NAME:T.U.ADARSH

ROLL NO:2023BCY0039.

DATA SCIENCE LAB...

BANK CUSTOMER CHURN PREDICTION REPORT...

1. Problem Statement
Customer churn occurs when clients close their accounts or stop using a company's services. For banks, managing churn is crucial because retaining an existing customer is more cost-effective than acquiring a new one.
The objective of this project is to build a machine learning model that predicts which customers are likely to leave the bank. The study also aims to identify the major factors contributing to churn so that the bank can develop data-driven retention strategies.

2. Dataset Overview
The dataset was obtained from Kaggle (Bank Customer Churn Dataset by Gaurav Topre). It contains 10,000 records and 12 columns, including demographic, financial, and behavioral details of customers. The target variable, **churn**, represents whether a customer left (1) or stayed (0) with the bank.

Feature summary:

- **Demographic:** Country, Gender, Age
- **Financial:** Credit Score, Balance, Estimated Salary
- **Behavioral:** Tenure, Number of Products, Active Member, Has Credit Card
- **Target:** Churn

3. Data Preparation
Initial data inspection confirmed that there were no missing or duplicate values.
Categorical features such as "gender" and "country" were encoded using Label Encoding and One-Hot Encoding.
Numeric features including "balance," "age," "credit_score," and "estimated_salary" were standardized using StandardScaler.
The dataset was split into 80% training and 20% testing sets with stratified sampling to maintain the original churn ratio (approximately 20%).

4. Exploratory Data Analysis
EDA was performed to understand relationships between customer characteristics and churn.
• Around 20% of customers had churned, confirming moderate imbalance.
• Churn rates increased sharply with age; older customers were more likely to leave.
• Customers with higher account balances showed a surprisingly higher churn tendency, suggesting dissatisfaction among premium clients.
• Customers with fewer bank products and low engagement (inactive members) churned more frequently.
• The correlation analysis showed that churn was positively related to age and balance, and negatively related to active membership and number of products.

5. Model Development and Evaluation

Three supervised learning models were developed and compared: Logistic Regression, Decision Tree, and Random Forest.

Model	Accuracy	ROC-AUC	Recall (Churners)	Precision (Churners)	F1-Score
Logistic Regression	0.714	0.777	0.70	0.39	0.50
Decision Tree	0.770	0.838	0.76	0.46	0.57
Random Forest	0.818	0.866	0.71	0.54	0.61

Interpretation

The Logistic Regression model served as a strong baseline with 71.4% accuracy and 0.777 ROC-AUC. It performed fairly well at identifying churners (recall 0.70) but with low precision (0.39).
The Decision Tree improved recall to 0.76 and F1-score to 0.57, but there was a slight risk of overfitting.
The Random Forest model achieved the highest overall performance, with **81.8% accuracy and a ROC-AUC of 0.866**, showing excellent separation between churners and non-churners. It maintained good recall (0.71) while improving precision to 0.54.
Hence, Random Forest was selected as the final model due to its superior generalization and interpretability.

6. Feature Importance Analysis

Feature importance scores from the Random Forest model revealed which variables most strongly influence churn.

Rank	Feature	Importance	Insight
1	Age	0.36	Older customers are significantly more likely to churn.
2	Products Number	0.25	Customers using fewer products are more likely to leave.
3	Balance	0.10	Higher balance customers tend to churn, indicating dissatisfaction among premium clients.
4	Country (Germany)	0.06	German customers showed slightly higher churn than others.
5	Active Member	0.05	Inactive members are more likely to churn.
6	Estimated Salary	0.05	Minor influence; not a strong predictor.
7	Credit Score	0.05	Slight negative influence on churn.
8	Tenure	0.03	Customers with shorter tenure tend to leave more.
9	Gender	0.02	Very small impact on churn.

Rank	Feature	Importance	Insight
10	Country (Spain)	0.01	Least influence.

Key Insight:
Age, product usage, and balance are the most influential churn drivers.
Older customers with high balances and fewer products are more likely to leave.
Banks should prioritize engagement strategies such as loyalty programs and cross-selling additional services to these segments.

7. Final Conclusion
The project successfully built a predictive model to identify customers likely to churn.
Among the three algorithms tested, **Random Forest performed best**, achieving 81.8% accuracy and 0.866 ROC-AUC.
The analysis showed that churn is primarily driven by behavioral and engagement factors rather than purely financial ones.
Customers who are older, hold higher balances, or use fewer bank products are at higher risk of leaving.
The bank should focus on increasing customer engagement, especially among premium and inactive clients, through personalized communication, reward programs, and product bundles.

THANK YOU.....