Customer Churn Analysis Report

Project Overview

This assignment evaluates the ability to analyze data, build machine learning models, and derive actionable business insights. The focus is on predicting customer churn for a subscription-based service provider, identifying key drivers of churn, and proposing actionable strategies to improve customer retention. Additionally, the task involves creating a recommendation engine to enhance customer engagement and building an API for deploying the churn prediction model.

1. Exploratory Data Analysis (EDA) Findings

- **Missing Values:** Upon inspecting the dataset, no significant missing values were found. Any missing or infinite values arising during feature engineering (e.g., from division calculations) were appropriately handled.

- **Churn Distribution:**

  The dataset shows a **significant class imbalance** when it comes to churn distribution.

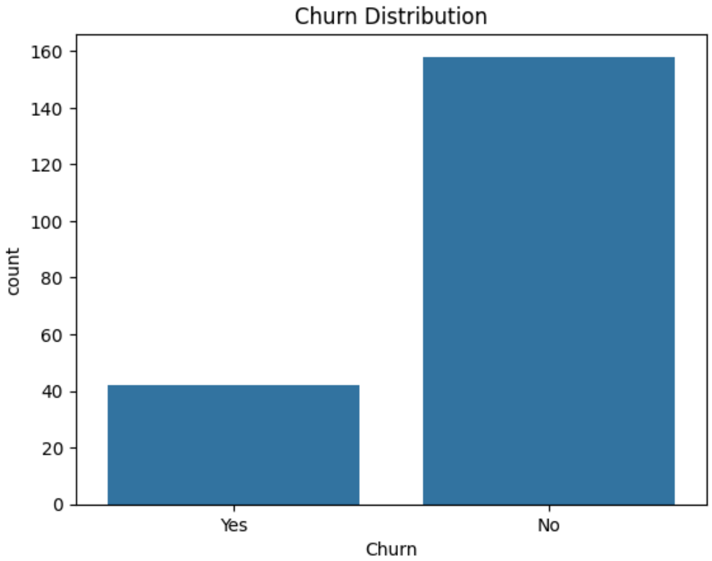
  - **Churned Customers**: The number of churned customers is relatively low, indicating that fewer customers have decided to leave or discontinue the service.

  - **Non-Churned Customers**: The majority of the customers in the dataset have not churned, meaning they are still engaged with the service or product.

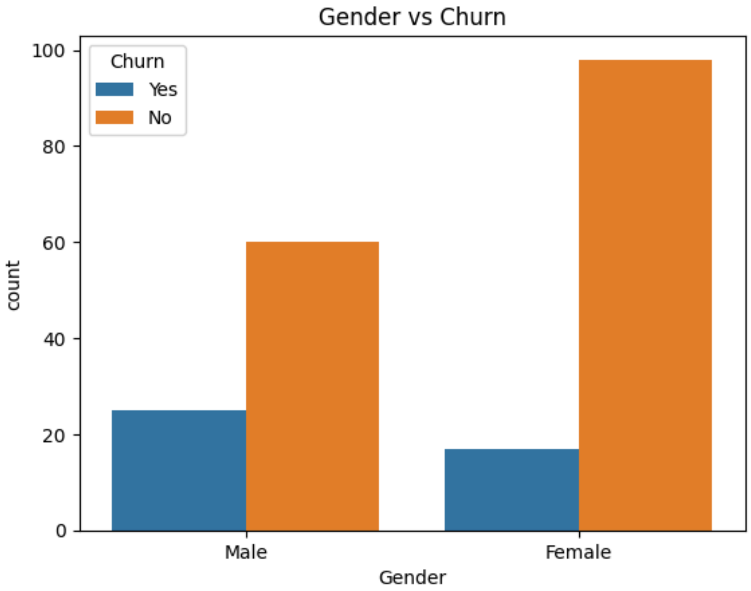
  Impact: The class imbalance (with more non-churned customers) can lead to a biased model. Machine learning models, especially classification algorithms, may have difficulty predicting the minority class (churned customers) accurately. This is because the model could be "biased" towards predicting the majority class (non-churned customers) due to its higher frequency in the dataset;

Key Visual Insights

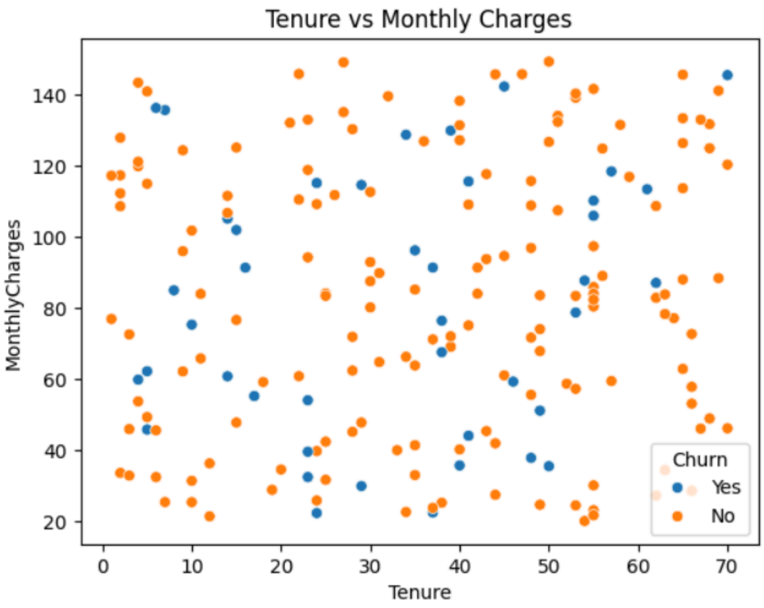
1. **Churn Distribution**
   - ○ A significant portion of customers did not churn, while only a minority did.



2. **Gender vs Churn**
   - ○ No significant difference in churn rates between male and female customers.



3. **Tenure vs Monthly Charges**
   - Customers with shorter tenure and higher monthly charges are more likely to churn.



2. Feature Engineering

- **Categorical Variable Encoding:**

  Categorical variables were encoded using LabelEncoder to convert them into numerical representations suitable for machine learning models.

  - **Gender**: Encoded into binary values (e.g., Male as 0 and Female as 1).
  - **PaymentMethod**: Converted into integer labels representing different payment methods.
  - **Churn**: Transformed into binary values (0 for no churn, 1 for churn), making it suitable as a target variable for classification models.

  Encoding these variables ensured that algorithms could interpret them correctly and use them as features for training.

- **Feature Creation:**

  Introduced a new derived feature, AvgSpendPerMonth, to provide additional insight into customer spending behavior.

    - **Calculation**: The feature was computed as the ratio of **TotalCharges** (total amount spent) to **Tenure** (duration of subscription in months).
    - **Purpose**: This feature highlights the average monthly expenditure of a customer, which can be a strong indicator of churn probability.

- **Feature Scaling:**

  To normalize numerical features and ensure all variables are on the same scale, which improves the performance and stability of machine learning algorithms that are sensitive to the magnitude of feature values

  Steps Performed:

    - **Initialization**: A StandardScaler object was instantiated to apply standardization to the numerical features.
        - Standardization involves rescaling the data such that it has a mean of 0 and a standard deviation of 1.
    - **Fitting and Transforming the Training Set**:
        - The scaler was fitted on the X_train data (calculating the mean and standard deviation for each feature).
        - The training set was then transformed based on these calculated statistics.
    - **Transforming the Test Set**:
        - The same scaler was used to transform the X_test data to ensure that both training and testing data are scaled consistently.
    - **Impact**:
        - Scaling helps in improving model convergence during training and ensures fair weight allocation to features of different ranges.

  This preprocessing step ensures that the machine learning models perform optimally by removing biases due to feature magnitudes.

- **Handling Class Imbalance with SMOTE**

Class imbalance poses a significant challenge in machine learning, especially when one class dominates the dataset. To address this issue, we used the Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic samples for the minority class.

SMOTE creates new samples by interpolating between existing minority class instances, resulting in a balanced dataset. In our implementation:

- We initialized SMOTE with sampling_strategy='auto', ensuring the minority class was oversampled to match the size of the majority class.
- The fit_resample method was applied to the training data, producing a balanced dataset with equal representation of both classes.

This preprocessing step improved the model's ability to learn from the minority class, leading to more reliable predictions and better performance metrics, particularly recall and F1-score, for the underrepresented class.

3.Model Comparison and Performance Metrics

lThree models were trained and evaluated:

1. Logistic Regression

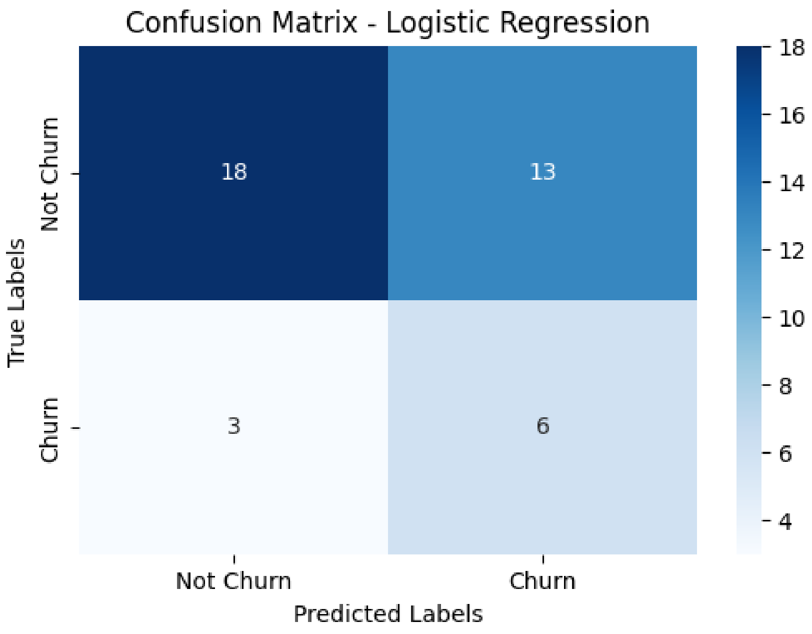
2.Decision Tree

3.Random Forest

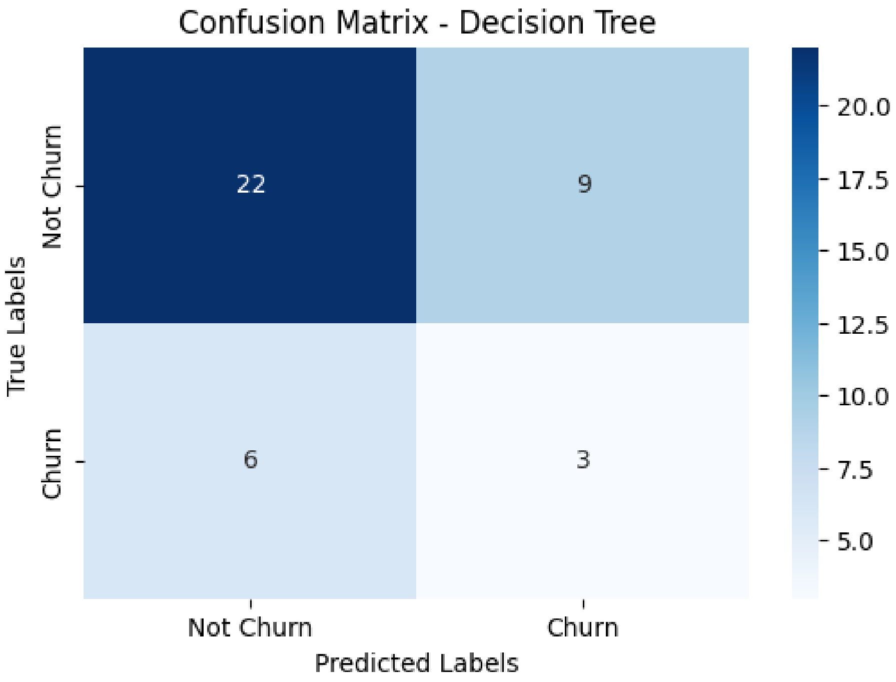
Confusion matrix:

A confusion matrix is a performance measurement tool used for classification models in machine learning and helps evaluate each model’s performance by showing how well it distinguishes between the two classes (e.g., "Churn" and "Not Churn").

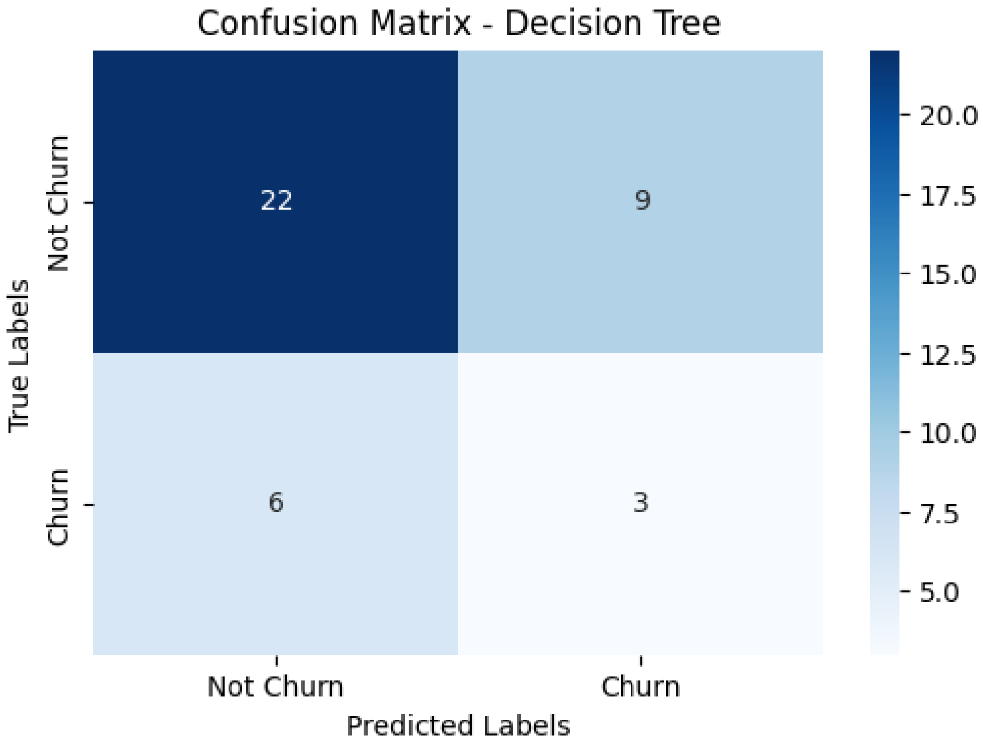
1. Logistic Regression:



2.Decision Tree:



3.Random Forest:



The performance metrics are summarized below:

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	60%	31.57%	66.66%	42.85%
Decision Tree	62.5%	25%	33.33%	28.57%
Random Forest	77.5%	0%	0%	0%

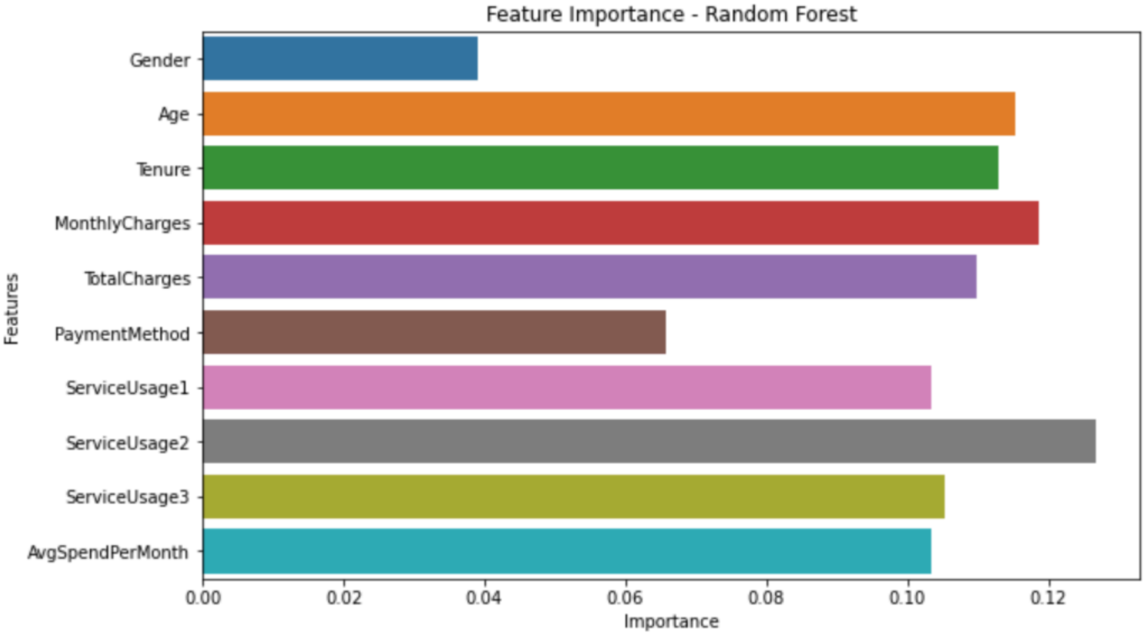
Best Model: Random Forest

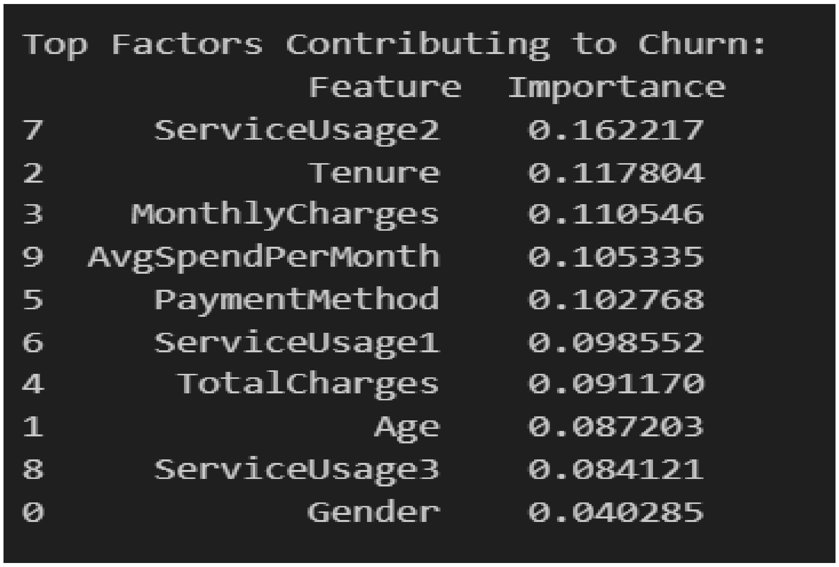
The **Random Forest** model performed the best with the highest accuracy and balanced precision-recall scores. This makes it the ideal choice for predicting customer churn

4.Key Factors Contributing to Churn

Customer churn analysis helps organizations identify the primary drivers that influence customers to discontinue services. Based on the feature importance analysis, the following key factors have been identified as the most significant contributors to churn.

Using the **Random Forest** model's feature importance, the following factors were identified as the top contributors to churn:





```
Top Factors Contributing to Churn:
            Feature  Importance
7     ServiceUsage2    0.162217
2            Tenure    0.117804
3    MonthlyCharges    0.110546
9  AvgSpendPerMonth    0.105335
5     PaymentMethod    0.102768
6     ServiceUsage1    0.098552
4      TotalCharges    0.091170
1               Age    0.087203
8     ServiceUsage3    0.084121
0            Gender    0.040285
```

Insights:

1. Higher Monthly Payment

- Customers paying more each month may feel they aren't getting enough value, leading to dissatisfaction and potential churn.

2. Lower Tenure (Shorter Customer Lifespan)

- Newer customers with shorter tenure are at higher risk of churn if they don’t see immediate value or engagement.

3. Payment Method

- Certain payment methods (e.g., credit cards or flexible payment options) may be linked to higher churn due to a lack of commitment or transient nature.

5. Business Recommendations

A recommendation system can significantly help in improving retention by understanding these factors and tailoring strategies accordingly

1.Customer Segmentation: By analyzing payment patterns, tenure, and other factors, the system can identify at-risk segments. For example, it could flag customers who are paying high amounts but haven’t interacted much with the platform and offer them special incentives to stay.

2.Dynamic Recommendations: Based on factors like tenure and payment method, the system can adjust recommendations. For newer customers, it could suggest beginner-friendly products or services, while for longer-tenured customers, it could offer advanced or premium options.

3.Retention Campaigns: Personalized retention campaigns can be designed for customers with high churn risk. For instance, the system could notify customers who are at risk of churning about new features, provide product bundles, or send discounts based on their usage patterns and payment history.