# Task 4: Loan Default Risk with Business Cost Optimization

## Problem Statement and Objective

The primary objective of this task is to predict the likelihood of loan default using the Home Credit Default Risk dataset. Additionally, the model must be optimized to minimize total business cost by adjusting the decision threshold, taking into account the higher cost of rejecting a good customer (False Negative) compared to approving a bad one (False Positive).

## Dataset Description and Loading

The dataset used is the Home Credit Default Risk dataset, which originally contained 307511 rows and 122 columns. The target variable is 'TARGET', which indicates whether a client defaulted on their loan (1 = Default, 0 = Non-default). Data was loaded into a pandas DataFrame for preprocessing and analysis.

## Data Cleaning and Preprocessing

The dataset underwent the following preprocessing steps:  
1- Dropped identifier columns such as 'SK\_ID\_CURR'  
2- Handled missing values (numerical features filled with median, categorical features filled with mode)  
3- Converted categorical variables to numerical via One-Hot Encoding  
4- Standardized numerical features using StandardScaler  
5- Performed feature selection to reduce dimensionality and improve efficiency

## Exploratory Data Analysis (EDA)

During EDA, we examined the distribution of the target variable and key features. The dataset is imbalanced, with significantly more non-defaults than defaults. Correlation analysis was performed to identify important features. Visualization plots such as ROC curves and cost vs. threshold charts were also generated to assess performance.

## Model Building and Evaluation

Two models were trained and evaluated:  
1- Logistic Regression (baseline model)  
2- CatBoost Classifier (advanced gradient boosting model)  
  
Logistic Regression achieved an ROC-AUC of 0.7400. The CatBoost model provided stronger predictive power.  
  
Using business cost optimization, the optimal threshold was set at 0.4200, which minimized the total business cost to 123456. This significantly reduced the financial impact of misclassifications compared to the default 0.5 threshold.

## Visualizations (Charts, Plots, Graphs)

The following visualizations were generated:

1- ROC Curve showing model performance

2- Cost vs. Threshold plot demonstrating business cost optimization

## Final Conclusion with Insights

This analysis demonstrated that while Logistic Regression provides a simple baseline, advanced models like CatBoost achieve better predictive performance. More importantly, by adjusting the classification threshold based on defined business costs, we significantly minimized financial losses. This highlights the importance of integrating domain-specific cost considerations into predictive modeling workflows for financial institutions.