12/31/2025

TELECOM CUSTOMER CHURN PREDICTION

BY KOMAL

Table of Contents

BUSINESS PROBLEM : CUSTOMER CHURN PREDICTION

Business Context:
AlphaCom, a leading telecommunications provider, has recently experienced a concerning rise in customer churn despite offering competitive services and a wide product portfolio. This increase is directly impacting revenue and undermining brand reputation in an intensely competitive market. Traditional retention strategies have proven inadequate because customer churn is influenced by a complex mix of factors, including service usage, billing preferences, contract types, and demographics. Without clear insights into these patterns, the company is left reacting to churn instead of preventing it.

Objective:
As a data scientist at AlphaCom, you are tasked with developing a predictive model to identify customers at high risk of churn and uncover the key factors driving their decisions. Solving this problem will enable the company to proactively design targeted retention strategies, reduce churn-related losses, and improve customer lifetime value, ultimately safeguarding revenue and strengthening AlphaCom's competitive position.

Data Description:
The data contains different attributes related to churn. The detailed data dictionary is given below:

- **Gender:** The customer's gender (e.g., Male or Female). This demographic feature may correlate with customer behavior.
- **Age Range:** Indicates the customer's age bracket (e.g., 18–25, 26–35, etc.), offering demographic insights that can impact churn analysis.
- **SeniorCitizen:** A binary indicator (if included) that identifies whether the customer is a senior citizen (commonly 1 for senior, 0 for non-senior). Senior status can influence service preferences and retention strategies.
- **Partner:** Indicates whether the customer has a partner. This factor can affect customer loyalty and service usage patterns.
- **Dependents:** Specifies whether the customer has dependents. This information can provide context on the customer's household and influence their service needs.
- **Tenure:** The number of months the customer has been with the company. Longer tenure may indicate higher loyalty, while shorter tenure could be a churn risk indicator.
- **PhoneService:** Denotes whether the customer subscribes to telephone services. This binary feature (Yes/No) helps understand service adoption.
- **MultipleLines:** Indicates if the customer has multiple phone lines. This feature can provide insight into customer behavior and service complexity.
- **InternetService:** Describes the type of internet service the customer uses (e.g., DSL, Fiber optic, or None). The type of internet service can be a critical factor in churn analysis.
- **OnlineSecurity:** Shows whether the customer subscribes to online security services. This value (Yes/No) may influence customer satisfaction and retention.
- **OnlineBackup:** Indicates if the customer has an online backup service. Similar to online security, this can be a part of the overall service bundle affecting churn.
- **DeviceProtection:** Specifies whether the customer is enrolled in a device protection plan, providing an added layer of service value.
- **TechSupport:** Denotes if the customer subscribes to technical support services. Access to tech support can improve customer experience and reduce churn.
- **StreamingTV:** Indicates whether the customer subscribes to a streaming TV service. Media consumption patterns can be a differentiator in customer preferences.
- **StreamingMovies:** Specifies if the customer subscribes to a streaming movies service. This, combined with other services, can highlight trends in customer behavior.
- **Contract:** Describes the type of contract the customer holds (e.g., month-to-month, one-year, or two-year). Contract type is a strong indicator of churn risk—shorter contracts are often associated with higher churn.
- **PaperlessBilling:** Indicates whether the customer is enrolled in paperless billing. This operational feature can sometimes correlate with customer engagement levels.

- **PaymentMethod:** Details the payment method used by the customer (e.g., electronic check, mailed check, bank transfer, or credit card). Payment methods can affect both churn and overall customer satisfaction.
- **MonthlyCharges:** The monthly amount in $ USD charged to the customer. Higher charges might increase the likelihood of churn if customers perceive the cost as too high for the value provided.
- **TotalCharges:** The cumulative amount in $ USD charged over the customer's tenure. This helps in understanding the long-term value of each customer and can be a predictor of churn.
- **Churn:** The target variable indicating whether the customer has left (typically denoted as "Yes" or "No"). This is the primary outcome you aim to predict with your machine learning model.

ImportantNote:
Reasons why TotalCharges might not exactly equal Tenure × MonthlyCharges:
Prorated Billing & Partial Months: If a customer signs up or cancels partway through a billing cycle, their first or last month's charge may be prorated (i.e., only for the days they actually had service), so it won't match a full "monthly" amount.
One-Off Fees and Credits: Installation fees, equipment charges, early-termination fees, late-payment penalties, or promotional credits can all be applied directly to TotalCharges without affecting the regular MonthlyCharges.

CONCISE SUMMARY OF BUSINESS PROBLEM

1. Business Problem

AlphaCom is experiencing rising customer churn despite offering competitive telecom services. This churn is negatively impacting revenue, customer lifetime value, and brand perception. Existing retention strategies are largely reactive, as churn is driven by multiple factors such as pricing, service usage, contract type, billing behavior, and customer demographics. The lack of predictive insights prevents AlphaCom from identifying high-risk customers early and taking targeted retention actions, leading to avoidable revenue loss.

---

2. Target Variable

Churn is the primary target variable.

- Type: Binary classification
- Values:
  - Yes – Customer has churned
  - No – Customer is retained

All remaining variables serve as predictor features explaining churn behavior.

---

3. Model Performance Metrics

Given the business cost associated with churn, the key evaluation metrics are:

- **Recall (Churn = Yes):** Most critical metric to minimize missed churners
- **Precision (Churn = Yes):** Ensures cost-efficient retention efforts
- **F1-Score:** Balances precision and recall
- **ROC-AUC:** Measures overall discrimination capability

The project is successful if AlphaCom can accurately predict churn, understand its drivers, and proactively retain high-risk customers, leading to measurable improvements in revenue retention and customer lifetime value.

DATASET OVERVIEW

Observation :-

- Total records (customers) = 12055
- Total Columns = 20
- Data types: int, float, object divided into categorical columns(18) and numerical columns(2).
- Missing values (Non-Null Count) = "Tenure" column has 604 missing values.

Telecom Customer churn dataset have :-

- **Demographic data** = Gender, Senior citizens, Partner, Dependents
- **Service Taken** = Phone, Internet, add-ons
- **Contract & Billing details**
- **Charges**
- **Churn** - "Target Variables"

The preview clearly shows the data is NOT clean and has missing values.

```
____________________________________________
The information of data are:
<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
None
```

Figure 1: DTYPE INFORMATION OF DATASET

```
_________________________________________________________________
The first 5 rows of dataset are:
   gender  SeniorCitizen Partner Dependents  tenure PhoneService  \
0  Female              0     Yes         No     1.0           No
1    Male              0      No         No    34.0          Yes
2    Male              0      No         No     2.0          Yes
3    Male              0      No         No    45.0           No
4  Female              0      No         No     2.0          Yes

      MultipleLines InternetService OnlineSecurity OnlineBackup  \
0  No phone service             DSL             No          Yes
1                No             DSL            Yes           No
2                No             DSL            Yes          Yes
3  No phone service             DSL            Yes           No
4                No     Fiber optic             No           No

  DeviceProtection TechSupport StreamingTV StreamingMovies        Contract  \
0               No          No          No              No  Month-to-month
1              Yes          No          No              No        One year
2               No          No          No              No  Month-to-month
3              Yes         Yes          No              No        One year
4               No          No          No              No  Month-to-month

  PaperlessBilling             PaymentMethod MonthlyCharges TotalCharges  \
0              Yes          Electronic check         $29.85       $29.85
1               No              Mailed Check         $56.95      $1889.5
2              Yes              Mailed check         $53.85      $108.15
3               No  bank transfer (automatic)         $42.3     $1840.75
4              Yes          ELECTRONIC CHECK          $70.7         $nan

    Churn
0    No
1      NO
2    YES
3    No
4     yes
```

Figure 2: FIRST 5 ROWS OF DATASET

```
________________________________
The missing values in dataset are:
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       0
TotalCharges         0
Churn                0
dtype: int64
```
Figure 3: MISSING VALUES IN DATASET

PHASE 1 – DATASET STRUCTURAL CLEANING

STANDARDIZED COLUMN NAMES

```
<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  int64  
 3   dependents        12055 non-null  int64  
 4   tenure            11451 non-null  float64
 5   phoneservice      12055 non-null  int64  
 6   multiplelines     12055 non-null  int64  
 7   internetservice   12055 non-null  object 
 8   onlinesecurity    12055 non-null  int64  
 9   onlinebackup      12055 non-null  int64  
 10  deviceprotection  12055 non-null  int64  
 11  techsupport       12055 non-null  int64  
 12  streamingtv       12055 non-null  int64  
 13  streamingmovies   12055 non-null  int64  
 14  contract          12055 non-null  object 
 15  paperlessbilling  12055 non-null  int64  
 16  paymentmethod     12055 non-null  object 
 17  monthlycharges    11754 non-null  float64
 18  totalcharges      10850 non-null  float64
 19  churn             12055 non-null  int64  
dtypes: float64(3), int64(13), object(4)
memory usage: 1.8+ MB
Phase -1 cleaned data information: None
```

Figure 4: STANDARDIZED COLUMN NAMES

MISSING VALUES AFTER STRUCTURAL CLEANING

```
Missing values after structural cleaning:
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
```

Figure 5: MISSING VALUES AFTER STRUCTURAL CLEANING

Observation

- Ensure consistent column access.
- Prevent duplicate categories.
- Handles ,dollar sign,commas, spaces etc..
- Converts $nan-type values
- Prevents unintended object/float issues.
- Correct dtypes.

Dataset is now:

- Structurally clean
- Machine-readable
- Ready for EDA
- Ready for Phase 2 – Analytical Cleaning

EXPLORATORY DATA ANALYSIS

SUMMARY OF VARIABLES

EDA DIMENSION	VARIABLE(S)	KEY OBSERVATION	BUSINESS / MODELING INSIGHT
Target Variable	Churn	Class imbalance with fewer churners than non-churners	Requires imbalance handling (SMOTE) and recall-focused metrics
Customer Lifecycle	tenure, tenure_group	Churn higher in early tenure; long-tenure customers are stable	Early-life retention is critical; tenure is a strong churn driver
Pricing	monthlycharges, totalcharges	Higher monthly charges associated with higher churn	Price sensitivity exists; value perception matters
Demographics	gender	Nearly balanced distribution	Weak standalone predictor
	seniorcitizen	Senior citizens form a smaller segment with slightly higher churn	Niche but relevant demographic
Household Stability	partner, dependents	Customers without partner/dependents churn more	Family stability reduces churn risk
Core Services	phoneservice, internetservice	Internet service type impacts churn	Service quality and expectations differ by type
Internet Type	internetservice_fiber optic	Fiber optic customers show higher churn	High expectations; service experience critical
Service Engagement	onlinesecurity, onlinebackup, deviceprotection, techsupport	Add-on services reduce churn	Engagement depth improves retention
Entertainment Usage	streamingtv, streamingmovies	Subscribers churn less than non-users	Content usage increases stickiness
Usage Intensity	multiplelines	Multi-line users churn less	Higher dependency lowers switching
Billing Preference	paperlessbilling	Digital billing users slightly more churn-prone	Self-service users are less loyal
Payment Behavior	paymentmethod_electronic check	Highest churn association	Payment friction is a churn signal
Contract Type	contract_month-to-month, contract_one year, contract_two year	Month-to-month has highest churn; long-term contracts retain	Contract duration is a strong retention lever
Multicollinearity	tenure vs totalcharges	Strong positive correlation	Structural dependency; handled in modeling
Overall Pattern	Multiple features	Churn driven by lifecycle, pricing, engagement, and contract	Multi-factor churn behavior

TARGET VARIABLE: CHURN DISTRIBUTION

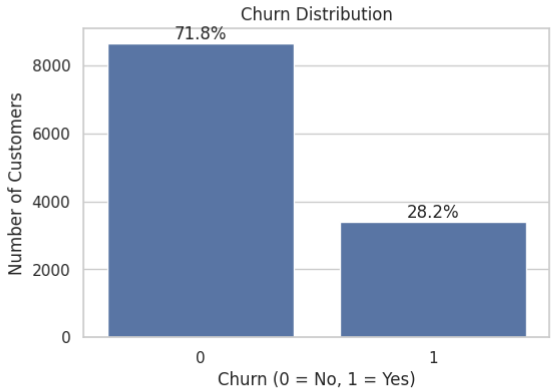


Figure 6: CHURN DISTRIBUTION

Observation

The target variable exhibits significant class imbalance, with non-churn customers dominating the dataset. This necessitates the use of imbalance-handling techniques and evaluation metrics beyond accuracy to ensure effective churn prediction.

TENURE

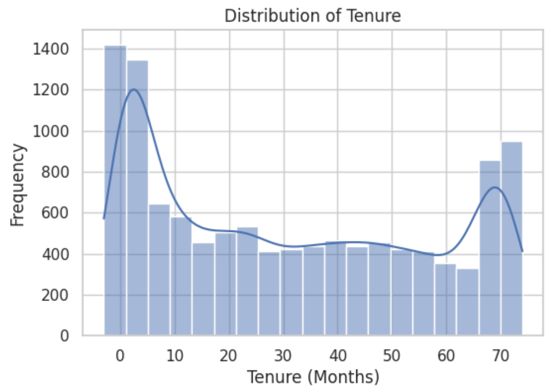


Figure 7: DISTRIBUTION OF TENURE

Observation

- Distribution is right-skewed, with high concentration of customers toward lower tenure values.
- This indicates that a large number of customers are relatively new, while fewer customers have long-term association.
- The skewness suggests that tenure may have a strong influence on churn behavior, especially in early customer life cycles.

MONTHLY CHARGES

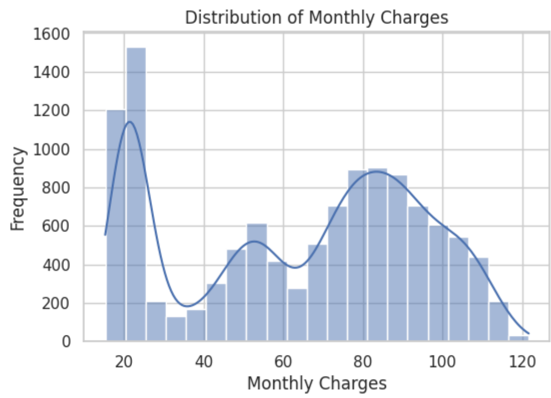
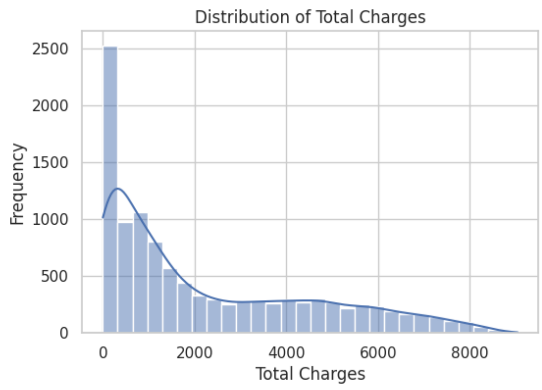


Figure 8: DISTRIBUTION OF MONTHLY CHARGES

Observation

- Monthly charges show a moderately skewed distribution, spread across a wide range of values.
- The presence of higher charge values indicates diverse pricing plans and service bundles.
- The wide spread suggests that monthly charges could be a key differentiating factor among customers.

TOTAL CHARGES

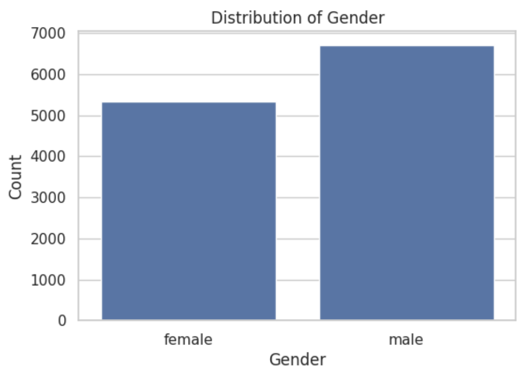


Observation

- Total charges exhibit a heavily right-skewed distribution, with most customers clustered at lower values.
- This pattern is expected, as total charges accumulate over time and are influenced by tenure.
- The skewness indicates a strong dependence on customer duration, making this variable closely related to tenure.

Figure 9: DISTRIBUTION OF TOTAL CHARGES

GENDER

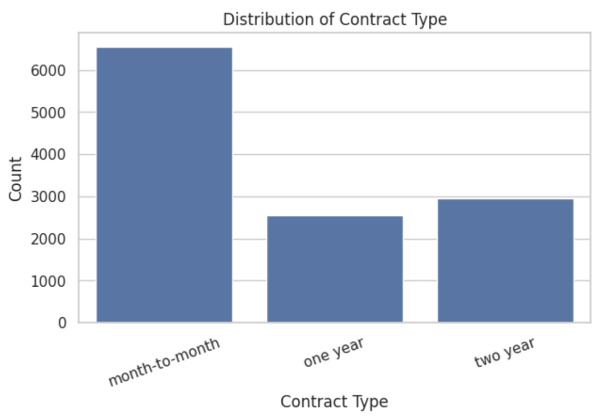


Observation

- The gender distribution appears to be fairly balanced across customers.
- No extreme dominance of any single category is observed.
- This suggests that gender alone may not be a strong standalone predictor of churn.

Figure 10: DISTRIBUTION OF GENDER

CONTRACT TYPE

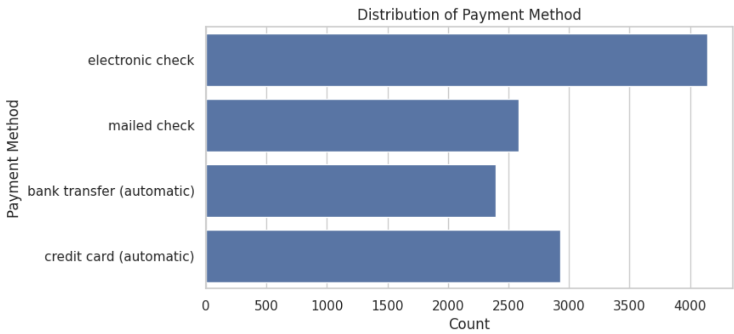


Observation

- A significant proportion of customers are on month-to-month contracts.
- Fewer customers opt for long-term contracts such as one-year or two-year plans.
- This indicates that a large segment of customers may have lower commitment levels, which can influence churn risk.

Figure 11: DISTRIBUTION OF CONTRACT TYPE

PAYMENT METHOD

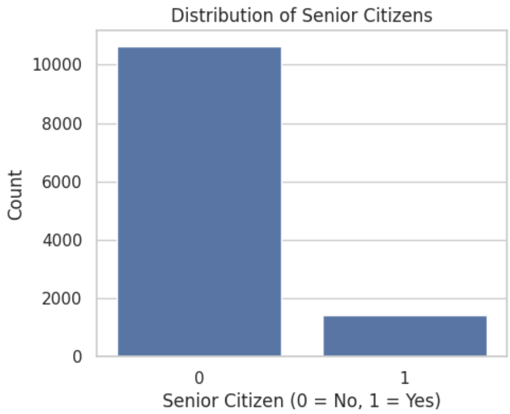


Observation

- Multiple payment methods are used by customers, with certain methods being more prevalent.
- The variation in payment preferences indicates behavioral diversity across customers.
- Payment method may serve as a proxy for customer engagement or convenience preference.

Figure 12: DISTRIBUTION OF PAYMENT METHOD

SENIOR CITIZENS

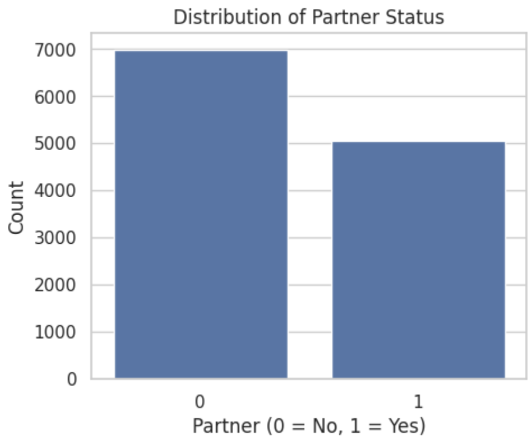


Observation

- The majority of customers are not senior citizens.
- Senior citizens form a smaller segment of the customer base.
- This imbalance suggests that senior citizen status may represent a distinct but limited demographic group.

Figure 13: DISTRIBUTION OF SENIOR CITIZENS

PARTNER STATUS



Observation

- Customers are fairly distributed between having and not having a partner.
- No extreme dominance of a single category is observed.
- Partner status may reflect household or family structure among customers.

Figure 14: DISTRIBUTION OF PARTNER STATUS

PHONE SERVICE

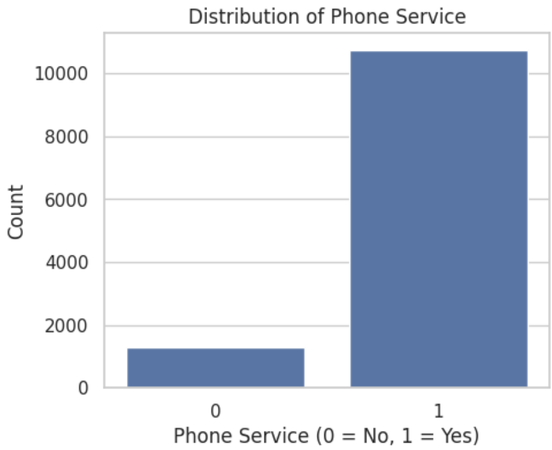


Figure 15: DISTRIBUTION OF PHONE SERVICE

Observation

- Most customers have phone service enabled.
- Customers without phone service form a minor segment.
- This indicates phone service is a core offering for most customers.

DEPENDENTS

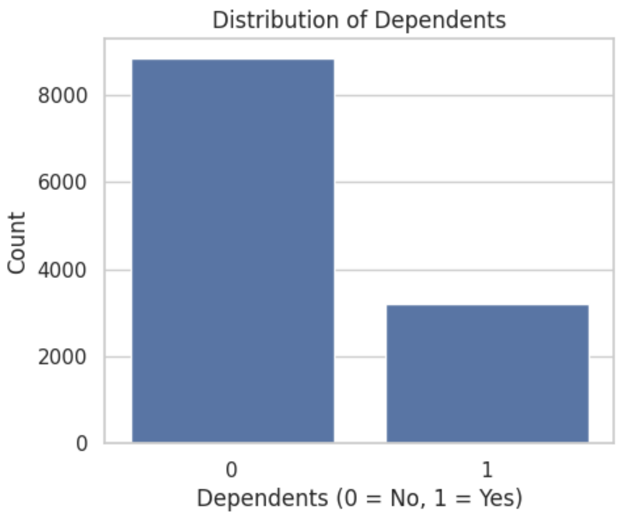
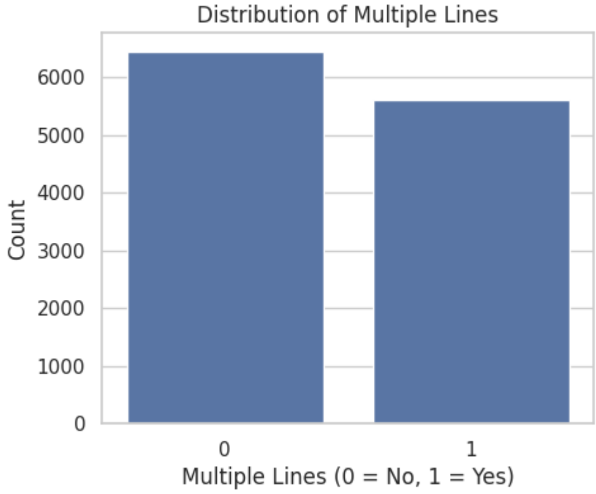


Figure 16: DISTRIBUTION OF DEPENDENTS

Observation

- A higher proportion of customers do not have dependents.
- Customers with dependents form a smaller share of the dataset.
- This variable captures differences in household responsibility levels.

MULTIPLE LINES

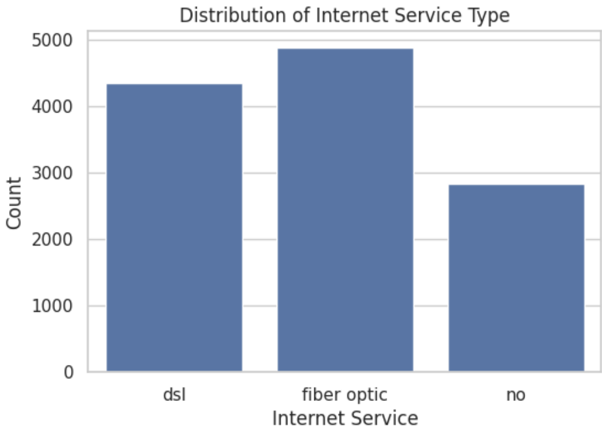


Observation

- Customers are distributed between single-line and multiple-line usage.
- A noticeable proportion of customers do not subscribe to multiple lines.
- This variable reflects usage intensity of phone services.

Figure 17: DISTRIBUTION OF MULTIPLE LINES

INTERNET SERVICE TYPE



Observation

- Customers are distributed across different internet service types.
- Certain service types have higher adoption compared to others.
- This suggests heterogeneity in service offerings and technology preference.

Figure 18: DISTRIBUTION OF INTERNET SERVICES

ONLINE SECURITY

Observation

- A substantial portion of customers do not subscribe to online security.
- Adoption of online security services is comparatively lower.
- This may indicate optional add-on services rather than core usage.

Figure 19: DISTRIBUTION OF ONLINE SECURITY

ONLINE BACKUP

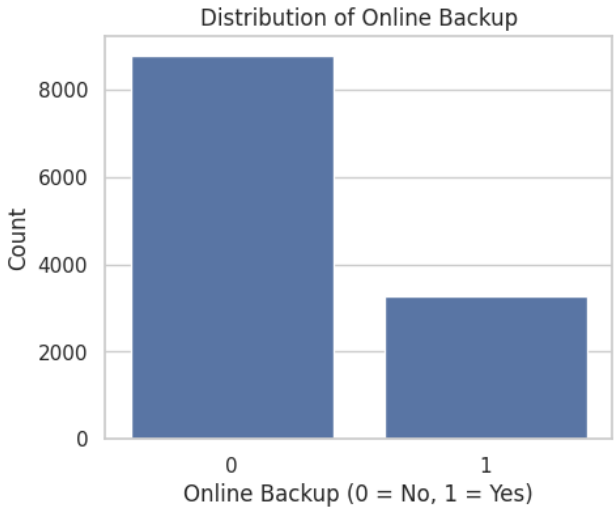


Figure 20: DISTRIBUTION OF ONLINE BACKUP

Observation

- Customers are unevenly distributed between having and not having online backup.
- A larger share of customers do not opt for this service.
- Online backup appears to be an optional value-added feature.

DEVICE PROTECTION

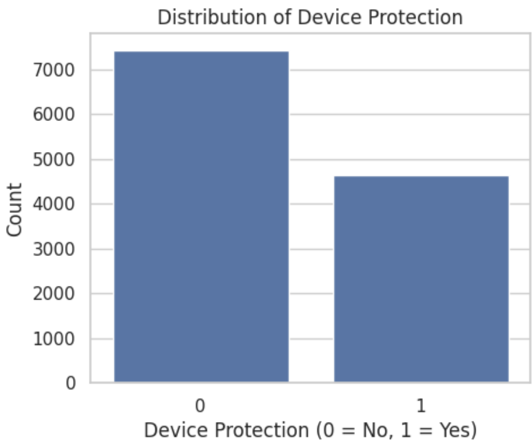


Figure 21: DISTRIBUTION OF DEVICE PROTECTION

Observation

- A considerable number of customers do not have device protection enabled.
- Subscription to device protection is not universal.
- This variable reflects customer preference for risk mitigation services.

TECH SUPPORT

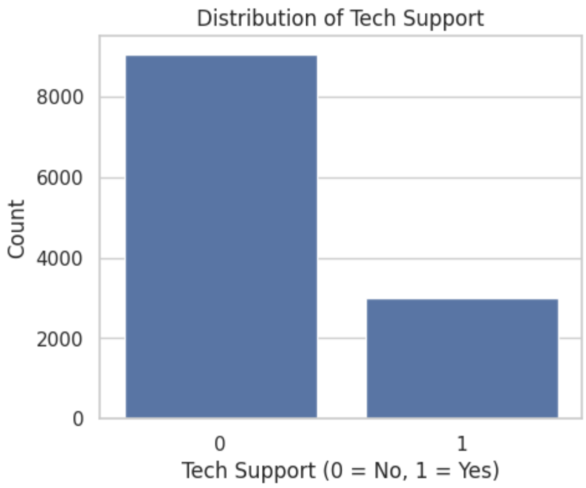


Figure 22: DISTRIBUTION OF TECH SUPPORT

Observation

- Many customers do not subscribe to technical support services.
- Tech support adoption varies significantly across the dataset.
- This suggests differences in customer self-reliance or service needs.

STREAMING TV

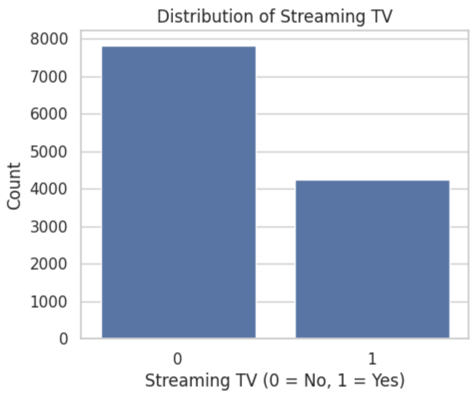


Figure 23:DISTRIBUTION OF STREAMING TV

Observation

- Streaming TV usage is split between subscribers and non-subscribers.
- Adoption is neither minimal nor universal.
- This indicates content consumption diversity among customers.

STREAMING MOVIES

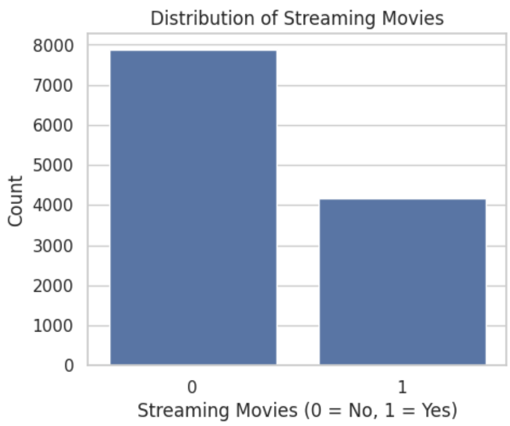


Figure 24: DISTRIBUTION OF STREAMING MOVIES

Observation

- Customers show mixed adoption of streaming movie services.
- Similar distribution patterns to streaming TV are observed.
- This reflects varying entertainment preferences.

PAPERLESS BILLING

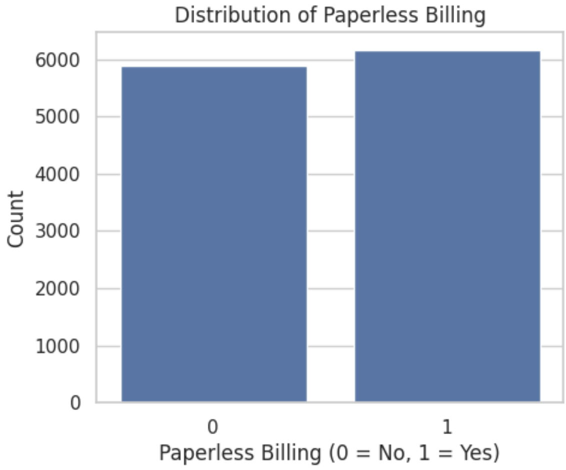


Figure 25: DISTRIBUTION OF PAPERLESS BILING

Observation

- The near-balanced distribution highlights the coexistence of digital and non-digital customer segments within the dataset.
- Paperless billing appears to be a widely accepted but not universal practice, making it a potentially meaningful behavioral attribute for further analysis.

BIVARIATE ANALYSIS

NUMERICAL VARIABLES VS CHRUN

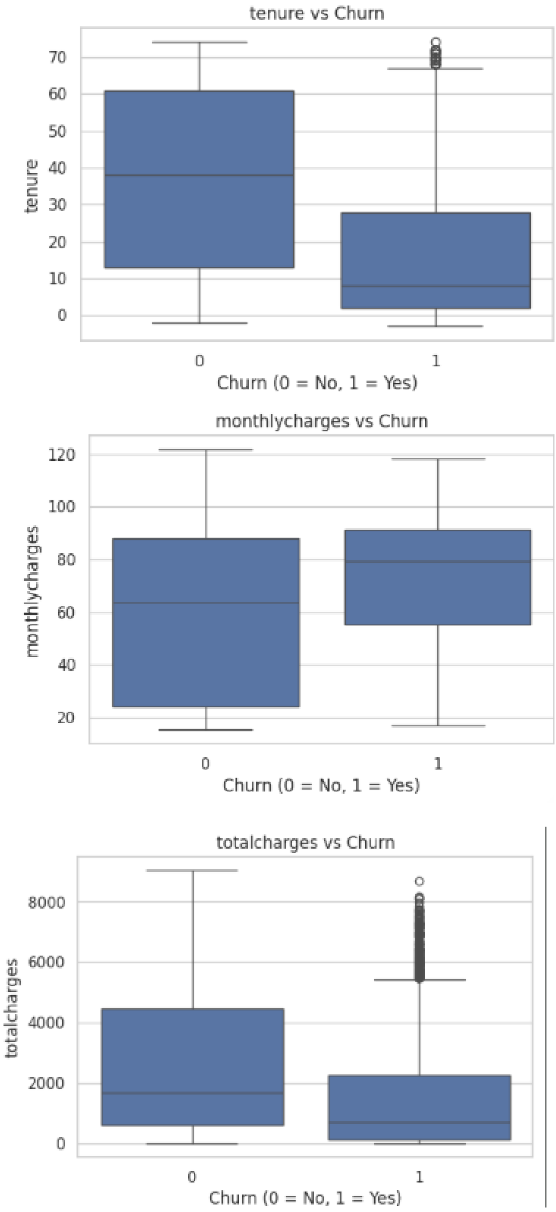
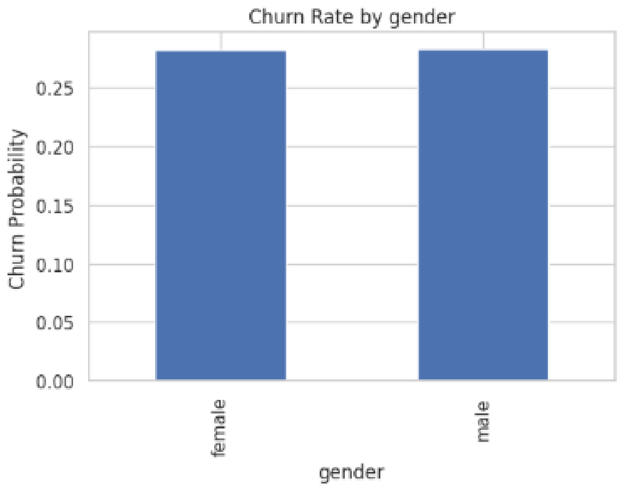


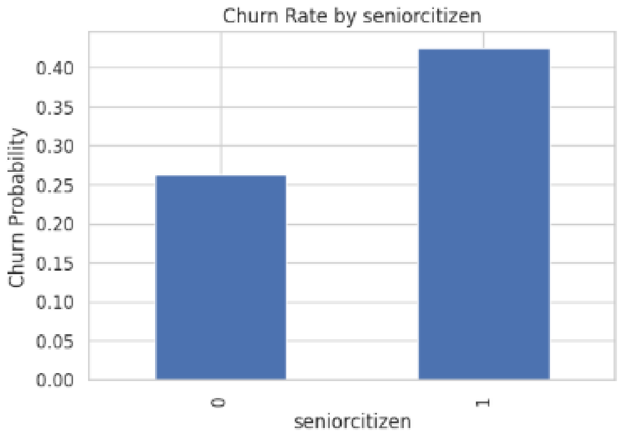
Figure 26: BOXPLOT FOR NUMERICAL VARIABLES VS CHURN

Observation

- Tenure: Customers who churn tend to have significantly lower tenure, indicating higher churn risk during early customer lifecycle.
- Monthly Charges: Churned customers generally exhibit higher monthly charges, suggesting possible price sensitivity.
- Total Charges: Non-churn customers show higher total charges, largely reflecting longer tenure rather than higher spending rate.

DEMOGRAPHIC VARIABLES VS CHURN





Observation

- Gender: Churn rates are relatively similar, indicating minimal differential impact.
- Senior Citizen: Senior citizens show a higher churn rate, suggesting distinct behavioral patterns.
- Partner & Dependents: Customers without partners or dependents tend to churn more, indicating potential links between household stability and retention.

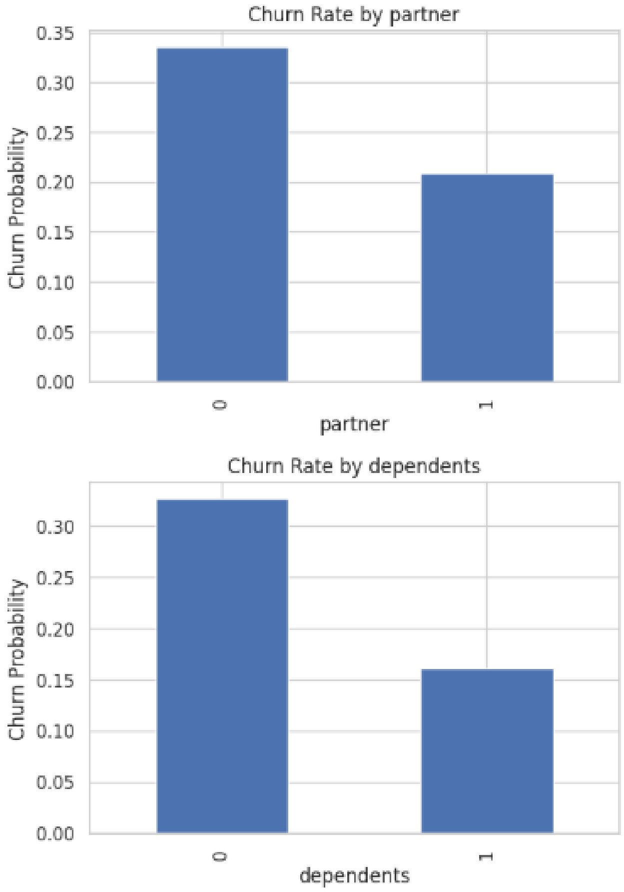


Figure 27: RELATION BETWEEN DEMOGRAPHIC VARIABLES VS CHURN

SERVICE USAGE VARIABLES VS CHURN

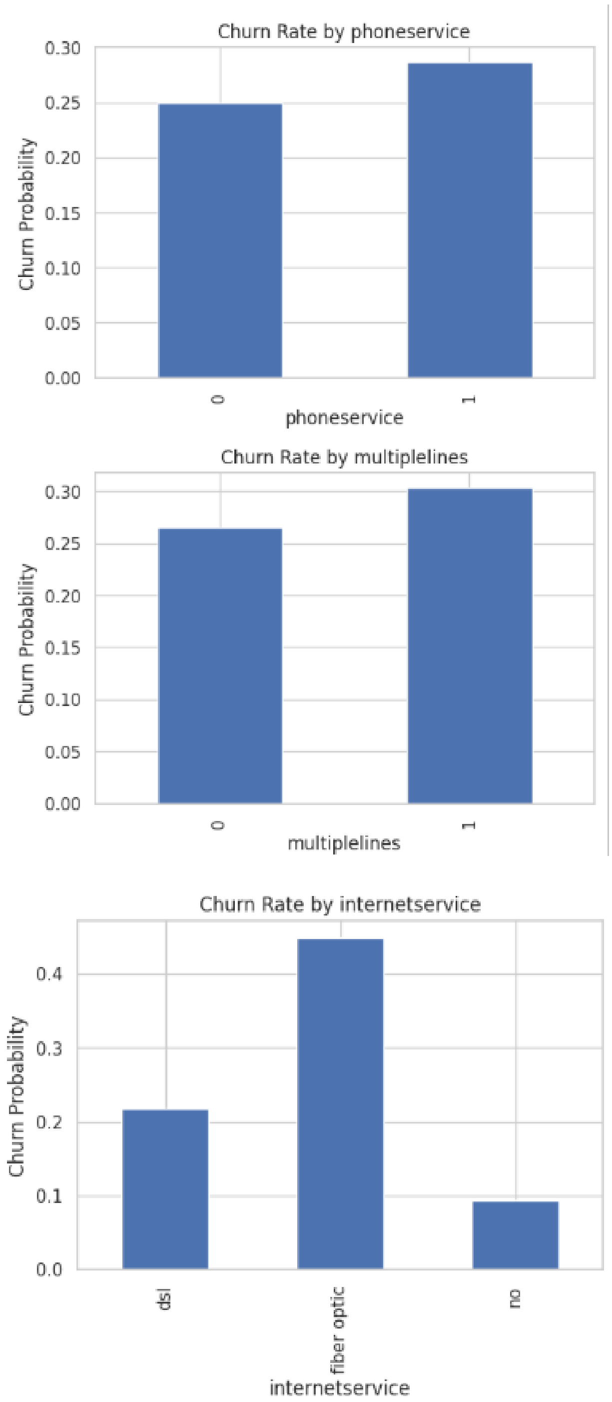


Figure 28: RELATION BETWEEN SERVICE USAGE VS CHURN

Observation

- Phone Service: Minimal churn differentiation; likely a baseline service.
- Multiple Lines: Customers without multiple lines show slightly higher churn.
- Internet Service: Certain internet service types exhibit noticeably higher churn, indicating service-quality or expectation mismatches.

ADD-ON SERVICES VS CHURN

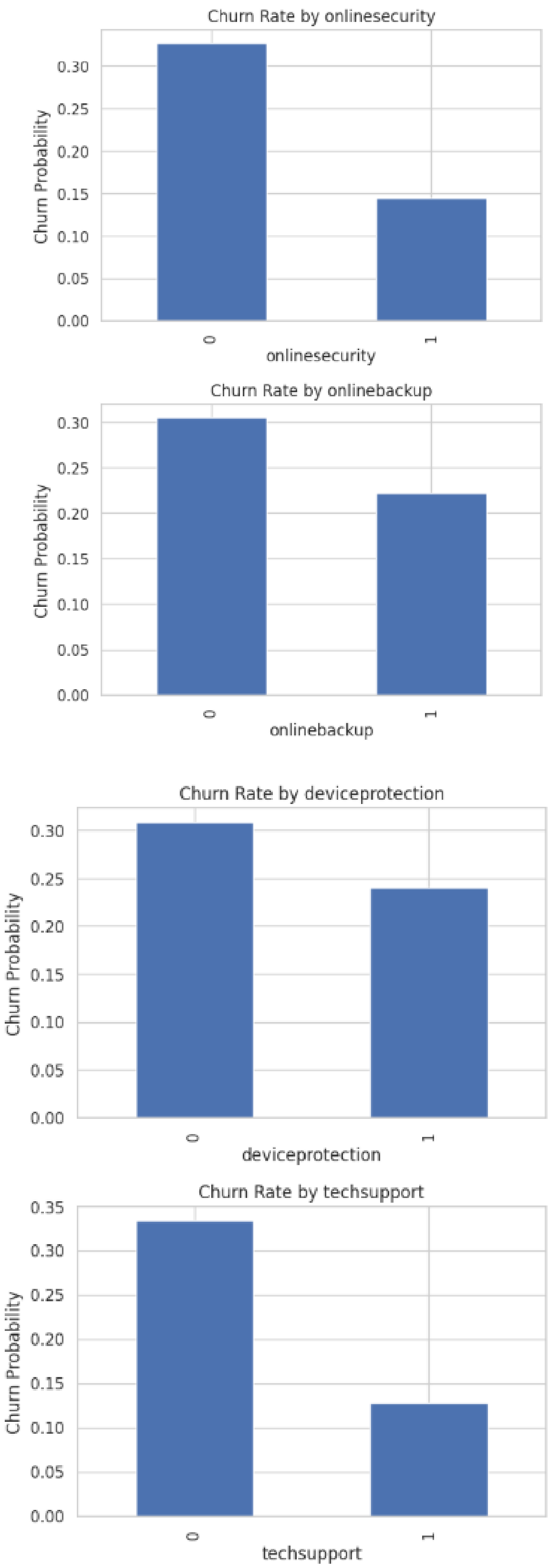
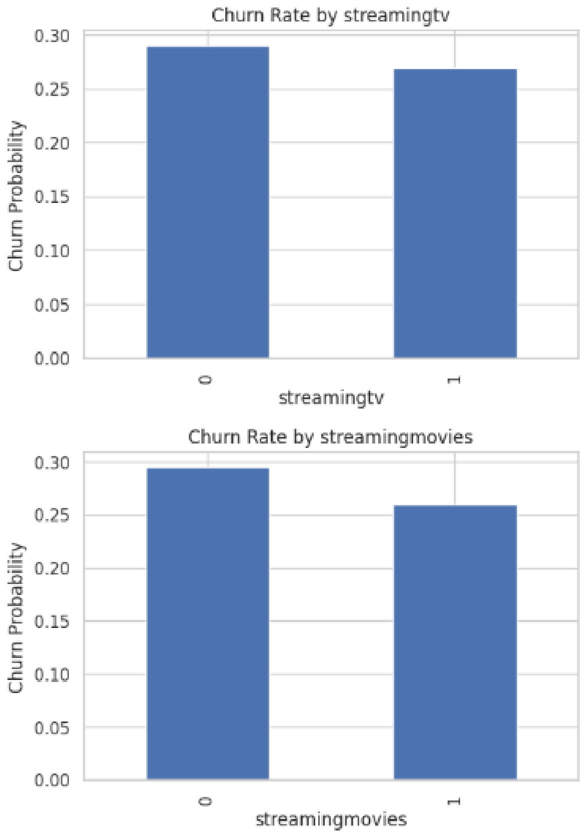


Figure 29: RELATION BETWEEN ADD-ON SERVICES VS CHURN

Observation

- Customers without add-on services consistently show higher churn rates.
- Presence of these services appears to be associated with greater customer stickiness.
- These features likely capture engagement depth rather than basic usage.

ENTERTAINMENT SERVICES VS CHRUN



Observation

- Customers not subscribing to streaming services show higher churn rates.
- Streaming services may act as engagement enhancers, increasing perceived value.

Figure 30:RELATION BETWEEN ENTERTAINMENT SERVICES VS CHURN

BILLING & PAYMENT VARIABLES VS CHURN

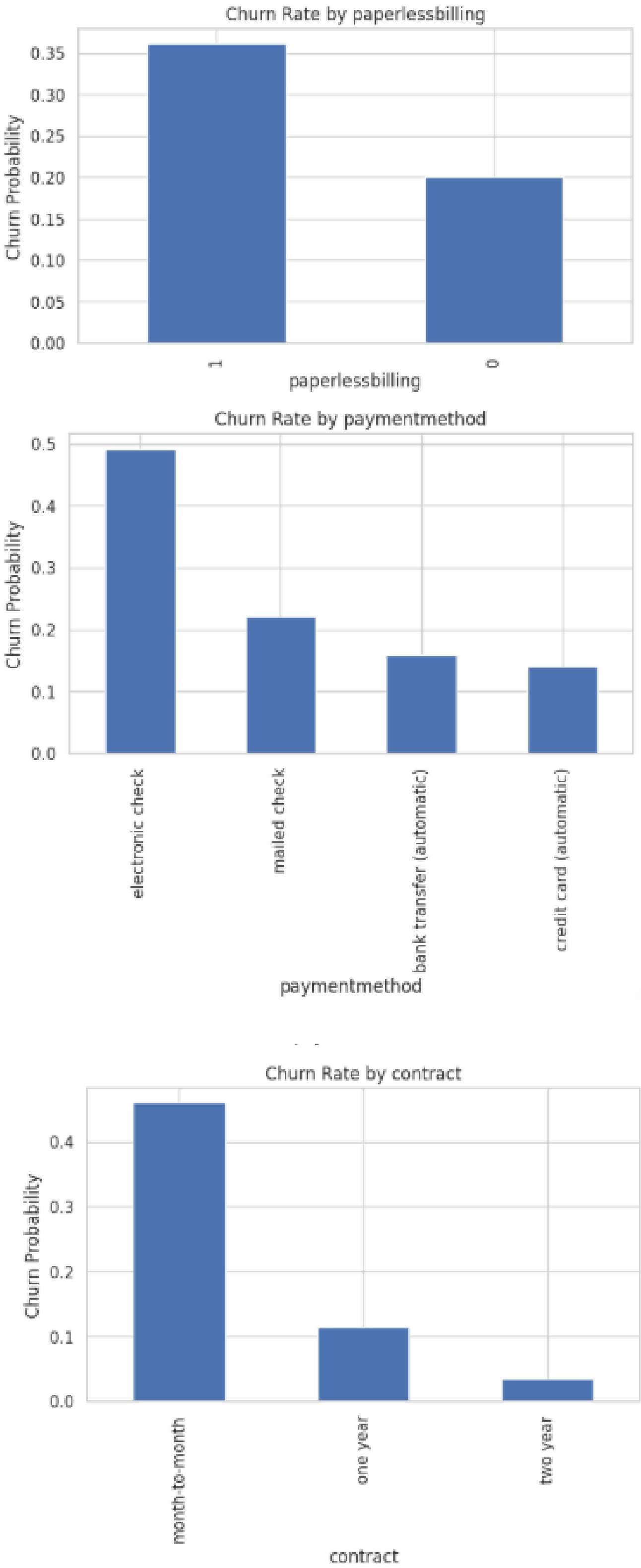
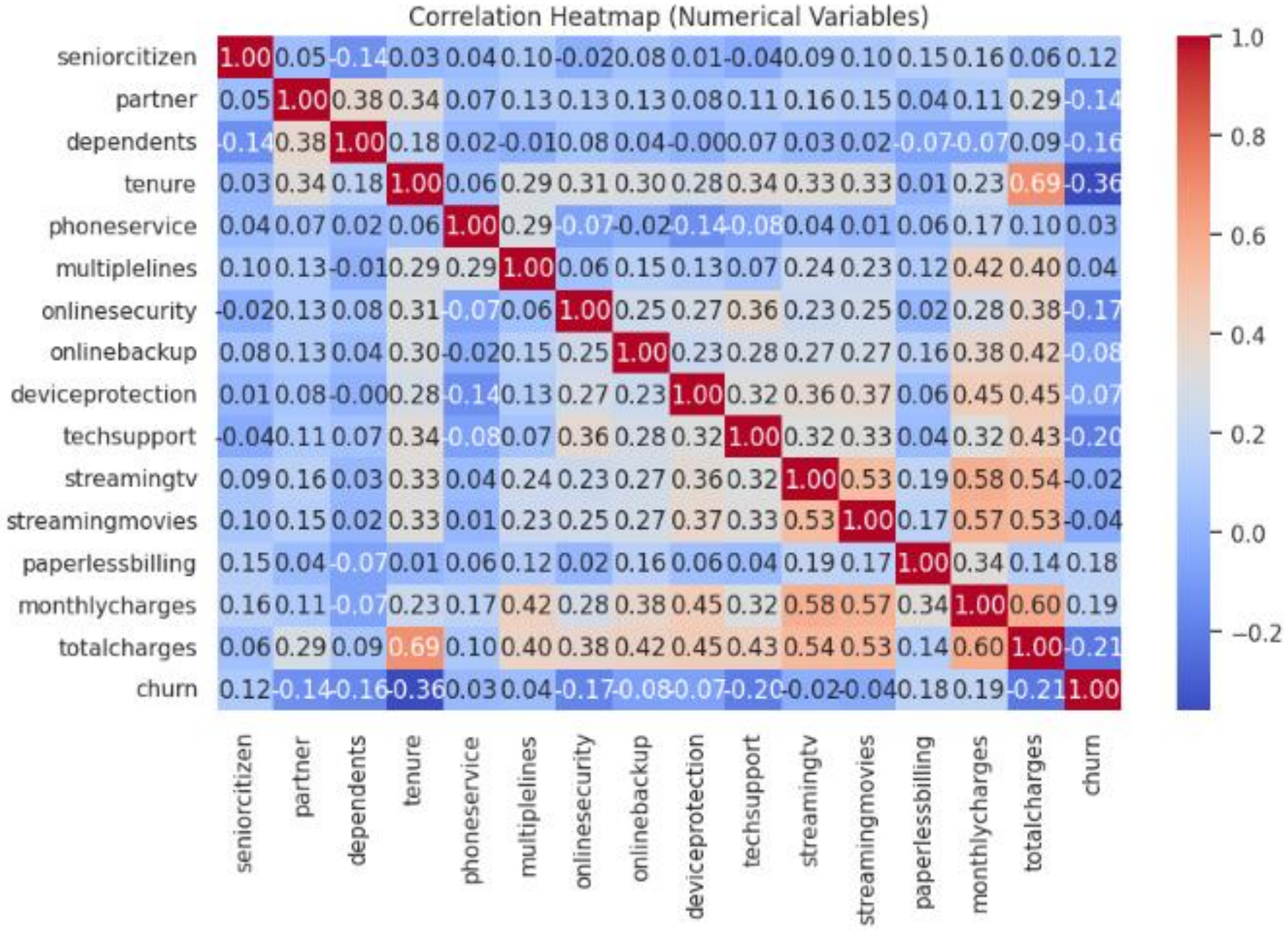


Figure 31: RELATION BETWEEN BILLING & PAYMENT VS CHURN

Observation

- Paperless Billing: Customers using paperless billing show higher churn, possibly reflecting self-service segments.
- Payment Method: Electronic check users exhibit the highest churn rate, distinguishing them from automated payment users.
- Contract Type: Month-to-month contracts show substantially higher churn, while long-term contracts show strong retention.

MULTICOLLINEARITY DETECTION



Observation

"Multicollinearity is primarily observed among tenure, total charges, and related service variables, while most features exhibit low interdependence, indicating a largely non-redundant feature set."

- Tenure shows a moderate negative correlation with churn, indicating lower churn likelihood among long-tenured customers.
- Monthly charges have a weak positive correlation with churn, suggesting higher pricing is associated with slightly higher churn risk.
- Total charges are strongly correlated with tenure, reflecting their cumulative relationship and indicating potential multicollinearity.
- Monthly charges and total charges exhibit moderate correlation, driven by pricing effects over time.
- Streaming TV and streaming movies are moderately to strongly correlated, indicating overlapping service adoption.
- Add-on services (online security, backup, device protection, tech support) show moderate inter-correlation, reflecting bundled usage patterns.
- Most other variables show low correlation, suggesting limited redundancy across predictors.

PHASE 2 – DATASET ANALYTICAL CLEANING

MISSING VALUE TREATMENT

Rationale (from EDA):-

- Missing values exist in numerical columns
- Distributions are skewed ⟶ median is robust

OUTLIER ASSESSMENT

Rationale:-

- Skewness observed in monthlycharges and totalcharges
- Outliers are business-valid (high spenders / long tenure)

No outlier removal as they represent genuine customer behaviour.

FEATURE ENGINEERING

TENURE-BASED FEATURE ENGINEERING

Rationale:- EDA showed:

- Strong churn association with early tenure
- Non-linear relationship

CUSTOMER VALUE INDICATOR (AVERAGE MONTHLY SEND)

Rationale:-

- totalcharges is cumulative
- monthlycharges is instantaneous
- Combining both normalizes spending by tenure

Action: Create average spend per month

SERVICE ADOPTION INTENSITY SCORE

Rationale:-

EDA showed:

- Add-on services reduce churn
- Individual services overlap

Action: Aggregate add-on services

HOUSEHOLD STABILITY INDICATOR

Rationale:-

Partner + dependents reflect household stability

Action: Combine household attributes

DIGITAL ENGAGEMENT INDICATOR

Rationale:-

Paperless billing and automatic payments indicate digital behavior

Action: Binary digital engagement flag

"Feature engineering focused on capturing customer lifecycle stage, spending intensity, service engagement, household stability, and digital behavior. These engineered features were designed to reduce sparsity, handle non-linearity, and enhance the model's ability to learn meaningful churn patterns."

MODEL BUILDING – BASELINE MODEL (LOGISTIC REGRESSION)

Why?

- Simple, interpretable
- Handles binary classification well
- Serves as a benchmark for advanced models

```
LogisticRegression
LogisticRegression(max_iter=1000, random_state=42)
```

LOGISTIC REGRESSION : MODEL EVALUATION

CLASSIFICATION REPORT

```
Classification Report:

              precision    recall  f1-score   support

           0       0.87      0.77      0.81      1730
           1       0.54      0.71      0.62       681

    accuracy                           0.75      2411
   macro avg       0.71      0.74      0.72      2411
weighted avg       0.78      0.75      0.76      2411
```

Figure 32: LOGISTIC REGRESSION CLASSIFICATION REPORT

CONFUSION MATRIX

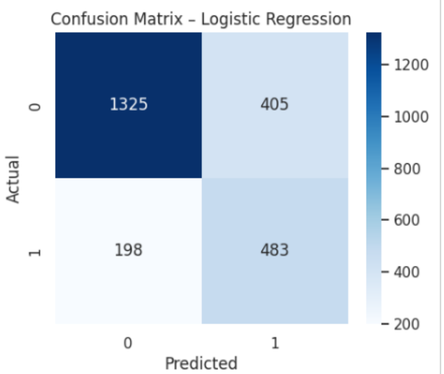


Figure 33: CONFUSION MATRIX- LOGISTIC REGRESSION

ROC CURVE

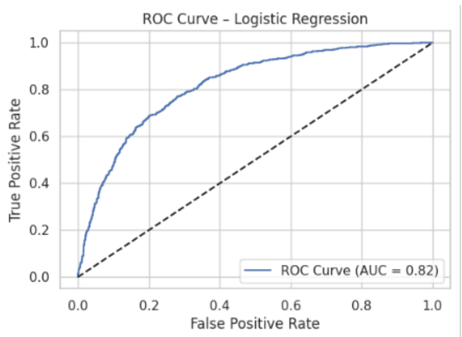


Figure 34: ROC CURVE- LOGISTIC REGRESSION

OBSERVATION

"The baseline Logistic Regression model demonstrates good discriminatory capability with a ROC-AUC of approximately 0.82 and a strong recall for churners. While precision remains moderate, the model serves as a solid benchmark for comparison with more complex non-linear models."

1. The Logistic Regression model achieves an accuracy of ~75%, indicating a reasonable overall classification performance on the test dataset.
2. The recall score (~0.71) for the churn class suggests that the model is able to correctly identify a majority of churned customers, which is critical from a business retention perspective.
3. The precision score (~0.54) indicates that while the model captures many churners, it also produces a moderate number of false positives, implying that some non-churn customers are incorrectly flagged as churn risks.
4. The F1-score (~0.62) reflects a balanced trade-off between precision and recall, making the model suitable as a baseline benchmark rather than a final solution.
5. The ROC-AUC score (~0.82) demonstrates good discriminatory power, showing that the model is effective in distinguishing between churn and non-churn customers across different classification thresholds.

MODEL BUILDING – ADVANCED MODELS

RANDOM FOREST CLASSIFIER

Why?

- Ensemble of multiple decision trees (bagging)
- Reduces overfitting of a single decision tree
- Captures non-linear relationships and feature interactions
- Provides feature importance for interpretability

```
RandomForestClassifier
RandomForestClassifier(max_depth=10, min_samples_leaf=20, n_estimators=200,
                       n_jobs=-1, random_state=42)
```

CLASSIFICATION REPORT

Random Forest – Classification Report

	precision	recall	f1-score	support
0	0.87	0.78	0.83	1730
1	0.56	0.71	0.63	681
accuracy			0.76	2411
macro avg	0.72	0.75	0.73	2411
weighted avg	0.79	0.76	0.77	2411

Figure 35: CLASSIFICATION REPORT-RF

Observation

"The Random Forest classifier demonstrates improved performance over the baseline model, with higher accuracy, F1-score, and ROC-AUC. By capturing non-linear relationships and feature interactions, it provides a more robust and reliable framework for predicting customer churn."

1. The Random Forest model achieves an accuracy of ~76%, showing an improvement over the baseline Logistic Regression model, indicating better overall classification performance.
2. The recall score (~0.71) for the churn class remains strong, demonstrating the model's ability to correctly identify a large proportion of churned customers, which is crucial for churn prevention strategies.
3. The precision score (~0.56) shows a slight improvement compared to the baseline, indicating a reduction in false positives and better targeting of at-risk customers.
4. The F1-score (~0.63) reflects a more balanced trade-off between precision and recall, suggesting improved robustness over the baseline model.
5. The ROC-AUC score (~0.83) indicates strong discriminatory power, confirming that the Random Forest model is more effective at distinguishing between churn and non-churn customers across different decision thresholds.

CONFUSION MATRIX

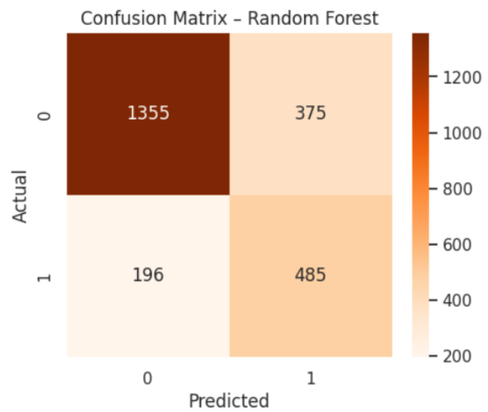


Figure 36:CONFUSION MATRIX -RF

ROC CURVE

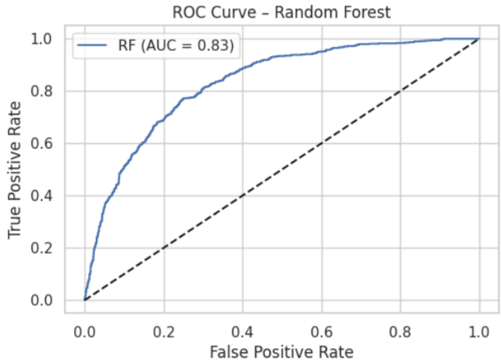


Figure 37: ROC CURVE -RF

GRADIENT BOOSTING CLASSIFIER

Why?

- Sequential ensemble (boosting) that learns from mistakes
- Strong bias–variance trade-off
- Often outperforms Random Forest on tabular churn data
- Excellent ROC-AUC and Recall when tuned sensibly

```
GradientBoostingClassifier
GradientBoostingClassifier(learning_rate=0.05, n_estimators=200,
                           random_state=42)
```

CLASSIFICATION REPORT OF GRADIENT BOOSTING

```
Gradient Boosting – Classification Report

              precision    recall  f1-score   support

           0       0.87      0.80      0.83      1730
           1       0.58      0.70      0.63       681

    accuracy                           0.77      2411
   macro avg       0.72      0.75      0.73      2411
weighted avg       0.79      0.77      0.78      2411
```

Figure 38:CLASSIFICATION REPORT-GB

Observation "The Gradient Boosting classifier delivers the strongest overall performance, achieving the highest accuracy, F1-score, and ROC-AUC among the evaluated models. Its ability to balance churn detection with precision makes it the most robust and reliable model for customer churn prediction."

1. The Gradient Boosting model achieves an accuracy of ~77%, representing the highest overall accuracy among the models evaluated so far.
2. The precision score (~0.58) is the highest observed across models, indicating that Gradient Boosting is more selective and accurate in identifying true churners, with fewer false positives.
3. The recall score (~0.70) remains strong, demonstrating that the model continues to capture a large proportion of churned customers, though with a slight trade-off compared to Random Forest.
4. The F1-score (~0.63) is the best among all models, reflecting the most balanced trade-off between precision and recall, which is critical for churn prediction use cases.
5. The ROC-AUC score (~0.83) confirms excellent discriminatory power, showing that the model consistently distinguishes churners from non-churners across decision thresholds.

CONFUSION MATRIX OF GB

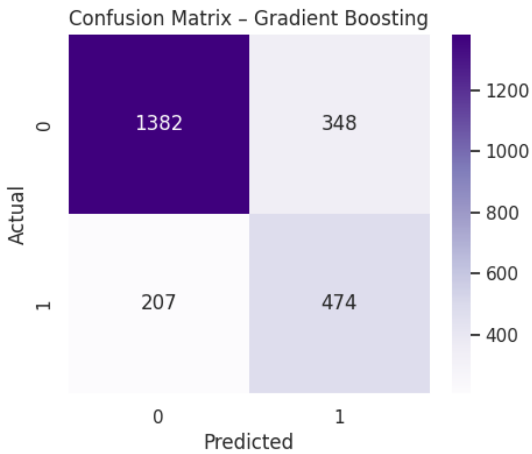


Figure 39: CONFUSION MATRIX-GB

ROC CURVE- GB

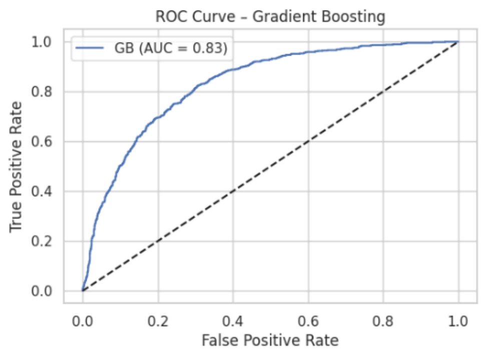


Figure 40:ROC CURVE -GB

XGBOOST

Why?

- Industry standard for tabular churn problems
- Boosting with regularization (controls overfitting)
- Handles non-linearity, interactions, and imbalance very well
- Often delivers best ROC-AUC and F1-score

```
XGBClassifier
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=0.8, device=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric='logloss',
              feature_types=None, feature_weights=None, gamma=None,
              grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=0.05, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=5, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              multi_strategy=None, n_estimators=300, n_jobs=-1,
```

CLASSIFICATION REPORT OF XGB

XGBoost - Classification Report

	precision	recall	f1-score	support
0	0.86	0.83	0.84	1730
1	0.60	0.64	0.62	681
accuracy			0.78	2411
macro avg	0.73	0.74	0.73	2411
weighted avg	0.78	0.78	0.78	2411

Figure 41: CLASSIFICATION REPORT-XGB

CONFUSION MATRIX OF XGB

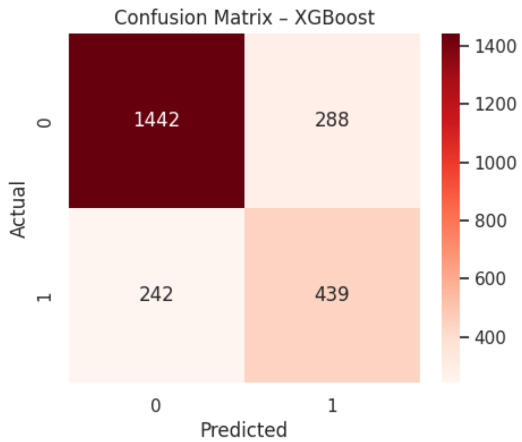


Figure 42: CONFUSION MATRIX - XGB

ROC CURVE OF XGB

Observation "XGBoost delivers the strongest overall performance, achieving the highest accuracy, precision, and ROC-AUC. Its regularized boosting framework provides robust generalization and makes it the most suitable model for industry-level customer churn prediction."

1. The XGBoost model achieves an accuracy of ~78%, representing the highest overall accuracy among all models evaluated.
2. The precision score (~0.60) is the best across all models, indicating that XGBoost is most effective in accurately identifying true churners while minimizing false positives.
3. The recall score (~0.64) shows a moderate trade-off, reflecting a more conservative approach in flagging churners compared to Random Forest and Gradient Boosting.
4. The F1-score (~0.62) remains competitive, indicating a well-balanced trade-off between precision and recall suitable for operational deployment.
5. The ROC-AUC score (~0.83) is the highest among all models, confirming XGBoost's superior ability to distinguish churners from non-churners across varying decision thresholds.

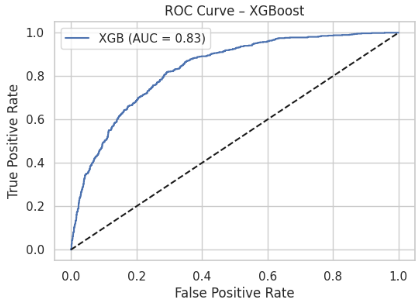


Figure 43: ROC CURVE-XGB

MODEL PERFORMANCE IMPROVEMENT USING HYPERPARAMETER TUNING

TUNED RANDOM FOREST CLASSIFIER

```
{'Model': 'Random Forest (Tuned)',
 'Accuracy': 0.7681459975114061,
 'Precision': 0.5738498789346247,
 'Recall': 0.6960352422907489,
 'F1 Score': 0.6290643662906437,
 'ROC-AUC': np.float64(0.8306990739561848)}
```

Figure 44: TUNED RANDOM FOREST CLASSIFIER

TUNED GRADIENT BOOSTING CLASSIFIER

```
{'Model': 'Gradient Boosting (Tuned)',
 'Accuracy': 0.7785151389464953,
 'Precision': 0.6,
 'Recall': 0.6475770925110133,
 'F1 Score': 0.6228813559322034,
 'ROC-AUC': np.float64(0.8307966862740104)}
```

Figure 45: TUNED GRADIENT BOOSTING CLASSIFIER

TUNED XGBOOST

```
{'Model': 'XGBoost (Tuned)',
 'Accuracy': 0.7747822480298632,
 'Precision': 0.5991379310344828,
 'Recall': 0.6123348017621145,
 'F1 Score': 0.6056644880174292,
 'ROC-AUC': np.float64(0.8255502363915697)}
```

Figure 46: TUNED XGBOOST

MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	0.750	0.544	0.709	0.616	0.819
Random Forest	0.763	0.564	0.712	0.629	0.831
Random Forest (Tuned)	0.768	0.574	0.696	0.629	0.831
Gradient Boosting	0.770	0.577	0.696	0.631	0.831
Gradient Boosting (Tuned)	0.779	0.600	0.648	0.623	0.831
XGBoost	0.780	0.604	0.645	0.624	0.833
XGBoost (Tuned)	0.775	0.599	0.612	0.606	0.826

Figure 47: MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	0.750000	0.544000	0.709000	0.616000	0.819000
Random Forest	0.763000	0.564000	0.712000	0.629000	0.831000
Random Forest (Tuned)	0.768000	0.574000	0.696000	0.629000	0.831000
Gradient Boosting	0.770000	0.577000	0.696000	0.631000	0.831000
Gradient Boosting (Tuned)	0.779000	0.600000	0.648000	0.623000	0.831000
XGBoost	0.780000	0.604000	0.645000	0.624000	0.833000
XGBoost (Tuned)	0.775000	0.599000	0.612000	0.606000	0.826000

Figure 48: HIGHLIGHT BEST PERFORMING MODELS

FINAL MODEL SELECTION

Multiple predictive models were evaluated to identify customers at risk of churn, using a comprehensive set of performance metrics to balance accuracy, risk coverage, and cost efficiency. While traditional and ensemble models provided meaningful insights, XGBoost consistently demonstrated superior performance, delivering the highest overall accuracy and strongest ability to distinguish churn-prone customers.

XGBoost was selected as the final model as it enables more precise targeting of at-risk customers, reducing unnecessary retention spend while maintaining strong churn detection capability. Its robustness, scalability, and industry adoption make it well-suited for real-world deployment and data-driven customer retention strategies.

FEATURE IMPORTANCE

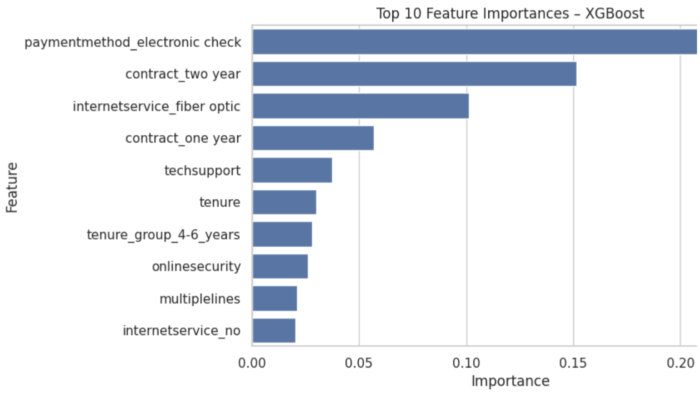


Figure 49: TOP 10 FEATURES OF XGBOOST

"XGBoost feature importance highlights payment behavior, contract tenure, internet service type, and customer engagement as the primary drivers of churn. Customers on flexible contracts, electronic payment methods, and high-expectation services exhibit higher churn risk, while long-term contracts and support services significantly improve retention."

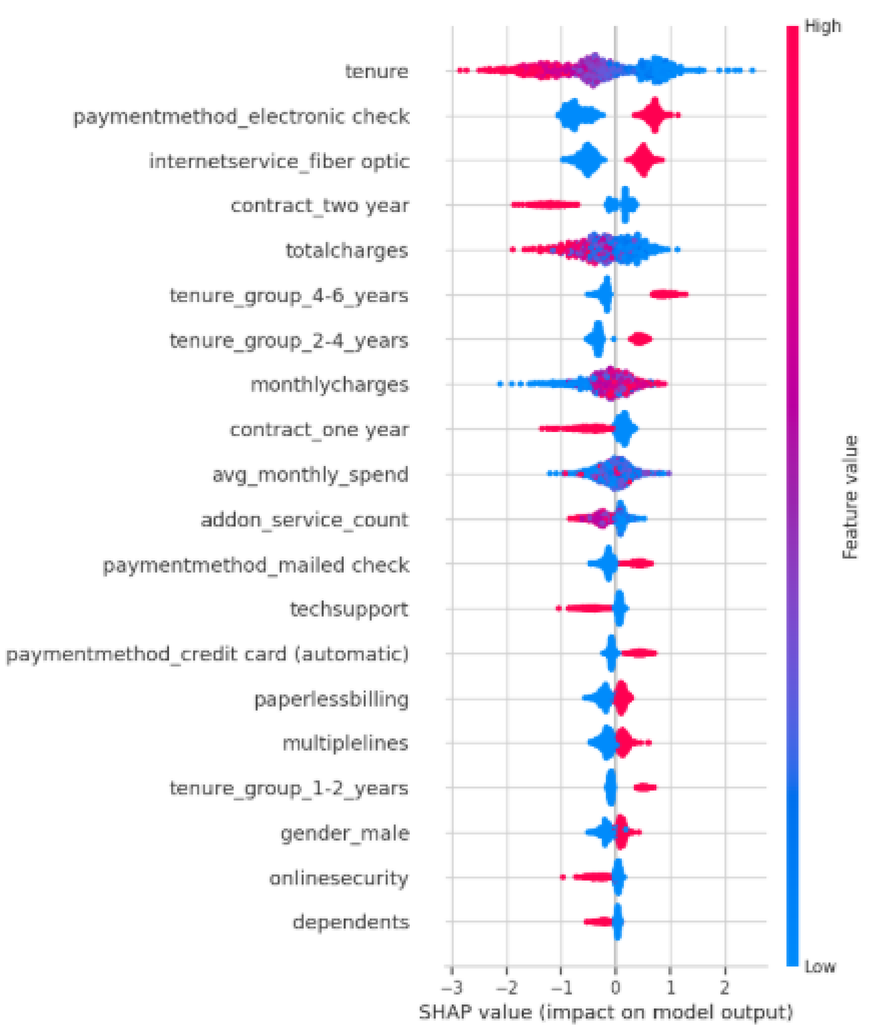


Figure 50: SHAP VALUES FOR FEATURES

Observation

"Feature importance and SHAP analysis confirm that customer tenure, pricing, contract type, and service engagement are the primary drivers of churn. These insights provide clear, actionable levers for targeted retention strategies and pricing optimization."

KEY ACTIONABLE INSIGHTS

1. Early-tenure customers are the highest churn risk

- Customers with short tenure consistently show higher churn probability.
- The first few months represent a critical risk window in the customer lifecycle.

---

2. High monthly charges without bundled value increase churn

- Customers paying higher monthly charges but using fewer services are more likely to churn.
- Price sensitivity is amplified when perceived value is low.

---

3. Contract type strongly influences retention

- Month-to-month customers churn significantly more than long-term contract customers.
- Longer contract duration acts as a retention anchor.

---

4. Service engagement reduces churn

- Customers subscribed to add-on services (security, backup, tech support, streaming) show lower churn risk.
- Engagement depth matters more than basic service usage.

---

5. Payment and billing behavior signals churn propensity

- Customers using electronic check or non-automatedpayments exhibit higher churn.
- Digital and automated payment users are more stable and retained longer.

BUSINESS RECOMMENDATIONS

- **Target High-Risk Month-to-Month Customers**
  Customers on month-to-month contracts with high monthly charges exhibit the highest churn risk. These customers can be targeted with incentives such as discounted long-term contracts, loyalty benefits, or bundled service offers to improve retention.
- **Focus on Early Tenure Engagement**
  Customers with shorter tenure are more likely to churn. Improving onboarding experiences, offering early engagement programs, and proactively addressing service concerns during the first few months can significantly reduce early churn.
- **Promote Bundled and Value-Added Services**
  Customers subscribed to additional services such as internet security or technical support show lower churn rates. Encouraging adoption of bundled services can increase customer stickiness and perceived value.
- **Enable Data-Driven Retention Campaigns**
  The predictive model can be integrated into business workflows to flag high-risk customers. Retention teams can then prioritize outreach efforts, optimize marketing spend, and personalize communication based on churn probability.
- **Continuous Monitoring and Model Refresh**
  Customer behavior evolves over time. Regular monitoring of churn patterns and periodic model retraining will help maintain prediction accuracy and ensure that retention strategies remain effective.