Interpretability & Explainability in AI Project

Credit Card Fraud Detection Using Machine Learning

Using : IEEE-CIS Dataset

By : Sama Sameh
Joumana Mohamed

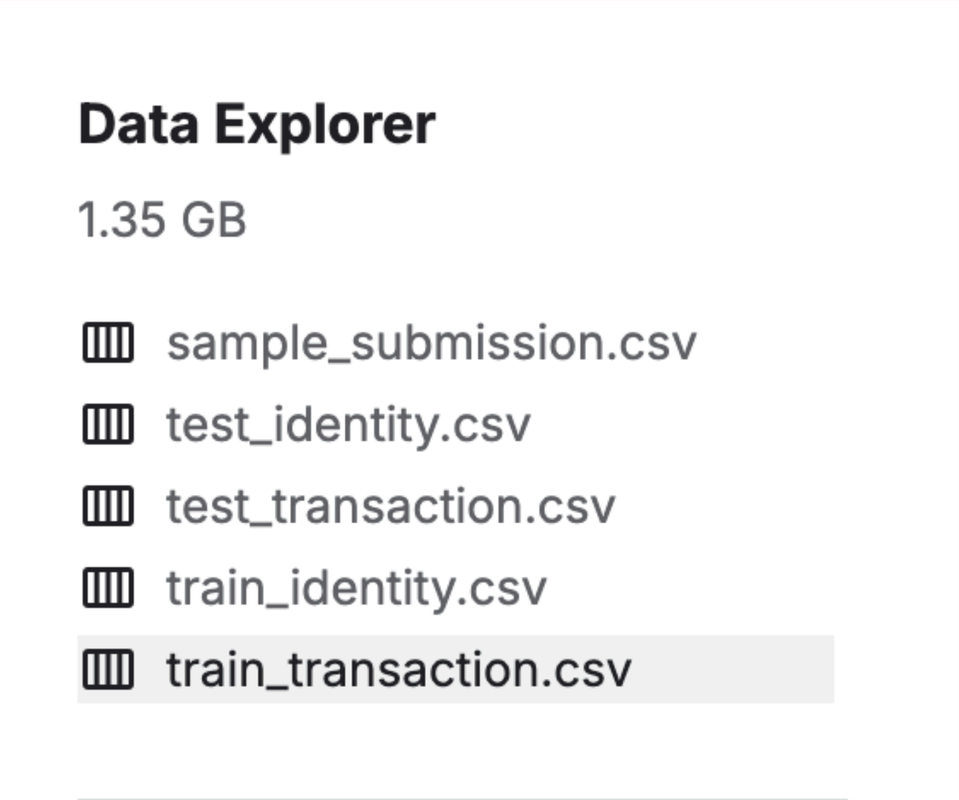
Overview

This project addresses the problem of fraudulent transaction detection using the IEEE-CIS Fraud Detection dataset.
We develop and evaluate various Machine Learning algorithms to detect fraud based on transaction and identity features.
In addition, we compare our results with findings from recent research papers that tackle the same classification problem but on different datasets, allowing us to analyze the generalizability and performance of these models across data sources.

Objectives

- Apply and fine-tune ML models (e.g., Logistic Regression, Random Forest, XGBoost, SVM , Naive Baise) to detect fraud
- Preprocess and explore the IEEE-CIS dataset for meaningful patterns
- Evaluate model performance using AUC, Precision, Recall, and F1-Score
- Compare our model performance with published results from related research papers
- Analyze the impact of dataset differences on fraud detection accuracy
- Interpret model predictions to highlight important features and patterns

DataSet Components



Dataset Overview:

- Source: IEEE-CIS (Kaggle)
- Goal: Predict whether a financial transaction is fraudulent or genuine.
- Type: Binary Classification Problem

Main Dataset Components:

- train_transaction.csv & test_transaction.csv:
  Contains data about the financial transactions.
- train_identity.csv & test_identity.csv :
  Contains data about the user's identity who made the transaction

https://www.kaggle.com/c/ieee-fraud-detection/data?select=train_transaction.csv

- **Methodology And Key Decisions: Feature Scenarios & Interpretations**

Scenario	Interpretation
“TransactionAmt” is very high	May indicate a suspicious transaction
Unusual “DeviceType”	Could be an unfamiliar device for the user
Very large “dist1"	May suggest travel or account compromise

Target : isFraud=1

or

Zero ?

- **This Table shows the most important features to predict fraud or not.**

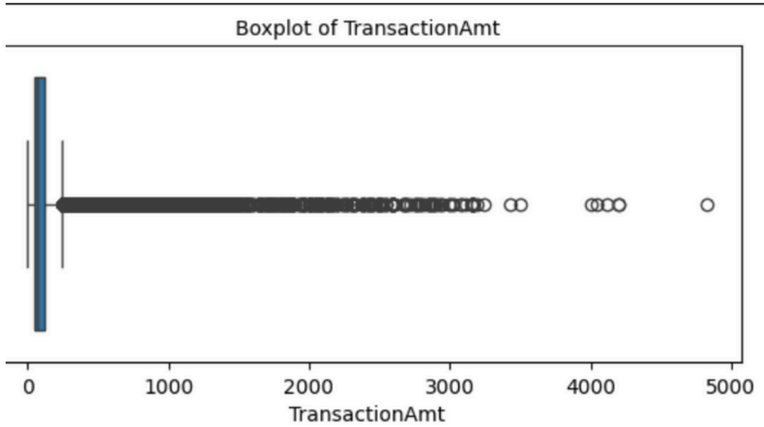
Challenges:

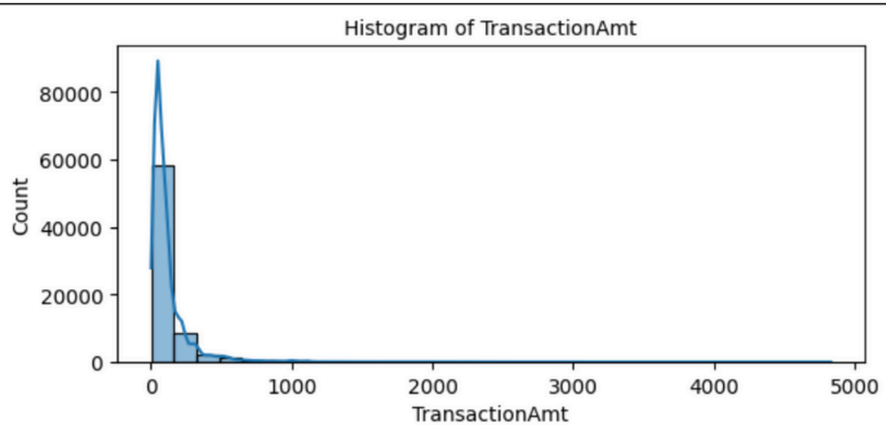
- **Imbalanced classes: Few fraud cases compared to normal ones**
- **Missing values: Many features contain missing data**
- **High feature count: Over 400 features to handle**

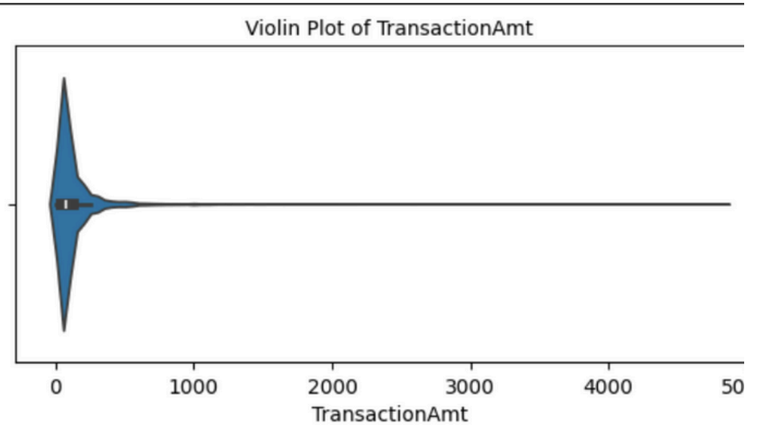
- **EDA Techniques**

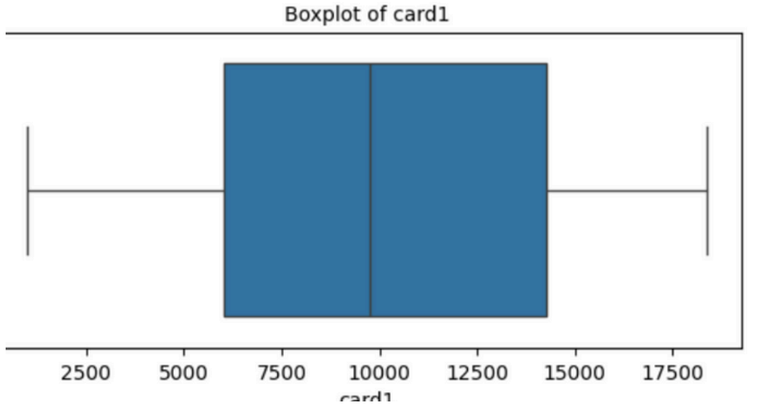
1-Univariate Analysis

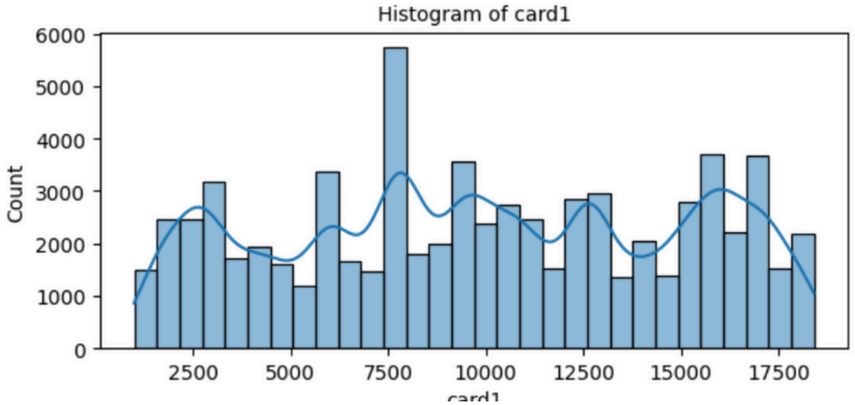
Goal: Understand the distribution of individual features.
Techniques: Histograms , Boxplots & Violin Plot

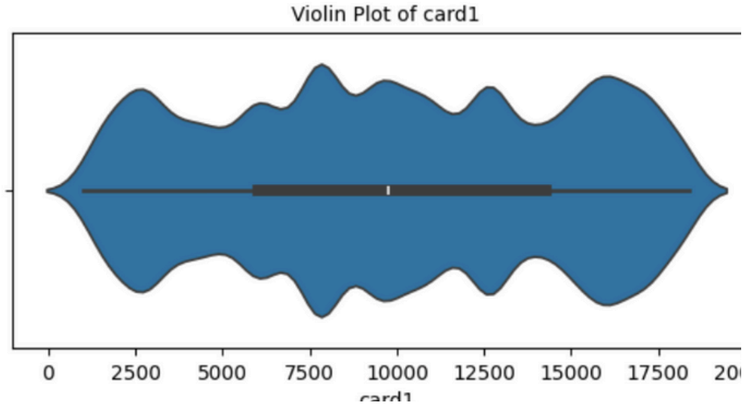






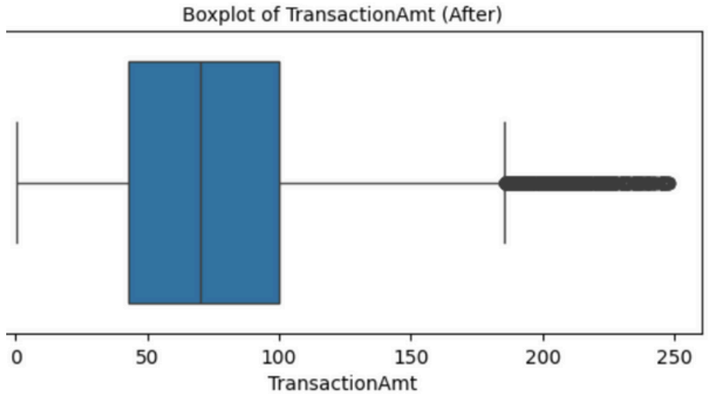


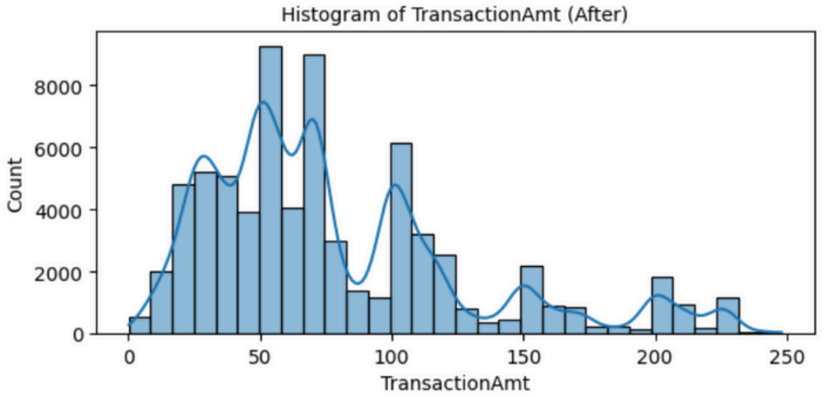


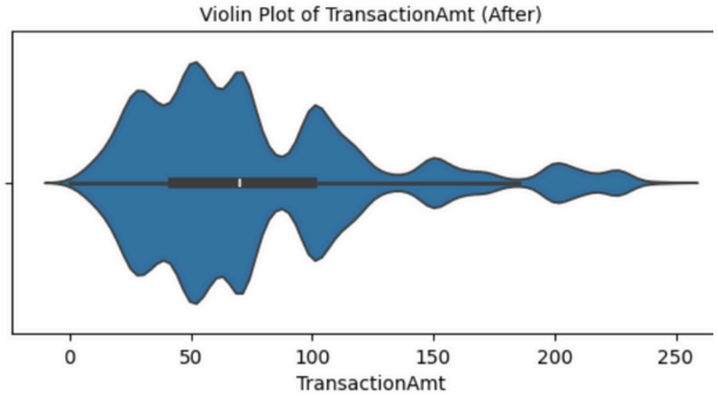


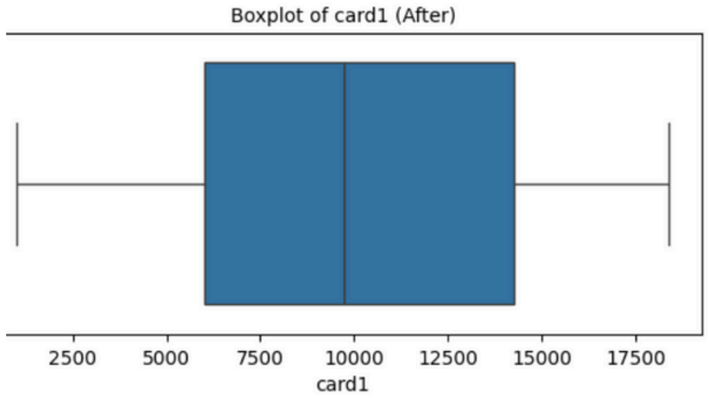
2- Outlier Detection

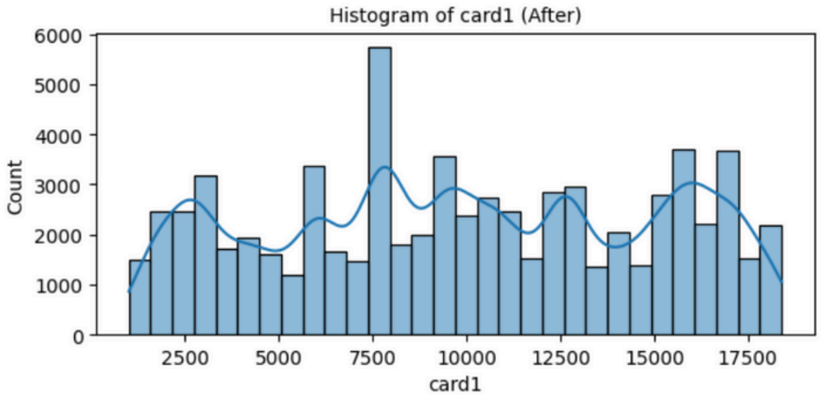
Techniques : Boxplots , IQR methods , Scatter plots

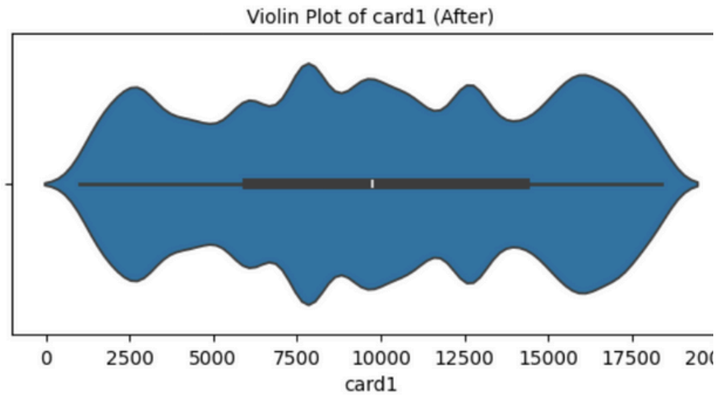






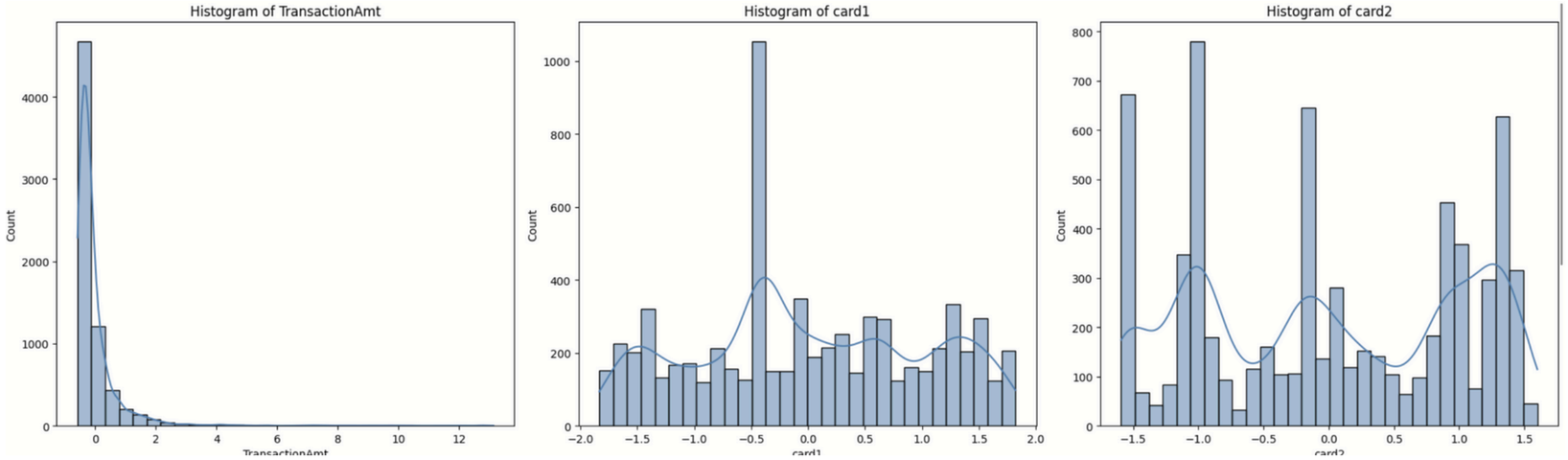




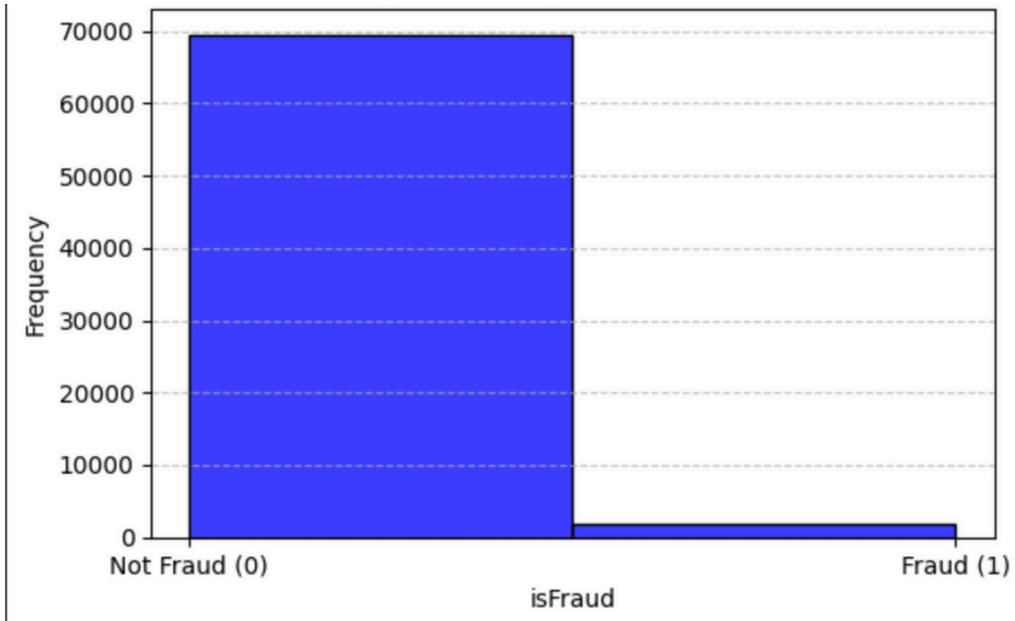


3-Visual insights into the distribution of key numeric and categorical features.

Useful for detecting : Skewed data , Outliers , Empty or sparse features



4-Histogram Plot Of The Target Class



Random Forest Classifier results:

Class 0 (Non-Fraud / Majority Class):

- Precision: 0.99 → Very few false positives.
- Recall: 1.00 → Model detects all non-fraud cases.
- F1-Score: 0.99 → Excellent balance.

Class 1 (Fraud / Minority Class):

- Precision: 0.78 → it's correct 78% to predict Fraud.
- Recall: 0.47 → Only detects 47% of actual fraud cases.
- F1-Score: 0.58 → Moderate, due to low recall.

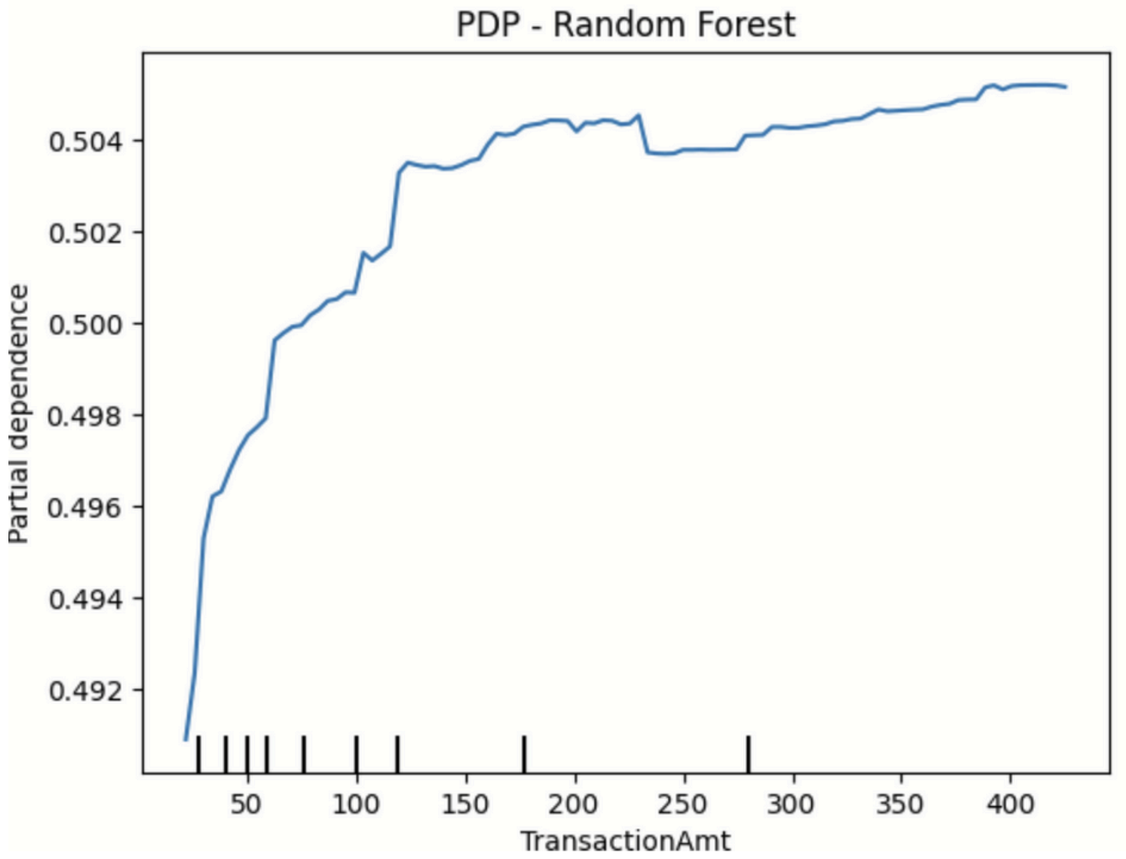
```
Accuracy: 0.9870633893919794
ROC-AUC: 0.8953386103781882

Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       758
           1       0.78      0.47      0.58        15

    accuracy                           0.99       773
   macro avg       0.88      0.73      0.79       773
weighted avg       0.99      0.99      0.99       773
```

Random Forest Classifier
XAI Tool 1 : PDP

- X-axis ->TransactionAmt.
- Y-axis ->Partial dependence
- Curve: Shows how the predicted probability of the positive class (typically fraud in fraud detection) changes as TransactionAmt increases, while keeping other features (averaged).

Interpretation:

- The plot reflect that while TransactionAmt increases, the predicted probability of the positive class slightly increases.
- The higher transaction amounts the higher likelihood of being(possibly fraudulent)

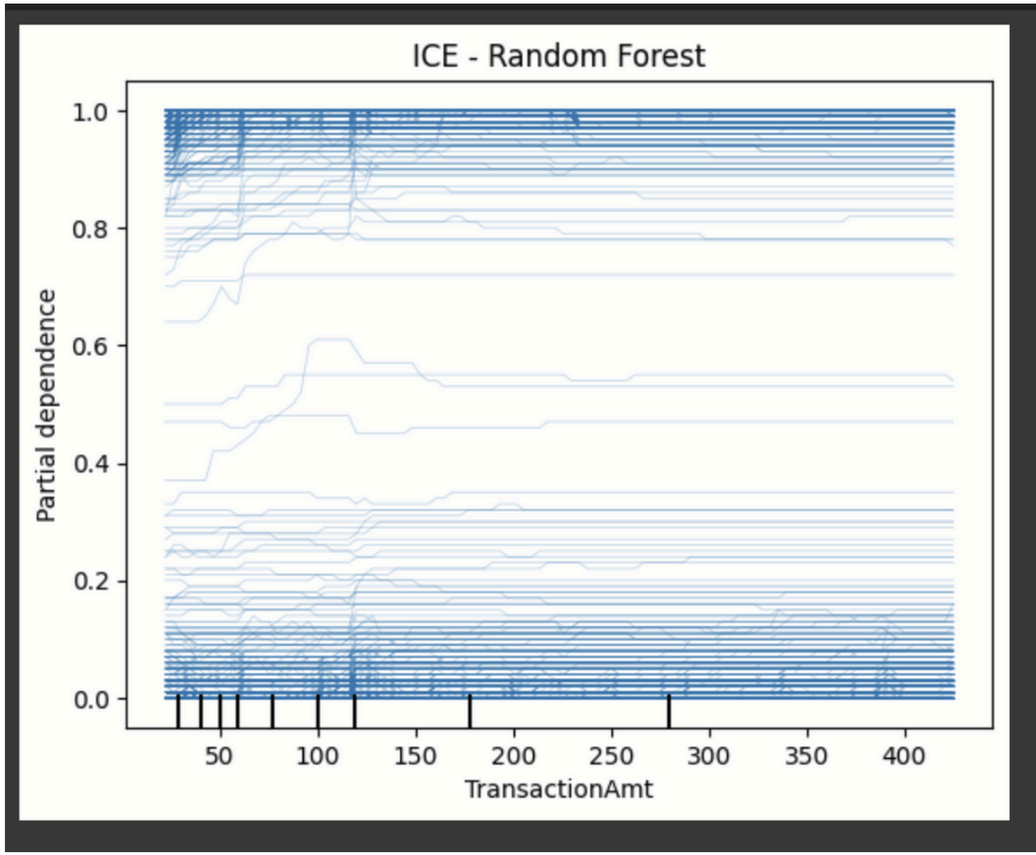


Random Forest Classifier
XAI Tool 2: ICE

- X-axis ->TransactionAmt.
- Y-axis ->Partial dependence
- Curve: A single instance for TransactionAmt varied while all other features kept fixed.

Interpretation:

- Steep upward lines → Higher amounts → higher fraud risk (per instance)
- Non-parallel lines → Feature interactions with TransactionAmt
- Flat lines at low values → Low impact for low-risk transactions
-

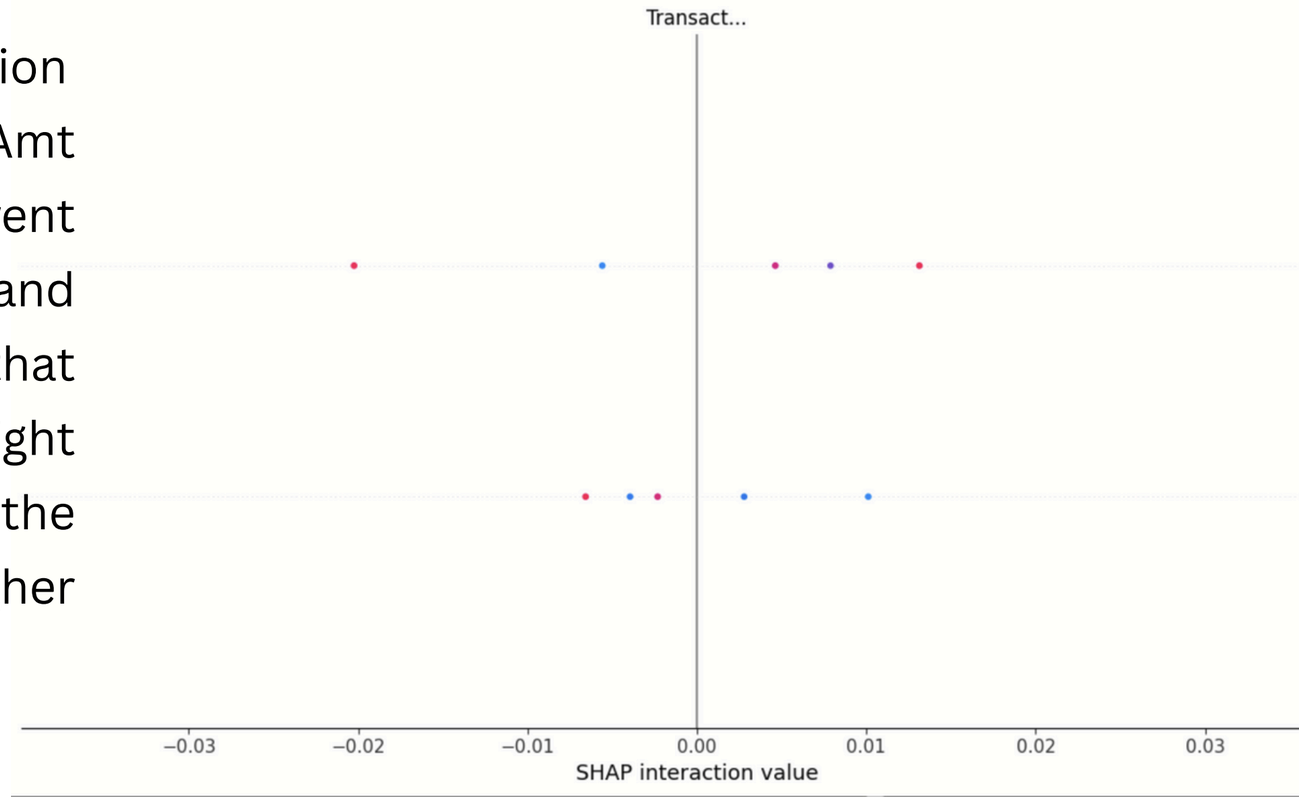


Random Forest Classifier
XAI Tool 3: LIME

- LIME goal is to explain the prediction of a single, specific data instance.
- Orange bars indicate fraud. For example, the value of "V294" is pointing strongly towards fraud.
- Blue bars indicate Not Fraud
- The chart below indicates the importance for each feature

Random Forest Classifier
XAI Tool 4: SHAP

- SHAP general Idea: Fair Payoffs Based on Contribution
- This plot shows how the value of TransictionAmt affects the model's prediction for different transactions. Each dot represents a transaction, and its horizontal position shows the impact of that feature's value on the prediction. Dots to the right mean a higher value of this feature pushes the prediction higher, while dots to the left mean a higher value pushes it lower.



XGBoost Classifier results:

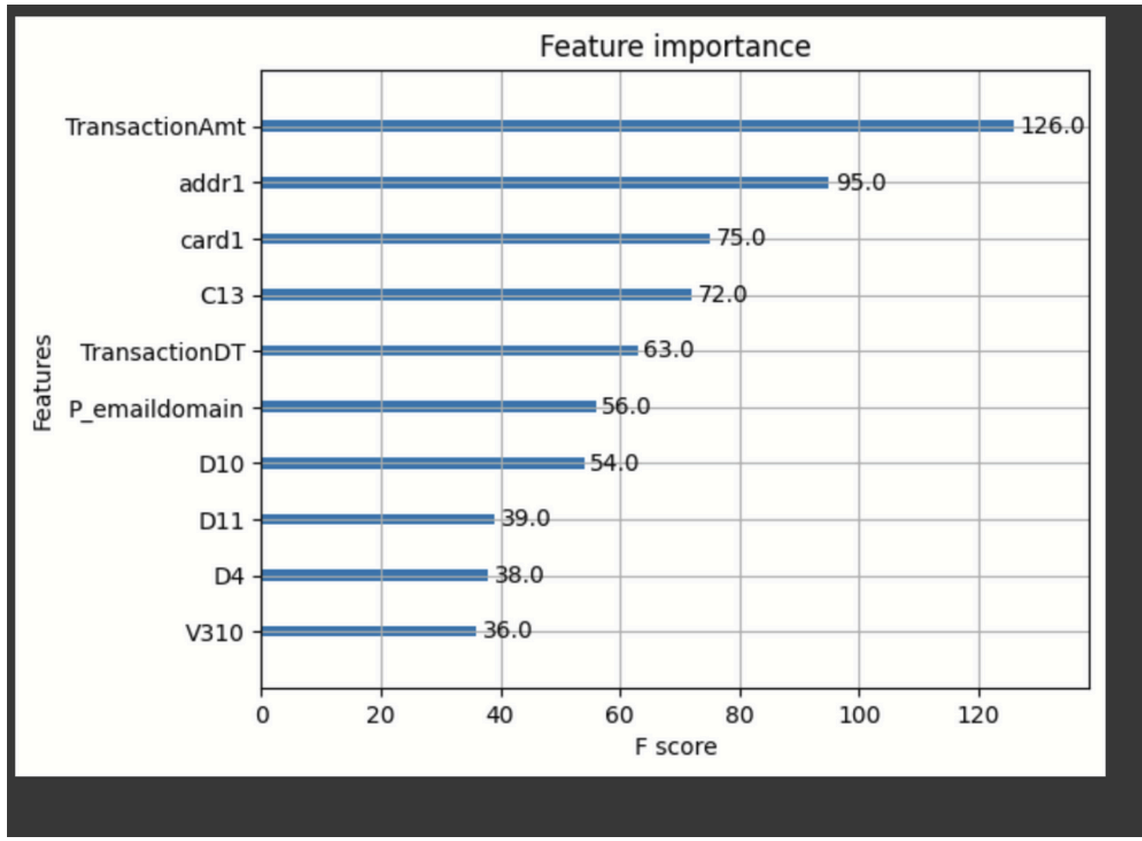
- Class 0 (Non-Fraud / Majority Class):
- Precision: 0.99 → Very few false positives.
- Recall: 1.00 → Model detects all non-fraud cases.
- F1-Score: 0.99 → Excellent balance.
- Class 1 (Fraud / Minority Class):
- Precision: 0.78 → it's correct 78% to predict Fraud.
- Recall: 0.47 → Only detects 47% of actual fraud cases.
- F1-Score: 0.58 → Moderate, due to low recall.

```
Accuracy: 0.9870633893919794
ROC-AUC: 0.8953386103781882

Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       758
           1       0.78      0.47      0.58        15

    accuracy                           0.99       773
   macro avg       0.88      0.73      0.79       773
weighted avg       0.99      0.99      0.99       773
```

XGBoost Classifier
XAI Tool 1: Feature Importance

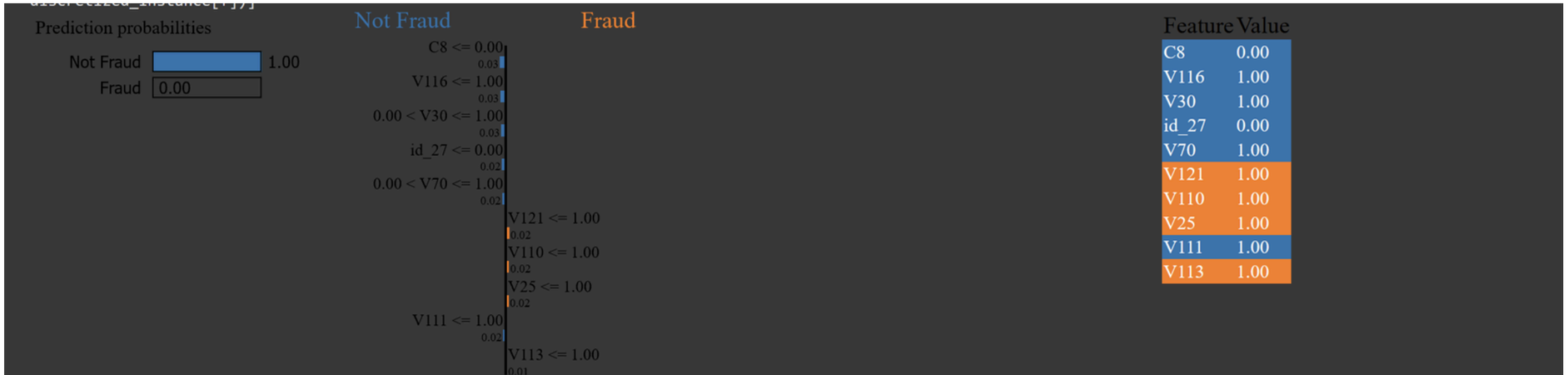
Interpretation:

- This plot shows the importance of each feature on the model performance
- TransictionAmt is the most important feature which has high effect on detecting the class(Fraud or Not)



XGBoost Classifier
XAI Tool 2: LIME

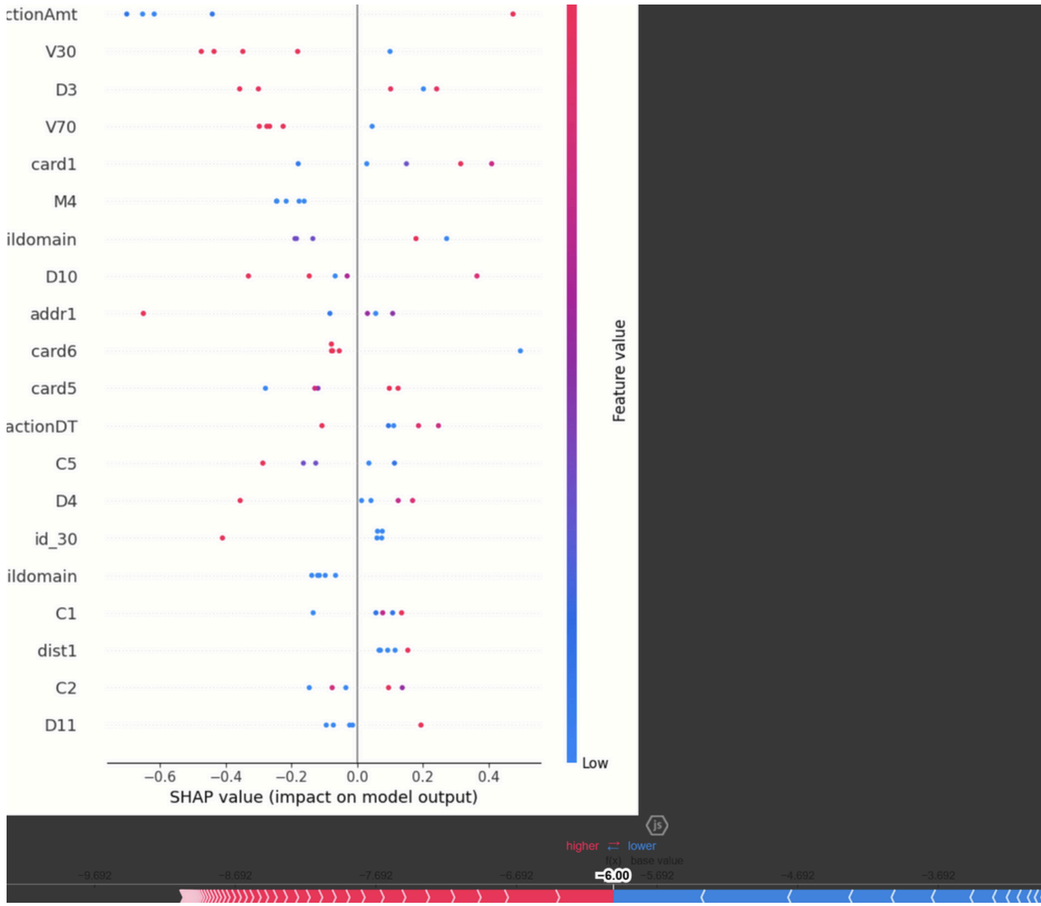
- LIME goal is to explain the prediction of a single, specific data instance.
- Orange bars indicate fraud. For example, the value of "V121" is pointing strongly towards fraud.
- Blue bars indicate Not Fraud
- The chart below indicates the importance for each feature for detecting if it is Fraud or Not



XGBoost Classifier
XAI Tool 3: SHAP

Interpretation:

- SHAP summaryplot shows which features are most important for the model's predictions and how they affect the output.
- The most important features are on the top
- Horizontal axis indicates the impact of a feature's value on the model's output for each transaction.
- TransactionAmt is a very important feature: Higher transaction amounts (red dots) tend to have a large positive impact on the model's output

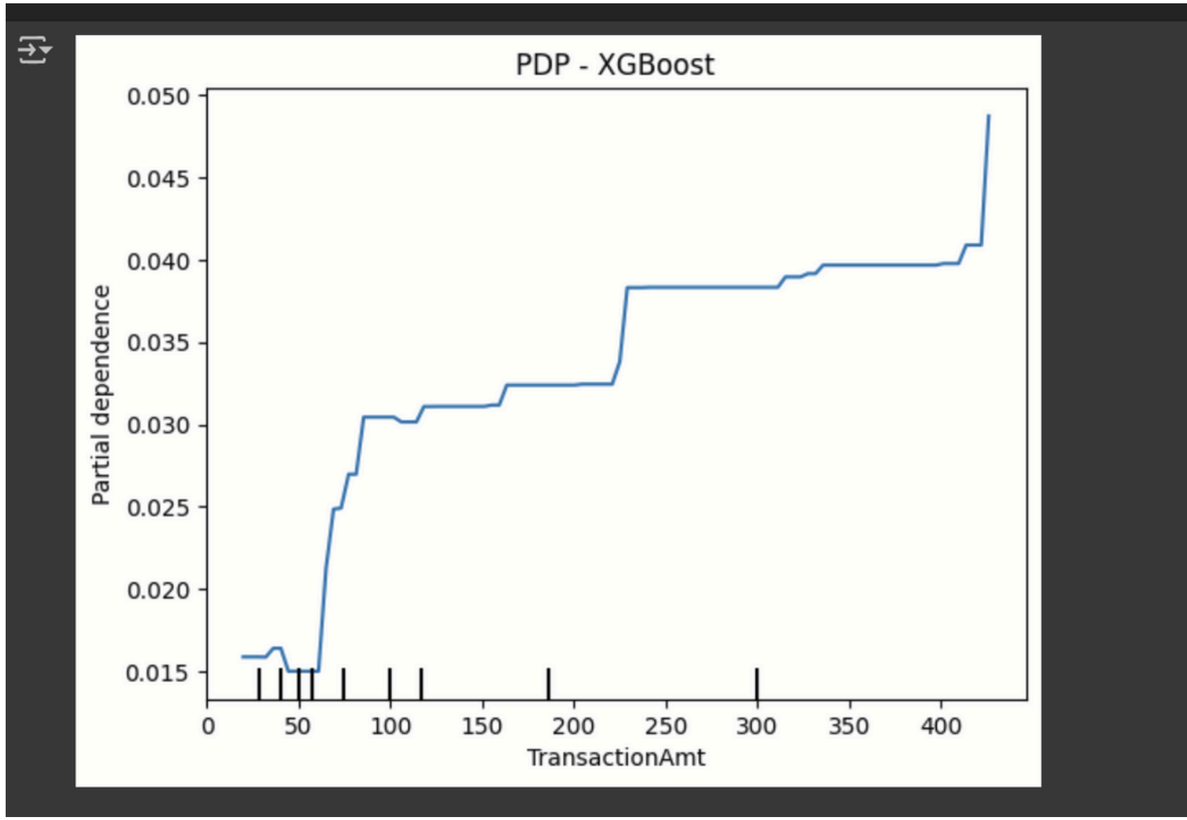


XGBoost Classifier
XAI Tool 4: PDP

- X-axis ->TransactionAmt.
- Y-axis ->Partial dependence
- Curve: Shows how the predicted probability of the positive class (typically fraud in fraud detection) changes as TransactionAmt increases, while keeping other features (averaged).

Interpretation:

- The plot reflect that while TransactionAmt increases, the predicted probability of the positive class slightly increases.



SVM Classifier results:

- High Overall Accuracy (93%): The model correctly predicts the outcome for most cases.
- Good Performance for Both Classes:
- Class 0: Correctly identified 95% of the actual cases (Recall), and 91% of its predictions were right (Precision).
- Class 1: Correctly identified 91% of the actual cases (Recall), and 95% of its predictions were right (Precision).
- Balanced Performance (F1-Score: 0.93)

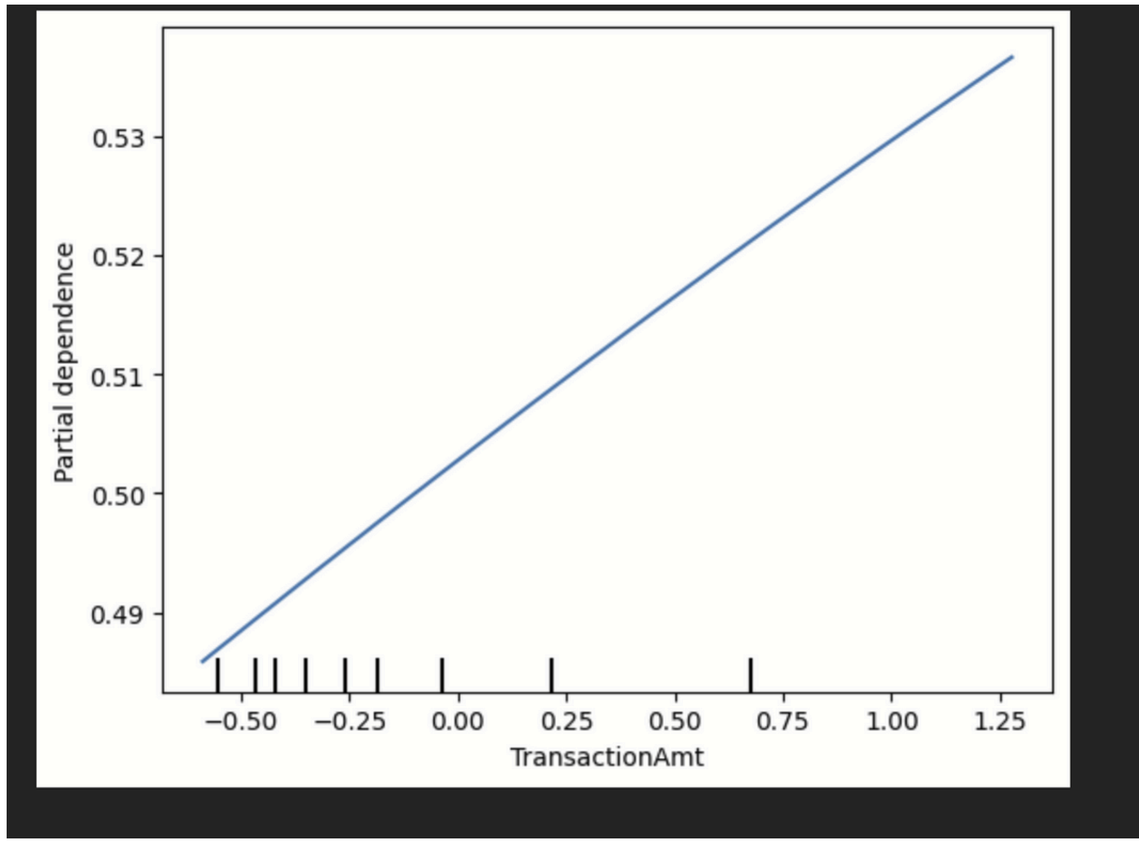
```
Accuracy: 0.9300595238095238

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.95      0.93      1339
           1       0.95      0.91      0.93      1349

    accuracy                           0.93      2688
   macro avg       0.93      0.93      0.93      2688
weighted avg       0.93      0.93      0.93      2688
```

SVM Classifier
XAI Tool 1: PDP

- X-axis ->TransactionAmt.
- Y-axis ->Partial dependence
- Curve: Shows how the predicted probability of the positive class (typically fraud in fraud detection) changes as TransactionAmt increases, while keeping other features (averaged).

Interpretation:

- The plot reflect that while TransactionAmt increases, the predicted probability of the positive class increases significantly

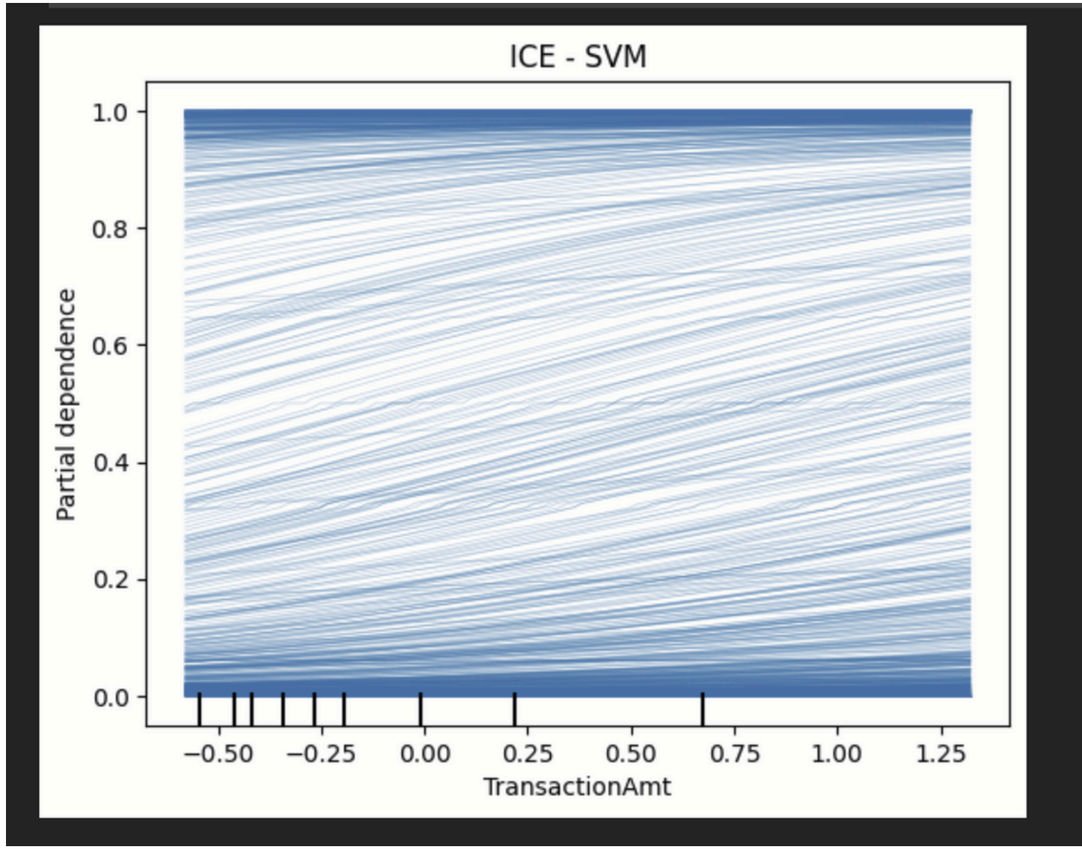


SVM Classifier
XAI Tool 2: ICE

- X-axis ->TransactionAmt.
- Y-axis ->Partial dependence
- Curve: A single instance for TransactionAmt varied while all other features kept fixed.

Interpretation:

- Steep upward lines → Higher amounts → higher fraud risk (per instance)
- Non-parallel lines → Feature interactions with TransactionAmt
- Flat lines at low values → Low impact for low-risk transactions

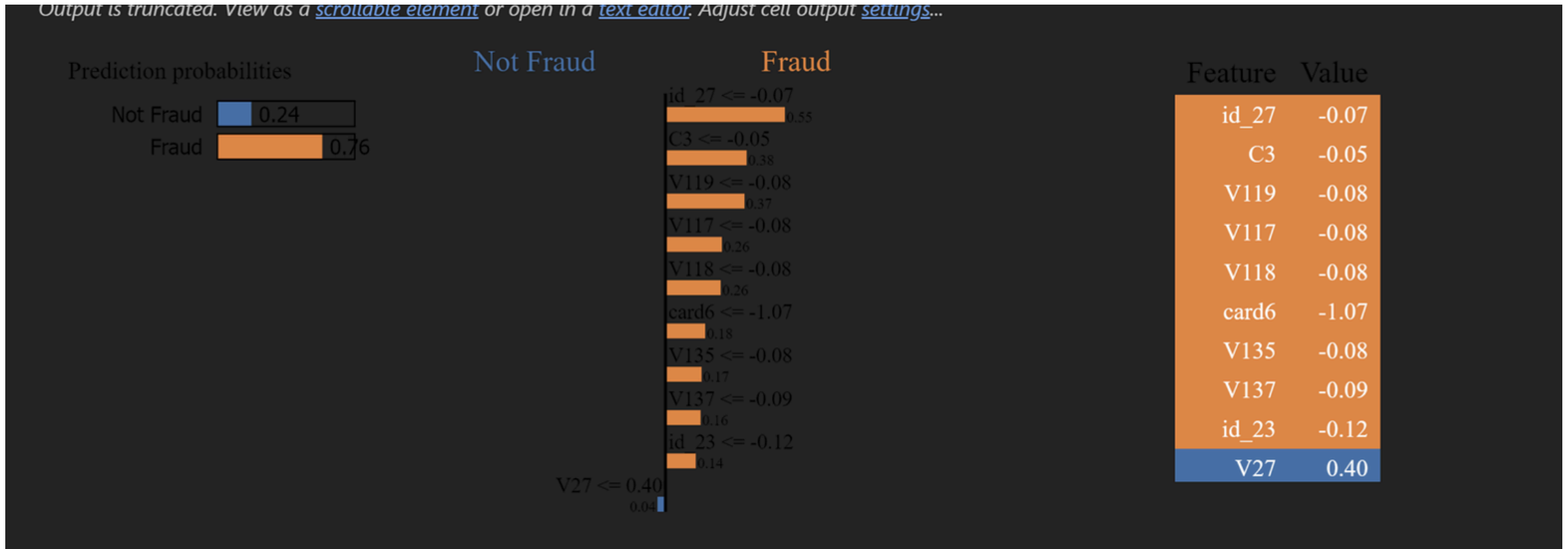


SVM Classifier
XAI Tool 3: LIME

Interpretation:

LIME goal is to explain the prediction of a single, specific data instance.Likely Fraud (76%): The model predicts a high probability of fraud for this specific case.

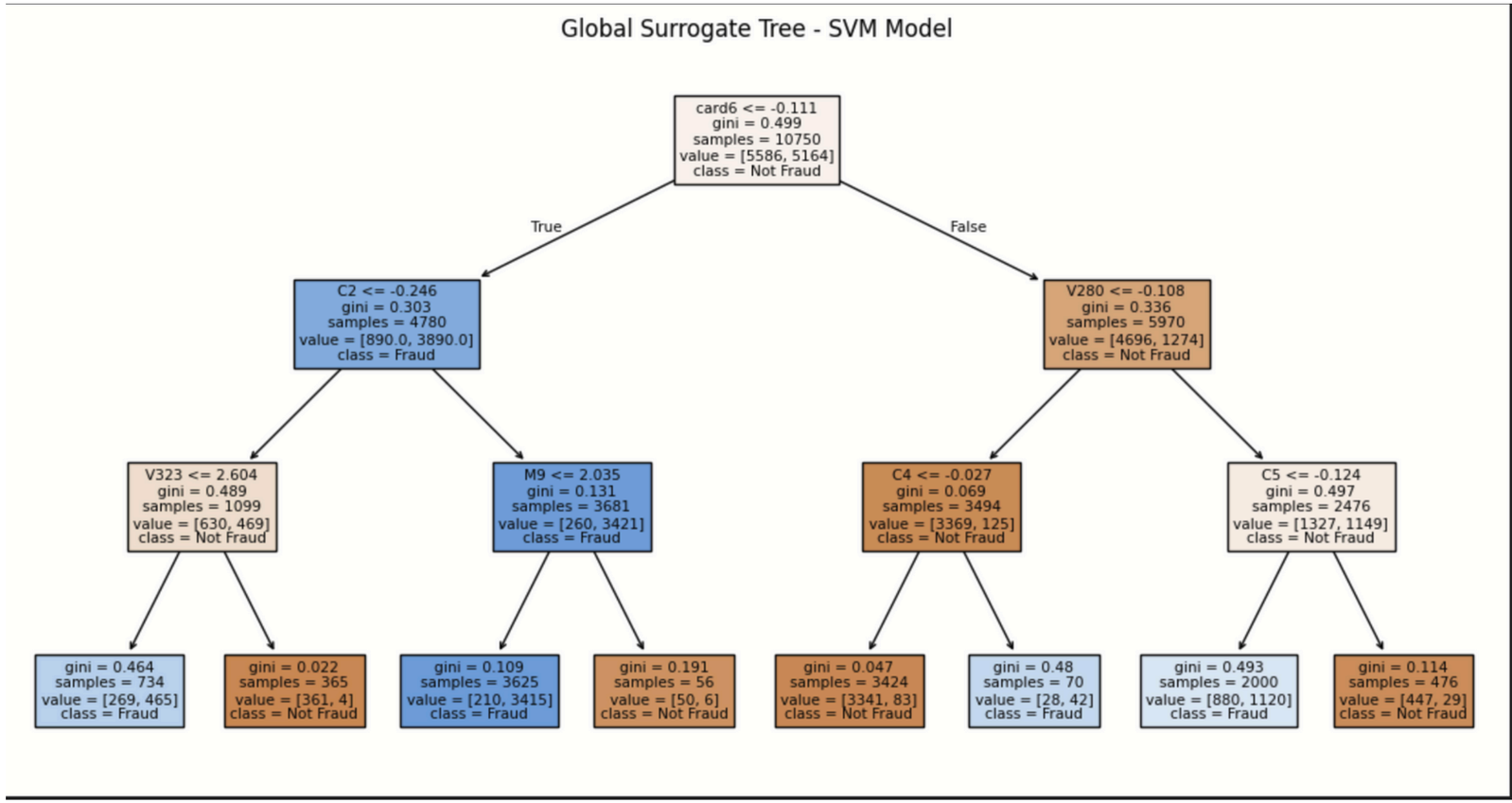
- Key Indicators of Fraud: Features like id_27, C3, V119, with their negative values, strongly suggest fraud.
- The positive value of V27 slightly reduces the likelihood of fraud but isn't enough to change the overall prediction.
- Overall Assessment: The combination of these factors leads the model to believe this transaction is likely fraudulent.



SVM Classifier
XAI Tool4: Global Surrogate Tree

Interpretation:

The tree shows a simplified way the model makes a fraud prediction by asking a sequence of questions(gini score) about different transaction features. By following the "True" or "False" answers down the tree, we reach a final prediction of whether a transaction is "Fraud" or "Not Fraud."



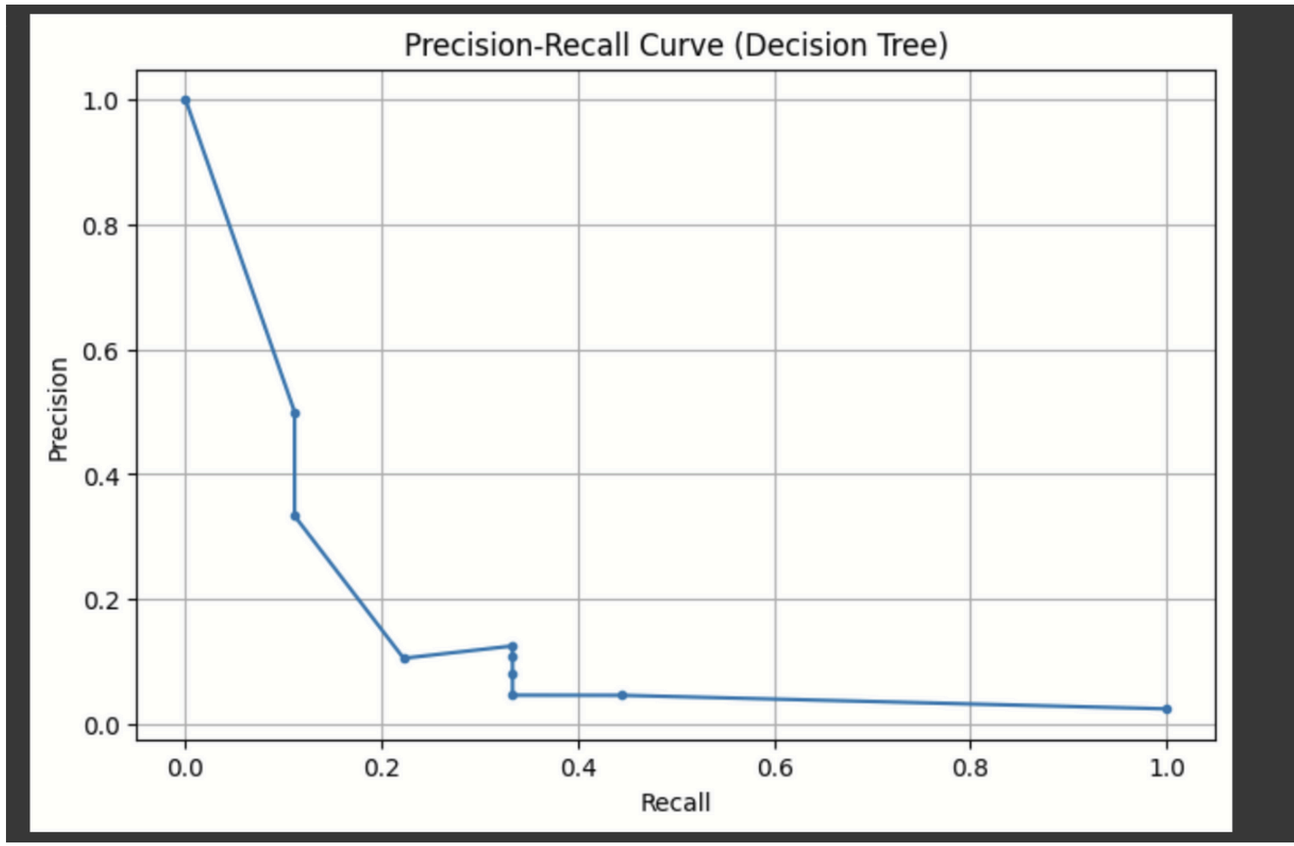
Desicision Tree Classifier results:

- Accuracy: 0.9353099730458
- Classification Report:
- Class 0:
  - Precision: It's correct 98% of the time. This is very good.
  - Recall: 0.95: The model correctly identifies 95% of all actual class 0 instances. This is also very good.
  - F1-score: 0.97: The harmonic mean of precision and recall is high, indicating a good balance for class 0.
- Class 1:
  - Precision: 0.11: When the model predicts class 1, it's only correct 11% of the time.
  - Recall: 0.22: The model only correctly identifies 22% of all actual class 1 instances.
  - F1-score: 0.14: The F1-score is also very low, indicating poor performance for class 1.

```
Accuracy: 0.9353099730458221

Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.95      0.97       362
           1       0.11      0.22      0.14         9

    accuracy                           0.94       371
   macro avg       0.54      0.59      0.55       371
weighted avg       0.96      0.94      0.95       371


Confusion Matrix:
 [[345  17]
 [  7   2]]
```

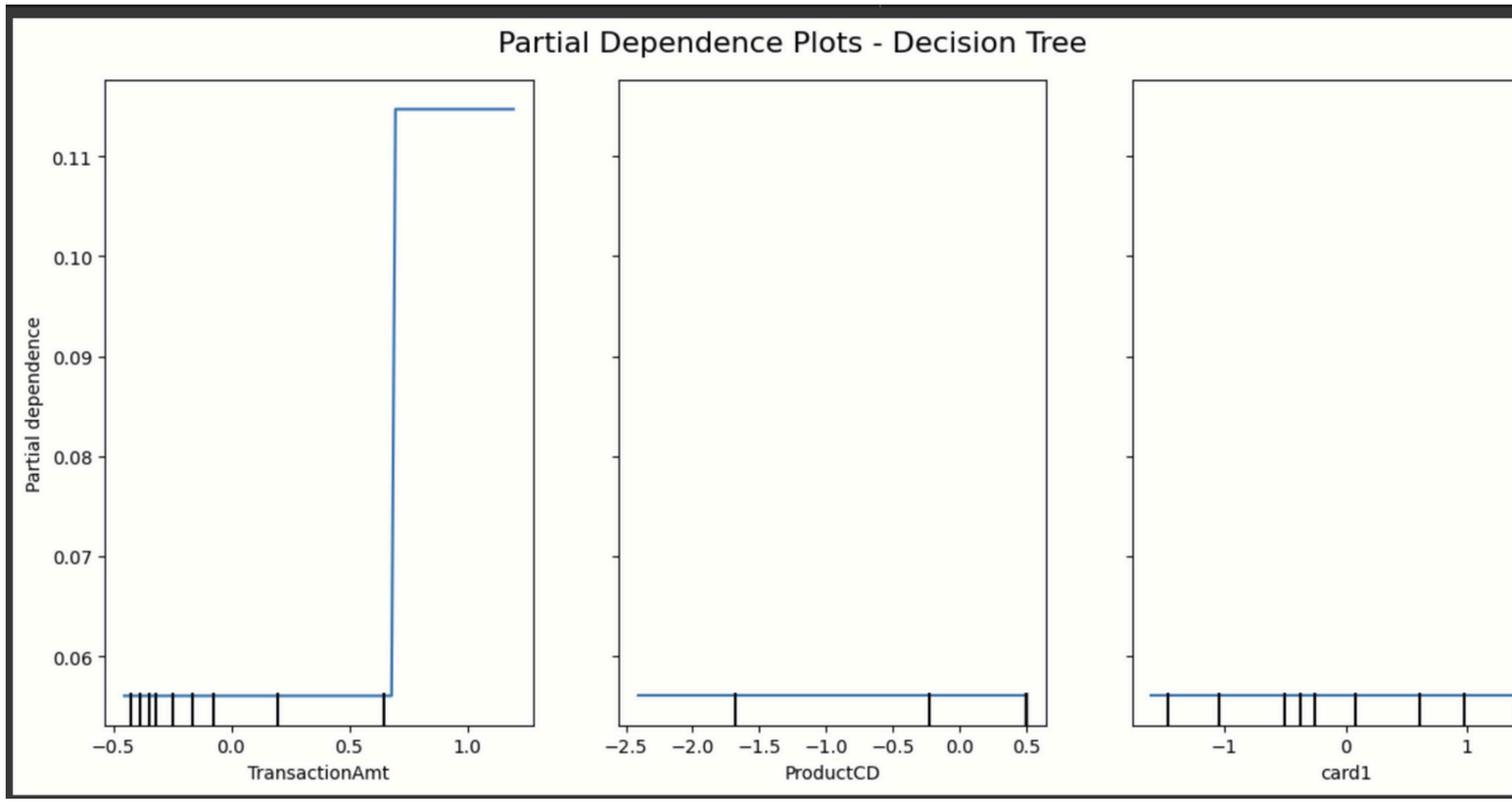
Desicision Tree Classifier results: Presision&Recall curve

- The curve shows the balance between Precision (how accurate positive predictions are) and Recall (how many actual positives are captured) at different thresholds.
- Poor Performance After Resampling: The curve stays mostly in the bottom-left area, indicating poor performance, especially in achieving both high precision and high recall simultaneously.



Desicision Tree Classifier
XAI tool 1: PDP

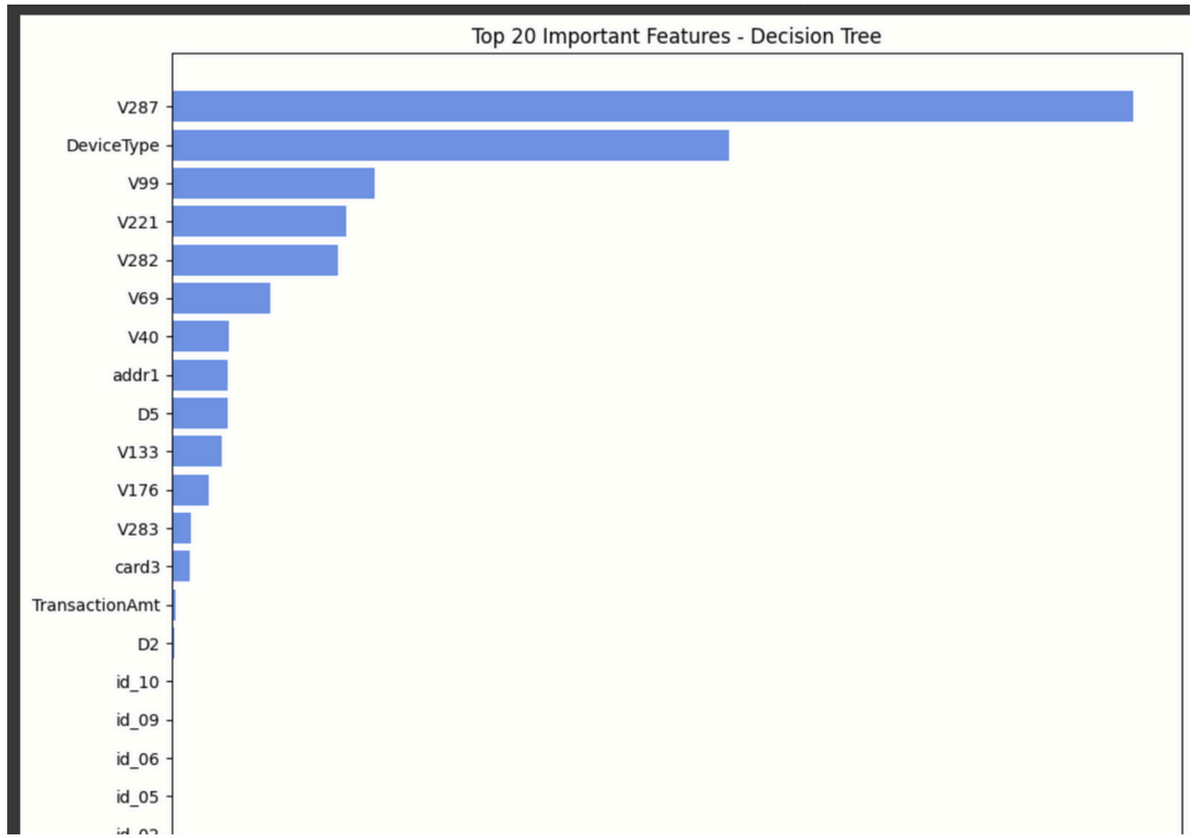
- TransactionAmt Impact: For TransactionAmt, the predicted outcome is low until (around 0.5). Above this, the predicted outcome signifcantly increases and then plateaus.
- ProductCDnImpact: The ProductCD plot shows a nearly flat line. This indicates that the ProductCD feature has very little influence on the Decision Tree's predictions.
- card1 Impact: Similarly, the card1 plot is also mostly flat. This suggests that the card1 feature has minimal impact on the Decision Tree's predictions.



Desicision Tree Classifier
XAI tool 2: Feature Importance

Interpretation:

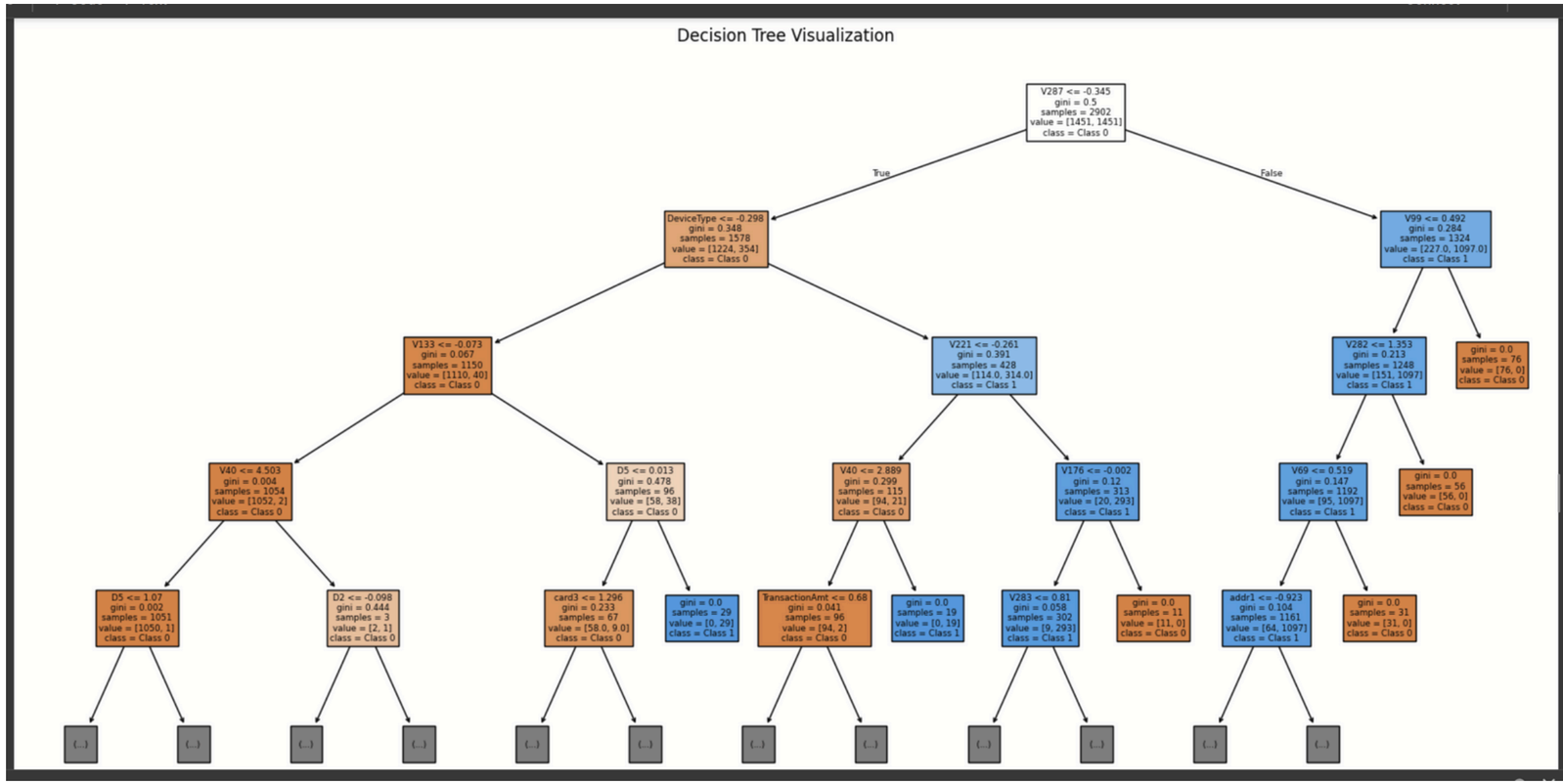
- This plot shows the importance of each feature on the model performance
  - V287 is the most important feature which has high effect on detecting the class(Fraud or Not) for the Decision Tree Model



Desicision Tree Classifier
XAI tool 3: Feature Importance

Interpretation for the DT model:

The tree shows a simplified way the model makes a fraud prediction by asking a sequence of questions(gini score) about different transaction features. By following the "True" or "False" answers down the tree, we reach a final prediction of whether a transaction is "Fraud" or "Not Fraud."



Desicision Tree Classifier
XAI tool 4: LIME

Interpretation:

LIME goal is to explain the prediction of a single, specific data instance.

- Strong Prediction for Class 0 (100%): The model is highly confident in predicting Class 0.
- Features Pushing Towards Class 1 (Orange): Features like V99, V69, V141, and V327 with their negative values contribute to increasing the likelihood of Class 1, but their influence is outweighed.
- Features Pushing Towards Class 0 (Blue): Features like V118, V138, V162, V108, and V119, with their negative values, strongly drive the prediction towards Class 0.

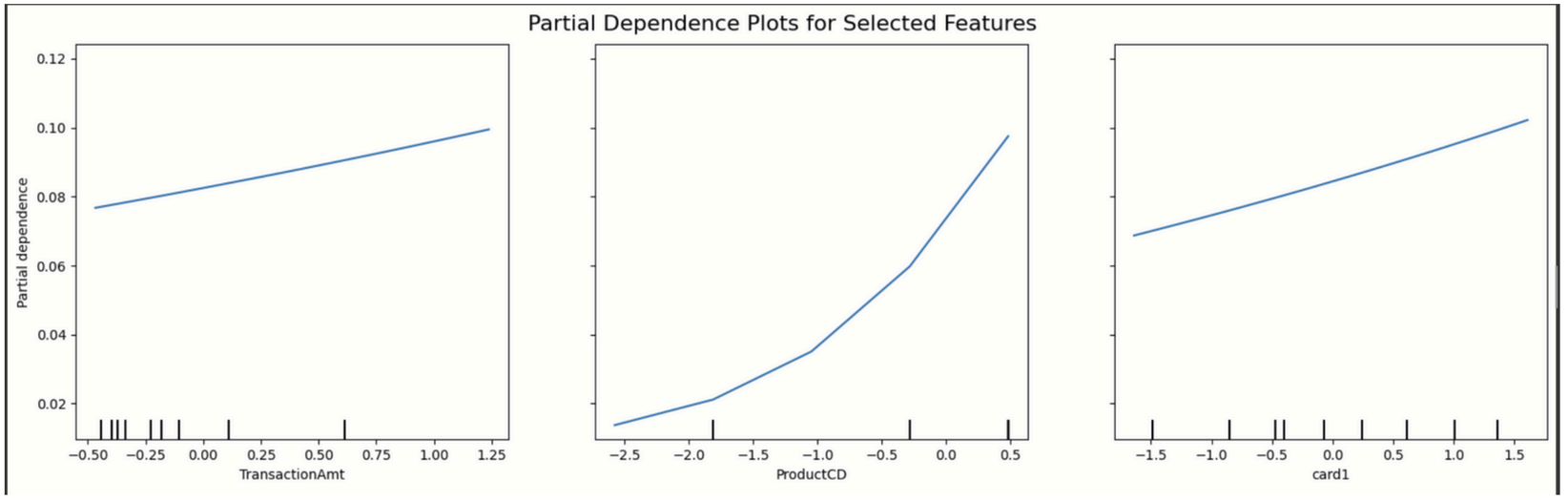
Logistic Regression Model result:

- The model has high accuracy (93%), but performs much better on class 0 (high precision and recall). For class 1, precision is low (18%), meaning many positive predictions are wrong, though recall is better (62%), indicating it captures more of the actual class 1 instances. The F1-score for class 1 is low (0.28), showing an imbalance in performance. Class 0 has significantly more support (556) than class 1 (13).
  -

```
Accuracy : 0.9279437609841827

Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.94      0.96       556
           1       0.18      0.62      0.28        13

    accuracy                           0.93       569
   macro avg       0.59      0.78      0.62       569
weighted avg       0.97      0.93      0.95       569
```

Logistic Regression Model
XAI tool 1:PDP

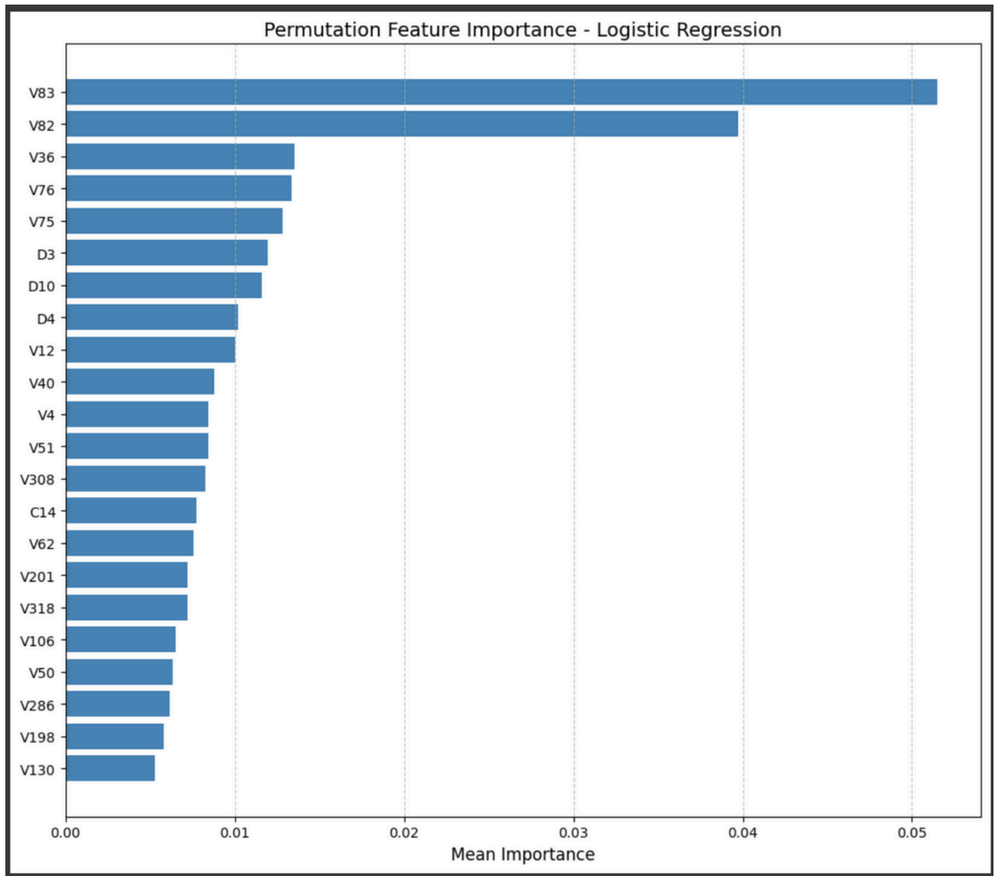
- TransactionAmt Positive Trend: Higher transaction amounts tend to slightly increase the predicted outcome.
- ProductCD Exponential Increase: The predicted outcome shows a strong exponential increase as ProductCD increases.
- card1 Moderate Positive Trend: Higher values of card1 are associated with a moderate increase in the predicted outcome.



Logistic Regression Model
XAI tool 2:Feature Permutation

Interpertation:

- V83 is Most Important: The feature V83 has the highest mean importance, meaning shuffling its values caused the biggest drop in model performance.
- Top Influential Features: V83, V82, V36, V76, and V75 are the most important features for the Logistic Regression model's predictions.
- Decreasing Importance: The importance of features decreases as we go down the list.
- Feature Relevance: Features with higher bars are more crucial for the model to make accurate predictions.

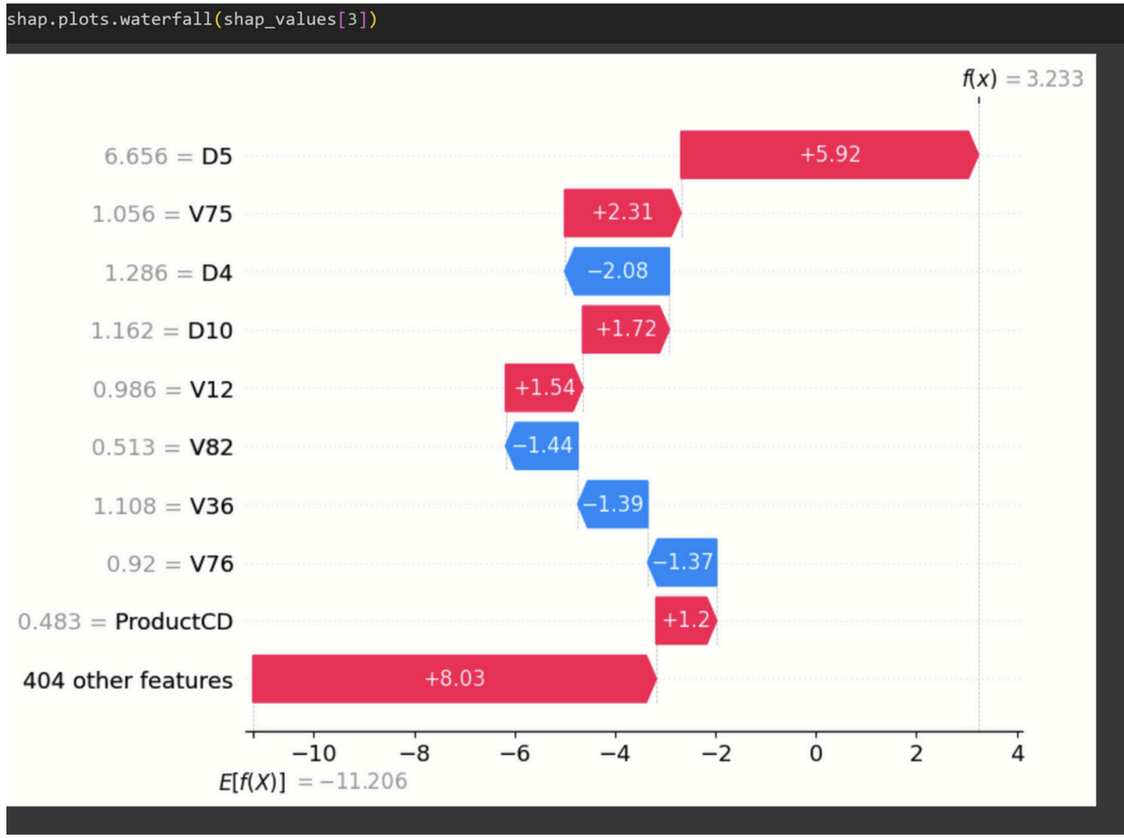


Logistic Regression Model
XAI tool 3:SHAP

Interpertation:

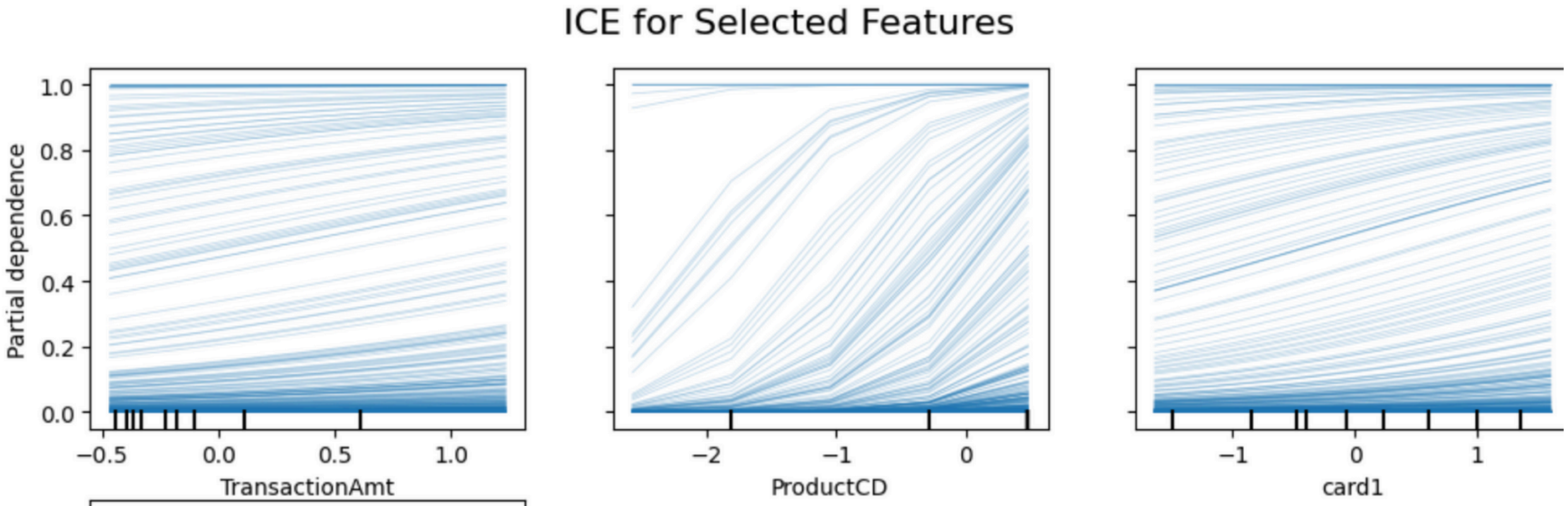
SHAPshows which features are most important for the model's predictions and how they affect the output

- Starts at -11.2 (average prediction).
- Red bars (like D5 and others) push the prediction up.
- Blue bars (like D4 and V82) push it down..



Logistic Regression
XAI Tool 4: ICE

- Individual Predictions:
- TransactionAmt Gradual Increase: For most transactions, the predicted outcome tends to gradually increase as the TransactionAmt increases.
- ProductCD :The impact of ProductCD is highly variable across transactions. For some, the prediction jumps sharply at a certain ProductCD value, while for others, it remains low.
- card1 Moderate Variation: The effect of card1 on the prediction shows some variation across transactions.



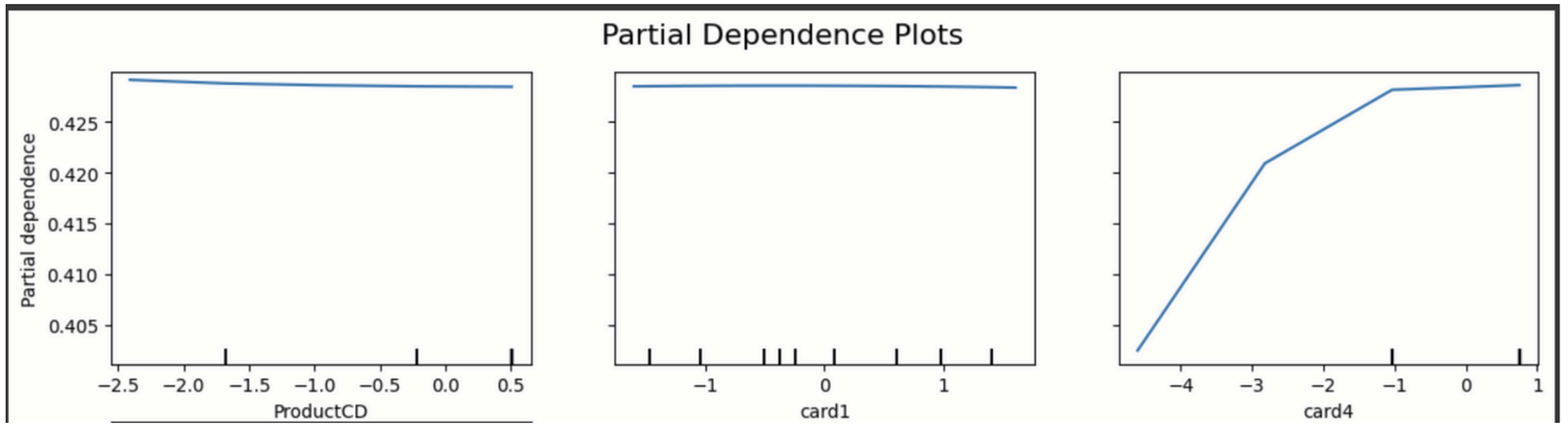
Navie Bayes Model result:

- Low Overall Accuracy (58%): The model is only correct about 58% of the time.
- Excellent for Class 0, Poor for Class 1:
- Class 0: Very high precision (99%) means when it predicts class 0, it's almost always right. Recall is moderate (58%), meaning it misses some actual class 0 instances.
- Class 1: Very low precision (4%) means most of its class 1 predictions are wrong. High recall (78%) means it captures a good portion of the actual class 1 instances, but with many false positives.
- Very Low F1-Score for Class 1 (0.08)

```
Accuracy: 0.5849056603773585

Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.58      0.73       362
           1       0.04      0.78      0.08         9

    accuracy                           0.58       371
   macro avg       0.52      0.68      0.41       371
weighted avg       0.97      0.58      0.72       371


Confusion Matrix:
 [[210 152]
 [  2   7]]
```

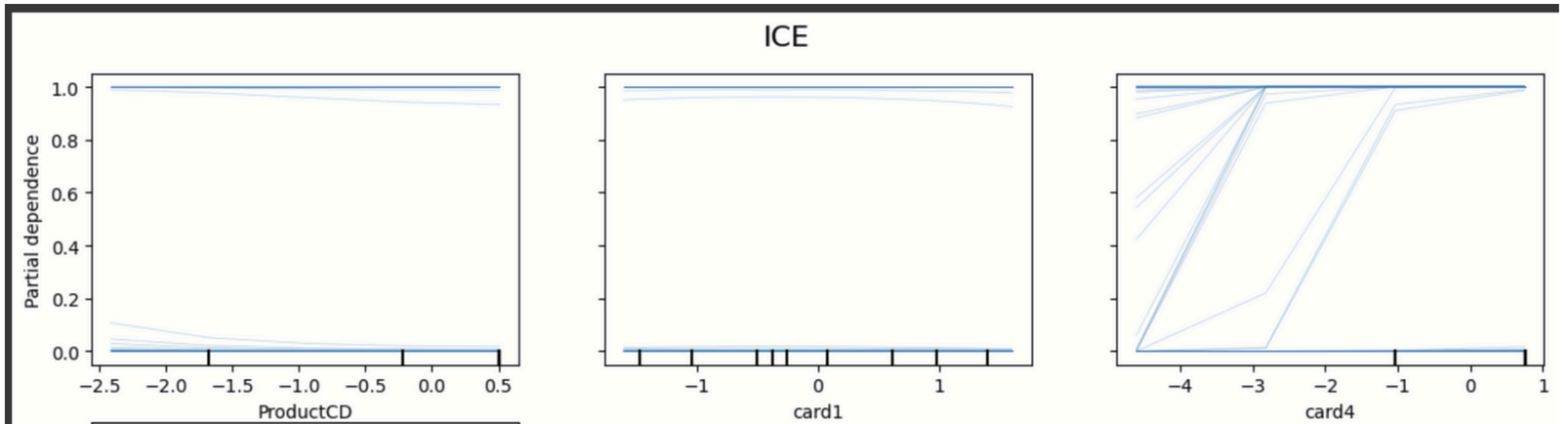
Navie Bayes Model
XAI tool 1: PDP

- ProductCD Minimal Impact: The predicted outcome shows very little change across different values of ProductCD.
- card1 Minimal Impact: Similarly, card1 has a negligible effect on the predicted outcome.
- card4 Increasing Influence: As card4 increases, the predicted outcome shows a noticeable upward trend, suggesting a positive relationship.
- Feature Importance: These plots indicate that card4 has a more substantial influence on the model's predictions compared to ProductCD and card1.



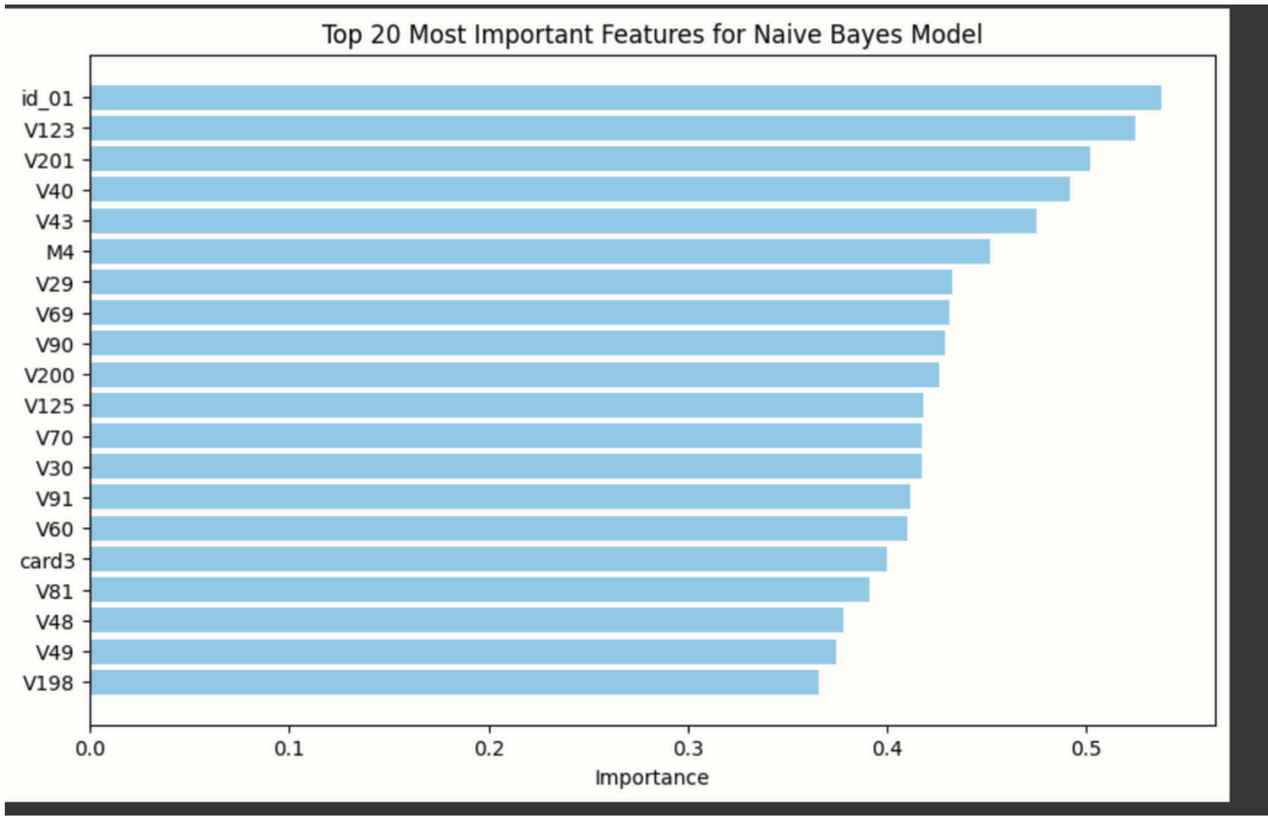
Navie Bayes Model
XAI tool 2: ICE

- ProductCD Stable Predictions: For most instances, the predicted outcome remains consistently high or low regardless of the ProductCD value.
- card1 Mostly Stable Predictions: Similar to ProductCD, the predictions for most instances are not significantly affected by changes in card1.
- card4 Heterogeneous Impact: The effect of card4 varies significantly across instances. For some, the prediction jumps sharply at a specific card4 value, while for others, it remains relatively stable.



Navie Bayes Model
XAI tool 3: PFI

- id_01 is Most Important: The feature id_01 caused the biggest drop in model performance when its values were shuffled, indicating it's the most important feature.
- Top Influential Features: id_01, V123, V201, V40, and V43 are the top 5 most important features for the Naive Bayes model.
- Decreasing Importance: The importance of features generally decreases as we go down the list.
- Naive Bayes Insights: This highlights which variables are most strongly influencing the Naive Bayes classifier's decisions.



Navie Bayes Model
XAI tool 4: LIME

- Strong Prediction for Class 0 : The model is absolutely certain this instance belongs to Class 0.
- Features Pushing Towards Class 1 (Orange): No features in the top contributors are pushing the prediction towards Class 1.
- Features Strongly Pushing Towards Class 0 (Blue): Features like V120, V112, V20, and V317 with their negative values strongly support the prediction of Class 0.

Conclusion

- Decision Tree: Highly interpretable due to its tree structure, allows to follow the decision path based on feature values. The Partial Dependence Plots also clearly show how individual features influence predictions based on sharp splits.
- Logistic Regression: Moderately interpretable. Permutation Feature Importance identifies important features.
- Naive Bayes: Less directly interpretable than Decision Trees or Logistic Regression. While Permutation Feature Importance shows feature importance, understanding the exact relationship and decision boundaries is more complex due to its probabilistic nature.
- XGBoost: More interpretable than Random Forest.
- Random Forest: Less interpretable than a single Decision Tree and generally slightly less interpretable than XGBoost

Conclusion

In conclusion, for the best overall interpretability of this dataset, the Decision Tree model likely stands out. Its structure is easy to visualize and understand, and the Partial Dependence Plots clearly illustrate the impact of individual features through its distinct decision boundaries. While SHAP provides excellent local interpretability, the Decision Tree offers a more readily global understanding of how the model makes decisions based on feature values.

Thank You