UNIVERSITAS
GADJAH MADA
Evaluasi Kuantitatif Fidelity Model Lokal Surrogate pada Stacking Ensemble untuk Penilaian Kredit
Rida Perwitasari, Afiahayati, S.Kom., M.Cs., Ph.D
Universitas Gadjah Mada, 2025 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

EVALUASI KUANTITATIF FIDELITY MODEL LOKAL SURROGATE PADA STACKING ENSEMBLE UNTUK PENILAIAN KREDIT

Oleh:

Rida Perwitasari
23/512302/PPA/06504

Explainable Artificial Intelligence (XAI) semakin menjadi elemen penting dalam aplikasi pendukung keputusan yang menuntut akurasi prediktif sekaligus transparansi. Dalam bidang seperti penilaian kredit, interpretabilitas berperan krusial untuk mendorong keadilan, akuntabilitas, serta kepatuhan terhadap regulasi. Meskipun demikian, bukti empiris yang menunjukkan sejauh mana penjelasan XAI benar-benar merefleksikan mekanisme pengambilan keputusan model yang kompleks masih terbatas. Penelitian ini mengkaji sejauh mana model surrogate dalam kerangka Local Interpretable Model-agnostic Explanations (LIME) mampu mengaproksimasi perilaku model ensemble stacking yang digunakan untuk evaluasi risiko kredit. Model *stacking* tersebut mengombinasikan beberapa pengklasifikasi dasar dengan sebuah *meta-learner* untuk menghasilkan prediksi akhir yang lebih akurat. Kualitas penjelasan dinilai berdasarkan seberapa dekat model *surrogate* merepresentasikan pengambilan keputusan model *stacking*.

Dengan menggunakan dataset Lending Club, dibangun sebuah model *ensemble stacking* dan diuji tiga model *surrogate*, yaitu *Decision Tree, Logistic Regression*, dan Naïve Bayes, berdasarkan metrik *fidelity* lokal seperti log-*loss*, *accuracy*, *F1-score*, dan korelasi Pearson. Evaluasi dilakukan pada *subset* data uji sebesar 5% yang dipilih secara terstratifikasi, kemudian dilanjutkan dengan validasi statistik menggunakan uji Anderson–Darling, Friedman, dan Wilcoxon. Hasil penelitian menunjukkan bahwa *Logistic Regression* mencapai tingkat *fidelity* tertinggi, yang mengindikasikan perilaku lokal yang lebih konsisten dan mudah diinterpretasikan dalam konteks aplikasi penilaian kredit.

Kata Kunci: penilaian kredit, *ensemble learning, stacking, explainable, surrogate model*, XAI, LIME

UNIVERSITAS
GADJAH MADA
Evaluasi Kuantitatif Fidelity Model Lokal Surrogate pada Stacking Ensemble untuk Penilaian Kredit
Rida Perwitasari, Afiahayati, S.Kom., M.Cs., Ph.D
Universitas Gadjah Mada, 2025 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

QUANTITATIVE EVALUATION OF FIDELITY IN LOCAL SURROGATE MODELS FOR STACKING ENSEMBLES IN CREDIT SCORING

By:

Rida Perwitasari
23/512302/PPA/06504

Explainable Artificial Intelligence (XAI) has become an increasingly important component in decision-support applications that require both predictive accuracy and transparency. In domains such as credit scoring, interpretability plays a critical role in promoting fairness, accountability, and regulatory compliance. Nevertheless, empirical evidence demonstrating whether XAI explanations truly reflect the decision-making mechanisms of complex models remains limited. This study investigates the extent to which surrogate models within the Local Interpretable Model-agnostic Explanations (LIME) framework can approximate the behavior of a stacking ensemble used for credit risk evaluation. The stacking model combines multiple base classifiers with a meta-learner to produce more accurate final predictions. The quality of explanations is assessed based on how closely the surrogate models represent the decision-making behavior of the stacking model.

Using the Lending Club dataset, a stacking ensemble was constructed and three surrogate models—Decision Tree, Logistic Regression, and Naïve Bayes—were evaluated using local fidelity metrics, including log-loss, accuracy, F1-score, and Pearson correlation. The evaluation was conducted on stratified subsets comprising 5% of the test data, followed by statistical validation using the Anderson–Darling, Friedman, and Wilcoxon tests. The results indicate that Logistic Regression achieved the highest level of fidelity, suggesting more consistent and interpretable local behavior in the context of credit-scoring applications.

Keywords: Credit scoring, ensemble learning, stacking, explainable, surrogate model, **XAI, LIME**