Approach Summary: Predicting Customer Churn

1. Data Preprocessing:

In the initial phase of data preprocessing, the following steps were performed:

Irrelevant columns ('Name', 'CustomerID') were dropped.

The 'Gender' column was encoded as binary values (1 for Male, 0 for Female).

2. Cleaning Phase 1:

Data was grouped by 'Location' to analyze churn rate and usage distribution by location. However, as location didn't appear to significantly affect the data, the column was dropped.

3. Cleaning Phase 2:

Feature engineering was conducted in this phase.

Two new features were created: 'Avg_Monthly_Usage' (calculated as Total Usage divided by Subscription Length) and 'Bill_Per_GB' (calculated as Monthly Bill divided by Avg_Monthly_Usage).

Outliers in 'Avg_Monthly_Usage' and 'Bill_Per_GB' were identified and removed to achieve better data distribution.

4. Correlation Matrix:

A correlation matrix was generated to analyze the relationships between features and the target variable ('Churn').

The top 5 features with the highest absolute correlation to 'Churn' were selected for the final analysis.

5. Train-Test Split:

The dataset was split into training and testing sets (85% training, 15% testing) for model evaluation.

6. Model Selection and Evaluation:

Three different models were trained and evaluated: K-Nearest Neighbors (KNN), Logistic Regression, and Neural Network (MLP).

For each model, cross-validation was used to estimate the model's performance.

The best performing KNN model was selected based on accuracy, achieving approximately 50% accuracy.

Logistic Regression was found to have similar accuracy but was slightly better after fine-tuning the probability threshold.

The Neural Network (MLP) was also evaluated and showed similar accuracy to the other models.

7. Performance Evaluation:

The performance of the selected KNN model was evaluated using metrics such as accuracy, precision, recall, and F1-score.

Given the nature of the data and features, the conclusion is that the dataset doesn't exhibit strong relationships between the provided variables and the 'Churn' target. Therefore, the prediction accuracy remains around 50%, which is roughly equivalent to random guessing.

8. Conclusion:

In conclusion, despite the efforts in preprocessing, feature engineering, and model selection, the dataset's inherent nature seems to limit the ability to predict customer churn using the provided variables. The models' accuracy aligns with random chance, indicating that other unobserved factors might be more influential in predicting churn.