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

University of Edinburgh Credit Scoring and Credit Control Conference XIX (Aug 2025)

Arijit Ganguly
Revolut Group Holdings Ltd, Bangalore, India
Global Credit Management – Data Science

August 27, 2025

Revolut

Context

Literature review

Methodology

Business Case & Results

Conclusions

References

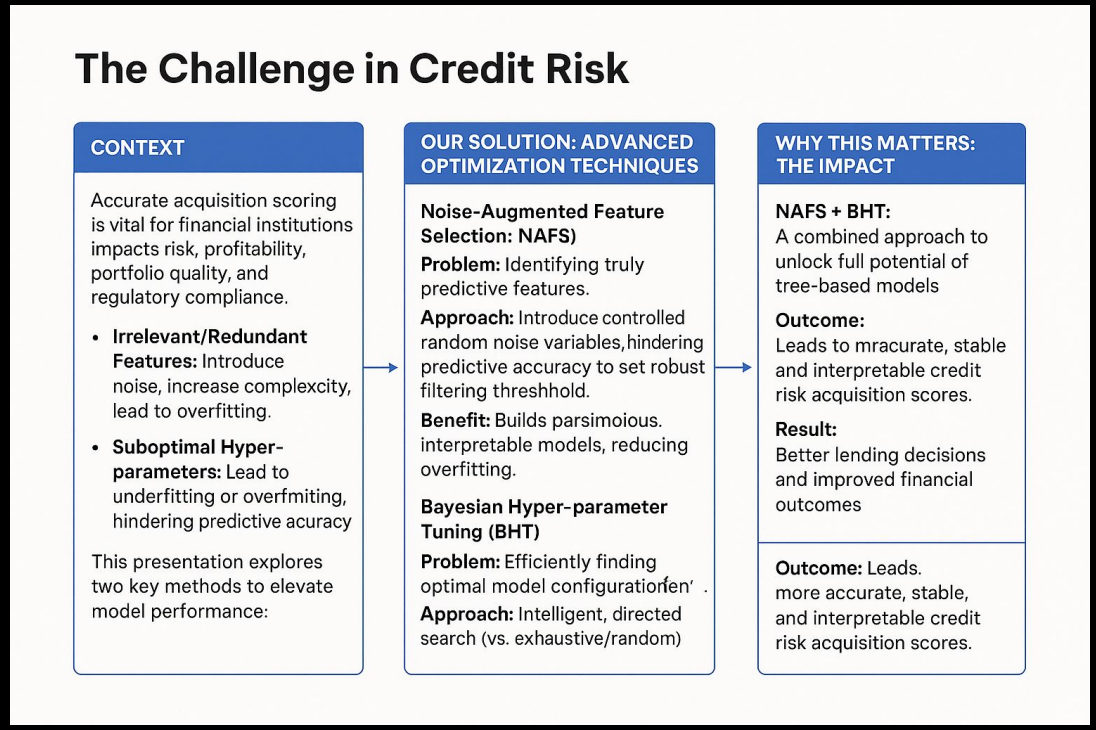
Context

The Challenge in Credit Risk

- Accurate acquisition scoring is vital for financial institutions: impacts risk, profitability, portfolio quality, and regulatory compliance.
- Tree-based models (XGBoost, LightGBM) are powerful but face challenges:
  - **Irrelevant/Redundant Features**: Introduce noise, increase complexity, lead to overfitting.
  - **Suboptimal Hyper-parameters**: Lead to underfitting or overfitting, hindering predictive accuracy.

Our Solution: Advanced Optimization Techniques

- **Noise-Augmented Feature Elimination (NAFE)**
  - **Problem**: Identifying truly predictive features.
  - **Approach**: Introduce controlled random noise variables.
  - **Mechanism**: Compare real feature importance against noise feature importance to set a robust filtering threshold.
  - **Benefit**: Builds parsimonious, interpretable models, reducing overfitting.
- **Bayesian Hyper-parameter Tuning (BHT)**
  - **Problem**: Efficiently finding optimal model configurations.
  - **Approach**: Intelligent, directed search (vs. exhaustive/random).
  - **Mechanism**: Builds a probabilistic model of the objective function to suggest optimal next parameters.
  - **Benefit**: Faster convergence to optimal settings, significantly reducing computational cost.



Why This Matters: The Impact

- NAFE + BHT: A combined approach to unlock full potential of tree-based models.
- Outcome: Leads to more accurate, stable, and interpretable credit risk acquisition scores.
- Result: Better lending decisions and improved financial outcomes.

Revolut

Context

Literature review

Methodology

Business Case & Results

Conclusions

References

Revolut

Literature Review

Impact of Feature Selection (FS)

- **Addresses High Dimensionality & Redundancy:** Credit risk models often face noise and irrelevant features leading to overfitting.
  - **Wong and Smith (2019)**: Emphasize feature engineering's critical role in optimizing ML models for credit risk.
  - **Patel and Brown (2021)**: Discuss interpretability challenges alleviated by feature selection.
- **Improves Interpretability & Performance**: Filtering features leads to more parsimonious and understandable models.
  - **MDPI (2023)**: Highlights various FS methods (univariate, RFE, feature importance, information value) for improving accuracy by reducing noise.
  - **arXiv (2023)**: Demonstrates positive contribution of FS on ML methods in credit scoring.
- **Noise-Augmented Feature Elimination (NAFE):** Reinforces filtering by establishing statistical thresholds against noise variables to identify truly informative features.
  - **Gharoun et al. (2023)**: Introduced "Noise-Augmented Boruta," an innovative approach incorporating noise into shadow variables for enhanced, accurate feature selection.

Impact of Bayesian Hyper-parameter Tuning (BHT)

- **Optimizes Tree-based Models (XGBoost, LightGBM)**: Addresses the challenge of tuning numerous hyper-parameters.
- **Enhances Efficiency & Performance**: Intelligently explores the hyper-parameter space, leading to faster convergence to optimal solutions.
  - **Preprints.org (2025)**: States hyperparameter optimization is essential for maximizing predictive performance and computational efficiency in credit risk modeling. Shows Optuna (Bayesian framework) significantly outperforms Grid/Random Search in speed for XGBoost/LightGBM.
  - **ResearchGate (2024)**: Applies Bayesian optimization (TPE) to tree-based models for fraud detection, with direct applicability to credit risk.
  - **Jenatton et al. (2017):** Discusses leveraging dependency structures in optimization domains for more efficient search, relevant for complex hyperparameter spaces.

Context

Literature review

Methodology

Business Case & Results

Conclusions

References

Revolut

Methodology: Noise Augmented Feature Elimination (NAFE)

I. Data Preparation & Noise Augmentation

- **Data Loading:** Load historical credit risk data, segmenting into training and validation sets. This includes relevant features and the target variable (e.g., default status).
- **Random Noise Feature Introduction:** Augment the dataset with a predefined number of synthetic random noise features. These features are generated with no inherent predictive power.

II. Noise-Augmented Feature Filtering

- **Initial Model Training:** Train an initial XGBoost or LightGBM model on the augmented training data (including real and noise features) using a predefined set of hyperparameters and early stopping.
- **Feature Importance Calculation:** Calculate feature importance for all features (real and noise). This can be based on:
  - **Gain**: The average gain of splits where the feature is used.
  - **SHAP (SHapley Additive exPlanations)**: Mean absolute SHAP values, providing a measure of feature contribution to model output.
- **Zero Importance Removal:** Remove all features with a feature importance of 0.
- **Noise Thresholding:** Identify features whose importance is less than or equal to the maximum importance observed among the random noise features. These features are considered non-informative and are marked for removal.

III. Correlation-Based Feature Elimination (Optional)

- **Correlation Analysis:** For the remaining non-categorical features, compute the pairwise correlation matrix.
- **Redundancy Handling**: If a pair of features exhibits a correlation above a defined threshold (e.g., 0.5), the feature with the lower importance (as determined in Step II.2) from that pair is identified for removal. This step helps reduce multicollinearity and further streamlines the feature set.

Methodology: Bayesian Hyperparameter Tuning with Optuna

I. Objective Function Definition

- **Model Agnostic Objective:** An objective function (objective_xgb or objective_lgbm) is defined to optimize either XGBoost or LightGBM models.
- **Hyperparameter Search Space**: Within this function, Optuna's trial.suggest_ methods are used to define the search range for key hyper-parameters (e.g., learning_rate, max_depth, subsample, colsample_bytree, regularization terms). This allows Optuna to intelligently explore the parameter space.
- **Cross-Validation:** To ensure robust evaluation, Stratified K-Fold Cross-Validation (N_FOLDS) is performed for each trial. The model is trained on multiple folds, and its performance is averaged.
- **Early Stopping**: Each model training within a trial incorporates early stopping based on the Area Under the Curve (AUC) on the validation set, preventing overfitting during the individual model training phase.
- **Performance Metric**: The objective function returns the average validation AUC across all folds, which Optuna aims to maximize.
- **Pruning Heuristics:** To accelerate the search and avoid unproductive trials, two pruning criteria are implemented:
  - **Overfitting Index:** Trials exhibiting an excessive difference between training AUC and validation AUC (indicating overfitting) are pruned.
  - **Coefficient of Variation (CV) of AUC**: Trials with high standard deviation of AUC across folds (indicating instability) are pruned.

II. Optuna Study Execution

- **Study Creation**: An Optuna study object is initialized with a direction="maximize" setting, as the goal is to maximize the validation AUC.
- **Optimization**: The study.optimize method is called, passing the objective function and the desired number of n_trials. Optuna intelligently selects hyper-parameter combinations for each trial based on the results of previous trials, using a Tree-structured Parzen Estimator (TPE) algorithm. Parallelization (n_jobs=NCPUS) is used to speed up the process.

III. Training with Best Hyper-parameters

- **Best Trial Extraction**: After the Optuna study completes, the optimal hyper-parameters are retrieved from the study.best_trial.
- **Final Model Training**: A final XGBoost or LightGBM model is trained on the entire training dataset (or on the combined train and test sets, as indicated in the code flow) using these best-found hyper-parameters and the feature set identified from the previous feature selection step.
- **Performance Evaluation**: The final model's performance (AUC and confidence intervals) is re-evaluated on both the training and a separate validation/test set to confirm generalization capabilities.

Context
Literature review
Methodology
Business Case & Results
Conclusions
References

Business Case PoC: Credit Card Acquisition Model

Objective: To enhance credit risk acquisition scoring for credit card applicants using advanced machine learning techniques, focusing on tree-based models.

Data Overview (Retro Sample)

- **Overall Sample:** 2 million records (UPLs, CCs, Overdrafts), including ~300k Revolut users.
- **Credit Card (CC) Sample**: 900k records, of which ~140k are Revolut users, with ~65k having opened a facility after joining Revolut.
- **Performance Window:** Facilities granted Jan 2020 - Jul 2022, with performance tracked up to 48 months.
- **Model Scope:** Models primarily built on the Credit Card (CC) Sample.

Methodology

- **Training Datasets Scenarios:**
  - All users (Revolut + Non-Revolut): With/Without eligibility checks (Bureau features).
  - Revolut users: With/Without eligibility checks (Bureau, Email/Device features).
  - Revolut users with internal history: With/Without eligibility checks (Bureau, Email/Device, Transaction features).
- **Sample Split:**
  - Train/Test (80/20): Data from August 2020 to March 2022*
  - Out-of-Time (OOT) Recent: Data from April 2022 to July 2022 for final validation.
- **Algorithms: XGBoost, LightGBM.**

Key Enhancements

- **Feature Selection:**
  - **Noise-Based Feature Selection**: Introduced 50 noise features; features with SHAP importance noise features were eliminated.
- **Hyper-parameter Tuning: Optuna**: Employed Bayesian optimization for efficient and robust hyper-parameter tuning of models.

Model Selection & Evaluation

Metrics: Performance evaluated using the Gini score on Train, Test, and Out-of-Time (OOT) samples and benchmarked against a vendor score.

* In-time Train includes origination cohort overlapping with Covid period partially due to limitations of retro data. However model performance is benchmarked on multiple out-of-time periods including pre-covid, covid and most-recent. Proposed solution outperformed benchmark on all oot periods.

Results: NAFE vs standard RFE

Compared Noise Augmented Feature Elimination (NAFE) against standard Recursive Feature Elimination (RFE) based on model discriminatory power and computational efficiency. To ensure a like-for-like comparison, steps were taken to ensure both models have similar number of features at the end of the feature selection process. The feature selection was compared for XgBoost and LightGBM algorithms on a pre-defined same of hyperparameters.

Key Findings

- **Higher discriminatory power** - The feature selection process employing NAFE results in a higher Gini value compared to the feature selection process using RFE.
  - The performance uplift is observed for both XgBoost and LightGBM.
  - For XgBoost, the model with NAFE outperforms the model with RFE on In-time Train (**+0.9% abs.**) and Out-of-Time Test (**+0.3% abs**). The overall absolute uplift is **+0.4%**
  - For LightGBM, the model with NAFE outperforms the model with RFE on In-time Train (**+5.4% abs.**) and Out-of-Time Test (**+1.0%**). The overall uplift is **+2.1%**
- **Lower computation time** - The NAFE process converges to similar number of features in a much less time compared to RFE. For XgBoost and LightGBM, the average reduction in feature selection time is **~75%**

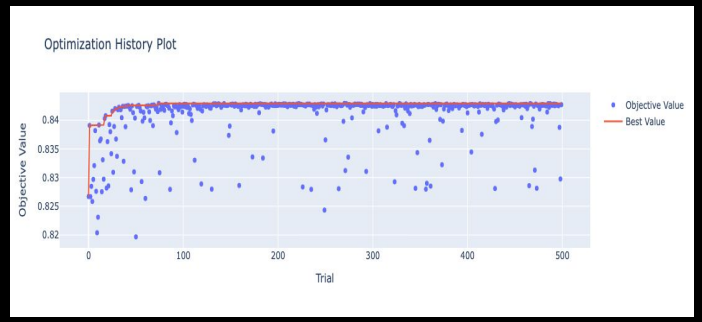
Discriminatory power assessment - Feature Selection					
Algorithm	**FS Process**	**Time to completion (mins)**	**Gini (In-Time Train)**	**Gini (In-Time Test)**	**Gini (Out-Of-Time Test)**
XGBOOST	NAFE	5.5	0.767	0.669	0.639
XGBOOST	RFE	23.0	0.758	0.669	0.636
LGBM	NAFE	5.0	0.750	0.655	0.624
LGBM	RFE	21.1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

Results: Insights into Bayesian Hyperparameter Tuning (BHT) with Optuna

Our implementation of Bayesian HPT using Optuna has significantly optimized the performance of our tree-based credit risk acquisition model. The visualization of the optimization process provides key insights into the hyper-parameter space and the efficiency of Optuna's search.

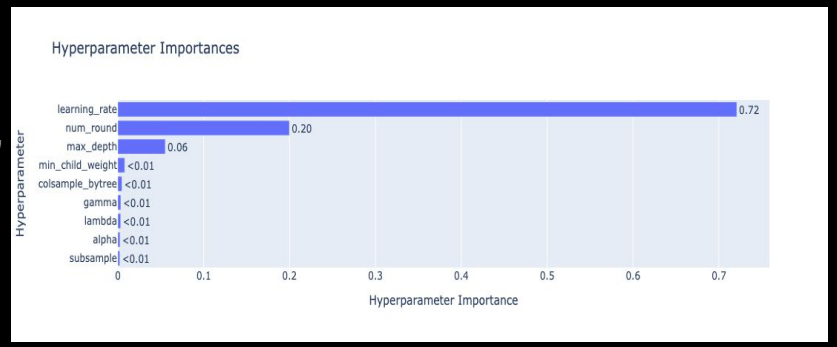
I. Optimization History

- **Rapid Convergence**: The "Optimization History Plot" demonstrates Optuna's ability to quickly converge towards the optimal objective value (AUC).
- **Stability**: After approximately 50-100 trials, the objective value stabilizes, suggesting that Optuna has effectively identified a highly optimized region.



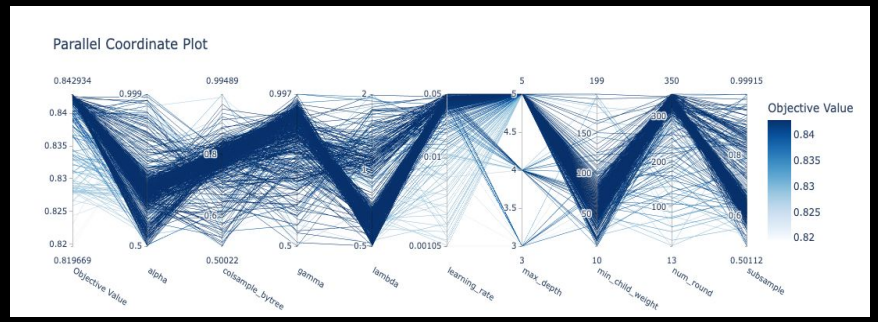
II. Hyperparameter Importances

- **Key Drivers**: The "Hyperparameter Importances" chart reveals that `learning_rate` and `num_round` are by far the most influential hyper-parameters, accounting for 72% and 20% of the importance respectively.



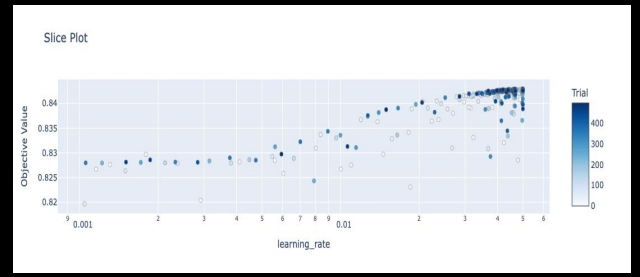
III. Parallel Coordinate Plot

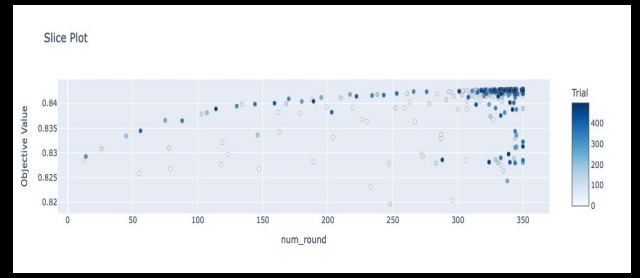
- Relationships between hyper-parameters and the objective value. Darker lines represent higher objective values.
- **Illustrative Trends**:
  - **Objective Value**: The highest AUCs are clustered around the upper end of the scale (e.g., 0.840 to 0.843).
  - **`learning_rate`**: Optimal `learning_rate` values appear to be in around 0.05.
  - **`num_round`**: Higher `num_round` values (towards 350) are associated with better performance, especially when paired with appropriate learning rates.
  - **`max_depth`**: Values around 5 seem to contribute to higher objective values.



IV. Slice Plots

- **Individual Parameter Impact**: For `learning_rate` and `num_round`, a clear trend shows that as these parameters approach their optimal ranges, the model performance improves.





Results: Enhanced performance using NAFE + BHT

Compared the combined impact of **Noise Based Feature Elimination** (**NAFE**) and B**ayesian Hyperparameter tuning (BHT) with Optuna** against the widely used **Recursive Feature Elimination** (**RFE**) combined with **Grid-Search** Hyperparameter tuning. For an even comparison, similar number of hyperparameter tuning trails (~500) was employed for Optuna and Grid-Search over the same range of hyperparameters. The model discriminatory is compared based on Gini. The rank-order scores were calibrated on a default@12M target using the same calibration technique.

Key Findings

- **Discriminatory power** - The modeling approach with NAFE + Optuna generates a higher discriminatory power compared to the modeling approach with RFE + Grid-Search.
  - This is evidence by higher Gini values on In-time Test (**+0.2% abs**) and Out-of-time Test (**+0.5% abs**).
- **Robustness** - The model with NAFE + Optuna has a better generalisation power and more robustness compared to the model with RFE + Grid-Search.
  - This is evidence by the fact that NAFE + Optuna has a lower Gini on In-time Train but higher Gini on IT Test and OOT Test. This shows that RFE + Grid-search overfits and has weak generalisation power, while NAFE + Optuna is more robust on unseen data.
- **PD Accuracy** - The accuracy of the acquisition PDs for both approaches is measured using Brier score and compared against each other.
  - The model with NAFS + Optuna has a higher Brier score on IT Train, but same score on IT Test and lower score on OOT Test.
  - This evidences that the proposed approach has a **better PD accuracy** than the approach with RFE + Grid-search on **unseen and out-of-time data**.

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

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

Impact Assessment: Model Performance & Business Uplift

Proposed credit risk acquisition model, leveraging NAFE and Bayesian HPT, demonstrates significant uplift in discriminatory power and tangible business benefits when compared against the benchmark score.

	Sample	Dataset					
		Overall	**Overall with eligibility**	**Rev Sample**	**Rev Sample with eligibility**	**Internal Features**	**Internal Features with eligibility**
XGBoost Gini	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 Gini	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%)
Gini 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%**

I. Discriminatory Power

- **Consistent Uplift in Gini**: Proposed model consistently outperforms the benchmark across all tested sample types. The uplift on **out-of-time** is consistently around **10%**. This indicates that our model is more effective at discriminating between good and bad credit risks.

II. Calibrated PD Scores: lower default rates

- **Default Rate Comparison (ODRs):** Proposed model, with a similar acceptance rate, maintains lower default rates. Among those accepted by both models, XGBoost has a **1.6%** ODR (Observed Default Rates), compared to benchmark's **2%**. This indicates efficient risk calibration.

Sample - OOT Recent	
XGBoost v/s Benchmark - Sales uplift	
Risk Grade	**Uplift (%)**
A	+52%
B	-29%
C	-5%
D	-29%
Total Uplift	**+10%**

III. Impact on Sales

- **Optimized Limit Allocation & Increased Sales Potential**: Proposed model distributes a higher proportion of customers to lower-risk grades with higher maximum limit compared to the benchmark, which drives the increased sales potential. This generates a **Sales Uplift of +10%** compared to the benchmark.

Revolut

Context
Literature review
Methodology
Business Case & Results
Conclusions
References

Conclusion: Enhancing Credit Risk Acquisition Scoring

Summary of Key Achievements

- **Higher Discriminatory Power:** The combined approach of Noise Augmented Feature Selection and Bayesian Hyper-parameter tuning resulted in a higher discriminatory power on unseen out-of-time data compared to the widely used approach - Recursive Feature Elimination + Grid-Search Hyperparameter tuning. This indicates the proposed approach led to a more robust model with better generalization.
- **Computational Efficiency**: The feature selection using NAFE converged to same number of input features in a much less time (**-75%)** compared RFE.
- **PD Accuracy**: The calibrated acquisition PD produced using the proposed approach had a better accuracy (lower Brier scores) on unseen out-of-time data leading to more accurate and robust PD estimates.
- **Tangible Business Impact**: All the above benefits translated into a **+10% uplift** in potential sales compared to the benchmark.

Challenges

Focus Area	Key Challenge	Alternate / Complementary Options
Noise-Augmented Feature Selection	Selecting the right number and distribution of noise features can 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

Future Steps

- **Refine the NAFE Process**
  - *Noise-Distribution Matching*: Generating synthetic noise that mirrors the statistical properties of each feature type (e.g., Gaussian for continuous variables, shuffled categories for one-hots).
  - *Noise-Quantity Tuning*: Systematically varying the number of noise variables to find the sweet spot between selection robustness and runtime.
- **Extend NAFE to MIV Logistic Regression**
  - Applying noise-augmented feature elimination coupling it with Marginal Information Value (MIV) in Logistic Regression, broadening applicability beyond tree ensembles.
- **Automate NAFE and BHT in Model Pipelines**
  - Embedding both feature-selection and hyper-parameter optimization modules directly into the end-to-end modeling workflow, ensuring every upgrade or new model benefits from consistent, scalable optimization with minimal manual intervention.

References

- 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.
- 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.
- Jemai, J., & Zarrad, A. (2023). Feature Selection Engineering for Credit Risk Assessment in Retail Banking. Information, 14(3), 200.
- Ko¸c, O., U˘gur, 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.
- 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).
- 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.
- 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.
- 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).

Thank you

Revolut