CREDIT RISK FOR BANKS, FINANCIAL INSTITUTIONS AND NBFCS.

GROUP MEMBERS

Krishna Mohan Panguluru

Kurti Dhabalia

Kundan Kumar

Problem statement

- Build a bank's internal end-to-end scoring mechanism, based on the application information, clubbed with the raw bureau information.
- To help home credit to identify the right customers using predictive models on the basis of past data applicant.
- To assist Home Credit in deciding which loan applications should be disbursed, and which should be rejected, based on the applicant's past behaviour and application information.

Goal

- First gather the information and clean it to make it usable.
- Apply 'Feature Engineering' techniques to roll up the information at applicant level, and thereby create manual features for model building.
- Build a classification model to differentiate applicants between approves and rejects.

Final Outcome

Leverage trade level information for Credit Bureaus by aggregating trade level information to applicant level in order to capture their payment behaviour, which can be used for decisioning, strategies and business insights for the bank.

Data understanding

To perform the credit risk analysis, we have two datasets i.e., Application_base & Bureau

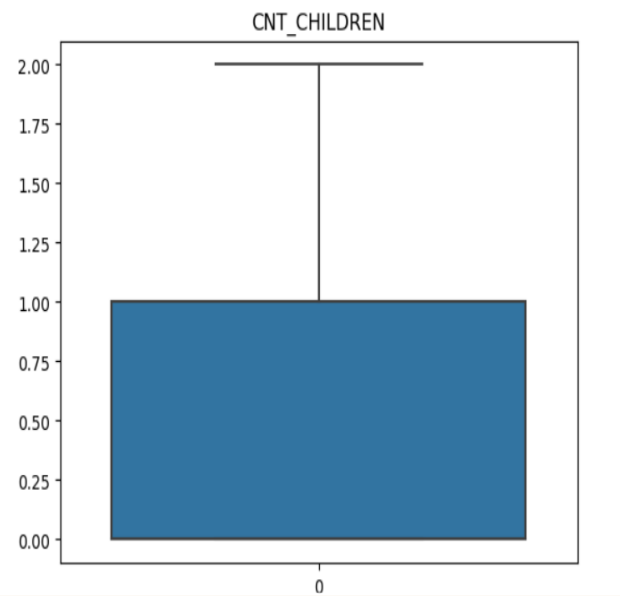
Application base data: This data set provide us the information on applicants and contains customer-level information on age, gender, income, marital status, type of contract for loan etc.

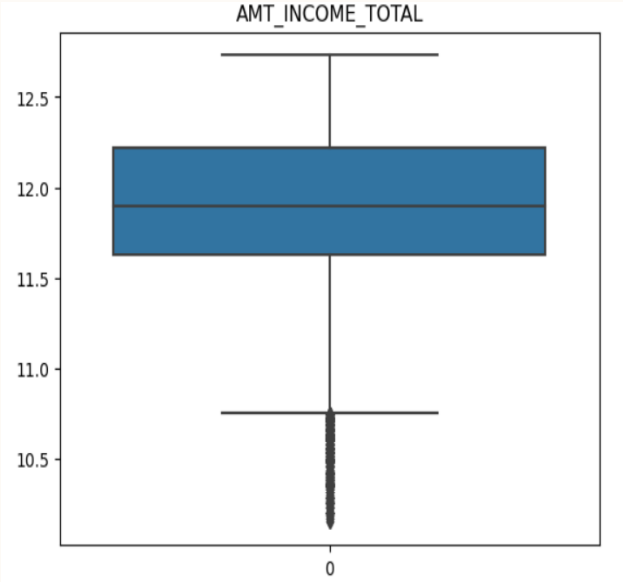
Credit bureau: This data set is taken from the credit bureau and contains variables such as credit active or not, no. of days for credit, how many days of credit overdue, type of credit, amount of credit and overdue amount if any', etc.

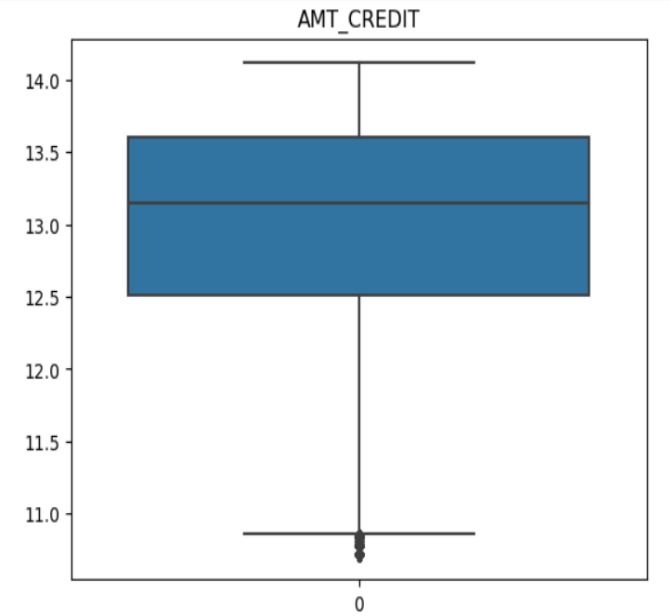
Data preparation:

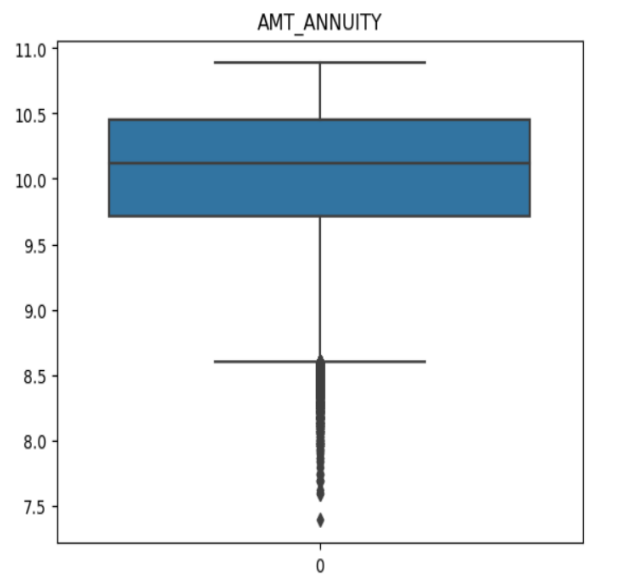
- ❑ Checking for duplicate values
- ❑ Checking all the columns for missing values & outliers
- ❑ Impute missing values
- ❑ Remove columns with >40% missing values
- ❑ Update num_cols after dropping columns
- ❑ Visualize boxplots to identify outliers
- ❑ Calculate IQR to quantify outliers
- ❑ Cap outliers at 95th percentile for numerical columns
- ❑ Apply log transformation for skewed financial variables (e.g., AMT_INCOME_TOTAL, AMT_CREDIT)
- ❑ Verify changes with updated boxplots
- ❑ Convert numerical columns to float
- ❑ Check missing values after removal

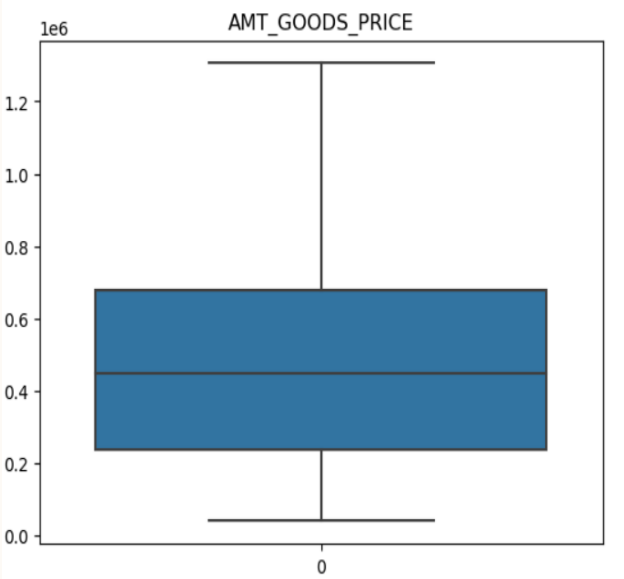
Detecting outliers by using boxplot (1/2):

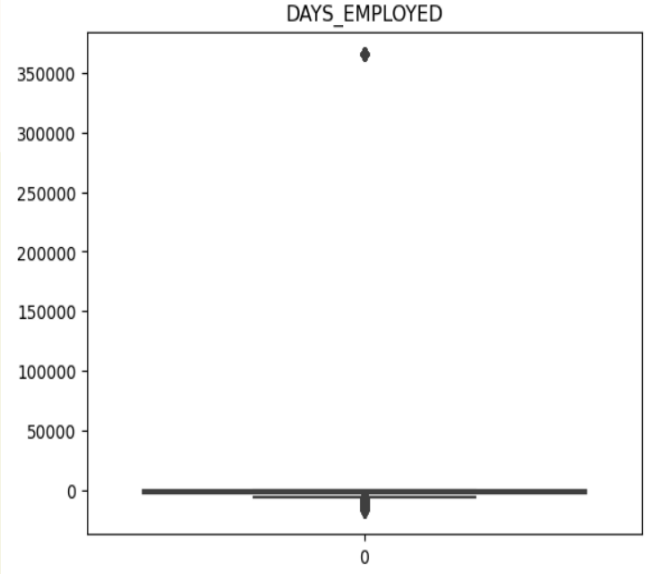




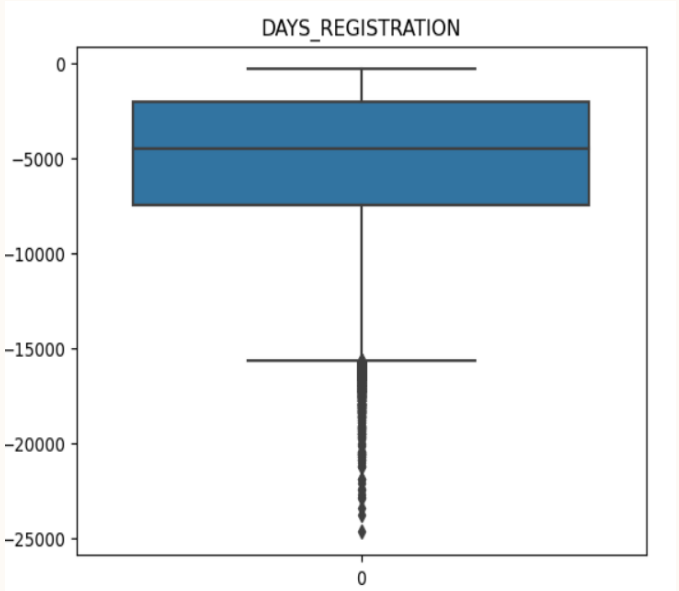


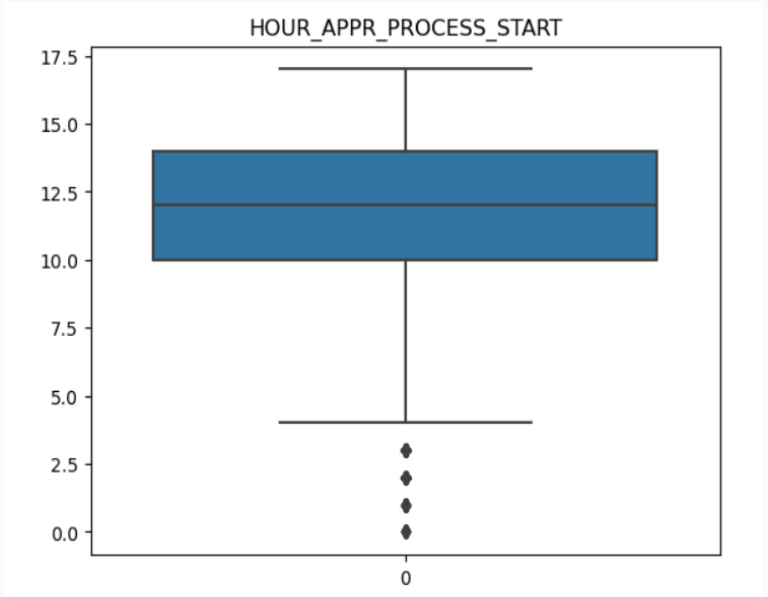


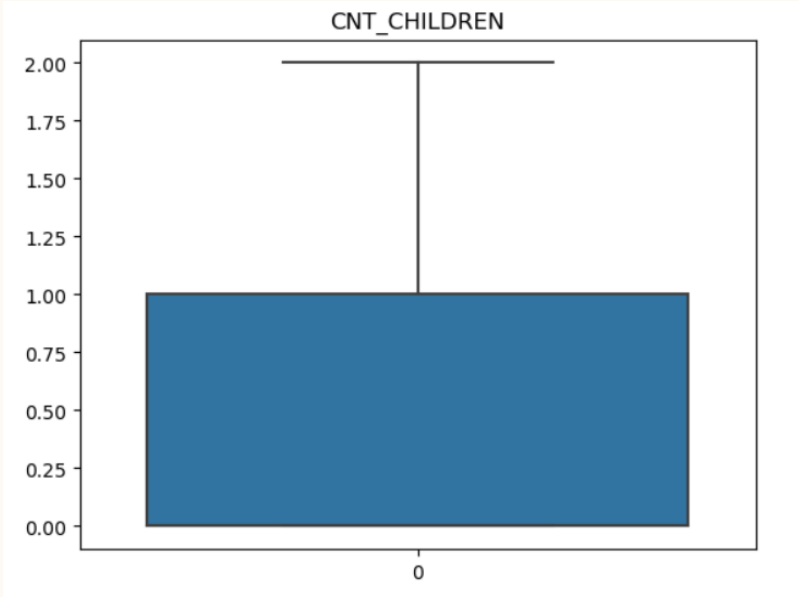


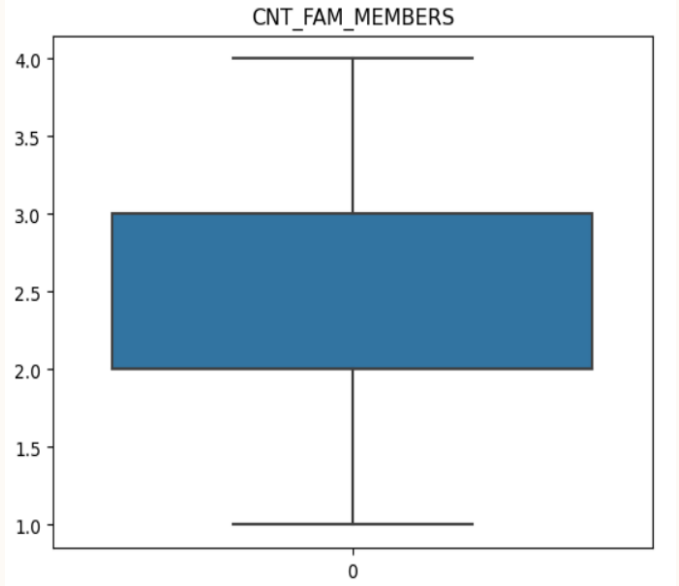


Detecting outliers by using boxplot (2/2):





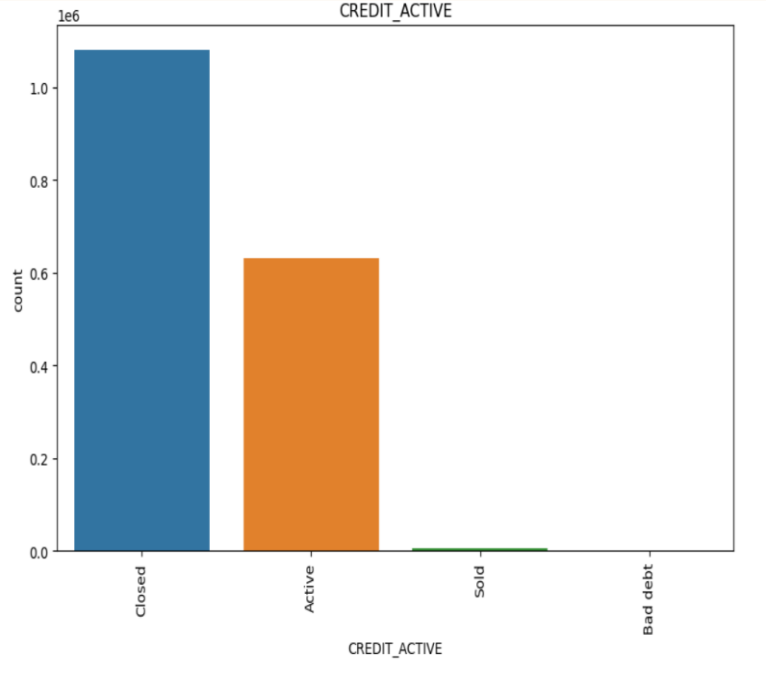


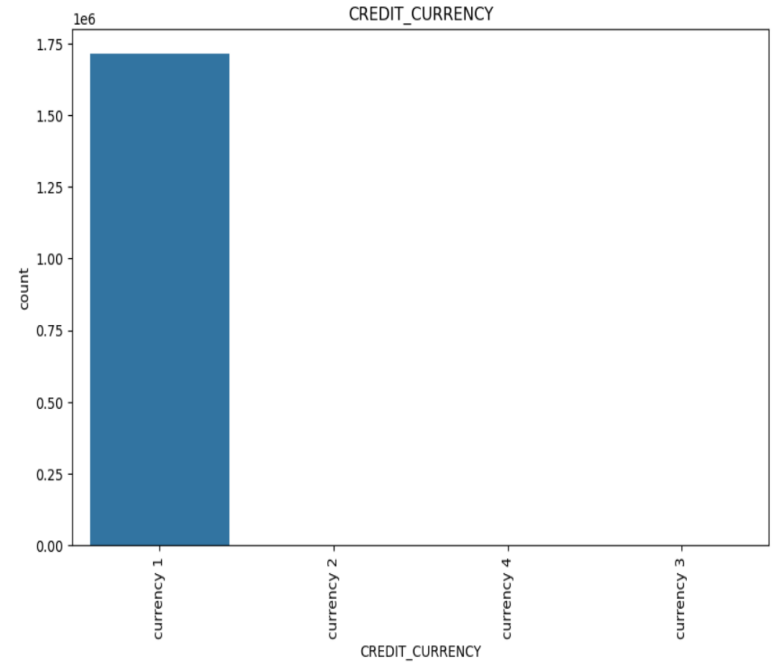


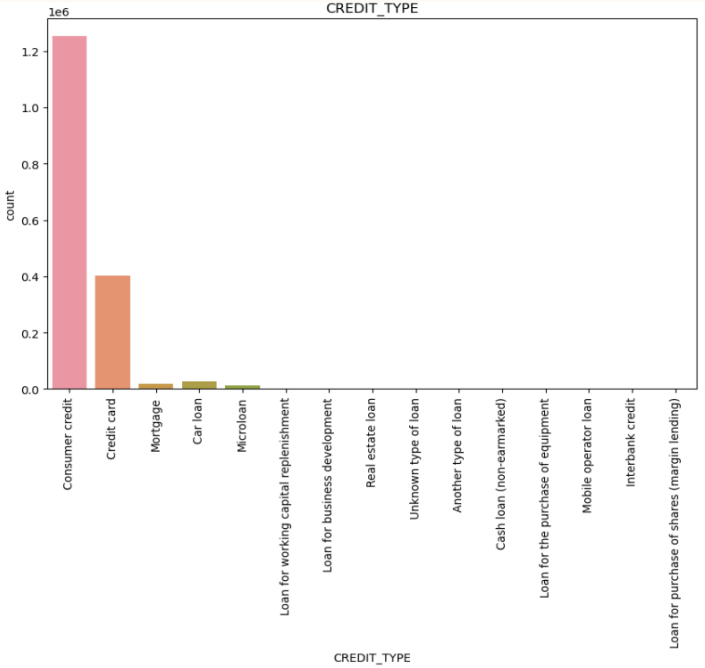
Observations

-Several numerical features exhibit outliers, especially in financial attributes like AMT_INCOME_TOTAL, AMT_CREDIT, and AMT_ANNUITY. -DAYS_EMPLOYED contains extreme negative values, which likely indicate anomalies (e.g., retired or unemployed individuals with unrealistic employment duration). -Variables such as CNT_CHILDREN and CNT_FAM_MEMBERS are heavily right-skewed, indicating that most applicants have a low number of dependents.

Segment Bureau data : Categorical variable distributions



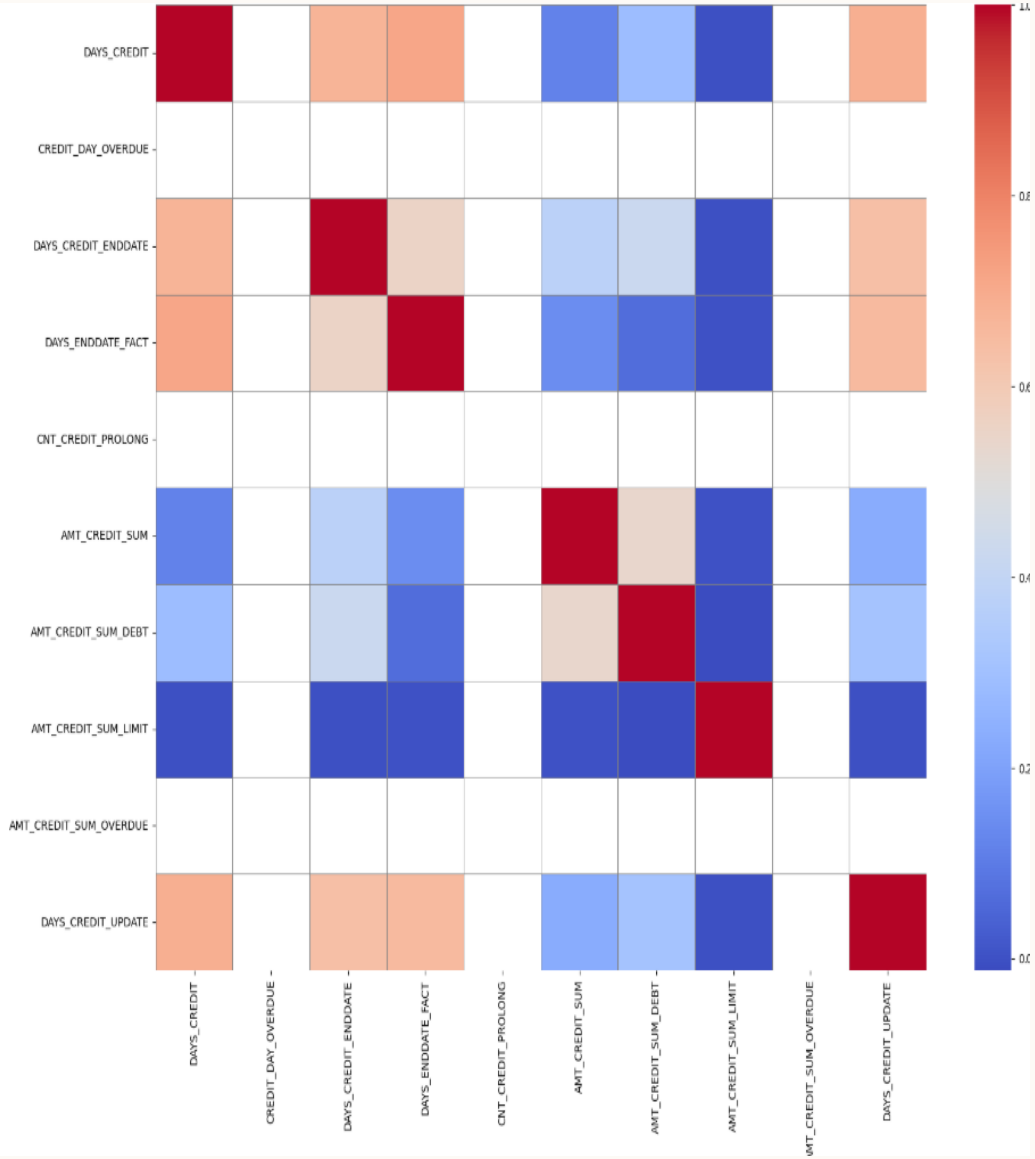




Observations

Most applicants have CREDIT_ACTIVE = Closed, meaning they have past credit history but no ongoing credit obligations. -Different types of credit (CREDIT_TYPE) are distributed across various categories, with consumer loans and credit cards being the most common. -CREDIT_CURRENCY shows most credits are in local currency, which simplifies risk evaluation.

Bureau correlation heatmap



Observations

1. **Strong correlations exist between:**

AMT_CREDIT_SUM, AMT_CREDIT_SUM_DEBT, and AMT_CREDIT_SUM_LIMIT → These variables are interrelated as they represent total credit, outstanding debt, and credit limits, respectively, indicating a cohesive picture of an applicant's credit utilization.

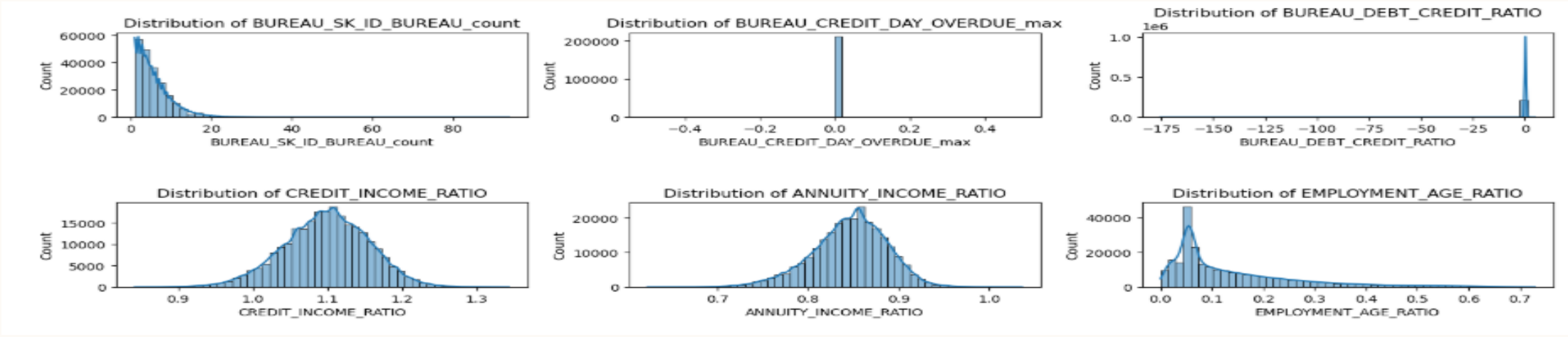
2. Moderate correlations observed between:

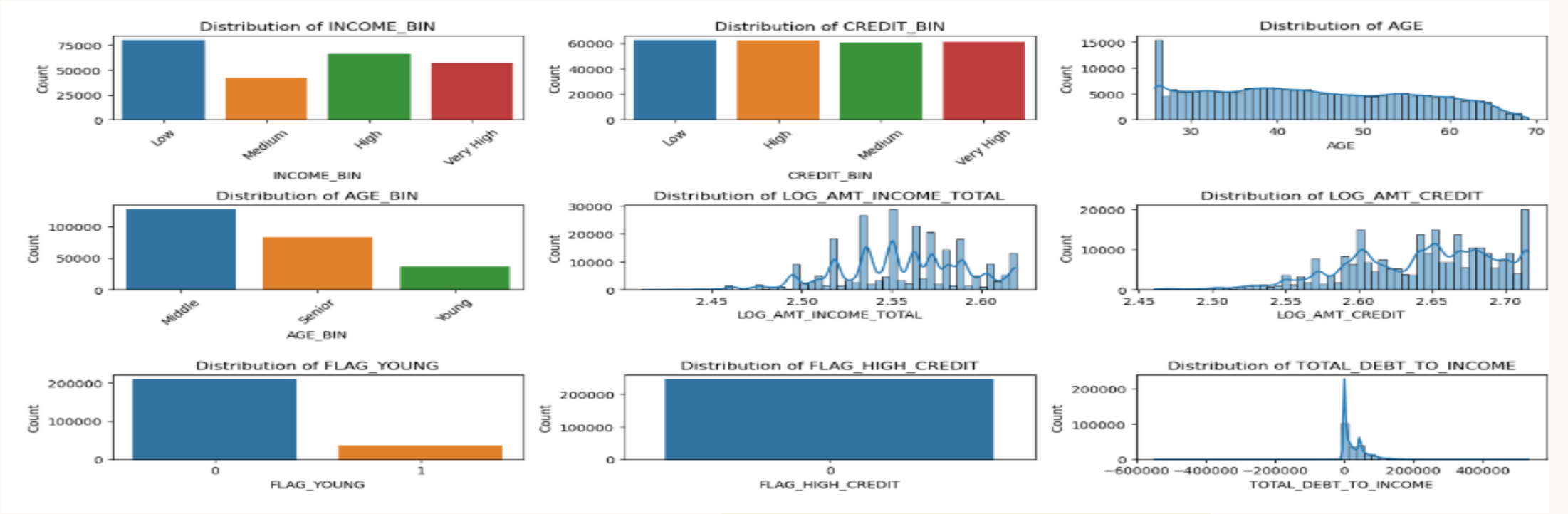
DAYS_CREDIT_UPDATE (last credit bureau update) and CREDIT_DAY_OVERDUE → Recent updates in credit history might be linked to borrower risk, suggesting that frequent updates could signal payment difficulties.

3. Notable patterns in loan behavior:

CNT_CREDIT_PROLONG (number of loan extensions) shows a moderate correlation with AMT_CREDIT_SUM_OVERDUE → Borrowers extending their loans frequently may be struggling financially, as extensions often accompany overdue payments.

Combine training data with target for EDA (1/2)



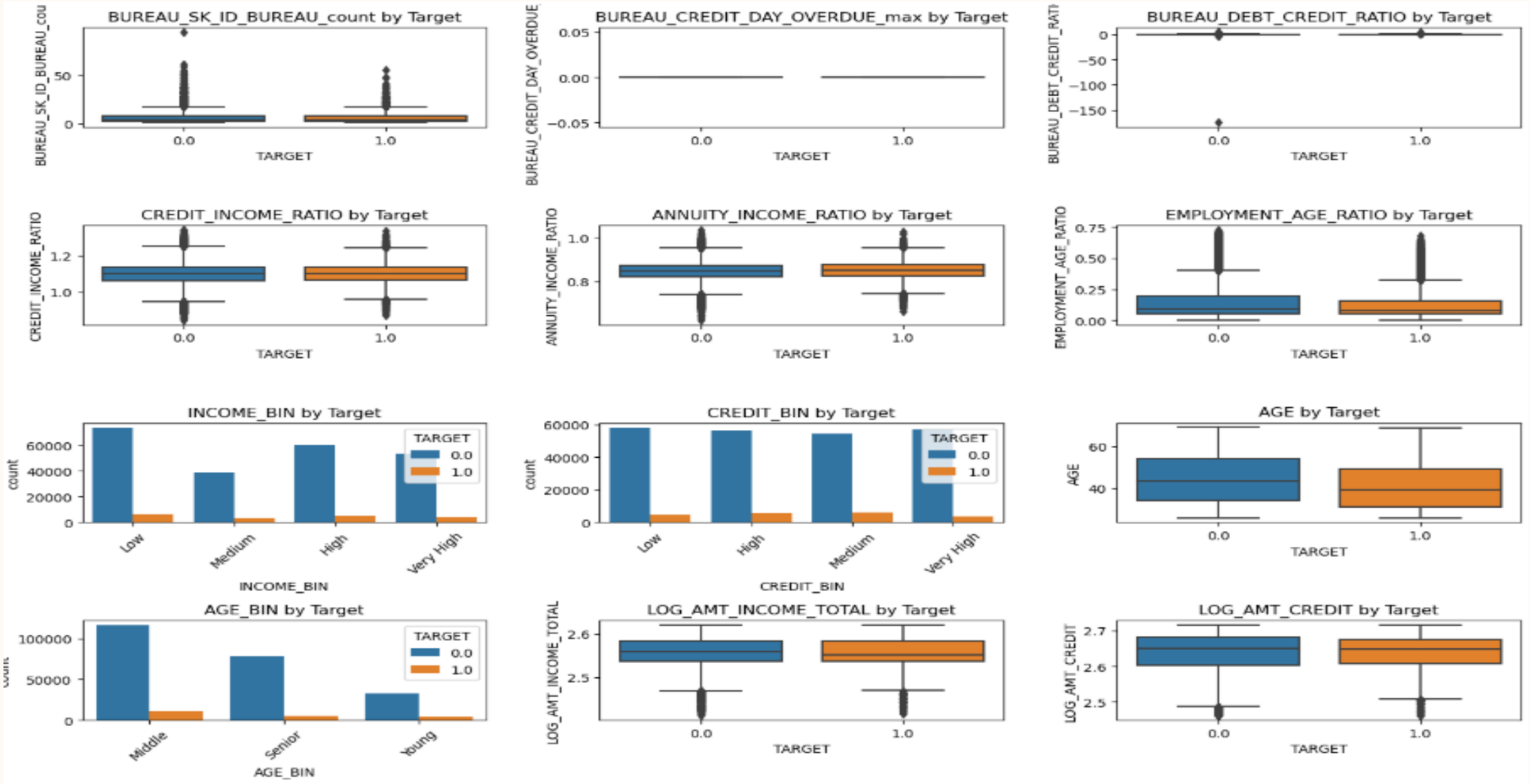


Combine training data with target for EDA (2/2)

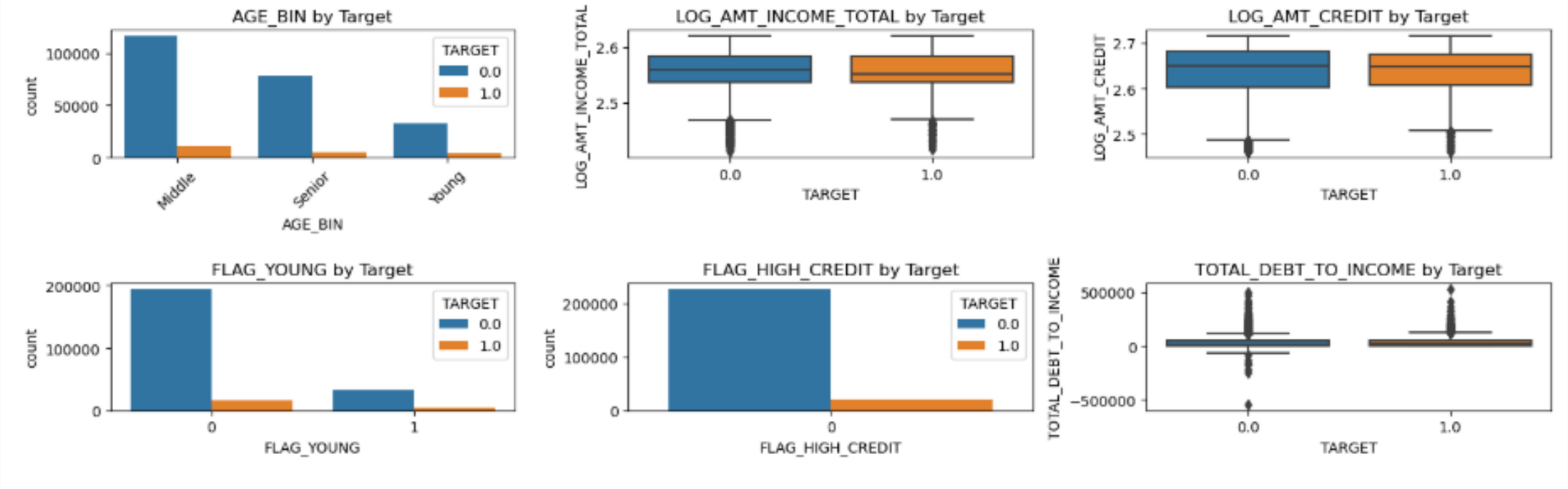
Observations

- Log_Amt_Income_Total and Log_Amt_Credit shows more systematic distributions compare to their original right-skewed forms.
- Income_Bin and Credit_Bin indicates that the majority of applicants fall into the 'Low' and 'Medium' categories, with fewer in 'High' and 'Very High'.
- Age_Bin shows a significant proportion of 'Young' applicants, which may correlate with higher risk.
- Age and Employment_Age_Ratio may exhibit bimodal behavior, suggesting distinct groups (e.g., young/newly employed vs. older/established workers).
- Days_Credit (from bureau data) continues to show bimodal patterns, indicating varied credit-seeking behaviors.
- After replacing Days_Employed extreme values (365243) with the median, the distribution appears more consistent, reducing the impact of anomalies.
- Flag_High_Credit highlights a small but notable subset (e.g., ~5-10%) of applicants with credit exceeding 4x their income.
- Flag_young identifies ~20-25% of applicants under 30 years, a group potentially at higher risk.
- Total_debt_to_income shows a right-skewed distribution, with a tail of applicants having high combined debt burdens, suggesting financial strain.

Boxplots of engineered features by target (1/2)



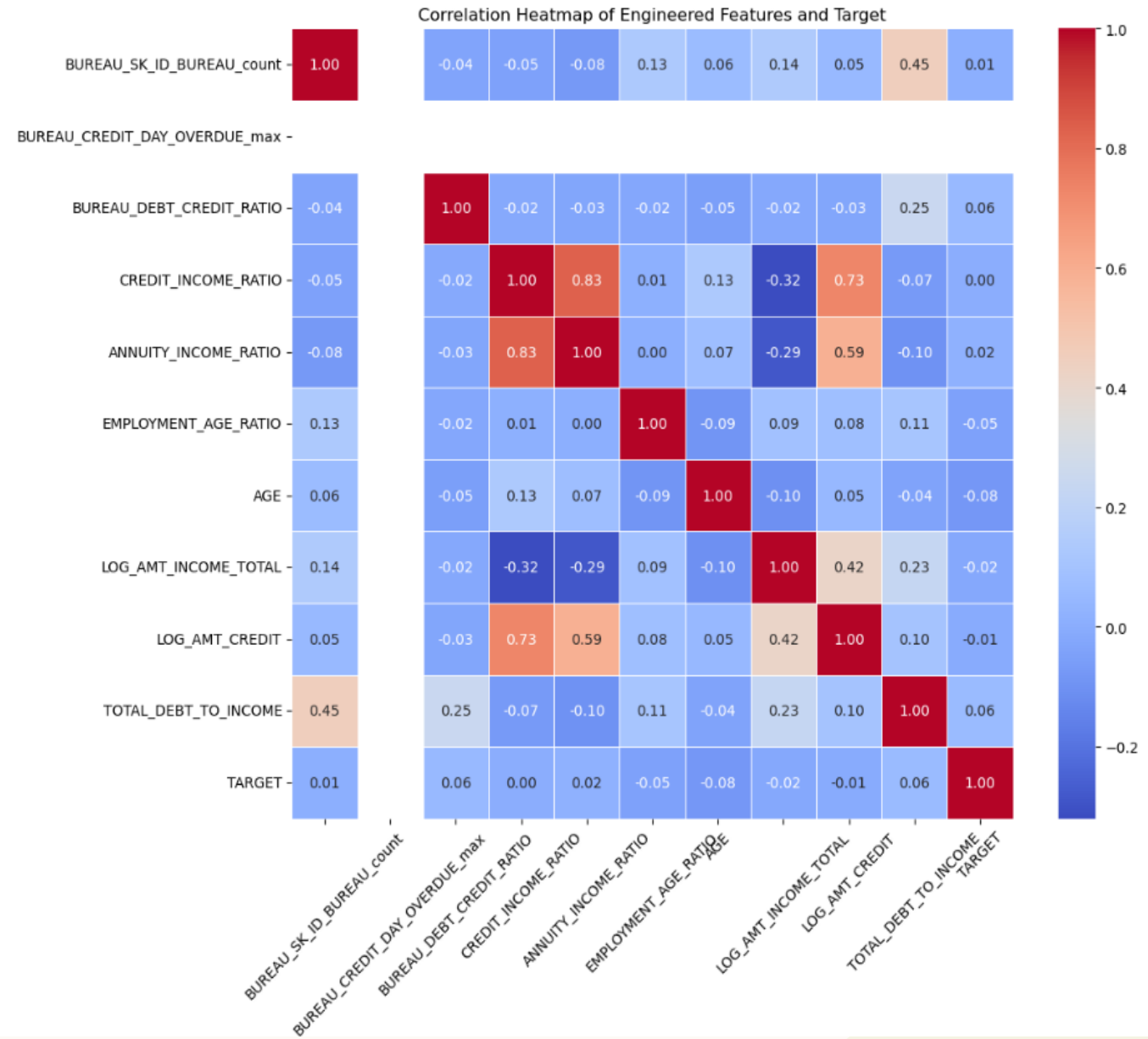
Boxplots of engineered features by target (2/2)



Observations

Borrowers with lower income or high overdue amounts are more likely to default. Credit history and repayment behavior play a key role in risk assessment. Features showing clear separation between TARGET = 0 and TARGET = 1 can be strong predictors for credit risk modeling.

Correlation heatmap for engineered features and target

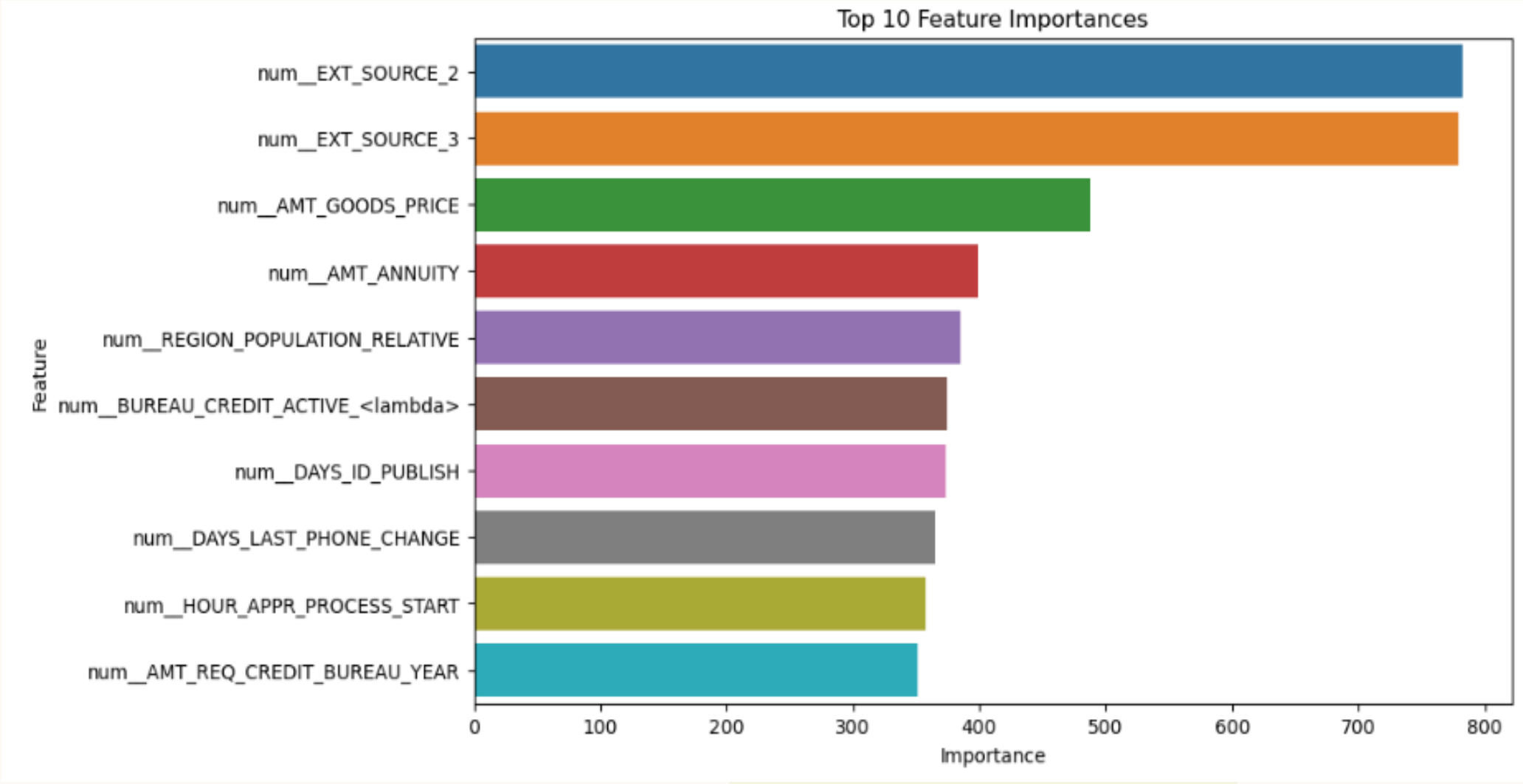


Observations

Strong correlations between some features suggest potential feature selection to avoid multicollinearity. Engineered features showing significant correlation with TARGET should be prioritized for predictive modeling. Weak correlations with TARGET indicate that additional feature engineering or non-linear modeling approaches may be needed to improve predictive power.

Business insights and recommendations (1/2)

Feature importance



Model development

Model 1: Logistic Regression with GridSearchCV

- Explanation of hyperparameter tuning:- Hyperparameter tuning uses GridSearchCV with 5-fold cross-validation and ROC-AUC scoring to optimize model performance.
- Baseline Logistic Regression ROC-AUC : 0.7499
- Best Logistic Regression model (after hyperparameter tuning)

Model 2: Random Forest with GridSearchCV

- Best Random Forest model (after hyperparameter tuning)
  Best Random Forest Parameters : {'max_depth': 10, 'n_estimators' : 100
  Best Random Forest ROC_AUC (Cross-Validation): 0.965519653193413

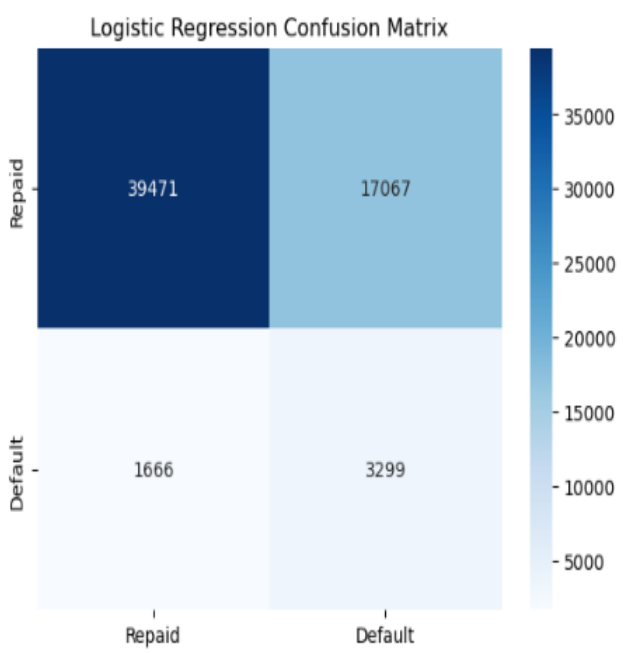
Model 3: LightGBM with GridSearchCV

- Best Random Forest model (after hyperparameter tuning)
  Best LightGBM Parameters: {'Learning_rate': 0.95, 'max_depth': 7, n_estimators': 500}
  Best LightGBM ROC_AUC (Cross-Validation): 0.9699017851879267

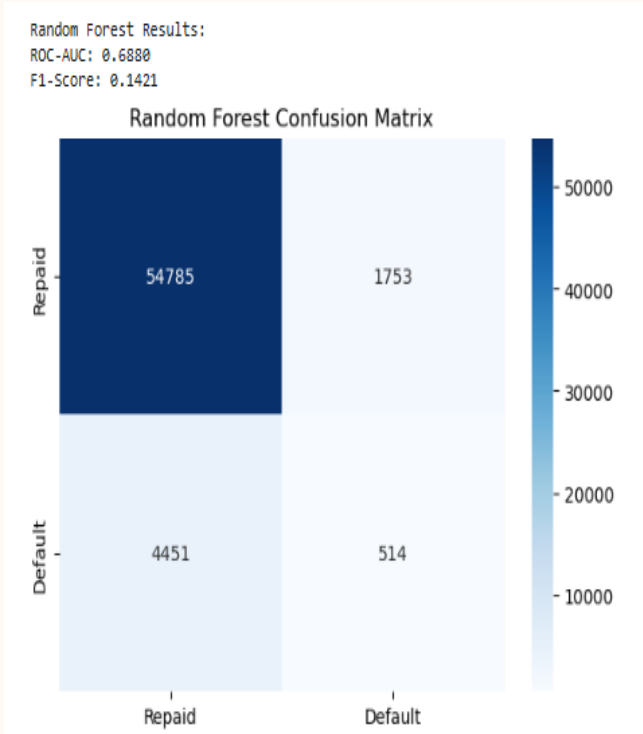
Model Evaluation

Compute ROC-AUC for each model on the test set

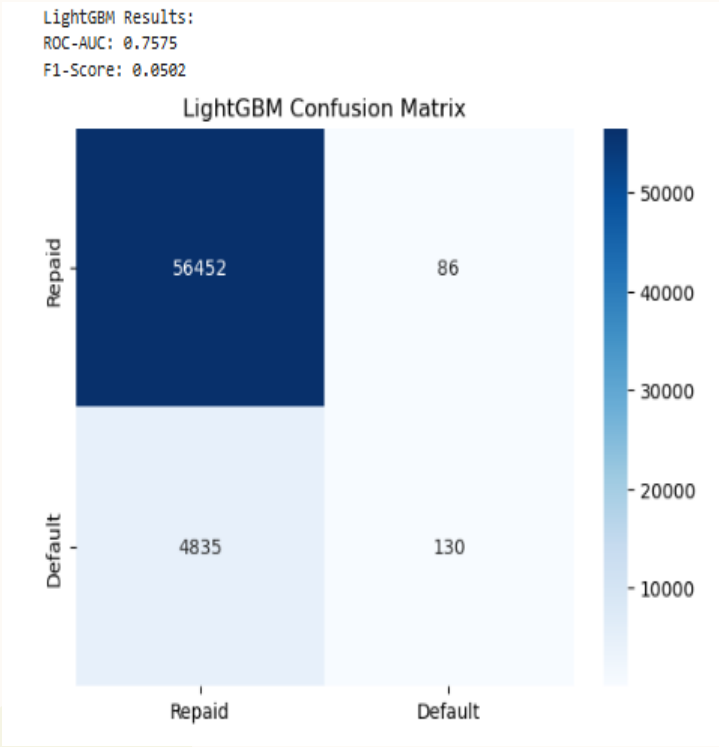
Logistic Regression Results:
ROC-AUC: 0.7457
F1-Score: 0.2605



Random Forest Results:
ROC-AUC: 0.6880
F1-Score: 0.1421



LightGBM Results:
ROC-AUC: 0.7575
F1-Score: 0.0502



Best Model (out of Logistic Regression, Random Forest, LightGBM):

LightGBM (ROC-AUC: 0.7575)

Business insights and recommendations (2/2)

Business Insights (Based on the Best Model):

- High 'BUREAU_CREDIT_DAY_OVERDUE_max' values strongly predict default risk, as seen in the EDA after merging.
- High 'CREDIT_INCOME_RATIO' and 'TOTAL_DEBT_TO_INCOME' indicate increased default likelihood, especially for applicants in the 'Low' INCOME_BIN.
- Young applicants (AGE_BIN = Young or FLAG_YOUNG = 1) with FLAG_HIGH_CREDIT = 1 show a higher default rate, suggesting age and credit burden as key risk factors.
- Active loans ('BUREAU_CREDIT_ACTIVE') and BUREAU_ANY_OVERDUE = 1 indicate repayment behavior patterns, with more active loans or any overdue history linked to higher risk. Recommendations:
- Reject applications with 'BUREAU_CREDIT_DAY_OVERDUE_max' > 30 days or BUREAU_ANY_OVERDUE = 1 to reduce default risk.
- Limit credit to 4x 'AMT_INCOME_TOTAL' and enforce a TOTAL_DEBT_TO_INCOME threshold (e.g., <40%) to mitigate risk for high CREDIT_INCOME_RATIO applicants.
- Prioritize applicants with fewer active loans and target young applicants (FLAG_YOUNG = 1) with additional income verification or lower credit limits if FLAG_HIGH_CREDIT = 1.
- Offer tailored loan terms (e.g., lower rates) to 'High' or 'Very High' INCOME_BIN applicants with no overdue history

Business insights for Board (1/2)

1. Risk-Based Loan Approval & Pricing

- ❑ Younger applicants (AGE_BIN = Young or FLAG_YOUNG = 1) exhibit higher default risk, especially when combined with FLAG_HIGH_CREDIT = 1 → Implement stricter loan-to-income ratios and lower credit limits for this segment.
- ❑ Unstable employment (reflected in bimodal EMPLOYMENT_AGE_RATIO) correlates with higher defaults → Adjust approval criteria with additional income stability checks and dynamic pricing.
- ❑ Property owners (FLAG_OWN_REALTY = 1) and applicants in 'High' or 'Very High' INCOME_BIN with no overdue history are lower risk → Offer preferential loan terms (e.g., lower interest rates).

2. Optimized Credit Limits & Loan Structuring:

- ❑ High CREDIT_INCOME_RATIO and TOTAL_DEBT_TO_INCOME (especially for 'Low' INCOME_BIN applicants) indicate significant risk → Use these as key factors to set credit limits and enforce a TOTAL_DEBT_TO_INCOME threshold (e.g., <40%).
- ❑ Frequent recent credit inquiries (DAYS_CREDIT_UPDATE trends) may signal financial stress → Introduce a risk flag and cap credit for affected applicants.

Business insights for Board (2/2)

3. Leveraging External Credit Scores

- ❑ Ext_source_2 remain strong risk predictors → Integrate these scores into the approval process for enhanced credit assessment

4. Proactive Risk Monitoring & Early Warning:

- ❑ Late payments (Bureau_Credit_Day_Overdue_max> 30 days) and Bureau_any_overdue = 1 should trigger interventions (e.g., reminders, restructuring offers).
- ❑ Monitor credit utilization (Total_debt_to_income) and multiple active loans (Bureau_credir_active) as early signs of default, with automated alerts for high-risk cases.

5. **Refinancing & Cross-Sell Opportunities**

- ❑ Closed credit accounts (Credit_active=closed) indicate potential for refinancing or cross-selling additional products.
- ❑ Enhanced feature engineering (e.g Log_amt_income_total, Income_bin, Total_debt_to_income) improves risk prediction, enabling personalized loan offers and refinancing options.

By integrating these insights, we can enhance risk management, optimize credit policies, and improve profitability.

THANK YOU