Grenze International Journal of Engineering and Technology, June Issue

Risk Analysis for Home Credit Default

T. Jalaja[1], T. Adailakshmi[2], S. Chaitanya Deepthi[3], D. Swetha[4] and E. Shailaja[5]
[1-5]Department of Computer Science & Engineering, Vasavi College of Engineering (Autonomous) Ibrahimbagh, Hyderabad-31, India

***Abstract*— Housing Loan prediction is a major challenge for many financial Organizations since predicting correct outcome can reduce the risk and improve the decision-making during the process of loan. This study includes a machine learning-based method that predicts home loans default based on real-time data such as borrowers, financial profiles, credit history, and loan amount. The preprocessing tasks like handling missing data and normalization of numerical features are performed to guarantee data quality. The process of model building increased prediction accuracy by combining several algorithms such as logistic regression, decision trees, and integration methods. Important performance metrics like accuracy, precision, recall, and F1 score are calculated to test the model performance. This analysis gives lenders a clearer picture of the most important factors that determine default rate.**

I. INTRODUCTION

Risk analysis for home credit default involves finding the probability that a borrower will fail to make payments on a home loan (i.e default) and the possible consequences on the lender. Financial institutions conduct risk analysis to decrease their losses, allocate suitable resources, and make clearer decisions about lending loans. It includes both qualitative and quantitative assessments, often making use of statistical models and data analytics. A wide range of socioeconomic, psychological, situational, and behavioral factors were explored, and their influence on predicting the expected outcome domain across various time points was assessed.

- Perform exploratory data analysis (EDA) to understand the distribution and characteristics of the home credit dataset.
- Identify key features influencing credit default through data visualization.
- Develop predictive models using machine learning algorithms to predict the likelihood of de-fault for home credit applicants.
- Evaluate the performance of the models using suitable evaluation metrics and techniques.
- Provide recommendations to reduce the credit default risk based on the analysis and modeling results.

II. LITERATURE SURVEY/ RELATED WORKS

The field of credit risk modeling has seen a important evolution, moving from simple statistical techniques to highly complex, data-driven machine learning algorithms.

A. Key Forces in current Risk Modeling: The Role of Systemic Shocks and Correlated Defaults

The most important push for the evolution of traditional statistical methods to advanced machine learning was the inefficiency of already created models during systemic risk events.

Grenze ID: 01.GIJET.12.2.62_24

The 2008 Global Financial Crisis provides us the essential case study, revealing the built-in instability of models that could not provide justification for sudden external changes and the correlated nature of default risk.

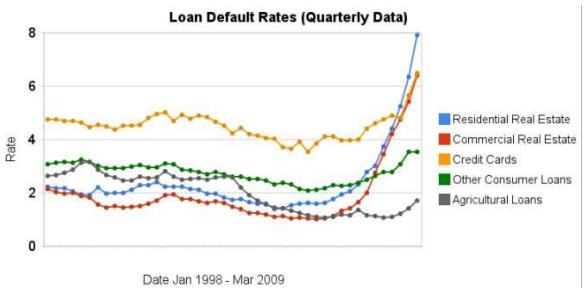

Figure 1: A chart showing the period of relative stability in default credit rates followed by the sharp hike in multiple credit sectors causing the Financial Crisis in 2008.

Analysis of Figure 1 The chart displays two economic systems. In the pre-crisis period (1998–2007), default rates among sectors operated within predictable limits. While risk levels varied (e.g., higher for credit cards), behavior was stable for static models like Logistic Regression. It's clear that we need adaptive models because the current state of the economy includes sudden shocks that show how static models doesn't work. Changes in the market can change the price of risk in the credit sector fastly.

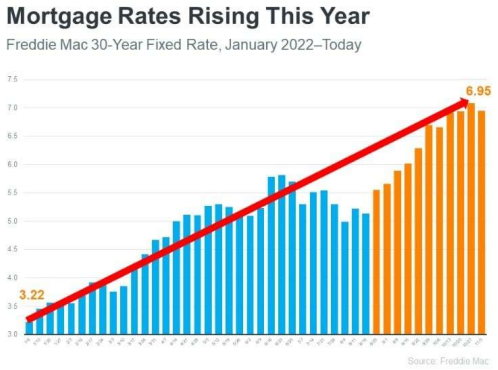

Figure 2: A chart showing the sudden increase in UK fixed mortgage rates in the immediate aftermath of the September 2022 "Mini-budget," showing the market's dependence to policy shocks.

Analysis of Figure 2 This above graph shows how an event in foreign policy can change a trend in the market that was expected. From October 2021 to early September 2022, mortgage rates has done up regularly. But the announcement of the "Mini-budget" caused a sudden break, and rates went up from 4.5% to more than 6.5%.
The drift phenomenon is a big problem for credit risk modeling because it changes the statistical properties of a model over time, making it less accurate. A model's training data is what it is based on, and if the financial system changes quickly, it may not work anymore.

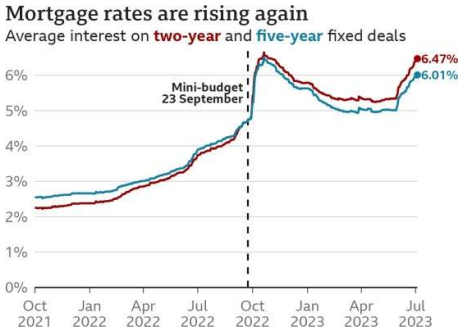

Figure 3: A chart determining the rapid increase of the Freddie Mac 30-Year Fixed Mortgage Rate in 2022, demonstrating a primary shift in the cost of credit.

Analysis of Figure 3 This chart shows a tight monetary policy cycle. The 30-year fixed mort-gage rate went from 3.22% in January to almost 7% by October, more than doubling in less than a year. This sudden rise caused a huge economic shock for people buying homes and changed the risk profile of borrowers.

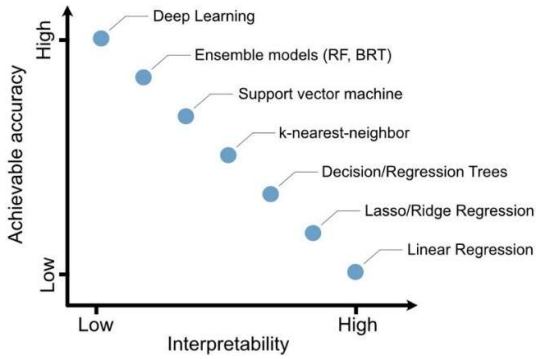

Figure 4: Accuracy versus interpretability in various models

III. Proposed System

In the financial sector, data is often hard to get and clean because it changes all the time. Home Credit, on the other hand, has a well-structured and fairly clean dataset with information about borrowers organized into several relational tables. It has information about the applicant, like their gender, age, family size, job, and credit history from both outside bureaus and the Home Credit Group.

One of the biggest problems is that there is a class imbalance, with most applicants not de-faulting and only a small number defaulting. This imbalance can change the results of a model and make standard classification algorithms less useful.

A. Data Specifications

The dataset consists of ten CSV files, each having information related to loan applications and credit history of clients of Home Credit

- application {train,test}.csv
- bureau.csv
- bureau balance.csv
- POS CASH balance.csv
- credit card balance.csv
- previous application.csv
- installments payments.csv

This dataset provides wide range of client's financial behavior and credit data, allowing clear risk analysis and model development for credit score calculation.

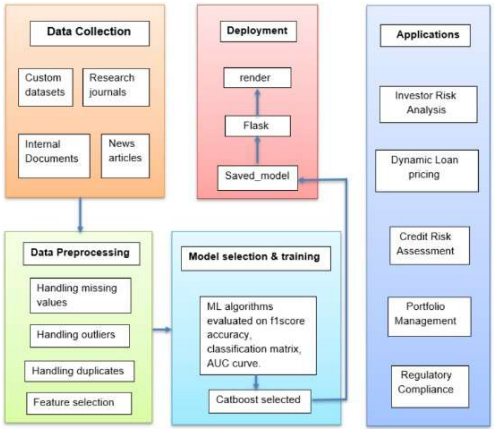

Figure 5: System Workflow showing the steps from data collection to deployment and usecases

IV. METHODOLOGY

The suggested method uses a structured pipeline to deal with the problems of the Home Credit dataset. The goal is to turn unprocessed data into useful features and teach models how to guess the TARGET variable. There are three steps in the process: Exploratory Data Analysis , Feature Engineering, Predictive Modeling, and Model Evaluation.

A. Exploratory Data Analysis (EDA) and Feature Engineering

This phase forms the foundation of the methodology, as data quality and feature relevance determine model performance.

Exploratory Data Analysis (EDA) EDA was performed to understand data structure and relationships before transformations. This includes:

- Univariate Analysis: Evaluating feature distributions to detect noise and outliers.
- Bivariate Analysis: Studying relationships between key features and the TARGET variable.
- Correlation Analysis: Identifying multicollinearity among numerical features to improve model stability.

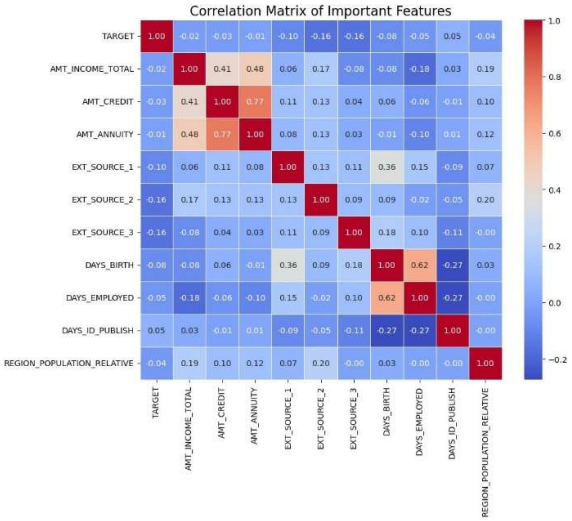

Figure 6: Heatmap of Pearson correlation coefficients in top numerical features

Data Integration and Aggregation Using SK ID CURR as the key, all of the relational tables were put together into one DataFrame. Aggregation methods were used for one-to-many relation-ships. To turn historical data into fixed features, numerical aggregates like COUNT, SUM, MEAN, MAX, MIN, and STD and categorical aggregates like NUNIQUE and MODE were calculated.

Feature Engineering Beyond aggregation, domain-specific features were also engineered, such as:

- Debt Burden Ratios: CREDIT INCOME PERCENT, ANNUITY INCOME PERCENT.
- Loan Structure Features: CREDIT TERM, GOODS PRICE RATIO.

Data Cleaning and Preprocessing

- Dealing with Missing Values: We filled in the numerical features with the median or mean and treated the categorical features as a ”missing” category.
- Categorical Encoding: Used One-Hot Encoding to stop false ordinal relationships.
- Outlier Management: We used the IQR and 3-standard deviation methods to limit outliers.
- Feature Scaling: We used StandardScaler to make sure that all the numerical features were the same.

B. Predictive Modeling

A set of classification models was used to calculate predictive performance, from simple to advanced methods:

- Logistic Regression: Used as a baseline for comparison and interpretability.
- Ensemble Models:
  - Random Forest & XGBoost: Used for improved predictive performance.
  - CatBoost: A gradient boosting method that builds trees sequentially to reduce errors.
- Multi-Layer Perceptron: A neural network used to learn complex feature representations.

C. Model Evaluation and Validation Strategy

A structured validation strategy and suitable metrics were used for imbalanced data.

- Validation: Data split into 80% training and 20% validation using stratified sampling; test set kept unseen.
- Metrics:
  - Precision, Recall, F1-Score: Focus on minority class, with Recall emphasized.
  - AUC-ROC: Primary metric for imbalanced classification.

The ROC curves in Figure 7 show the superior performance of ensemble models. They plot True Positive Rate (TPR) vs False Positive Rate (FPR) across thresholds. Curves closer to the top-left indicate better differentiation between defaulters and non-defaulters.

The ROC curves show that CatBoost (AUC = 0.75), Logistic Regression (0.74), and DNN (0.74) outperform others with higher TPR across FPR values. In contrast, Decision Tree (0.55) and SVM (0.61) perform poorly, with curves close to the diagonal, indicating near-random performance.

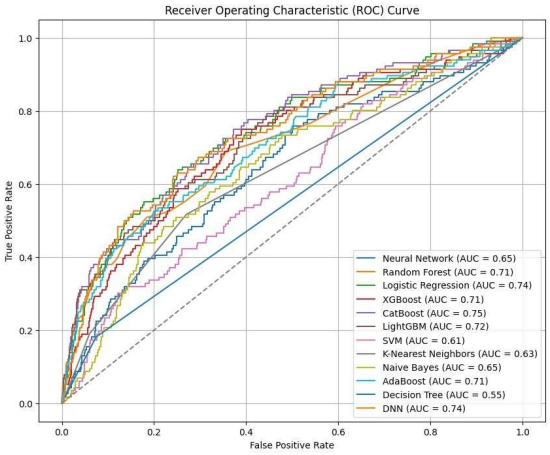

Figure 7: ROC curves comparison across models. The curve closest to the top-left corner with the highest AUC indicates the strongest model in distinguishing defaulters from non-defaulters.

TABLE I: PERFORMANCE COMPARISON OF ML ALGORITHMS

ML Algorithm	F1 Score	Accuracy	AUC
Neural Network	0.1156	90.36	0.65
CatBoost	0.0496	92.75	0.75
Logistic Regression	0.0168	92.68	0.74
LightGBM	0.1085	92.75	0.72
Random Forest	0.0333	92.69	0.71
XGBoost	0.0896	92.31	0.71
Decision Tree	0.1707	87.15	0.55

The results in Table 1 show that ensemble models beat Logistic Regression. CatBoost achieved the highest accuracy of [0.9275%], making it the most effective model, with a clear improvement over the baseline.

Random Forest also performed well, but CatBoost showed a slight advantage due to its boosting approach. The DNN outperformed the baseline but did not exceed ensemble models, which is common for structured data.

V. CONCLUSION

This study shows a risk analysis of home credit default using EDA, feature engineering, and machine learning. It recognize key patterns in borrower demographics, financial behavior, and loan history that effect default.

Various models were evaluated, including linear, tree-based, ensemble, and deep learning meth-ods. Ensemble models like CatBoost and LightGBM performed best across metrics such as accuracy, F1-score, and AUC-ROC, showing their effectiveness for complex and imbalanced data.

The study also introduced a credit scoring mechanism mapped to FICO ranges, providing both classification and an interpretable measure of creditworthiness, which is important for stakeholders. Overall, the results highlight the potential of machine learning to improve credit risk assessment, reduce financial losses, and enhance transparency in lending decisions. from loan defaults, and improve the transparency of lending decisions. The

integration of exploratory analysis, modern predictive modeling techniques, creates a strong foundation for operational credit default prediction
systems.

FUTURE WORK

The results show a lot of promise, but more work is required before they can be used in real life.

A. Optimizing Hyperparameters

Using grid search, random search, and Bayesian optimization to tune hyperparameters will make the model work much better.

B. Fairness and Bias Reduction

A fairness audit will look at how well people of different races, genders, and ages do their jobs, and bias mitigation techniques will make sure that decisions are made only in an ethical way.

C. Cloud Deployment That Can Grow

Using Docker and Kubernetes to deploy on cloud platforms like AWS, GCP, and Azure will make it possible to scale and run applications in real time.

REFERENCES

[1] B. A. C¸ allı and E. Co¸skun, "A Longitudinal Systematic Review of Credit Risk Assessment and Credit Default Predictors," SAGE Open, vol. 11, no. 4, Oct. 2021. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/21582440211061333

[2] W. Koehrsen, "Start Here: A Gentle Introduction," Kaggle. [Online]. Available: https://www. kaggle.com/code/willkoehrsen/start-here-a-gentle-introduction

[3] "Featured Prediction Competition," Kaggle, Home Credit Default Risk. [Online]. Available: https://www.kaggle.com/competitions/home-credit-default-risk/discussion/59347

[4] U. Aslam, H. I. T. Aziz, A. Sohail, and N. K. Batcha, "An Empirical Study on Loan De-fault Prediction Models," Journal of Computational and Theoretical Nanoscience, vol. 16, no. 8, pp. 3483–3488, Aug. 2019, doi: 10.1166/jctn.2019.8312. [Online]. Available: https: //www.researchgate.net/publication/335966806