Bank Customer Churn Prediction

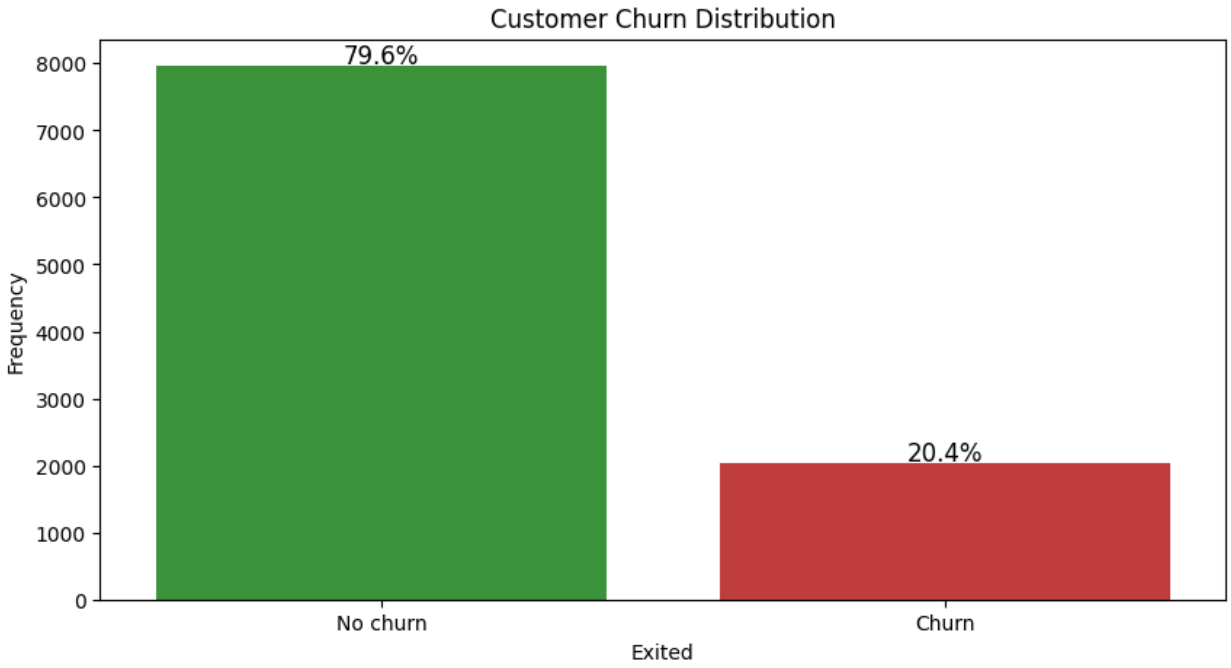
Overview

In the banking industry, it is important to predict and understand when customers might decide to end their relationship with the bank, known as customer churn. When customers leave, it can lead to financial losses and impact the bank's reputation. By identifying customers who are likely to churn, the bank can proactively take measures to retain them and minimize revenue loss. Therefore, the goal of this project is to develop a system that can accurately predict customer churn in order to take proactive steps to retain these customers.

Objective

The main objective of this project is to create a model that can predict which bank customers are likely to leave in the near future. By identifying these at-risk customers, the bank can implement strategies to keep them engaged and satisfied, ultimately reducing churn rates and improving customer retention.

Customer Churn Distribution

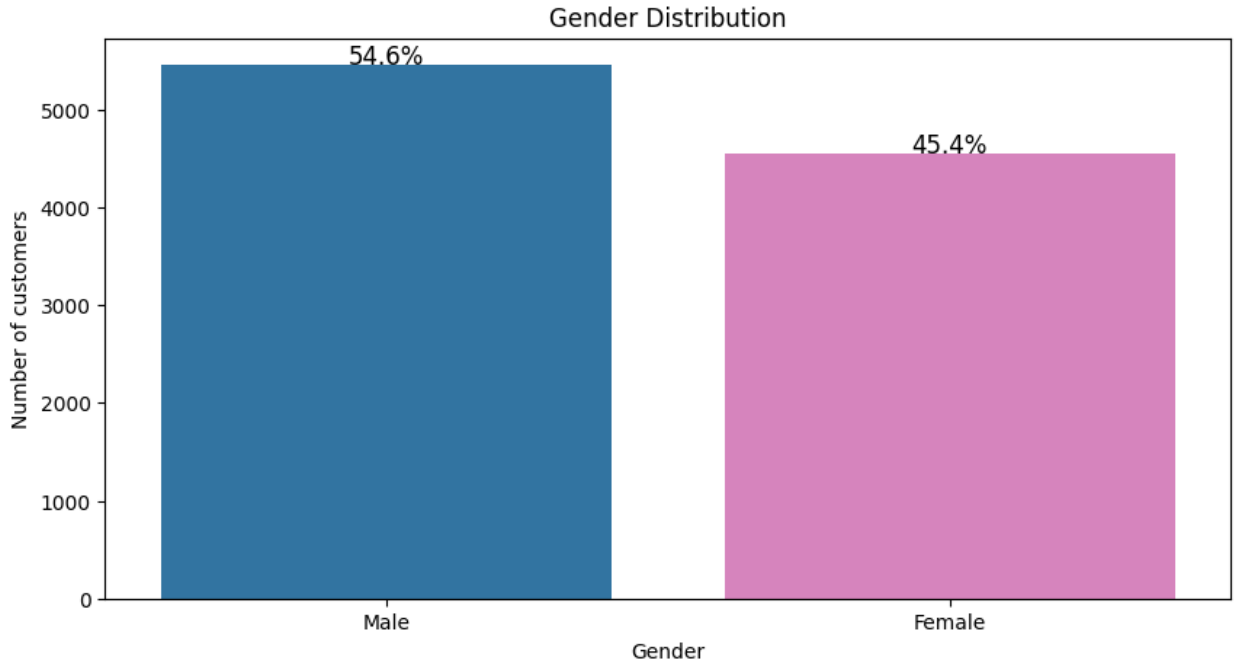


Business Implications:

1. **Retention Focus**:
    - While churn represents only **20.4%**, it is critical to focus on this group since retaining a customer is often more cost-effective than acquiring a new one.
    - A targeted retention strategy can significantly reduce churn and improve revenue.
2. **Impact of Churn**:

- Churned customers might represent lost revenue, making them the priority for predictive modeling and retention strategies.

Gender Distribution



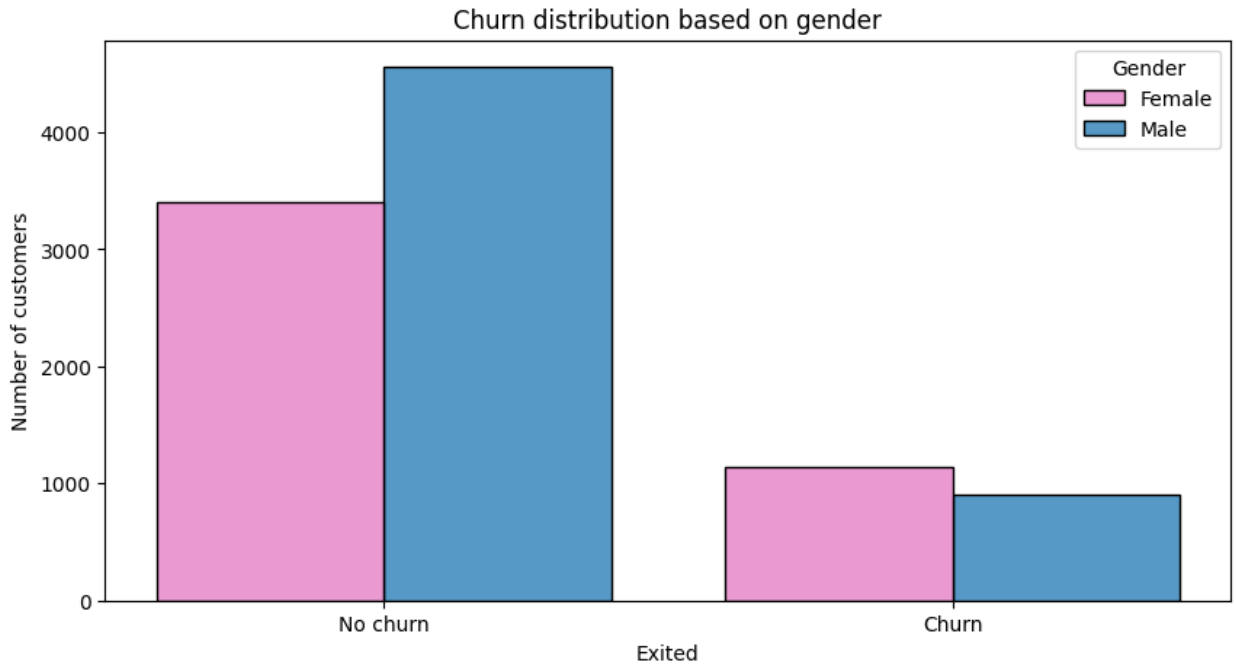
Key Observations:

1. **Male Customers (54.6%)**:
    - The **blue bar** represents male customers, accounting for **54.6%** of the dataset.
    - Male customers form the slight majority in the dataset.
2. **Female Customers (45.4%)**:
    - The **pink bar** represents female customers, making up **45.4%** of the dataset.
    - Female customers are a substantial portion but slightly fewer than males.
3. **Balanced Representation**:
    - The dataset shows a fairly balanced distribution between male and female customers, which reduces the risk of bias in gender-based analysis.

Business Interpretation:

1. **Understanding Gender Trends**:
    - The gender distribution helps businesses understand the composition of their customer base.
    - Insights into gender-specific behavior, preferences, or churn rates can guide targeted marketing strategies.
2. **Segmentation Opportunities**:
    - If subsequent analysis reveals significant differences in churn rates or behaviors between genders, businesses can develop gender-specific retention campaigns.
    - Example: If female customers are more likely to churn, tailored campaigns can address their specific needs or concerns.
3. **Strategic Planning**:
    - This distribution can also guide product development, service improvements, or marketing strategies based on gender demographics.

Churn Distribution Based on Gender



Key Observations:

1. **No Churn**:
    - **Male Customers**:

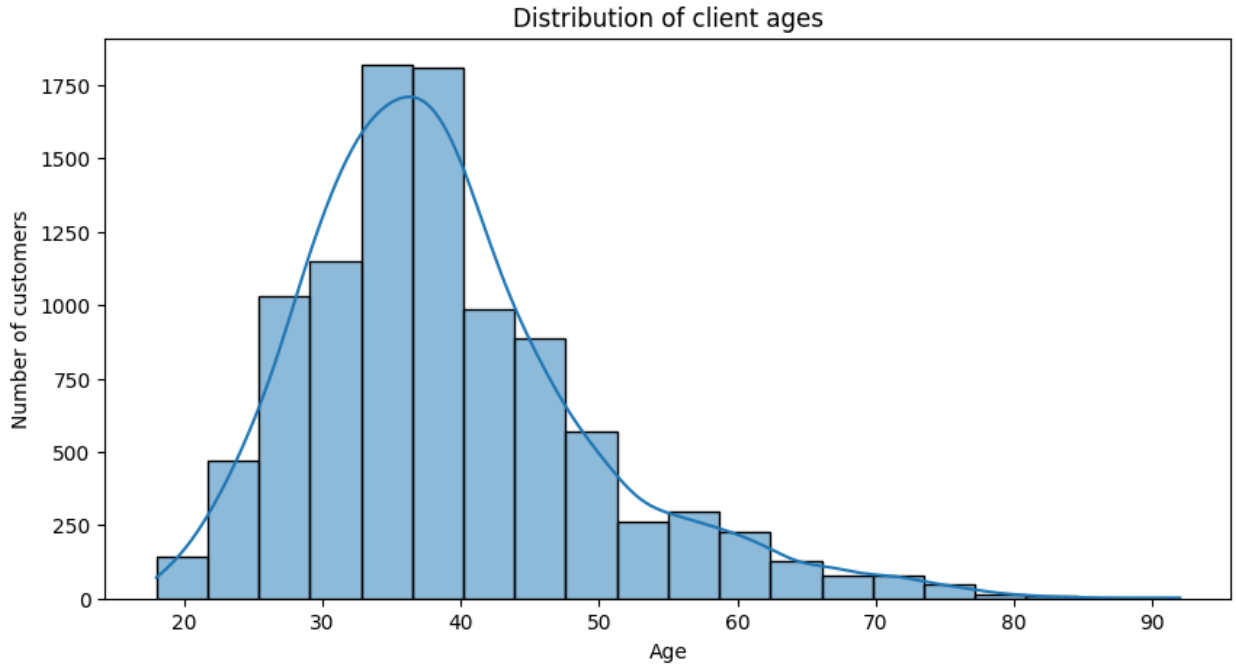
        - Represent the majority in the **No Churn** category, with a higher count compared to female customers.
    - **Female Customers**:
        - Fewer female customers remain active compared to male customers, but they still form a significant portion of the retained customer base.

2. **Churn (Exited)**:
    - **Female Customers**:
        - Female customers have a slightly higher count in the **Churn** category compared to males.
    - **Male Customers**:
        - Although fewer males churn compared to females, they still make up a substantial number in the churn group.

Business Interpretation:

1. **Retention Strategies**:
    - **Female Customers**:
        - Focus on understanding and addressing the specific reasons female customers churn more often.
        - Introduce tailored campaigns to improve engagement and retention.
    - **Male Customers**:
        - Maintain the loyalty of male customers through rewards programs, personalized offers, and improved customer experience.
2. **Actionable Insights**:
    - Identify features that influence churn for each gender (e.g., product preferences, pricing sensitivity, or customer service issues).
    - Address gender-specific pain points to reduce churn.
3. **Segmented Retention Campaigns**:
    - Design campaigns targeting female customers with incentives like loyalty rewards, personalized communication, or enhanced customer service.
    - For male customers, ensure continued satisfaction to maintain the high proportion of retained customers.

Distribution of Client Ages



Key Observations:

1. **Age Range**:
    - The customer ages range from approximately **20 to 90 years**, with the majority of customers concentrated in the **25 to 50-year range**.
2. **Peak Age Group**:
    - The distribution peaks around the age of **35 to 40**, indicating that this age group contains the largest number of customers.
3. **Skewness**:
    - The distribution is **slightly right-skewed**, with fewer customers in the older age groups (above 60).
4. **Density Plot**:
    - The density plot confirms that the bulk of the customer base is in the middle-aged segment, with a gradual decline in the number of customers as age increases.

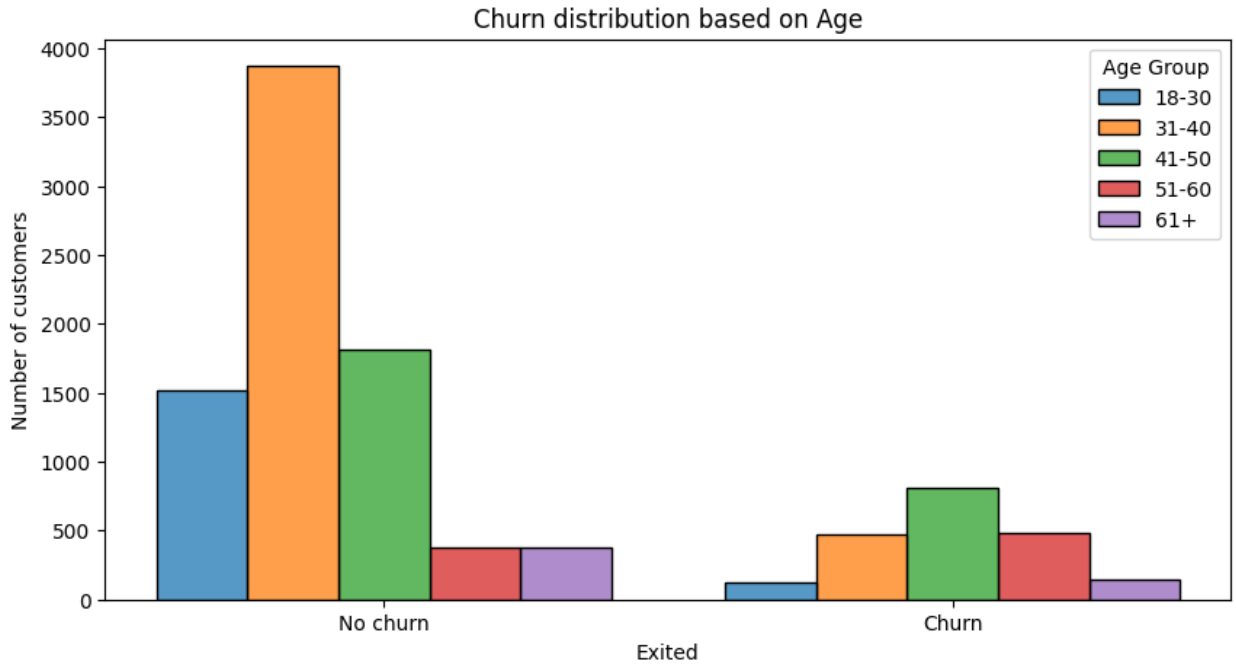
Business Interpretation:

1. **Customer Demographics**:
    - The business primarily caters to middle-aged customers (25–50 years old), who form the largest segment of the customer base.

- The declining numbers in older age groups suggest that this demographic is less engaged or less targeted by the business.

2. **Marketing and Product Strategies**:
    - Focus marketing campaigns on the largest segment (middle-aged customers) to maximize engagement and revenue.
    - Design products or services tailored to the needs and preferences of the 25–50 age group.
3. **Growth Opportunities**:
    - The younger (20–25) and older (60+) segments are underrepresented.
    - Opportunity to expand into these age groups by offering products, services, or campaigns tailored to their needs.

Churn Distribution Based on Age



Key Observations:

1. **Age Group with the Largest Customer Base**:
    - The **31-40 age group (Orange)** has the highest number of customers in both **No Churn** and **Churn** categories.
    - This indicates that this age group forms the core of the customer base.
2. **Churn Trends Across Age Groups**:

- **41-50 age group (Green)** has the highest number of churned customers, followed closely by the **31-40 age group (Orange)**.
- The **18-30 age group (Blue)** and **61+ age group (Purple)** have relatively fewer churned customers, reflecting their smaller customer base.

3. **Older Age Groups (51-60 and 61+)**:
    - Customers in these age groups have lower overall numbers but exhibit noticeable churn, especially in the **51-60 age group (Red)**.
4. **Younger Age Groups (18-30)**:
    - The youngest group has the lowest churn, likely due to their smaller proportion in the overall customer base.

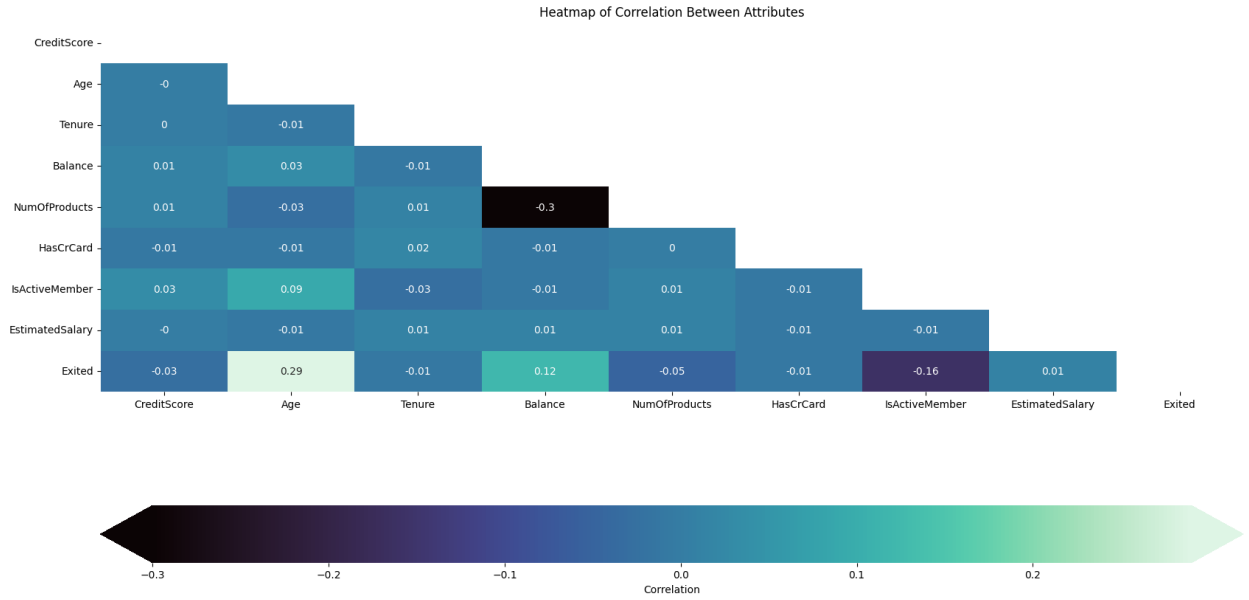
Business Interpretation:

1. **High-Churn Age Groups**:
    - **41-50 age group**:
        - This group shows the highest churn numbers, signaling a need for targeted retention strategies.
        - These customers might be at a life stage where competing priorities or service dissatisfaction lead to churn.
    - **51-60 age group**:
        - Churn in this group may indicate unmet needs or disengagement due to product or service relevance.
2. **Retention Strategies**:
    - **41-50 and 51-60 age groups**:
        - Implement loyalty programs or personalized offers to re-engage these customers.
        - Address their specific needs through targeted communication and product offerings.
    - **31-40 age group**:
        - While this group has high retention, their churn numbers are also significant. Prevent churn by maintaining consistent engagement and satisfaction levels.
3. **Younger and Older Customers**:
    - **18-30 age group**:

- Though churn numbers are low, this group has growth potential. Focus on acquisition and engagement to expand this segment.

- **61+ age group**:
  - Improve relevance and accessibility of products or services to better cater to this smaller but valuable segment.

Correlation Between Attributes



This heatmap represents the **correlation matrix** of various attributes in the dataset. Correlation values range from **-1 to 1**:

- A value close to **1** indicates a strong positive correlation.
- A value close to **-1** indicates a strong negative correlation.
- A value near **0** implies no correlation.

Key Observations:

1. **Age and Exited (Churn)**:
   - **Positive correlation (0.29)**: Age is positively correlated with churn. Older customers are more likely to churn than younger ones.
2. **IsActiveMember and Exited**:
   - **Negative correlation (-0.16)**: Being an active member reduces the likelihood of churn, indicating active engagement plays a role in retention.
3. **Balance and Exited**:

- **Positive correlation (0.12)**: Customers with higher account balances are slightly more likely to churn, possibly due to dissatisfaction or unmet expectations.

4. **Other Attributes and Exited**:
    - Attributes such as **Tenure**, **CreditScore**, and **EstimatedSalary** show very weak correlations with churn, indicating they have limited influence.
5. **Attribute-to-Attribute Correlations**:
    - **Balance and NumOfProducts (-0.3)**:
        - A moderate negative correlation suggests customers with more products tend to have lower account balances.
    - **Age and IsActiveMember (0.09)**:
        - A slight positive correlation indicates older customers are more likely to be active members.

Business Interpretation:

1. **Customer Churn Insights**:
    - Age is a significant predictor of churn, with older customers requiring specific retention strategies.
    - Active engagement is a strong retention factor. Customers who interact with the company more frequently are less likely to churn.
2. **High Balance Customers**:
    - Customers with higher balances are more likely to churn. This group could be dissatisfied with service or finding better alternatives. Special attention is needed for this segment.
3. **Products and Balance**:
    - Customers with more products are less likely to have high balances. This could indicate product diversification strategies may help retain customers.

Strategic Recommendations:

1. **Retention for Older Customers**:
    - Design loyalty programs and campaigns targeting older customers to reduce churn.
2. **Engagement Strategies**:
    - Focus on increasing customer activity through personalized communication, incentives, or rewards.
3. **High Balance Retention**:

- Monitor high-balance customers closely and provide exclusive benefits or concierge services to retain them.

4. **Cross-Selling Opportunities**:
    - Leverage the negative correlation between NumOfProducts and Balance to promote product bundling or cross-selling.

Model Performance Summary:

Model Performance Summary

	Model	Train Accuracy	Train Precision	Train Recall	Train F1-Score	Test Accuracy	Test Precision	Test Recall	Test F1-Score
0	Baseline Decision Tree	1.000000	1.000000	1.000000	1.000000	0.784000	0.454545	0.496183	0.474453
1	Second Decision Tree	1.000000	1.000000	1.000000	1.000000	0.795500	0.480000	0.488550	0.484237
2	Logistic Regression	0.697845	0.702008	0.687539	0.694698	0.717500	0.381868	0.707379	0.495986
3	KNeighborsClassifier	0.938326	0.895964	0.991819	0.941458	0.746500	0.397482	0.562341	0.465753
4	Baseline Random Forest	1.000000	1.000000	1.000000	1.000000	0.859000	0.675079	0.544529	0.602817
5	XGBoost Classifier	0.960903	0.952572	0.970107	0.961260	0.830000	0.561201	0.618321	0.588378

To select the best model from this summary, let's evaluate the trade-offs and criteria based on the problem of **customer churn prediction**. The goal is to balance generalization (test performance) and focus on metrics most relevant to predicting churn.

Key Criteria for Model Selection:

1. **F1-Score:** Since churn prediction is typically a class-imbalanced problem, F1-Score is a better metric than accuracy. It balances **precision** (correct positive predictions) and **recall** (capturing all actual positives).
    - Higher F1-Score is better for ensuring a balance between false positives (predicting churn when it hasn't occurred) and false negatives (failing to predict churn).
2. **Recall (Sensitivity):** Important for churn prediction because missing a customer who might churn (false negatives) is often costlier than wrongly predicting churn (false positives).
3. **Generalization:** Models that overfit (perfect train scores but low test scores) should be avoided.

Summary Analysis of Models:

1. Baseline Random Forest (Best Test F1-Score: 0.60):

- **Pros:**
    - Highest test F1-Score (0.60).
    - Balanced precision (0.67) and recall (0.54), making it effective at identifying churners without too many false alarms.
- **Cons:**

    - Overfits training data (perfect train metrics).
- **Verdict:** A strong contender and best-performing model in this analysis.

2. XGBoost Classifier (Test F1-Score: 0.59):

- **Pros:**
    - Competitive F1-Score (0.59).
    - Slightly better recall (0.61) than Random Forest, which is critical for minimizing false negatives.
    - Less overfitting compared to Random Forest.
- **Cons:**
    - Slightly lower precision (0.56), meaning it predicts more false positives.
- **Verdict:** Close second; a good choice if minimizing false negatives (high recall) is prioritized.

3. Second Decision Tree (Test F1-Score: 0.48):

- **Pros:**
    - Improved over baseline decision tree but still subpar performance.
- **Cons:**
    - Moderate overfitting.
    - Poor test performance compared to Random Forest and XGBoost.
- **Verdict:** Not recommended.

4. Logistic Regression (Test F1-Score: 0.50):

- **Pros:**
    - Simpler model, easier to interpret.
- **Cons:**
    - Underfits both train and test data.
    - Low precision (0.38) and moderate recall (0.70), leading to unreliable predictions.
- **Verdict:** Not suitable for this task.

5. KNeighborsClassifier (Test F1-Score: 0.46):

- **Pros:**
    - Moderate training performance.
- **Cons:**

    - Overfits training data.
    - Poor test performance and not competitive with Random Forest or XGBoost.
- **Verdict:** Not recommended.

Final Recommendation:

- **Primary Choice: Random Forest Classifier**
    - Best balance of F1-Score, precision, and recall for generalization to unseen data.
    - Suitable for both recall-sensitive and balanced needs.
- **Alternative Choice: XGBoost Classifier**
    - Better recall than Random Forest, making it ideal if minimizing false negatives is crucial (e.g., prioritizing retention of all potential churners).