Model Development Document (MDD)

Model Name: Elder Financial Abuse (EFA) Classifier

Version: v1.0

Date: 2025-04-11

Developer(s): NLP Model Development Team

Model Owner: Compliance Analytics Group

Business Line: Fraud and Financial Crime Risk Management

# 1. Executive Summary

This model is developed to automatically classify bankers' free-text notes from customer interactions as either containing Elder Financial Abuse (EFA) indicators or not.   
It is intended to serve as a screening mechanism to assist compliance analysts in identifying possible EFA cases from narrative documentation. The model uses LightGBM as a binary classifier trained on labeled data reviewed and annotated by EFA domain experts.

# 2. Conceptual Soundness Evaluation

## 2.1 Data Quality and Suitability

Data Sources:  
- The dataset includes 2,000 anonymized bankers’ notes extracted from the internal CRM platform, spanning the period from January 2023 to January 2024.  
- Each note was manually labeled by trained compliance officers as either “EFA” or “Non-EFA.”  
  
Data Completeness:  
- No missing records were observed in the dataset. Each instance contains both a narrative and an expert-assigned label.  
  
Data Accuracy:  
- Data was verified by compliance leads to ensure labeling consistency and removal of ambiguous or multi-label notes.  
  
Data Consistency:  
- All text records were processed using consistent UTF-8 encoding. Notes were normalized for spacing, casing, and special characters.  
  
Data Relevance:  
- Narrative notes are the primary source of language signals for identifying EFA; all entries directly relate to customer interactions.  
  
Outlier Handling:  
- Outliers in length (e.g., extremely short or long notes) were flagged but retained due to their potential importance in EFA detection.

## 2.2 Input Design and Control

Feature Engineering:  
- Input text was preprocessed using tokenization, stopword removal, lemmatization, and TF-IDF vectorization.  
- Additional features: presence of EFA-specific phrases, word count, and sentiment score.  
  
Embedding Approaches:  
- TF-IDF served as the primary representation method; no deep learning embeddings were applied in this version.  
  
Feature Selection:  
- Features were selected based on importance derived from preliminary LightGBM runs and domain review of top contributing terms.  
  
Input Control:  
- Input pipeline includes validation checks for nulls, minimum length, and tokenization errors. Sensitivity analysis showed stable predictions with +/-10% TF-IDF variance.

## 2.3 Model Design, Methodology, and Assumptions

Model Type and Justification:  
- LightGBM was selected due to its balance of interpretability and performance on tabular TF-IDF data. It provides feature importance outputs and rapid retraining.  
  
Model Assumptions:  
- Assumes independence between TF-IDF feature dimensions.  
- Assumes the label annotations are accurate representations of ground truth.  
  
Benchmarking:  
- Compared to Logistic Regression and Naive Bayes baselines; LightGBM outperformed both by 6–8% in F1 score.  
- A small BERT-based classifier was tested but did not outperform LightGBM.  
  
Sound Practices:  
- 80/20 stratified train-test split.  
- Performance metrics validated using 5-fold cross-validation.

## 2.4 Explainability and Interpretability

Interpretability Strategy:  
- LightGBM provides inherent interpretability via feature importance.  
  
Global Interpretability:  
- Top 10 TF-IDF terms contributing to classification were reviewed by SMEs and matched known EFA indicators.  
  
Local Interpretability:  
- SHAP analysis was used to explain individual prediction outcomes.  
  
Adverse Action Readiness:  
- Model provides human-readable term importance for flagged notes, suitable for audit and compliance traceability.

## 2.5 Parameter and Hyperparameter Optimization

Parameter Estimation:  
- LightGBM model trained using default parameters optimized for binary classification with log-loss.  
  
Hyperparameter Tuning:  
- Conducted via random search across learning\_rate, max\_depth, and num\_leaves.  
- Optimal values found: learning\_rate=0.05, max\_depth=7, num\_leaves=31  
  
Stability Testing:  
- Model tested with five different random seeds and yielded stable precision/recall within 2% variance.

# 3. Appendices

## 3.1 Feature Dictionary

| Feature Name | Description | Source | Type | Transformation | Role |  
|---------------------|------------------------------------------|----------------|----------|----------------|----------|  
| tfidf\_<term> | Term Frequency-Inverse Document Frequency | NLP pipeline | Numeric | TF-IDF | Predictor|  
| word\_count | Number of words in the note | Derived | Integer | None | Predictor|  
| has\_efa\_keywords | Binary flag for known EFA terms present | Derived | Boolean | Keyword match | Predictor|  
| sentiment\_score | Sentiment polarity score | NLP pipeline | Float | TextBlob | Predictor|  
| label | EFA or Non-EFA | Human expert | Categorical| None | Target |

## 3.2 Acronyms and Abbreviations

EFA – Elder Financial Abuse   
TF-IDF – Term Frequency-Inverse Document Frequency   
SME – Subject Matter Expert   
SHAP – SHapley Additive exPlanations

## 3.3 References

- SR 11-7: Supervisory Guidance on Model Risk Management, Federal Reserve, 2011.   
- Sudjianto, A. & Zhang, A. (2024). Model Validation Practice in Banking: A Structured Approach for Predictive Models.   
- LightGBM Documentation: https://lightgbm.readthedocs.io