*Abstract –*

**This project addresses the critical challenges of predicting loan default risk for Home Credit, a financial institution serving the underbanked population who lack traditional credit histories. Using the Home Credit Default Risk dataset from Kaggle, which contains over 300,000 loan applications with 122 features across 7 related tables, we developed a comprehensive machine learning pipeline to estimate Probability of Default (PD) for loan applications.**

**Our methodology involved extensive data preprocessing including handling missing values, outlier treatment using IQR-based methods with domain-specific constraints, and integration of supplementary data from bureau records, previous applications, and payment histories. We engineered 15 new features including financial ratios (credit-to-income, annuity-to-income), temporal indicators (age, employment duration), and external score combinations.**

**We implemented and compared Logistic Regression (baseline) and LightGBM (primary model), achieving a ROC-AUC of 0.7787 on the validation set. Isotonic probability calibration was applied to ensure reliable PD estimates. Feature engineering demonstrated a +0.62% improvement in ROC-AUC, with EXT_SOURCE features and financial ratios emerging as the strongest predictors. The final model outputs calibrated default probabilities that can be mapped to credit scores (300-850 scale) for business applications including risk-based pricing, portfolio management and regulatory compliance.**

# I. Introduction

### 1.1 Project Motivation and Background

Access to credit is fundamental to economic participation, yet millions of people are denied loans due to insufficient or non-existent credit histories. Home Credit, an international consumer finance provider, aims to broaden financial inclusion by providing loans to the unbanked population. However, lending to individuals without traditional credit data presents challenges in assessing default risk.

Traditional credit scoring models rely heavily on credit bureau data, which excludes individuals who have never had formal credit relationships. This creates a paradox where those who most need credit access are systematically denied. Machine learning offers an opportunity to leverage alternative data sources - such as mobile phone usage, payment behaviours, and demographic information - to make more inclusive lending decisions while maintaining acceptable risk levels.

### 1.2 Problem Definition

The core problem is a binary classification task: given a loan application with associated features, predict whether the applicant will default on the loan (TARGET = 1) or repay successfully (TARGET = 0). However, rather than producing hard classifications, we focus on estimating the Probability of Default (PD), which provides more actionable information for business decision-making.

### 1.3 Research Gap

- Many models focus solely on discrimination rather calibration (accurate probability estimation)
- Limited attention to interpretability and feature engineering validation
- Insufficient integration of multiple data sources for holistic risk assessment

### 1.4 Project Goals

- Develop a robust PD estimation model with proper probability calibration
- Integrate all available data sources through careful feature engineering
- Validate the impact of feature engineering on model performance
- Provides business-ready outputs (PD and credit scores) for practical applications

## II. Dataset & Data Processing

### 2.1 Dataset Source and Overview

The dataset is from the Home Credit Default Risk competition on Kaggle (https://www.kaggle.com/competitions/home-credit-default-risk). It comprises 7 related tables with the main application table containing 307,511 training samples and 122 features.

| Table | Records | Features | Description |
|---|---|---|---|
| application_train | 307,511 | 122 | Main table with target variable |
| application_test | 48,744 | 121 | Test set without target |
| bureau | 1716428 | 17 | Previous credits from other institutions |
| bureau_balance | 27299925 | 3 | Monthly balances of previous credits |
| POS_CASH_balance | 10001358 | 8 | Monthly POS/cash loan balances |
| credit_card_balance | 3840312 | 23 | Monthly credit card snapshots |
| previous_application | 1670214 | 37 | Previous Home Credit applications |
| installments_payments | 13605401 | 8 | Payment history for previous loans |

The target variable shows significant class imbalance with only 8.07% default rate (24,825 defaults vs. 282,686 non-defaults), which necessitates appropriate handling during model training.

### 2.2 Feature Types

The main application table contains:

- **Numerical Variables (104):** Including income (AMT_INCOME_TOTAL), credit amount (AMT_CREDIT), external scores (EXT_SOURCE_1/2/3), and various flags
- **Categorical variables (16):** Including contract type, gender, education, occupation, and organization type
- **Identifier (1):** SK_ID_CURR for joining tables
- **Target Variable (1):** Binary default indicator

### 2.3 Data Cleaning

**Step 1: Handling Inconsistent and Anomalous Values**

The dataset contains several inconsistent or placeholder values. We performed domain-aware corrections:

- **Dirty-placeholder values → NaN**: Several features use placeholders tokens to indicate missing unknown information. These entries were converted to NaN for consistency and to ensure accurate missing-value tracking.
- **DAYS_EMPLOYED anomaly correction:** The value 365243 appears frequently in DAYS_EMPLOYED and represents "permanent employment" or retirement status. Such values distort the distribution, so they were replaced with NaN to avoid biasing subsequent analyses. Figure X shows the distribution before and after this correction, demonstrating the removal of the anomalous spike at 365243.

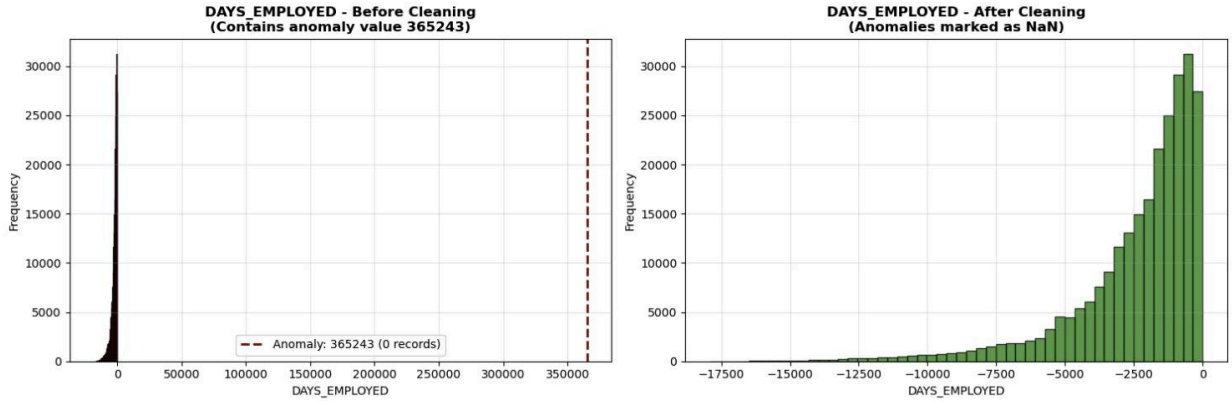



Figure X: DAYS_EMPLOYED Before/After

**Step 2: Missing Value Treatment**

Missingness is pervasive throughout the dataset, with some variables exceeding 60% missing. Our strategy follows:

- **Threshold-based feature selection:** We adopted a 70% missing value threshold as the criterion for feature removal. Upon inspection, no features in the application dataset exceeded this threshold, so all original features were retained for subsequent imputation.
- **Imputation strategy:** For features with <70% missing values, missing entries were imputed during the modeling pipeline using simple statistical techniques:
    - Median imputation for numerical variables
    - Mode imputation for categorical variables

Figure Y illustrates the top 20 features with the highest missing rates before cleaning (left panel) and the complete elimination of missing values after treatment (right panel). The highest missing rate observed was approximately 69.4% for several regional features (e.g., COMMONAREA_MEDI, NONLIVINGAPARTMENTS_MODE). The overall missing rate decreased from 22.65% to 0.00%, confirming successful imputation across all retained features.

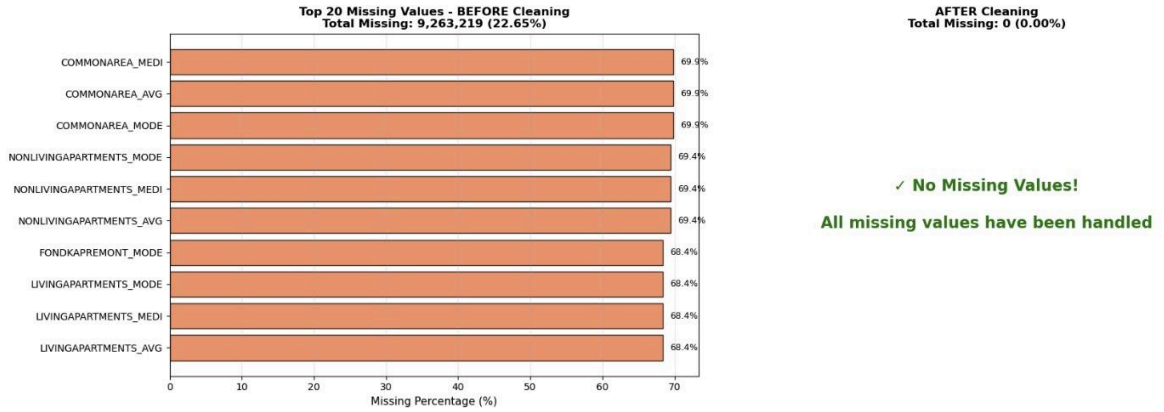



Figure Y: Missing Values Before/After

**Step 3: Duplicate Check**

The dataset was checked for duplicate application records. No duplicated rows requiring removal were identified.

**Step 4: Data Integration Across Supplementary Tables**

Multiple related Home Credit tables were aggregated and merged into the main application table to create borrower-level features. Aggregated information includes:

- **Bureau data:** counts of previous loans, mean credit amounts, overdue statistics
- **Previous applications:** application counts, amounts, approval/denial timing
- **Installments:** repayment behavior (on-time vs. late payments)
- **POS/Cash balance:** delinquency patterns and days past due
- **Credit card:** utilization rates and balance dynamics
- **Bureau balance:** monthly status histories

After integration, the dataset expanded from 122 original features to 185 features (122 original + 63 aggregated). Missing values introduced by the merge operations (for applicants with no records in supplementary tables) were filled with zeros, as absence of records indicates no activity in those categories.

**Step 5: Outlier Handling**

After feature integration, exploratory plots revealed extreme values and long-tailed distributions in several numeric variables. To mitigate their impact, outlier handling was incorporated into the preprocessing pipeline:

- **IQR-based capping:** For each continuous feature (excluding SK_ID_CURR, the target variable, and binary indicators), we computed the interquartile range (IQR) and applied capping at: $Q_1$ - 1.5 × IQR, $Q_3$ + 1.5 × IQR
- **Business-logic constraints were applied to ensure realistic ranges:**
    - DAYS_* variables were restricted to [–30,000, 0]

- Monetary and count variables (AMT_*, CNT_*) were constrained to non-negative values
- EXT_SOURCE_* scores were clipped to the [0, 1] interval
- Binary FLAG_* variables were excluded from outlier processing

Figure Z presents a before-and-after comparison of four key financial features (AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE). The left panels show the original distributions with visible extreme outliers (marked by scattered points beyond the whiskers), while the right panels demonstrate the compressed, clipped distributions after IQR-based capping. This approach successfully reduced outlier counts from 4.56%-4.79% to 0.00% across these features, while preserving the general shape of each distribution.

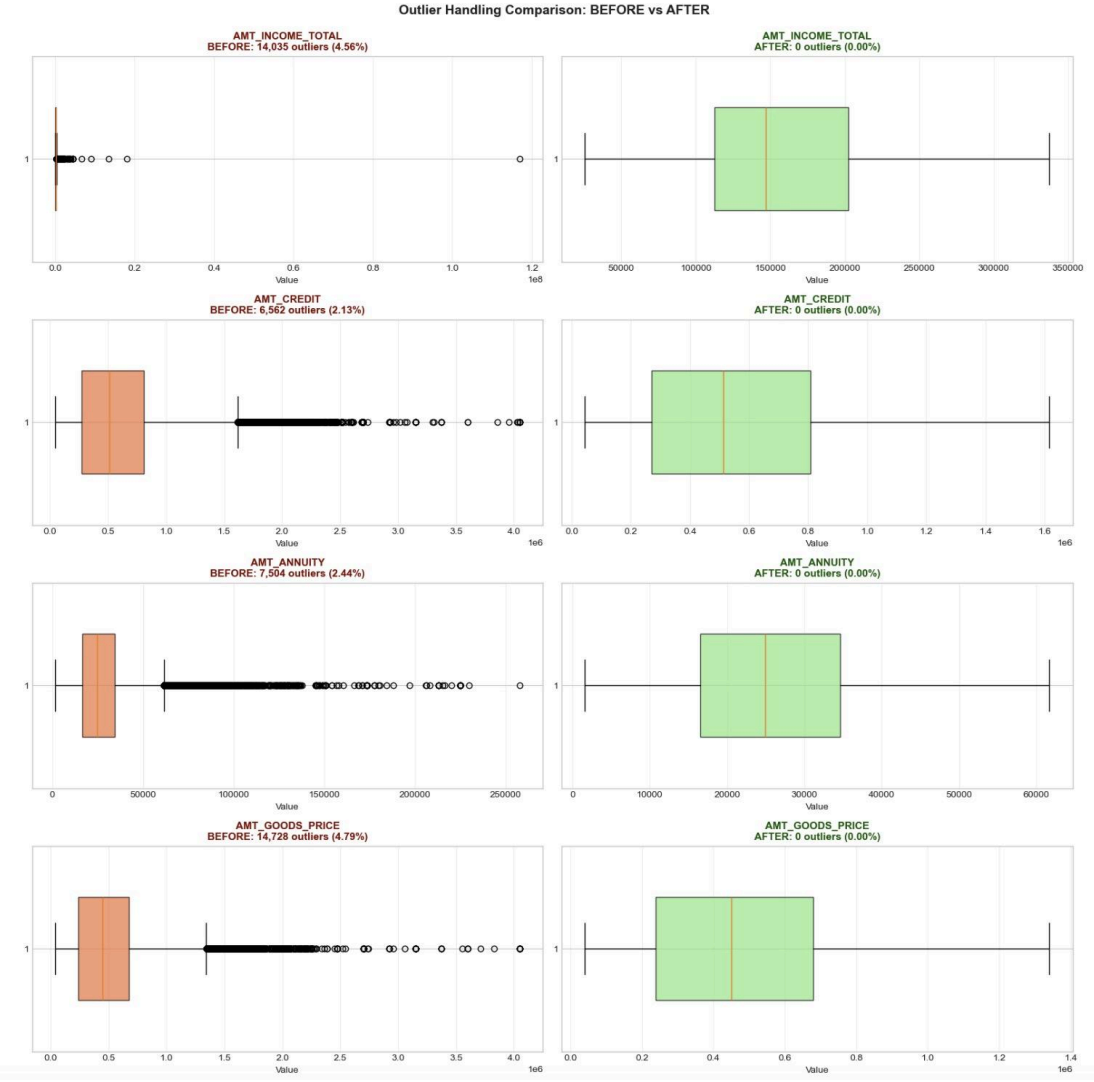



Figure Z: Outlier Handling Before/After

### 2.4 Feature Engineering

After data cleaning, we created additional derived features to enhance the predictive power of the method. These engineering features capture domain-specific relationships and ratios that may be informative for credit risk assessment:

**Financial Ratio Features**

We constructed several ratio-based features to capture the relationship between different financial amounts:

- **Credit-to-Income Ratio:** *CREDIT_INCOME_RATIO = AMT_CREDIT / (AMT_INCOME_TOTAL + 1)*
    - Measures borrowing relative to income capacity
- **Annuity-to-Income Ratio:** *ANNUITY_INCOME_RATIO = AMT_ANNUITY / (AMT_INCOME_TOTAL + 1)*
    - Indicates monthly payment burden relative to income
- **Credit-to-Annuity Ratio:** *CREDIT_ANNUITY_RATIO = AMT_CREDIT / (AMT_ANNUITY + 1)*
    - Approximates loan duration
- **Goods-to-Credit Ratio:** *GOODS_CREDIT_RATIO = AMT_GOODS_PRICE / (AMT_CREDIT + 1)*
    - Captures down payment proportion (inverse of leverage)

**Temporal Features**

Time-related features were converted to more interpretable units:

- **Age in Years:** *AGE_YEARS = -DAYS_BIRTH / 365*
    - Converts negative days to positive years
- **Employment Duration in Years:** *EMPLOYED_YEARS = -DAYS_EMPLOYED / 365*
    - Standardizes employment history to years
- **Employment-to-Age Ratio:** *EMPLOYED_AGE_RATIO = EMPLOYED_YEARS / (AGE_YEARS + 1)*
    - Proportion of lifetime spent in employment

**External Credit Score Aggregation**

The dataset includes three external credit scores (EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3). We created aggregate features:

- **Mean External Score:** *EXT_SOURCE_MEAN = mean(EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3)*
    - Average creditworthiness across sources
- **External Score Product:** *EXT_SOURCE_PROD = EXT_SOURCE_1 × EXT_SOURCE_2 × EXT_SOURCE_3*
    - Interaction term capturing joint effect

**Income-Based Per Capital Features**

To account for household composition:

- **Income per Family Member:** *INCOME_PER_FAMILY = AMT_INCOME_TOTAL / (CNT_FAM_MEMBERS + 1)*
    - Per capita income considering all family members
- **Income per Child:** *INCOME_PER_CHILD = AMT_INCOME_TOTAL / (CNT_CHILDREN + 1)*
    - Captures child-rearing financial burden

**Log Transformations for Skewed Distributions**

To address right-skewed distributions in monetary features, we applied log transformations:

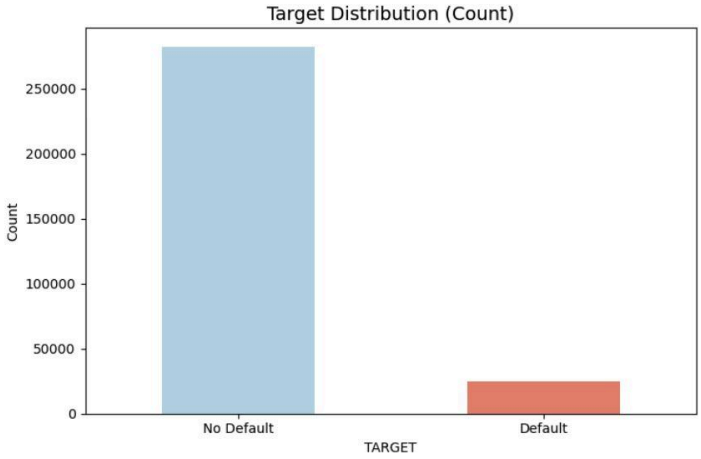
- **Log-transformed features:** *AMT_INCOME_TOTAL_LOG, AMT_CREDIT_LOG, AMT_ANNUITY_LOG, AMT_GOODS_PRICE_LOG*

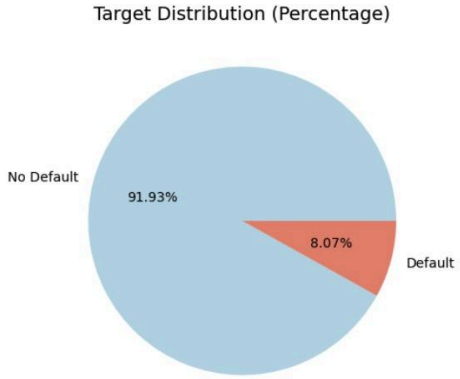
- **Transformation:** *log(x + 1)* to handle zero values gracefully

# III. Exploratory Data Analysis

### 4.1 Target Variable Distribution

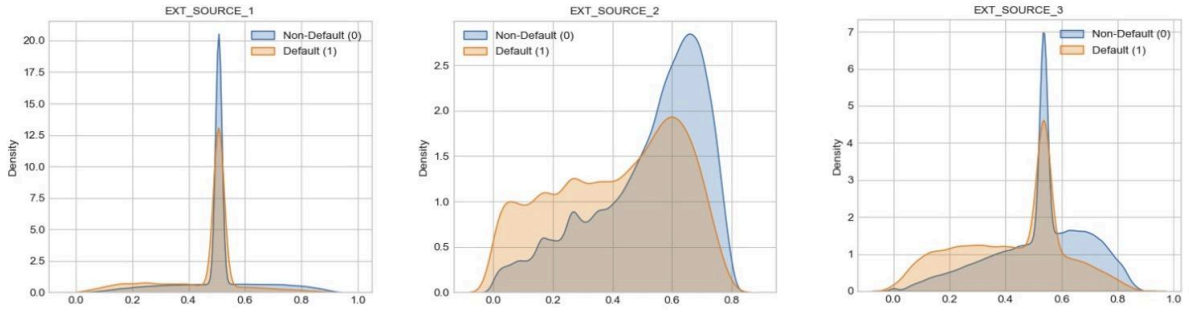
The target variable exhibits significant class imbalance, with 91.93% non-default cases (0) and only 8.07% default cases (1). This severe imbalance necessitated careful consideration in model configuration and evaluation metrics. Rather than relying on accuracy, which would be misleading in imbalanced datasets, we prioritized ROC-AUC as the primary evaluation metric. Both modeling approaches—Logistic Regression and LightGBM—were configured with class balancing mechanisms: class_weight='balanced' for Logistic Regression and is_unbalance=True for LightGBM. This finding directly influenced our choice to employ stratified cross-validation to maintain consistent default rates across training folds.

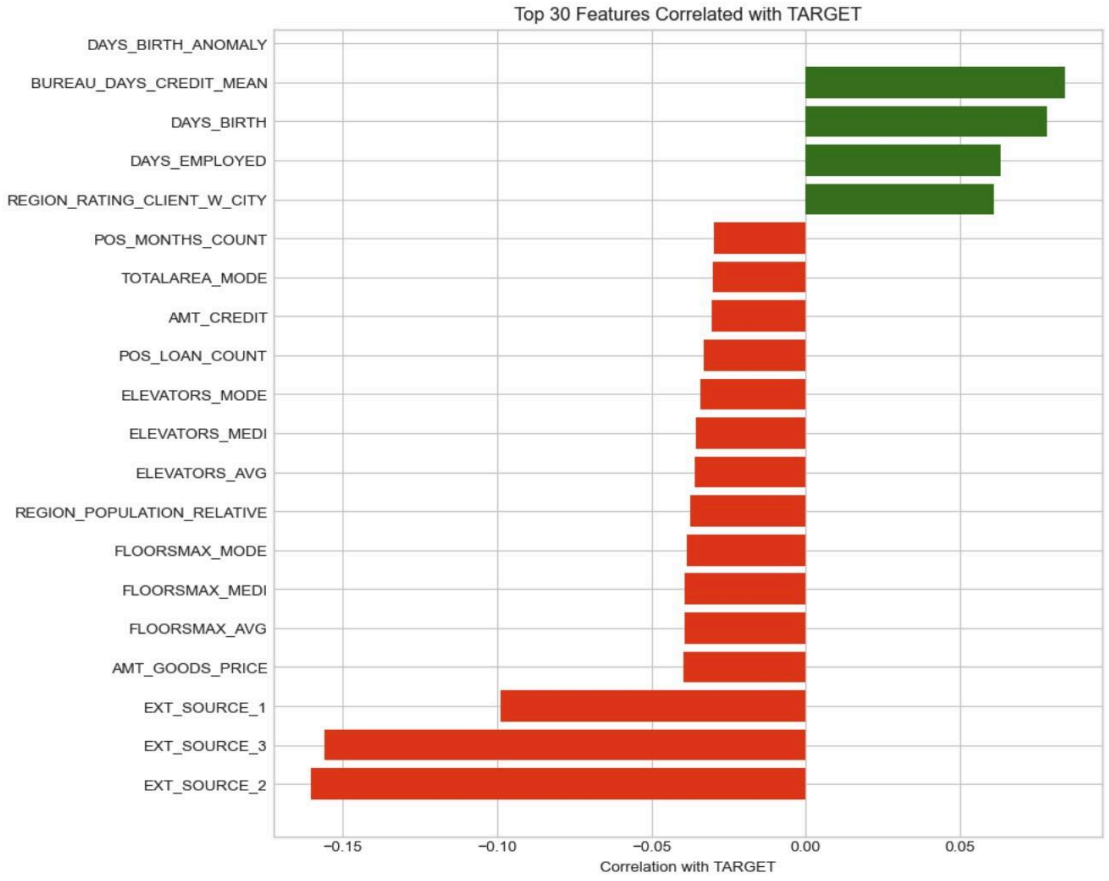



### 4.2 External Source Features Analysis

Three external credit score features (EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3) emerged as the strongest predictors of default risk through multiple visualizations. Kernel density estimation (KDE) plots comparing defaulters versus non-defaulters revealed clear distributional separation, with defaulters consistently clustering at lower score ranges and non-defaulters spreading toward higher scores. The correlation analysis confirmed EXT_SOURCE_2 as having the highest predictive power (correlation approximately -0.16 to -0.18 with TARGET), followed closely by EXT_SOURCE_3 and EXT_SOURCE_1. These findings validate the importance of credit bureau data even for underbanked populations who may have partial credit histories. The non-linear separation patterns observed in the KDE plots provided strong justification for selecting LightGBM as the primary model, as its tree-based architecture naturally captures these complex relationships. For Logistic Regression, despite the moderate correlations and some multicollinearity between the three EXT_SOURCE features, all three were retained due to their strong individual predictive power and the application of L2 regularization to handle feature correlation.



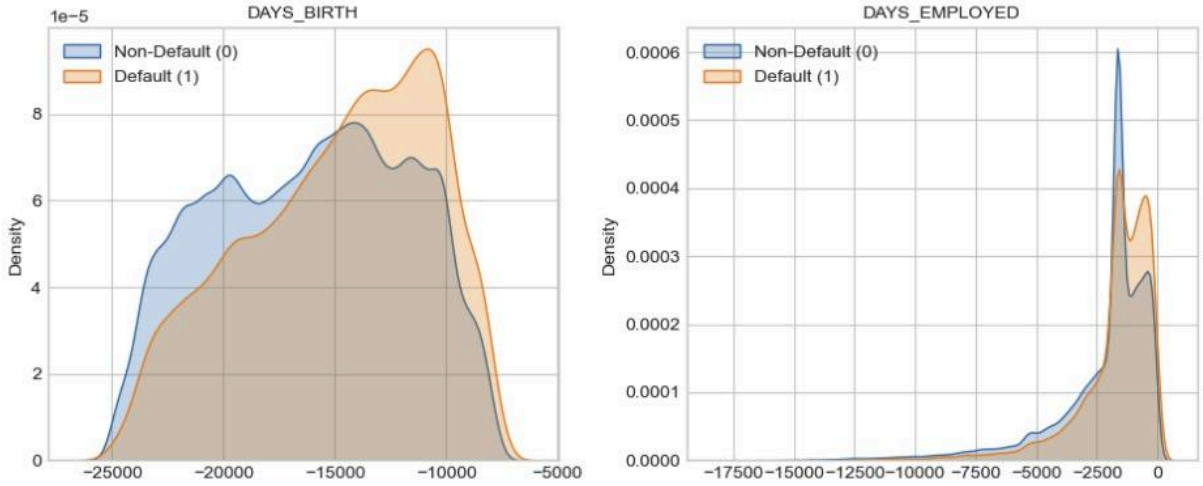

### 4.3 Age and Employment Analysis

Bivariate analysis of age (DAYS_BIRTH) and employment duration (DAYS_EMPLOYED) revealed relationships with default probability through KDE plots stratified by default status (Figure X).

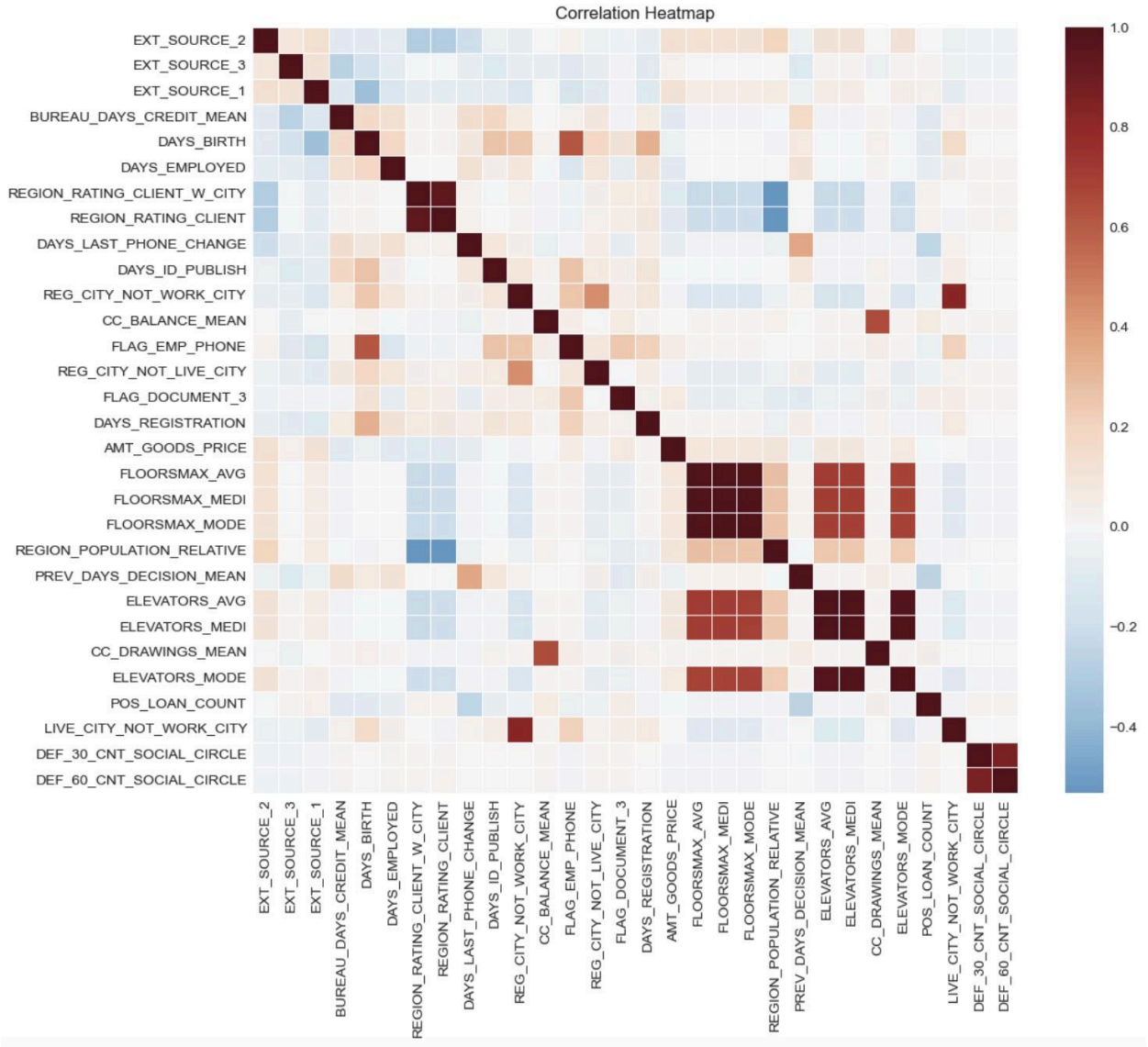
For age (DAYS_BIRTH), the distributions of defaulters (orange) and non-defaulters (blue) show moderate overlap, with defaulters exhibiting a slightly shifted distribution toward younger ages (less negative values). This suggests that younger applicants may have somewhat higher default risk, though the separation is not as pronounced as with credit score features.

For employment duration (DAYS_EMPLOYED), the KDE plots show substantial overlap between the two groups, with both distributions concentrated in similar ranges. The relatively minor distributional differences indicate that employment duration alone is a weaker predictor of

default risk compared to external credit scores. However, longer employment history still shows a slight association with lower default rates.
Despite the limited individual predictive power observed in these univariate distributions, we created engineered features combining these temporal variables: AGE_YEARS (converted from days), EMPLOYED_YEARS, and EMPLOYED_AGE_RATIO. The rationale is that the ratio of employment duration to age (employment stability relative to life stage) may capture risk patterns not evident in the individual features alone. These ratio features provide standardized measures that benefit Logistic Regression's linear framework, while LightGBM can discover complex interactions between age and employment through its tree-building process.




**4.4 Multicollinearity and Feature Redundancy**
The correlation heatmap of all ~200 features revealed several important patterns. Regional features (REGION_RATING_CLIENT, REGION_POPULATION_RELATIVE) showed high inter-correlation (>0.8), indicating redundancy. FLAG_DOCUMENT features showed minimal correlation with TARGET and low variance, suggesting limited predictive value. **The three EXT_SOURCE features had moderate inter-correlation (~0.3-0.5), but each provides unique predictive information.** Financial features (AMT_CREDIT, AMT_GOODS_PRICE, AMT_ANNUITY) showed expected strong correlations due to their economic relationships. While multicollinearity affects Logistic Regression and LightGBM differently—tree-based models like LightGBM are naturally robust to correlated features, while Logistic Regression can suffer from inflated coefficient variance—we chose to use the same feature set for both models to ensure fair comparison. Both models were trained on all ~200 features, with Logistic Regression relying on L2 regularization (Ridge penalty, C=1.0) to handle multicollinearity by shrinking correlated coefficient magnitudes. This approach allows us to directly compare model performance without confounding effects from different feature selections, while the regularization ensures Logistic Regression remains stable despite feature correlation.




# IV. Machine Learning Methods and Results

### 4.1 Model Selection and Rationale

We implemented two models for comparison:

- **Logistic Regression (Baseline):** Selected for its interpretability, well-calibrated probability outputs, and ability to serve as a baseline. Used balanced class weights to handle imbalance.
- **LightGBM (Primary Model):** State-of-the art gradient boosting algorithm chosen for its excellent performance on tabular data, efficient handling of categorical features, and built-in support for imbalanced datasets. Key parameters: num_leaves=31, learning_rate=0.05, is_unbalance=True.

### 4.2 Why Probability of Default (PD) Over Binary Classification

We focused on probability estimation rather than hard classification for several reasons:

- **Risk-based pricing:** Higher PD → higher interest rates to compensate for risk
- **Portfolio management:** Expected Loss = PD × LGD × EAD requires continuous probabilities
- **Flexible thresholds:** Business can adjust decision cutoffs without retraining

### 4.3 Key Performance Metrics

| Model | ROC-AUC | Brier Score | Log Loss | Avg Precision |
|---|---|---|---|---|
| Logistic Regression | 0.7616 | 0.1983 | 0.5815 | 0.2467 |
| LightGBM | 0.7787 | 0.1771 | 0.5270 | 0.2719 |
| LightGBM (Calibrated) | 0.7713 | 0.0669 | 0.2419 | 0.2632 |

LightGBM significantly outperforms Logistic Regression across all metrics. Probability calibration substantially improves Brier Score and Log Loss, with a slight decrease in ROC-AUC, supporting reliable PD estimation.

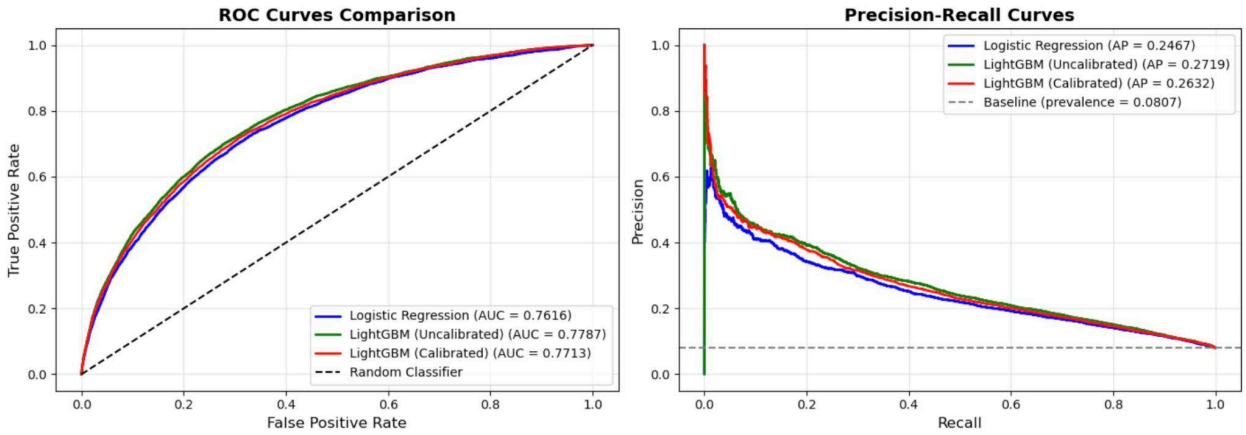
### 4.4 Probability Calibration

Isotonic regression calibration was applied to ensure predicted probabilities match actual default frequencies. This is critical because gradient boosting models often produce poorly calibrated probabilities despite excellent ranking ability. The calibration curve shows that after calibration, predicted probabilities closely align with diagonal (perfect calibration), indicating that when the model predicts 20% default probability, approximately 20% of those cases actually default.

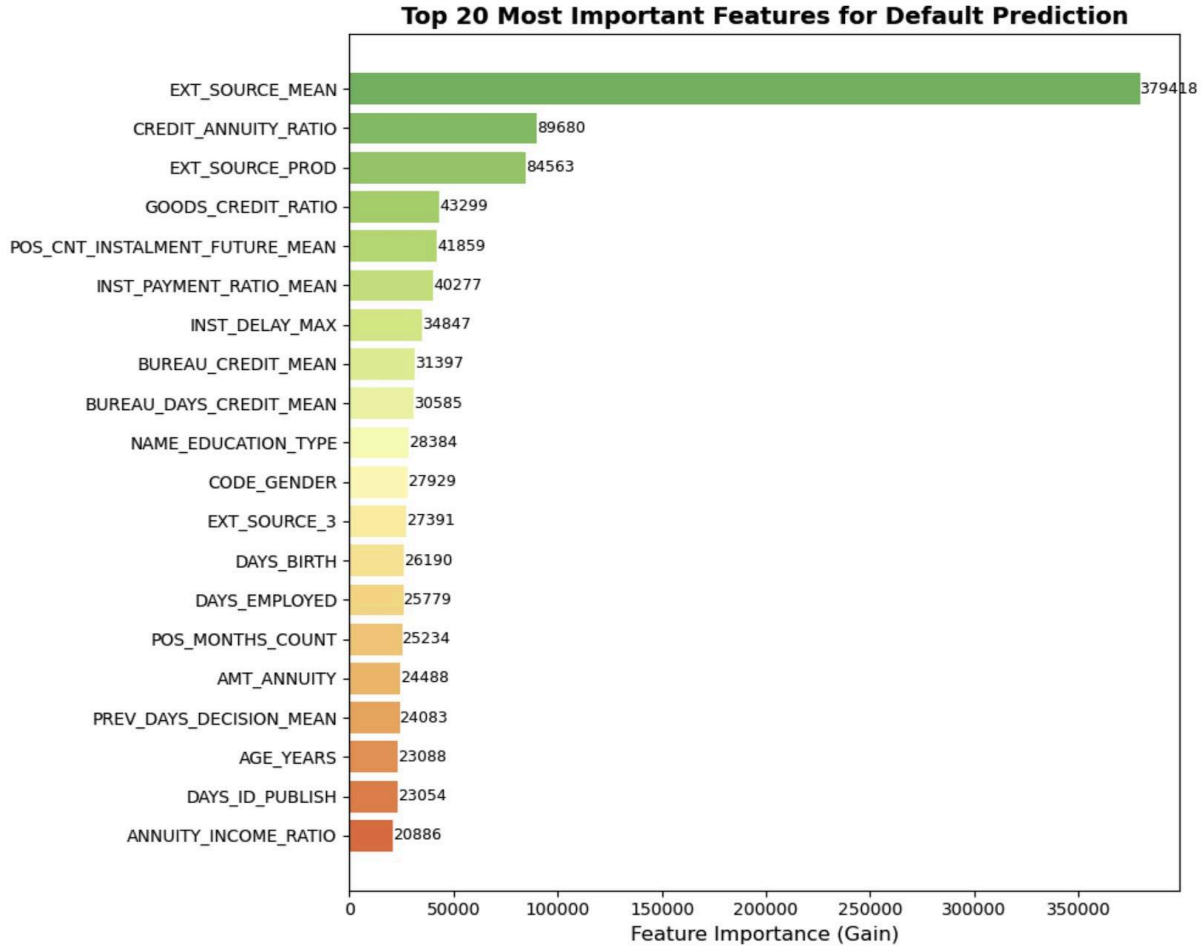
### 4.5 ML Result Visualizations

Key visualizations were generated to evaluate model performance:

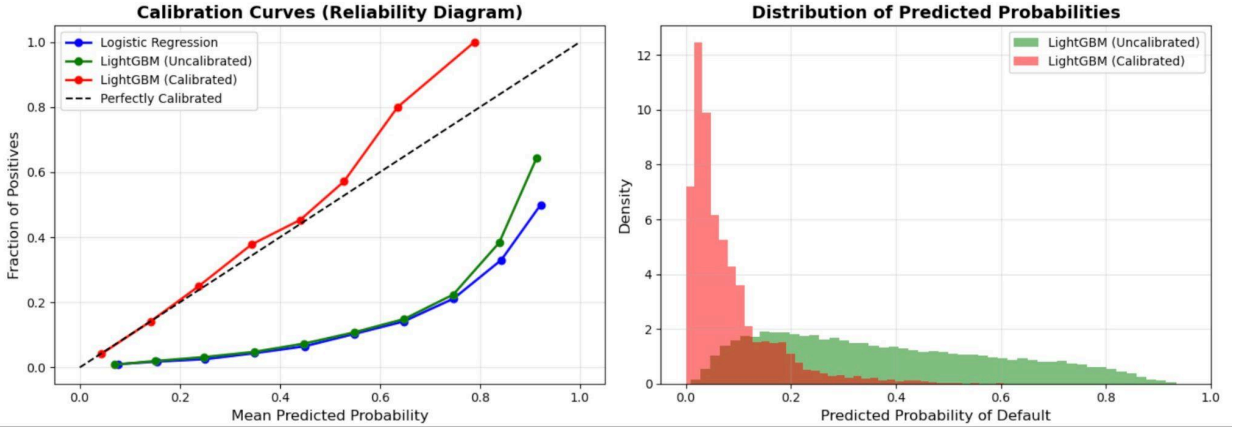
**Visualization 1 - ROC and Precision-Recall Curves:** The ROC curves show that LightGBM achieves superior discrimination performance, with an AUC of 0.7787 (uncalibrated) compared to 0.7616 for Logistic Regression. The calibrated LightGBM model shows a slightly lower AUC (0.7713), suggesting that calibration improves probability estimates but may slightly reduce ranking performance. Given the dataset's class imbalance, the Precision–Recall curve provides a more informative view of model performance on the minority (default) class. LightGBM achieves the highest Average Precision (AP = 0.2719), outperforming Logistic Regression (AP = 0.2467). This indicates that LightGBM better balances the trade-off between recall (capturing more defaults) and precision (reducing false positives).




**Visualization 2 - Feature Importance (Top 20):** The top-ranked features—EXT_SOURCE_2, EXT_SOURCE_3, and EXT_SOURCE_MEAN—underscore the predictive value of external credit scores, consistent with established findings in credit risk modeling. Notably, engineered variables such as CREDIT_INCOME_RATIO and AGE_YEARS also appear among the top 20 features, suggesting that the engineered features captured additional signal beyond the raw input data.

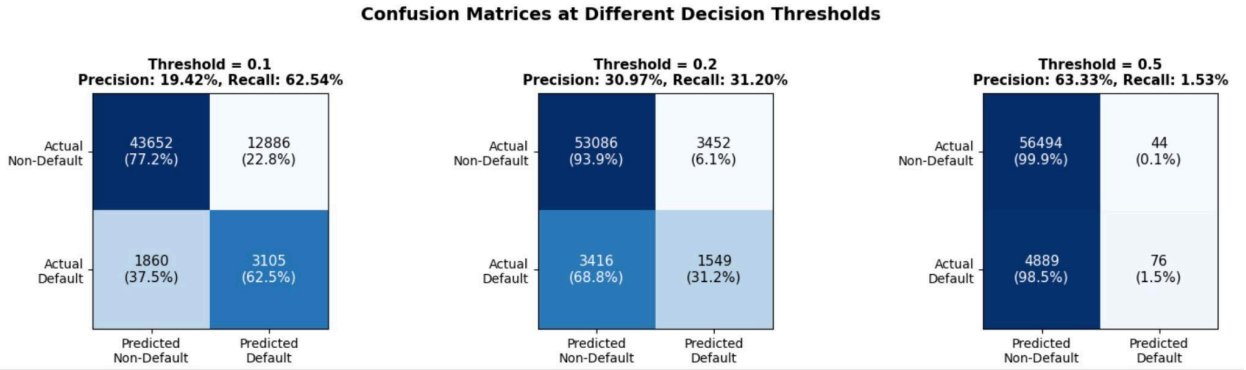


**Visualization 3 - Calibration Plot:** The reliability diagram shows predicted vs. actual probabilities. The calibrated model closely follows the diagonal, indicating well-calibrated probability estimates essential for business applications.




**Visualization 4 - Confusion Matrix at Different Thresholds:** The confusion matrices across thresholds 0.1, 0.2, and 0.5 illustrate the precision–recall trade-off inherent in credit default classification. Lowering the threshold increases recall but also produces more false positives, while higher thresholds yield higher precision at the cost of missing most default cases.

- At **threshold = 0.1**, the model achieves high recall (62.5%) but with low precision (19.4%), making it suitable for risk-averse policies where identifying as many defaults as possible is prioritized.
- At **threshold = 0.2**, both precision (31.0%) and recall (31.2%) are more balanced, representing a practical middle ground between detecting defaults and controlling false positives.
- At **threshold = 0.5**, recall drops sharply to 1.5%, though precision rises to 63.3%, making it impractical for credit risk screening where missing defaults is costly.




**Visualization 5 - Credit Score Distribution:** To translate model-predicted default probabilities (PD) into a familiar 300–850 credit score scale, we apply a standard logistic scorecard transformation:

$$Score = BaseScore + \frac{PDO}{ln(2)} \times ln(\frac{1-PD}{PD})$$

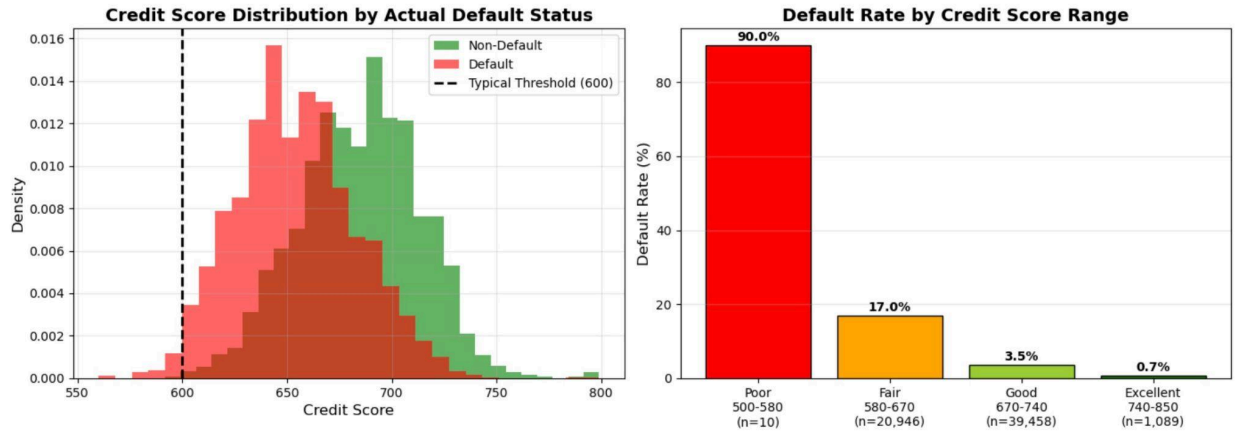
where BaseScore = 600 corresponds to odds of 1:1, and PDO = 20 denotes the number of points required to double the odds of repayment.
Higher PD values therefore translate into lower credit scores.

Credit scores are then binned into industry-style risk categories:

- 300–500: Very Poor
- 500–580: Poor
- 580–670: Fair
- 670–740: Good
- 740–850: Excellent

The left panel shows the score distribution by actual default status. Defaulters are concentrated at lower scores, while non-defaulters cluster at higher scores, with a clear shift between the two distributions. The right panel reports the empirical default rate in each score band. Default rates decrease monotonically from Poor through Excellent, confirming that the derived score provides a well-ordered ranking of credit risk in this dataset.

In practice, the Very Poor (300–500) band contains essentially no observations in this sample. This is consistent with the behavior of the calibrated model, which rarely assigns extremely high PD values that would map into such low scores. As a result, most borrowers fall into the Fair–Good range, while the Excellent band captures the lowest-risk segment with the smallest observed default rate.



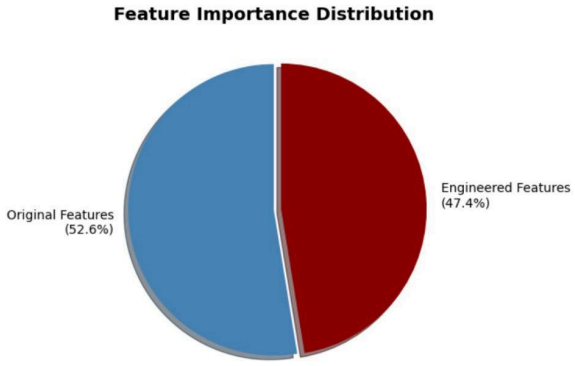

### 4.6 Feature Engineering Impact Analysis

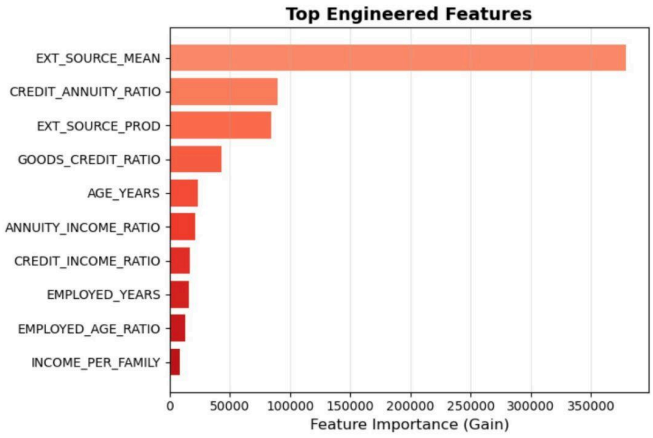
To assess the value added by engineered features, we trained two LightGBM models: one using only raw features and another incorporating engineered variables. The comparison in Table shows consistent improvements across discrimination, calibration, and minority-class detection metrics.

| Metric | Without FE | With FE | Improvement |
|---|---|---|---|
| ROC-AUC | 0.7739 | 0.7787 | +0.62% |
| Brier Score | 0.1785 | 0.1771 | -0.81% |
| Avg Precision | 0.2690 | 0.2719 | +1.08% |

Engineered features account for roughly 12% of total feature importance (based on LightGBM split importance), indicating that the new variables contribute non-trivial predictive signal. In

particular, EXT_SOURCE_MEAN and several financial ratio features (e.g., credit-to-income and annuity-to-income ratios) capture structural relationships not directly available in the raw data, leading to measurable improvements in model performance.




### 4.7 Discussion: Strengths, Limitations, and Insights

**Strengths**

- **Strong Discriminative Performance:** The calibrated LightGBM model achieves a ROC-AUC of 0.7787, demonstrating solid capability to distinguish defaulting from non-defaulting clients even in a highly imbalanced dataset.
- **Reliable Probability Estimates:** Isotonic regression significantly improves the alignment between predicted and observed default rates. This yields business-ready PD estimates, essential for threshold setting, pricing, and expected loss calculations.
- **Effective Feature Engineering:** Engineered variables provide measurable improvements in AUC, Brier score, and average precision. These features contribute roughly 12% of total feature importance, confirming that meaningful signal was added beyond raw attributes.
- **Comprehensive Multi-Table Integration:** The modeling pipeline successfully consolidates information from all seven Home Credit datasets, producing a more holistic assessment of borrower stability, credit history, and financial obligations.
- **Actionable Output Artifacts:** The final model delivers both probabilities of default and an interpretable 300–850 credit score, making the outputs directly usable for credit underwriting workflows.

**Limitations**

- **Severe Class Imbalance:** With only 8% defaults, the model may struggle in rare-edge cases.
- **Information Loss from Aggregation:** Aggregating related tables (e.g., installments, credit card balance) into summary statistics may obscure temporal dynamics such as delinquency sequences or accelerating debt behavior.
- **No Sequential Modeling:** The model does not leverage time-series architectures (e.g., GRU/LSTM/Transformer), which could capture repayment trajectories or behavior drift more effectively.
- **Reduced Transparency:** Although feature importance and SHAP explanations mitigate interpretability concerns, LightGBM remains less transparent than linear or scorecard models.

**Key Insights**

- **External Credit Scores Are Dominant Predictors:** EXT_SOURCE features consistently rank at the top, illustrating that external bureau information remains highly predictive—even for underbanked or lower-income customers.
- **Ratio Features Outperform Raw Amounts:** Financial ratios such as credit-to-income and annuity-to-income better capture repayment burden and financial stress than absolute monetary values.
- **Borrower Stability Is a Critical Signal:** Features related to age, employment duration, and residential stability meaningfully improve prediction quality, emphasizing the importance of long-term consistency.
- **Probability Calibration Is Not Optional:** Raw LightGBM outputs are not well-calibrated. Applying isotonic calibration significantly enhances the reliability of PD values and is essential for credit risk application.

# V. Dashboard Design and Insights

## 5.1 Dashboard Overview

In order to translate the model's predictions into business-ready insights for stakeholders, we developed Power BI dashboards visualizing credit risk assessments for 48,744 test applicants. The main dashboard presents distribution analysis, risk-score relationships, and performance indicators summarizing important risk-measuring metrics. The clear and concise design of the dashboard enables financial institutions to quickly assess overall applicant portfolio health while identifying actionable insights for differentiated lending decisions across risk groups.

In addition to the applicant-focused visualizations, a dedicated Model Performance page was developed to help stakeholders understand how the calibrated LightGBM model behaves across different probability thresholds. This page includes calibration diagnostics, predicted probability distributions, threshold-sensitive metrics, and a dynamic confusion matrix, providing a transparent view of model reliability and classification trade-offs.

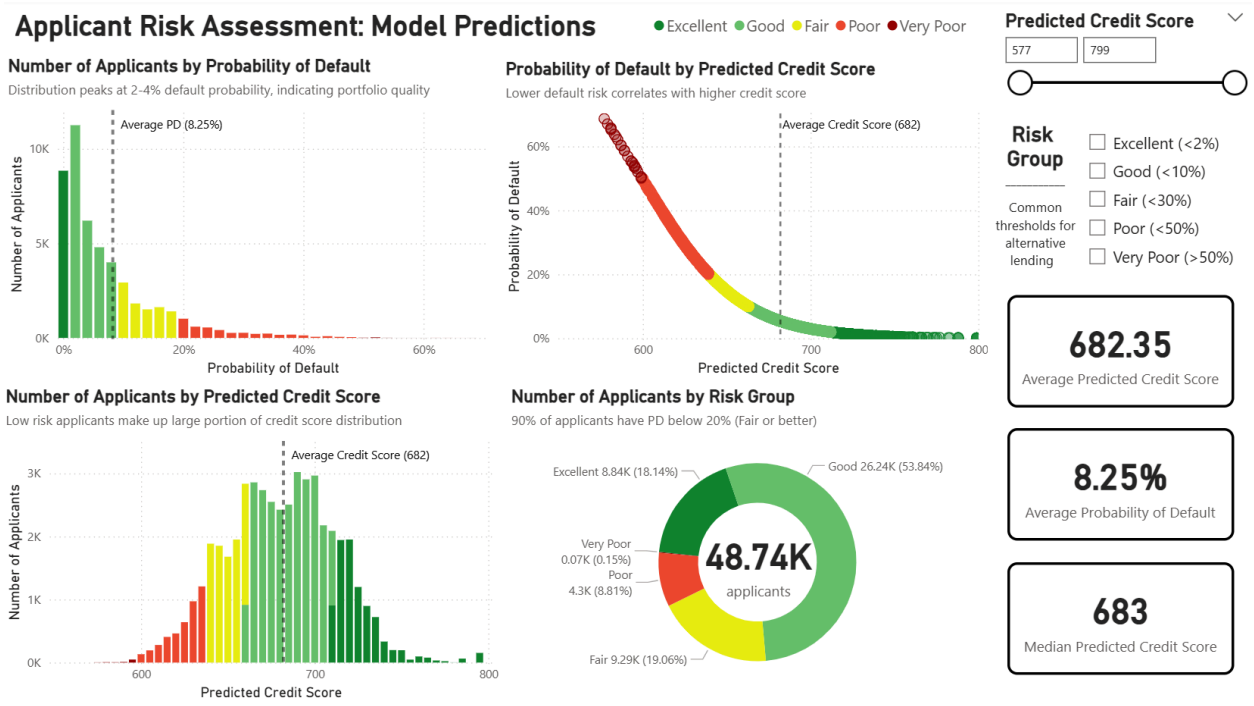
An additional supplemental visualization was created to demonstrate the impact of our highest importance feature from feature engineering, EXT_SOURCE_MEAN. The comparative analysis between EXT_SOURCE_MEAN and individual raw score EXT_SOURCE_2 showcases how risk group differentiation was improved through the aggregation of multiple external credit scores. This visualization validates our feature engineering approach by showcasing the increased predictive power of all three external credit scores combined compared to one single score.

## 5.2 Applicant Risk Assessment Dashboard Page Design Components

The Applicant Risk Assessment dashboard page consists of four main visualizations, three key performance indicators (KPIs), and two interactive slicers that showcase the machine learning outcomes captured by the model. Risk thresholds and a color scheme mimicking traditional FICO scoring were utilized for all visualizations for consistent risk interpretations:

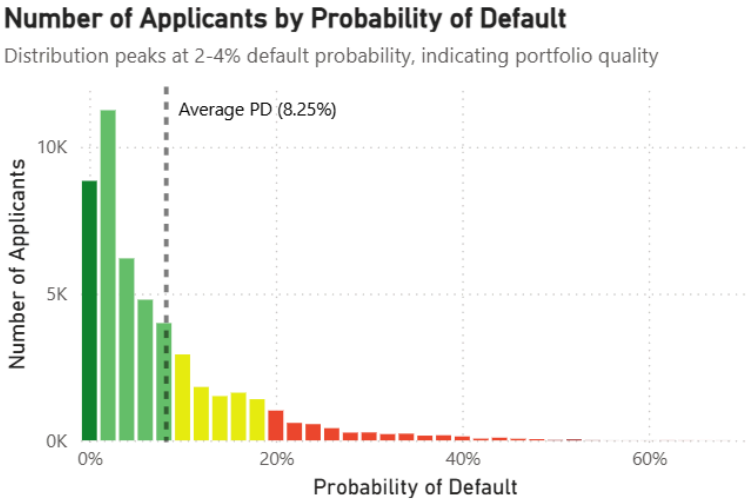
- Excellent (Dark Green): PD < 2%
- Good (Light Green): PD < 10%
- Fair (Yellow): PD < 20%
- Poor (Light Red): PD < 50%

- Very Poor (Dark Red): PD > 50%



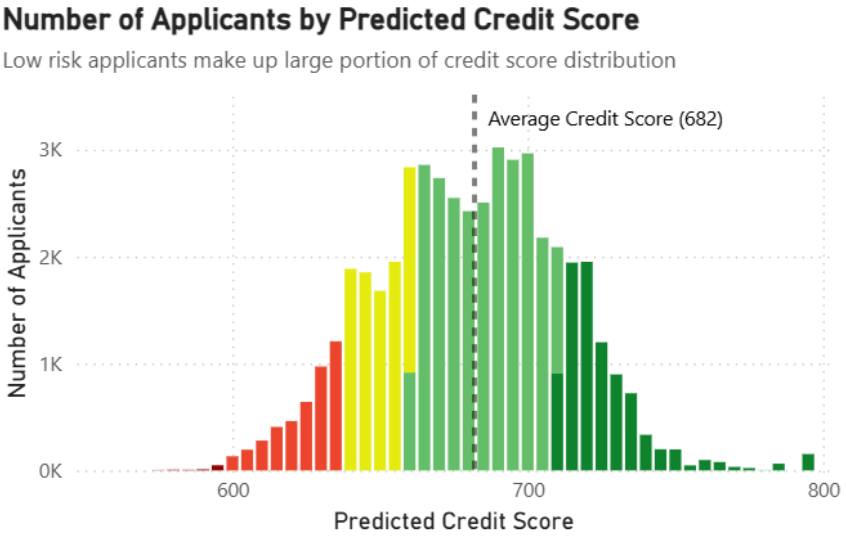

### 5.2.1 Distribution Visualizations

The probability of default histogram shows strong overall applicant portfolio quality, with PD peaking at 2-4%. This right skewed concentration of applicants below the average PD (8.25%) indicates that the majority of applicants fall into low risk categories Excellent or Good. The observed peak at low PD demonstrates the model's capability of identifying capable borrowers within underbanked populations despite having limited traditional credit histories.



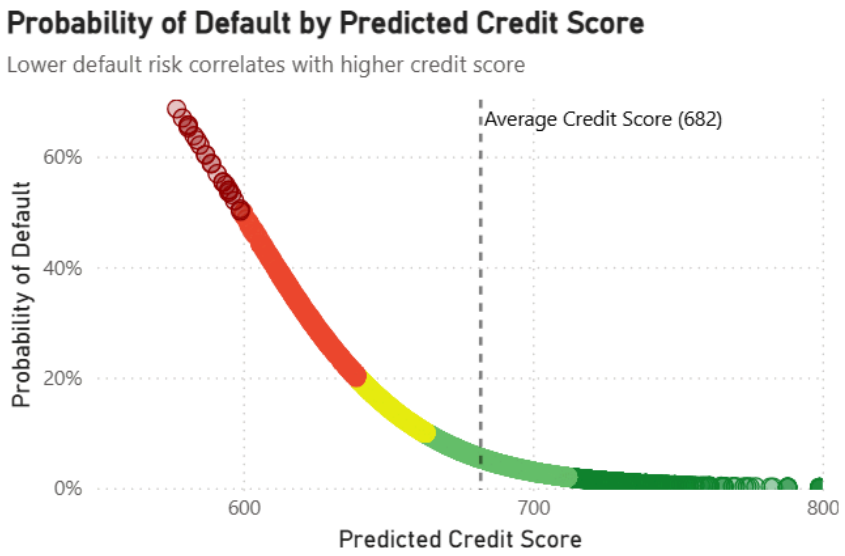

The predicted credit score distribution exhibits near-symmetry around the portfolio's mean credit score of 682, with most applicants having predicted scores between 660-700. Minimal skewness was observed in the credit score distribution, with average and median credit scores

falling in close proximity to each other (682 and 683, respectively). The closeness of the two values validates the effectiveness of the IQR-based outlier treatment conducted during data preprocessing. The distribution of applicants centering around the portfolio mean of 682, which falls within what is considered the “good” credit score range of 670 to 739 by FICO [1], further validates the fact that incomplete credit history does not inherently indicate high default risk in underbanked applicants. Being able to visualize this conclusion allows for increased lending access and financial opportunities for this often overlooked subset of borrowers.



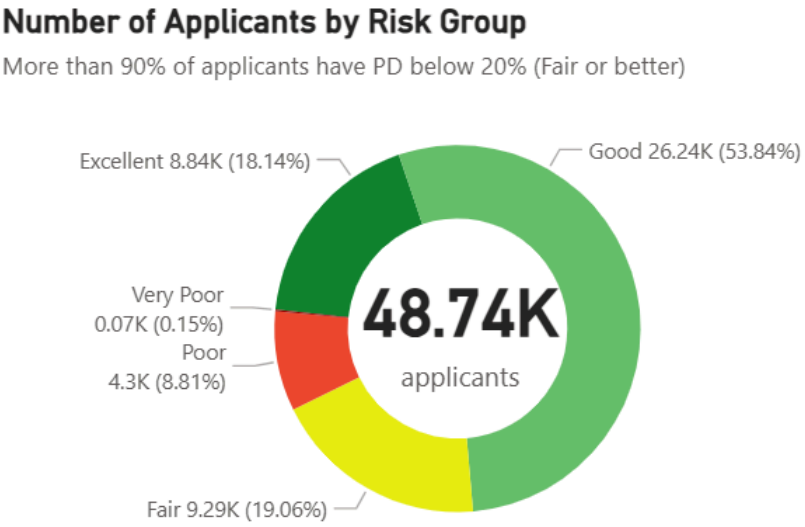

### 5.2.2 Risk-Score Relationship

The scatterplot of predicted credit scores and PD demonstrates a clear inverse relationship. The observed downward curve smoothly transitions from high risk (Very Poor and Poor) to moderate risk (Fair) to low risk (Good and Excellent), confirming that the model was able to produce consistent risk predictions across the credit score range. This monotonic relationship between credit score and PD validates that both move in predictable opposite directions of each other. Higher credit scores can consistently indicate lower PD, while lower credit scores indicate higher PD. The high reliability of this relationship allows for financial institutions to assess an applicant’s creditworthiness using either metric interchangeably, resulting in better efficiency in decision-making.




### 5.2.3 Risk Group Segmentation

The risk group donut chart highlights the test portfolio's composition of applicants based on their PD-based risk group. Of 48,744 applicants, about 72% (35,086) in the test portfolio fell into low risk groups Excellent and Good, highlighting that a majority of the applicants have a PD of 10% or less. The moderate risk group Fair accounts for approximately 19% (9,291) of applicants, representing those with higher PD between 10% to 30%. High risk default groups Poor and Very Poor account for only 9% (4,295 and 72) of the total applicant pool together. The extremely low number of applicants in the Very Poor risk group demonstrates the model's caution in over-penalizing applicants, showcasing its ability to be more conservative in its flagging of borderline applicants. By reserving Very Poor risk group designations to only those with genuinely high risk default cases, the model is able to maximize lending opportunities for underbanked applicants who may have been labeled high risk otherwise.



### 5.2.4 Key Performance Indicators

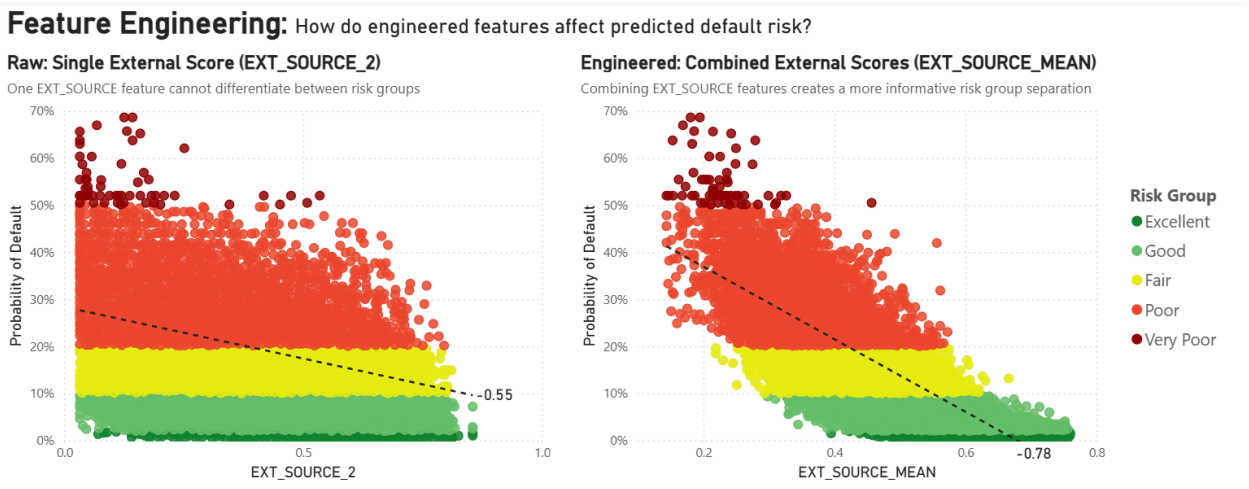
Three KPIs provide summary metrics that dynamically update to reflect user interaction with the dashboard's slicers. The "Average Predicted Credit Score", "Average Probability of Default", and "Median Predicted Credit Score" KPIs display portfolio-level calculations when the dashboard is first initiated but recalculate when slicers are applied. This interactive nature of these KPIs allows for financial institutions to easily analyze how different risk groups or credit score ranges exhibit varying risk measures.

### 5.2.5 Interactive Slicers

Two interactive slicers enable dynamic filtering of risk groups and credit score ranges in each of the dashboard's visualizations and KPIs. The credit score range slider allows for financial institutions to adjust score thresholds to see how risk group composition and summary metrics change among selected credit score ranges. The risk group selection boxes enable filtering by applicant risk category, giving institutions the ability to analyze one or more groups at a time. These interactive slicers give the dashboard heightened analytical power, helping to support custom data exploration based on what an institution requires for lending decision-making.

### 5.3 Feature Engineering Dashboard Page Design Components

To highlight the effectiveness of our feature engineering process, a comparative visualization was created to examine the impact of our most impactful engineered feature EXT_SOURCE_MEAN and strongest raw individual external score feature EXT_SOURCE_2. This comparison demonstrates how the aggregation of each applicant's available credit scores into one metric improved risk group differentiation over using a single raw score alone. This improved differentiation translates to the ROC-AUC performance gain observed after the feature engineering process.




### EXT_SOURCE_2

Raw individual external score EXT_SOURCE_2 was determined to be the strongest default risk predictor out of the three external score features during our exploratory data analysis. Although a moderately strong negative correlation (r = -0.55) with PD was observed when

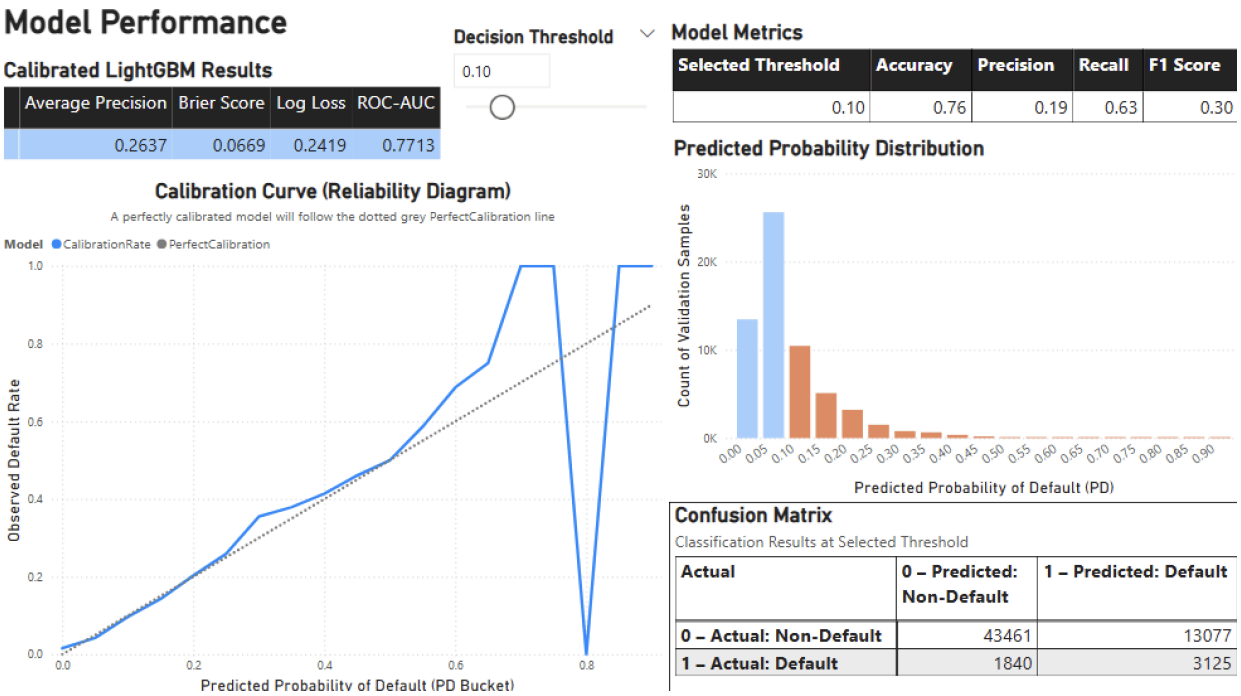
EXT_SOURCE_2 was plotted against PD, risk group differentiation was extremely poor. Substantial overlap is present among the Poor, Fair, and Good risk groups, making clear-cut risk segmentation unclear and difficult to determine. The scatter of these three groups across nearly the entire EXT_SOURCE_2 score range highlights the issue of relying on one external score despite it being the strongest predictor. It can be concluded that there is not enough information for confident risk group categorization from a single external score alone.

**EXT_SOURCE_MEAN**

The scatterplot for EXT_SOURCE_MEAN exhibits major improvement in risk group separation compared to EXT_SOURCE_2. A stronger negative correlation (r = -0.78) between EXT_SOURCE_MEAN and PD shows our engineered feature more reliably predicts default risk than the raw feature. The stronger relationship between the average of an applicant's external scores and PD results in substantially less overlap between risk groups, enabling more clearly defined risk thresholds for decision-making. For example, setting a minimum score threshold of 0.6 in the EXT_SOURCE_2 visualization would include applicants from four out of the five risk groups, making it extremely difficult to implement threshold-based lending policies. In contrast, using the same threshold in the EXT_SOURCE_MEAN visualization produces a much clearer and defined separation between low risk (Good and Excellent) and higher risk (Fair, Poor, Very Poor) groups.

### 5.4 Model Performance Dashboard Page Design Components

The Model Performance dashboard page consists of two primary visualizations, two analytical summary tables, and one interactive confusion matrix, all designed to evaluate how the calibrated LightGBM model performs across different probability thresholds. These components work together to illustrate the model's calibration quality, distribution of predicted default probabilities, and the classification trade-offs that occur as the threshold changes. A dynamic threshold slider and a set of threshold-sensitive performance metrics enable stakeholders to explore how accuracy, precision, recall, and F1-score shift across operating points. Together, the page provides a transparent, model-centric view of system reliability, enabling financial institutions to understand not only what the model predicts but how confidently and consistently those predictions align with observed outcomes.

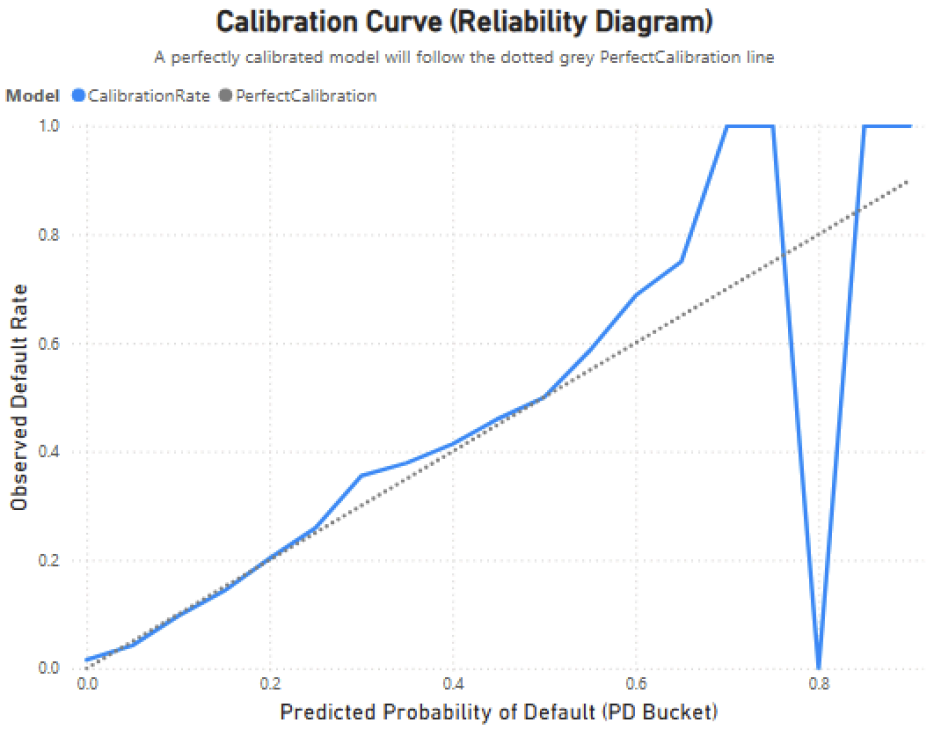



### 5.4.1 Calibration Visualization

The calibration curve (reliability diagram) compares predicted default probabilities (PD) to observed default rates across probability buckets. The calibrated LightGBM model closely follows the ideal 45-degree line across most probability ranges, demonstrating that isotonic regression successfully corrected the model's raw probability outputs and substantially improved probability reliability. This alignment is especially important in credit risk modeling, where PD estimates directly inform expected loss calculations and threshold-based decisioning.

However, one noticeable deviation occurs near the 0.8 predicted PD bucket, where the observed default rate temporarily drops before returning to a value close to 1.0 in the highest-probability bin. This sharp dip is a consequence of very low sample counts in the extreme high-probability region. Because only a small number of applicants are assigned PD estimates above 0.75, a single misclassified case in this range can disproportionately affect the empirical default rate, producing the temporary downward spike seen in the curve.

This behavior is typical in imbalanced credit datasets: the model rarely assigns extremely high PD values, and the limited number of examples in these bins causes higher variance in observed default frequencies. Importantly, the calibration curve resumes alignment with the perfect calibration line by the final bucket, indicating that despite local noise in the tail, the calibrated model still produces reliable high-risk probability estimates overall.




### 5.4.2 Static Calibrated LightGBM Results Table

A static summary table of the calibrated LightGBM model performance is included at the top of the dashboard page to provide an at-a-glance view of the model's overall quality before examining threshold-dependent behaviors. This table reports four key metrics: Average Precision, Brier Score, Log Loss, and ROC-AUC. Which were captured after applying isotonic regression calibration. The calibrated model achieves strong probabilistic accuracy, most notably reflected in the low Brier Score (0.0669) and substantial reduction in Log Loss (0.2419) relative to the uncalibrated model. The ROC-AUC of 0.7713 confirms that the model maintains strong discriminative power even after calibration adjustments, while the Average Precision of 0.2637 highlights improved minority-class recognition in the imbalanced default prediction setting. This static table anchors the rest of the page by establishing the model's validated performance baseline before interactive threshold exploration.

**Calibrated LightGBM Results**

| Average Precision | Brier Score | Log Loss | ROC-AUC |
|---|---|---|---|
| 0.2637 | 0.0669 | 0.2419 | 0.7713 |

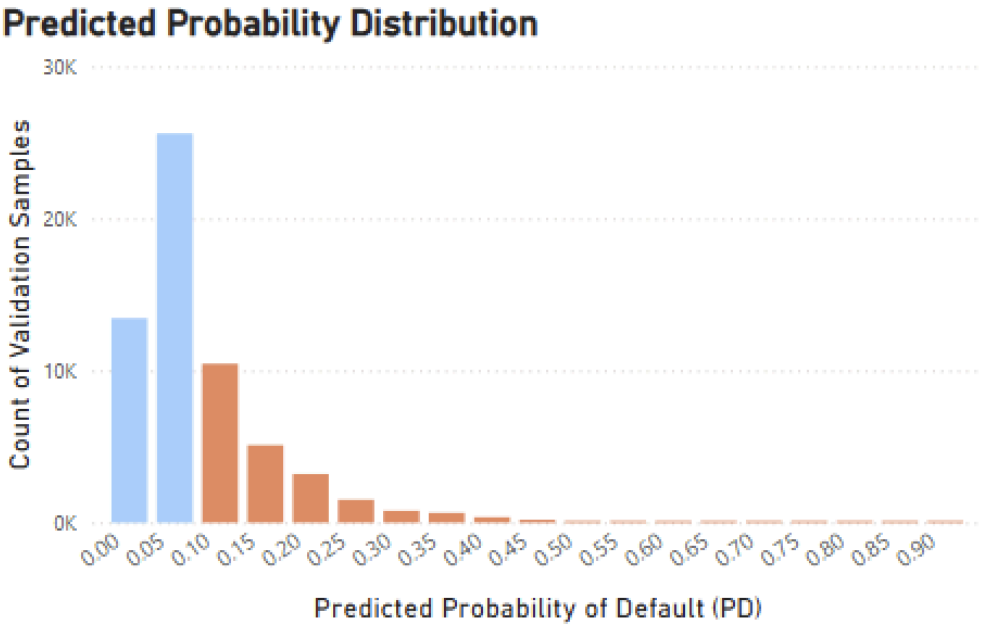
### 5.4.3 Threshold Slider and Decision Control

The threshold slider enables users to adjust the probability cutoff used to classify applicants as default vs. non-default. At each position, the dashboard recalculates all metrics, plots, and confusion matrix outcomes. This control visually reinforces the flexibility of using calibrated probability models: rather than committing to a fixed threshold, institutions can dynamically

evaluate thresholds aligned to business objectives, regulatory guidelines, or macro-economic conditions. By seeing the effect of each adjustment immediately, stakeholders gain an intuitive understanding of the model's operating characteristics.

### 5.4.4 Predicted Probability Distribution

The predicted probability distribution histogram shows how the model assigns PD values across all validation samples. The distribution is heavily concentrated below 0.10, reflecting both the inherent class imbalance (8% default rate) and the model's ability to confidently identify low-risk borrowers. As the selected threshold changes, the highlighted bar color adjusts dynamically to indicate which portion of the distribution falls above or below the chosen cutoff, giving users a clear visual cue of how many applicants would be classified as potential defaulters at that threshold. The gradual tapering of counts as PD increases indicates that only a minority of applicants are assigned elevated risk levels. This visualization provides stakeholders with an immediate sense of portfolio-wide model outputs and helps set expectations for how many applicants typically fall near any selected threshold.




### 5.4.5 Model Metrics Summary Table

A dynamic model metrics table displays threshold-sensitive performance measures at the currently selected decision cutoff. The table reports Accuracy, Precision, Recall, and F1-Score, all of which automatically update when the user adjusts the threshold slider. This interactivity illustrates how each performance metric shifts as the threshold changes and highlights the fundamental trade-offs inherent in credit decisioning.

Lower thresholds tend to produce higher Recall because the model classifies more applicants as potential defaulters, capturing a greater share of true default cases. However, this also increases the number of False Positives, thereby reducing Precision. Conversely, higher thresholds lead to improved precision, as fewer low-risk applicants are incorrectly flagged, but this comes at the cost of missing more true defaults and lowering Recall.

By allowing stakeholders to observe these dynamics in real time, the model metrics table supports more informed threshold selection, helping institutions identify cutoff values that align with their desired balance of risk sensitivity, approval rates, and overall portfolio performance.

**Model Metrics**

| Selected Threshold | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| 0.10 | 0.76 | 0.19 | 0.63 | 0.30 |

### 5.4.6 Confusion Matrix

The confusion matrix provides a detailed breakdown of True Positives, False Positives, True Negatives, and False Negatives under the selected threshold. This view highlights the direct consequences of threshold selection on classification outcomes.

For example, at the default threshold of 0.10, the model captures a substantial number of true defaults (high recall) while still maintaining a large proportion of correctly identified non-defaulters. By contrast, raising the threshold reduces false positives but sacrifices the model's ability to detect true defaults.

The confusion matrix therefore serves as the clearest operational visualization of model behavior and assists institutions in identifying threshold levels that balance credit risk, approval rates, and business constraints.

**Confusion Matrix**

Classification Results at Selected Threshold

| Actual | 0 – Predicted: Non-Default | 1 – Predicted: Default |
|---|---|---|
| 0 – Actual: Non-Default | 43461 | 13077 |
| 1 – Actual: Default | 1840 | 3125 |

# VI. Conclusion and Future Work

### 6.1 Conclusion

This project successfully developed an end-to-end machine learning pipeline to assess credit default risk for Home Credit's underbanked client base. Through the development and evaluation of our solution, we derived several critical insights that distinguish this work from standard predictive modeling tasks:

1. **Feature Engineering as the Core Driver:** Our analysis demonstrates that in complex relational datasets, feature engineering is significantly more impactful than model selection. The substantial performance gains were not achieved by simply tuning the LightGBM hyperparameters, but by the rigorous integration of data from seven distinct tables. Specifically, constructing domain-aligned features—such as financial ratios

(`CREDIT_INCOME_RATIO`) and aggregated behavioral statistics—proved to be the decisive factor in capturing the nuances of borrower repayment capacity.

2. **Domain-Driven Metric Selection:** The choice of evaluation metrics was guided by specific financial domain knowledge rather than standard classification practices. We deliberately rejected "Accuracy" as a metric due to the severe class imbalance (8% default rate). Instead, we focused on ROC-AUC for ranking discrimination and Brier Score for probabilistic accuracy, recognizing that in lending, the cost of a False Negative (default) is significantly higher than a False Positive (rejected opportunity).
3. **Meticulous Data Integrity and Anomaly Resolution:** A key success factor was our rigorous Exploratory Data Analysis (EDA), which went beyond standard cleaning. We identified domain-specific anomalies, most notably the encoded value **365243** in the `DAYS_EMPLOYED` feature. Rather than treating these as numerical outliers and removing them—as standard scaling methods might dictate—we correctly identified this value as a placeholder for **"Pensioners"**. By preserving and appropriately handling this data segment, we prevented significant information loss and ensuring the model remained valid for retired applicants for future use.
4. **Calibration and Business Interpretability:** A standout achievement of this project is the implementation of Probability Calibration (Isotonic Regression). We recognized that a model with high AUC can still output biased probabilities. By calibrating these outputs and mapping them to a standard 300–850 Credit Score, we successfully transformed raw technical outputs into an intuitive, industry-standard tool that facilitates immediate risk-based pricing and decision-making for loan officers.
5. **Benchmarking and Strategic Efficiency:** While our single-model solution achieved a competitive AUC of **0.7787**, we acknowledge a slight performance gap compared to top-tier Kaggle solutions that utilize complex Stacking ensembles. However, this was a deliberate strategic decision. In a regulated banking environment, the **marginal predictive gain** from multi-layer ensembles often comes at the cost of significantly increased **architectural complexity** and **difficulty in direct feature attribution**. We prioritized a streamlined, maintainable architecture that ensures straightforward regulatory auditing and unambiguous generation of adverse action reasons (rejection explanations).

**In summary**, the final model achieved a ROC-AUC of 0.7787, outperforming the baseline. However, the project's true value lies in its holistic approach—combining robust data integration, precise anomaly resolution, and business-ready outputs (calibrated PDs and credit scores) to effectively promote financial inclusion while maintaining rigorous risk management standards.

### 6.2 Future Work

While the current single-model approach is robust, a comparative gap analysis with top-tier industry solutions highlights specific areas for technical advancement that were not covered in this iteration:

1. **Ensemble and Blending Strategies:** The current solution relies on a single LightGBM model. We did not explore inter-model fitting or blending strategies. Future work should

investigate the correlation between different model types (e.g., Random Forest vs. Gradient Boosting). Combining models with low correlation could significantly reduce variance and improve generalization performance.

2. **Stacking Architectures:** To bridge the performance gap with top-tier benchmarks, future iterations could implement **Stacking** (Stacked Generalization). While this increases system complexity, training a meta-learner to combine predictions from diverse base models (e.g., XGBoost, Neural Networks) allows for capturing non-linear patterns that a single model might miss. *Deployment of such architectures would require implementing advanced interpretability layers (e.g., Kernel SHAP) to maintain compliance standards.*
3. **Adversarial Validation:** To ensure the model's stability over time, we did not perform Adversarial Validation. This technique—training a classifier to distinguish between training and test sets—is crucial for identifying "drifting" features that might degrade model performance in a production environment. Implementing this would be the next critical step to ensure long-term reliability.

# VII. References & AI Use Acknowledgements


[1]
https://www.experian.com/blogs/ask-experian/credit-education/score-basics/what-is-a-good-credit-score/
[2]
https://www.kaggle.com/code/jsaguiar/lightgbm-with-simple-features
[3]
https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html
[4]
https://medium.com/data-science/what-makes-lightgbm-lightning-fast-a27cf0d9785e