




PREDICTING THE RISK OF DEFAULT ON HOME CREDIT

A DATA SCIENCE APPROACH TO SUPPORTING CREDIT GRANTING DECISIONS



BACKGROUND

Problem



The banking industry needs to accurately predict credit risk. Home Credit faces the challenge of assessing borrowers without a credit history, so it uses a comprehensive dataset to build a predictive model that is more precise than conventional methods.

This project aims to reduce bad debt and expand credit access. Using machine learning methods such as neural networks and ensemble methods, complex data patterns are transformed into intelligence to optimize credit decisions.

PROBLEM STATEMENT

How can we accurately predict the probability of loan default using applicant information and historical credit data?



OBJECTIVES

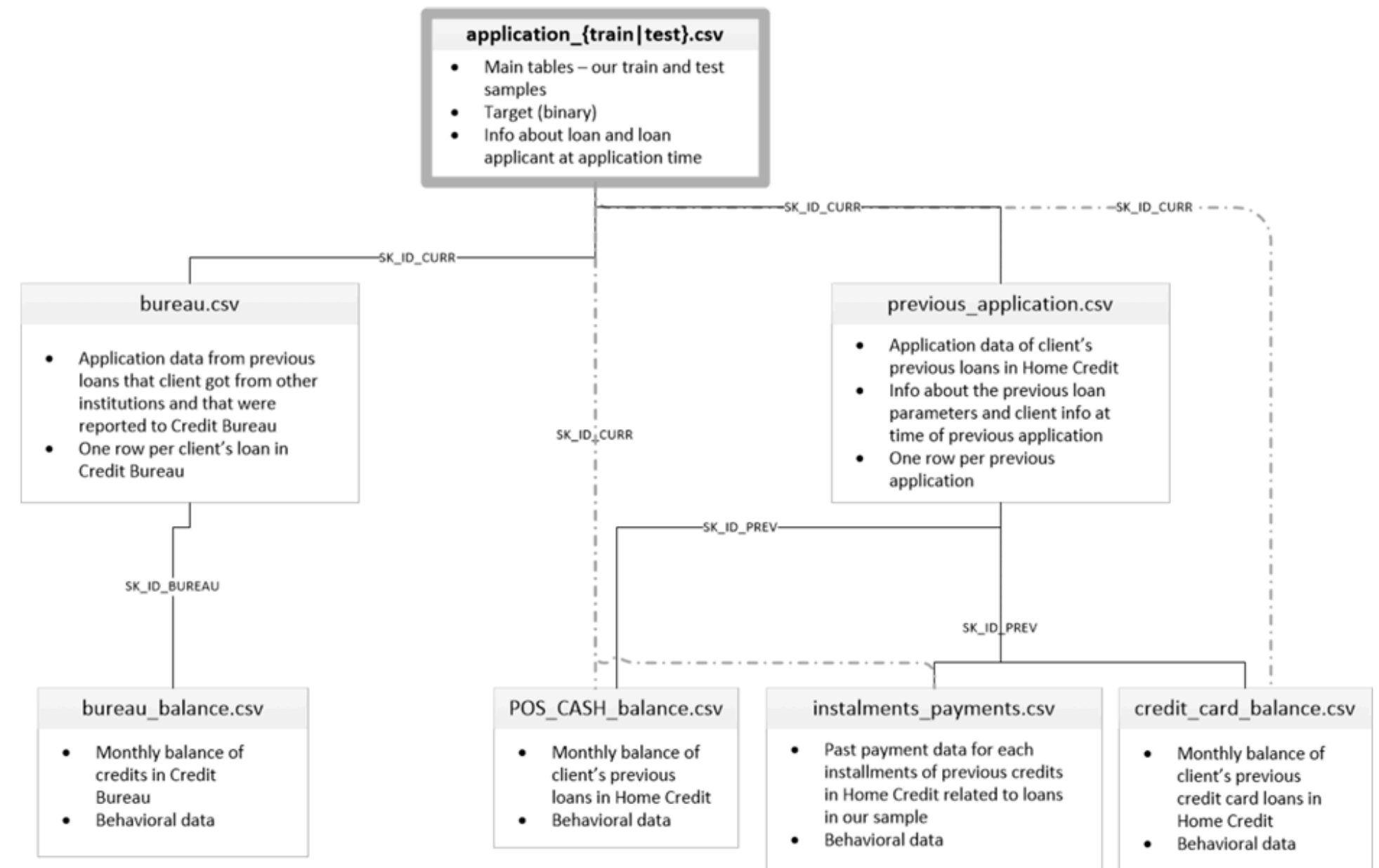
- 01 Identify potential borrowers at risk of default early.
- 02 Improve prediction accuracy to minimize risk.
- 03 Provide practical recommendations to support business decision-making.



DATASET OVERVIEW

Home Credit Dataset

- **Data source:** Home Credit (beberapa tabel: application, bureau, credit card balance, previous application, Pos Cash Balance, and installments payments).
- **Training set:** 307511 data.
- **Testing dataset:** 48744 data.
- **Total dataset size:** 2559.36 MB



METHODOLOGY



Analytical Approach

01 Data Preprocessing

- **Missing value imputation** (median for numeric, 'MISSING' for categorical)
- **Feature engineering** (Age, Income ratios, Credit utilization)
- **Multi-table aggregation** for enriched features

02 Feature Engineering

- **Derived variables:** AGE, YEARS_EMPLOYED, INCOME_PER_FAMILY
- **Financial ratios:** CREDIT_TO_INCOME, ANNUITY_TO_INCOME
- **Bureau aggregates:** Historical credit behavior metrics

03 Model Strategy

- **Traditional ML** (Random Forest, LightGBM)
- **Deep Learning** (Simple NN, Deep NN)
- **Hyperparameter tuning** for optimization

04 Model Evaluation

- AUC, Accuracy, Precision, Recall, F1

ANALYSIS RESULT

01

Basic Risk Overview

Overall, we found that the average Home Credit customer has a default rate of around 8.07%. This is a baseline figure that we need to try to reduce through improved prediction systems.

02

The Most Risky Customer Segment

Data analysis shows that risk is not evenly distributed. The two highest-risk customer groups are:

- Customers who list their income type as 'Maternity Leave', who have a very high default rate of 40.00%.
- The young age group (0 to 25 years old) has the highest default rate of all age groups, reaching 12.29%.

ANALYSIS RESULT

03

Key Risk Determinants (EXT_SOURCE)

We found that the most important factor determining whether a borrower will default is the External Credit Score (represented by features EXT_SOURCE_1, 2, and 3). The better the external score, the less likely the borrower is to default.

04

Best Model Performance

Of the various prediction models developed (including Random Forest and Neural Networks), the LightGBM model demonstrated the best risk prediction performance. It achieved a risk prediction accuracy (AUC) score of 0.7628, making it the most reliable tool for identifying risky borrowers.

RECOMENDATION

1. Prioritize External Data

Since External Credit Score (EXT_SOURCE) has proven to be the most important predictor, companies must ensure that this data is always integrated, valid, and up-to-date in the credit approval process.

2. Risky Segment Management Strategy

Stricter policies or adjustments to offerings for high-risk segments are needed.

- For younger applicants (0-25 years old), consider offering smaller loan limits or charging slightly higher interest rates to offset the risks.
- Applicants with the "Maternity Leave" income status should be reviewed, given their very high default rate.

3. Automated Decision Optimization

When implementing the best model (LightGBM), we need to be careful in determining the risk prediction threshold. Although LightGBM is accurate overall (AUC 0.7628), the model currently tends to focus more on identifying as many default cases as possible (high Recall) than on being absolutely certain that the default prediction is correct (low Precision, 0.1893). We need to set this threshold to balance minimizing losses (by rejecting truly risky applicants) with not losing too many good potential customers.

CONCLUSION

01.

Key Achievements

This project successfully identified and built the best loan risk prediction model. The LightGBM model outperformed all other methods, including Random Forest models and more complex artificial neural networks (Deep NN).

02.

Critical Insight

This analysis confirms that external data and age are the main drivers of default risk.

03.

Model Comparison

Interestingly, we found that sophisticated traditional Machine Learning models like LightGBM still outperform (are more accurate) than complex Deep Learning models for this type of customer data, even though Deep Learning is able to capture complex non-linear patterns.



THANK YOU

● FOR YOUR NICE ATTENTION