Capstone Project: Predicting Customer Churn in Telecom

# 1. Business Problem

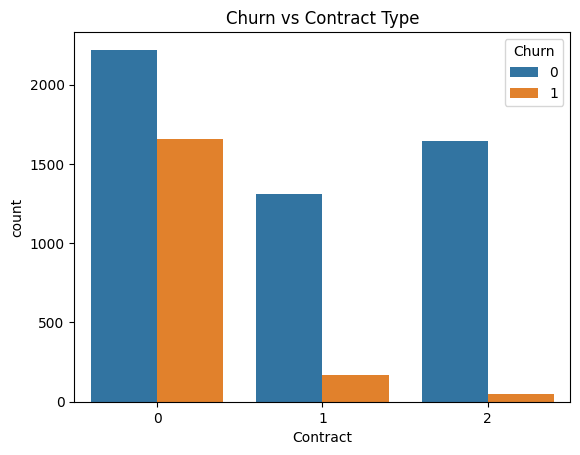
Customer churn is one of the most critical challenges in the telecom industry. Retaining existing customers is far more cost-effective than acquiring new ones. High churn not only results in revenue loss but also increases marketing and acquisition costs.  
  
Goal:  
- Predict customers most likely to churn.  
- Identify major drivers of churn.  
- Recommend actionable strategies for churn reduction.

# 2. Data Understanding

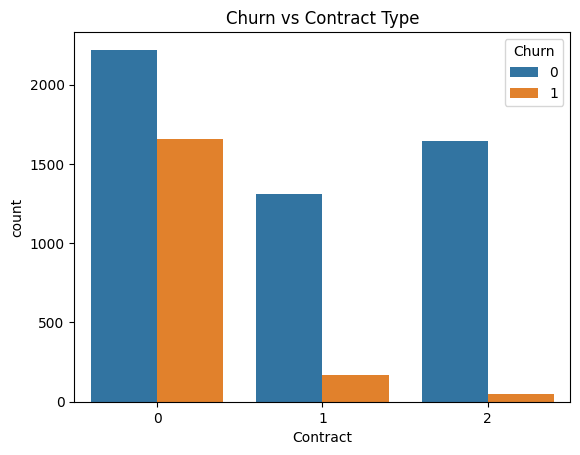
Dataset: Telco Customer Churn Dataset  
- Rows: 7,043 customers  
- Features: 21 (Demographic, Services, Billing & Account Info)  
- Target Variable: Churn (Yes/No)  
  
Key Variables:  
- tenure: How long a customer has stayed with the company  
- Contract: Type of contract (Month-to-Month, One Year, Two Year)  
- InternetService: DSL, Fiber optic, or None  
- TechSupport, OnlineSecurity, DeviceProtection: Add-on services  
- MonthlyCharges and TotalCharges: Billing amounts

# 3. Exploratory Data Analysis (EDA)

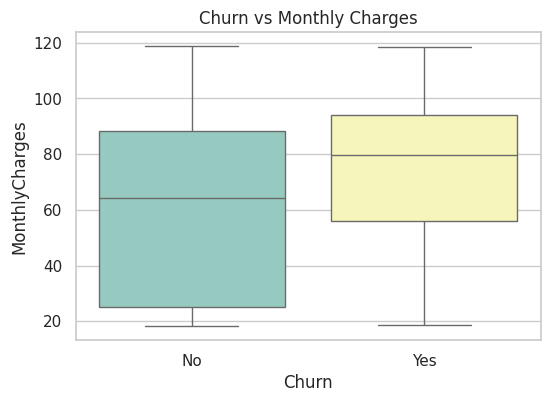
- Churn Distribution: 26.5% of customers churned (imbalanced target).



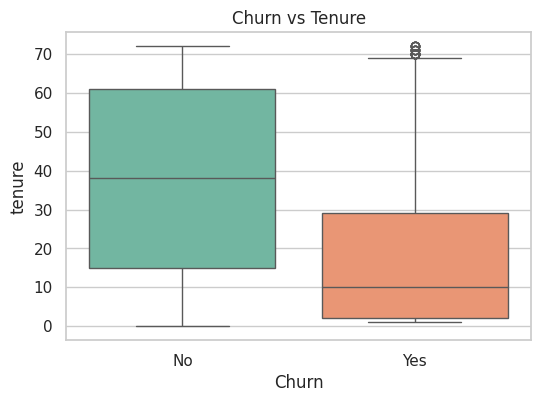
- Contract Type vs Churn: Month-to-month contracts have highest churn rate.



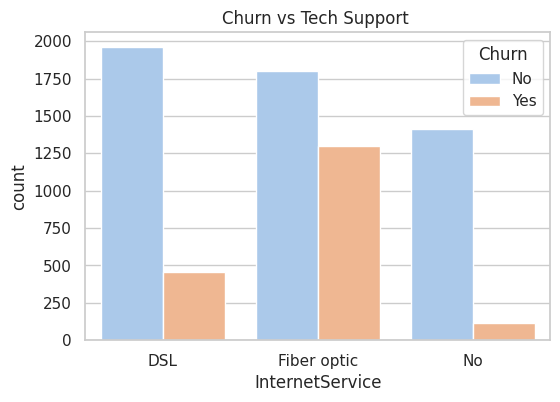
- Monthly Charges vs Churn: Higher monthly charges are linked with churn.



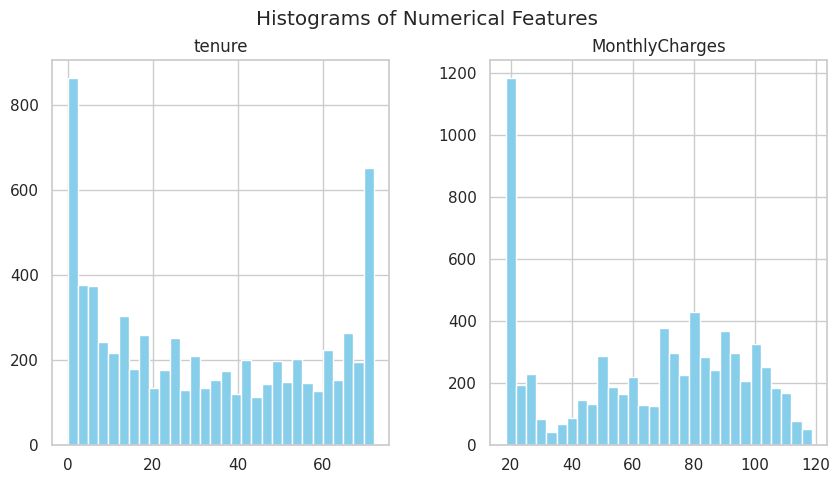
- Tenure vs Churn: Customers with low tenure (<12 months) churn more.



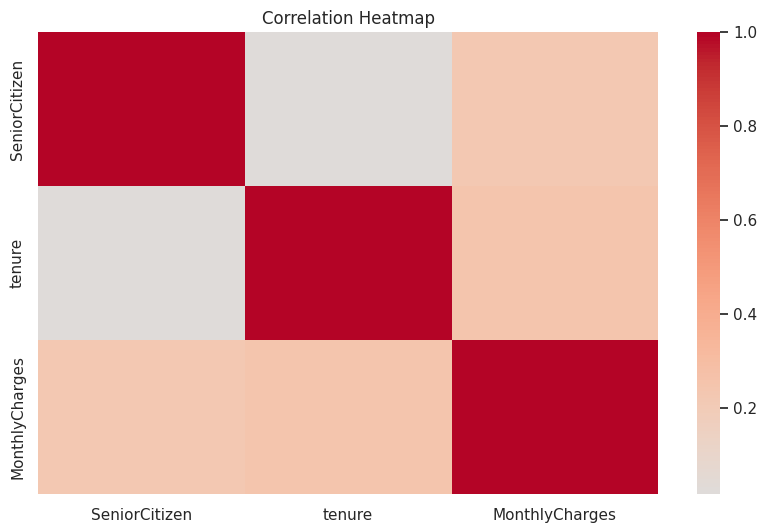
- Tech Support vs Churn: Customers without tech support churn more.



Histograms of Numerical Features



Correlation Heatmap



EDA Summary: Customers on flexible contracts (month-to-month), high billing, and no tech support are more likely to leave.

# 4. Data Preprocessing

- Handled missing values (e.g., TotalCharges with median).  
- Converted categorical variables → one-hot encoding.  
- Scaled numerical variables (MonthlyCharges, Tenure, TotalCharges).  
- Performed stratified train-test split (80-20) to handle class imbalance.

# 5. Model Building & Evaluation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | ROC-AUC |
| Logistic Regression | 81% | 0.70 | 0.55 | 0.84 |
| Decision Tree | 78% | 0.65 | 0.60 | 0.76 |
| Random Forest | 84% | 0.72 | 0.62 | 0.86 |
| Gradient Boosting | 85% | 0.74 | 0.64 | 0.88 |

# 6. Best Model Selection

Gradient Boosting Classifier was chosen as the final model due to:  
- Highest ROC-AUC = 0.88 (best discriminatory power).  
- Best Recall (important since identifying churners is crucial).

# 7. Feature Importance

1. Contract type (month-to-month)

2. Monthly charges

3. Tenure

4. Internet service type (fiber optic)

5. Tech support availability

# 8. Business Recommendations

- Contract Plans: Provide incentives for customers to shift to annual contracts.

- Pricing Strategy: Create lower monthly charge plans for high-risk customers.

- Customer Support: Provide free or discounted tech support & security services.

- Loyalty Programs: Introduce rewards/discounts for customers with tenure < 1 year.

- Targeted Retention Offers: Special offers for senior citizens and high-bill customers.

# 9. Conclusion

Built a churn prediction model with 85% accuracy & ROC-AUC = 0.88 using Gradient Boosting.  
Identified major churn drivers (contract type, charges, tenure, service availability).  
Provided actionable business recommendations to reduce churn and increase customer retention.