Report on fine tuning a LightGBM model using hyperparameter search of scikit-learn's GridSearch

Richard Hajek

May 2025

Contents

1 Fine-Tuning LightGBM for Credit Card Fraud Detection

1.1 Objective

The objective of this work is to reproduce and improve the credit card fraud detection model presented in the Kaggle notebook by gpreda. The model is based on LightGBM and targets the highly imbalanced binary classification problem inherent to fraud detection. This report documents the reproduction, tuning, and evaluation process.

1.2 Dataset

- Source: Kaggle (mlg-ulb/creditcardfraud)
- Content: 284,807 transactions with 492 fraud cases (~0.17%)
- Features: 30 input features including anonymized PCA components (V1–V28), `Time`, and `Amount`
- Target: `Class` (0 = non-fraud, 1 = fraud)

1.3 Initial Model Reproduction

- Library: LightGBM
- Split:
  - Training, validation, and test sets using `train_test_split`
- Hyperparameters:
  - `learning_rate`: 0.05
  - `num_leaves`: 7
  - `max_depth`: 4
  - `scale_pos_weight`: 150 (to counteract class imbalance)
- Evaluation Metric: AUC
- Training: Early stopping with patience of 100 rounds (`2 * EARLY_STOP`)

- Result: Model trained with LightGBM's native API using `lgb.Dataset`

1.4 Fine-Tuning Approach

- Converted to `LGBMClassifier` API for compatibility with `GridSearchCV`
- Same base hyperparameters used as in reproduction
- Fit using `fit()` on training set with early stopping on val set
- Warnings from `DataConversionWarning` suppressed

1.5 Hyperparameter Optimization

- Method: Grid Search Cross-Validation (3-fold)
    - Parallelization: Utilized all cores (`n_jobs=-1`)
    - Parameters Tuned:
        * `learning_rate`: [0.05, 0.01, 0.005]
        * `num_leaves`: [7, 5]
        * `max_depth`: [5, 4, 3]
        * `min_child_samples`: [100, 200]
        * `max_bin`: [100, 200]
        * `subsample`: [0.9, 0.8]
        * `colsample_bytree`: [0.7]
    - Other Parameters: Held constant due to dataset characteristics (e.g., `scale_pos_weight=150`)
- Class imbalance addressed via `scale_pos_weight` and stratified splits
- Early stopping and validation sets incorporated for generalization control
- Grid search improved model precision through deeper hyperparameter exploration
- Custom context manager effectively handled logging noise

1.5.1 Noise Suppression

`stderr` output suppressed using `contextlib` and `os.devnull` to reduce verbosity during grid search

1.6 Evaluation

- Best Estimator: Obtained from `GridSearchCV`
- Test Set Performance: Evaluated using `score()` method on `X_test`, `y_test`
- Metric: AUC

1.7 Conclusion

The original model by gpreda was effectively reproduced and fine-tuned. Grid search with parallelization provided an optimized LightGBM model with better hyperparameters. Evaluation on the test set confirms the operational viability of the tuned classifier in high-imbalance fraud detection scenarios.