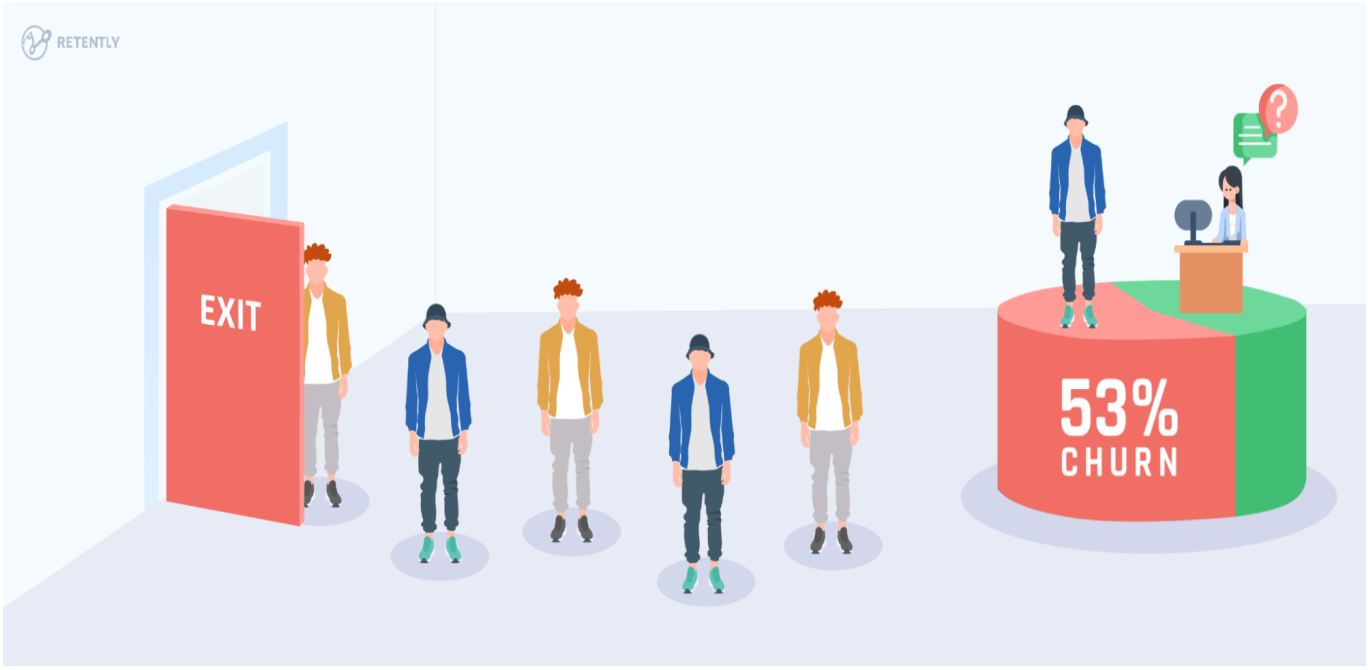
Customer Churn Prediction



— Ruchi Rikta

Contents

List of Figures

Defining the Problem Statement

Customer churn, when a customer stops using a company's service, is a critical issue for telecom companies operating in highly competitive markets.
Acquiring a new customer is 5–7 times more expensive than retaining an existing one. Therefore, predicting churn in advance can help the company take proactive steps (e.g., personalized offers, improved service quality) to maintain customers.

Problem Statement:

The goal of this project is to develop a data-driven model that predicts customer churn in a telecom company and identifies the key factors influencing customer retention. This will enable the company to implement targeted retention strategies, reduce revenue loss, and improve customer loyalty.

Need for the Study / Project

- Customer churn is a major challenge in the telecom industry, where intense competition and similar service offerings make it difficult to retain customers. Acquiring a new customer costs five to six times more than retaining an existing one, and the success rate of selling to existing customers (60–70%) is far higher than to new ones (5–20%). This makes retention a far more cost-effective strategy for sustaining profitability.
- Acquiring new customers is significantly more expensive than retaining existing ones. Studies indicate that the cost of acquiring a new customer is 5 to 6 times higher than that of retaining an existing customer. Furthermore, the success rate of selling new products or services to existing customers is between 60% and 70%, compared to only 5% to 20% for new customers. This highlights the importance of focusing on retention strategies — not only to reduce acquisition costs but also to maximize the lifetime value of current customers.
- Customer churn is not just about the loss of a single customer — it directly impacts the bottom line, profits, and company reputation. According to studies, increasing customer retention by just 5% can boost profits by 25% to 95%. Conversely, U.S. companies collectively lose an estimated $136.8 billion annually due to preventable customer churn. Therefore, developing a predictive model to understand the factors influencing churn is not only a data science exercise but a strategic business initiative. The insights gained from this project will enable the company to improve retention, reduce acquisition costs, and strengthen its competitive advantage in the telecom industry.

Understanding the Business / Social Opportunity

Business Opportunity:

Reducing churn by even a few percentage points can lead to substantial revenue growth. For instance:

- Retaining just 5% more customers can increase profits by 25–95%, according to various industry studies.
- The insights from churn prediction can guide the design of loyalty programs, customer experience improvements, and pricing strategies.

Social Opportunity:

By improving service quality and customer satisfaction, telecom providers contribute to better communication access, connectivity, and digital inclusion — all of which have broader social benefits, especially in emerging markets.

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.
- **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 impact both customer churn and overall 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.

Data Overview

Methodology

Missing values in numerical columns were imputed by filling “0”. Duplicate records were checked. The target variable Churn was converted to binary (1 = Churn, 0 = No Churn). All the Charges were converted into USD. Converted negative values in ‘tenure’ and ‘TotalCharges’ to positive values, assuming it is a typo.

- **1st 5 rows of the Dataset**

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection
0	Female	0	Yes	No	1.0	No	No phone service	DSL	No	Yes	No
1	Male	0	No	No	34.0	Yes	No	DSL	Yes	No	Yes
2	Male	0	No	No	2.0	Yes	No	DSL	Yes	Yes	No
3	Male	0	No	No	45.0	No	No phone service	DSL	Yes	No	Yes
4	Female	0	No	No	2.0	Yes	No	Fiber optic	No	No	No

TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
No	No	No	Month-to-month	Yes	Electronic check	$29.85	$29.85	No
No	No	No	One year	No	Mailed Check	$56.95	$1889.5	NO
No	No	No	Month-to-month	Yes	Mailed check	$53.85	$108.15	YES
Yes	No	No	One year	No	bank transfer (automatic)	$42.3	$1840.75	No
No	No	No	Month-to-month	Yes	ELECTRONIC CHECK	$70.7	$nan	yes

- **Statistical Summary**

	count	mean	std	min	25%	50%	75%	max
SeniorCitizen	12055.0	0.117959	0.322573	0.00	0.0000	0.00	0.0000	1.00
tenure	11451.0	31.237796	25.027111	-3.00	6.0000	28.00	54.0000	74.00
MonthlyCharges	11754.0	64.366212	30.332938	15.29	30.3125	71.35	89.3775	121.67
TotalCharges	10850.0	2291.757286	2277.461224	-197.00	383.6500	1328.75	3959.0750	9039.92

- **Null values**

  Null values were found in “tenure”, “MonthlyChargers”, and “TotalCharges” columns.

● Observations and Insights

- The dataset has 12055 rows and 20 columns.
- There are ‘MonthlyCharges’ and ‘TotalCharges’ columns showing as object type; converted them to float.
- There are 2 Duplicate values that are not exactly duplicates.
- There are null values in the ‘tenure’, ‘MonthlyCharges’, and ‘TotalCharges’ columns.
- There are unique values in the ‘Churn’ and ‘PaymentMethod’ columns.
- Prices in ‘MonthlyCharges’ and ‘TotalCharges’ are in USD as well as GBP in the dataset.
- There is a huge difference in the min and max of ‘TotalCharge’, which indicates outliers.
- Around 11.8% of customers are Senior Citizens, indicating most users are non-senior.
- Average tenure is ~31 months, with 75% of customers below 54 months of service.
- Tenure ranges from -3 to 74, where -3 is invalid and needs data cleaning.
- High standard deviation in TotalCharges (≈2277) shows a large variation in spending patterns.
- Median TotalCharges (1328.75) < Mean (2291.76) shows the distribution is right-skewed, with a few high-spending customers.
- Found Negative values in 'tenure' and 'TotalCharges' and converted them into positive values, assuming it was a typo.

Exploratory Data Analysis

Methodology

Exploratory Data Analysis was performed to understand customer demographics and behavior. Univariate analysis was conducted using histograms and count plots to explore the distribution of features like tenure, monthly charges, and payment method. Bivariate analysis was performed to study relationships between features and churn rate, such as churn vs. contract type and churn vs. internet service. Correlation analysis was used to identify multicollinearity among numerical variables.

1. Univariate Data Analysis

1.1. Churn Distribution

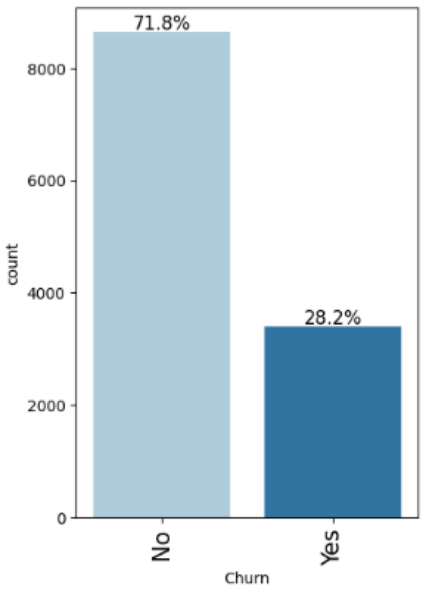


Fig.1.1: Univariate Analysis on Churn Distribution

- The dataset shows that 71.8% of customers did not churn, while 28.2% of customers churned.
- This indicates that the majority of customers remain with the company, whereas a smaller proportion discontinue the service.
- The dataset is moderately imbalanced, with the non-churned group being more than twice the size of the churned group.

1.2. Distribution of Tenure

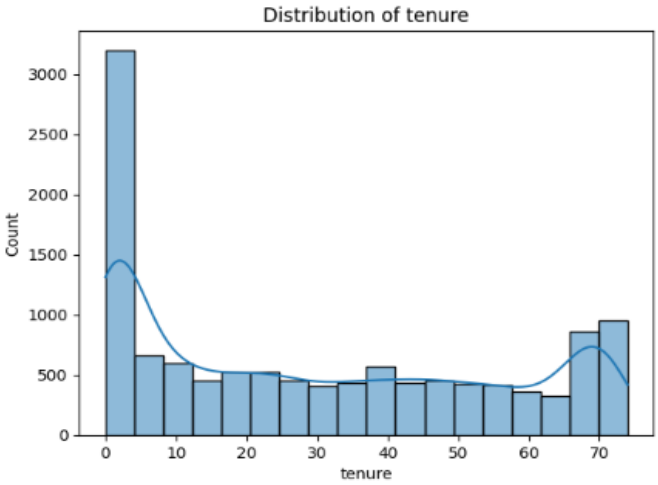


Fig.1.2: Univariate Analysis on Tenure

- The highest frequency occurs in the first bin (tenure 0-5), with a count of over 3000. This indicates that a large number of individuals have a very short tenure.
- After the initial peak, the frequency drops significantly. The count for the second bin (tenure 5-10) is less than 700
- The frequency of individuals with tenure between 5 and 65 is relatively low and consistent, with counts generally fluctuating between 400 and 600.
- The graph suggests a high turnover rate for individuals with short tenure, as well as a significant number of long-tenured individuals, with a smaller proportion in the middle range.

1.3. Distribution of Contract

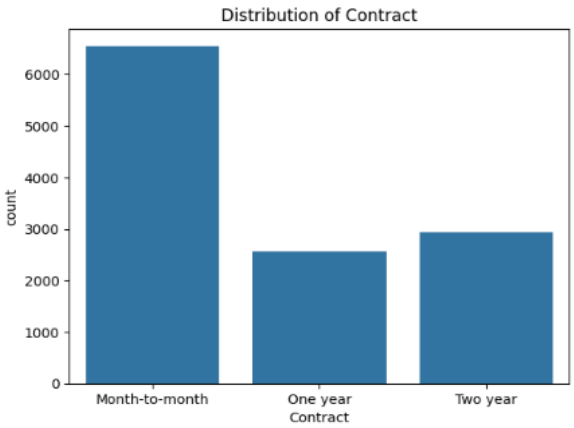


Fig.1.3: Univariate Analysis on Contract

- The majority of the contracts are month-to-month.
- The number of month-to-month contracts is the highest, with a count of more than 6,000.
- The number of one-year contracts is the lowest, with a count of approximately 2,800.
- The number of two-year contracts is slightly higher than the one-year contracts, with a count of approximately 3,000.
- The number of contracts decreases as the contract duration increases from month-to-month to one-year.

1.4. Distribution of Monthly Charges

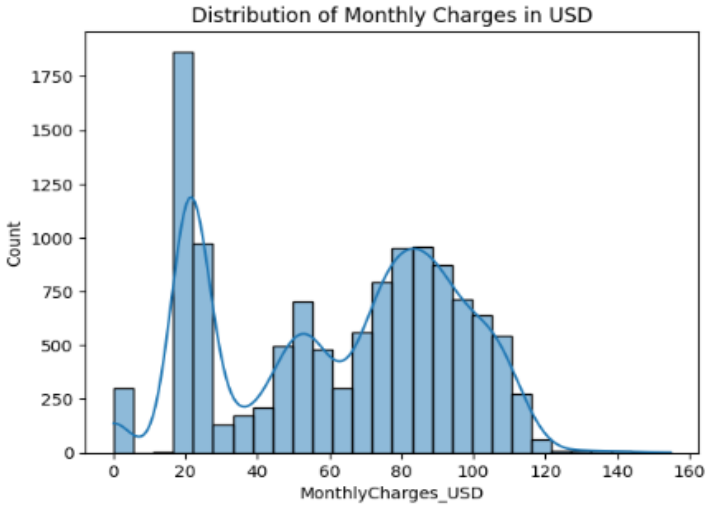


Fig.1.4: Univariate Analysis on Monthly Charges

- There are very few customers with monthly charges above $120.
- The first and most significant peak is centered around the $20-$30 range, with a count of nearly 1800. This indicates a large number of customers have monthly charges in this range.
- The second, broader peak is located in the $70-$110 range, suggesting another group of customers with higher monthly charges.
- The lowest monthly charges are between $0 and $10, with a count of approximately 300.
- The distribution is not symmetric and is skewed to the left, with a long tail extending towards higher monthly charges.

1.5. Distribution of Total Charges

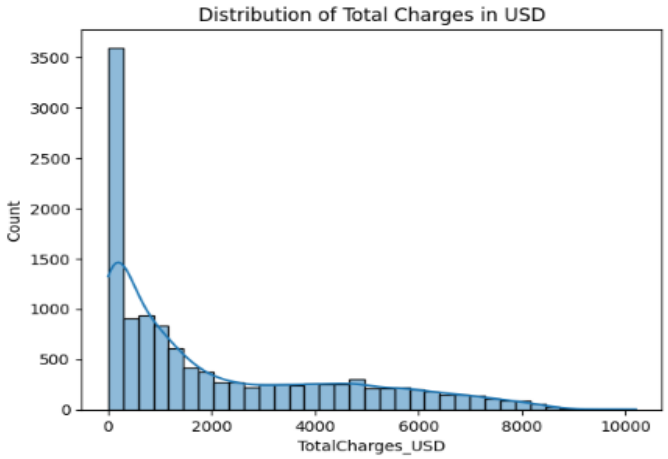


Fig.1.5: Univariate Analysis on Total Charges

- Right-skewed distributions were noted, indicating a need for scaling before modeling.
- The majority of total charges are concentrated in the lower range, specifically between $0 and $1000.
- As the total charge amount increases, the frequency (count) of occurrences decreases significantly.
- There are very few instances of high total charges, with the count dropping off sharply after $2000.

For continuous variables like Tenure, MonthlyCharges, and TotalCharges, the mean and median were compared across churned and non-churned groups to identify separation, while distributions were checked for skewness that could impact model performance. The Graph is plotted in the code file.

For categorical variables such as Contract, InternetService, and PaymentMethod, bar charts were used to compare churn proportions across categories. Variables showing clear differences in churn rates were identified as potential predictors for modeling.

2. Bivariate Data Analysis

2.1. Correlation Between Numeric Columns

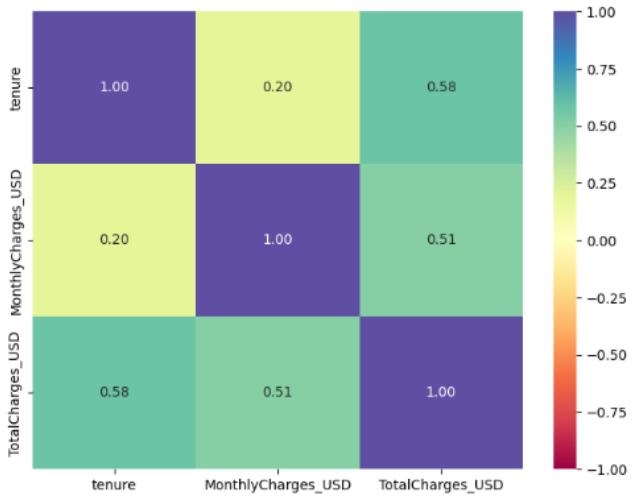


Fig.2.1: Correlation Between Numerical Values

- There is a strong positive correlation between tenure and TotalCharges_USD, with a correlation coefficient of 0.58.
  This suggests that as a customer's tenure increases, their total charges also tend to increase significantly.
- There is a moderately strong positive correlation between MonthlyCharges_USD and TotalCharges_USD, with a correlation coefficient of 0.51. This indicates that customers with higher monthly charges also tend to have higher total charges.
- There is a weak positive correlation between tenure and MonthlyCharges_USD, with a correlation coefficient of 0.20. This suggests a very slight tendency for monthly charges to increase with tenure, but the relationship is not strong.
- All three variables show a positive correlation with each other, with the strongest relationship being between tenure and TotalCharges_USD.

2.2. Churn Vs Contract

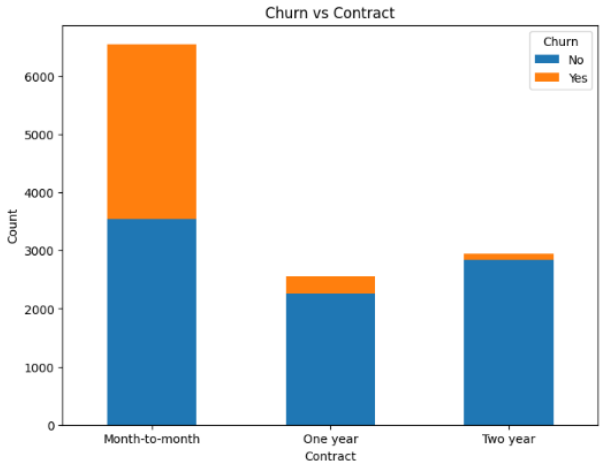


Fig.2.2: Bivariate Analysis on Churn and Contract

- Customers with a two-year contract have the lowest churn rate, with a very small orange "Yes" segment compared to the blue "No" segment.
- The number of customers who did not churn ("No") is highest for month-to-month contracts, followed by two-year and then one-year contracts.
- The number of customers who churned ("Yes") is significantly higher for month-to-month contracts compared to one-year and two-year contracts combined.

2.3. Churn Rate Vs Tenure Group

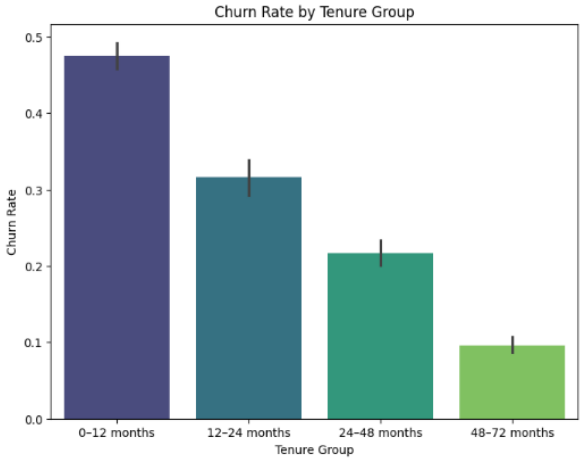


Fig.2.3: Bivariate Analysis on Churn Rate and Tenure Group

- The churn rate is highest for the "0-12 months" tenure group, with a rate of approximately 0.47.
- The churn rate decreases as the tenure of the customer increases.
- The "12-24 months" tenure group has the second-highest churn rate, at roughly 0.31.
- The lowest churn rate is observed in the "48-72 months" tenure group, at about 0.10.

2.4. Churn Rate Vs Monthly Charges

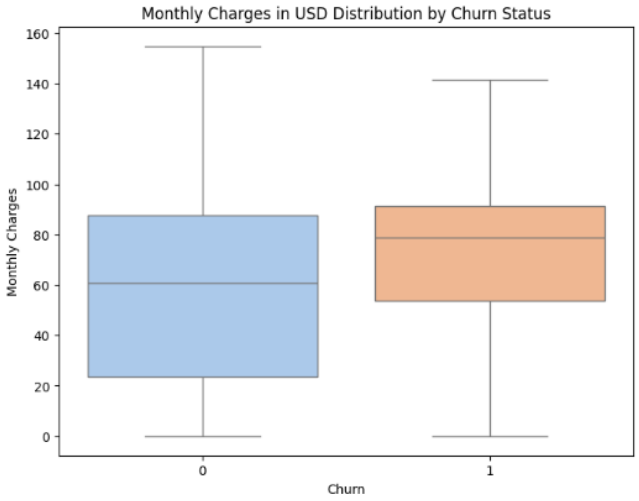


Fig.2.4: Bivariate Analysis on Churn Rate and Monthly Charges

- The median monthly charge for customers who did not churn (Churn = 0) is approximately $60.
- The median monthly charge for customers who did churn (Churn = 1) is approximately $75.
- The distribution of monthly charges for customers who did not churn is roughly symmetric, with the median line close to the center of the box.

3. Multivariate Data Analysis

3.1. Monthly Charges vs Tenure by Churn

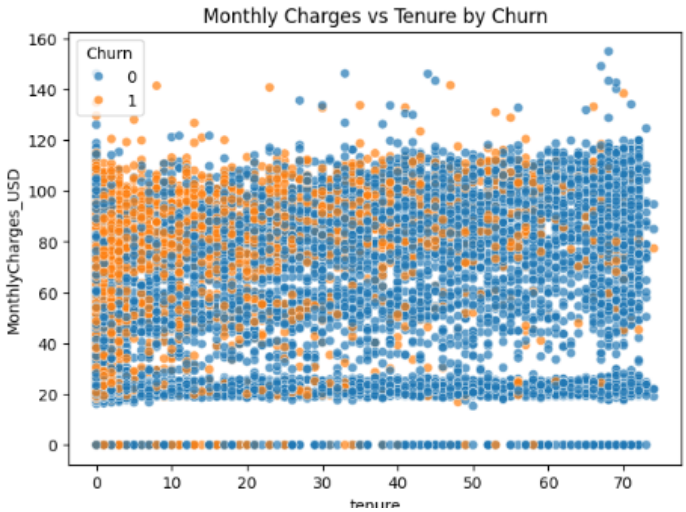


Fig.3.1: Multiivariate Analysis on Monthly Charges vs Tenure by Churn

- A significant number of orange dots are concentrated at the lower end of the "Tenure" axis (0 to 10 months). This indicates that a large portion of customer churn occurs early in their subscription period.
- There is a noticeable cluster of orange dots in the upper range of "MonthlyCharges_USD" (above $100). This suggests that customers with high monthly charges are more likely to churn.
- While most churn occurs early, there are still some orange dots present at higher tenures (e.g., 50-70 months). This suggests that some customers with long-term subscriptions also churn, but they are fewer in number compared to new customers.
- The graph shows that churn is less common among customers with low monthly charges, as there are fewer orange dots in the lower range of the "MonthlyCharges_USD" axis.

The complete detailed EDA for all variables is included in the code file.

Data Processing

Methodology

Outliers in the dataset were identified and addressed. Categorical variables were transformed using one-hot encoding, and the dataset was subsequently split into training and testing sets for model development and evaluation.

- **Outlier Detection**

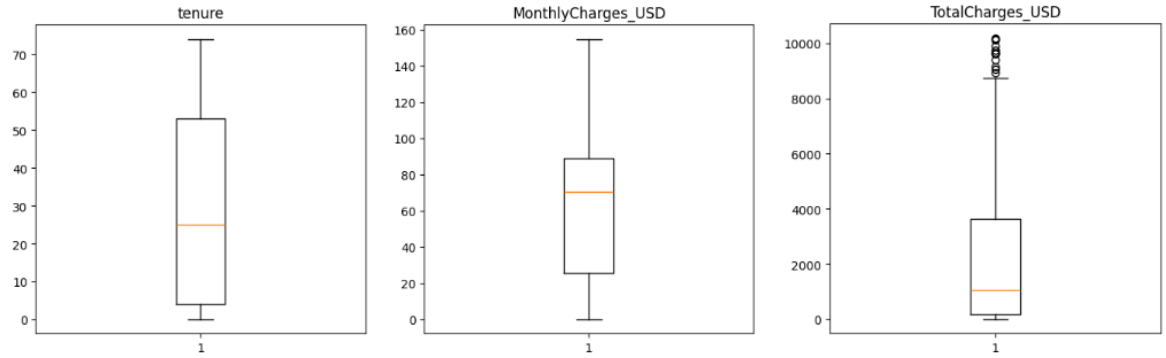


Fig.4: Outlier Detection

- Although outliers are there in 'TotalCharges_USD', we will keep them as they may contain some valuable input.

- **Feature Engineering**

  Applied one-hot encoding to the categorical columns to create dummy variables, enabling clearer representation and analysis of each category.

Data Preparation for Modeling

- **Categorical Encoding:** Used pd.get_dummies() to convert categorical variables into a numerical format suitable for machine learning models.
- **Data Splitting:** Divided the dataset into training (70%) and testing (30%) sets, containing 8,438 and 3,617 samples, respectively, with 31 features each.
- **Class Distribution:** The target variable Churn is imbalanced, with around 72% non-churn (class 0) in both sets, indicating a consistent split.
- **Feature Scaling:** Applied standardization (Z-score scaling) to continuous variables (Tenure, MonthlyCharges, TotalCharges) to ensure equal feature contribution and improve model performance.

Baseline Model Building

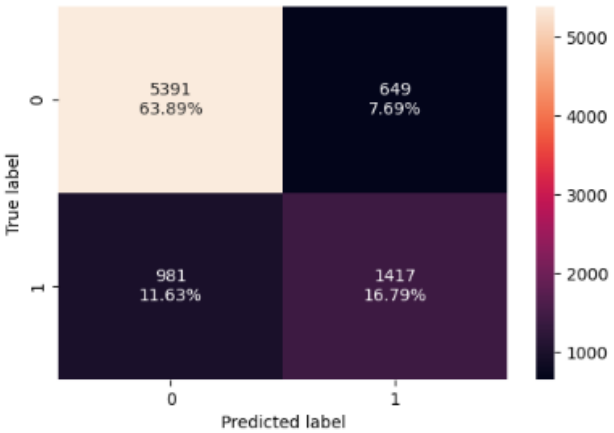
Methodology

Logistic Regression was selected as the baseline model because it provides interpretability and a solid linear benchmark for binary classification problems. The model was trained on standardized numerical data and dummy-encoded categorical features. Model performance was evaluated using Accuracy, Precision, Recall, and F1-score. Recall chosen as the primary metric since it focuses on minimizing the False Negatives.

Observations on baseline Model:

- The model was run on 8438 observations. The Pseudo R-squared is 0.2989, and the LLR p-value is 0.000, suggesting the model is statistically significant.
- **Negative Impact on Churn:** Longer tenure, higher total charges, having a partner, having dependents, using online security, using online backup, using tech support, having a one-year or two-year contract, and using a credit card for payment are all associated with a lower likelihood of churn.
- **Positive Impact on Churn:** Having multiple lines, using fiber optic internet service, and using an electronic check for payment are associated with a higher likelihood of churn.
- p-value of a variable indicates if the variable is significant or not. If we consider the significance level to be 0.05 (5%), then any variable with a p-value less than 0.05 would be considered significant. However, these variables might contain multicollinearity, which will affect the p-values.
- But these variables might contain multicollinearity, which will affect the p-values.
- We have to remove multicollinearity from the data to get reliable coefficients and p-values.

- **Confusion Matrix**



Training performance:

	Accuracy	Recall	Precision	F1
0	0.807	0.591	0.686	0.635

Observations:

- The model has an overall accuracy of approximately 80.7%, indicating that it correctly classified about four out of every five instances.
- The recall for class 1 is approximately 59.1%, which means the model only correctly identified about 59% of all actual class 1 instances.
- The precision for class 1 is approximately 68.6%, meaning that when the model predicted class 1, it was correct about 68.6% of the time.
- The F1 score, which is the harmonic mean of precision and recall, is approximately 63.5%. This metric is often more reliable than accuracy for imbalanced datasets and shows a moderate performance for the positive class.
- The dataset appears to be imbalanced, as there are significantly more instances of class 0 (5391 TN + 649 FP = 6040 total) than class 1 (981 FN + 1417 TP = 2398 total).

● Checking Multicollinearity

- None of the numerical variables shows multicollinearity.
- If VIF is equal to or exceeds 10, it shows signs of high multicollinearity.

● After treating multicollinearity

Training performance:

	Accuracy	Recall	Precision	F1
0	0.807	0.589	0.687	0.634

● Dropping high p-value variables

- Some of the dummy variables have a p-value > 0.05. So, they are not significant, and we will drop them.
- But sometimes p-values change after dropping a variable. So, we'll not drop all variables at once
- Instead, we will do the following:
    - Build a model, check the p-values of the variables, and drop the column with the highest p-value
    - Create a new model without the dropped feature, check the p-values of the variables, and drop the column with the highest p-value.
    - Repeat the above two steps till there are no columns with p-value > 0.05

- **Variables with a high p-value were dropped.**

- ## Coefficient interpretations

- A negative coefficient, such as for tenure (-0.0202), TotalCharges_USD (-9.851e-05), and Contract_Two year (-1.5392), suggests that as the value of the independent variable increases, the log-odds of a customer churning decrease. This implies that longer tenure, higher total charges, and having a two-year contract are associated with a lower probability of churning.
- A positive coefficient, such as for MultipleLines_Yes (0.2845), InternetService_Fiber optic (0.9166), and PaymentMethod_Electronic Check (0.4254), suggests that as the value of the independent variable increases, the log-odds of a customer churning increase. This implies that having multiple lines, using fiber optic internet, and paying with an electronic check are associated with a higher probability of churning.

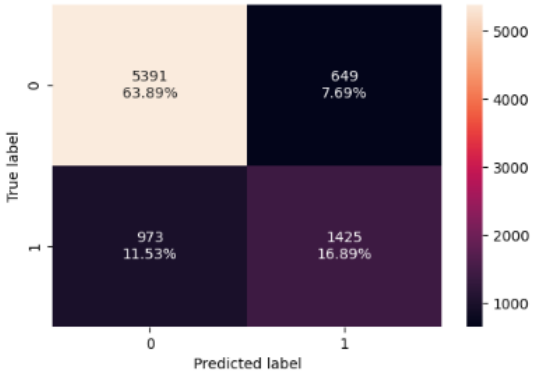
- **Converted coefficients to odds**

- ## Coefficient interpretations

- Positive Coefficients – Increased Churn Likelihood:
  Variables with positive coefficients indicate a higher likelihood of churn. Customers with Fiber Optic Internet, Electronic Check payment method, Multiple Lines, and Paperless Billing show increased churn probabilities. For instance, having fiber optic service increases the odds of churning by about 150%, while using electronic check payments raises it by 53%, having multiple lines by 33%, and opting for paperless billing by 35%.
- Negative Coefficients – Reduced Churn Likelihood:
  Variables with negative coefficients indicate a lower likelihood of churn. Customers with One-year or Two-year contracts, those without Internet Service, and those with Technical Support are less likely to leave. A two-year contract reduces churn odds by approximately 78.6%, a one-year contract by 56.3%, a lack of internet service by 65.3%, and having technical support by 42.8%.

Performance of New Model

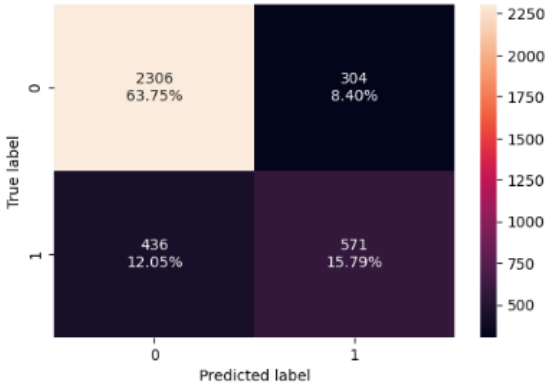
- **Training set performance**



Training performance:

	Accuracy	Recall	Precision	F1
0	0.808	0.594	0.687	0.637

- **Test set performance**



Test performance:

	Accuracy	Recall	Precision	F1
0	0.795	0.567	0.653	0.607

- **Observations and Insight**

- The performance is consistent across training and test datasets, confirming that the Logistic Regression model generalizes well and does not overfit.
- The slight drops in recall and F1-score on the test set are expected, as real-world data tends to be more variable than training data.
- Accuracy (~80%) and F1-score (~0.61) indicate that the model provides a strong baseline for predicting customer churn.
- Precision dropped slightly, meaning the model makes a few more false positive churn predictions on the test data. However, it still maintains good reliability in its churn predictions.
- The Recall was chosen as the primary metric because the business objective is to minimize False Negatives.

- The Recall on the test set (0.567) is close to the training score (0.594), indicating that the model generalizes well and does not exhibit significant overfitting.
- The small decrease in test Recall is expected and acceptable for a baseline logistic regression model, as some performance drop is typical when moving forward from training to unseen data.

- Overall, the model shows reasonable predictive power and can serve as a solid baseline before experimenting with more complex models (e.g., Random Forest, Gradient Boosting) to improve recall and overall churn detection.

Model Performance Improvement

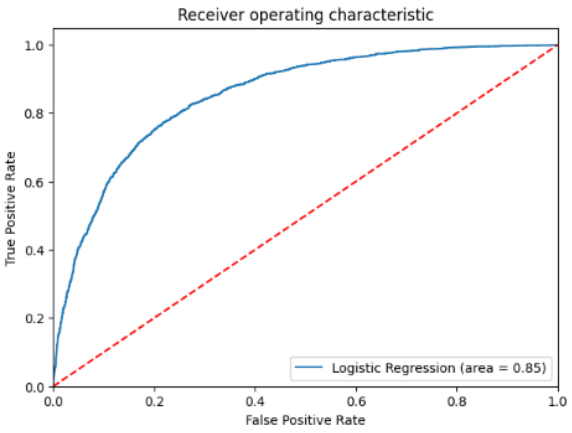


Fig.5: ROC-AUC(Training Set)

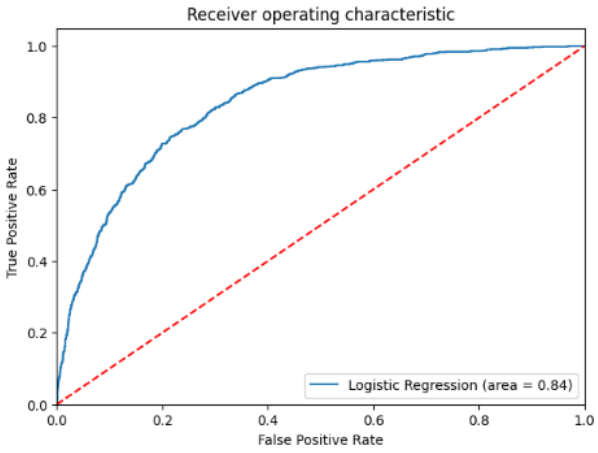
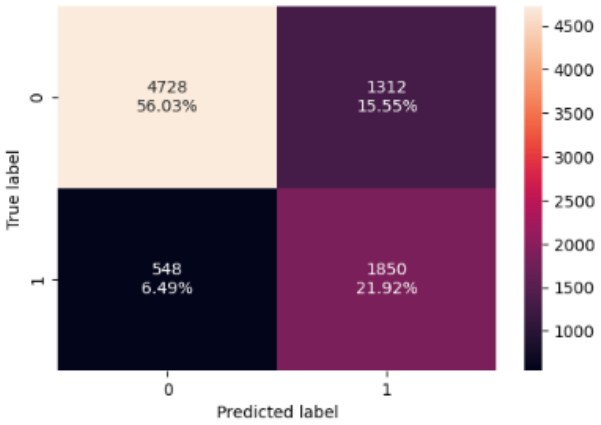


Fig.6: ROC-AUC(Test Set)

- ROC-AUC score of 0.85 on training is quite good.

Optimal threshold using AUC-ROC curve

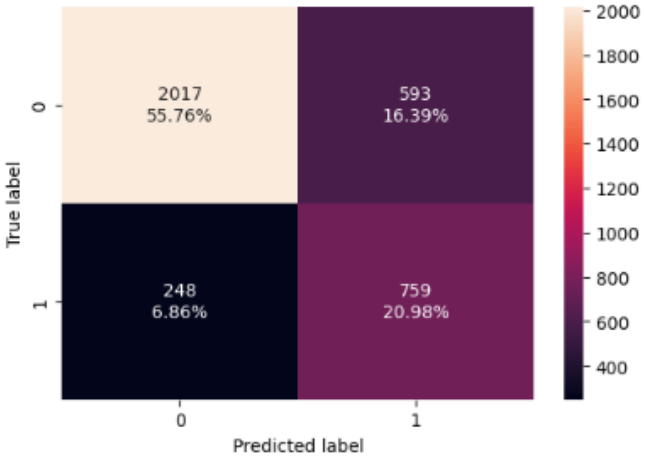
- **Checking model performance on training set**



Training performance:

	Accuracy	Recall	Precision	F1
0	0.780	0.771	0.585	0.665

- **Checking model performance on training set**



Test performance:

	Accuracy	Recall	Precision	F1
0	0.767	0.754	0.561	0.643

- Recall and F1 score of the model have increased, but the other metrics have reduced.
- The model is still giving a good performance.

Precision-Recall Curve

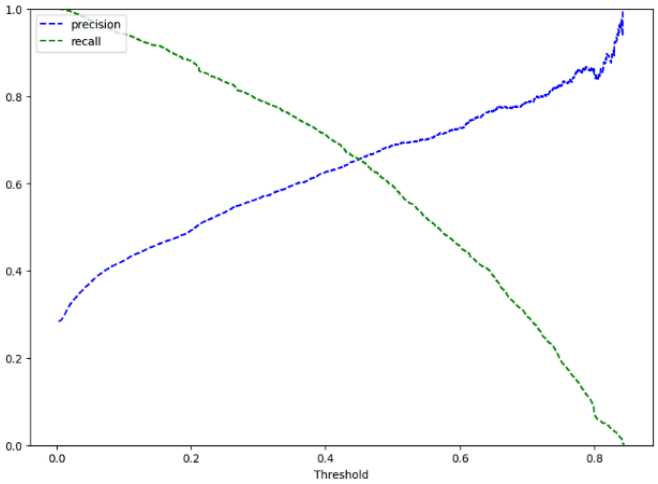
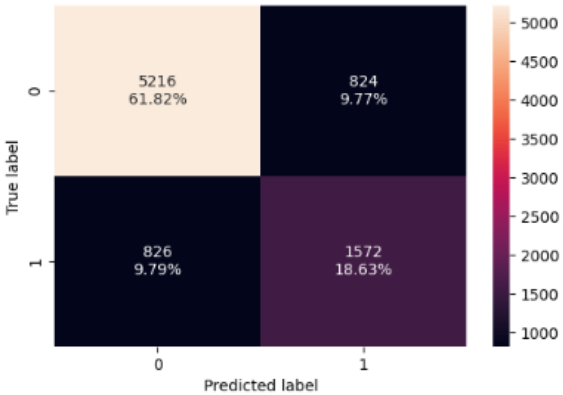


Fig.7: Precision-Recall Curve

- At a 0.45 threshold, we get a balanced precision and Recall.

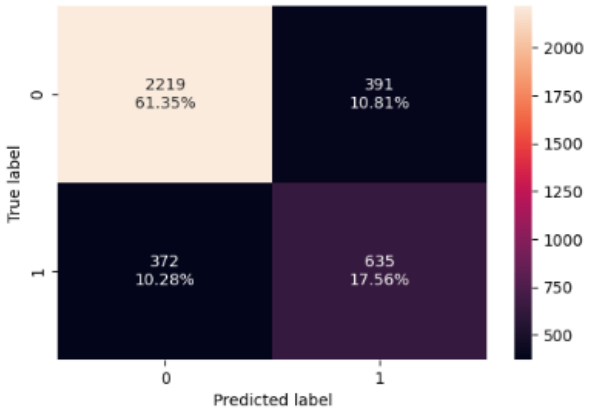
- **Checking model performance on training set - With 0.45 as Threshold.**



Training performance:

	Accuracy	Recall	Precision	F1
0	0.804	0.656	0.656	0.656

- **Checking model performance on test set - With 0.45 as Threshold**



Test performance:

	Accuracy	Recall	Precision	F1
0	0.789	0.631	0.619	0.625

- Recall and F1 score have reduced, but other metrics have increased in both the training and test sets.

Model performance threshold comparison summary

Training performance comparison:

	Logistic Regression-default Threshold (0.5)	Logistic Regression-0.33 Threshold	Logistic Regression-0.45 Threshold
Accuracy	0.808	0.780	0.804
Recall	0.594	0.771	0.656
Precision	0.687	0.585	0.656
F1	0.637	0.665	0.656

Observations:

- Lowering the threshold from 0.5 to 0.33 significantly increases Recall on both training (0.594 to 0.771) and test sets (0.567 to 0.754). This means the model becomes more sensitive, it catches more churners, reducing False Negatives.
- As Recall rises, Precision decreases (training: 0.637- 0.585, test: 0.653-0.561). The model classifies more customers as churners, including some non-churners (False Positives).
- At threshold = 0.45, the model achieves a good balance between Recall and Precision.
- Accuracy remains stable across thresholds (around 0.77–0.80) for both training and test sets, indicating no significant overfitting.
- Lowering the decision threshold improves Recall but slightly reduces Precision.

- By using the 0.33 threshold, the company now successfully identifies 75.4% of all customers who are going to churn. This gives the Retention Team a much larger pool of at-risk customers to target with intervention offers.

Model Building with Original Data

```
Cross-Validation performance on training dataset with original data:

Bagging: 0.5517595998487054
Random forest: 0.5941795884539951
GBM: 0.6161607488256934
Adaboost: 0.5965974815540285
Xgboost: 0.5978118551265321
dtree: 0.509623870389989

Validation Performance:

Bagging: 0.5520702634880803
Random forest: 0.5873697118332312
GBM: 0.6233292831105711
Adaboost: 0.6363091671659676
Xgboost: 0.597604790419617
dtree: 0.528281164195497
```

Insights

- **AdaBoost** achieved the highest validation Recall (0.636) among all models, indicating its superior ability to correctly identify churners in the original (imbalanced) dataset.
- **GBM** (0.623) also performed competitively, showing slightly lower Recall but still strong generalization compared to other ensemble methods.
- **Random Forest and Bagging** models achieved moderate Recall (~0.58), suggesting that while they capture some churn patterns, they are less sensitive to minority (churn) cases in the imbalanced data.
- **Decision Tree and XGBoost** models performed comparatively lower, indicating possible overfitting or difficulty in handling class imbalance without resampling.
- The difference between training and validation Recall is small across all models, showing stable generalization and no significant overfitting despite class imbalance.
- Overall, AdaBoost is the best-performing model on the original dataset, making it a strong baseline before applying resampling techniques (like SMOTE or undersampling) for further improvement.

Model Building with Oversampled Data

```
Cross-Validation performance on training dataset with oversampled data:

Bagging: 0.838366122646135
Random forest: 0.8525850440165579
GBM: 0.8563423833198556
Adaboost: 0.8485199936968678
Xgboost: 0.8464997663089004
dtree: 0.8082812970718404

Validation Performance:

Bagging: 0.5561694290976059
Random forest: 0.5959232613908872
GBM: 0.6414662084765178
Adaboost: 0.6656084656084656
Xgboost: 0.6
dtree: 0.5392781316348195
```

Insights

- **AdaBoost** achieved the highest validation Recall (0.683) among all models, showing the best generalization and ability to identify churners after applying oversampling.
- **GBM (Gradient Boosting)** followed closely with a validation Recall of 0.609, performing consistently across training and validation with a moderate gap (0.25).
- **Random Forest, Bagging, and Decision Tree** models achieved near-perfect Recall on training data (0.98–0.99) but saw a steep drop (0.49–0.54) on validation data, indicating significant overfitting.
- **XGBoost** also showed signs of overfitting, high training Recall (0.865), but a much lower validation Recall (0.551).
- Among all models, AdaBoost demonstrated the best trade-off between Recall and overfitting, as the difference between training and validation Recall (0.172) was the smallest.
- The overall performance improvement in Recall (compared to models on the original data) confirms that oversampling effectively helped balance the dataset, improving the model's ability to detect churners.

Model Building with Undersampled Data

```
Cross-Validation performance on training dataset with undersampled data:

Bagging: 0.7297977392570025
Random forest: 0.7615405414694457
GBM: 0.7797625923700584
Adaboost: 0.7758083170071262
Xgboost: 0.7508336804567382
dtree: 0.6846880682485577

Validation Performance:

Bagging: 0.6106280193236715
Random forest: 0.6418950301904319
GBM: 0.6534296028880866
Adaboost: 0.6551111111111111
Xgboost: 0.634849167041873
dtree: 0.5617582417582417
```

Insights

- **AdaBoost** achieved the highest validation Recall (0.655) among all models, followed closely by GBM (0.653) and Random Forest (0.642). This indicates these ensemble models perform well even when trained on reduced, balanced data.
- The training and validation **Recall** values are close for all models, showing good generalization and no overfitting, unlike models trained on oversampled data.
- **Decision Tree** performed the worst (validation Recall 0.561), confirming that single weak learners are less effective compared to ensemble methods in handling class imbalance.
- **GBM and AdaBoost** show the best balance between Recall and model stability; they perform consistently across training and validation datasets.
- Compared to the original data models, undersampling improved Recall scores across all models, confirming that balancing the dataset helped the models detect churners more effectively.

Hyperparameter Tuning

Methodology

Hyperparameter tuning was performed to optimize model performance and prevent overfitting. To identify the best combination of hyperparameters, a Randomized Search Cross-Validation (RandomizedSearchCV) approach was employed using scikit-learn. This technique randomly samples a defined number of parameter combinations from the specified search space and evaluates model performance using cross-validation to ensure robustness and efficiency.

Models that might perform better after tuning:-

Based on the goal of maximizing Recall and minimizing False Negatives, we must choose the models that show the highest Recall on the Validation Set, as this is the best indicator of real-world performance.
Here are the models chosen for hyperparameter tuning and why, based on the validation performance.

- AdaBoost with Undersampled data
- Gradient Boost with Undersampled data
- AdaBoost with Original Data
- Random Forest with Undersampled Data

1. Tuning AdaBoost with Undersampled data

Reasoning - AdaBoost consistently gave the highest Recall (0.65–0.67) across both original and undersampled datasets.
It generalized well on both training and validation, Recall scores were close (small difference), indicating low overfitting.

Performance on training set

	Accuracy	Recall	Precision	F1
0	0.768	0.846	0.731	0.785

Performance on validation set

	Accuracy	Recall	Precision	F1
0	0.730	0.826	0.514	0.634

Observations:

- **Overall Model Performance:**
  The AdaBoost model performed reasonably well, achieving training accuracy of 0.768 and validation accuracy of 0.730, indicating good generalization with only a minor drop between the two sets.
- **Recall:**
  The model maintained a high recall on both training (0.846) and validation (0.826) sets, showing a strong ability to correctly identify customers who are likely to churn. This is beneficial for customer retention strategies.
- **Precision:**
  Precision decreased from 0.731 (training) to 0.514 (validation), suggesting that while the model identifies churners effectively, it also produces some false positives, predicting churn for customers who actually stay.
- **F1-Score:**
  The F1-score dropped from 0.785 to 0.634, showing moderate overfitting. The model still maintains a reasonable balance but can be improved by fine-tuning parameters or exploring resampling techniques like SMOTE.
- **Interpretation:**
  The model is recall-oriented, meaning it prioritizes identifying as many churners as possible, a desirable trait for churn prediction. However, the lower precision indicates a need for further refinement to reduce false positives without losing recall strength.

2. Tuning Gradient Boost with Undersampled data

Reasoning - GBM consistently showed strong Recall (0.61–0.65) and stable performance across all datasets (original, oversampled, undersampled). It has high potential for improvement through parameter tuning (learning rate, number of trees, depth, subsampling).

Performance on training set

	Accuracy	Recall	Precision	F1
0	0.793	0.825	0.776	0.800

Performance on validation set

	Accuracy	Recall	Precision	F1
0	0.759	0.783	0.552	0.647

Observations:

- **Overall Model Performance:**
  The Gradient Boosting model achieved a training accuracy of 0.793 and a validation accuracy of 0.759, indicating strong and consistent performance with minimal performance drop between the two datasets, suggesting good generalization.
- **Recall:**
  The model maintained a high recall on both training (0.825) and validation (0.783) sets, demonstrating its ability to correctly identify a large proportion of churned customers. This is crucial in churn prediction, where missing churners can lead to direct revenue loss.
- **F1-Score:**
  The F1-score decreased from 0.800 to 0.647, indicating some overfitting. The model performs slightly better on the training data than on unseen data. However, it still maintains a reasonable balance between recall and precision on validation.

3. Tuning AdaBoost with Original Data

Reasoning - Adaboost (Original Data) is the superior choice for initial deep tuning because of its exceptionally low overfitting. It has a higher Validation Score (0.6511). It provides the most robust starting point with the lowest risk, making it the primary tuning candidate for a production-stable mode.

Performance on training set

	Accuracy	Recall	Precision	F1
0	0.793	0.479	0.693	0.566

Performance on validation set

	Accuracy	Recall	Precision	F1
0	0.800	0.471	0.725	0.571

Observations:

- **Overall Model Performance:**
  The AdaBoost model trained on the original dataset demonstrated balanced and consistent performance across both training and validation sets, with training accuracy of 0.793 and validation accuracy of 0.800. The close alignment between these values indicates minimal overfitting and strong model stability.
- **Recall:**
  The recall remained low on both training (0.479) and validation (0.471) datasets, suggesting that the model struggled to correctly identify churned customers. While this reduces false positives, it increases the risk of missing potential churners.
- **Precision:**
  Precision was relatively high and consistent, at 0.693 for training and 0.725 for validation. This means that when the model predicted churn, it was usually correct, reflecting a precision-focused behavior suitable for avoiding unnecessary retention offers.
- **F1-Score:**
  The F1-scores of 0.566 (training) and 0.571 (validation) show a balanced trade-off between precision and recall. Though overall performance is moderate, the model generalizes well without significant loss of accuracy between training and validation sets.

4. Tuning Random Forest with Undersampled Data

Reasoning - Random Forest performed strongly on undersampled data (Recall 0.64–0.76), with high accuracy and balanced performance. It showed less sensitivity to overfitting compared to models trained on oversampled data.

Performance on training set

	Accuracy	Recall	Precision	F1
0	0.874	0.910	0.849	0.878

Performance on validation set

	Accuracy	Recall	Precision	F1
0	0.770	0.773	0.568	0.655

Observations:

- **Overall Model Performance:**
  The Random Forest model achieved a training accuracy of 0.874 and a validation accuracy of 0.770, showing strong predictive capability with a moderate performance drop between training and validation sets. This indicates some overfitting, but the model still generalizes reasonably well.
- **Recall:**
  The model recorded a high recall of 0.910 on the training set and 0.773 on the validation set, demonstrating its strong ability to identify churned customers correctly. High recall is valuable in churn prediction, as it ensures most potential churners are captured.
- **Precision:**
  Precision decreased from 0.849 (training) to 0.568 (validation), suggesting an increase in false positives on unseen data. While the model effectively identifies churners, it also misclassifies some non-churners, which could lead to unnecessary retention offers.
- **F1-Score:**
  The F1-score dropped from 0.878 to 0.655, indicating a noticeable reduction in balanced performance between the training and validation datasets. This shows that while the model performs excellently on training data, it loses some precision–recall balance on validation data due to minor overfitting.

Model Performance Comparison And Final Model Selection

- **Training performance comparison**

Training performance comparison:

	Adaboost tuned with undersampled data	Gradient Boost tuned with undersampled data	Adaboost tuned with original data	Random Forest tuned with undersampled data
Accuracy	0.768	0.793	0.793	0.874
Recall	0.846	0.825	0.479	0.910
Precision	0.731	0.776	0.693	0.849
F1	0.785	0.800	0.566	0.878

Validation performance comparison:

	Adaboost tuned with undersampled data	Gradient Boost tuned with undersampled data	Adaboost tuned with original data	Random Forest tuned with undersampled data
Accuracy	0.730	0.759	0.800	0.770
Recall	0.826	0.783	0.471	0.773
Precision	0.514	0.552	0.725	0.568
F1	0.634	0.647	0.571	0.655

The **AdaBoost model tuned with undersampled data** is the best-performing model. It achieves the highest recall, maintains consistent performance across datasets, handles class imbalance effectively, and aligns directly with the business goal of identifying as many potential churners as possible to minimize customer loss.

- It achieved the highest recall (0.826) on the validation set, outperforming all other models.
- While precision (0.514) is moderate, this trade-off is acceptable in churn prediction where missing churners (false negatives) are more costly than contacting a few extra non-churners (false positives).
- Its validation accuracy (0.73) and F1-score (0.634) are also reasonable and consistent with business priorities.

The performance of the best model on the test set

	Accuracy	Recall	Precision	F1
0	0.740	0.824	0.525	0.641

Observations:

- **High Recall (0.824):**
  The model successfully identifies about 82% of customers who actually churn, making it highly effective for retention campaigns aimed at minimizing customer loss.
- **Moderate Precision (0.525):**
  Around 52% of customers predicted as churners are truly at risk. While this means there are some false positives, it's an acceptable trade-off because, in business terms, false negatives (missed churners) are far more costly.
- **Balanced F1-Score (0.641):**
  The F1-score indicates a good balance between recall and precision, showing that the model maintains overall reliability without heavily sacrificing precision for recall.
- **Stable Accuracy (0.74):**
  The test accuracy confirms consistent performance and generalization from the validation stage, with minimal overfitting observed.

Feature Importance

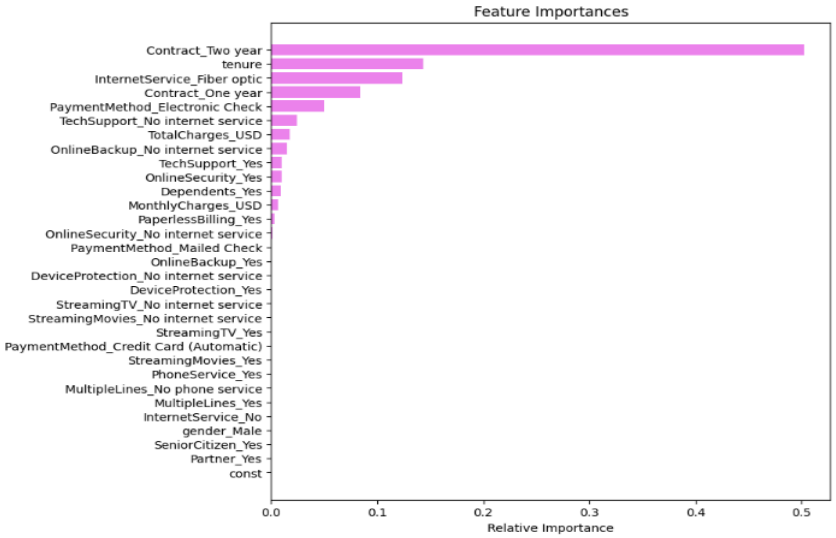


Fig.8: Feature Importance

- Contract duration and tenure strongly influence churn — customers with long-term contracts and longer relationships are far less likely to leave.
- Fiber optic users and electronic check payers show higher churn, highlighting dissatisfaction with service quality or billing experience.
- Demographics have minimal impact, confirming that churn is mainly driven by contract type, service quality, and payment behavior.

Actionable Insights & Recommendations

- **Encourage Long-Term Contracts**
  - Customers with month-to-month contracts show the highest churn, while those with one- or two-year contracts are far more loyal.
  - **Recommendation:** The company should incentivize long-term contracts through loyalty rewards, bundled offers, or discounted pricing for one- and two-year commitments. This will improve customer stability and reduce revenue volatility.

- **Focus on New and Short-Tenure Customers**
  - Customers with longer tenure are less likely to churn, while new customers show higher attrition risk.
  - **Recommendation:** Focus on early-stage customer retention programs such as onboarding support, welcome discounts, or follow-up calls during the first 3–6 months of service. Building trust and satisfaction early can increase lifetime value and retention.

- **Address Fiber Optic Customer. Churn**
  - Customers with longer tenure are less likely to churn, while new customers show higher attrition risk.
  - **Recommendation:** Conduct customer satisfaction surveys or feedback campaigns targeting fiber optic users. Offer customized service upgrades, pricing reviews, or improved technical support to address pain points and reduce churn in this high-value segment.

- **Optimize Billing and Payment Experience**
  - Customers using electronic checks and paperless billing exhibit higher churn rates, suggesting potential frustration with billing transparency or process complexity.
  - **Recommendation:** Simplify billing processes and improve communication around charges. Provide alternative payment options (auto-pay, credit card, digital wallets) and incentives for secure payment modes to improve convenience and trust.

- **Promote Value-Added Services**
    - Customers who subscribe to Tech Support, Online Security, or Device Protection show lower churn rates, as these services enhance perceived value and satisfaction.
    - **Recommendation:** Promote bundled service plans combining these add-ons at a discounted rate. Highlight the benefits of these services through personalized marketing to encourage adoption and deepen customer engagement.

- **High Monthly Charges Increase Churn Probability**
    - Customers with higher monthly bills tend to churn more frequently, likely due to perceived low value-for-money or pricing dissatisfaction.
    - **Recommendation:** Introduce tiered pricing plans, customizable packages, or loyalty-based discounts for long-term customers. Transparent communication about value-added features can also help justify higher costs.

- **Improve Customer Feedback and Support Channels**
    - Many churners, such as dissatisfaction with service or billing, can be mitigated through better communication.
    - **Recommendation:** Strengthen customer touchpoints by implementing 24/7 support chat, feedback forms, and post-service follow-ups. Rapid resolution of issues can directly reduce churn risk.

Appendix

- **Code File -** HTML Code File