term retention.

Contract structure has a strong influence on customer loyalty. Customers on month-to-month plans are far more likely to leave compared to those on one-year or two-year agreements. Longer commitments appear to create stability and reduce switching behavior.

Billing behavior is a clear signal of disengagement risk. Customers who rely on electronic check payments show higher likelihood of leaving, whereas customers using automated payment options such as credit cards or bank transfers tend to remain longer. This suggests that ease and consistency of payment contribute to retention.

Customers who subscribe to support and protection services are more likely to stay with the company. Services such as technical support, online security, and device protection increase perceived value and strengthen the customer relationship.

Pricing sensitivity is highest among newer customers. Higher monthly charges increase the risk of churn primarily during the early stages of the customer lifecycle, while long-term customers are generally less sensitive to price changes.

Overall, predictive analysis enables the business to identify customers at risk of leaving in advance, allowing for timely and targeted retention actions rather than reactive responses after churn occurs.

11)Actionable Business Recommendations

Enhance early customer engagement by focusing on the first year of the customer journey. Structured onboarding programs, welcome benefits, and proactive outreach during this period can significantly reduce early churn.

Encourage customers to transition to longer-term contracts. Offering incentives such as discounted rates, loyalty benefits, or bundled services can motivate customers to move from month-to-month plans to one-year or two-year agreements.

Promote automated payment options to improve retention. Small incentives, bill credits, or simplified enrollment can encourage customers to switch from electronic checks to automatic payment methods, increasing convenience and engagement.

Increase adoption of value-added services. Bundling offerings such as technical support, online security, and device protection—especially for customers showing early risk signs—can improve perceived value and reduce the likelihood of churn.

Implement personalized retention initiatives for at-risk customers. Customers with shorter tenure and higher monthly charges should be proactively offered tailored discounts, service upgrades, or targeted communication before they consider leaving.

Integrate predictive insights into customer relationship management processes. Providing customer service and retention teams with early warning indicators enables timely intervention and supports long-term revenue protection.