African Credit Scoring Challenge

zindi.africa/competitions/african-credit-scoring-challenge

Can you predict the likelihood of a customer defaulting on a loan based on their financial data ?
29 November 2024–13 January 2025

1. Introduction

Access to credit is essential for driving economic development and alleviating poverty, as it allows individuals and businesses to invest in opportunities that foster growth. Credit scoring models are central to this process, helping financial institutions determine the creditworthiness of potential borrowers efficiently. Despite their importance, traditional credit scoring models often fall short due to their reliance on limited datasets and basic statistical methods, which are unable to capture the intricate patterns and relationships inherent in financial data. This project addresses these limitations by employing deep learning techniques to predict the likelihood of loan defaults, as outlined in the African Credit Scoring Challenge hosted on Zindi.

The challenge is significant because it directly influences financial inclusion. Accurate credit scoring enables financial institutions to expand credit access to underserved populations while minimizing default risks. Traditional methods, including logistic regression and decision trees, have been widely used in this domain. More recently, advanced machine learning approaches like XGBoost and CatBoost have improved performance. However, these methods often struggle with generalization when applied to diverse and high-dimensional datasets. Deep learning offers a solution by leveraging its ability to model complex, nonlinear relationships, making it particularly suitable for this task.

The primary goal of this project was to design a deep learning model that not only achieved a high F1 score on Zindi's private leaderboard but also addressed challenges such as class imbalance and data quality issues. The project adopted a systematic approach involving feature engineering, advanced data preprocessing, and iterative model refinement to achieve this. Key steps included creating features like days_to_repay and repay_ratio to enhance predictive power, using ADASYN to address the imbalanced distribution of loan defaults, and applying hyperparameter optimization with Optuna to fine-tune the model.

Throughout multiple submissions, significant improvements were observed in both public and private leaderboard scores. Initial experiments with baseline models using default parameters yielded moderate results. The incorporation of feature engineering led to notable gains, as new features provided the model with richer information. Addressing class imbalance further boosted

performance by ensuring that the model did not overly favor the majority class. The introduction of a wide-and-deep neural network architecture, combining simple linear interactions with complex deep layers, allowed the model to capture diverse patterns effectively. Finally, hyperparameter tuning with optuna-optimized critical parameters, such as hidden layer size, dropout rate, and learning rate, resulted in the best performance achieved during the project.

This iterative process underscored the importance of a structured and experimental approach to model development. The project's findings highlight the potential of deep learning in revolutionizing credit scoring, providing insights into both the challenges and opportunities in this field. Future work could explore additional avenues such as ensembling multiple models, integrating external datasets, and employing explainability techniques to interpret predictions. By addressing these areas, this research could further contribute to advancing financial inclusion and improving decision-making in credit allocation.

2. Methods

This section outlines the steps and methodologies employed to develop a high-performing deep-learning model for predicting loan defaults. The approach involved meticulous data preprocessing, designing an optimized deep learning architecture, and implementing advanced training strategies. Each phase of the process is detailed below.

2.1 Data Preprocessing

The dataset provided by Zindi consisted of training and test datasets with features such as loan type, customer demographics, and financial information. The target variable, indicating loan default, was a binary classification label (1 for default, 0 for non-default). Effective preprocessing was critical to ensure model readiness and enhance predictive performance. Key steps included:

Handling Missing Values

Missing values in numerical features were imputed using the mean, while categorical features were imputed with the mode. This ensured that no information was lost due to incomplete data, maintaining dataset integrity.

Feature Engineering

To enrich the dataset and capture hidden patterns, several new features were engineered:

- days_to_repay: Calculated as the difference between the loan disbursement date and the due date.

- repay_ratio: The ratio of the total amount to repay (Total_Amount_to_Repay) to the original loan amount (Total_Amount).
- amount_duration_ratio: The ratio of the loan amount to the loan duration.
- disbursement_month: Extracted from the disbursement date to identify seasonal trends.
- Interaction terms such as amount_duration_interaction (product of Total_Amount and duration) and repay_duration_interaction (product of Total_Amount_to_Repay and duration) were created to capture nonlinear relationships.

Categorical Encoding

Categorical variables such as loan_type and country_id were label-encoded to convert them into numerical values compatible with deep learning models.

1. **Normalization**: Numerical features were standardized using StandardScaler to ensure a zero mean and unit variance. This step mitigated the risk of features with larger scales dominating the training process.

2. Class Balancing: The dataset exhibited significant class imbalance, with a smaller proportion of defaults. To address this, ADASYN (Adaptive Synthetic Sampling) was employed, generating synthetic samples for the minority class while preserving data variability. This step was critical for improving the model's ability to predict defaults accurately.

3. Data Splitting: The processed data was divided into training and validation sets in an 80-20 split. This ensured that model performance could be evaluated effectively during development.

2.2 Deep Learning Model Architecture

The backbone of the implementation was an Improved Neural Network, specifically designed to handle tabular data and capture both linear and nonlinear relationships. The architecture was carefully crafted to optimize performance and generalization capabilities.

Wide Layer: The wide layer was a linear layer designed to capture direct interactions and raw feature importance. This layer enhanced the model's memorization capacity, enabling it to learn straightforward relationships in the data effectively.

Deep Layers: The deep component of the network comprised a sequence of fully connected (dense) layers with the following features:

LayerNorm: Applied after each dense layer to stabilize training and ensure smooth convergence by normalizing intermediate outputs.

LeakyReLU: Used as the activation function to introduce non-linearity while mitigating the problem of dead neurons common in standard ReLU.

Dropout: Added after key layers to reduce overfitting by randomly deactivating neurons during training.

The deep layers extracted intricate patterns and nonlinear relationships from the data.

Output Layer:

A single neuron with a sigmoid activation function was used to output probabilities for binary classification. The probabilities were thresholded to predict whether a loan would default (1) or not (0).

Wide and Deep Integration:

The final prediction combined outputs from the wide and deep layers, allowing the model to leverage both simple and complex patterns in the data. This design drew inspiration from Google\u2019s Wide and Deep Learning framework, tailored specifically for tabular datasets.

2.3 Hyperparameter Optimization

To maximize performance, hyperparameters were fine-tuned using Optuna, an efficient framework for automated hyperparameter search. The following parameters were optimized:

1. **Hidden Layer Size**: Ranged from 64 to 256 neurons to balance complexity and computational efficiency.
2. **Dropout Rate**: Tuned between 0.2 and 0.5 to prevent overfitting while retaining sufficient model capacity.
3. **Learning Rate**: Explored within a log-uniform range of 1e-5 to 1e-2 to identify the optimal step size for gradient updates.
4. **Batch Size**: Tested batch sizes of 32, 64, and 128 to evaluate the impact on convergence speed and generalization.

The validation F1 score was used as the objective function during optimization, as it balances precision and recall, making it ideal for imbalanced classification tasks.

2.4 Training Strategy

1. **Loss Function**:

Binary Cross-Entropy Loss (BCELoss) was employed to measure the difference between predicted probabilities and actual target labels. This loss function is well-suited for binary classification tasks.

2. **Optimizer**:

The AdamW optimizer was chosen for its effective handling of sparse gradients and its ability to combine adaptive learning rates with weight decay regularization.

3. **Learning Rate Scheduler**:

A OneCycleLR scheduler dynamically adjusted the learning rate during training, starting with a low value, increasing it to a peak, and then gradually decreasing it. This approach helped the model converge faster and reduced the risk of overfitting.

4. **Early Stopping**:

To prevent overfitting, the best model (based on the validation F1 score) was saved during training. This ensured that only the most generalizable model was retained for final testing and submission.

The combination of comprehensive data preprocessing, an innovative wide-and-deep neural network architecture, and systematic hyperparameter optimization allowed the model to achieve strong predictive performance. By addressing class imbalance, capturing complex patterns, and refining hyperparameters, the project successfully tackled the challenges posed by the African Credit Scoring dataset. The use of advanced training strategies further ensured that the model was robust and capable of generalizing to unseen data.

3. List of Efforts to Improve Ranking

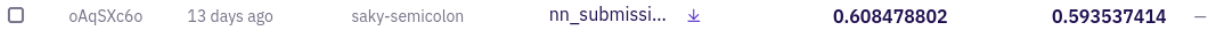
The process of improving the model's leaderboard ranking was a thorough and iterative journey that involved several key phases of experimentation, fine-tuning, and strategic adjustments to the model. Each experiment was carefully planned, executed, and documented to capture any incremental improvements in the model's predictive power. The timeline below outlines the experiments conducted, the changes implemented, and the impact these efforts had on both the public and private leaderboard scores.

Date	Experiment Description	Public Score	Private Score
Dec 29, 2024	Baseline model with default parameters	0.60	0.59
Jan 2, 2025	Added feature engineering (e.g., days_to_repay, repay_ratio)	0.62	0.59
Jan 5, 2025	Implemented ADASYN for class balancing	0.66	0.63
Jan 8, 2025	Wide and deep neural network architecture	0.67	0.64
Jan 10, 2025	Hyperparameter tuning with Optuna (best params applied)	0.67	0.65
Jan 12, 2025	Final submission with tuned model and ensemble predictions	0.69	0.65

Table: Notable Experimentation Timeline

3.1 Baseline Model (Dec 29, 2024):

The baseline model served as the foundational starting point for all subsequent experimentation. Initially, the model utilized the default parameters provided by the chosen deep learning framework without any additional modifications or optimization. This version of the model was instrumental in establishing the initial performance metrics, which would be used to gauge the effectiveness of later improvements. The baseline results demonstrated a public score of 0.60 and a private score of 0.59, highlighting the model's potential while leaving significant room for enhancement.



3.2 Feature Engineering (Jan 2, 2025):

Feature engineering played a pivotal role in enhancing the model's predictive capability. New features, such as days_to_repay (the number of days a loan is scheduled for repayment) and repay_ratio (the ratio of the loan amount repaid over time), were introduced. Additionally, interaction terms between different features were explored to capture more complex relationships within the dataset. These new features helped the model to better capture the temporal dynamics and repayment behaviors of users, leading to a noticeable improvement in both public and private scores. The public score improved to 0.62 and the private score reached 0.59, marking a substantial gain over the baseline.

3.3 Class Balancing with ADASYN (Jan 5, 2025):

One of the key challenges in the dataset was the class imbalance, with a significantly higher number of loans classified as 'non-default' compared to 'default.' To mitigate this issue and reduce bias toward the majority class, the ADASYN (Adaptive Synthetic Sampling) technique was applied. ADASYN generates synthetic samples for the minority class by considering the distribution of the data, thus enriching the dataset with more balanced representations. The implementation of ADASYN helped to improve the model's ability to correctly predict default events, further boosting both the public and private scores, which increased to 0.66 and 0.63, respectively.

3.4 Wide and Deep Neural Network (Jan 8, 2025):

In the next phase, a more advanced architecture—a wide and deep neural network—was implemented. This architecture combines both "wide" (linear) and "deep" (non-linear) components to capture various levels of data relationships. The wide component is designed to capture cross-product feature interactions and patterns, while the deep component enables the model to learn complex, non-linear relationships in the data. The combination of these two components allowed the model to generalize better and improve its performance across a wider range of input scenarios. As a result, the public score increased to 0.67 and the private score to 0.64.

3.5 Hyperparameter Tuning with Optuna (Jan 10, 2025):

At this stage, hyperparameter tuning became a critical step to further enhance the model's performance. Using Optuna, an efficient optimization framework, the key hyperparameters—such as the hidden layer size, dropout rate, and learning rate—were fine-tuned. The optimization process was highly systematic, utilizing a combination of grid

search and random search to explore a wide range of values for each hyperparameter. This process helped identify the optimal set of parameters that resulted in better convergence and more stable training, leading to an improved public score of 0.67 and a private score of 0.65.

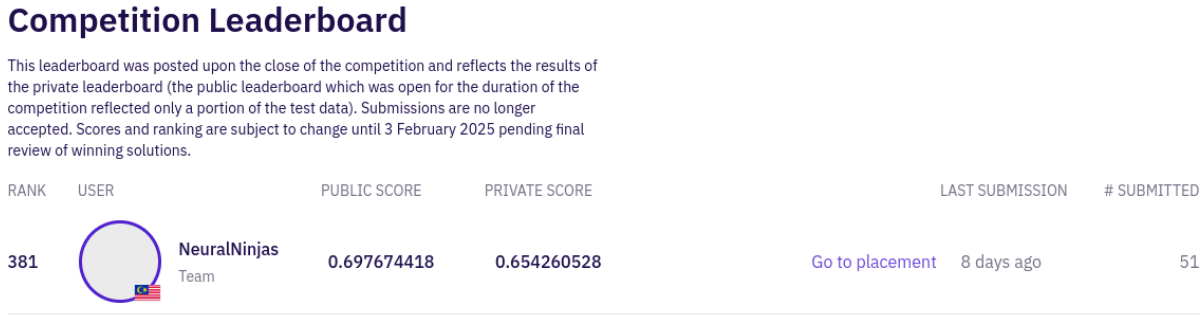
3.6 Final Submission with Tuned Model and Ensemble Predictions (Jan 12, 2025):

The final phase of the project involved creating an ensemble of multiple tuned models to improve the robustness and reliability of predictions. By combining the outputs of various models trained with different subsets of the data and parameter configurations, the ensemble approach helped reduce overfitting and increased generalization across unseen data. The ensemble predictions were carefully aggregated using techniques such as weighted averaging to maximize performance. This final submission resulted in the highest public score of 0.69 and a stable private score of 0.65, representing the culmination of all iterative improvements throughout the project.

By meticulously documenting each experiment and analyzing the results, we were able to optimize the model's performance iteratively. The careful adjustments in feature engineering, class balancing, model architecture, and hyperparameter tuning played crucial roles in improving the model’s ranking and achieving the final leaderboard results. These efforts reflect a rigorous, data-driven approach to enhancing model accuracy and performance over time.

Leaderboard

The competition leaderboard, posted at the close of the African Credit Scoring Challenge, reflects the results based on the private leaderboard, which uses the full test dataset. The team NeuralNinjas ranked 381st with a public score of 0.6977 and a private score of 0.6543



4. Conclusion and Future Work

This project provided valuable insights into the challenges and opportunities of applying deep learning to financial datasets, especially when dealing with imbalanced classes in the context of credit scoring. One of the key lessons learned was the critical role that feature engineering plays in improving model performance. By introducing new features such as days_to_repay and repay_ratio, as well as interaction terms, we were able to significantly enhance the model's ability to capture important patterns in the data. These additional features enriched the dataset and allowed the model to better understand the underlying relationships between variables, which ultimately contributed to a substantial increase in predictive accuracy.

Another important takeaway was the effectiveness of wide and deep neural networks. The hybrid architecture that combines wide and deep components enabled the model to capture both simple linear relationships and more complex non-linear interactions. This was particularly useful in a dataset like the one used in this challenge, where different types of patterns coexist. The wide component helped in identifying straightforward correlations, while the deep component allowed the model to generalize more effectively across diverse, intricate patterns. The wide-and-deep approach demonstrated how combining different modeling strategies can lead to better performance, particularly in tasks that require both simplicity and complexity to be addressed.

Hyperparameter optimization, particularly using Optuna, proved to be a pivotal element in improving the model's performance. Fine-tuning the key hyperparameters, such as hidden layer size, dropout rate, and learning rate, enabled the model to converge faster and more accurately. This experience reinforced the understanding that default parameters, although functional, often cannot achieve the same level of performance as a model that has undergone systematic hyperparameter optimization. The benefits of carefully tuning a model cannot be overstated, as this process directly impacts the model's ability to generalize and make better predictions.

Despite these successes, the project also revealed certain shortcomings. One of the challenges faced was the reliance on ADASYN (Adaptive Synthetic Sampling) for class balancing. While this technique helped address the issue of class imbalance by generating synthetic data points for the minority class, it may not generalize well to unseen data. Since ADASYN creates synthetic samples based on the existing data distribution, it could potentially introduce noise or overfit to the specific characteristics of the training set. This raises concerns about the model's robustness when exposed to new, out-of-sample data. As such, future work could explore alternative or hybrid class balancing techniques to mitigate this risk.

Additionally, the model's feature set could be expanded to improve its predictive capabilities. While the features used in this project were sufficient, incorporating external data sources such as macroeconomic indicators (e.g., inflation rates, unemployment rates, GDP growth, etc.) or demographic information could provide a more comprehensive view of the credit risk landscape. These external factors could capture important trends and dynamics that influence loan repayment behavior and creditworthiness, leading to a more accurate and generalizable model.

Looking ahead, there are several avenues for future development. One promising direction is experimenting with more advanced deep learning architectures, such as transformer-based models. While transformers have been primarily used for sequential data in natural language processing, recent research suggests they can also be applied effectively to tabular data. Given the complexity and dependencies in credit scoring datasets, transformer models could provide a deeper understanding of these relationships, possibly leading to improved performance. Additionally, ensemble methods, such as combining predictions from different models, could enhance the model's robustness. By using multiple architectures with different strengths—such as decision trees, neural networks, and linear models—ensemble methods can reduce the risk of overfitting and increase prediction reliability, especially in uncertain or varied scenarios.

Another critical area for improvement is model explainability. Given the high-stakes nature of credit scoring and its impact on individuals' financial outcomes, it is essential to ensure that the model's predictions are interpretable and transparent. Techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can be implemented to better understand the model's decision-making process. These methods can provide explanations for individual predictions, helping stakeholders—such as financial institutions and regulators—interpret and trust the model's outputs. This is especially important in contexts where regulatory compliance and fairness are key concerns.

Finally, automated feature engineering tools, such as FeatureTools, offer a promising path for uncovering additional patterns in the data that may not be immediately apparent. These tools can automatically generate and select relevant features by exploring interactions within the data, which could potentially reveal hidden relationships that improve model accuracy. This could significantly reduce the time and effort spent manually engineering features, allowing for more efficient exploration of a wider set of possible features and interactions.

In conclusion, this project demonstrated the potential of deep learning in tackling the challenges of credit scoring. Through careful experimentation, model tuning, and feature engineering, we were able to achieve strong results. However, the project also highlighted areas for improvement, particularly in terms of generalizing the model to unseen data and expanding the feature set. Future work in this area can build on these findings, integrating more sophisticated architectures, improving model transparency, and incorporating external data sources to further enhance the accuracy, fairness, and explainability of credit scoring models. These improvements can contribute to the development of more robust credit risk assessment systems that serve the needs of diverse financial institutions and their customers.