Team Runtime Terror
ML Model Design

Segregation of Data:

- We test out data, column by column and row by row. This is pin-point our problem in the dataset.
- This is achieved through narrowing down the amount on isFraud and isFraudFlagged.
- Once this is done, we separate out these rows from the entire DB to have better focus on them.

Sampling of Data:

- Heavily imbalanced data is to bias machine learning. We need to use traditional way to process data, i.e. under-sampling method, over-sampling, SMOTE
- We need to over-sample the fraud transactions, or under-sample the clean ones to level the playing field for our models. We can do this using the imbalance-learn library.

Splitting of Data:

- We split the dataset into **Training sets** and **Testing sets**.
- This is achieved through **sklearn.model_selection import train_test_split.** With test_size = 0.2 (20% validation data and 80% training data.), random_state = 42(produce the same results across a different run).
- The reason we do it this way is that our model should not just give us good results with a part of the training dataset, but should also provide good results with data that we have never seen before**.**

Logistic regression:

- We use a simple **linear logistic regression classification model**.
- It is the go-to method for binary classification problems (problems with two class values). In this post you will discover the logistic regression algorithm for machine learning. It is quite simple to implement in comparison to other models.
- Through sklearn, we obtain- **model predictions**, **classification report** and **confusion matrix**, **accuracy score** etc.