Predicting Home Credit Default Risk Using Random Forest and SHAP Feature Importance Analysis

Hanzhe CUI
Department of Civil and Environmental Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
hcuiah@connect.ust.hk

Qun ZHENG
Department of Civil and Environmental Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
qzhengam@connect.ust.hk

Jingying WANG
Department of Civil and Environmental Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
jwangmi@connect.ust.hk

Wen HUANG
Department of Civil and Environmental Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
whuangbp@connect.ust.hk

Abstract

Modeling credit default risk from multi-source financial data is central to modern credit scoring. While large feature sets may capture complex relationships, they also introduce redundancy and reduce interpretability. In this study, we develop a reproducible and explainable approach for the Home Credit Default Risk problem on Kaggle. The overall workflow comprises four major components: Data Processing, Random Forest Modeling, Model Interpretability Using SHAP, and Top-K Feature Selection and Retraining. Starting from the main applicant table and five auxiliary relational tables, we construct 4,166 aggregated features and train a Random Forest classifier. Using SHAP values, we evaluate the global importance of features. Subsequently, a sensitivity analysis is conducted by varying the number of top-ranked features (top-K) and comparing the corresponding AUC values to identify the optimal feature subset. Results show that retaining about 3,600 SHAP-ranked features yields the best validation AUC, with a Kaggle submission score of 0.7582.Our ranking on Kaggle is approximately 4000. This study demonstrates how model explainability can guide effective feature selection in large-scale tabular datasets, providing a reproducible approach for credit risk prediction. For our code, please refer to https://github.com/hanzhe910/MATH5470_CUI_Huang_Zheng_Wang and https://www.kaggle.com/code/hcuiah/math5470-cui-huang-zheng-wang.

Preprint. Under review.

1 Introduction

Predicting an applicant's ability to repay a loan is central to credit scoring, where financial institutions aim to identify both high-risk and reliable borrowers. The Home Credit Default Risk dataset provides a realistic benchmark for such tasks. The information in this dataset includes clients, previous credits, and external bureau records. The key challenge is effectively integrating data from multiple sources while ensuring the result is both robust and interpretable.

Methods such as Random Forests (RF) and gradient boosting machines, which rely on tree-based ensembles, have demonstrated strong predictive performance on tabular datasets. However, with huge features derived from categorical encodings and aggregated data, it can be challenging to select the most important ones and interpret the model. Traditional feature ranking methods (e.g., Gini importance) may be biased by feature cardinality, while wrapper methods are computationally expensive for such high-dimensional settings.

To address these challenges, we employ the fasttreeshap package, a scalable variant of SHAP (SHapley Additive exPlanations), to quantify the contribution of each feature to the RF model output. SHAP values provide a consistent, theoretically grounded measure of feature importance. By ranking features according to their mean absolute SHAP values, we systematically evaluate how the validation results of the random forest model change as we retain different numbers of top-ranked features.

This report demonstrates how a Random Forest model, combined with SHAP-based feature analysis, can be used to predict the likelihood of repaying a loan while identifying the most influential features for interpretability.

2 Data and Methodology

2.1 Data Processing

The Home Credit Default Risk dataset integrates multiple relational tables describing demographic attributes, financial records, previous loan applications, and external bureau statistics. All tables are merged through multi-level aggregations (mean, sum, count, and ratios) using the unique identifier. Categorical variables are encoded via one-hot encoding, while continuous features are retained in their original scale. This process results in three datasets, including training input (`X_train`), training label (`y_train`), and test dataset (`X_test`). We will use the `X_train` and `y_train` datasets to train and validate our model, and then use `X_test` to obtain the results and submit them to Kaggle. `X_train` has a total of 307511 samples and 4166 features.

2.2 Random Forest Modeling

We first employ an RF classifier to train a model to predict the probability of loan repayment. The training dataset (`X_train` and `y_train`) is further divided into two parts, with 80% of the samples used for model training and the remaining 20% reserved for validation. Missing numerical values are imputed using the median computed from the training set. The model is configured with 1000 decision trees and utilizes 256 threads for parallel computation.

Following model training, performance is assessed on the validation set using the Area Under the ROC Curve (AUC). The model produces predicted probabilities of loan repayment rather than binary labels, enabling a more informative ranking of applicants according to their likelihood of repayment. Furthermore, the AUC is used to evaluate models built with different numbers of top features ranked by SHAP, helping to determine the optimal feature subset.

2.3 Model Interpretability Using SHAP

To interpret the trained RF model, we compute SHAP values using the fasttreeshap package, which efficiently handles tree ensembles. A random subset from the test set is used for explanation to balance memory efficiency and computational cost. Missing values are imputed using the median of the training set, and features are cast to float64 for numerical stability.

The SHAP explainer is applied to compute the contributions of each feature when predicting the results. Global feature importance is quantified as the mean absolute SHAP value across samples,

and features are ranked accordingly. This analysis highlights the features with the greatest influence on the probability of loan repayment, guiding subsequent feature selection.

2.4 Top-K Feature Selection and Retraining

Using the SHAP-based ranking, we construct multiple Top-K feature subsets by retaining the top 200, 400, ... up to 4,000 features. For each subset, an RF classifier is retrained on the original training data with the same split.

Validation AUC is used to evaluate each model, enabling a comparison of predictive performance as a function of the number of retained features. This procedure identifies the feature subset that balances model complexity and accuracy. The selected Top-K features are then used to train a final model on the full training set and generate probabilistic predictions on the test set for evaluation and submission.

3 Results

3.1 Model Prediction Performance

We trained a Random Forest model using all the features obtained and then utilized SHAP to rank the feature importance. Subsequently, we selected different numbers of top features to train the model and found that the model achieved the highest AUC value when the top 3,600 features were utilized (Table 1). The figure 1 reports AUC = 0.7592 with a 95% confidence interval [0.7526, 0.7658], indicating good discrimination. The overall regression fit results is good, suggesting a strong monotonic relationship; that is, groups with higher predicted probabilities correspond to higher actual default rates, demonstrating good ranking ability of the model. In the extremely low probability range (<0.02), the points closely align with the line y=x, and as the probability increases, some points deviate slightly. In Figure 2b, the axes are scaled to approximately 0.14–0.15, covering about 90% of the bin points (as noted in the title). Local fitting reveals also a good results, indicating that the probability characterization for the "low-risk population" is relatively reliable. In the range of 0.05–0.12, many points fall above the line y=x (with the vertical coordinate greater than the horizontal), suggesting that the actual default rates are slightly higher than the predicted probabilities provided by the model. This indicates that the model tends to "underestimate risk" or is somewhat "conservative" in the mid–to–high probability segments.

Table 1: AUC of different Top-K features

Top - K features	Validation AUC
200	0.75840
400	0.75828
600	0.75995
800	0.75956
1000	0.75990
1200	0.76007
1400	0.76051
1600	0.75783
1800	0.75964
2000	0.76006
2200	0.75951
2400	0.76063
2600	0.76044
2800	0.75827
3000	0.75920
3200	0.75909
3400	0.75932
3600	0.76080
3800	0.75835
4000	0.75830

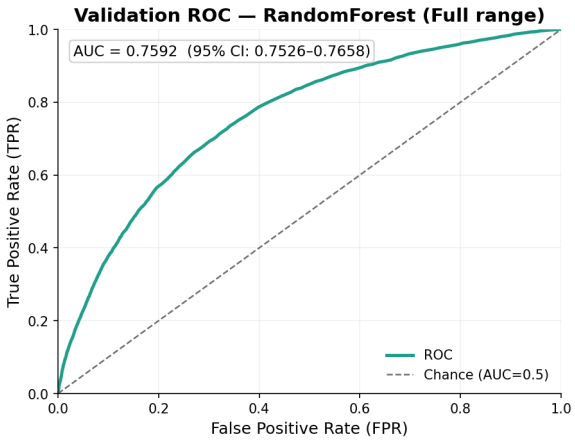

Figure 1: Validation ROC curve (Home Credit Default Risk). The x-axis is the false positive rate (FPR) and the y-axis is the true positive rate (TPR, recall). The teal line shows the model's operating points across thresholds; the grey dashed line is the random baseline (AUC=0.5). Axes span 0–1 with a clean, Nature-style layout for readability across false-positive operating regions.

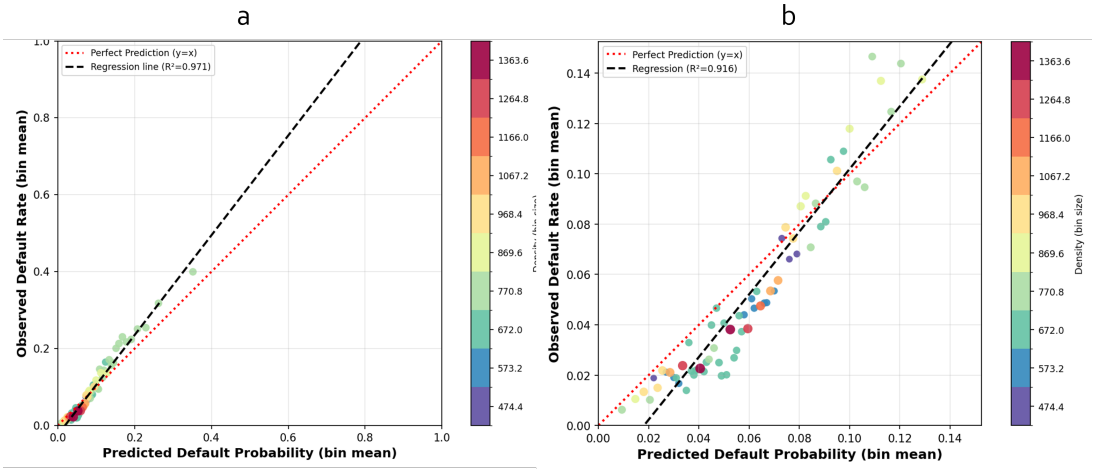

Figure 2: Binned calibration scatter plots (validation set). (a) Validation samples are split into quantile bins by predicted probability. Each dot represents one bin with the mean predicted probability on the x-axis and the empirical default rate (mean label) on the y-axis. Dot color and size encode the bin size. The red dotted line denotes the perfect calibration line (y=x); the black dashed line is the linear fit to the binned points with R^2 = [0.971]. Both axes share the same units and an equal aspect ratio to visualize deviations from y=x. (b) Axes zoomed to the 90th percentile of the binned predicted probabilities (covering [90.0%] of bins). Linear fit is recomputed within the zoomed range with R^2 = [0.916]. Dot color and size indicate bin size. Both panels use equal-aspect axes for a fair comparison against y=x.

3.2 SHAP-Based Feature Interpretation

The target variable is what we are asked to predict: either a 0, indicating that the loan was repaid on time, or a 1, indicating that the client had payment difficulties. The SHAP plot illustrates the average impact of each feature on the credit default risk prediction results. Specifically, the larger the absolute value of the mean SHAP value, the stronger the feature's average influence on the model output. The top 20 features broadly characterize the client's repayment ability from five aspects: historical credit application records, external credit assessments, historical repayment performance, consumption and transaction behavior, and personal basic characteristics.

Among these, the most important feature is PREV_NAME_CONTRACT_STATUS_Approved_MEAN, which represents the mean number of approved historical contracts and accounts for the highest contribution to the results (close to 6%). This indicates that the average status of approved historical contracts is the core factor affecting repayment risk. Additionally, PREV_NAME_CONTRACT_STATUS_Refused_SUM and PREV_NAME_CONTRACT_STATUS_Canceled_VAR, which represent the total number of historical contract refusals and the variance of canceled contracts in historical applications, rank 6th and 13th in feature contributions, respectively, but are less important than the mean number of approved historical contracts. These three indicators all belong to the category of historical credit application records. The next key factor is EXT_SOURCES_MIN (representing the minimum value of external credit scores), which contributes 5%, indicating that the level of external scoring has a significant impact on repayment risk assessment. Following that are BUREAU_BB_STATUS_1_MEAN_VAR (variance of overdue performance across different loans) and CREDIT_TERM (the pressure of monthly payments relative to the loan amount), each contributing 4%, suggesting that expected loan performance and monthly payment pressure have a significant influence on repayment risk, just behind the two previously mentioned factors. The contributions of the remaining factors range from 2% to 4%, including personal characteristics of the client, such as DAYS_BIRTH_ABS (age) and NAME_FAMILY_STATUS_Married (marital status); client transaction and consumption behaviors, including POSCURR_POS_SK_ID_CURR_COUNT_SUM (total number of monthly POS records), POSCURR_POS_MONTHS_BALANCE_MEAN_SUM (total of monthly balance means), and POSCURR_POS_SK_ID_CURR_SUM_MEAN (mean of total transaction records); and historical repayment performance, including INSTALCURR_INSTAL_DBD_VAR_MIN (variance of early repayment days) and INSTALCURR_INSTAL_DPD_VAR_VAR (variance of overdue days for each installment) among other variables.

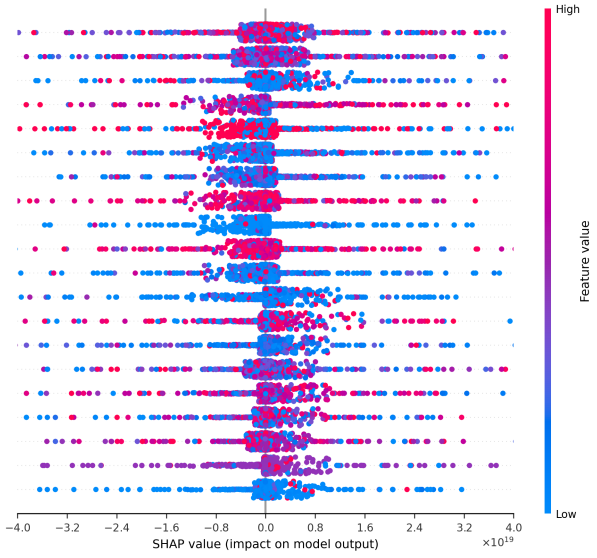

Figure 3: SHAP Feature Contribution Scatter Plot for Credit Risk Predictions

In the correlation matrix, we observe that EXT_SOURCE_MIN shows a high correlation with EXT_SOURCE_1 and EXT_SOURCE_3. This correlation arises from the fact that EXT_SOURCE_MIN is derived from the three external scoring variables, including these two, which also exhibit a certain positive correlation. DAYS_BIRTH_ABS demonstrates a notably high positive correlation with EXT_SOURCE_1, EXT_SOURCE_3, and EXT_SOURCE_MIN, indicating a relationship between age-related features and external data sources. Furthermore, the correlation between PREV_NAME_CONTRACT_STATUS_Approved_MEAN and PREV_NAME_CONTRACT_STATUS_Refused_SUM, as well as PREV_NAME_CONTRACT_STATUS_Canceled_VAR, shows a strong negative correlation. This is logically sound since an increase in approvals naturally leads to fewer refusals and cancellations. For all other variables, the absolute correlation values are generally below 0.2, suggesting low linear correlation between these features.

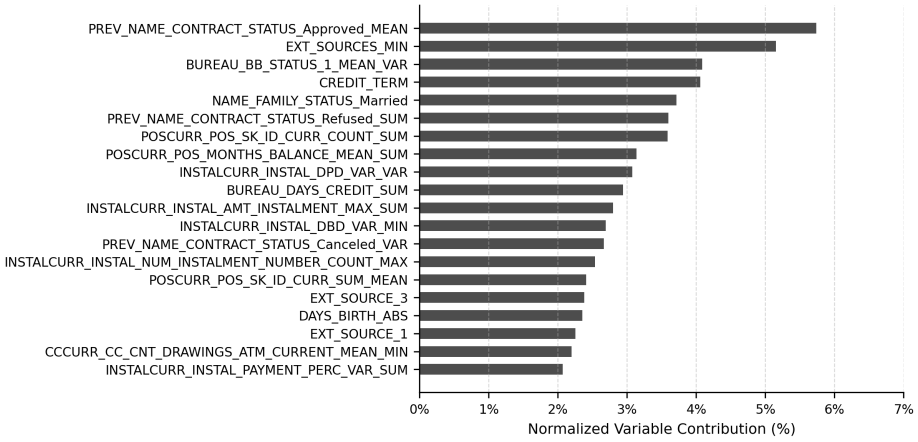

Figure 4: Feature Importance in Credit Risk Prediction (Normalized Variable Contribution)

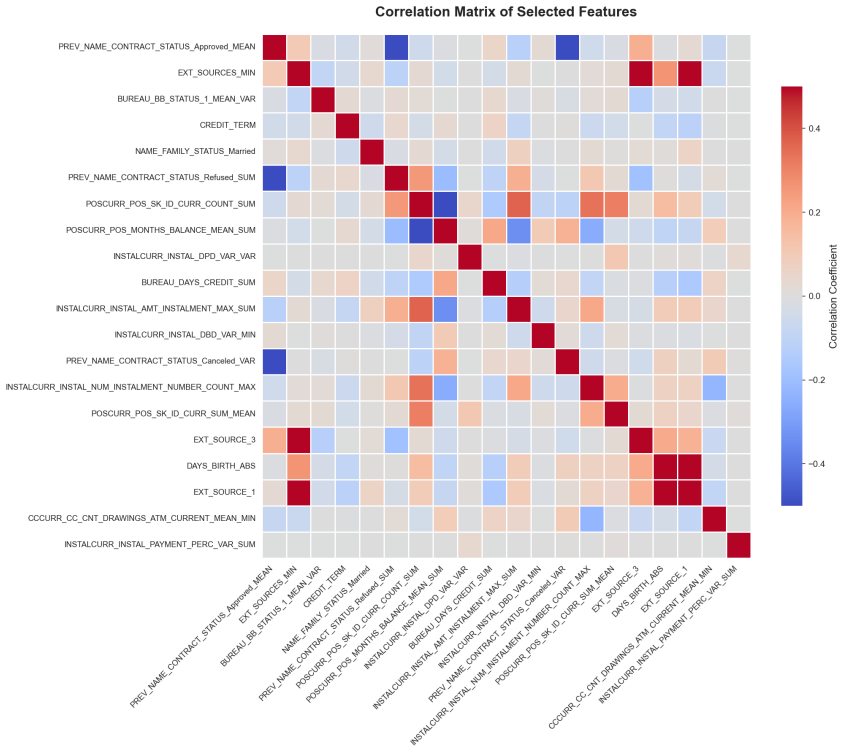

Figure 5: correlation matrix between top 20 features

4 Conclusion and discussion

Utilizing the Random Forest model, we identified the top 3,600 features based on SHAP values to forecast credit default risk, the target variable, achieving an AUC of 0.7582. The predicted target values predominantly fell within the range of 0–0.15. Subsequent to this, a detailed analysis and explanation were carried out on the significance of the top 20 indicators and the rationale behind their substantial impact on credit default risk. Additionally, a correlation matrix was constructed for these 20 features to identify variables exhibiting high correlations, followed by an in-depth analysis of the underlying reasons.

Based on our comprehensive feature analysis, we recommend that business applications, when assessing customer repayment risk, should prioritize the top 20 features we have identified. Furthermore, for features with significant impact, if they are easily accessible and interpretable within a business context, we suggest deriving more segmented indicators around these features (such as credit bureau status across different time periods) to enhance the model's predictive accuracy in evaluating credit default risk. Additionally, for customers with lower scores, it is advisable to consider offering more relaxed credit products, while for customers with higher scores, strengthening default risk management measures is recommended.

5 Contribution

Hanzhe CUI: Wrote code, reviw the report and coordinated the work among the team

Qun ZHENG: draw the figures, wrote the report for sections 3.1

Jingying WANG: wrote the report for sections 1-2

Wen HUANG: wrote the report for sections 3-4