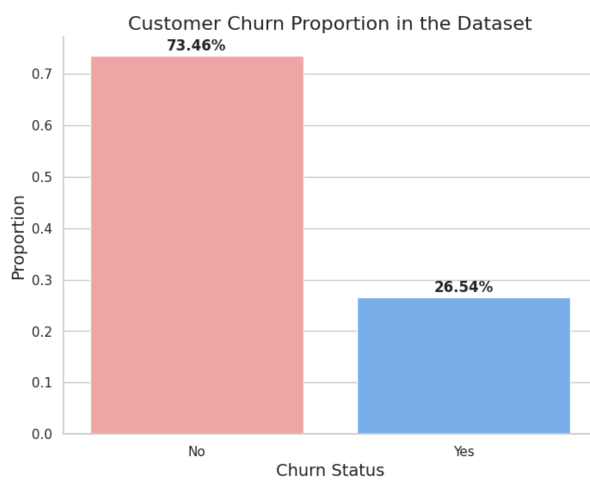
# Project Update 1: Customer Churn Prediction in the Telecom Industry

**Team Members:** Seokhoon Shin, Joshua Nahm, Jiwon Choi, Shinyeong Park, Jaejoong Kim

## 1. Tasks Completed So Far

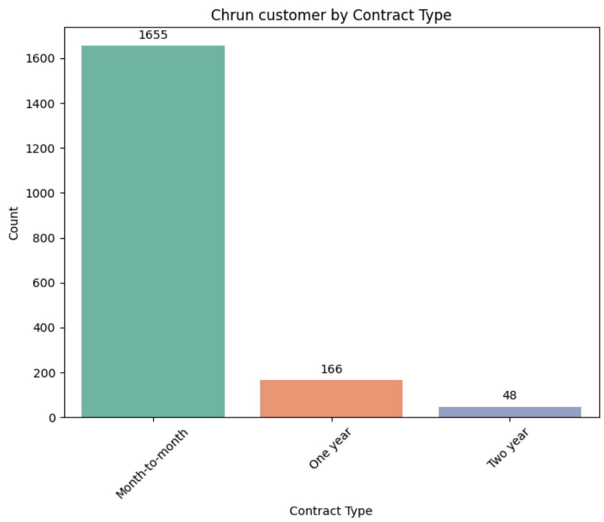
Our primary focus has been on preparing the data and building preliminary models to better understand the patterns associated with churn. Below are the key tasks accomplished:

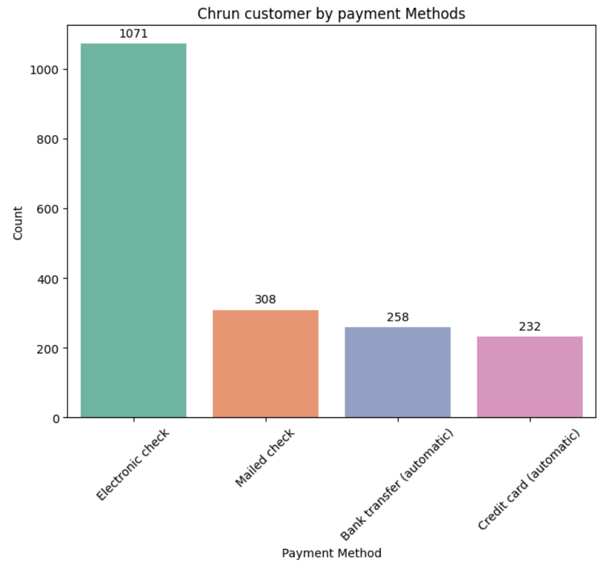
1. **Data Exploration and Cleaning:**
    a. Loaded the **Telco Customer Churn dataset** from Kaggle, which contains 7,043 records and 21 variables related to customer demographics, contracts, and services.
    b. Conducted an initial check for missing data, finding some missing values in the 'TotalCharges' column, which were either imputed or removed to maintain data integrity.
    c. Dropped unnecessary columns such as 'customerID' and replaced irrelevant values (e.g., 'No phone service' was standardized to 'No').
2. **Exploratory Data Analysis (EDA):**
    a. Created visualizations to analyze the relationship between churn and customer features, such as gender, contract type, and payment method.

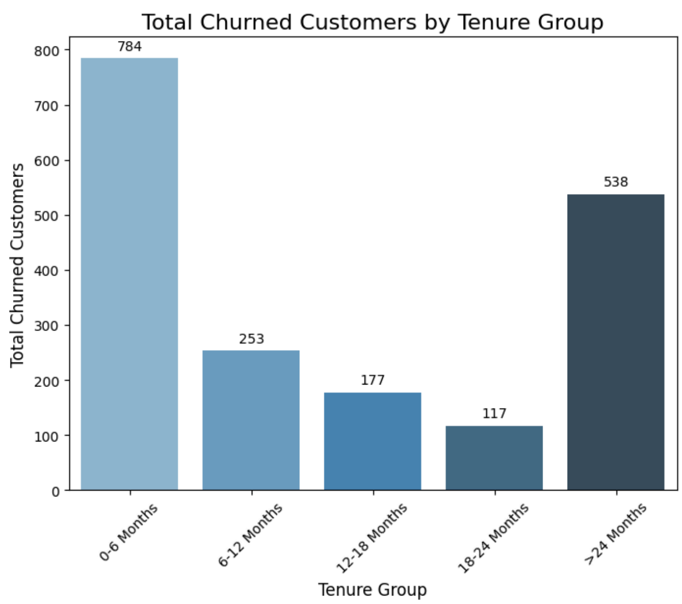



    b. Key findings so far include:
        i. **Month-to-month contracts** are associated with higher churn rates, indicating a potential lack of long-term engagement.
        ii. Customers who use **electronic checks** tend to churn more, possibly hinting at dissatisfaction with the payment method.

iii. Customers with shorter tenures (e.g., **0-6 months**) are more likely to churn compared to long-term customers.




3. **Feature Engineering:**
   a. Encoded categorical variables using **Label Encoding** to make them compatible with machine learning algorithms.
   b. Converted binary responses (e.g., 'Yes/No') into **0 and 1** values for numerical processing.
   c. Normalized features such as 'MonthlyCharges' and 'TotalCharges' to ensure fair model performance.

4. **Model Building and Evaluation:**
    a. We experimented with several machine-learning models:
        i. **Logistic Regression** – Achieved an accuracy of **82%**.
        ii. **Random Forest** – Performed slightly better with an accuracy of **84%**.
        iii. **XGBoost** – Yielded an accuracy of **83%**.
    b. To address the **class imbalance** in the dataset, we applied **SMOTE** (Synthetic Minority Oversampling Technique), which helped ensure that the models were not biased toward non-churning customers.
    c. We evaluated the models using confusion matrices and classification reports to understand their precision, recall, and F1 scores.

## 2. Challenges Faced

1. **Class Imbalance:**
   The original dataset had more non-churning customers than churning ones, making it challenging for the models to predict churn accurately. While **SMOTE** improved balance, it may have introduced some noise into the data, potentially impacting model performance.
2. **Encoding Complex Categorical Features:**
   Encoding variables like **Contract Type** and **Payment Method** using **Label Encoding** was straightforward but could lead to unintended bias. We need to evaluate whether **one-hot encoding** might yield better results for these features.
3. **Feature Correlation:**
   Some features, such as 'MonthlyCharges' and 'TotalCharges', are highly correlated, which might result in **multicollinearity**. This issue could affect model stability and needs to be further investigated.

## 3. Questions and Areas for Further Exploration

1. **Hyperparameter Tuning:** Should we implement **grid search or random search** to fine-tune parameters for models like Random Forest and XGBoost?
2. **Feature Selection:** Would it be beneficial to explore **feature selection techniques** (e.g., PCA or RFE) to remove redundant variables and improve model performance?
3. **Ensemble Learning:** Should we explore additional ensemble models such as **LightGBM** and **CatBoost**, or try **stacking classifiers** to further boost performance?

## 4. Plan for the Next Steps (By November 7)

1. **Hyperparameter Tuning:**
   We will implement **grid search or random search** to optimize the performance of our models, particularly Random Forest and XGBoost.
2. **Finalizing the Best Model:**
   After comparing models based on key metrics (accuracy, precision, recall, and F1-score), we aim to select the best-performing model for the final report.
3. **Handling Multicollinearity:**
   We will further analyze feature correlations to identify and remove highly correlated variables to improve the stability of our models.
4. **Creating Actionable Insights and Reports:**
   We plan to generate insights from the models (e.g., identifying the top predictors of churn) and compile these findings into a **dashboard or report**. This will help stakeholders focus on retaining vulnerable customer segments.

## 5. Conclusion

As we have successfully built multiple models and gained meaningful insights from the data. Moving forward, our focus will be on fine-tuning the models, addressing any data-related challenges, and delivering actionable insights that can support telecom companies to minimize customer churn.