HOME CREDIT

Default Risk Assessment

Yuri Croci

Personal Project

Project Overview

Context & Challenge:

- Home Credit provides loans to customers with limited or no credit history, typically rejected by traditional banks.
- Traditional credit scoring is insufficient, making it difficult to identify creditworthy applicants and manage default risk.

Project Objective:

- Build a predictive model to assess loan default risk and inform lending decisions.
- Segment customers by risk level to enable differentiated lending strategy.

Data Overview

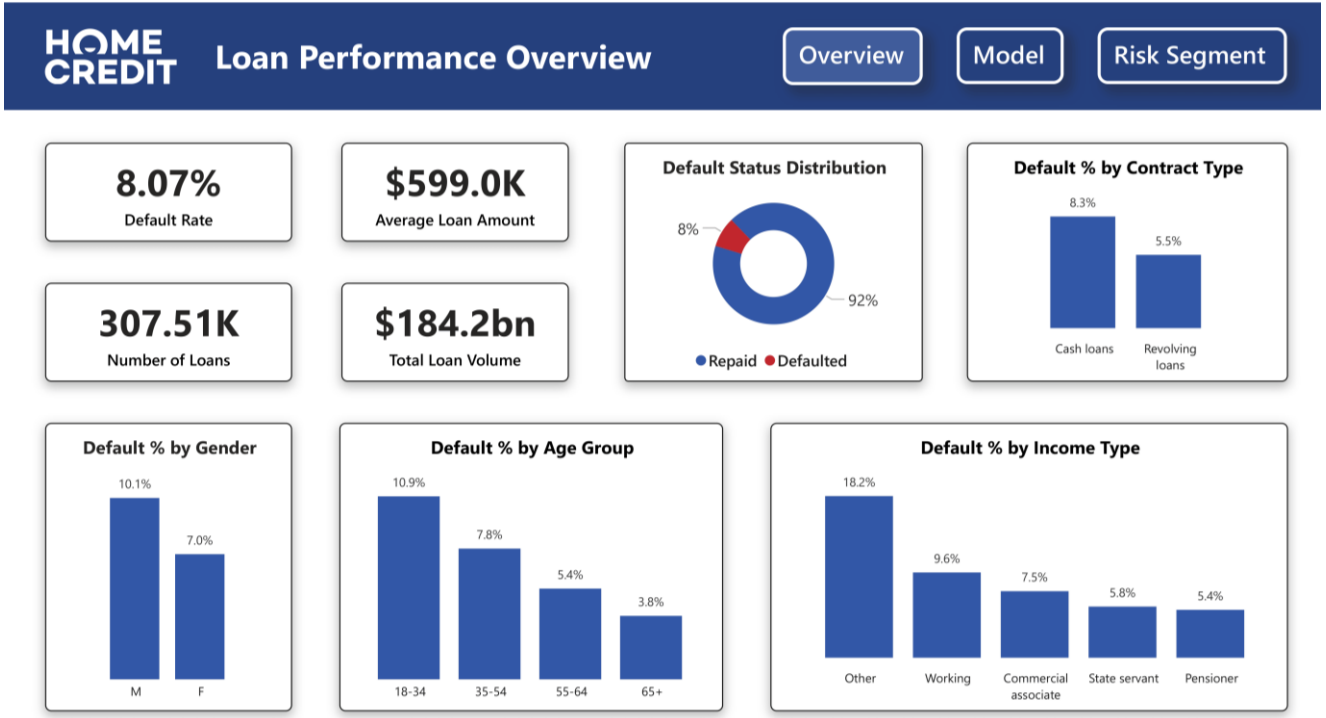
Dataset:

- 307,511 loan applications containing application details and applicant profile characteristics.
- Additional data from external sources including credit bureau history and previous Home Credit applications, decisions, and transactions.

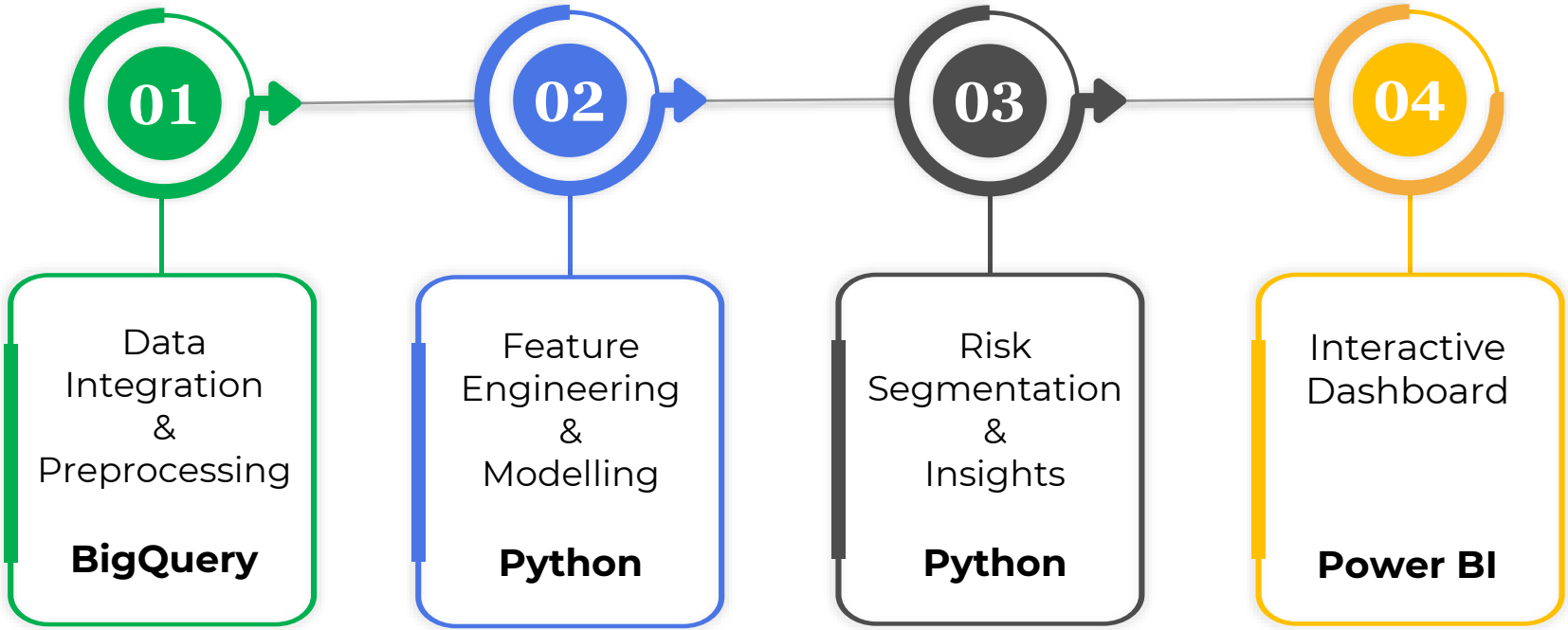
Key characteristics:

- Highly unbalanced class with 8% defaulted application over the 92% repaid.
- High data sparsity with complex missing value patterns where some missing information carries predictive signal.
- Mixed data types including categorical, numerical, and binary variables.

Home Credit Overview



Methodology Overview



Data Preparation & Feature Engineering

Data Preparation:

- Data cleaning including outlier treatment, sentinel value normalization, and category consolidation.
- Missing value handling with informative flags, temporal feature transformation, and data type conversions.

Feature Engineering:

- Multi-table integration aggregating bureau and previous application data to construct historical credit behavior profiles.
- Feature selection using phi-k correlation for mixed variable types followed by multicollinearity analysis, reducing from 100+ to 36 relevant features.

Modelling Development

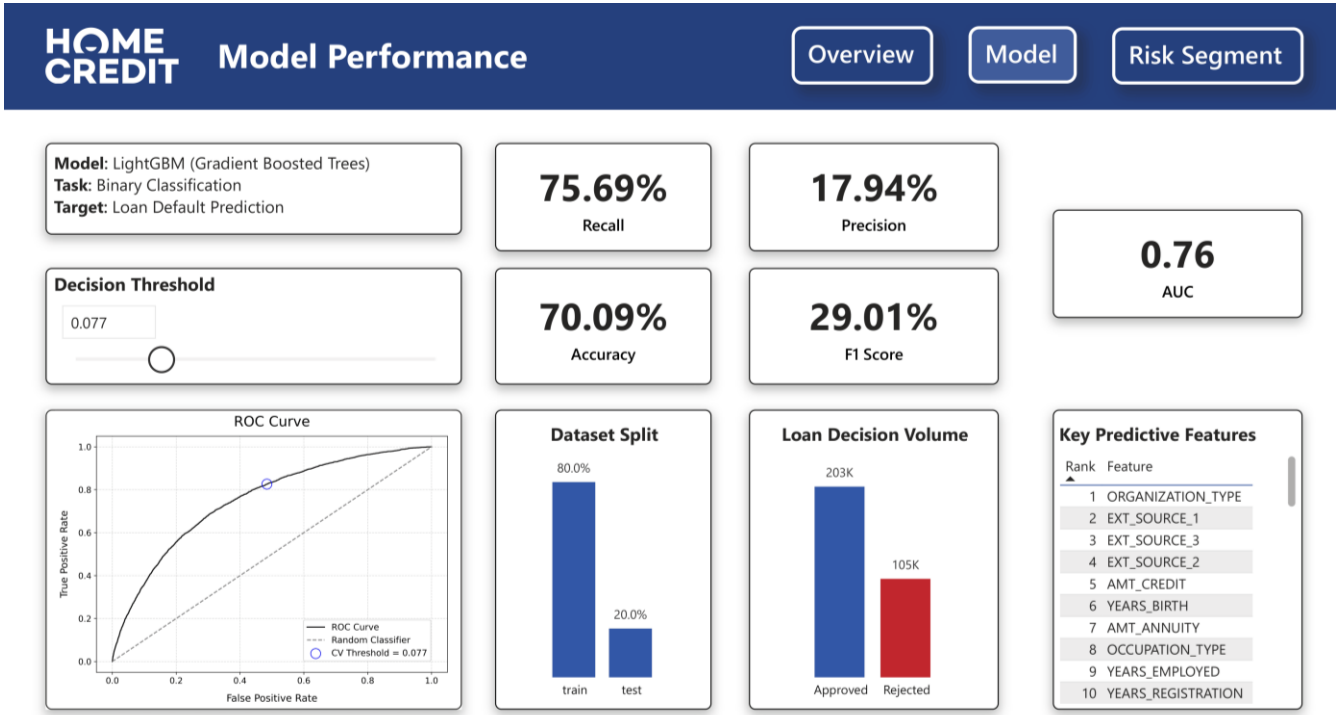
Modelling Setup:

- 80/20 stratified split with cross-validation for hyperparameter tuning to ensure robust performance estimates.
- Baseline logistic regression benchmarked against LightGBM for native handling of missing values and mixed data types.

Model Selection:

- Threshold optimization via cross-validation folds to prevent data leakage.
- Evaluation prioritizing recall and AUC given asymmetric costs of misclassification in default prediction.
- LightGBM selected for improved performance and computational efficiency.

Model Performance



Risk Segmentation

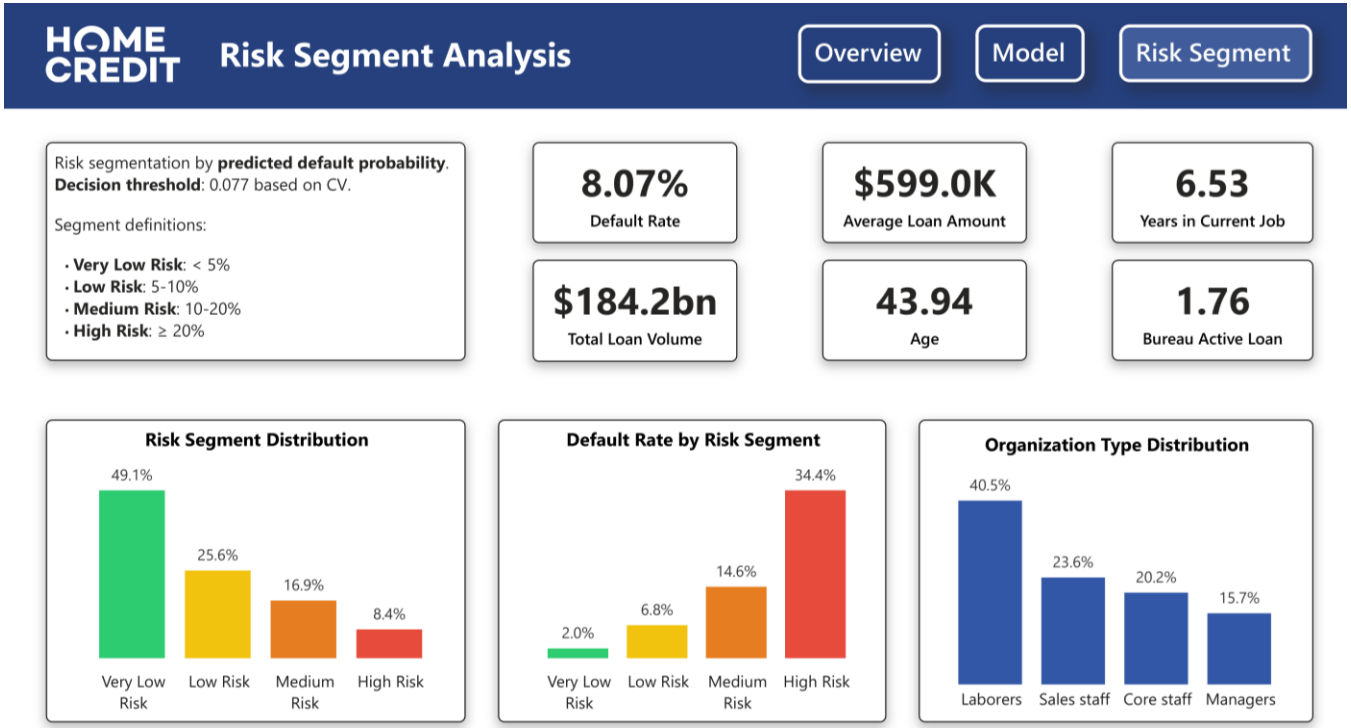
Segmentation Approach:

- Four-tier risk classification with Very Low (<5%), Low (5-10%), Medium (10-20%), and High (≥20%) based on model predicted probabilities of default.
- Feature profiling of the risk segment across demographic, employment, and credit dimensions.

Key Findings:

- Increasing average default rates across segments validating risk separation.
- Negative correlation between age, employment tenure, education level, and default risk.
- Positive correlation with prior loan refusals and multiple active bureau loans.

Risk Segment Analysis



Business Recommendations

1. **Streamline approval process** through automation for the Very Low Risk segment (49% of applications, 2.0% default rate) while concentrating manual screening resources on Medium and High Risk tiers requiring deeper evaluation.

2. **Implement risk-based pricing strategy** with differentiated interest rates and flexible repayment schedules across segments to enable approvals for higher-risk segments through appropriate risk compensation.

3. **Enhance credit evaluation workflows** by integrating insights on feature-risk relationships to support decision-making in borderline or uncertain application cases.