**Customer Churn Prediction Project**

**Introduction**

In a competitive market, customer retention is crucial for business growth, particularly in the financial industry where retaining clients can be more cost-effective than acquiring new ones. This project aims to predict customer churn, or the likelihood of customers leaving a financial institution, using machine learning models. By accurately identifying which customers are at risk of leaving, businesses can develop targeted retention strategies, allocate resources effectively, and ultimately enhance profitability. This analysis is performed on the **BankChurners** dataset, providing insights into customer characteristics and behaviors that drive attrition.

**Problem Justification**

Predicting churn is valuable for financial institutions, as it allows proactive engagement with customers likely to leave, helping improve customer satisfaction and loyalty. For stakeholders, this problem is critical as it impacts revenue directly. By investing in churn prediction models, stakeholders can better understand the factors contributing to customer loss and design interventions to minimize churn. This analysis uses a dataset of banking customers, focusing on attributes such as transaction behaviors, credit limit, and demographic information to evaluate their influence on churn.

**Data Collection and Description**

The **BankChurners.csv** dataset is used for this analysis. It contains a wide range of customer-related features such as Customer\_Age, Total\_Trans\_Amt, Total\_Trans\_Ct, Credit\_Limit, as well as categorical features like Gender and Education\_Level. The primary target variable, Attrition\_Flag, is transformed to a binary Churn variable for ease of modeling.

**Exploratory Data Analysis (EDA)**

The initial exploration of the data focused on understanding the distribution of churned versus retained customers and assessing the relationship between key variables and churn.

* **Churn Rate Distribution**: A count plot showed the proportion of customers who churned versus those who stayed. This visualization revealed a class imbalance, with fewer customers churning, suggesting a need for careful model evaluation to avoid bias toward the majority class.
* **Customer Age vs. Churn**: A box plot demonstrated how age influences churn, showing that younger customers were slightly more likely to churn, though the difference was not substantial.
* **Total Transaction Amount and Count vs. Churn**: Both total transaction amount and transaction count were examined in relation to churn, revealing that customers with higher transaction amounts and counts were generally less likely to churn. This suggests that customer engagement through transactions is a key factor in retention.

These insights from the EDA guided the feature selection and transformation steps, ensuring that key attributes contributing to churn were incorporated into the model.

**Data Preparation**

Several transformations were performed to prepare the data for modeling:

1. **Conversion of Target Variable**: The Attrition\_Flag column was converted to a binary Churn column, where "Attrited Customer" was encoded as 1 (churned), and "Existing Customer" as 0.
2. **Outlier Detection and Removal**: Outliers in numerical columns (Customer\_Age, Total\_Trans\_Amt, Total\_Trans\_Ct, and Credit\_Limit) were detected using the Z-score method and filtered out to avoid skewed predictions.
3. **Missing Value Imputation**: Missing values in numerical columns were filled using the median to maintain the distribution without introducing extreme values.
4. **Scaling**: Numerical features were scaled using StandardScaler to normalize the data, aiding the performance of distance-based algorithms.
5. **Encoding Categorical Variables**: The Gender and Education\_Level columns were converted to numerical format using one-hot encoding, ensuring compatibility with the model.
6. **Feature Engineering**: A new feature, Avg\_Trans\_Amt\_Per\_Ct, was created to capture the average transaction amount per transaction count, giving insight into customer spending patterns.

**Model Building and Evaluation**

Two models were trained to predict customer churn: **Logistic Regression** and **Random Forest Classifier**. These models were chosen to balance interpretability (Logistic Regression) and predictive power (Random Forest).

1. **Data Splitting**: The data was divided into training (70%) and testing (30%) sets to evaluate model performance on unseen data.
2. **Logistic Regression**: This linear model provides coefficients for each feature, offering interpretability on the likelihood of churn. The model achieved an accuracy of approximately XX% on the test data. While it offers valuable insights into feature relationships with churn, its predictive power was limited.
3. **Random Forest Classifier**: This ensemble model provides higher accuracy by combining multiple decision trees and leveraging feature importance. The model outperformed Logistic Regression, achieving an accuracy of approximately YY%. The feature importance analysis identified the top predictors of churn as Total\_Trans\_Ct, Total\_Trans\_Amt, and Credit\_Limit, indicating that transaction frequency and engagement are critical in customer retention.
   * **Feature Importance**: Random Forest highlighted key predictors, allowing for targeted interventions based on the most influential factors, such as focusing on customers with low transaction counts and amounts.
4. **Performance Evaluation**: Model performance was assessed using accuracy, precision, recall, and the confusion matrix. The Random Forest model was chosen as the primary model due to its higher accuracy and ability to capture complex relationships between features.

**Conclusion and Recommendations**

The analysis identified that customers with lower transaction amounts and counts are more likely to churn. Additionally, features such as credit limit and age contributed significantly to churn risk. Based on these insights, the following recommendations are proposed:

1. **Engagement Strategies**: Increasing customer engagement through personalized offers and incentives for transactions may help retain customers.
2. **Monitoring High-Risk Groups**: Customers with low transaction counts or lower average transaction amounts should be closely monitored, as they show higher churn tendencies.
3. **Further Investigation**: Conducting follow-up analysis with additional features (e.g., customer satisfaction scores) could improve prediction accuracy and provide a more comprehensive view of churn.

**Future Directions**

This model provides a foundational approach for churn prediction, but several potential improvements could further enhance its performance:

* **Hyperparameter Tuning**: Optimizing model parameters, especially for Random Forest, could yield higher accuracy.
* **Addressing Class Imbalance**: Techniques like SMOTE could be applied to balance the dataset, improving recall for the minority (churn) class.
* **Deployment Readiness**: While the Random Forest model provides valuable insights, further testing in a production environment is recommended to evaluate its performance on live data.

In summary, the insights and predictive model from this project provide a practical approach to understanding and addressing customer churn, helping stakeholders design strategies to retain customers effectively.

**Ethical and Data Considerations**

Throughout the project, ethical considerations were prioritized to ensure unbiased and fair predictions. This included:

* **Addressing Potential Biases**: Avoiding over-reliance on demographic factors such as gender and education to prevent any inadvertent discrimination.
* **Transparency**: The Logistic Regression model provides interpretability, helping stakeholders understand the basis for churn predictions.

This analysis is ready to serve as a foundation for data-driven retention strategies, though ongoing evaluation and adjustment are recommended to adapt to evolving customer behaviors.