**CREDIT CARD FRAUD DETECTION**

We have to build a classification model to predict whether a transaction is fraudulent or not.

1. **Problem Statement**

A credit card is one of the most used financial products to make online purchases and payments. Though the Credit cards can be a convenient way to manage your finances, they can also be risky. Credit card fraud is the unauthorized use of someone else's credit card or credit card information to make purchases or withdraw cash.

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

1. **Dataset Description**

**The dataset used in this project is creditcard.csv, which contains the following columns:**

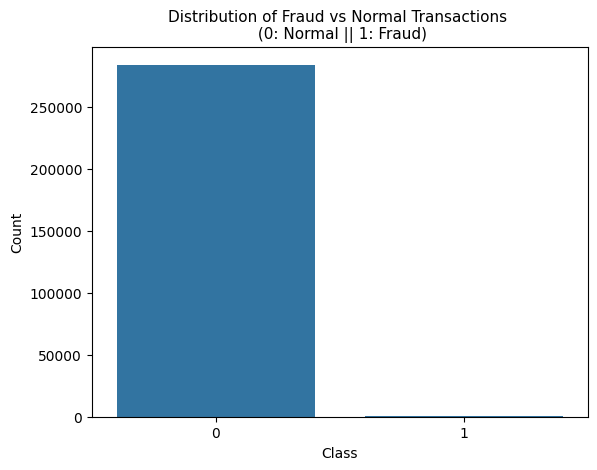
* **Time**: The time elapsed since the first transaction in the dataset.
* **V1, V2, …, V28**: PCA-transformed features representing transaction data.
* **Amount**: The amount of the transaction.
* **Class**: The class label where:
* 0 indicates a normal transaction.
* 1 indicates a fraud transaction.

1. **Model Implementation**
2. **Exploratory Data Analysis**: Analyse and understand the data to identify patterns, relationships, and trends in the data by using Descriptive Statistics and Visualizations.
3. **Data Cleaning**: This might include standardization, handling the missing values and outliers in the data.
4. **Dealing with Imbalanced data**: This data set is highly imbalanced. The data should be balanced using the appropriate methods before moving onto model building.
5. **Feature Engineering:** Create new features or transform the existing features for better performance of the ML Models.
6. **Model Selection**: Choose the most appropriate model that can be used for this project.
7. **Model Training:** Split the data into train & test sets and use the train set to estimate the best model parameters.
8. **Model Validation:** Evaluate the performance of the model on data that was not used during the training process. The goal is to estimate the model's ability to generalize to new, unseen data and to identify any issues with the model, such as overfitting.
9. **Model Deployment**: Model deployment is the process of making a trained machine learning model available for use in a production environment.
10. **Data Insights**

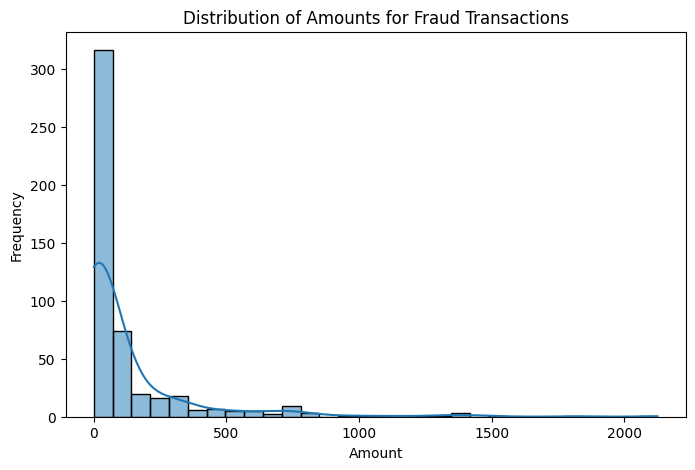
The creditcard.csv dataset contains details of 284,807 transactions out of which 492 are fraudulent transactions.

The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

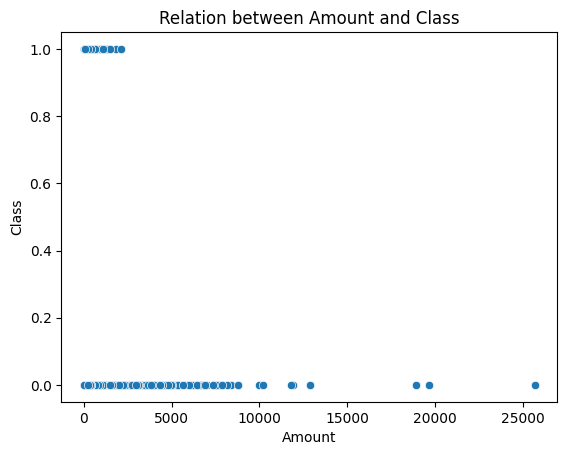
* Data has 31 features from V1-V28 & Time, Amount and Class
* Input features: V1-V28, Time and Amount
* **Target variable: Class**
  1. **Transaction Distribution by class**



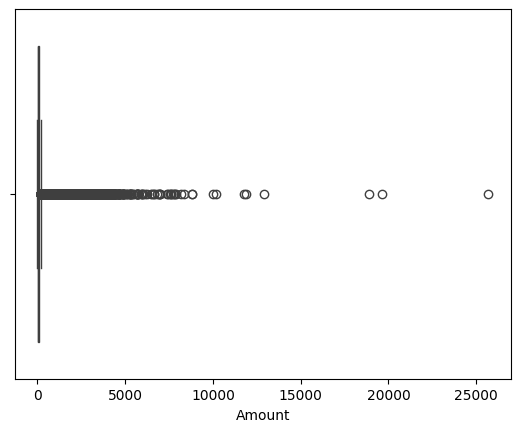
* 1. **Distribution of amount for "Fraud" Transactions**
* **High Frequency of Small Amounts:** Most fraud transactions involve smaller amounts, as indicated by the highest bar (just above 300) for amounts close to zero.
* **Decreasing Frequency with Higher Amounts:** As the transaction amount increases, the frequency of fraud transactions decreases.



* The fraud transaction amounts seem less than 2500.
  1. **Relation between Amount and Class**

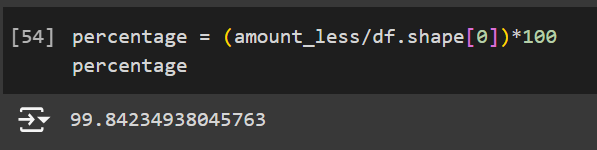


As we can predict from the above scatterplot, the Amount variable has some outliers.



* 1. **Fraud Percentage**

Approximately 0.168% of transactions were identified as fraudulent, indicating a low overall fraud rate.



1. **Model Used**

We have used pycaret module to identify the best model for this project. The following are the 5 promising models based on accuracy metric:

* 1. Extreme Gradient Boosting
  2. Random Forest Classifier
  3. Extra Trees Classifier
  4. Logistic Regression
  5. Light Gradient Boosting Machine

1. **Outcome**

Random Forest Classifier turns out to be the best fit for this model (among these 5 models). With high test accuracy (99.96%), it also has high precision (97%) and good Recall (80%) and F1 – Score (88%).

Uploaded a detailed outcome report (Credit card fraud report.pdf) for the same.

1. **Future Work**

In future, this work will be extended to hyperparameter tuning for further boosting the performance of the model. The additional features based on domain knowledge and different data techniques to overcome class imbalance will also be explored. We will try to test this model on different datasets as well to see how it works on unseen data. Finally, we will focus on enhancing model interpretability to better understand the decision-making process behind predictions.

1. **Conclusion**

The project involved developing models as mentioned earlier and addressing the imbalance in the data using techniques such as SMOTE and feature scaling. The results indicated significant improvement over the baseline models in terms of accuracy, precision, and recall, demonstrating the effectiveness of our approach. However, the current study has limitations. The overall performance of the models may be influenced by data quality, data collection, data size, and modelling, so further research is necessary. Overall, this project highlights the importance of reducing class imbalance and improving model performance in predictive analytics, laying the foundation for robust and dependable applications in future research.