Customer Churn Prediction Report

Introduction

The telecommunication company plans to predict the likelihood of customers leaving. It has collected information from 7,043 customers to better understand their demographics and preferences, aiming to improve services and increase the retention rate. The dataset includes basic demographic information, such as gender and age range; account features, such as contract type and charge amounts; services subscribed to; and whether the customers continue using the services. This report aims to provide an in-depth analysis of customer patterns and model the prediction of the churn rate, offering insights and strategic recommendations to enhance customer retention.

Data Cleaning

The raw dataset contains 7,043 entries (customers) and 21 columns (features). These features can be categorized into four groups to describe the information:

1. **Customer Demographic Information:** Includes customer ID, gender, age range (indicated as senior citizen), and whether the customer has a partner or dependents.
2. **Account Information:** Describes the customer's tenure (how long they have been using the services), contract type, payment method, paperless billing status, monthly charges, and total charges.
3. **Signed-Up Services:** Includes details about subscribed services such as phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming for TV and movies.
4. **Churn:** Indicates whether the customer left within the last month.

To clean the raw dataset, we first remove the unnecessary column ‘CustomerID.’ Next, we examine the dataset to review the data types and find that ‘TotalCharges’ is stored as an object. We convert it to a numerical column for proper analysis. Additionally, we convert the columns ‘Partner,’ ‘Dependent,’ ‘PaperlessBilling,’ and ‘Churn’ to binary format to ensure they are appropriately used in the analysis.

Missing values

We check the missing values of the whole dataset and find there are 11 missing values in the ‘TotalCharges’. Since the distribution shows right-skewed and the number of missing values is a relatively small fraction of the whole, we impute it with the median.

Adjusting inconsistency

For the signed-up services, there are two main categories: phone services and internet services. Each category includes three possible entries: ‘Yes,’ ‘No,’ and ‘No services.’ If customers sign up for a main service, their entries for the dependent branch services are marked as ‘Yes’ or ‘No.’ However, if customers do not sign up for the main service, their entries for the dependent branch services are recorded as ‘No services.’ Since there is no significant difference between

‘No’ and ‘No services’ in detecting its effect on the retention rate, we standardized the entries by adjusting ‘No services’ to ‘No.’

Detecting outliers

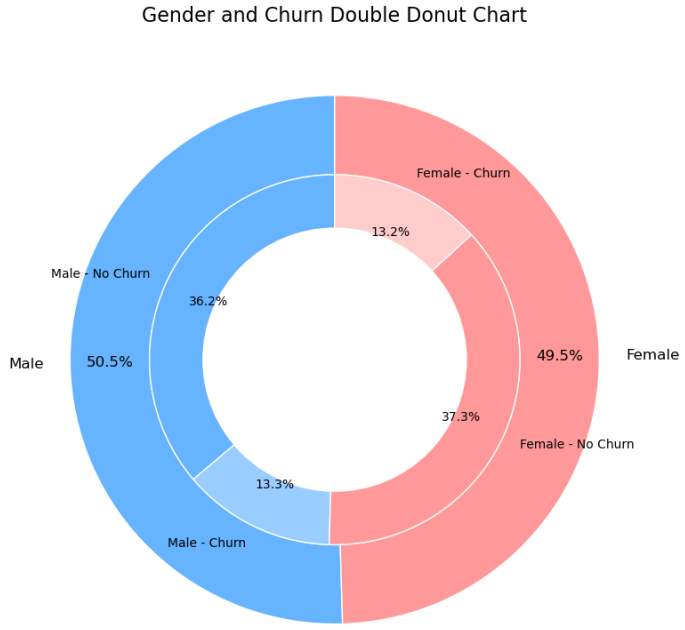
We apply the Interquartile Range (IQR) method to identify potential outliers and use boxplots to visualize the results for the numerical columns: ‘Tenure,’ ‘MonthlyCharges,’ and ‘TotalCharges.’ Both the IQR analysis and boxplots confirm that there are no outliers in the dataset.

Exploratory Data Analysis

Customer demographics

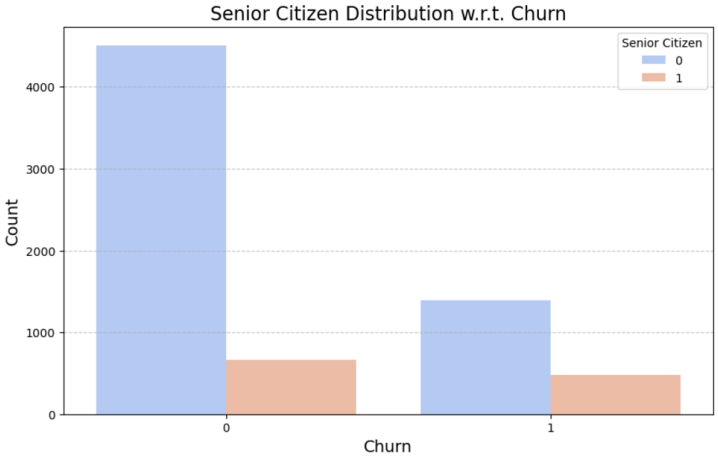
Gender

Customers are 49.5 % female and 50.5 % male. 26.5 % of customers switched to another firm.
The churn rates for male and female customers are very similar, indicating that gender is not a significant factor in determining churn behavior.



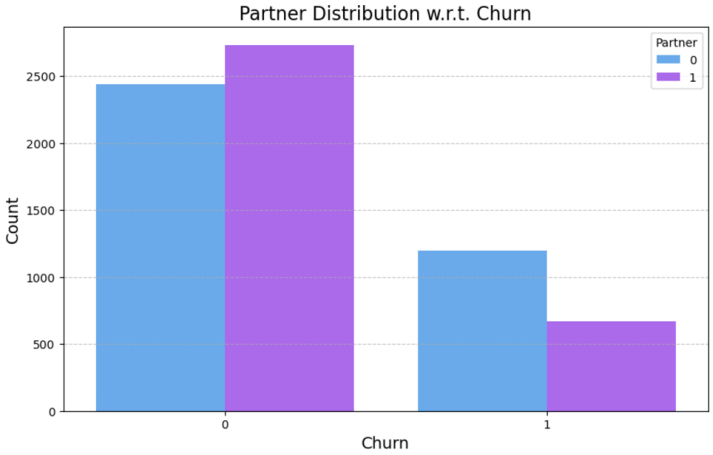
Senior citizens

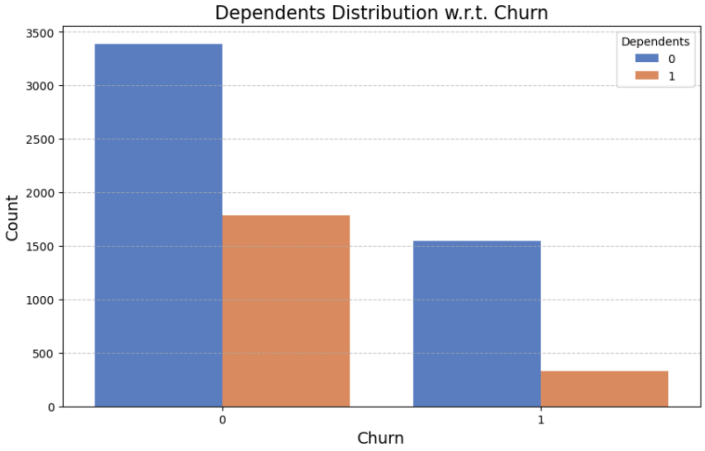
Senior citizens have a higher probability of churning, but only about 16% of the customers are senior citizens, they constitute a small fraction of the customer base.



Partner and Dependents

Approximately 48.3% of customers have a partner. About 30% of the customers have dependents. Customers without a partner or dependents are more likely to churn. This suggests that individuals may find it easier to switch service providers.

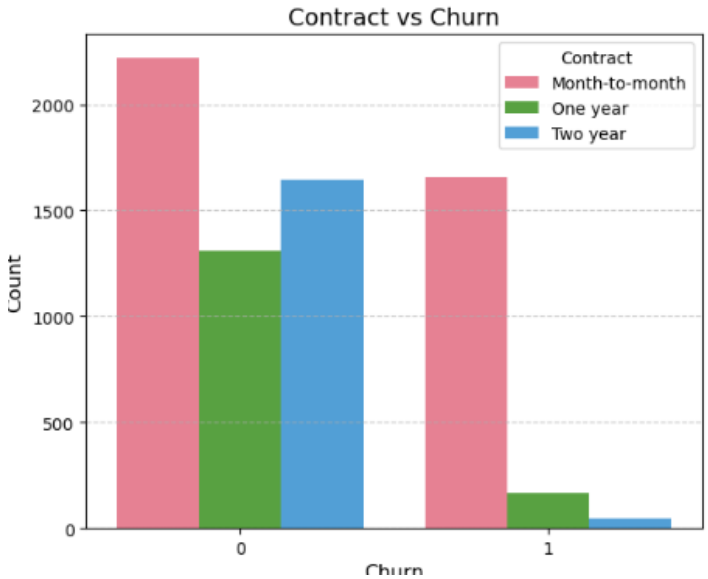




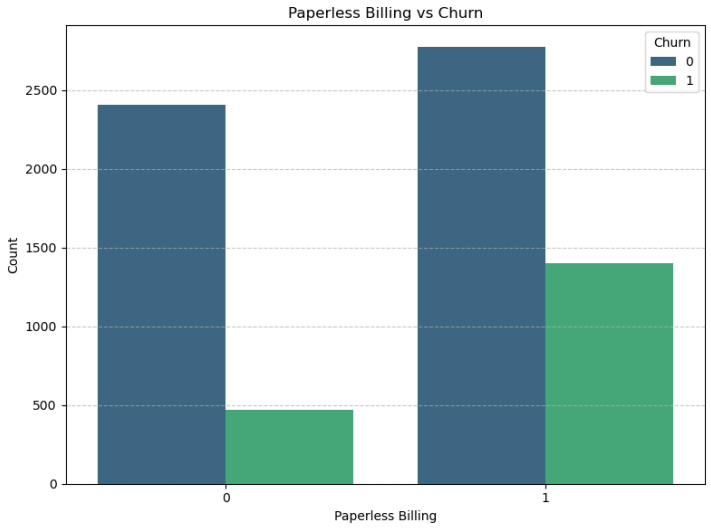
Account Features

Contract Type

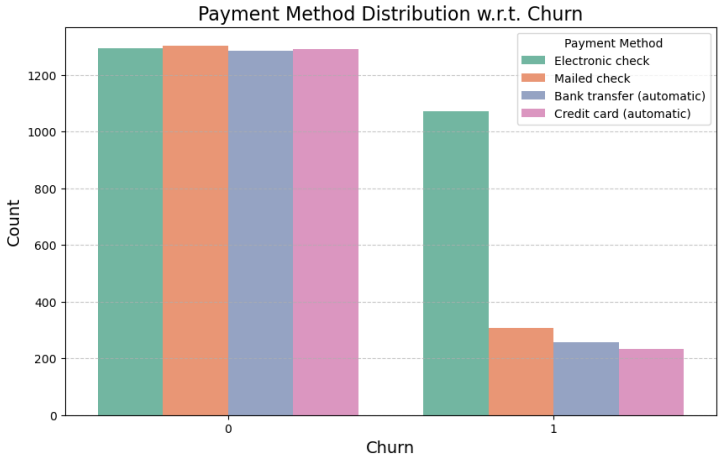
Most customers have month-to-month contracts. However, Month-to-month contracts have higher churn compared to longer-term contracts. About 89% of customers with Month-to-Month Contract move out to other providers as compared to 9% of customers with One Year Contract and 2.6% with Two Year Contract.



Paperless billing
More customers use paperless billing compared to paper billing. Customers with paperless billing show a higher churn rate, indicating dissatisfaction with the convenience or user experience associated with paperless billing.



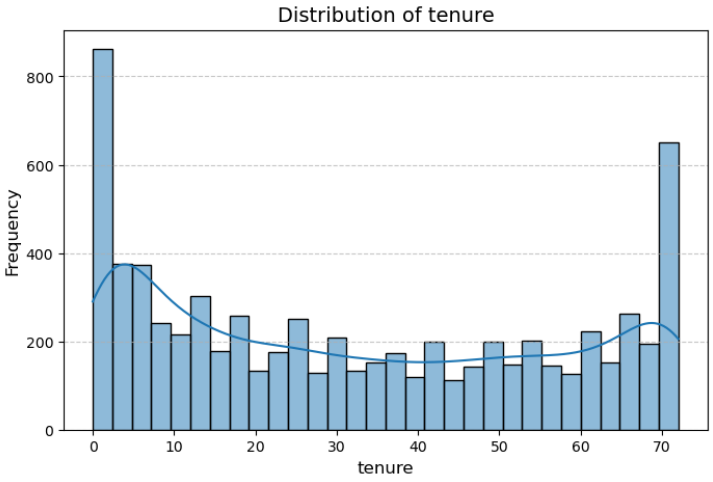
Payment Methods
Electronic check is the most popular way to pay. But customers paying via electronic check are the most likely to churn.

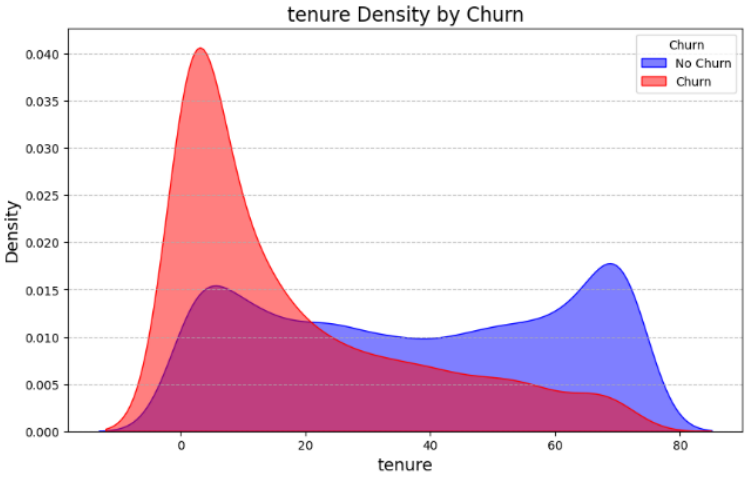


Tenure

This distribution is bimodal, with peaks at the low and high tenure ranges. The middle range (10–60 months) has relatively even and lower frequency, fewer customers in this tenure range.

Customers with short tenure (<10 months) are more likely to churn, indicating the importance of onboarding and early engagement strategies.

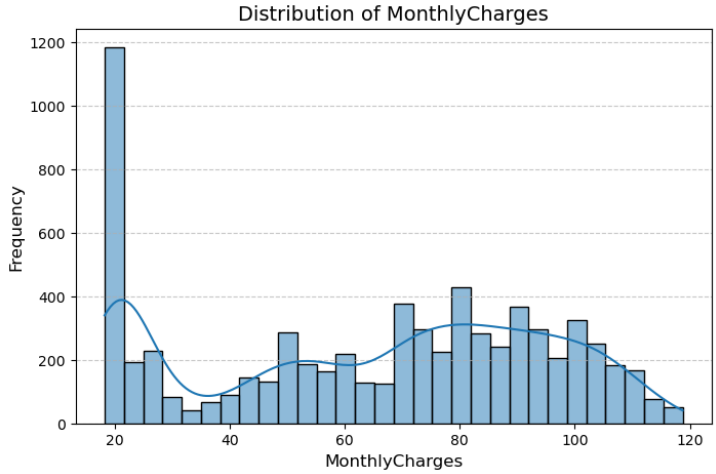


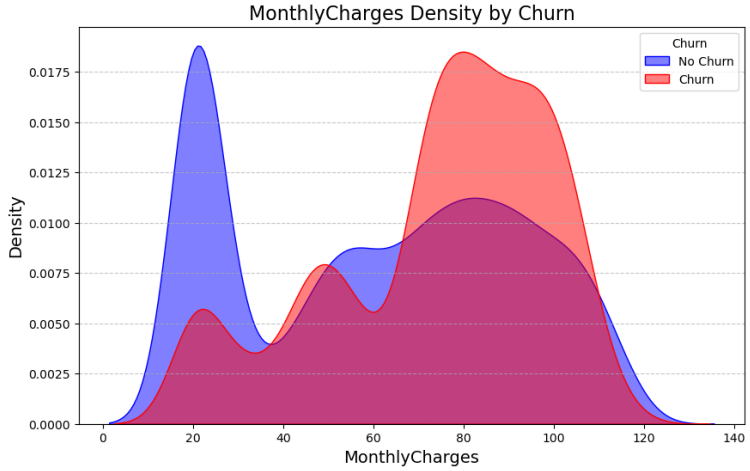


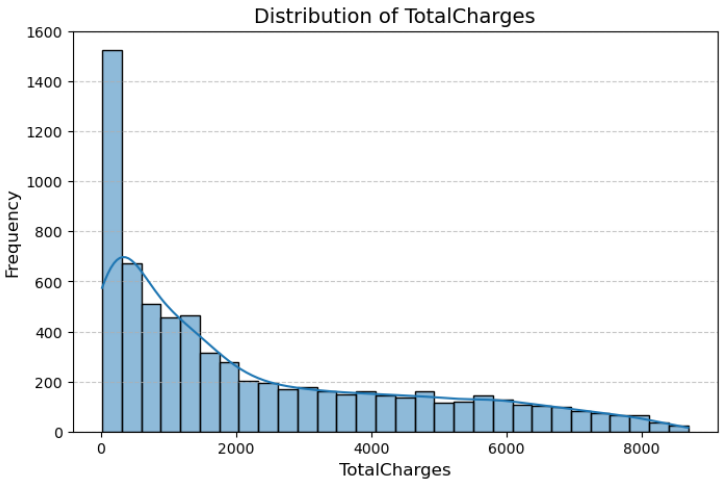
Monthly charges and total charges

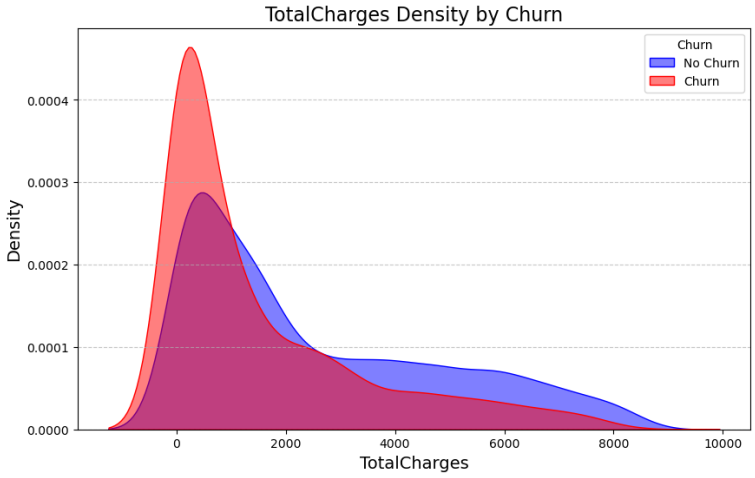
The monthly charges distribution is slightly right skewed. There is a concentration around 20-30 and 70-100, this indicates most customers have lower monthly charges. Churn likelihood increases as monthly charges increase, with the highest density observed between $70 and $100.

Total charges' distribution is heavily concentrated at the lower range around $0–500. The distribution is right-skewed, showing a clear long tail extending toward very high total charges. The density is higher for lower total charges (below $1,000), indicating newer customers are at higher risk of churn. Lower total charges correlate with shorter tenures, reinforcing the importance of customer retention strategies early in their lifecycle.





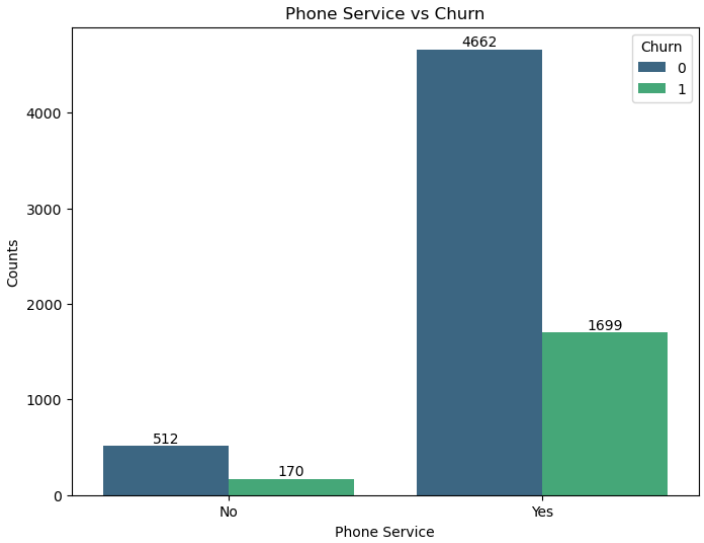


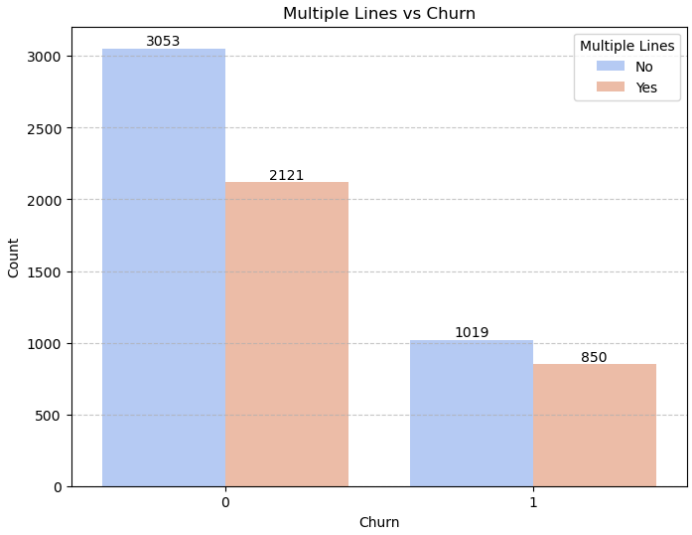


Usage Patterns

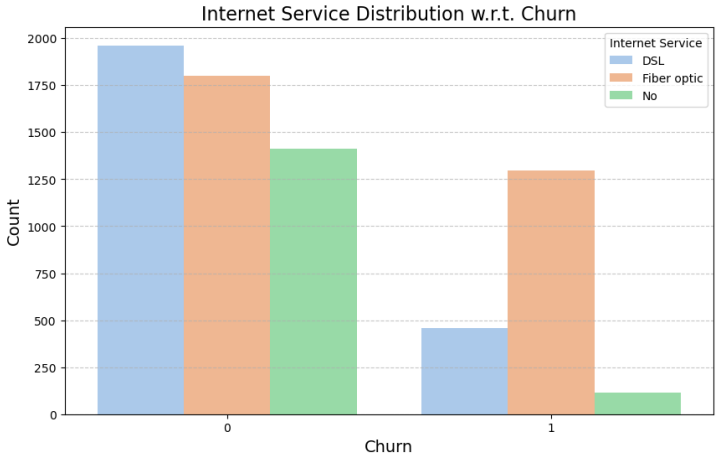
Phone Service and Multiple Lines

A very small fraction of customers doesn't have a phone service. Customers with phone services are more likely to churn, especially when they have multiple lines. When customers don't have phone service, the churn rate is 24.9%; when customers have service, the churn rate is 26.7%, under this circumstance, if they have single line, their churn rate is 25%, but if they have multiple lines, their churn rate is 28.6%.

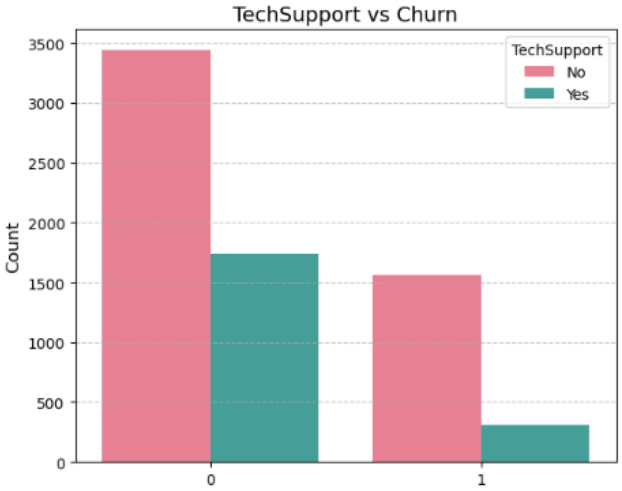




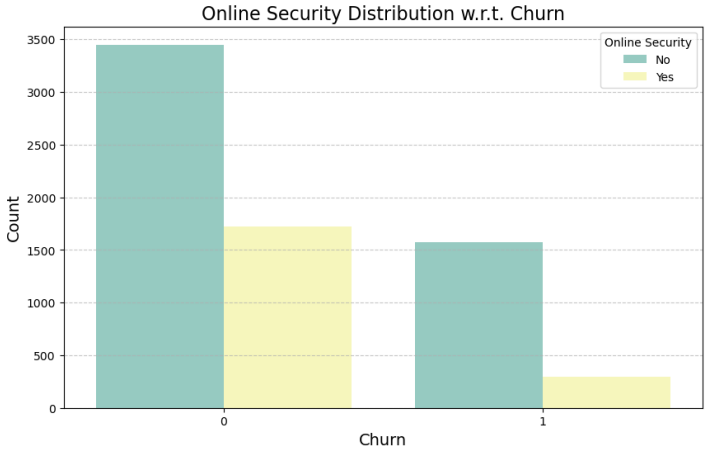
Internet Service
Customers who choose Fiber optic are more than DSL, but customers who use Fiber optic have higher churn rate than DSL and no service, this might suggest a dissatisfaction with this type of internet service.



TechSupport
The number of customers who don't have tech support is more than those who have tech support. Customers without tech support are most likely to migrate to another service provider

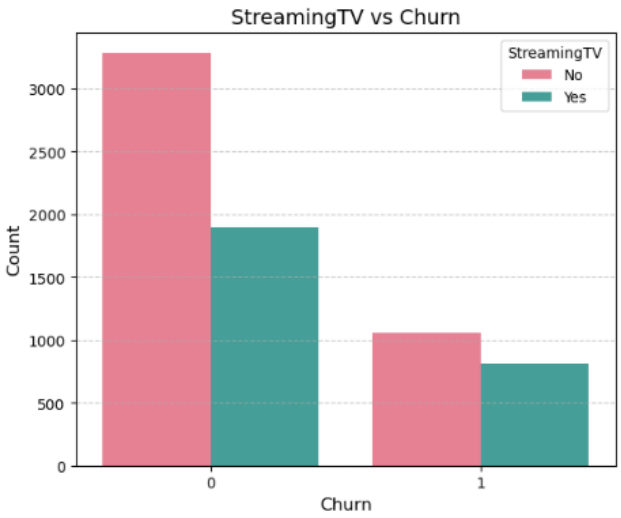


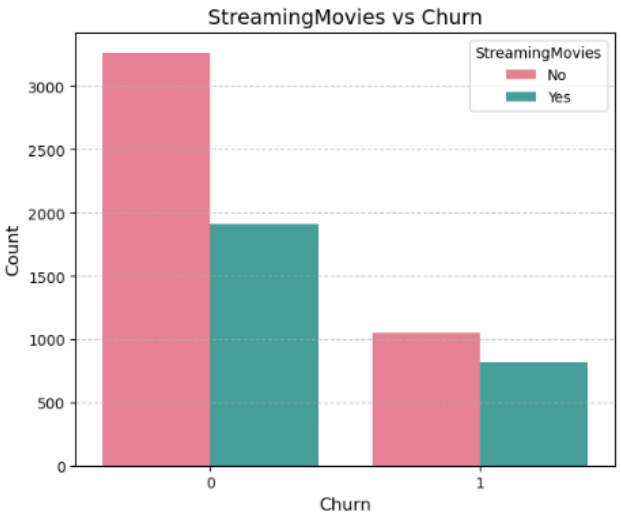
Online security
More customers don't have online security. Customers without online security have a higher churn rate.



Streaming Services (TV and Movies)

Less customers subscribe to streaming TV and movies than subscribers. Customers subscribed to streaming services (both TV and movies) are slightly more likely to churn. This could indicate a mismatch between the services offered and customer expectations.





Modeling prediction

Model Selection

The logistic regression model was selected for its simplicity, interpretability, and effectiveness in binary classification problems. The goal is to identify customers at high risk of leaving and to develop strategic actions to improve retention. Categorical variables were encoded using One-Hot Encoding, and numerical features were standardized using StandardScaler. The dataset was split into 80% training and 20% testing for robust evaluation. The model's performance was measured using Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

Feature Analysis

Based on the Odds ratio, key influential features include TotalCharges and InternetService_Fiber Optic, both of which positively correlate with churn. Customers with higher total charges are 1.865 times more likely to churn compared to those with lower total charges. Fiber optic users are 1.702 times more likely to churn compared to customers using other types of internet services. This suggests potential dissatisfaction with the quality, reliability, or cost of fiber optic services.

Conversely, longer tenure and two-year contracts significantly reduce churn likelihood. For every unit increase in tenure, the odds of churn decrease by a factor of 0.266. This indicates that longer-tenured customers are significantly more loyal and less likely to leave. Longer-term contracts also provide stability and often include incentives that reduce churn likelihood. Retention efforts should focus on maintaining satisfaction for these long-term customers.

The seemingly contradictory relationship between TotalCharges (high odds ratio indicating high churn risk) and tenure (low odds ratio indicating low churn risk) can be interpreted through the lens of context and customer behavior. TotalCharges is a cumulative value dependent on MonthlyCharges and tenure. However, customers with higher MonthlyCharges over shorter tenures can accumulate high TotalCharges, especially if they use premium or multiple services. Therefore, high TotalCharges may represent customers who pay more monthly but leave early due to dissatisfaction, perceived overpricing, or lack of perceived value.

Model Evaluation

Based on the confusion matrix and ROC curve, the logistic regression model achieved an accuracy of 82% and a ROC-AUC score of 0.86, indicating strong overall performance. Class 0 (Non-Churners) exhibited high precision, recall, and F1-Score, while Class 1 (Churners) had a lower recall of 0.60, highlighting the need for improvement in identifying high-risk customers.

Confusion Matrix	
933(TN)	103(FP)
149(FN)	224(TP)

	Precision	Recall	F1- Score	Support
0 (Non-Churners)	0.86	0.90	0.88	1036
1 (Churners)	0.69	0.60	0.64	373
Accuracy			0.82	1409
Macro Avg	0.77	0.75	0.76	1409
Weighted Avg	0.82	0.82	0.82	1409

Insights and recommendations

EDA and classification help the company identify high-churn segments, improving retention and reducing customer churn. We find four key predictors: (1) TotalCharges**:** With an odds ratio of 1.865, customers with higher total charges are significantly more likely to churn. (2) InternetService_Fiber Optic: An odds ratio of 1.702 suggests that fiber optic service users are at a higher risk of churn, possibly due to dissatisfaction with service quality or price. (3) Tenure: An odds ratio of 0.266 emphasizes that longer tenure strongly reduces churn likelihood, which means customer loyalty is very important. (4) Contract_Two year: An odds ratio of 0.542 demonstrates that two-year contracts decrease churn risk, indicating that longer commitments enhance retention. Here are targeted retention strategies.

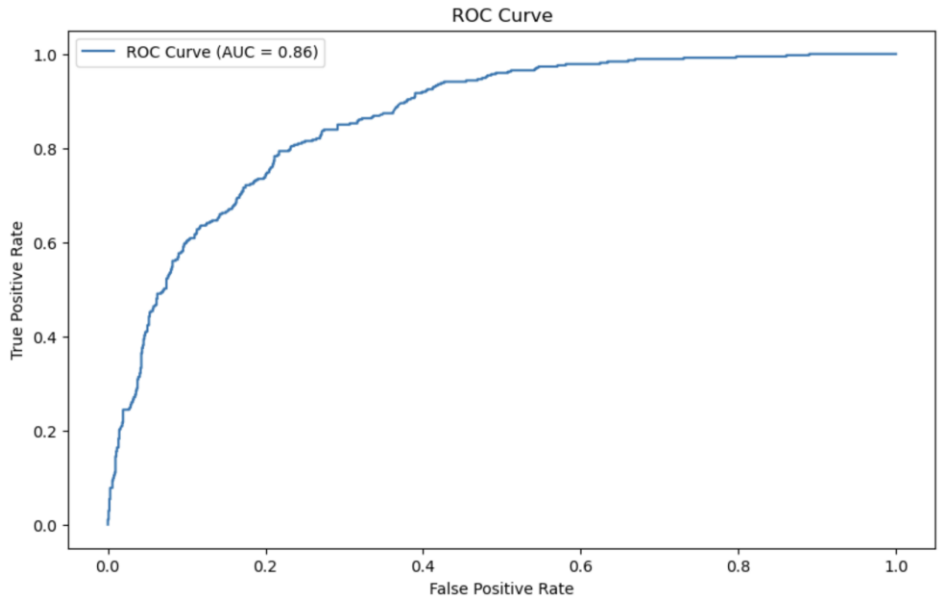
1. **Focus on High-Value Customers**: Create a retention strategy tailored for high-value customers. For customers with high total charges who pay more monthly but leave early, companies should conduct satisfaction surveys to identify dissatisfied drivers. Offer discounts, exclusive benefits or loyalty points for keeping these valuable customers.
2. **Enhance Service Reliability**: Address quality or other concerns related to fiber optic services through targeted performance improvements or value-added features.
3. **Improve Onboarding and Early Engagement**:Keeping customers at an early stage is essential to improving tenure and contract duration. Company need to develop a structured onboarding process for new customers, focusing on quick issue resolution, usage tips, and personalized recommendations. Also, implementing proactive follow-ups within the first few months can ensure customer satisfaction.
4. **Promote Long-Term Contracts**: Provide incentives for transitioning to one- or two-year contracts, such as discounts, loyalty rewards, or additional perks, to encourage long-term commitments.

5. **Run A/B Testing for Interventions**: Continuously test the effectiveness of different interventions (e.g., discounts, bundling, communication strategies) to identify the most impactful approaches for reducing churn.

Appendix: Statistics Summary and Model Output

	SeniorCitizen	Partner	Dependents	tenure	PaperlessBilling	MonthlyCharges	TotalCharges	Churn
count	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000
mean	0.162147	0.483033	0.299588	32.371149	0.592219	64.761692	2281.916928	0.265370
std	0.368612	0.499748	0.458110	24.559481	0.491457	30.090047	2265.270398	0.441561
min	0.000000	0.000000	0.000000	0.000000	0.000000	18.250000	18.800000	0.000000
25%	0.000000	0.000000	0.000000	9.000000	0.000000	35.500000	402.225000	0.000000
50%	0.000000	0.000000	0.000000	29.000000	1.000000	70.350000	1397.475000	0.000000
75%	0.000000	1.000000	1.000000	55.000000	1.000000	89.850000	3786.600000	1.000000
max	1.000000	1.000000	1.000000	72.000000	1.000000	118.750000	8684.800000	1.000000

	Feature	Coefficient	Odds Ratio	Predicted Probability
6	TotalCharges	0.623635	1.865698	0.651045
10	InternetService_Fiber optic	0.532133	1.702560	0.629981
17	StreamingMovies_Yes	0.194683	1.214926	0.548518
4	PaperlessBilling	0.163782	1.177958	0.540854
21	PaymentMethod_Electronic check	0.152621	1.164884	0.538081
9	MultipleLines_Yes	0.151060	1.163067	0.537693
16	StreamingTV_Yes	0.144517	1.155482	0.536067
0	SeniorCitizen	0.058474	1.060218	0.514614
1	Partner	0.027348	1.027725	0.506836
14	DeviceProtection_Yes	0.011120	1.011182	0.502780
22	PaymentMethod_Mailed check	-0.012211	0.987863	0.496947
7	gender_Male	-0.025731	0.974597	0.493568
20	PaymentMethod_Credit card (automatic)	-0.034944	0.965659	0.491265
13	OnlineBackup_Yes	-0.061559	0.940298	0.484615
2	Dependents	-0.073888	0.928776	0.481536
8	PhoneService_Yes	-0.113404	0.892790	0.471679
15	TechSupport_Yes	-0.137397	0.871624	0.465705
12	OnlineSecurity_Yes	-0.174481	0.839893	0.456490
18	Contract_One year	-0.267780	0.765076	0.433452
5	MonthlyCharges	-0.407462	0.665337	0.399521
11	InternetService_No	-0.451280	0.636813	0.389056
19	Contract_Two year	-0.611867	0.542337	0.351633
3	tenure	-1.324222	0.266010	0.210117


ROC Curve
ROC Curve (AUC = 0.86)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0