# Report on Credit Card Fraud Detection Dataset

## 1. Overview of the Dataset

The credit card fraud detection dataset contains 284,807 transactions, of which only 492 (0.172%) are fraudulent. The dataset is highly imbalanced, with 30 PCA-transformed features, along with `Time`, `Amount`, and `Class` as additional variables. The target variable, `Class`, is binary (0 = Genuine, 1 = Fraud). The data was collected over two days from European cardholders and is anonymized to protect privacy.  
  
 - Highly imbalanced: Fraud cases constitute a very small fraction of the dataset.  
 - Time feature: Represents seconds since the first transaction.  
 - Feature anonymization: Original features were transformed via PCA.  
 - No missing data: Ensures model readiness without preprocessing for null values.

## 2. Business Understanding

Credit card fraud poses significant financial risks and undermines consumer trust. Detecting fraud in real-time is critical for minimizing losses. The goal of this analysis is to build a machine learning model that accurately identifies fraudulent transactions while minimizing false positives, which can unnecessarily block legitimate user activity.  
Challenges due to the dataset:  
- \*\*Imbalanced data\*\*: Standard accuracy metrics are not meaningful due to class imbalance.  
- \*\*High-stakes predictions\*\*: Missing a fraud case is costly, and false positives can disrupt user experience.

## 3. Exploratory Data Analysis (EDA) and FE

Key Insights:

- Fraudulent transactions tend to have smaller average transaction amounts compared to genuine ones.  
- The `Time` variable does not show strong correlation with fraud occurrence.  
- Certain PCA-transformed features exhibit distinct value ranges for fraudulent and genuine transactions, which could be used for classification.

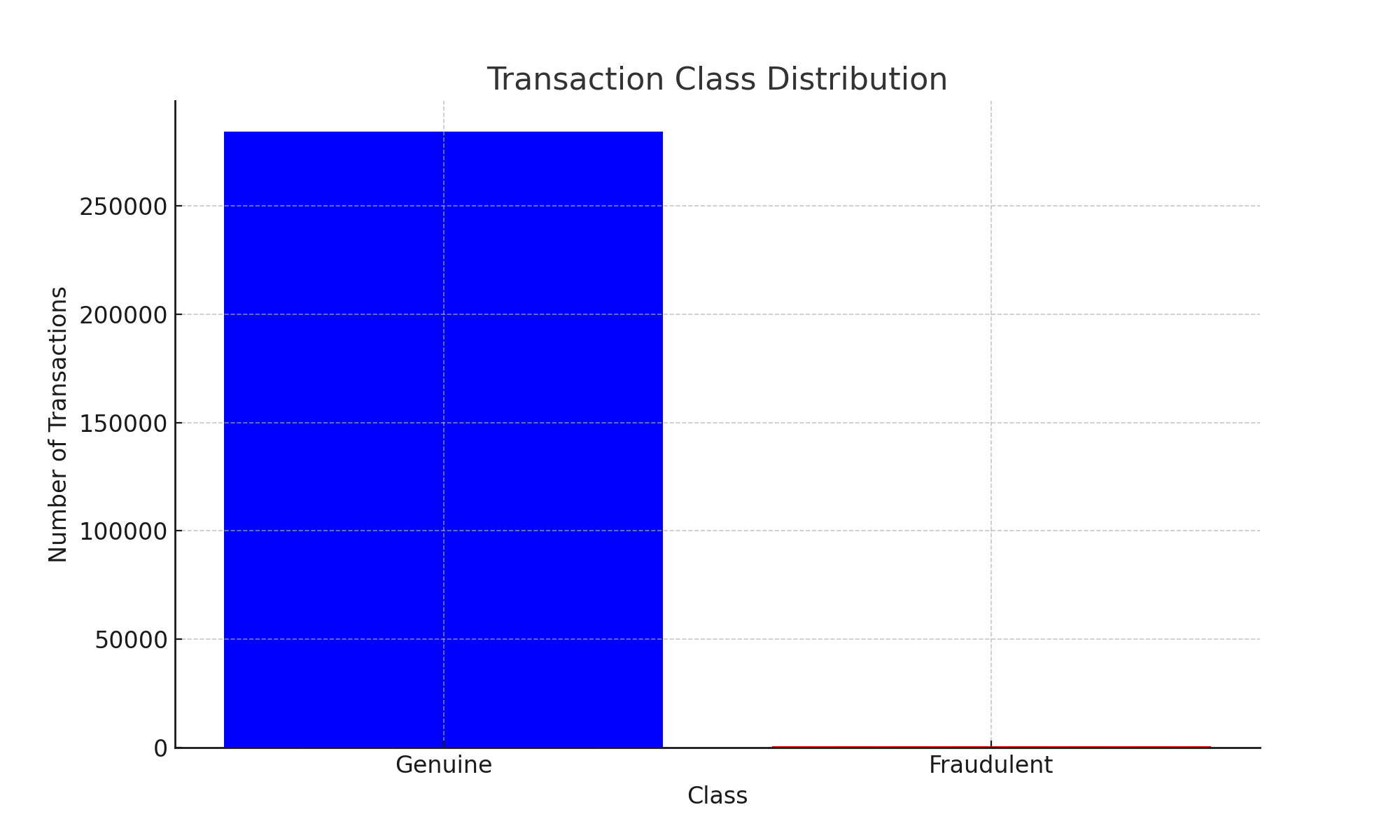
A screenshot of a graph

Description automatically generatedA screenshot of a computer

Description automatically generated

A graph of a graph

Description automatically generated with medium confidence



**Data Splitting**:

* The data is initially split into training-validation (X\_train\_val, y\_train\_val) and testing sets (X\_test, y\_test) using a **70-30 split**.
* Further, the training-validation set is split into training (X\_train, y\_train) and validation sets (X\_validate, y\_validate) using a **20% split** of the training-validation set.

**Feature Scaling**:

* A scaler is applied to standardize the features across the training, validation, and testing datasets.
* This ensures that all features have the same scale, improving the performance of machine learning models.

## 4. Machine Learning Models

Four machine learning models were evaluated on their ability to classify fraudulent transactions. Evaluation focused on the \*\*fraudulent class\*\*, as detecting fraud accurately is the primary objective.

**Model Metrics for Legit Class:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| RandomForestClassifier | 99.96 | 0.999719 | 0.999883 | 0.999801 |
| CatBoostClassifier | 99.96 | 0.999707 | 0.999906 | 0.999807 |
| LGBMClassifier | 99.5 | 0.999247 | 0.995721 | 0.997481 |
| XGBoost | 99.96 | 0.999707 | 0.99993 | 0.999818 |
| **Fraudulent Class:** |  |  |  |  |
| Model | Accuracy | Precision | Recall | F1-Score |
| RandomForestClassifier | 99.96 | 0.918033 | 0.823529 | 0.868217 |
| CatBoostClassifier | 99.96 | 0.932773 | 0.816176 | 0.870588 |
| LGBMClassifier | 99.5 | 0.16476 | 0.529412 | 0.251309 |
| XGBoost | 99.96 | 0.948718 | 0.816176 | 0.87747 |

**Most Accurate Model for Each Metric**

* **Precision (Fraudulent Class): XGBoost (0.948718)**
  + Indicates the model is the most accurate in predicting fraud without significant false positives.
* **Recall (Fraudulent Class): RandomForestClassifier (0.823529)**
  + Shows it captures more actual fraud cases than others.
* **F1-Score (Fraudulent Class): XGBoost (0.87747)**
  + Balances precision and recall effectively.

**Final Model**

* **XGBoost: Based on its strong balance across all metrics, especially in terms of F1-Score, it is the most effective for detecting fraud.**

**Model Analysis**

1. **RandomForestClassifier**:
   * Strengths: High recall ensures most fraudulent cases are detected.
   * Weaknesses: Lower precision indicates more false positives compared to XGBoost.
2. **CatBoostClassifier**:
   * Strengths: Handles imbalanced data natively; competitive precision and F1-score.
   * Weaknesses: Slightly lower recall compared to Random Forest.
3. **LGBMClassifier**:
   * Strengths: Efficient for large datasets; computationally fast.
   * Weaknesses: Fails to capture the minority class effectively, leading to poor recall and F1-score.
4. **XGBoost**:
   * Strengths: Best F1-score and precision, indicating a strong balance between detecting fraud and minimizing false positives.
   * Weaknesses: Computationally expensive compared to LGBM and Random Forest.

## 5. Conclusion

The \*\*XGBoost model\*\* is the most effective for detecting fraud in this dataset. It achieves the highest F1-score (0.87747), indicating a strong balance between precision and recall for the fraudulent class. While Random Forest and CatBoost also performed well, their slightly lower F1-scores make XGBoost the preferred model.