📋 Final Modeling Report – Home Credit Default Risk

## 1. Objective

Predict the probability that a loan applicant will experience financial difficulties using data from Home Credit. This model supports lending decisions by identifying high-risk borrowers with greater accuracy than traditional scoring methods.

## 2. Business Context

In lending, false negatives (approving high-risk clients) are more costly than false positives (denying credit to low-risk clients). Thus, \*\*recall\*\* for the positive class (defaults) is a critical metric, though precision and AUC are also monitored.

## 3. Data Summary

The main dataset contains 307,511 training rows and 122 columns. Six additional relational tables (e.g., credit history, previous applications) were merged and aggregated by `SK\_ID\_CURR`. The final dataset used for modeling includes ~500 engineered features.

## 4. Feature Engineering

Engineered features include:  
- `CREDIT\_TO\_INCOME\_RATIO`, `ANNUITY\_TO\_INCOME\_RATIO`, `EXT\_SOURCES\_MEAN`  
- Aggregates from `bureau`, `installments`, `POS`, `previous\_application`  
- Time delay features, DPD scores, approval ratios  
All new features were tested for correlation, missingness, and predictive power.

## 5. Modeling Strategy

Primary model: \*\*CatBoostClassifier\*\*  
- Handles categorical variables natively  
- Optimized via \*\*Optuna\*\* for hyperparameters: `depth`, `learning\_rate`, `l2\_leaf\_reg`  
- \*\*Stratified K-Fold CV\*\* used (k=5) to account for class imbalance  
- Custom class weights tuned for recall on minority class

## 6. Evaluation Metrics

Best model performance (based on CV):  
- AUC: ~0.80  
- Recall: ~0.74 (threshold optimized)  
- Precision: ~0.32  
- F1 Score: ~0.45  
- Confusion matrix and SHAP plots support results  
Threshold was tuned to maximize recall while controlling false positives.

## 7. Explainability

SHAP summary plots show most important features:  
1. `EXT\_SOURCE\_2`, `EXT\_SOURCE\_3`  
2. `DAYS\_BIRTH`, `DAYS\_EMPLOYED\_PERC`  
3. `CREDIT\_TO\_INCOME\_RATIO`, `ANNUITY\_TO\_INCOME\_RATIO`  
Clients with younger age and lower external scores are higher risk.  
Local explanations via SHAP force plots were used to evaluate specific predictions.

## 8. Final Model Output

The model was retrained on 100% of `application\_train.csv` using the tuned parameters.  
Predictions were generated on `application\_test.csv`, outputting probabilities of default (`TARGET\_PROB`).

## 9. Reproducibility

Environment managed with Conda (`environment.yml`).  
Code runs in sequence:  
1. `1\_EDA\_FeatureEngineering.ipynb`  
2. `2\_Modeling\_CatBoost.ipynb`  
3. `3\_Evaluation\_Thresholding\_SHAP.ipynb`  
4. `4\_TestPrediction\_Submission.ipynb`

## 10. Conclusions

- The final model achieves high recall and reasonable AUC using ensemble methods and rich feature sets.  
- Business stakeholders can adjust the threshold to fine-tune risk appetite.  
- Further improvements possible with temporal validation and more complex models like TabNet or LightGBM ensembles.  
- Model explanations are sufficiently transparent for credit risk environments.