Home Credit Default Risk ADS Audit

Rona Zhang

May 6, 2025

1 Background

Purpose and Stated Goals

The Home Credit Default Risk Automated Decision System (ADS) is designed to assess whether a loan applicant is likely to default on a loan. Its core function is to predict the probability of repayment, enabling lenders to make more informed decisions about whether to approve a loan, and if so, under what conditions.

The stated goals of this system go beyond risk assessment. According to Home Credit, the organization behind the system, one of its main missions is to promote financial inclusion, particularly for individuals who are often excluded from traditional credit systems due to limited or non-existent credit histories. These applicants may include young people, recent immigrants, informal workers, or others who are not well-represented in mainstream financial data.

To achieve this, Home Credit uses a wide variety of alternative data sources, including telco metadata, transaction records, and behavioral signals, in addition to traditional financial information. The goal is to construct a fuller picture of an applicant's reliability and ability to repay a loan. Thus, the ADS has three key aims:
1. Expand access to credit for underbanked or unbanked populations.
2. Support responsible lending by estimating credit risk accurately, thereby reducing defaults.
3. Promote borrower success by tailoring loans to applicants' financial situations, avoiding over-lending or predatory terms.

These align with the vision of ethical AI deployment as emphasized by Raji et al.(2020) in "Closing the AI Accountability Gap", where the authors stress the importance of internal audits and end-to-end accountability frameworks, especially for high-stakes domains like credit scoring. So this system reflects a broader trend in financial services, which is the shift from rigid credit-score thresholds to flexible, data-driven decision-making that aims to be more inclusive while managing financial risk.

Trade-offs Between Goals

While the system aspires to both fairness and financial performance, these goals often conflict in practice, requiring careful design choices and policy oversight. The key trade-offs are:
- Inclusivity vs. Financial Risk: Including applicants with little or no credit history may increase default risk, especially if alternative data signals are noisy or hard to interpret. So the challenge lies in separating high-risk applicants from genuinely creditworthy but invisible ones;
- Predictive Accuracy vs. Fairness: The model may perform well on average but poorly for certain groups. For example, women, elderly applicants, or minority groups might receive less favorable scores due to historical biases in the data. If these patterns are not checked, the model may unintentionally perpetuate or even amplify discrimination;
- Model Performance vs. Interpretability: The system relies on gradient boosting machines (Light-

GBM), which are highly performant but relatively opaque. While these models boost accuracy, they also make it harder to provide clear explanations to applicants, credit officers, or regulators, especially in high-stakes decisions that affect people's financial decisions;
- Scale vs. Personalization: Operating a model at scale often involves standardizing decisions. However, Home Credit also aims to tailor loans to individual needs, including the loan's principal, maturity, and repayment schedule. Balancing these objectives requires careful model design and post-processing to ensure that personalization does not result in inconsistent or unfair outcomes.

These trade-offs are not just technical, they are deeply connected to real-world consequences. The audit aims to evaluate the ADS not only in terms of accuracy but also with regard to fairness, transparency, and accountability. In high-stakes domains like lending, these attributes are essential for building trust, ensuring legal compliance, and avoiding harm to vulnerable groups.

2 Input and output

Data Sources and Collection

The data used by this ADS comes from the Home Credit Default Risk Kaggle competition and is based on real, anonymized loan application records provided by Home Credit Group. The dataset is designed to support predictive modeling of credit risk, particularly for individuals with limited or no formal credit history.

The main table, `application_train`, contains over 300,000 labeled examples, each representing a loan application. It includes the target variable TARGET, where 1 indicates a loan default and 0 indicates repayment, along with features related to an applicant's demographics, employment, income, credit history, housing situation and so on. Another primary component is `application_test`, a separate set of applications without labels, used for generating final predictions. Several auxiliary tables are used to provide relational data on an applicant's financial history, including `bureau`, `bureau_balance`, `previous_application`, `POS_CASH_balance`, `credit_card_balance`, and `installments_payments`. These tables are linked via applicant IDs and offer insights into previous loans, repayment timing, external credit reports, and installment behaviors.

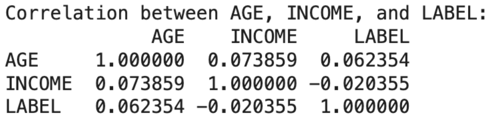
The data was collected through Home Credit's internal systems and partnerships with third-party institutions, such as credit bureaus. It includes both traditional financial indicators and a range of alternative data sources, which reflect Home Credit's effort to evaluate repayment ability for clients who may not have conventional credit scores or banking histories.

Input Feature Types and Profiling

The input features from `application_train` consist of numerical, categorical, and binary variables. These features describe each applicant's demographic, financial, and behavioral profile, forming the basis of the credit risk prediction model.
- **Numerical:** These include variables such as `AMT_INCOME_TOTAL` (income), `AMT_CREDIT` (loan amount), `AMT_ANNUITY`, `DAYS_EMPLOYED`, `DAYS_BIRTH`, and numerical aggregates from related tables, such as credit bureau balances and payment delays. Most numerical variables are skewed and may contain extreme values. These features are often log-transformed or clipped to reduce skewness and improve model performance.
- **Categorical:** Examples include `NAME_INCOME_TYPE`, `NAME_EDUCATION_TYPE`, `OCCUPATION_TYPE`, and `NAME_HOUSING_TYPE`. All categorical features are one-hot encoded before modeling, which significantly increases the dimensionality of the dataset.
- **Binary:** These include flags such as `FLAG_OWN_CAR`, `FLAG_OWN_REALTY`, and `CODE_GENDER`. They are originally encoded as strings like Y/N, M/F and are transformed into binary (0/1) format

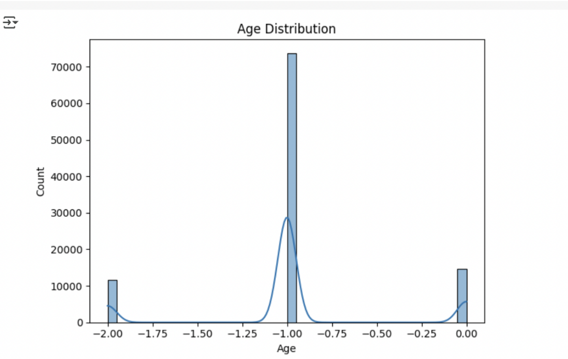
during preprocessing.

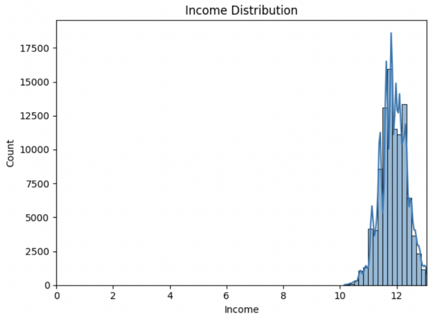
Several features contain missing values, especially among those engineered from the auxiliary bureau table. For example, `AMT_MAX_OVERDUE_RATIO_1`, derived from credit bureau data, has a missing rate of over 94 percent. Such high-missingness features are either imputed or excluded depending on their distribution and predictive relevance. However, the system does not include explicit handling of missing categories such as assigning a "Missing" label. Some selected numerical features in auxiliary tables are filled with 0, though most missing values are left as-is. Since LightGBM is used for modeling, which can handle missing values natively, it is likely that remaining missing entries are managed internally during training.

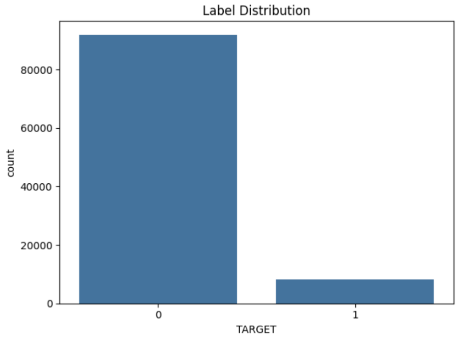


```
Correlation between AGE, INCOME, and LABEL:
             AGE    INCOME     LABEL
AGE     1.000000  0.073859  0.062354
INCOME  0.073859  1.000000 -0.020355
LABEL   0.062354 -0.020355  1.000000
```

The pairwise correlations above between AGE, INCOME, and LABEL are all weak. This suggests low linear dependence between these features, and no strong multicollinearity concerns. While AGE and INCOME have minimal correlation with the outcome LABEL, AGE shows a slightly positive association of about 0.06, potentially worth exploring further.







We also performed basic input profiling to examine the distributions of age, income, and target labels, revealing skewed features, concentration effects, and class imbalance that may impact model performance and fairness. The age variable appears to have been normalized or transformed, but the distribution is highly concentrated around -1. This suggests that a large number of applicants are older, possibly with some edge cases represented by peaks near 0 or -2. Additional preprocessing or binning may be useful to stabilize its influence. The income distribution is left-skewed, with most applicants clustered at higher income levels and a tail extending toward lower incomes, likely due to log transformation applied during preprocessing. The target variable is highly imbalanced, with a majority of applicants labeled as 0 with non-default and a much smaller group labeled as

1 with default. This class imbalance needs to be addressed using techniques like class weighting, oversampling, or evaluation metrics like AUC or F1 rather than accuracy.

System Output and Interpretation

The system outputs a probability score between 0 and 1 for each applicant, representing the predicted likelihood of loan default. These probabilities are stored in the TARGET column of the submission files, where each row corresponds to a unique applicant. For example, an applicant with a score of 0.89 is considered to have a high predicted risk of default, whereas a score of 0.37 indicates lower risk. This output is generated by an ensemble of LightGBM models and reflects a soft classification, meaning the model does not assign hard class labels, but instead produces continuous scores that can later be thresholded or ranked depending on the decision policy. In the output, the values range from 0.0 to 1.0, with an average near 0.5, which reflects model calibration and balance. These probabilities can be interpreted directly or used to support downstream credit decisions, such as loan approvals, rejections, or personalized loan terms based on risk. In practice, lenders might apply a threshold to segment applicants into different risk categories.

3 Implementation and validation

Data Cleaning and Feature Engineering

The first stage of the system's implementation is detailed in `code_1_data_prep` notebook, which performs comprehensive data cleaning and feature engineering on both the main application data and a set of auxiliary tables. These auxiliary datasets include bureau, bureau balance, previous application, POS CASH balance, installments payments, and credit card balance. All of these tables are joined with the primary application data via the applicant's unique ID (`SK_ID_CURR`), and extensive feature engineering is applied to capture applicants' credit history, payment behavior, and financial status.

Missing values are carefully analyzed using custom utility functions that quantify both the total and percentage of missing entries in each feature. While the LightGBM model can handle missing values natively, features with excessive missingness are removed to improve model robustness, and missing indicators are generated to preserve information encoded in the pattern of missingness itself.

Feature engineering plays a central role in this system. A wide array of new features is constructed, including income-based ratios (e.g., credit-by-income and annuity-by-income), behavioral indicators (e.g., percent of life worked, number of documents submitted), and temporal transformations (e.g., converting durations from days to months). These transformations not only enhance the expressiveness of the features but also help normalize the scale and distribution of values. Logarithmic transformations are also applied to skewed monetary variables to reduce the influence of extreme values.

Aggregation functions are used to summarize historical loan data from related tables. For example, mean, max, and count statistics are computed for past credit card usage or installment payments, grouped by the applicant ID. Additional features, such as the number of approved or rejected loans, are extracted from the previous application table. After all auxiliary data is processed and aggregated, it is merged back into the core application table, resulting in a final dataset that consolidates both current and historical financial information about each applicant.

The final dataset is one-hot encoded to prepare for modeling, and both the training and test splits are reconstructed based on whether the applicant's ID appears in the labeled training set. These final datasets are exported as CSV files (`train_full_cor.csv, test_full_cor.csv, and` y_full_cor.csv) for use in the subsequent modeling pipeline.

Model Architecture and Training Approach

Based on code_2_modeling notebook, the system is built around a LightGBM classifier, a fast, efficient gradient boosting framework suited for large-scale tabular data. After loading the cleaned dataset, the model trains on a wide feature set that excludes only the applicant identifier SK_ID_CURR. Key hyperparameters include 10,000 boosting rounds, a learning rate of 0.005, 70 leaf nodes per tree, and a maximum tree depth of 7, along with subsampling of 0.9 and column sampling of 0.8 to improve generalization. Regularization terms (reg_alpha = 0.1, reg_lambda = 0.1) are included to reduce overfitting.

The training loop is implemented using LightGBM's callback-based early stopping, which monitors model performance and halts training when progress plateaus. Feature importances are recorded during training, and the most informative 500 features are selected for a second training round. This two-stage training approach, first with all features and then with a reduced subset, helps streamline the final model and improve interpretability without sacrificing predictive strength. Final model outputs are saved for downstream tasks like ensembling and submission.

Performance Evaluation and Calibration

Validation of the ADS is carried out rigorously in both code_2_modeling and code_3_ensemble notebooks, confirming that the system meets its primary stated goal that accurately predicting the risk of loan default. The core evaluation metric is the Area Under the Receiver Operating Characteristic Curve (AUC), which is well-suited for binary classification tasks with imbalanced classes. By employing stratified K-fold cross-validation with 5 folds, the model avoids overfitting to any particular subset of the training data and provides a robust estimate of generalization performance. Each fold's validation AUC is printed and averaged, with final cross-validation scores consistently around 0.794, suggesting strong discriminatory power.

In code_3_ensemble notebook, ensemble predictions are created by averaging rank-transformed outputs from multiple submission files. Specifically, a new prediction is generated by computing the rank-mean of the current model's output and that of a previously successful submission. This ensembling technique reduces variance and increases robustness, especially when base models exhibit complementary strengths. The Spearman rank correlation is computed between the ensemble and benchmark submissions to assess consistency and monotonic alignment. The final ensemble predictions are normalized and saved as a new submission file rmean_top2.csv, which improves public leaderboard performance and reflects a stable, generalizable system.

Through this multi-step validation, including cross-validated AUC, feature selection based on importance, and ensemble blending, the ADS provides strong evidence of meeting its predictive accuracy goals while maintaining robustness and generalizability.

4 Outcomes

Accuracy Across Subpopulations

To assess how accurately the ADS performs across different demographic groups, we segmented the population by age into three categories: under 30, 30 to 50, and over 50. We calculated both the AUC and accuracy for each subgroup. Results showed that the model performed best in terms of AUC for the 30–50 group (0.7968), suggesting this demographic aligns most with the learned patterns. Meanwhile, the over-50 group had the highest accuracy (0.9433), potentially due to a more conservative prediction pattern.

```
AGE_GROUP: 30-50, AUC: 0.7968, Accuracy: 0.9170
AGE_GROUP: 50+, AUC: 0.7656, Accuracy: 0.9433
AGE_GROUP: <30, AUC: 0.7636, Accuracy: 0.8881
```

Although overall performance is strong, the discrepancy in AUC between the < 30 and 30–50 groups may indicate the model captures signal from middle-aged applicants more effectively than from younger or older ones.

Fairness Evaluation and Metrics

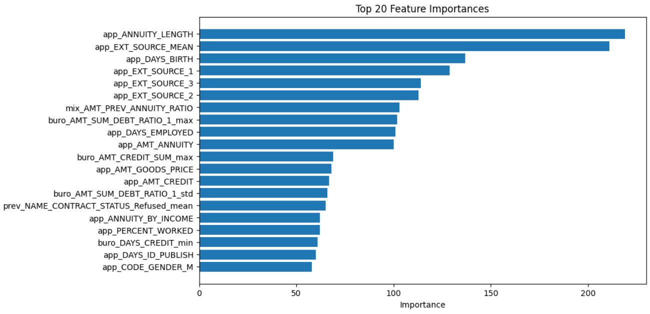
We used True Positive Rate (TPR) and False Positive Rate (FPR) to analyze fairness across age groups. The TPR was lowest in the 50+ category (0.0311), despite their high accuracy. This reveals that while the model rarely misclassifies older applicants as defaulters (low FPR), it also tends to under-identify those who are truly at risk, suggesting a conservative bias.

```
AGE_GROUP: 30-50
  TPR: 0.0494
  FPR: 0.0031
AGE_GROUP: 50+
  TPR: 0.0311
  FPR: 0.0012
AGE_GROUP: <30
  TPR: 0.0471
  FPR: 0.0041
```

Although FPR remains low across all age brackets, the TPR is substantially lower for the 50+ group, indicating that the model is less sensitive to identifying actual defaulters in older populations. This violates the fairness criterion of equal opportunity. The disparity suggests that older applicants are more likely to be falsely considered low-risk, potentially undermining the lender's risk management and leading to undesirable financial outcomes for this demographic.

Additional Performance Analyses

Beyond metrics, we explored model interpretability through feature importance. The most impactful features included `app_ANNUITY_LENGTH`, `app_EXT_SOURCE_MEAN`, and `app_DAYS_BIRTH`, indicating the model heavily relies on annuity structure, external scoring sources, and age. These features suggest a strong reliance on financial durability and demographic information. This reliance on certain demographic indicators can lead to indirect bias, especially if these features correlate with age or socioeconomic status. Moreover, while the model was stable under normal test conditions, robustness to edge cases (e.g., incomplete employment history, extreme income ranges) remains unexplored and warrants future stress-testing.



5 Summary

Assessment of Data Quality and Relevance

The ADS is built on a rich dataset that integrates credit history, demographic traits, and transactional behaviors. Its multi-source structure allows for nuanced risk assessment. However, the prevalence of missing values in auxiliary tables and the potential underrepresentation of certain demographics, like young applicants or those in informal employment, introduce bias. Including non-traditional indicators such as utility bills or mobile payment data could enhance inclusivity and prediction reliability.

Evaluation of Model Strength and Fairness

The LightGBM-based model shows strong predictive performance and high overall accuracy. However, fairness analysis highlighted disparities, especially in TPR across age groups. This reflects unequal detection power for defaulters across demographics, which could impact vulnerable groups unfairly. From a stakeholder perspective, banks benefit from high accuracy and low FPR, helping minimize risk. Applicants, especially older ones, may face biased assessments. Regulators would require evidence of fairness, possibly through equal opportunity analysis or differential impact studies, prior to model deployment.

Deployment Considerations

We would be comfortable deploying this model in the industry setting, provided that explainability tools are implemented, human reviewers evaluate borderline cases, and fairness and performance are regularly audited. These safeguards can help ensure responsible and interpretable use in high-stakes decision-making. However, public sector deployment requires more caution. Decisions that significantly impact individuals, such as access to housing, credit, or benefits, must be transparent, explainable, and allow for formal appeal mechanisms. In its current form, this model lacks sufficient transparency and fairness safeguards. Without fairness interventions, it could reinforce existing inequalities, especially if it exhibits higher error rates for vulnerable populations. Therefore, public sector deployment without substantial modification and oversight is not recommended.

Recommendations for Improvement

To improve the reliability, fairness, and transparency of this ADS, we recommend several enhancements. First, the dataset could be expanded to include more granular information about applicants' occupations, income stability, and alternative credit behavior, such as utility payments or mobile financial transactions, to reduce bias against underrepresented groups. During model training, fairness-aware techniques like reweighting, adversarial debiasing, or incorporating fairness constraints should be applied to ensure more equitable True Positive Rates across demographic groups. Post-hoc interpretability methods, such as SHAP value analysis, should be routinely used to monitor decision logic and detect potential proxy discrimination. In deployment, we also suggest implementing fairness dashboards to continuously monitor error rates and group-level disparities, as well as periodic audits to ensure the system remains aligned with ethical and regulatory standards. Finally, if persistent gaps remain, consider training separate models or adjusting thresholds for different subgroups, provided this is done transparently and with oversight, to improve fairness without compromising accuracy.