Credit Card Fraud Detection using Data Science

2.1 Short Explanation about the Problem Statement:

Credit card fraud poses a significant threat to both financial institutions and cardholders. The primary challenge is to distinguish fraudulent transactions from legitimate ones within a vast dataset of transactions. The goal is to create a machine learning model capable of real-time fraud detection by leveraging data science techniques.

2.2 Dataset source and details:

We have collected the dataset containing legitimate and fraud transactions from the duration 1st Jan 2019 - 31st Dec 2020. It covers credit cards of 1000 customers doing transactions with a pool of 800 merchants from Kaggle. This dataset contains the sequence of transaction information like

trans_date_trans_time, cc_num, merchant, category, amt, firstname, lastname, gender

street, city, state, zip, latitude, longitude, city_pop, job, trans_num, unix_time, merch_latitude,

merch_longitude is_fraud.

https://www.kaggle.com/datasets/kartik2112/fraud-detection

2.3 Columns details:

trans_date_trans_time => it consist of transaction date and time

merchant => it contains merchant name to whom the transaction is being processed.

Category => it consists of a genre of transaction like recharge shopping etc..,

Amt => amount of transaction.

And other parameters like firstname, lastname, gender

street, city, state, zip, latitude, longitude, city_pop, job, trans_num, unix_time, merch_latitude,

merch_longitude is_fraud,cc_num.

2.4 Details of Libraries Needed for Data Prediction and How to Download:

To implement the credit card fraud detection solution, the following Python libraries are essential:

Required Libraries:

1. **Pandas**: For data manipulation and analysis.
2. **NumPy**: For numerical operations.
3. **Scikit-Learn**: Provides machine learning algorithms and evaluation metrics.
4. **Matplotlib and Seaborn**: For data visualization.

5. **Imbalanced-Learn**: Useful for addressing class imbalance.
6. **TensorFlow**: For building neural network models.

How to Download:

You can download these libraries using Python's package manager, pip, by executing the following commands:

pip install pandas numpy scikit-learn matplotlib seaborn imbalanced-learn tensorflow

2.5 How to Train and Test

The process of training and testing a credit card fraud detection model involves the following steps:

1.Data Splitting: Divide your dataset into two parts, typically a training set (e.g., 70-80% of the data) and a testing set (20-30% of the data). You can use Scikit-Learn's `train_test_split` function for this purpose.

2.Model Selection:

The choice of a machine learning algorithm or neural network architecture depends on the nature of the dataset and the complexity of the fraud detection problem. The following dataset parameters influence model selection:

- **Transaction Amount (`amt`)**: The transaction amount is a crucial parameter that may have a significant impact on model selection. Algorithms like gradient boosting or decision trees may handle this parameter well.
- **Merchant Information (`merchant`)**: The category or type of merchant can be a relevant feature for the model, influencing the choice of algorithms.
- **Transaction Date and Time (`trans_date_trans_time`)**: Temporal patterns may be important, affecting whether time series models or algorithms like recurrent neural networks (RNNs) are considered.

3.**Model Training**:

During model training, the dataset parameters help build the model's understanding of fraudulent and genuine transactions. These parameters play a role in feature engineering and data preparation:

- **Transaction Amount (`amt`)**: This parameter may be normalized or scaled to ensure it aligns with other features.
- **Merchant Information (`merchant`)**: One-hot encoding or label encoding may be applied to handle categorical data, depending on the choice of algorithm.
- **Transaction Date and Time (`trans_date_trans_time`)**: Temporal features like time of day, day of the week, or holiday indicators can be engineered from this parameter.

4.**Model Testing and Evaluation**:

The effectiveness of the model in accurately identifying fraudulent transactions is evaluated using various metrics. The dataset parameters influence the model's ability to make predictions:

- **Transaction Amount (`amt`)**: This parameter can directly impact precision and recall, as fraud detection often involves setting thresholds for transaction amounts.
- **Merchant Information (`merchant`)**: The type of merchant may influence the accuracy of the model, especially if certain merchant categories are more prone to fraud.
- **Transaction Date and Time (`trans_date_trans_time`)**: Temporal patterns can significantly impact the ROC-AUC score, as different times of day or days of the week may exhibit varying fraud rates.

These metrics and dataset parameters collectively determine the model's effectiveness in accurately identifying fraudulent transactions while minimizing false positives. It's essential to analyze how these parameters interact with the chosen algorithm during both model selection and training to optimize the detection system.

2.6 Explanation of the Solution

The solution to the credit card fraud detection problem involves the creation of a predictive model that, when given a new transaction, can classify it as either genuine or fraudulent. This is achieved through a machine learning model or neural network that has been trained on a dataset containing historical transaction data.

2.7 Metrics Used for the Accuracy Check

The accuracy of the credit card fraud detection model is assessed using the following evaluation metrics:

1. **Accuracy**: The proportion of correctly classified transactions (both fraudulent and genuine) out of the total.
2. **Precision**: The ratio of true positive predictions (correctly identified fraudulent transactions) to all positive predictions.
3. **Recall**: The ratio of true positive predictions to all actual fraudulent transactions.
4. **F1-Score**: A balanced measure of precision and recall, computed as 2 * (precision * recall) / (precision + recall).
5. **ROC-AUC (Receiver Operating Characteristic - Area Under the Curve)**: A metric that measures the model's ability to distinguish between genuine and fraudulent transactions across different thresholds.

These metrics collectively assess the effectiveness of the credit card fraud detection model in correctly identifying fraudulent transactions while minimizing false positives.

This document provides an overview of the steps involved in solving the credit card fraud detection problem using data science techniques. It covers libraries required for data prediction, data splitting, model selection, model training and testing, and the evaluation metrics used to gauge the model's accuracy.