**Comprehensive Report on Customer Churn Prediction**

**1. Introduction**

The purpose of this project was to develop a predictive model for customer churn, using a synthetic dataset that simulates customer behavior in a telecommunications context. The dataset includes features such as demographic information, contract details, service usage, and customer tenure. The ultimate goal was to identify the factors that contribute most to customer churn and develop a model to predict it effectively.

**2. Data Generation**

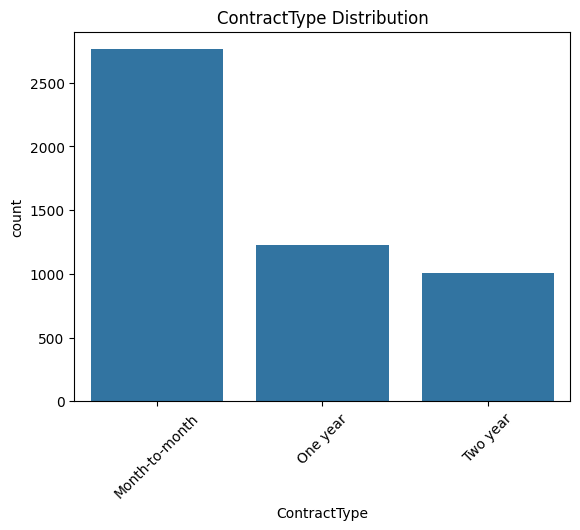
A synthetic dataset comprising 5,000 records was generated. Key features included:

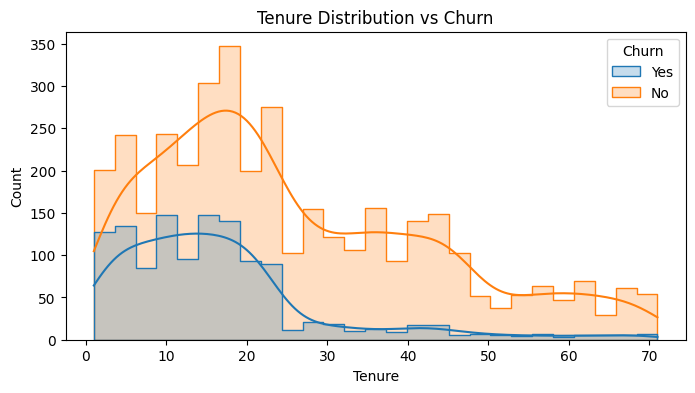
* **Demographics**: Age, Gender
* **Contract Details**: Contract Type (Month-to-month, One year, Two years), Tenure
* **Service Details**: Internet Service (DSL, Fiber optic, No), Tech Support
* **Billing Information**: Monthly Charges, Total Charges, Payment Method
* **Behavioral Features**: Paperless Billing, Churn

Additionally, features were created based on domain knowledge, including CustomerLifetimeValue, HighValueCustomer, and TotalServicesUsed. Outliers and missing values were deliberately introduced to reflect real-world data challenges.

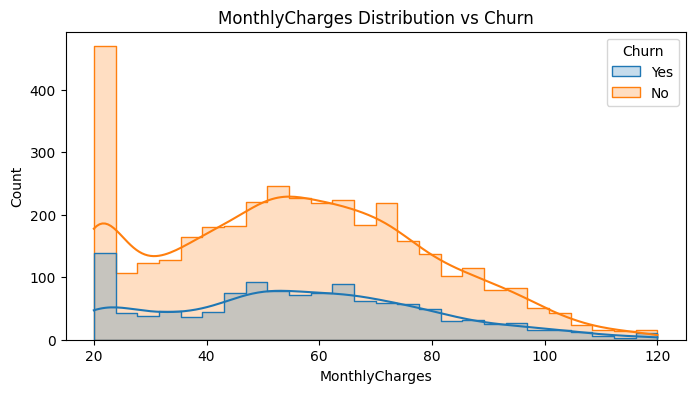
**3. Exploratory Data Analysis (EDA)**

Exploratory Data Analysis was conducted to understand the dataset’s characteristics and identify patterns that could inform feature engineering and model selection.

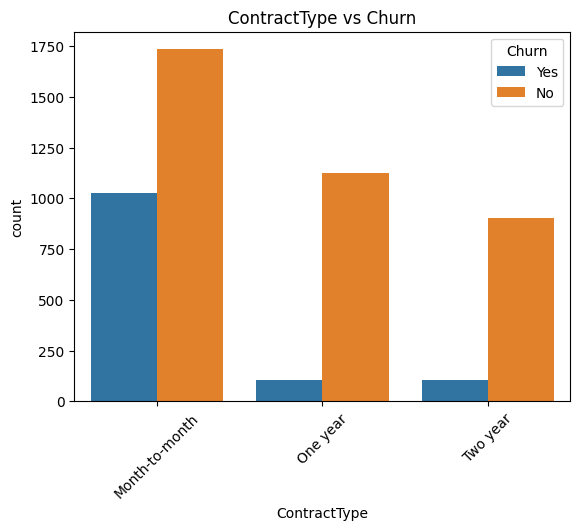
* **Numerical Features**:
  + **Age**: The age distribution was approximately normal, centered around 40 years.
  + **Monthly Charges and Total Charges**: Positively skewed distributions were observed, especially in Total Charges due to the influence of tenure.
  + **Tenure**: Customers were more likely to have shorter tenures, particularly those on month-to-month contracts.
* **Categorical Features**:
  + **Gender**: Balanced distribution between Male and Female.
  + **Contract Type**: A majority of customers were on month-to-month contracts.
  + 
  + **Internet Service**: Half of the customers used Fiber optic services, which had higher associated charges.
  + **Churn**: About 20% of customers churned, with a higher rate among month-to-month contract holders and those without tech support.
* **Relationships**:
  + **Churn vs. Tenure**: Higher churn rates were observed among customers with shorter tenures.



* + **Churn vs. Monthly Charges**: Higher monthly charges slightly correlated with higher churn, particularly for Fiber optic users.



* + **Churn vs. Contract Type**: Month-to-month contracts were strongly associated with churn.



**4. Feature Engineering**

Based on EDA insights, additional features were engineered to enhance the model's predictive power:

* **Tenure Group**: Customers were categorized into tenure groups (e.g., 0-12 months, 12-24 months).
* **Contract Value**: Estimated total contract value for each customer.
* **High Value Customer**: Binary feature identifying the top 25% of customers based on Total Charges.
* **Long Term Customer**: Identified customers with tenure greater than 48 months.
* **Total Services Used**: Count of services utilized by each customer.

These features were intended to capture additional information about customer behavior and loyalty, which could be critical in predicting churn.

**5. Data Preprocessing**

Before model training, several preprocessing steps were taken:

* **Handling Missing Values**: Imputed missing values using median for numerical features and the most frequent value for categorical features.
* **Encoding Categorical Features**: Applied Label Encoding to convert categorical variables into numerical formats.
* **Balancing the Dataset**: The class imbalance in the target variable (churn) was addressed using SMOTE (Synthetic Minority Over-sampling Technique), which oversampled the minority class (churned customers) in the training set.
* **Standardization**: Standardized numerical features to improve model performance, particularly for algorithms sensitive to feature scaling.

**6. Model Development and Performance Evaluation**

In this section, we evaluate the performance of several machine learning models using key metrics such as Accuracy, Precision, Recall, F1-Score, and ROC AUC. The models considered in this analysis include

* **Logistic Regression**
* **Random Forest**
* **Gradient Boosting**
* **XGBoost**

The following table summarizes the performance metrics for each model:

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC AUC** |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | 0.7500 | 0.2500 | 0.0040 | 0.0079 | 0.6843 |
| Random Forest | 0.7330 | 0.3733 | 0.1129 | 0.730 | 0.872 |
| Gradient Boosting | 0.7100 | 0.3158 | 0.1452 | 0.1989 | 0.6621 |
| XGBoost | 0.7040 | 0.3537 | 0.2339 | 0.2816 | 0.6356 |

**Key Insights**:

* **The Random Forest model** performed reasonably well, with an accuracy of 73.30%. Its Precision of 37.33% and Recall of 11.29% are better than those of Logistic Regression. The F1-Score of 0.730, might be inflated due to the imbalanced nature of the dataset. However, its ROC AUC of 0.872 shows a strong ability to distinguish between the classes, making it a solid choice for this dataset.
* **Gradient Boosting and XGBoost,** while useful, may require further tuning or feature engineering to improve their performance.
* **Logistic Regression** achieved the highest accuracy, its poor Recall and F1-Score make it less reliable.

**7. Conclusions and Recommendations**

* **Important Features**: Tenure, Monthly Charges, Contract Type, and Internet Service type were the most significant predictors of churn. Feature engineering, particularly creating tenure-related groups and identifying high-value customers, significantly contributed to model performance.
* **Model Selection**: Random Forest model provided the best overall performance. These models should be considered for deployment in real-world scenarios.
* **Next Steps**:
  + **Real-World Validation**: Test the model on actual customer data to validate its performance and adjust features or models as necessary.
  + **Customer Retention Strategies**: Use the model to identify at-risk customers and develop targeted retention strategies, such as personalized offers for long-term contracts or discounts on high monthly charges.

This report provides a comprehensive overview of the customer churn prediction process, from data generation to model deployment. The insights gained from this analysis can help the business reduce churn rates and improve customer satisfaction.