**Bank Customer Churn Dataset**

**Project Overview: Bank Customer Churn Prediction**

**Project Motivation**

The banking industry faces significant challenges due to customer churn, which refers to the loss of customers over time. As a finance and banking graduate, I have a keen interest in understanding customer behavior and implementing strategies to retain customers. The motivation behind this project is to leverage machine learning to predict whether a customer is likely to exit (churn) based on various features such as demographics, account balance, and satisfaction scores. By accurately predicting churn, banks can implement targeted retention strategies, thereby enhancing customer satisfaction and profitability.

**Steps Taken in the Project**

Data Collection:

The dataset was sourced from a bank's customer records, containing information about customer demographics, account details, and behavior metrics. Data Cleaning:

Removed unnecessary columns such as RowNumber, CustomerId, and Surname that did not contribute to the predictive capabilities of the model.

Checked for missing values and duplicates. Missing data can lead to inaccurate predictions, so I ensured to address any inconsistencies before proceeding.

Feature Engineering:

Categorical columns were transformed for better representation. For instance, converting the Gender and Card Type columns into numerical format allows the model to process them effectively.

I created new features that could enhance model performance, such as customer tenure and the total number of products held.

**Data Visualization:**

I utilized visualization techniques to explore the dataset, revealing patterns and relationships among the features. For instance, visualizing churn rates against customer demographics helped identify key segments of customers likely to exit. Graphs and charts were created using libraries like Seaborn and Matplotlib to represent churn rates, customer balances, and satisfaction scores visually.

**Model Selection:**

After analyzing the dataset, I decided to use Logistic Regression for this project. This decision was based on the following considerations:

The nature of the problem is binary classification (churn or not). Logistic regression is effective for such problems and provides probabilities of class membership, allowing for easy interpretation. Given the size of the dataset, logistic regression would efficiently handle the data without requiring extensive computational resources.

**Model Training and Evaluation:**

Split the data into training and testing sets to validate the model's performance. Trained the Logistic Regression model on the training dataset.

Evaluated the model using accuracy, classification reports, and confusion matrices, achieving an accuracy of 0.99. The classification report showed precision, recall, and F1 scores for both classes, confirming that the model performed well in predicting customer churn.

**Insights and Results**

Accuracy:

0.999 means your model correctly predicts 99.9% of the instances in the test set, which is outstanding. Classification Report:

Precision:

For class 0 (not exited): 1.00 (100% of predicted non-churn customers were correct). For class 1 (exited): 1.00 (100% of predicted churn customers were correct). Recall: For class 0: 1.00 (the model identified 100% of actual non-churn customers). For class 1: 1.00 (the model identified 100% of actual churn customers). F1-Score: Both classes have an F1 score of 1.00, indicating a perfect balance between precision and recall.

**Confusion Matrix:**

[[1606 1]

[ 1 392]]

The confusion matrix shows:

True Negatives (TN): 1606 (non-churn customers correctly predicted as non-churn).

False Positives (FP): 1 (a non-churn customer incorrectly predicted as churn).

False Negatives (FN): 1 (a churn customer incorrectly predicted as non-churn).

True Positives (TP): 392 (churn customers correctly predicted as churn).

Considerations: While your model performs exceptionally well, a few points should be considered:

Potential Overfitting: The very high accuracy and perfect precision/recall might indicate overfitting, especially if the dataset is imbalanced. If your dataset has many more non-churn than churn customers, the model may not generalize well to unseen data.

Cross-Validation: To ensure that your model is robust, consider using cross-validation. This will give you a better idea of how the model performs on different subsets of the data.

Evaluate with AUC-ROC: Plotting the ROC curve and calculating the AUC (Area Under Curve) can help you assess how well the model discriminates between the two classes.

Check for Class Imbalance: If there is a significant imbalance between churned and non-churned customers, techniques like SMOTE (Synthetic Minority Over-sampling Technique) can help improve the model's performance on the minority class.

Feature Importance: Assessing which features contribute most to your predictions can provide insights into customer behavior and aid in business strategies.

**Conclusion**

This project has provided valuable insights into customer behavior within the banking sector. By implementing predictive analytics, banks can develop strategies to mitigate churn and foster long-term customer relationships. The skills acquired during this project, including data cleaning, feature engineering, and model evaluation, will be instrumental in my future endeavors in the finance industry.