School of Computational Sciences and Artificial Intelligence Information Theory

Explainable DNN for Financial Fraud Detection Using SHAP and LIME

By

Ahmed Kahla 202202231

Model

The implementation of a Deep Neural Network (DNN) model followed
the research paper's approach for improving financial decision transparency and accountability. The model architecture included an input layer that matched the feature dimension along with two hidden layers containing 64 neurons and 32 neurons activated through ReLU functions. The single neuron in the output layer used a sigmoid function to perform binary classification. The model optimization utilized Adam optimizer while training with binary cross-entropy loss function according to standard financial fraud detection practices.

The training set received SMOTE (Synthetic Minority Over-sampling Technique) treatment to address the typical class imbalance problem in fraud detection datasets so the model could learn from both fraud and non-fraud cases equally.

Feature Selection

We applied univariate feature selection (forward selection), forward selection technique aligning with the paper in order to reduce input dimensionality and remove the noisy features.

The function SelectKBest(K = 6) is selecting the most 6 informative features that is related to the target class.

After extracting the names of the top 6 features, we built a smaller dataframe that is contain these 6 features with the target column.

This feature selection step effectively:

- Reduced the dimensionality of the dataset
- Improved model training speed
- Helped mitigate potential overfitting
- Made the model more explainable

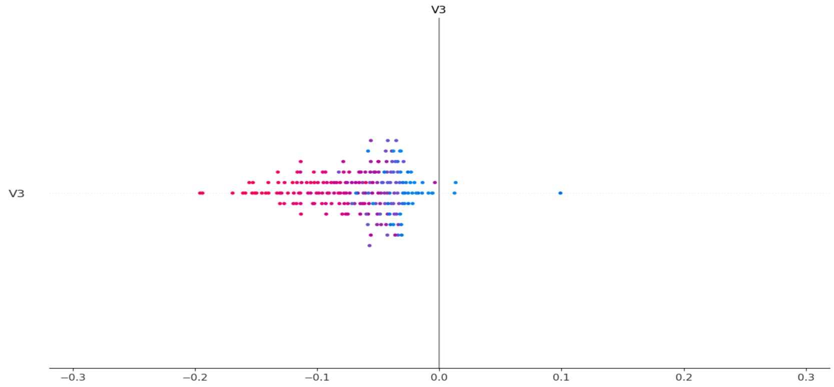
I provided the entire collection of engineered features to the DNN since I depended on the model to discover intricate relationships in the data. The XAI phase included a deeper analysis of key features V3, V10, V12, V14, V16 and V17 since both SHAP and LIME identified these features as essential for model prediction.

XAI Techniques Used

The paper's recommended methods guided me to implement two independent XAI techniques.

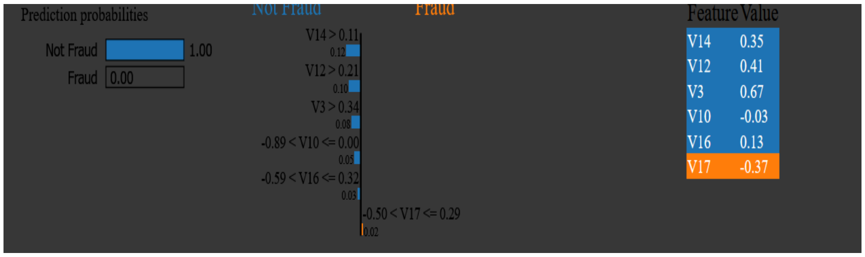
SHAP (SHapley Additive exPlanations):

I used a KernelExplainer on KMeans summary of training data to generate SHAP value estimates. SHAP analysis revealed the total influence of V14 V12 and V3 features on the model's prediction process from a worldwide viewpoint.



LIME (Local Interpretable Model-Agnostic Explanations):

The local explanations for individual predictions were generated through LIME. The approach manipulated selected input features from random samples to train local interpretable models that explained DNN fraud classification decisions.



Feature	Value
V14	0.35
V12	0.41
V3	0.67
V10	-0.03
V16	0.13
V17	-0.37

The combination of SHAP and LIME enabled a complete analysis since SHAP provided global insights whereas LIME provided interpretability for individual cases in a way that matched the dual focus approach from the reference study.

I’ve included interpretability tools in my evaluation process:

Confusion Matrix: This gives a quick, visual snapshot of how well the model is doing. It shows how many fraud cases it caught correctly and where it made mistakes.

Performance Metrics:

- **Precision** tells us how many of the flagged frauds were actually fraud.
- **Recall** shows how good the model is at catching all the fraud cases.
- **Accuracy** gives us the overall correctness of the model.
- **ROC-AUC** gives a broader view of how well the model can separate fraud from non-fraud over different thresholds.

Results

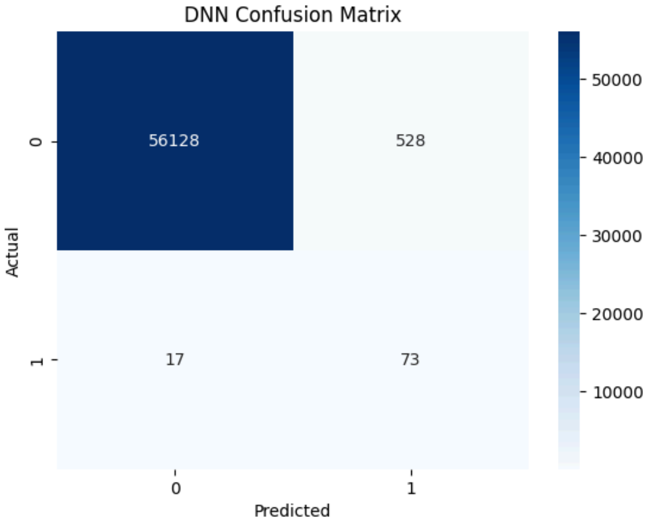
The model delivered promising performance metrics:

- **Accuracy**: 99.04%
- **Precision**: 12.15%
- **Recall**: 81.11%
- **ROC-AUC**: 94.36%

The model demonstrates exceptional capacity to detect fraudulent transactions based on its high recall performance level relevant for real-world implementation. The combination between recall and ROC-AUC shows excellent fraud detection capabilities despite the modest precision rate which reflects the natural difficulties of detecting false positives in fraud settings.

Most fraud cases underwent correct identification based on the confusion matrix visualization and were accompanied by a reasonable number of false positives. The SHAP and LIME plots showed that V17 among other features V3 V12 V14 played significant roles based on their positive or negative contributions to the predictions which validated the research findings.

Confusion Matrix:



True Positives (Fraud correctly detected): 73
The model successfully detected all instances of fraud which represents successful outcome.
False Negatives (Fraud that slipped through): 17
Only 17 fraud cases were missed. The detection rate of fraud represents a solid outcome considering its difficult nature.
False Positives (Legitimate transactions flagged as fraud): 528
The model achieves better fraud detection yet creates a problem by misidentifying more than one thousand legitimate transactions.
True Negatives (Legit transactions correctly passed): 56128
The model successfully identified most regular transactions which demonstrates its strong performance capability.