# Project: Customer Churn Prediction in Banking (Classification) Description:

The banking industry faces a significant challenge in retaining high-value customers, as customer churn can have a direct impact on profitability. This project focused on developing a classification model to predict the likelihood of customers leaving the bank based on various factors such as account balances,

loan repayments, transaction history, and other customer behaviors. By identifying at-risk customers, the bank could intervene proactively with tailored retention strategies. The project contributed to a reduction in churn and improved customer retention using data-driven insights.

## Roles and Responsibilities:

### Data Collection and Connection:

* + - Connected to the bank’s database using Python to retrieve relevant customer data such as account balances, loan repayments, and transaction frequencies.
    - Pulled and cleaned the data from Oracle databases for preprocessing in Python.

### Data Preprocessing:

* + - Cleaned, normalized, and transformed the raw banking data to prepare it for model training.
    - Handled missing values, outlier detection, and performed feature scaling (MinMax Scaling and Standard Scaling).
    - Performed feature engineering by extracting key customer attributes like transaction frequency, loan repayment history, and overall financial behaviors.

### Model Building:

* + - Developed and tuned classification models using Random Forest and XGBoost to classify customers based on their likelihood of churn.
    - Performed feature selection and optimization to ensure that the most relevant variables were being used.
    - Used precision, recall, F1 score, and accuracy to evaluate model performance.

### Handling Data Imbalance:

* + - Addressed imbalanced data using SMOTE (Synthetic Minority Over-sampling Technique) to ensure the model could accurately predict churn for minority classes (churning customers).

### Model Validation and Evaluation:

* + - Applied K-Fold cross-validation to ensure model stability and generalization across different data splits.
    - Tuned hyperparameters using grid search to improve model performance and reduce overfitting.

### Collaboration with Business Teams:

* + - Worked closely with the customer retention team to interpret model outputs and develop actionable strategies for retaining at-risk customers.
    - Delivered detailed insights that helped the bank create targeted campaigns to reduce customer churn.

### Model Deployment:

* + - Deployed the model using Flask to integrate it into the bank’s CRM system, enabling real-time churn predictions and alerts for customer service teams.
    - Provided continuous monitoring of the model post-deployment to ensure its performance with new data.

## Tools and Techniques Used:

* + - **Languages/Packages:** Python, pandas, scikit-learn, cx\_Oracle.
    - **Algorithms:** Random Forest, XGBoost, Classification Models.
    - **Data Balancing:** SMOTE.
    - **Evaluation Metrics:** Precision, Recall, F1 Score, Accuracy, K-Fold Cross-Validation.
    - **Deployment:** Flask.
    - **Model Saving:** Pickle.

## Impact and Outcome:

By implementing this customer churn prediction model, the bank was able to effectively reduce the churn rate. The retention team could intervene with personalized offers and services based on the predicted risk, leading to improved customer loyalty and a noticeable increase in customer retention. The overall accuracy of the model was around 87%, with significant business value added through the use of data-driven strategies.

Although this was an unsupervised learning project, I noticed an imbalance in the customer distribution across different income and credit score groups. This was handled by carefully selecting representative features and refining the clustering parameters to ensure proper segmentation.

## Model Evaluation and Insights:

After creating the clusters, I analyzed the characteristics of each group:

* **Cluster 1:** High-income, low-risk customers – ideal for premium loan offers.
* **Cluster 2:** Moderate-income, moderate-risk customers – good candidates for standard loans with flexible terms.
* **Cluster 3:** Low-income, high-risk customers – potential churners or those who may need small loans with higher interest.
* **Cluster 4:** High-income, high-risk customers – require tailored offers with stricter repayment terms.

## Model Deployment:

After validating the model's effectiveness with various business stakeholders, I proceeded with the deployment:

* 1. **Visualization:** Used Matplotlib and Seaborn to visualize the customer clusters and present the segmentation results to stakeholders.
  2. **Actionable Insights:** Collaborated with the marketing team to translate these segments into actionable marketing campaigns. Each cluster was targeted with personalized loan offers.
  3. **Deployment:** I packaged the model as a pickle file and integrated it with a Flask API. The API was hosted on AWS EC2, allowing real-time customer segmentation for loan offers.

## Results and Impact:

* + - The clustering model improved the targeting of personalized loan offers, which led to a 20% increase in loan approval rates.
    - The marketing campaigns based on these customer segments showed a marked improvement in customer engagement and retention.
    - The model helped optimize the marketing budget by targeting high-value customer segments.

## Model Metrics and Evaluation:

Since clustering models don’t use typical accuracy metrics, I evaluated the model based on:

* + - **Cluster Cohesion (Within-Cluster Sum of Squares).**
    - **Cluster Separation (Between-Cluster Sum of Squares).**
    - **Silhouette Score** to measure how similar customers within each cluster are compared to those in other clusters.