2024

Task 1 : Exploratory Data Analysis on Customer Churn

BY: Sheema Masood

EcodeCamp Software Company

Introduction

The dataset used in this analysis is the Telco Customer Churn dataset, sourced from Kaggle. This dataset contains information on a fictional telecommunications company's customer base, including customer demographics, service usage metrics, and a churn indicator, which signifies whether a customer has left the company or not.

Features Overview:

∗ Customer Demographics: Information such as gender, age, seniority, and whether the customer has dependents.

∗ Service Usage Metrics: Details about the services used by customers, including phone, internet, and streaming services.

∗ Additionally, it covers the type of contracts customers have, payment methods, and monthly and total charges.

∗ Churn Indicator: A binary feature that indicates whether a customer has churned (left the company) or not.

Purpose of the Analysis:

The primary objective of this analysis is to identify patterns and gain insights that can help predict customer churn. By understanding the factors that contribute to churn, the company can develop strategies to retain customers and reduce churn rates, ultimately improving business performance and customer satisfaction.

<u>Defining Target and Features for Customer Churn Analysis</u>

This analysis aims to predict customer churn in a telecommunications (telco) company. We'll use a dataset containing information about customer demographics, service usage, account details, and financial data to build a predictive model.

Target Variable:

Churn: This binary variable indicates whether a customer has churned (left the service) or remained a customer.

- o 1: Customer has churned.
- o 0: Customer is still active.
- o

Features (Predictor Variables):

1. Account and Service Usage:

- **tenure:** Length of time the customer has been with the telco (in months).
- **PhoneService:** Whether the customer has phone service (binary: Yes/No).
- **MultipleLines:** Whether the customer has multiple phone lines (categorical: No phone service, Yes, No).
- **InternetService:** Type of internet service the customer has (categorical: DSL, Fiber optic, No).
- **OnlineSecurity:** Whether the customer has online security service (binary: Yes/No).
- **OnlineBackup:** Whether the customer has online backup service (binary: Yes/No).
- **DeviceProtection:** Whether the customer has device protection (binary: Yes/No).
- **TechSupport:** Whether the customer has tech support (binary: Yes/No).
- **StreamingTV:** Whether the customer has streaming TV service (binary: Yes/No).
- **StreamingMovies:** Whether the customer has streaming movie service (binary: Yes/No).

2. Customer Demographics:

- **gender:** Gender of the customer (likely binary: Male/Female).
- **SeniorCitizen:** Indicates if the customer is a senior citizen (binary: 0 = No, 1 = Yes).
- **Partner:** Whether the customer has a partner (binary: Yes/No).
- **Dependents:** Whether the customer has dependents (binary: Yes/No).

3. Account Details:

- **Contract:** Type of contract the customer has (categorical: Month-to-month, One year, Two year).
- **PaperlessBilling:** Whether the customer uses paperless billing (binary: Yes/No).
- **PaymentMethod:** How the customer pays their bills (categorical: Electronic check, Mailed check, Bank transfer, Credit card).

4. Financial:

- **MonthlyCharges:** Amount the customer pays each month.
- **TotalCharges:** Total amount the customer has paid to the telco (numeric).

Note:

- Categorical features will need to be encoded (e.g., one-hot encoding) before being used in machine learning models.
- We may explore creating additional features through feature engineering to improve model performance.

Data preprocessing

1- **Handling Missing Values:**

There are no missing values in dataset as we can see in df.info

2- **Transforming Data:**

- **Convert Sinior citizen column into str :**

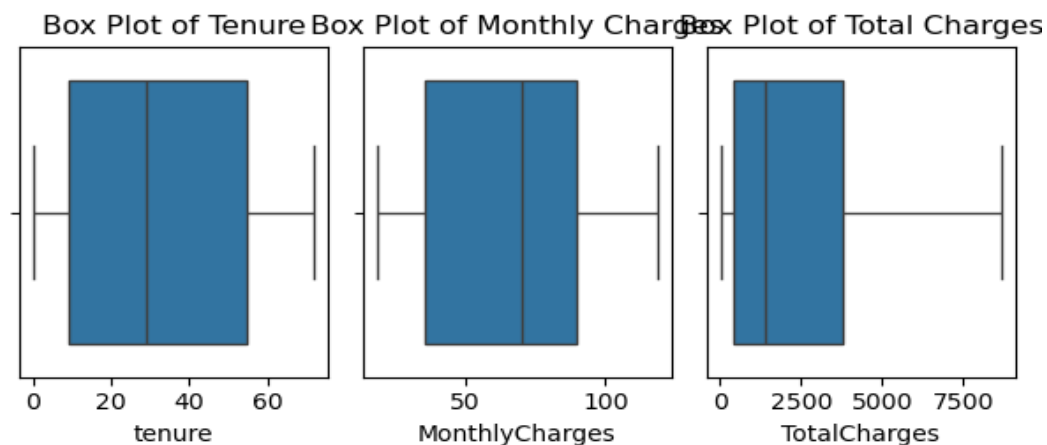Since Sinior citizen is stored as an int, convert it to str , because it represents categorial data.

- **Convert TotalCharges to Numeric:**

Since TotalCharges is stored as an object, convert it to a numeric data type (float64).

3- Managing Outliers:

Check for outliers that might distort the analysis or model performance.

Visualization of Any outliers in dataset:

Key Findings:

1- Tenure:

The box plot for tenure shows no significant outliers. The data appears well-distributed within the whiskers, so no action is required here.

2- MonthlyCharges:

Similar to tenure, the MonthlyCharges box plot does not show significant outliers, indicating no extreme values that need to be addressed.

3- TotalCharges:

no significant outliers

Exploratory Data Analysis

Churn: Target Column

The pie chart represents the distribution of churned and non-churned customers in your dataset. The key findings and insights based on this chart are:

1. **Churn Rate**: 26.5% of customers have churned (represented in orange). This indicates that approximately one-quarter of the customers have left the service or discontinued their relationship with the business.
2. **Retention Rate**: 73.5% of customers have not churned (represented in blue). This shows that a significant portion of the customer base is still retained, suggesting customer satisfaction or loyalty for the majority.

Churn Distribution
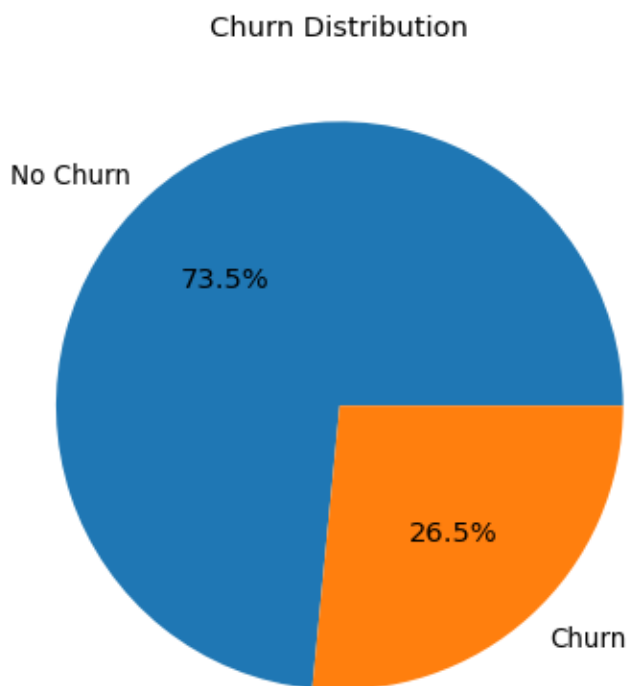
Business Implications:

- The churn rate of 26.5% is relatively high, which could be concerning for the business. Reducing this churn percentage should be a focus area to retain more customers.
- Retention strategies, customer satisfaction initiatives, and targeted retention campaigns could be employed to bring the churn rate down.

Identify Patterns and Potential Predictors

Visualizing the relationships between features and churn using various plots.

1. **Tenure and Churn:**

Key Findings:

- Customers with a longer tenure are less likely to churn, as depicted by the higher median and wider spread for those who did not churn.
- Customers who churned tend to have a shorter tenure, indicating that newer customers are more likely to leave compared to those who have been with the company for a longer time.
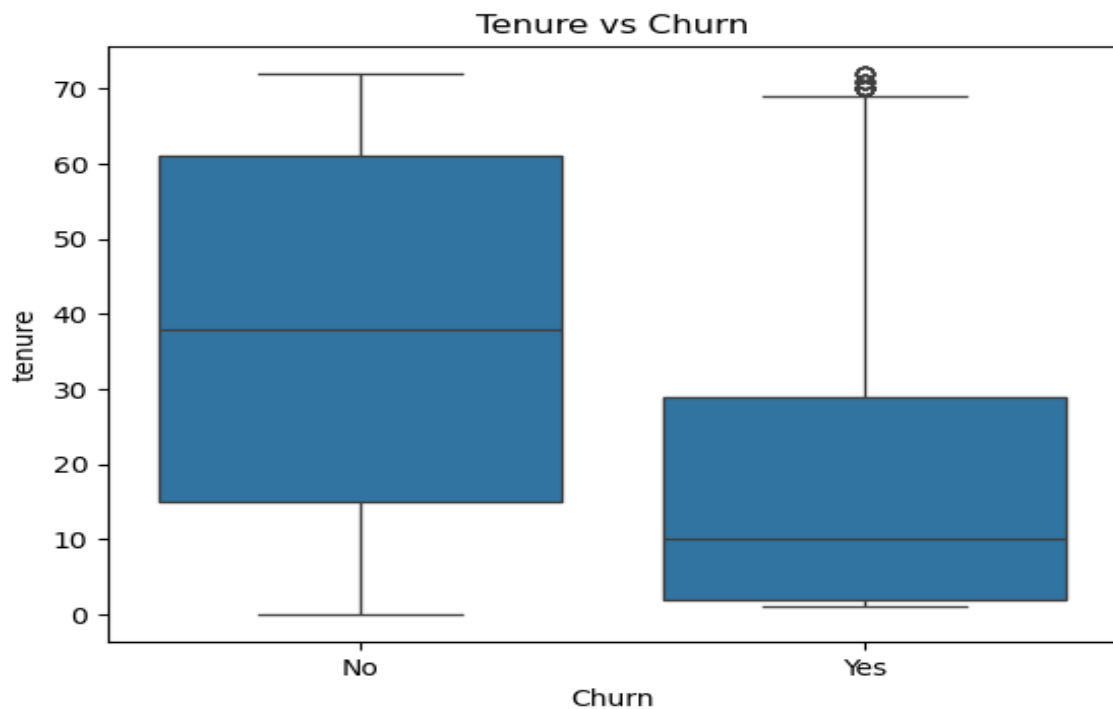
Insights:

- Implementing onboarding and early engagement strategies could help in retaining new customers during the critical early months.

## 2. Relationship of monthly & total charges with Churn:
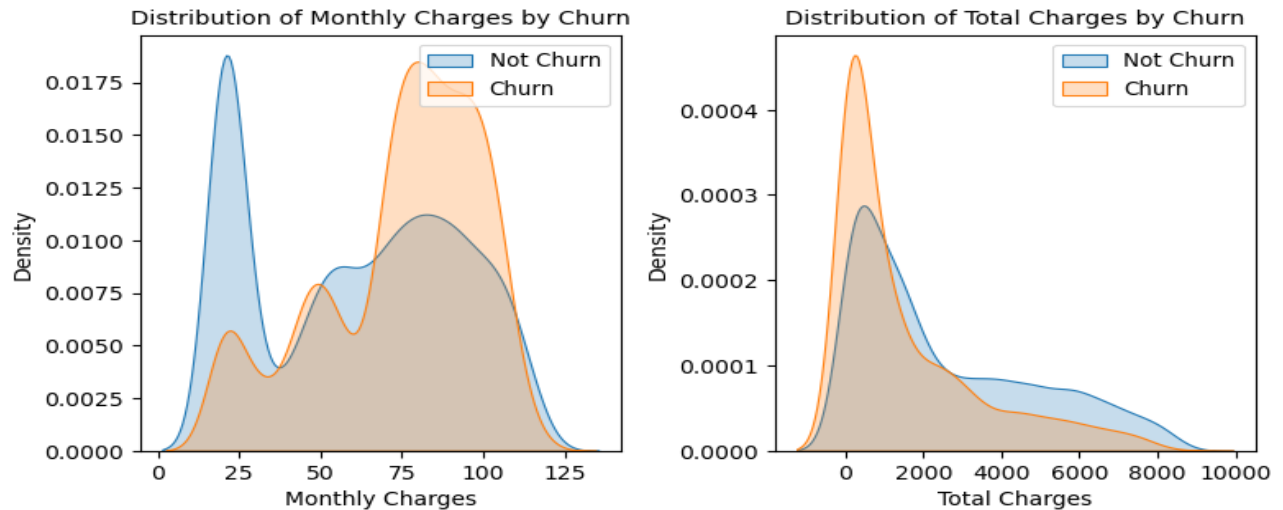
Key Findings

- The density plot shows that customers who churn (orange) tend to have higher monthly charges compared to those who do not churn (blue).
- For total charges, the density plot reveals that churned customers (orange) generally have accumulated lower total charges compared to non-churned customers (blue).

Insights:

- Higher monthly charges are correlated with a higher likelihood of churn, indicating that pricing could be a key factor in customer retention.
- This suggests that customers who leave early (churn) tend to accumulate fewer total charges, likely because they have a shorter tenure or were dissatisfied early in the relationship.

.

3. "Distribution of Monthly and Total Charges by Churn"*
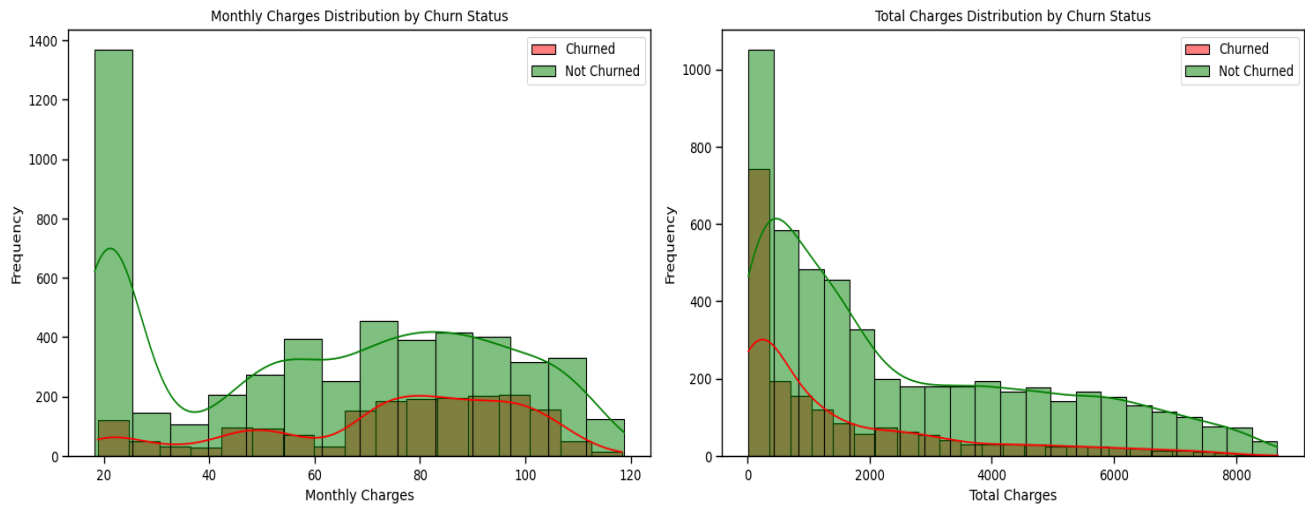
Key Findings:

- The histogram shows that customers with low monthly charges (around $20) rarely churn. However, as the monthly charges increase, the number of churned customers (red) also increases, especially in the mid-range charges.
- The histogram indicates that churned customers generally have lower total charges. Non-churned customers (green) tend to have higher total charges, suggesting they have been with the company longer or have more services.

Insights:

- Customers with higher total charges, likely due to longer tenure or more services, are less likely to churn. This underscores the importance of customer satisfaction and long-term engagement.
- While low charges retain customers, mid-to-high charges may lead to dissatisfaction, prompting churn. This highlights the need to align pricing with perceived value.

General Conclusion:

Pricing and Retention:

Both sets of charts suggest that while higher monthly charges can drive churn, customers who stay longer and accumulate higher total charges tend to remain loyal. Pricing strategies should consider customer value perception, particularly for higher-tier services.

4. Relationship between type of Internet services with churn
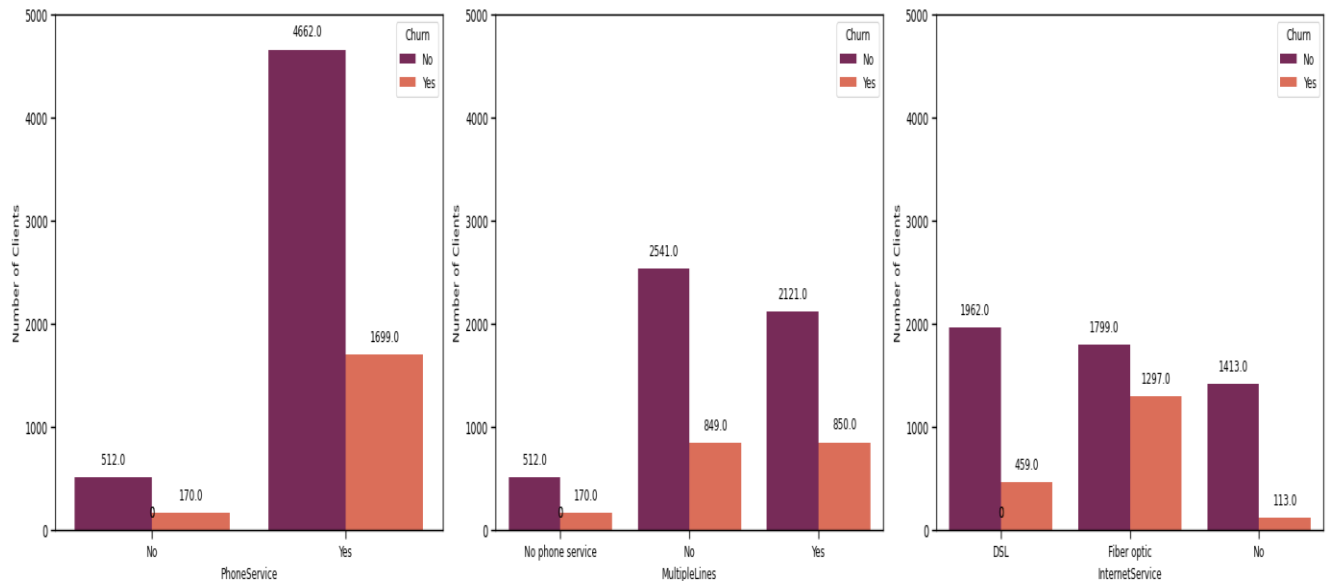
Key Findings:

1. Churn by Phone Service
   - Higher Churn with Phone Service: Customers with PhoneService have a significantly higher churn rate compared to those without it.
2. Churn by Multiple Lines
   - Moderate Churn with Multiple Lines: Customers who have MultipleLines tend to churn at a similar rate whether they use them or not. However, those with no phone service do not churn at all, which might indicate this group is less engaged or reliant on the service.
3. Churn by Internet Service
   - Highest Churn with Fiber Optic: Customers using Fiber optic internet service have a higher churn rate compared to those using DSL or no internet service.
4. Lowest Churn with No Internet:
   - Customers without an internet service have the lowest churn rate, suggesting they might have less dependency on the service, or are more satisfied with basic offerings.

Insights:

- Focus on upgrading DSL customers to fiber optic services, where available, to reduce churn.
- Promoting phone service as part of a bundled offer might help in reducing churn.

5. Relationships between Services Subscription with churn

OnlineSecurity	TechSupport	OnlineBackup	DeviceProtection	
Yes	Yes	Yes	Yes	90.918511
No	Yes	Yes	Yes	90.428906
	No	Yes	Yes	89.195122
Yes	No	Yes	Yes	85.139462
No	Yes	No	Yes	80.545833
	No	No	Yes	79.551233
Yes	Yes	No	Yes	79.336008
No	Yes	Yes	No	77.975121
Yes	No	No	Yes	77.539429
		Yes	No	75.801636
	Yes	Yes	No	75.446575
No	No	Yes	No	74.605945
	Yes	No	No	69.214050
	No	No	No	67.774783
Yes	No	No	No	67.725974
	Yes	No	No	65.901796

Key Findings:

1. When we look at the chart, it's clear that having all four services—OnlineSecurity, TechSupport, OnlineBackup, and DeviceProtection—really makes a difference. Customers with a full set of these services score the highest, reaching up to 90.92. This suggests that when people have all their bases covered, they're not just satisfied but truly happy with their experience.
2. On the flip side, if customers don't have any internet services, their scores drop dramatically to just 21.08. This shows how crucial these services are—without them, the experience is significantly less positive.
3. For those who have some services but not all, the values fall in between. For example, having OnlineSecurity and TechSupport but missing out on OnlineBackup and DeviceProtection still gives a decent score, but it's not as high as having the full package. This tells us that each service adds value, but missing just one or two can make a noticeable difference.
4. In short, the more services you offer, the happier your customers seem to be. If you're looking to boost satisfaction or performance, focusing on providing a complete suite of services might be the key to achieving the best results.
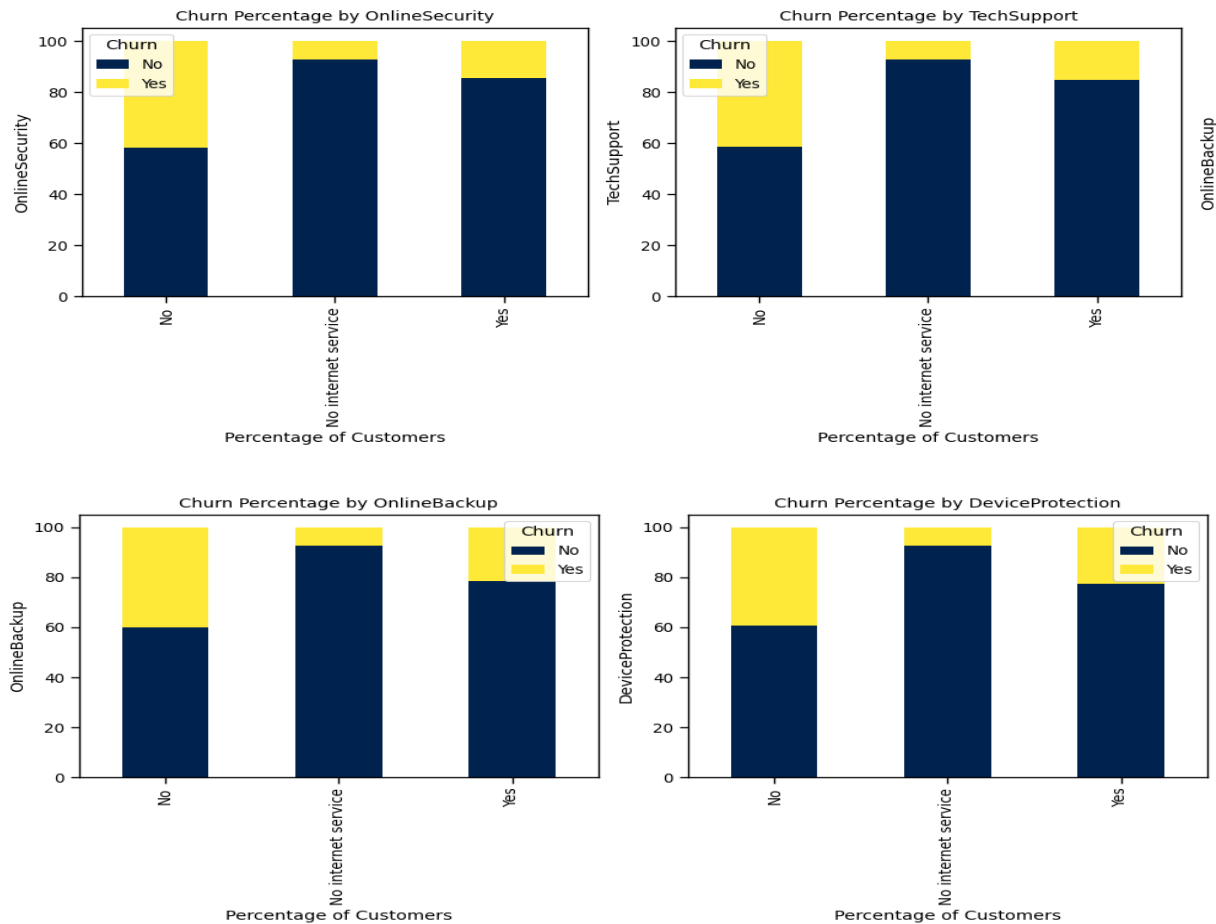
Key Findings "Proportion of Customers":

- It highlights that a higher proportion of customers without OnlineSecurity tend to churn compared to those with the service.
- The chart indicates that customers without TechSupport have a higher churn rate than those with TechSupport.
- It shows that customers without OnlineBackup have a higher churn percentage compared to those who have this service.
- customers who do not have DeviceProtection are more likely to churn compared to those who do.

Insights:

Bundling additional services with attractive pricing or offering trial periods might encourage customers to stay

6. Churn w.r.t payments method
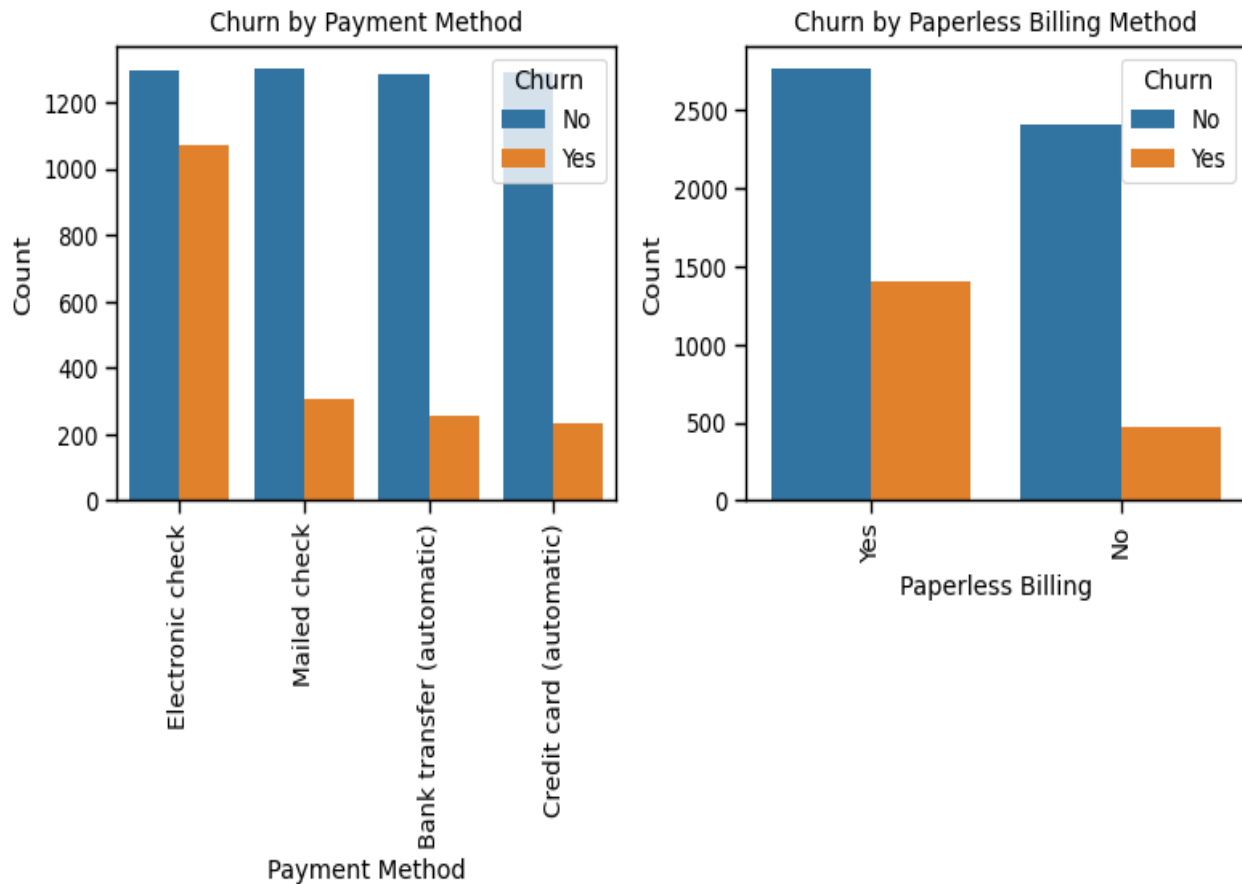
Key Findings "Billings & Churn":

1. **Churn by Payment Method**
   - High Churn with Electronic Checks: Customers paying via Electronic check are more likely to churn compared to those using other payment methods.
   - Lower Churn with Automatic Payments: Bank transfer and Credit card automatic payments have the lowest churn rates, indicating greater customer retention.

2. **Churn by Paperless Billing**
   - Increased Churn with Paperless Billing: Customers opting for Paperless Billing are more prone to churn than those receiving paper bills.

Key Takeaways

- **Focus Areas for Retention:** The business should target interventions to improve satisfaction among `Electronic check` users and those with `Paperless Billing` to reduce churn.

7. According to Customers Demographics:
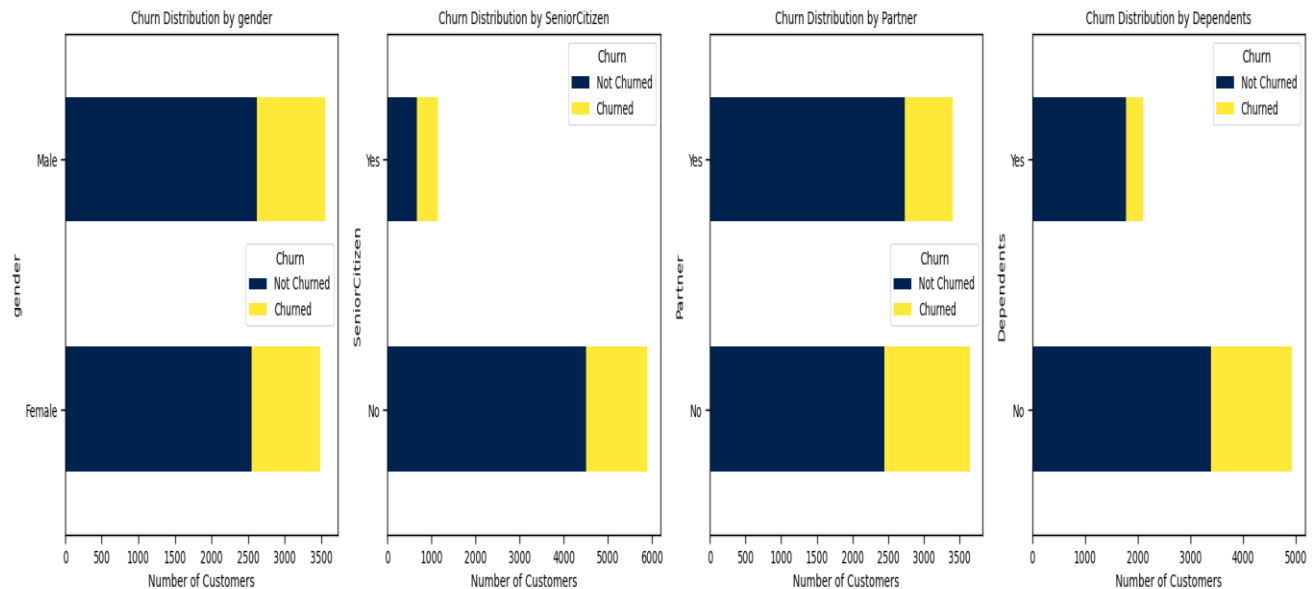
Key Findings:

1. **Gender:**
   - The churn rate is relatively similar for both Male and Female customers.

2. **SeniorCitizen:**
   - This suggests that older customers may be more likely to leave the service.

3. **Partner:**
   - Customers who do not have a partner exhibit a higher churn rate compared to those who do.
   - This could imply that having a partner may contribute to customer retention.

4. **Dependents:**
   - Customers without dependents have a higher churn rate than those with dependents.
   - This suggests that customers with dependents might find the service more valuable or necessary, leading to lower churn.

Insights:

- Understanding the needs of customers with dependents and catering to them could help maintain this customer segment.
- Consider offering tailored services or incentives for senior customers to improve retention.

8. Customer Distribution by Streaming Movies, streaming TV and Churn Status
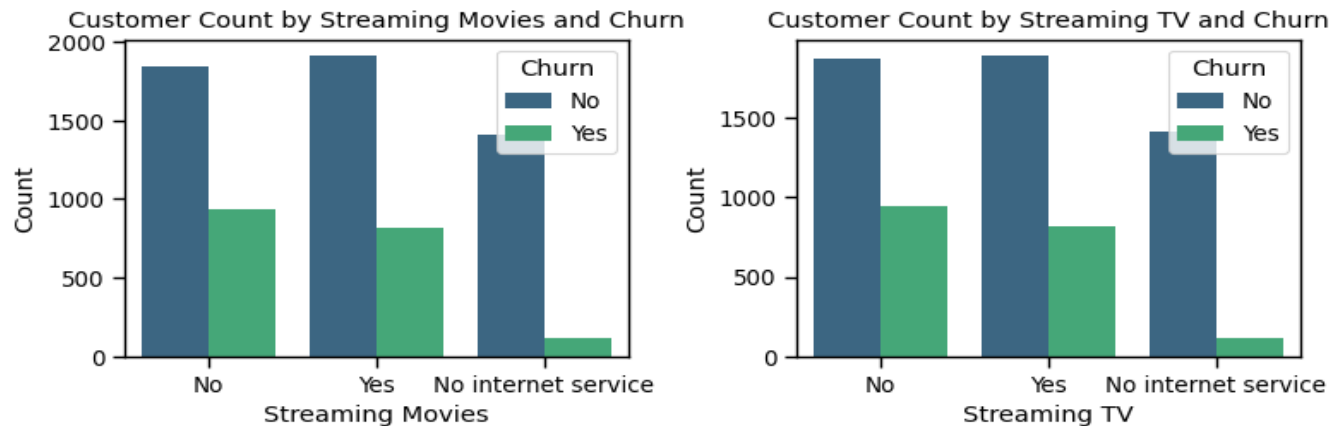
Key Findings:

- **Churn and Streaming Services:**
  Customers who do not use streaming services for movies or TV are less likely to churn compared to those who actively stream. This indicates that streaming service engagement could be linked to higher churn rates.

- **No Internet Service Impact:**
  Customers without internet service show significantly lower churn rates, suggesting that the absence of these services may create more stability in customer retention. This could be due to fewer options for entertainment leading to fewer triggers for dissatisfaction.

- **Similar Churn Patterns for Movies and TV:**
  The trends in both movies and TV show that customers who stream tend to churn more, while those who don't use these services or have no internet service are more likely to stay with the provider.

Insights:

- **Improve Streaming Services**: Enhance content quality and personalization to reduce churn among active streaming users.
- **Upsell to Non-Streaming Customers**: Offer streaming services with clear value to customers who don't currently use them, without compromising their retention.
- **Enhance Internet Service Value**: Bundle internet services with other stable services to attract customers who currently do not use internet services.
- **Tailor Retention Strategies**: Develop targeted strategies for streaming customers based on their specific needs and preferences to reduce churn.

9. Distribution of TotalCharges per Contract type, with hue of Churn

Key Findings from Box Plot of TotalCharges by Contract Type and Churn:

1- High Churn in Month-to-Month Contracts:
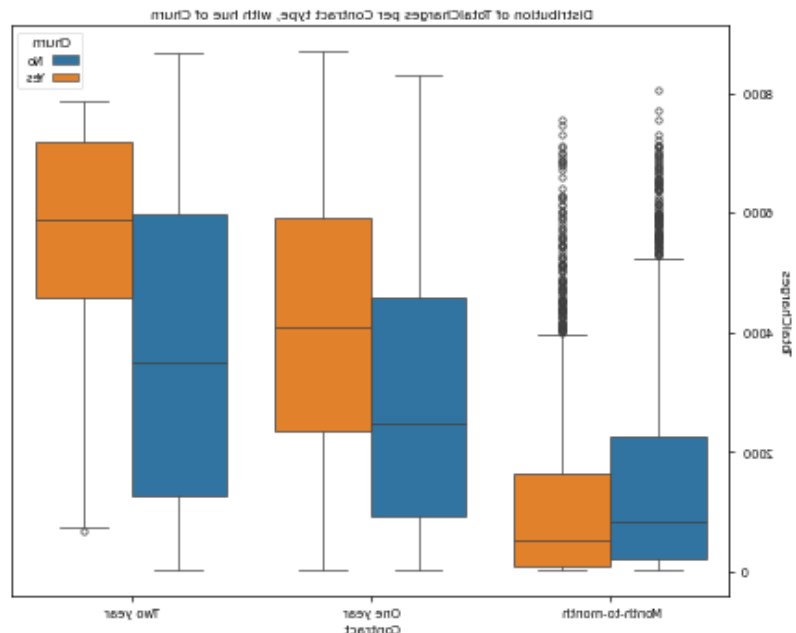
2- Moderate Churn in One-Year Contracts:
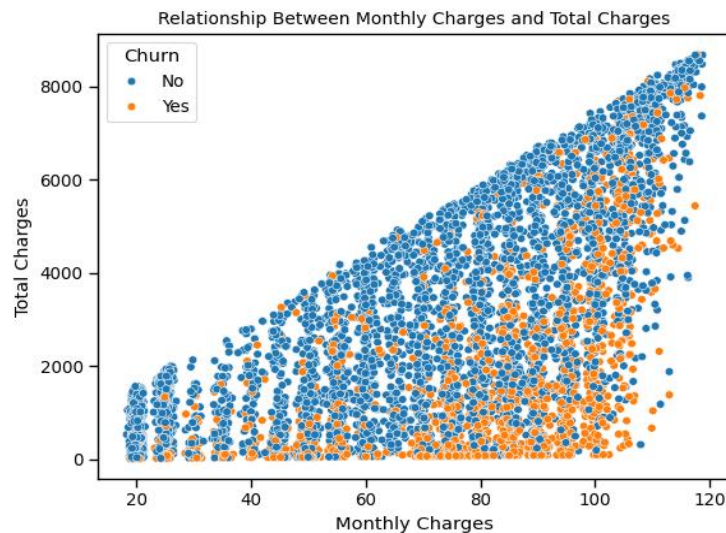
3- Low Churn in Two-Year Contracts:

InShort:

Longer contract terms significantly reduce churn and are associated with higher customer spending, while customers on month-to-month contracts are more prone to churn and tend to generate lower total charges.

Relationship Between Montly & Total charges with the hue of Churn:

- A clear linear relationship is observed between monthly charges and total charges, suggesting that customers with higher monthly charges .accumulate higher total charges over time.

- Churned customers (orange dots) are distributed throughout the range of charges but are more concentrated in the lower range of total charges, indicating that customers with lower total charges might be more prone to churn.

- The spread of churned customers among various levels of monthly charges suggests that while cost may be a factor, other variables could influence churn behavior.

10 . Correlation of features with Churn

Key Findings:

1- **Strong Positive Correlation with Churn:**

- Fiber Optic Connection: Customers using fiber optic services have a high tendency to churn, suggesting dissatisfaction with this service or competition offering similar services.

- Electronic Payment Methods: Users opting for electronic check payments also show a strong positive correlation with churn, which may indicate a need to explore possible reasons such as billing convenience or issues with the payment method.
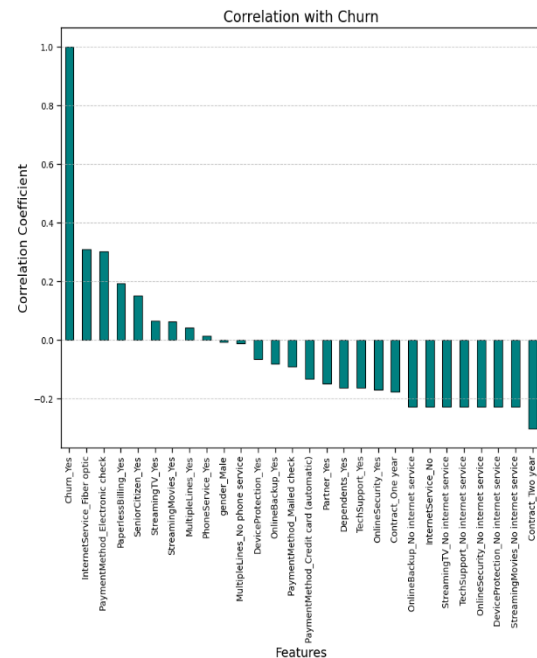
2- **Moderate Positive Correlation with Churn:**

- Billing-Related Services: Billing-related features such as "Billing_Yes" and "Senior Citizen_Yes" also demonstrate positive correlations, indicating that certain customer demographics and service setups might be linked to a higher churn rate.

- Streaming Services: Customers subscribed to premium services like movies or online streaming also show a moderate positive correlation with churn. This could imply that certain value-added services do not meet customer expectations.

3- Negative Correlation with Churn:

- Contract Terms: Longer-term contracts such as "Two-year" contracts are strongly negatively correlated with churn, indicating that customers who commit to long-term plans are more likely to stay. This suggests that offering incentives for long-term contracts could reduce churn rates.

- Dependents and Partners: Customers who have dependents or partners show a slight negative correlation with churn, implying that individuals with family responsibilities may have higher retention rates.

4- **Neutral Factors**:

- Phone Services and Additional Features: Basic phone services and other add-ons such as tech support show minimal correlation with churn, indicating these are not significant factors in a customer's decision to leave.

10. Correlation between features

Key Findings from Correlation Heatmap (Numeric Features)

Correlation Coefficient Heatmap (Numeric Features)

- **Strong Positive Correlation Between Tenure and TotalCharges:**

    The correlation coefficient between "tenure" and "TotalCharges" is 0.83, indicating a strong positive relationship. This suggests that customers who have been with the company longer (higher tenure) tend to have higher total charges, likely due to the accumulation of charges over time.
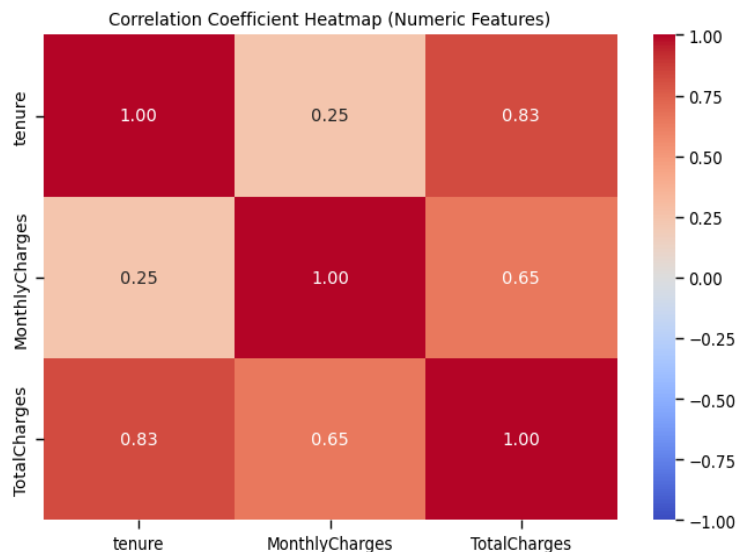
- **Moderate Positive Correlation Between MonthlyCharges and TotalCharges:**

    The correlation between "MonthlyCharges" and "TotalCharges" is 0.65, showing a moderate positive correlation. This suggests that customers with higher monthly charges tend to accumulate higher total charges over time.

- **Weak Positive Correlation Between Tenure and MonthlyCharges:**

    The correlation between "tenure" and "MonthlyCharges" is 0.25, indicating a weak positive relationship. This suggests that while customers with longer tenure tend to have slightly higher monthly charges, the relationship is not as strong as with total charges.

Statistical Analysis Of NUMERIC FEATURES:

	TENURE	MONTHLY CHARGES	TOTALCHARGES
COUNT	7043.000000	7043.000000	7043.000000
MEAN	32.371149	64.761692	2281.916928
STD	24.559481	30.090047	2265.270398
MIN	0.000000	18.250000	18.800000
25%	9.000000	35.500000	402.225000
50%	29.000000	70.350000	1397.475000
75%	55.000000	89.850000	3786.600000
MAX	72.000000	118.750000	8684.800000

Evaluating the Models:

Classification Report & Confusion Matrix:

Logistic Regression Evaluation:				
Accuracy: 82%				
	precision	recall	f1-score	support
No	0.86	0.90	0.88	1036
Yes	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

k-NN Confusion Matrix
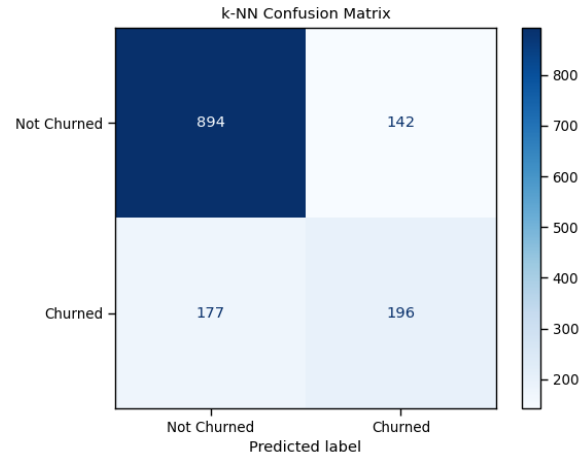
Decision Tree Evaluation:				
Accuracy: 70%				
	precision	recall	f1-score	support
No	0.80	0.80	0.80	1036
Yes	0.45	0.46	0.45	373
accuracy			0.71	1409
macro avg	0.63	0.63	0.63	1409
weighted avg	0.71	0.71	0.71	1409

Decision Tree Confusion Matrix
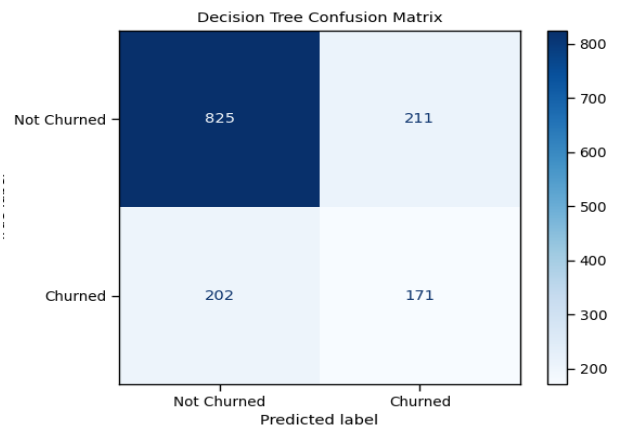
Random Forest Evaluation:				
Accuracy: 79%				
	precision	recall	f1-score	support
No	0.83	0.91	0.87	1036
Yes	0.65	0.49	0.56	373
accuracy			0.80	1409
macro avg	0.74	0.70	0.71	1409
weighted avg	0.78	0.80	0.79	1409

Random Forest Confusion Matrix
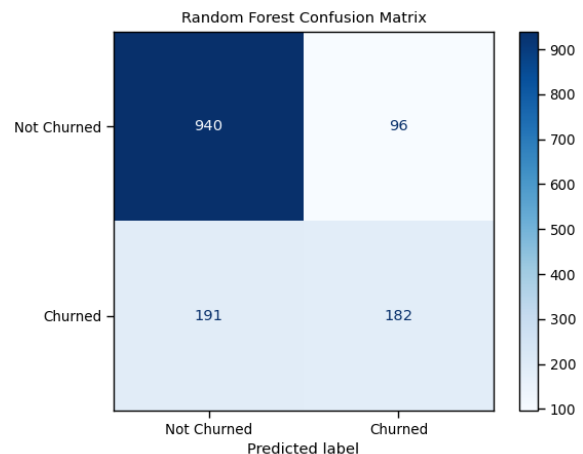
k-NN Evaluation:

Accuracy:77%

	precision	recall	f1-score	support
No	0.83	0.86	0.85	1036
Yes	0.58	0.53	0.55	373
accuracy			0.77	1409
macro avg	0.71	0.69	0.70	1409
weighted avg	0.77	0.77	0.77	1409

k-NN Confusion Matrix
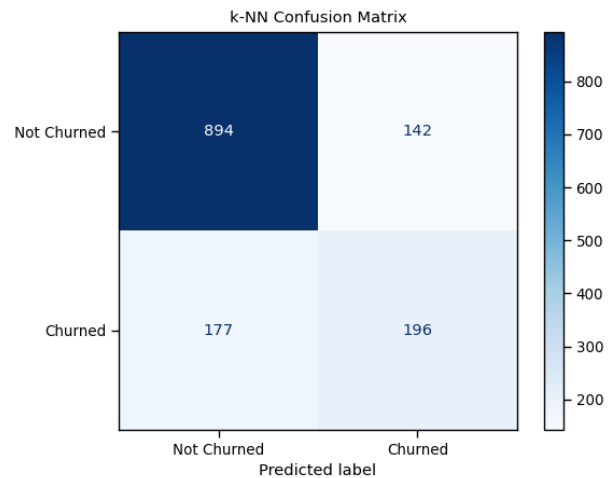
SVM

Evaluation:

Accuracy: 81%

|	precision	recall	f1-score	support
No	0.84	0.92	0.88	1036
Yes	0.70	0.52	0.60	373
accuracy			0.81	1409
macro avg	0.77	0.72	0.74	1409
weighted avg	0.80	0.81	0.80	1409

SVM Confusion Matrix
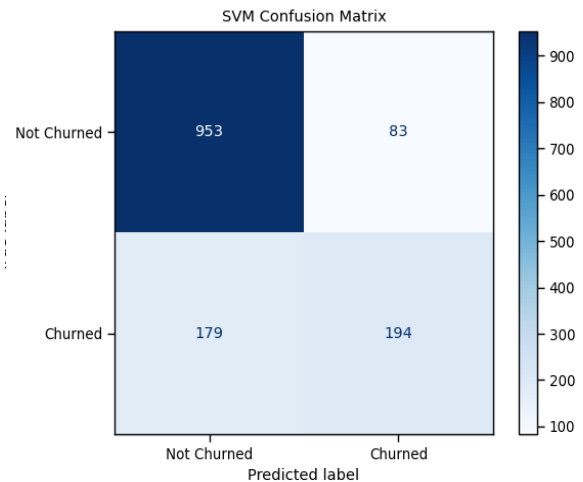
Best Model

Logistic Regression & SVM both are performing well , both have same values for accuracy and F1 score:

- Highest Accuracy: 0.82
- Highest F1-score: 0.88 (weighted avg)
- Balanced Performance: Good precision and recall for both classes

Important Features from Logistic Regression:

Feature	Importance
Contract_Two year	1.393039
tenure	1.338958
InternetService_Fiber optic	0.992466
Contract_One year	0.636088
TotalCharges	0.626486
PhoneService_Yes	0.515758
OnlineSecurity_Yes	0.402374
StreamingMovies_Yes	0.367158
PaperlessBilling_Yes	0.332317
PaymentMethod_Electronic check	0.323456
TechSupport_Yes	0.321197
MonthlyCharges	0.318686
MultipleLines_Yes	0.290413
StreamingTV_Yes	0.266704
Dependents_Yes	0.161684
SeniorCitizen_Yes	0.159905
TechSupport_No internet service	0.148723
StreamingTV_No internet service	0.148723
InternetService_No	0.148723
DeviceProtection_No internet service	0.148723
OnlineBackup_No internet service	0.148723
OnlineSecurity_No internet service	0.148723
StreamingMovies_No internet service	0.148723
OnlineBackup_Yes	0.142697
PaymentMethod_Credit card (automatic)	0.088579
MultipleLines_No phone service	0.083187
Partner_Yes	0.053098
gender_Male	0.052033
PaymentMethod_Mailed check	0.030606
DeviceProtection_Yes	0.006848

Visualizing the feature importances:
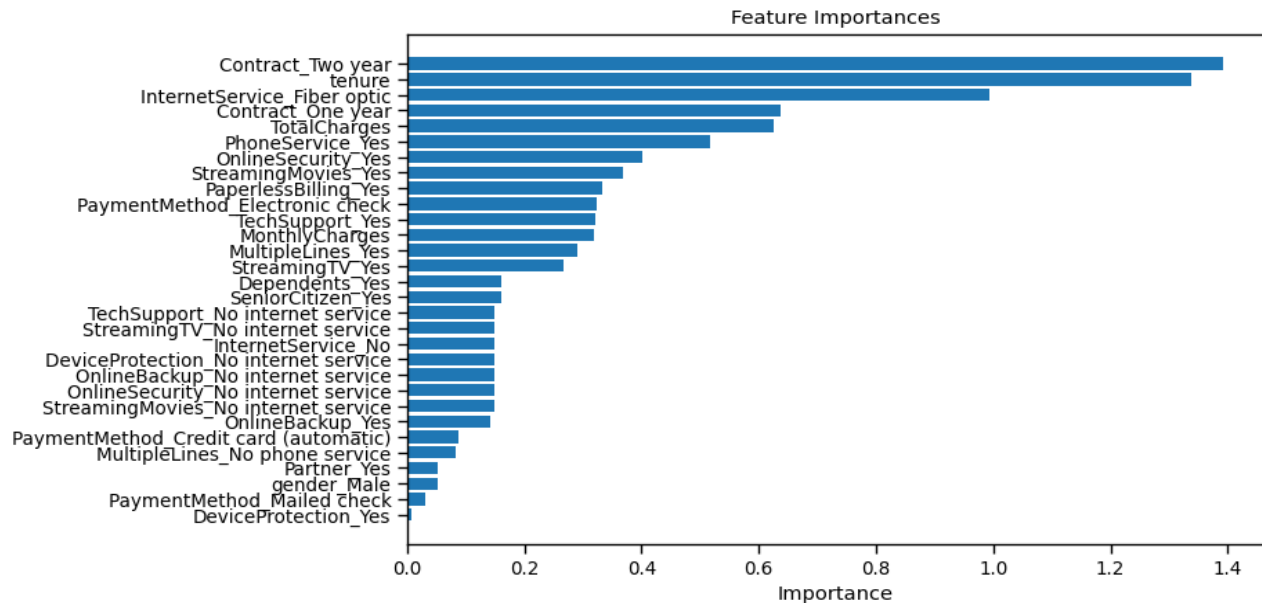
High-Importance Features:

1. **Contract_Two year (1.397018):** This feature is the most important, indicating that customers with a two-year contract have a significant impact on the model's predictions.
2. **tenure (1.362531**): Customer tenure is also highly influential, suggesting that the length of time a customer has been with the company is crucial for predicting churn.
3. **InternetService_Fiber optic (0.996214**): The type of internet service, particularly fiber optic, is a key factor.
4. **TotalCharges (0.653440**): The total amount charged to customers has a strong effect on the model.

Moderate-Importance Features:

5. **Contract_One year (0.637260):** Similar to the two-year contract but slightly less impactful.
6. **PhoneService_Yes (0.514002):** Having phone service affects predictions moderately.
7. **OnlineSecurity_Yes (0.402304**): Whether a customer has online security is moderately important.

Low-Importance Features:

8. **DeviceProtection_Yes (0.007401):** This feature has the least importance, suggesting it has minimal impact on the model's predictions.

Perform Cross-Validation

Cross-Validation Accuracy Scores:

[0.821 , 0.787 , 0.806 , 0.801 , 0.801]

Mean Accuracy: 0.803

Standard Deviation: 0.0111

Interpreting the Performance for our Model

1- **Model Performance:** model has a good average accuracy of about 80.5%, which suggests it performs well on the customer churn dataset.

2- **Consistency:** The low standard deviation (1.14%) indicates that the model's performance is consistent across different folds of the data. This consistency is a positive sign, as it suggests that the model generalizes well.

Key Findings and Recommendations Recap

1. **Churn Predictors**:
   - **Customer Tenure and Contract Type**: The analysis revealed that customers with shorter tenures and month-to-month contracts are significantly more likely to churn. This suggests a need for targeted retention efforts, such as offering incentives for long-term contracts or loyalty programs for newer customers.
   - **Internet Service and Tech Support**: Customers with fiber-optic internet service and those who did not receive tech support also showed higher churn rates. Improving tech support accessibility and customer satisfaction with fiber-optic services could reduce churn.
2. **Demographic Influences**:
   - **Senior Citizens**: Senior citizens have a higher likelihood of churning, possibly due to their lower usage of premium services. Tailoring service packages to this demographic or providing more personalized customer support could mitigate churn in this group.
   - **Monthly Charges**: Higher monthly charges correlate with increased churn. This points to the importance of competitive pricing strategies or bundling offers that provide better value for customers.
3. **Service Usage Patterns**:
   - **Dependence on Streaming Services**: Customers who heavily use streaming services are less likely to churn, indicating that enhancing the quality and variety of streaming services could be a strategic focus for retaining customers.

Targeted Retention Strategies:

1. **Focus Areas**:
   - Develop targeted retention strategies based on identified churn factors.
   - Prioritize retention efforts for customers with high monthly charges, short tenure, and month-to-month contracts.
   - Service Package Enhancements
2. **Service Upgrades**:
   - Encourage customers to transition from DSL to fiber optic services.
3. **Bundling Offers**:
   - Offer attractive service bundles to enhance customer satisfaction and foster loyalty.
   - Pricing Strategies and Incentives
4. **Pricing Adjustments**:
   - Review and adjust pricing strategies, particularly for high-charge customers.
5. **Contract Incentives**:
   - Provide incentives for customers to choose longer contract terms and offer discounts for those with higher cumulative charges.
   - Payment Method Optimization
6. **Process Improvements**:
   - Streamline and improve the electronic check payment process to minimize churn associated with this payment method.
   - Personalized Customer Engagement
7. **Tailored Communication**:
   - Customize communication and promotional efforts based on customer demographics, service usage patterns, and preferred payment methods.

Further Analysis:

Further analysis could include:
- **Advanced Feature Engineering:**
  - Exploring more complex feature interactions and transformations.
- **Model Ensembling**:
  - Combining multiple models to improve prediction accuracy.
- **Deep Learning Techniques:**
  - Utilizing deep learning models to potentially capture more intricate patterns.

Conclusion:

The analysis revealed key factors impacting customer churn, providing valuable insights for developing targeted retention strategies. By addressing the identified areas of concern, the Telco company can significantly improve customer loyalty and reduce churn rate.