****

**Aim of the Project:**

The problem to be solved with data mining is online payment fraud detection. The goal is to develop a machine learning model that can accurately identify fraudulent transactions in online payments.

**Benefits of Payment Fraud Detection:**

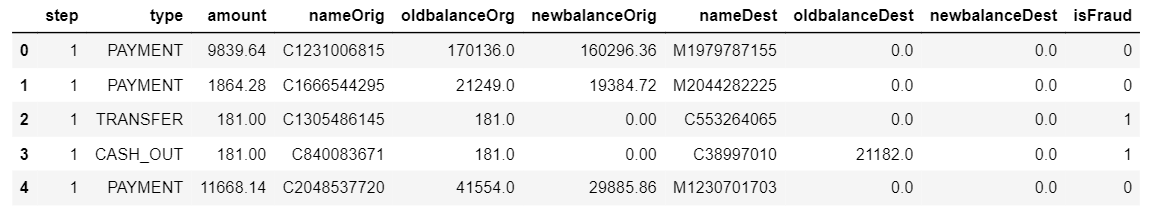
By detecting and preventing fraudulent transactions, the business can benefit in several ways:

* Minimize financial losses: Online payment fraud can result in significant financial losses for businesses. By accurately identifying and preventing fraudulent transactions, the business can minimize the financial impact of fraudulent activities, reducing potential revenue losses.
* Enhance customer trust: Fraudulent transactions can undermine customer trust in the business's online payment system. By implementing an effective fraud detection solution, the business can demonstrate its commitment to safeguarding customer transactions and enhance customer trust and confidence in using their payment platform.
* Protect reputation: Instances of online payment fraud can harm the business's reputation and brand image. By proactively detecting and preventing fraud, the business can maintain a positive reputation, reassuring customers that their transactions are secure and protecting the brand's integrity.
* Improve operational efficiency: Dealing with fraudulent transactions can be time-consuming and resource-intensive for businesses. By automating the fraud detection process with a machine learning model, the business can streamline its operations, reduce manual efforts, and allocate resources more efficiently.
* Regulatory compliance: Compliance with regulations and legal requirements is essential for businesses operating in the online payment industry. Implementing a robust fraud detection system helps the business adhere to regulatory guidelines and demonstrate compliance, ensuring a secure and trustworthy payment environment for customers.

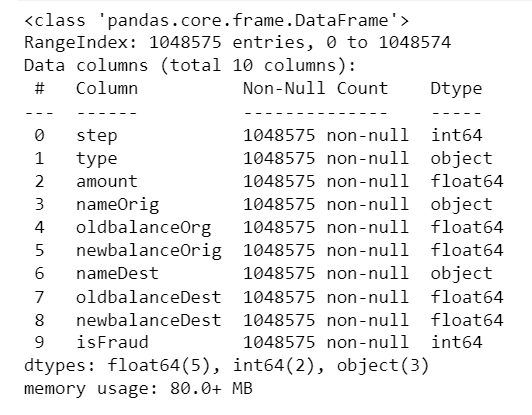
Overall, an effective fraud detection solution powered by data mining can help the business safeguard its financial interests, maintain customer trust, protect its reputation, improve operational efficiency, and ensure compliance with regulatory standards.

**First Step:**

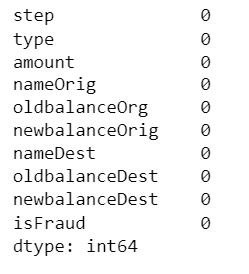
* Read in the data



* Check for information about the dataset

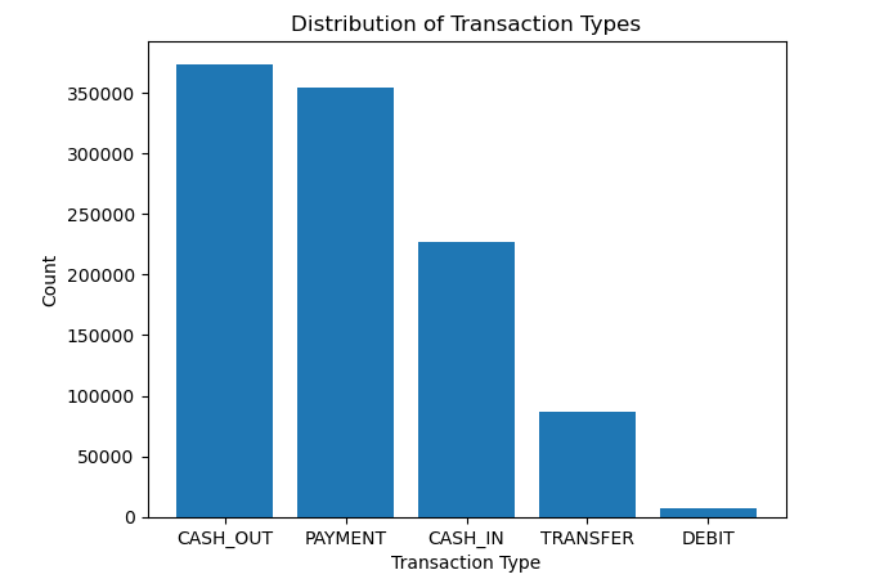


* Check for null values : No null values

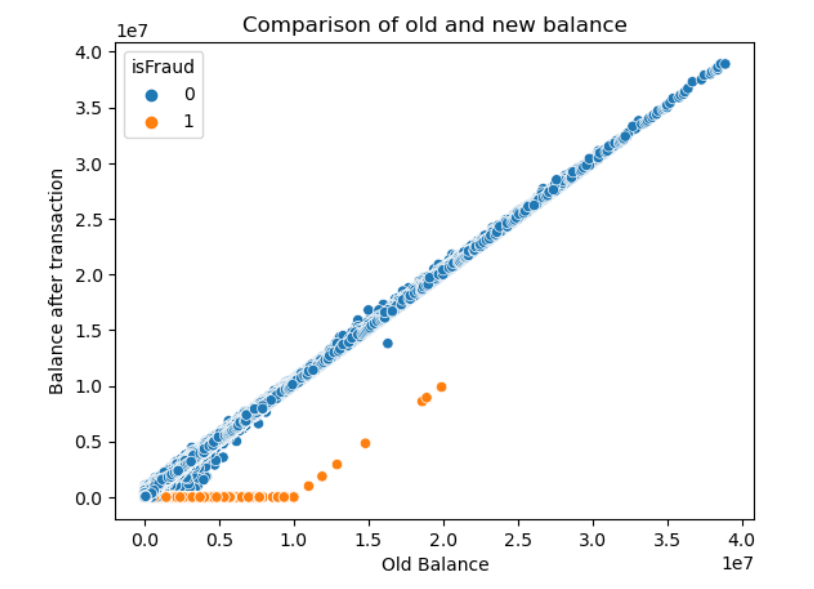


**Exploratory Data Analysis**

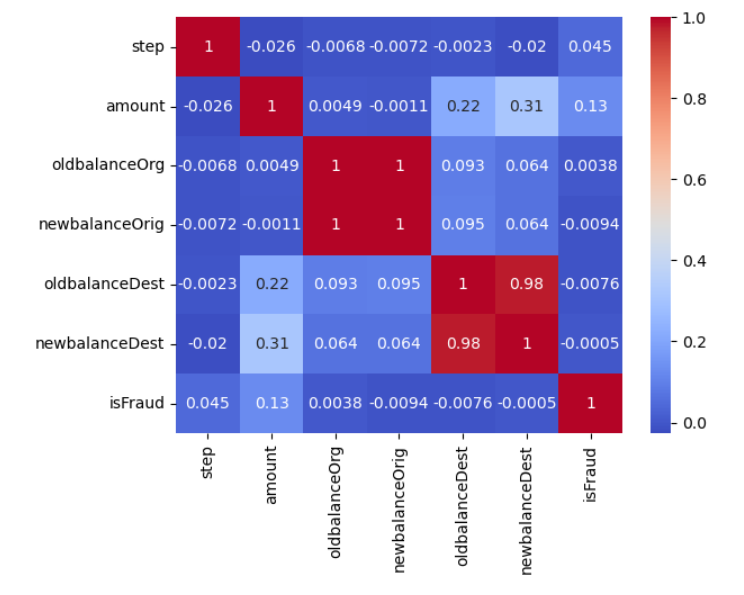
**Univariate Analysis**



**Bivariate Analysis**

****

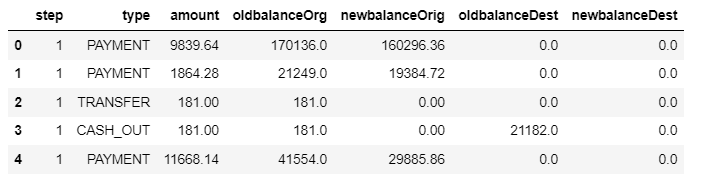
**Correlation Matrix**

****

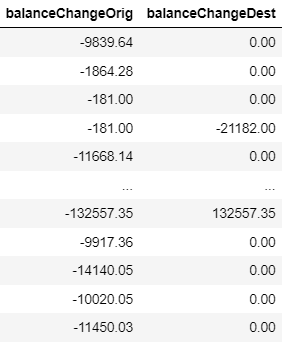
**Data Preprocessing Steps:**

* **Feature engineering**

1. Separated the target variable from the dataset and dropped redundant variables. The target variable ‘isFraud’ was dropped alongside ‘nameOrig’ and ‘nameDest’.

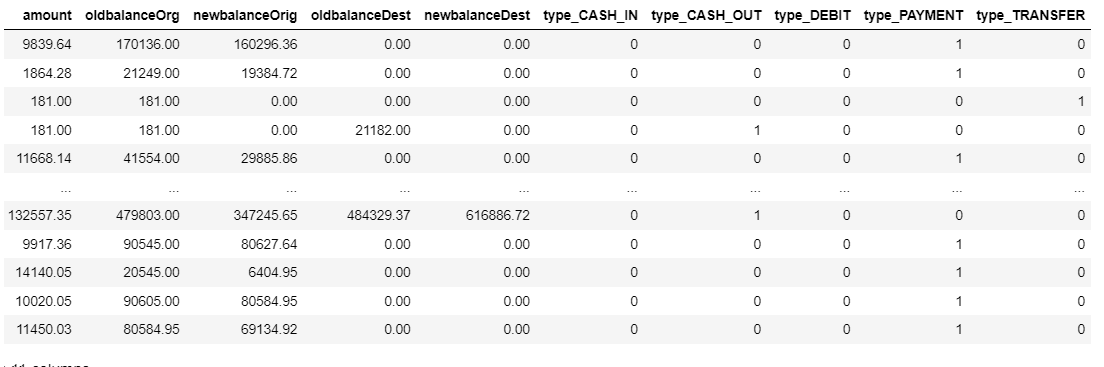


1. Created a new column to show the differences between new balance and old balance



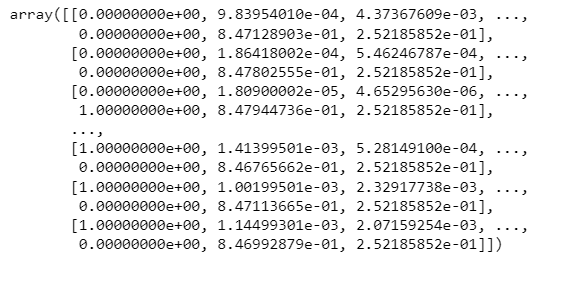
* **Encoding**

The One-hot encoding Style was implemented to change categorical variables into numerical variables. The ‘type’ column has been encoded as shown below



* **Scaling**

The dataset was scaled using MinMaxScaler to deal with high variability or high variance in a dataset. This will compress the data in the dataset while still maintaining the information in the data set.



**Model Building**

* The dataset is split in train and test. The test\_size=0.2 parameter specifies that 20% of the data will be used for testing, while the remaining 80% will be used for training. The random\_state=42 parameter sets a specific random seed to ensure reproducibility.
* First Model Implemented: **Logistics Regression Model**

1. I created a LogisticRegression object and instantiated it. Then I fit the model on top of subset of the dataset kept for training
2. Next step is to predict on the test dataset and pass the subset of the dataset kept for testing

**Model Evaluation (Performance Metrics)**

1. Area Under the ROC Curve (AUC-ROC): 0.5781226132626223
2. Accuracy score: 0.9990940085353933
3. F1 Score: 0.26923076923076916
4. Precision: 0.9722222222222222
5. Recall Score: 0.15625

* Second Model Implemented: **Random Forest Regressor Model**

**Model Evaluation (Performance Metrics)**

**R2 Score: 0.7777093013690195**

* Third Model Implemented: **Decision Tree Regressor Model**

**R2 Score: 0.6776032968314205**

* Fourth Model Implemented: **Decision Tree Classifier**

**Model Evaluation (Performance Metrics)**

1. Area Under the ROC Curve (AUC-ROC): 0.9284735723389413
2. Accuracy score: 0.9996519085425458
3. F1 Score: 0.8402625820568927

**More Experimental Models**

* **Random Forest Classifier**

Accuracy Score: 0.999766349569654

* **Support Vector Machine (SVM)**

Accuracy Score: 0.9991321555444294

* **XGBoost (Extreme Gradient Boosting)**

Accuracy Score: 0.9998378752115967

* **LightGBM classifier**

Accuracy Score: 0.9985265717759817

### Confusion Matrix (Evaluation of the Predictions)

