## **Documentation: Fraud Detection Model Evaluation**

### **Overview**

In this document, we outline the preprocessing steps and the evaluation of different classification models used for fraud detection. The models evaluated include Random Forest, Decision Tree, and Artificial Neural Network (ANN). We also highlight the precision values obtained from these models and discuss the significance of features like fraud count and no fraud count.

### **Preprocessing Steps**

#### **1. Gender Preprocessing**

* **Original Data:** The gender column contained values such as 'M' and 'F'.
* **Transformation:** Converted 'M' to 1 and 'F' to 0. Removed any other values that were not 'M' or 'F'.

#### **2. Age Preprocessing**

* **Original Data:** The age column had mixed types, including non-numeric characters.
* **Transformation:** Extracted numeric values, converted them to integers, and removed non-numeric entries.

#### **3. Category Label Encoding**

* **Original Data:** The category column contained categorical text data.
* **Transformation:** Applied Label Encoding to convert categories into numeric values.

#### **4. Fraud and No Fraud Count Calculation**

* **Original Data:** Used to compute how many times each customer was involved in fraud versus non-fraudulent activities.
* **Transformation:** Added two new columns: fraud\_contact (count of frauds per customer) and no\_fraud\_contact (count of non-frauds per customer).

#### **5. Merchant and Customer Encoding**

* **Original Data:** The merchant and customer columns contained categorical text data.
* **Transformation:** Applied Label Encoding to convert these categorical variables into numeric values.

#### **6. Amount Transformation**

* **Original Data:** The amount column was in raw form.
* **Transformation:** Applied a logarithmic transformation (amount\_log) to handle skewed data. Dropped the original amount column along with customer and merchant.

#### **7. Standardization**

* **Transformation:** Applied StandardScaler to standardize the features, ensuring they have zero mean and unit variance.

### **Model Evaluation**

#### **Random Forest**

* **Precision:**
  + Class 0: **1.00**
  + Class 1: **0.92**
* **Recall:**
  + Class 0: **1.00**
  + Class 1: **0.83**
* **F1-Score:**
  + Class 0: **1.00**
  + Class 1: **0.87**

**Confusion Matrix:**[[151819 130]

[ 321 1521]]

#### **Decision Tree**

* **Precision:**
  + Class 0: **1.00**
  + Class 1: **0.81**
* **Recall:**
  + Class 0: **1.00**
  + Class 1: **0.82**
* **F1-Score:**
  + Class 0: **1.00**
  + Class 1: **0.81**

**Confusion Matrix:**[[151602 347]

[ 338 1504]]

* **Best Parameters:**
  + Criterion: gini
  + Max Depth: 10
  + Min Samples Split: 10

#### **ANN**

* **Precision:**
  + Class 0: **1.00**
  + Class 1: **0.89**
* **Recall:**
  + Class 0: **1.00**
  + Class 1: **0.74**
* **F1-Score:**
  + Class 0: **1.00**
  + Class 1: **0.81**

**Confusion Matrix:**  
[[151776 173]

[ 478 1364]]

### **Insights on Feature Importance**

The preprocessing and feature selection steps revealed the following key insights:

1. **Fraud Count and No Fraud Count:**
   * These features are critical in distinguishing between fraudulent and non-fraudulent transactions.
   * The significant role of these features is evident from their correlation with fraud detection. Higher fraud contact counts strongly indicate fraudulent behavior, while higher non-fraud contact counts suggest legitimate transactions.
2. **Amount and Merchant Encoding:**
   * The amount\_log transformation proved essential in reducing skewness and improving model performance.
   * Encoding the merchant and customer columns facilitated better model learning by converting categorical data into numerical form, which is crucial for models like Random Forest and ANN.

### **Conclusion**

Among the models evaluated, **Random Forest** demonstrated the highest precision for detecting fraud, especially with its precision for class 1 (fraud) at **0.92**, indicating its effectiveness in minimizing false positives while accurately identifying fraud. The models also showed high recall and F1-scores, validating their ability to detect fraud effectively.

The preprocessing steps, especially the calculation of fraud and non-fraud counts, and feature transformations like amount\_log, significantly contributed to the improved performance of the models. These features, along with effective encoding and scaling, ensured that the models could accurately classify fraudulent transactions.

This comprehensive approach to data preprocessing and model evaluation highlights the importance of feature engineering and appropriate model selection in achieving accurate and reliable fraud detection.