Enhanced Credit Risk Acquisition Scoring via Noise-Augmented Feature Selection and Bayesian Hyper-parameter Tuning

Arijit Ganguly
Revolut Group Holdings Ltd, Bangalore, India
Global Credit Management – Data Science
Data Scientist
arijit.ganguly@revolut.com

August 27, 2025

Abstract

Effective credit risk acquisition scoring is crucial for financial institutions, where gradient-boosted models, such as XGBoost and LightGBM, are increasingly prevalent. However, their performance is contingent upon appropriate feature selection and hyper-parameter tuning. Feature selection identifies pertinent predictors, reducing complexity and enhancing interpretability, while hyper-parameter tuning optimizes model configurations, mitigating overfitting and maximizing predictive accuracy. This study investigates advanced feature selection and hyper-parameter optimization techniques to improve gradient-boosted model performance in credit risk acquisition.

A novel feature selection method, employing random noise variables, was explored. The method involves generating artificial noise features and training the gradient-boosted models on the augmented dataset. By comparing the feature importance of the original features with that of the noise features, less informative features are identified and eliminated. This technique offers an improvement over recursive feature elimination (RFE), a widely used method, by providing a more robust assessment of feature importance through comparison with noise variables, thus aiding in filtering out features that may appear important but do not truly contribute to the model's predictive power. This was tested on an acquisition model development with an initial set of 6000 bureau features. The noise-based technique yielded a more optimal feature subset, achieving a significant dimensionality reduction from over 6,000 features to fewer than 200 in a single step, whereas RFE required significantly more computational time to reach a comparable, yet less refined, outcome. This direct comparison underscores the proposed method's ability to identify truly relevant features, improving model performance, interpretability and efficiency.

Subsequently, Bayesian hyper-parameter tuning, utilizing Optuna, was investigated. Optuna leverages past trial results to intelligently explore the hyper-parameter space, accelerating convergence to optimal configurations, unlike grid search's exhaustive approach. To empirically validate this advantage, a stratified 5-fold cross-validation scheme was implemented, maximizing average AUC-ROC, with overfitting controlled by cross-validation standard deviation and a custom overfit index. Results confirmed Optuna's efficacy in rapidly identifying optimal hyper-parameters, supporting its superiority over grid search.

The paper will show how the combined application of the proposed robust feature selection method and Bayesian hyper-parameter optimization enhances predictive accuracy, robustness, and computational efficiency of gradient-boosted models in credit risk acquisition scoring.

1 Introduction

The accurate assessment of creditworthiness for new applicants is paramount for sustainable growth and effective risk management within the financial sector. This challenge in credit risk involves precise acquisition scoring, which is vital for financial institutions as it directly impacts profitability, portfolio quality, and adherence to regulatory compliance. Gradient-boosted models, such as XGBoost and LightGBM, represent powerful ensemble learning methods that sequentially construct a strong predictive model from a series of weaker models. These models have become increasingly prevalent in credit risk assessment due to their high predictive power and inherent ability to capture complex non-linear relationships within data.

Despite their significant strengths, gradient-boosted models encounter notable challenges. Firstly, the presence of irrelevant or redundant features can introduce noise into the model, escalate its complexity, and frequently lead to overfitting. Overfitting occurs when a model learns the training data too meticulously, including its noise and outliers, which results in poor generalization to new, unseen data. This phenomenon often manifests as high accuracy on training data but significantly lower accuracy on validation or test sets. Secondly, suboptimal hyper-parameters can lead to either underfitting, where a model is too simplistic to capture the underlying patterns in the data, or overfitting, thereby hindering predictive accuracy. The process of finding the right balance of hyper-parameters is crucial to ensure the model's robustness and generalization capabilities in real-world applications. The implications of these challenges extend beyond mere technical performance; they represent a substantial hidden cost to financial institutions. Poor risk assessment due to overfit or inaccurate models can lead to higher default rates on granted loans, missed revenue opportunities from incorrectly declined applicants, and potential non-compliance with stringent financial regulations.

The proposed solution involves the application of two advanced techniques: Noise-Augmented Feature Selection (NAFS) and Bayesian Hyper-parameter Tuning (BHT). NAFS specifically addresses the problem of identifying truly predictive features by introducing random noise variables into the dataset. The core mechanism involves comparing the feature importance of the original features against that of these synthetic noise features. A robust filtering threshold is then established, allowing less informative features—those whose importance is comparable to or less than that of random noise—to be identified and eliminated. The inherent benefit of NAFS is its systematic retention of features with high predictive power, which intrinsically reduces the risk of overfitting by focusing exclusively on the most relevant predictors.

Complementing NAFS, Bayesian Hyper-parameter Tuning (BHT) tackles the problem of efficiently finding optimal model configurations. Unlike traditional exhaustive (grid search) or random search methods that explore the hyper-parameter space without learning from past trials, BHT employs an intelligent, directed search strategy. It constructs a probabilistic model of the objective function (e.g., model performance) based on the results of previous hyper-parameter combinations. This probabilistic model is then utilized to suggest the next set of parameters to evaluate, leading to faster convergence to optimal settings and significantly reducing the computational cost associated with tuning complex models.

The combined approach of NAFS and BHT is designed to unlock the full potential of tree-based models. By first refining the input features to ensure only truly predictive variables are utilized, and then meticulously fine-tuning the model's internal configurations, this integrated methodology results in credit risk acquisition models with higher discriminatory power and enhanced accuracy. This ultimately leads to more informed and effective lending decisions, improved portfolio quality, and increased sales potential for financial institutions. This integrated approach directly mitigates the business risks associated with suboptimal model performance, offering a pathway to more reliable and profitable credit operations.

2 Literature Review

Feature selection and hyper-parameter tuning are foundational steps in developing high-performing machine learning models, particularly in complex domains like credit risk assessment. This section reviews existing literature relevant to these two critical areas, highlighting their impact and the advancements that inform the proposed methodology.

2.1 Impact of Feature Selection (FS)

Feature selection is a crucial preprocessing step in machine learning that directly addresses the challenges posed by high dimensionality and redundancy in datasets. In the context of credit risk models, the presence of numerous features, many of which may be noisy, irrelevant, or highly correlated, frequently leads to overfitting and reduced model interpretability. Effective feature selection aims to identify the most pertinent predictors, thereby reducing model complexity and enhancing the model's ability to generalize to unseen data. This process of filtering features through various methods consistently improves both interpretability and performance, resulting in more parsimonious and understandable models.

Wong and Smith (2019) emphasize the critical role that robust feature engineering, encompassing selection, plays in optimizing machine learning models for credit risk assessment. Furthermore, Patel and Brown (2021) discuss how the interpretability challenges inherent in complex machine learning models can be significantly alleviated through judicious feature selection, leading to more transparent and explainable credit decisions. Various research efforts underscore the benefits of feature selection: MDPI (2023) highlights a range of techniques, including univariate methods, Recursive Feature Elimination (RFE), feature importance-based selection, and Information Value (IV), all of which are shown to improve model accuracy by effectively reducing noise and focusing on informative variables. Similarly, research presented in arXiv (2023) demonstrates the substantial positive contribution of various feature selection methods on the overall performance of machine learning models employed in credit scoring.

Noise-Augmented Feature Elimination (NAFE) represents an advanced approach that reinforces this filtering process by establishing statistical thresholds against synthetic noise variables. This method is specifically designed to identify truly informative features by comparing their importance against that of randomly generated noise, thus providing a more robust criterion for feature retention. The term ”robustness” in this financial context is not merely a technical descriptor; it signifies stability, reliability, and resilience of the model to fluctuations in market conditions or shifts in data patterns. This is paramount for regulatory compliance and long-term model performance, where even minor errors can lead to significant financial consequences. The emphasis on noise-based methods implicitly addresses the need for models that perform consistently under varying, real-world conditions, not solely on historical training data. Gharoun, Yazdanjoe, Khorshidi, and Gandomi (2023) introduced ”Noise-Augmented Boruta,” an innovative approach that incorporates noise into shadow variables for enhanced and more accurate feature selection, further validating the utility of noise-based methods.

2.2 Impact of Bayesian Hyper-parameter Tuning (BHT)

Bayesian Hyper-parameter Tuning (BHT) is a sophisticated optimization technique that addresses the inherent challenge of tuning numerous hyper-parameters in complex tree-based models like XGBoost and LightGBM. These models, while powerful, often possess a large number of configurable parameters that significantly influence their performance. Traditional tuning methods, such as exhaustive grid search or random search, explore the hyper-parameter space without leveraging information from previous evaluations, making them computationally expensive and often inefficient, especially for models with many parameters or costly objective functions.

BHT enhances efficiency and performance by intelligently exploring the hyper-parameter space. It operates by constructing a probabilistic model (often a Gaussian Process or Tree-structured Parzen Estimator) of the objective function (e.g., cross-validation performance like AUC) based on the results of previously evaluated hyper-parameter combinations. This probabilistic model is then used to intelligently suggest the next set of parameters to evaluate, aiming to maximize the expected improvement in the objective function. This directed search strategy leads to significantly faster convergence to optimal solutions compared to non-informed search methods.

The effectiveness of BHT is well-documented in recent literature. Preprints.org (2025) explicitly states that hyper-parameter optimization is essential for maximizing predictive performance and computational efficiency in credit risk modeling, and their work demonstrates that Optuna, a popular Bayesian framework, significantly outperforms both Grid Search and Random Search in terms of speed and effectiveness for tuning XGBoost and LightGBM. Research conducted on ResearchGate (2024) further supports this by applying Bayesian optimization (specifically TPE) to tree-based models for fraud detection, with findings that are directly applicable to credit risk assessment. Furthermore, Jenatton, Popov, and Bach (2017) discuss the broader concept of leveraging dependency structures in optimization domains for more efficient search, which is highly relevant for navigating the complex and often interdependent hyper-parameter spaces of advanced machine learning models. The emphasis on robust optimization methods in the literature review underscores the need for models that are stable and reliable under varying real-world conditions, a critical requirement for financial applications.

3 Methodology

This section outlines the detailed experimental design and procedures employed in this study, focusing on the implementation of Noise-Augmented Feature Elimination (NAFE) and Bayesian Hyper-parameter Tuning (BHT) with Optuna. These specific methodological choices are deliberate and designed to mitigate known machine learning challenges such as overfitting, computational cost, and model instability, which are particularly critical in a high-stakes financial modeling environment.

3.1 Noise Augmented Feature Elimination (NAFE)

The NAFE process is structured into three primary phases: data preparation and noise augmentation, noise-augmented feature filtering, and an optional correlation-based feature elimination.

3.1.1 Data Preparation & Noise Augmentation

Initially, historical credit risk data was loaded and subsequently segmented into distinct training and validation sets. This comprehensive dataset included all relevant features as well as the crucial target variable, typically representing the default status of an applicant. A critical step involved augmenting the existing dataset through the introduction of a predefined number of synthetic random noise features. These artificially generated features were meticulously designed to possess no inherent predictive power, serving as a robust baseline for comparison against the predictive power of the original features.

3.1.2 Noise-Augmented Feature Filtering

Following data preparation, an initial gradient-boosted model, specifically either an XGBoost or LightGBM model, was trained on the augmented training data. This preliminary training process utilized a predefined set of hyper-parameters and incorporated early stopping mechanisms to prevent overfitting during this phase. Early stopping is a crucial technique that monitors model performance on a validation set and halts training when performance no longer improves, effectively preventing the model from learning noise in the training data too closely.

Subsequent to initial model training, the importance of all features, encompassing both the real and the introduced noise features, was meticulously calculated. This calculation could be based on either the ’Gain’ metric, representing the average gain of splits where the feature was utilized, or on SHAP (SHapley Additive exPlanations) values, which provide a measure of each feature’s contribution to the model’s output. A critical step involved the removal of all features that exhibited a calculated feature importance of zero, as these features were deemed to have no contribution to the model’s predictions. This ensures that only features with at least some perceived relevance are carried forward. The core of NAFE lies in its noise thresholding. Features whose calculated importance was less than or equal to the maximum importance observed among the synthetic random noise features were specifically identified. These features were then marked for subsequent removal, as their predictive contribution was considered negligible or non-informative when compared to random chance. This rigorous, data-driven approach to feature selection goes beyond simple importance ranking by establishing a clear statistical baseline, aiming to identify truly predictive features and intrinsically reducing the risk of overfitting by focusing only on the most relevant predictors.

3.1.3 Correlation-Based Feature Elimination (Optional)

For the subset of remaining non-categorical features, a pairwise correlation matrix was computed. This analysis aimed to identify highly correlated features that might introduce redundancy or multicollinearity. In instances where a pair of features exhibited a correlation above a defined threshold (e.g., 0.5), the feature with the lower importance (as previously determined in Step 3.1.2) from that specific pair was identified for removal. This strategic step was implemented to further reduce multicollinearity within the feature set and streamline its overall composition. This comprehensive methodology was effective in generating a robust feature set for credit risk acquisition scoring.

3.2 Bayesian Hyper-parameter Tuning with Optuna

The approach leveraged Optuna, an open-source hyper-parameter optimization framework, to efficiently discover optimal configurations for XGBoost and LightGBM models.

3.2.1 Objective Function Definition

An objective function, specifically named *objective_xgb* or *objective_lgbm*, was meticulously defined. This function's primary purpose was to optimize the performance of either XGBoost or LightGBM models, making the optimization process adaptable to different gradient-boosting algorithms. Within the confines of this objective function, Optuna's *trial.suggest_* methods were extensively utilized to define the comprehensive search range for key hyper-parameters. These parameters included, but were not limited to, *learning_rate*, *max_depth*, *subsample*, *colsample_bytree*, and various regularization terms. This systematic definition allowed Optuna to intelligently explore the vast and complex parameter space.

To ensure a robust and reliable evaluation of model performance, a Stratified K-Fold Cross-Validation (with N_FOLDS folds) was rigorously performed for each trial. The model was trained on multiple folds of the data, and its performance metrics were subsequently averaged across these folds to provide a more stable assessment. During each individual model training phase within a trial, early stopping mechanisms were incorporated. This critical technique, based on the Area Under the Curve (AUC) metric calculated on the validation set, was employed to prevent overfitting and ensure that the model did not learn the training data's noise too closely. The objective function was designed to return the average validation AUC across all folds, which served as the primary performance metric that Optuna aimed to maximize during the optimization process. To accelerate the search and efficiently avoid unproductive trials, two distinct pruning criteria were implemented. Trials exhibiting an excessive difference between training AUC and validation AUC (indicating overfitting) were pruned. Additionally, trials with a high standard deviation of AUC across folds (indicating model instability) were also pruned, ensuring that only robust and generalizable configurations were considered. These pruning heuristics are proactive measures to ensure that the search focuses on stable and generalizable model configurations, which is vital for models deployed in a financial context where erratic performance is unacceptable.

3.2.2 Optuna Study Execution

An Optuna study object was meticulously initialized with the *direction="maximize"* setting. This configuration explicitly indicated that the primary goal of the optimization process was to maximize the validation AUC, guiding Optuna's search strategy. The *study.optimize* method was then invoked, with the previously defined objective function and the desired number of *n_trials* being passed as arguments. Optuna intelligently selected subsequent hyper-parameter combinations for each trial based on the results obtained from previous trials, leveraging a Tree-structured Parzen Estimator (TPE) algorithm. Parallelization, specified by *n_jobs=NCPUS*, was also utilized to significantly speed up the entire optimization process.

3.2.3 Training with Best Hyper-parameters

Upon the completion of the Optuna study, the optimal hyper-parameters that yielded the best performance were meticulously retrieved from *study.best_trial.* A final XGBoost or LightGBM model was subsequently trained on the training dataset using these optimally found hyper-parameters. This training also incorporated the refined feature set that had been identified during the preceding feature selection step. The final model's performance, specifically its AUC and associated confidence intervals, was re-evaluated on both the training/validation dataset and a separate, unseen test set. This comprehensive evaluation was performed to definitively confirm the model's generalization capabilities and its robustness on new data. This detailed methodology implicitly conveys the researchers' understanding of practical challenges in machine learning deployment and their proactive steps to build a robust system, not just a high-performing one on training data.

4 Business Case & Results

The practical application of the proposed methodology was demonstrated through a business case proof of concept (PoC) focused on enhancing credit risk acquisition scoring for credit card applicants using tree-based models. This section details the data utilized, the experimental design, and the empirical findings.

4.1 Data Overview (Retro Sample)

The dataset utilized for this study comprised a retro sample of historical credit risk data. The overall sample size encompassed 2 million records, which included various financial products such as Unsecured Personal Loans (UPLs), Credit Cards (CCs), and Overdrafts. Within this larger dataset, approximately 300,000 records pertained to Revolut users. For the specific scope of credit card acquisition modeling, a dedicated Credit Card (CC) sample was extracted, consisting of 900,000 records. Of these, approximately 140,000 records were associated with Revolut users, and among them, about 65,000 users had opened a credit facility after joining Revolut. The performance window for the granted facilities spanned from January 2020 to July 2022, with performance being tracked for up to 48 months. The models developed within this study were primarily built using the data from the Credit Card (CC) Sample.

4.2 Experimental Design

Various training dataset scenarios were considered to thoroughly evaluate the model's performance under different conditions and data availability. These scenarios included:

- **All users (Revolut + Non-Revolut) with/without eligibility checks (Bureau features):** This scenario utilized the full dataset encompassing both Revolut and non-Revolut users. Models were trained both with and without the inclusion of eligibility checks, which typically involve features derived from credit bureau data. Bureau features include information such as credit scores, historical payment behavior, existing credit lines, and public records, all crucial for assessing an applicant's external credit risk profile.

- **Revolut users with/without eligibility checks (Bureau, Email/Device features):** This scenario focused specifically on data pertaining to Revolut users. Training was conducted both with and without eligibility checks. The feature set for Revolut users was expanded to include not only bureau features but also internal data such as email and device-related features.

- **Revolut users with internal history with/without eligibility checks (Bureau, Email/Device, Transaction features):** This was the most comprehensive scenario for Revolut users, incorporating their internal transaction history in addition to bureau, email, and device features. Models were again trained with and without eligibility checks. Transaction features could include metrics such as average transaction value, frequency of transactions, spending categories, balance fluctuations, and international transaction patterns, offering a rich understanding of a user's financial behavior within the Revolut ecosystem.

The dataset was partitioned into distinct subsets to facilitate robust model training and evaluation. A Train/Test (80/20) split was applied to data spanning from August 2020 to March 2022. This portion of the data was used for the primary training and in-time testing of the models. It is noted that the in-time training data included an origination cohort that partially overlapped with the Covid period, a consideration necessitated by limitations of the retro data. However, model performance was benchmarked on multiple out-of-time periods including pre-Covid, Covid, and most-recent, and the proposed solution outperformed the benchmark on all Out-of-Time (OOT) periods. As mentioned, a separate dataset, comprising data from April 2022 to July 2022, was specifically reserved for out-of-time validation. This OOT sample was crucial for assessing the model's generalization capabilities on data collected after the primary training period, ensuring its performance robustness over time.

XGBoost and LightGBM were the algorithms employed for model development within this study. Key enhancements applied included Noise-Based Feature Selection, where 50 noise features based on a Gaussian distribution were introduced, and features with SHAP importance less than the maximum importance of noise features were eliminated. Hyper-parameter tuning utilized Optuna (Bayesian optimization) for efficient and robust optimization. Model selection and evaluation primarily relied on the Gini score on Train, Test, and Out-of-Time (OOT) samples, benchmarked against a vendor score.

4.3 Results

The empirical results demonstrate the effectiveness of the proposed Noise-Augmented Feature Elimination (NAFE) and Bayesian Hyper-parameter Tuning (BHT) approaches, both individually and in combination.

4.3.1 Noise-Augmented Feature Elimination (NAFE) versus Standard Recursive Feature Elimination (RFE)

The performance of Noise-Augmented Feature Elimination (NAFE) was rigorously compared against that of standard Recursive Feature Elimination (RFE) based on two primary criteria: model discriminatory power and computational efficiency. To ensure a direct and equitable comparison, measures were taken to ensure that both models ultimately utilized a similar number of features following their respective feature selection processes. This comparative analysis was conducted for both XGBoost and LightGBM algorithms, with a predefined set of hyperparameters being applied consistently across all evaluations.

The findings indicate that the feature selection process employing NAFE consistently resulted in a higher Gini coefficient compared to the feature selection process utilizing RFE. This performance uplift was observed for both the XGBoost and LightGBM algorithms. For XGBoost, the model developed with NAFE demonstrated superior performance over the model utilizing RFE on the In-time Train dataset (an absolute increase of +0.9%) and on the Out-of-Time Test dataset (an absolute increase of +0.3

Table 1: Discriminatory power assessment - Feature Selection

Algorithm	**FS Process**	**Time to completion (mins)**	**Gini (In-Time Train)**	**Gini (In-Time Test)**	**Gini (Out-Of-Time Test)**
XGBOOST	NAFE	5.5	0.767	0.669	0.639
XGBOOST	RFE	23.0	0.758	0.669	0.636
LGBM	NAFE	5.0	0.750	0.655	0.624
LGBM	RFE	21.1	0.696	0.656	0.614

4.3.2 Insights into Bayesian Hyper-parameter Tuning (BHT) with Optuna

The implementation of Bayesian Hyper-parameter Tuning (BHT) using Optuna significantly optimized the performance of the tree-based credit risk acquisition model. Visualizations of the optimization process provided key insights into the exploration of the hyper-parameter space and the efficiency of Optuna's search methodology. The "Optimization History Plot" clearly demonstrated Optuna's capability for rapid convergence towards the optimal objective value, represented by the Area Under the Curve (AUC). A notable stabilization of the objective value was observed after approximately 50-100 trials, which suggested that a highly optimized region within the hyper-parameter space had been effectively identified by Optuna. This rapid convergence highlights the efficiency of BHT compared to exhaustive search methods.

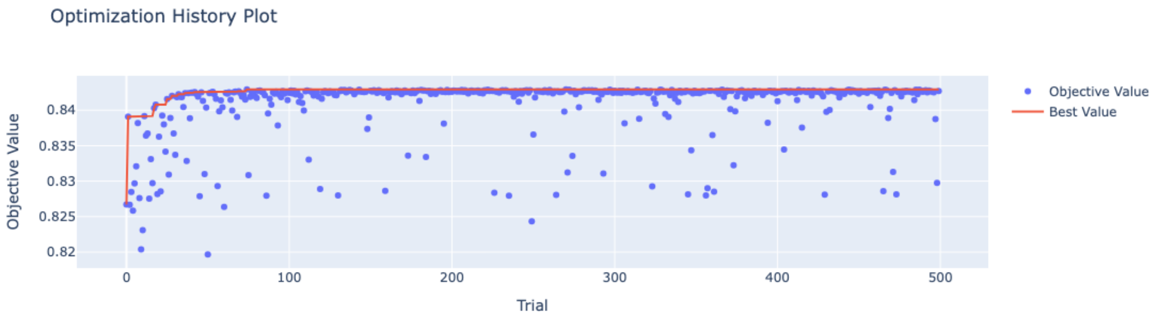

Figure 1: Optimization History Plot

The "Hyperparameter Importances" chart revealed that *learning_rate* and *num_round* (representing the number of boosting rounds) were identified as the most influential hyper-parameters. These two parameters collectively accounted for a substantial proportion of the total importance, specifically 72% and 20% respectively. This indicates that focusing optimization efforts on these parameters yields the greatest performance gains.

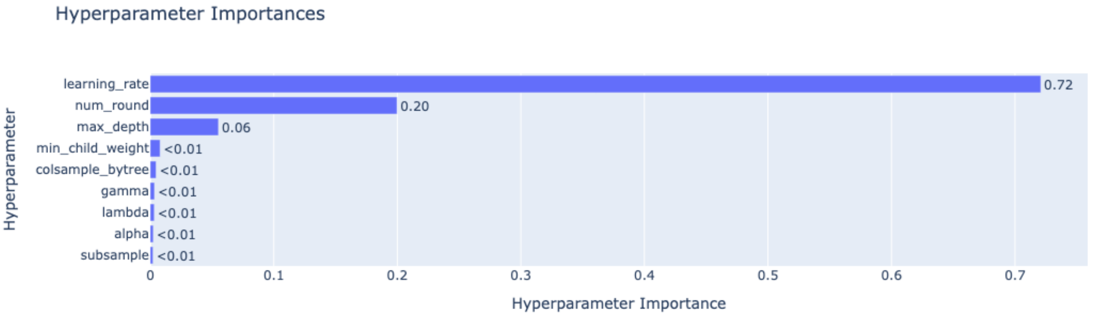

Figure 2: Hyperparameter Importances

The "Parallel Coordinate Plot" visually illustrates the intricate relationships between various hyper-parameters and the corresponding objective value. Darker lines within the plot represented higher objective values, indicating more optimal model configurations. Illustrative trends showed that the highest Area Under the Curve (AUC) values were observed to be clustered within the upper end of the scale, ranging approximately from 0.840 to 0.843. Optimal values for the *learning_rate* hyper-parameter consistently appeared to be concentrated around 0.05. Higher *num_round* values, specifically those approaching 350, were found to be strongly associated with improved model performance, particularly when appropriately paired with suitable learning rates. Values for *max_depth* around 5 were observed to contribute positively to higher objective values.

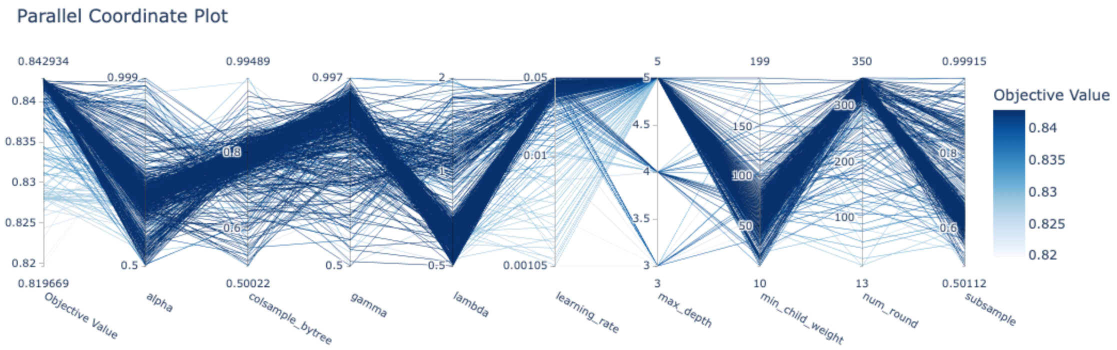

Figure 3: Parallel Coordinate Plot

Finally, for both *learning_rate* and *num_round*, the ”Slice Plots” exhibited a clear trend: as these parameters approached their respective optimal ranges, a consistent improvement in model performance was observed. These visualizations collectively confirm Optuna’s intelligent search strategy, efficiently navigating the complex hyper-parameter landscape to identify optimal configurations.

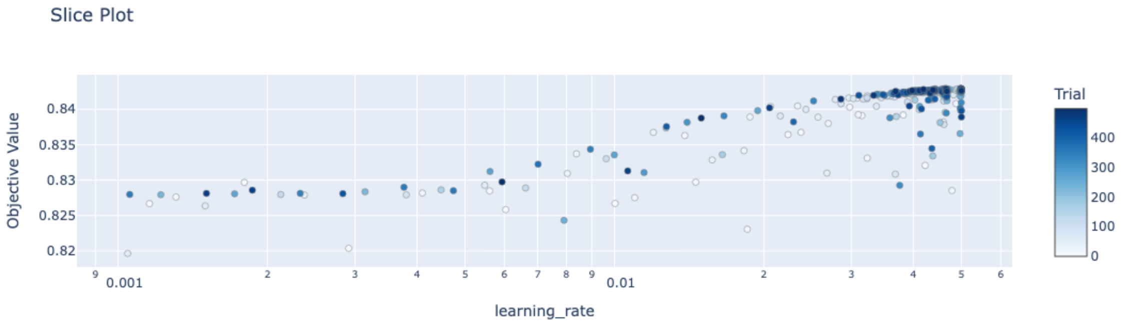

Figure 4: Slice Plot - Learning Rate

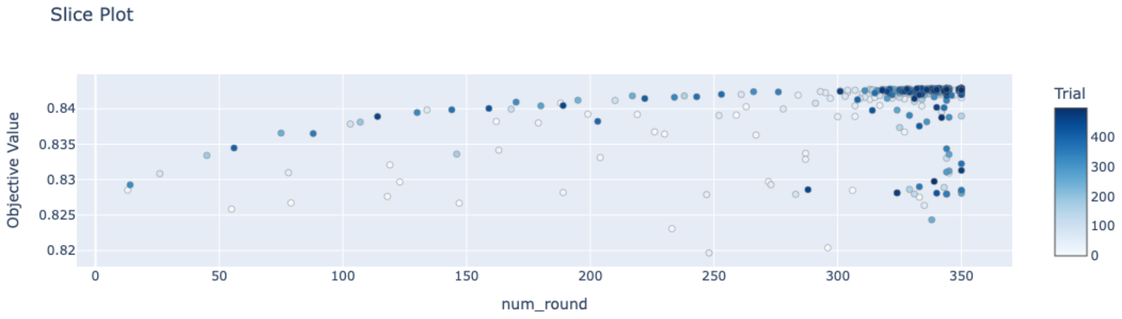

Figure 5: Slice Plot - Number of Rounds

4.3.3 Performance using NAFE + BHT

The combined impact of Noise-Based Feature Elimination (NAFE) and Bayesian Hyperparameter Tuning (BHT) with Optuna was comprehensively compared against the widely used combination of Recursive Feature Elimination (RFE) and Grid-Search Hyperparameter Tuning. To ensure a fair and equitable comparison, a similar number of hyperparameter tuning trials (approximately 500) were employed for both Optuna and Grid-Search across the identical range of hyperparameters. Model discriminatory power was assessed based on the Gini coefficient, and the rank-order scores were calibrated on a default target using a consistent calibration technique.

The modeling approach integrating NAFE and Optuna consistently generated higher discriminatory power when compared to the modeling approach utilizing RFE and Grid-Search. This was evidenced by higher Gini coefficient values observed on both the In-time Test dataset (an absolute increase of +0.2%) and the Out-of-Time Test dataset (an absolute increase of +0.5%). This table is crucial for demonstrating the synergistic effect of combining the proposed feature selection and hyper-parameter tuning methods, providing evidence that the NAFE + Optuna combination leads to models that generalize better and are more robust.

Table 2: Discriminatory power assessment - FS + HPT

Model	**FS Process**	**HPT Process**	**Gini (IT-Train)**	**Gini (IT-Test)**	**Gini (OOT-Test)**
XGBOOST	NAFS	Optuna	0.734	0.672	0.644
XGBOOST	RFE	GridSearchCV	0.875	0.670	0.639

The model developed with NAFE and Optuna demonstrated superior generalization power and enhanced robustness when compared to the model employing RFE and Grid-Search. This was substantiated by the observation that the NAFE + Optuna model exhibited a lower Gini coefficient on the In-time Train dataset but achieved higher Gini coefficients on both the In-time Test and Out-of-Time Test datasets. This pattern indicates that the RFE + Grid-Search approach tended to overfit the training data, resulting in weak generalization power, whereas the NAFE + Optuna approach proved to be more robust on unseen data. The accuracy of the acquisition Probability of Defaults (PDs) for both approaches was measured using the Brier score and compared against each other. The model with NAFS + Optuna had a higher Brier score on In-Time Train, but the same score on In-Time Test and a lower score on Out-Of-Time Test. This evidence shows that the proposed approach has better PD accuracy than the approach with RFE + Grid-search on unseen and out-of-time data. This table provides critical evidence for the calibration aspect of the model. A lower Brier score indicates better accuracy of PD estimates.

Table 3: PD Accuracy Assessment

Model	**FS Process**	**HPT Process**	**Brier Score (IT Train) Model**	**Brier Score (IT Test) Model**	**Brier Score (OOT Test) Model**	**Baseline IT Train**	**Baseline IT-Test**	**Baseline OOT Test**
XGBOOST	NAFS	Optuna	0.041	0.038	0.047	0.049	0.046	0.055
XGBOOST	RFE	GridSearchCV	0.039	0.038	0.049	0.049	0.046	0.055

4.4 Impact Assessment: Model Performance & Business Uplift

As part of the impact assessment, the proposed credit risk acquisition model leveraging NAFE and Bayesian HPT was compared against a vendor benchmark score. The model was compared against the benchmark on discriminatory power. Subsequently, the calibrated PD was used to compare the observed default rates using the proposed model with the observed default rates using the benchmark. Finally, the sales impact of the proposed model was compared against the sales impact of the benchmark.

The proposed model consistently outperforms the benchmark across all tested sample types, demonstrating a consistent uplift in Gini. The uplift on out-of-time data is consistently around 10%. This indicates that the developed model is more effective at discriminating between good and bad credit risk applications. Outperforming a vendor benchmark signifies a competitive advantage and superior risk-return profiles.

Table 4: Discriminatory power comparison (Gini)

	Sample	**Dataset**	**Overall**	**Overall with eligibility**	**Rev Sample**	**Rev Sample with eligibility**	**Internal Features**	**Internal Features with eligibility**
XGBoost	IT Train		70.3% (69.9%, 70.7%)	70.2% (69.6%, 70.7%)	67.8% (66.9%, 68.7%)	67.8% (66.6%, 69.1%)	69.3% (67.3%, 71.3%)	69.3% (67.3%, 71.3%)
	IT Test		67.9% (67.2%, 68.7%)	70.2% (69.2%, 71.2%)	69.0% (67.2%, 70.8%)	69.4% (67.0%, 71.8%)	69.3% (66.3%, 72.3%)	71.4% (67.6%, 75.2%)
	OOT Test		68.4% (67.0%, 68.4%)	68.4% (67.5%, 69.3%)	64.8% (62.0%, 65.2%)	64.8% (62.8%, 66.7%)	64.8% (62.6%, 66.9%)	65.4% (62.7%, 68.1%)
Benchmark	IT Train		63.6% (63.2%, 64.0%)	62.7% (62.1%, 63.5%)	61.5% (60.5%, 62.5%)	60.5% (59.2%, 61.8%)	63.9% (62.3%, 65.4%)	62.3% (60.2%, 64.4%)
	IT Test		62.9% (63.7%, 64.5%)	63.0% (61.9%, 64.1%)	62.7% (60.8%, 64.6%)	62.7% (60.1%, 65.2%)	63.5% (60.3%, 66.6%)	66.0% (61.9%, 70.1%)
	OOT Test		62.0% (61.3%, 62.8%)	61.8% (60.8%, 62.7%)	58.3% (56.7%, 60.0%)	58.7% (56.6%, 60.7%)	59.5% (57.2%, 61.7%)	59.6% (56.7%, 62.4%)
Uplift (%)	IT Train		10.5%	11.9%	10.1%	12.2%	9.4%	11.2%
	IT Test		8.0%	11.4%	10.1%	10.7%	9.1%	8.2%
	OOT Test		9.1%	10.6%	9.1%	10.3%	8.9%	9.7%

Regarding default rate comparison (Observed Default Rates - ODRs), the proposed model, with a similar acceptance rate, maintains lower default rates. Among those accepted by both models, XGBoost has a 1.6% ODR, compared to the benchmark's 2%. This indicates efficient risk calibration, meaning the model is better at identifying and managing risk within its accepted population. Furthermore, the proposed model demonstrates optimized limit allocation and increased sales potential. It distributes a higher proportion of customers to lower risk-grades with higher maximum limits compared to the benchmark, which directly drives the increased sales potential. This ultimately generates an uplift in sales of approximately 10% compared to the benchmark. These findings collectively highlight the substantial benefits of implementing the proposed NAFE and BHT methodology in a real-world credit risk acquisition scenario.

Table 5: Sample - OOT Recent: XGBoost vs Benchmark Sales Uplift

Risk Grade	**Uplift (%)**
A	+52%
B	-29%
C	-5%
D	-29%
Total Uplift	**+10%**

5 Conclusion

This study successfully demonstrated the significant advantages of integrating Noise-Augmented Feature Selection (NAFE) with Bayesian Hyper-parameter Tuning (BHT) for enhancing credit risk acquisition scoring using gradient-boosted models.

5.1 Summary of Key Achievements

The combined approach of Noise-Augmented Feature Selection and Bayesian Hyper-parameter tuning resulted in higher discriminatory power on unseen out-of-time data compared to the widely used Recursive Feature Elimination + Grid-Search Hyperparameter tuning approach. This indicates that the proposed approach leads to a more robust model with superior generalization capabilities. Computational efficiency was a notable achievement, with feature selection using NAFE converging to the same number of input features in significantly less time (approximately 75% faster) compared to RFE. The calibrated acquisition PDs produced using the proposed approach exhibited better accuracy (lower Brier scores) on unseen out-of-time data, leading to more accurate and robust PD estimates. Ultimately, all these benefits translated into a tangible business impact, specifically a 10% uplift in potential sales compared to the benchmark. This demonstrates the direct financial value derived from the enhanced modeling approach.

5.2 Challenges

While the proposed methodology offers significant advantages, certain challenges were identified that warrant further exploration.

Table 6: Challenges

Focus Area	**Key Challenge**	**Alternate / Complementary Options**
Noise-Augmented Feature Selection	Selecting the right number and new distribution of noise features could be subjective and compute-intensive on large datasets.	Boruta, Permutation Importance
Bayesian Hyper-parameter Tuning	Still demanding for extremely expensive objective functions or vast, poorly understood search spaces, leading to potential overfitting.	Hyperband, Population-Based Training

5.3 Future Steps

Building upon the success of this study, several avenues for future research and development are identified:

- The NAFE process can be further refined by focusing on two key aspects: noise-distribution matching and noise-quantity tuning. Noise-distribution matching would involve generating synthetic noise that mirrors the statistical properties of each feature type (e.g., Gaussian for continuous variables, shuffled categories for one-hot encoded variables). Noise-quantity tuning would entail systematically varying the number of noise variables to find the optimal balance between selection robustness and runtime efficiency.

- Beyond tree ensembles, the NAFE methodology can be extended to Marginal Information Value (MIV) Logistic Regression, broadening its applicability to other widely used modeling techniques in credit risk.
- A crucial future step involves automating NAFE and BHT within comprehensive model pipelines. Embedding both feature-selection and hyper-parameter optimization modules directly into the end-to-end modeling workflow would ensure that every model upgrade or new model benefits from consistent, scalable optimization with minimal manual intervention.

References

[1] Gharoun, H., Yazdanjoe, N., Khorshidi, M. S., & Gandomi, A. H. (2023). Noise-Augmented Boruta: The Neural Network Perturbation Infusion with Boruta Feature Selection. arXiv preprint arXiv:2309.09694.

[2] Jenatton, R., Popov, P., & Bach, F. (2017, August). Bayesian optimization with tree-structured dependencies. Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR 70:1633–1642.

[3] Jemai, J., & Zarrad, A. (2023). Feature Selection Engineering for Credit Risk Assessment in Retail Banking. Information, 14(3), 200.

[4] Koç, O., Uğur, O., & Kestel, A. S. (2023). The Impact of Feature Selection and Transformation on Machine Learning Methods in Determining the Credit Scoring. arXiv preprint arXiv:2303.05427.

[5] Patel, S., & Brown, M. (2021). Application of AI in Credit Evaluation: Challenges and Opportunities. Proceedings of the International Conference on Artificial Intelligence Applications (pp. 142–156).

[6] Souadda, L. I., Halitim, A. R., Benilles, B., Oliveira, J. M., & Ramos, P. (2025). Optimizing Credit Risk Prediction for Peer-to-Peer Lending Using Machine Learning.

[7] Lim, Z. Y., Pang, Y. H., Kamarudin, K. Z. B., Ooi, S. Y., & San Hiew, F. (2024). Bayesian optimization driven strategy for detecting credit card fraud with Extremely Randomized Trees. MethodsX, 13, 103055.

[8] Wong, E., & Smith, M. (2019). A Comparative Analysis of Machine Learning Techniques for Credit Risk Assessment. Proceedings of the International Conference on Artificial Intelligence and Applications (pp. 76–89).