Akshay Kumar
Date: -15-09-2022

Fraud Detection using Machine Learning

1. **Problem Statements**
   - Fraud detection is a challenging problem. The fact is that fraudulent transactions are rare and they represent a very small fraction of activity within an organization. The challenge is that a small percentage of activity can quickly turn into big losses.

   - Fraud has been a major issue in sectors like banking, medical, insurance, and many others. Due to the increase in online transactions through different payment options, such as credit/debit cards, PhonePe, Gpay, Paytm, etc., fraudulent activities have also increased. Moreover, fraudsters or criminals have become very skilled in finding escapes so that they can loot more.

   - No system is perfect and there is always a loophole them, it has become a challenging task to make a secure system for authentication and preventing customers from fraud. So, Fraud detection algorithms are very useful for preventing frauds.

2. **Market/Customer Need Assessment**
   - Many people are not well versed with the online payment system because of which the fraudster take the advantage and do online scams.

   - For making trust on the online payment, we have to provide security to the people.

   - Today lots of people in the shopping use cash on delivary method because they don't have trust on the online payment method.

   - By providing the fraud detection services we can get the trust of Customer of online payment.

3. **Target Specification and characterization**
   - To identify normal customer patterns in order to be able to identify abnormal or fraudulent ones.
   - Generating a Casual model corresponding to a user.
   - Predict expected behavior of user during a next event in the account using casual model.
   - Generating fraud event parameters using fraud model.
   - Generating a risk score of next event using expected event parameters and fraud event parameters.

4. **External Search (information sources)**
   - The dataset can be found on the kaggle. Link: -Dataset Link
   - The dataset consists following information that is given below. The screenshot consists top 5 rows of dataset.

```
In [3]: df=pd.read_csv("Fraud.csv")
        df.head()
```

Out[3]:

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0	0
1	1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0	0
2	1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.0	0.0	1	0
3	1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.0	0.0	1	0
4	1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0	0

– Shape of the datasets and all columns' data type are given below in screenshot.

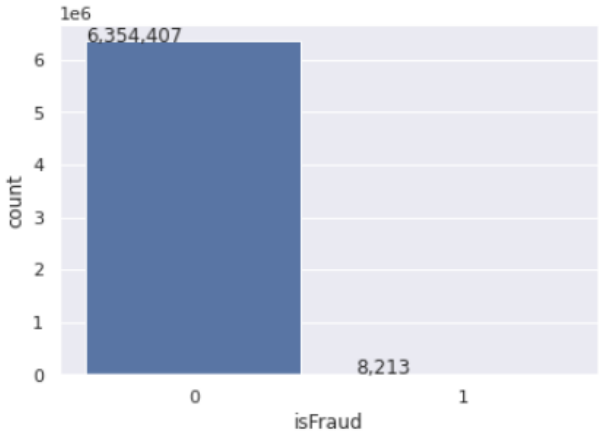
```
In [4]: df.shape
Out[4]: (6362620, 11)
```

Finding Data Types of all Columns

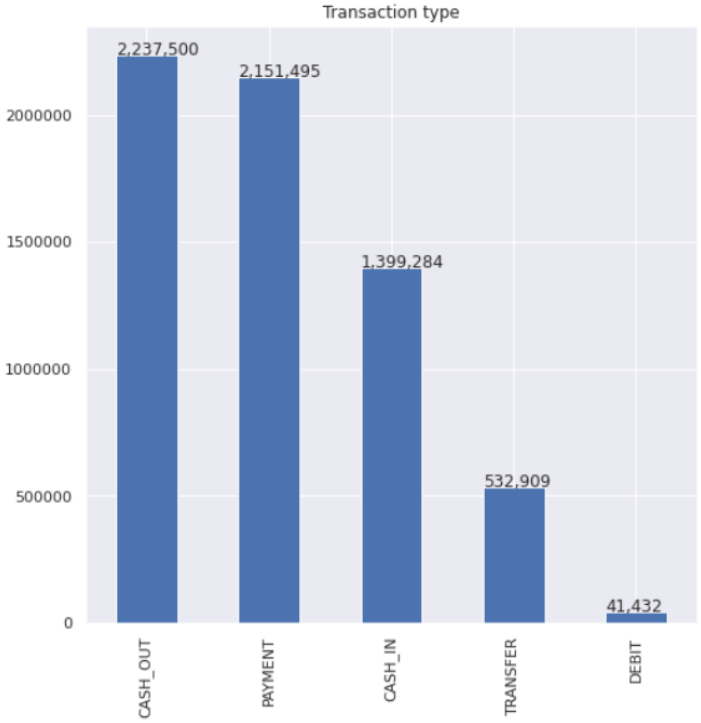
```
In [5]: df.dtypes
Out[5]: step              int64
        type             object
        amount          float64
        nameOrig         object
        oldbalanceOrg   float64
        newbalanceOrig  float64
        nameDest         object
        oldbalanceDest  float64
        newbalanceDest  float64
        isFraud           int64
        isFlaggedFraud    int64
        dtype: object
```

5. Benchmarking

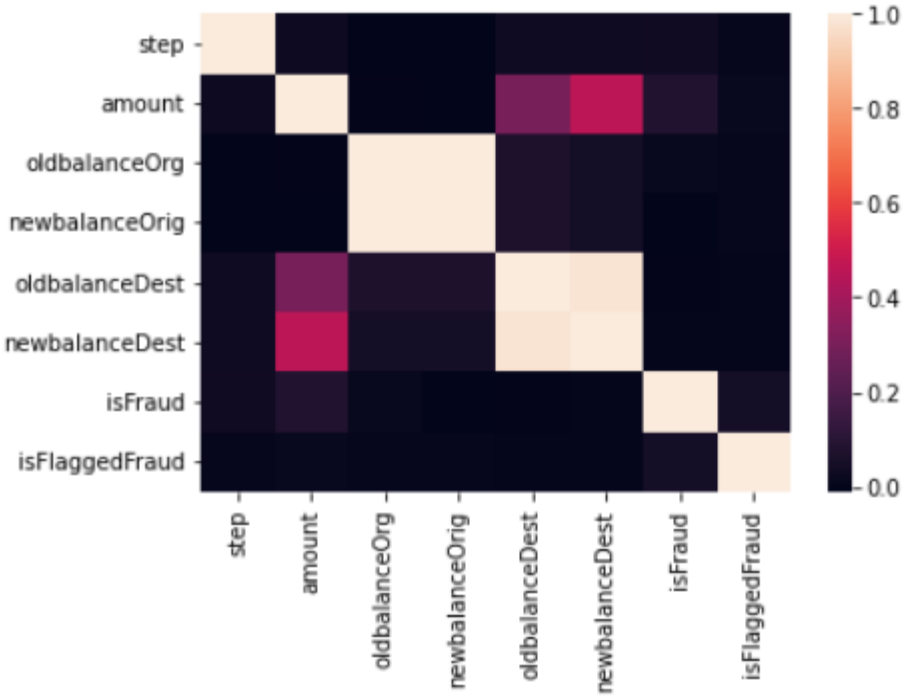
– Total Number of Fraud vs Not Fraud



– Total Number of Transaction happens by each method is given in below graph.

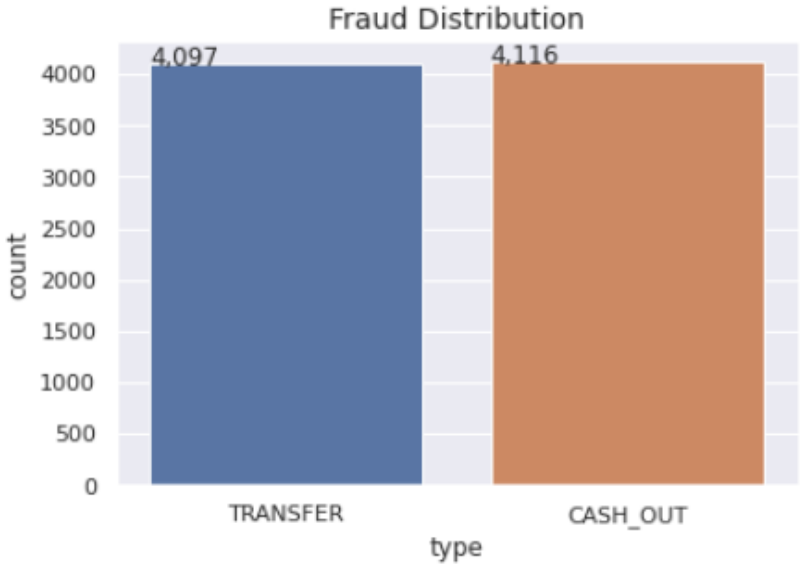


– Payment Correction Heat Map for all Payment methods



–

– Total fraud is happened only by the method of ‘TRANSFER’ and ‘CASH_OUT’. Number of counts is also levels in the graph.



–

6. Applicable Patents

- This is the link of Applicable Patents: - Patent Link
- Patent Application Publication on Jan. 22, 2015: -US 2015/0026027 A1

7. Applicable Regulations

- These regulations may be called the Securities and Exchange Board of India (Prohibition of Fraudulent and Unfair Trade Practices relating to Securities Market) Regulations, 2003.
- Link of the Regulation is given below: -
  PDF link

8. Applicable Constraints

- No person shall directly or indirectly—
  a) buy, sell or otherwise deal in securities in a fraudulent manner;
  b) use or employ, in connection with issue, purchase or sale of any security listed or proposed to be listed in a recognized stock exchange, any manipulative or deceptive device or contrivance in contravention of the provisions of the Act or the rules or the regulations made thereunder;
  c) employ any device, scheme or artifice to defraud in connection with dealing in or issue of securities which are listed or proposed to be listed on a recognized stock exchange;
  d) engage in any act, practice, course of business which operates or would operate as fraud or deceit upon any person in connection with any dealing in or issue of securities which are listed or proposed to be listed on a recognized stock exchange in contravention of the provisions of the Act or the rules and the regulations made thereunder.

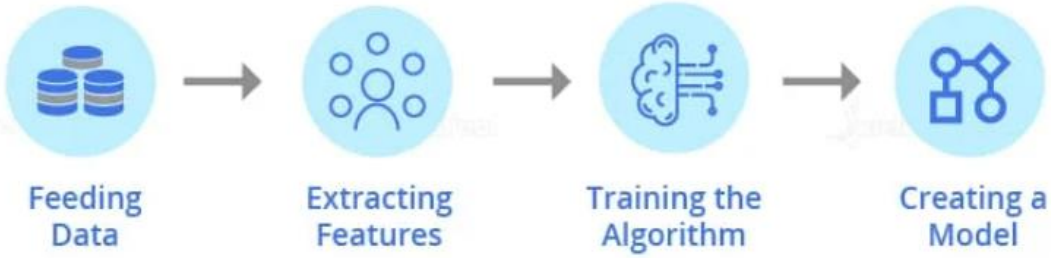
9. Business Opportunity

- Customers want fast, efficient interaction when transacting with companies, but they also want to ensure their identities, payment card number and account information are guarded and protected.

- We can give the service to the company (startup) they are just growing their business and don't have protection and fraud detection system.

- The Federal Trade Commission (FTC) says that fraudulent business opportunities consistently rank in the top 10 categories in its database of consumer fraud complaints.

10. Concept Generation



- **Feeding Data:** First, the data is fed into the model. The accuracy of the model depends on the amount of data on which it is trained, more data better the model performs.

- **Extracting Features:** Feature extraction basically works on extracting the information of each and every thread associated with a transaction process. These can be the location from where the transaction is made, the identity of the customer, the mode of payments, and the network used for transaction.

    1. *Identity***:** This parameter is used to check a customer's email address, mobile number, etc. and it can check the credit score of the bank account if the customer applies for a loan.
    2. *Location***:** It checks the IP address of the customer and the fraud rates at the customer's IP address and shipping address.
    3. *Mode of Payment:* It checks the cards used for the transaction, the name of the cardholder, cards from different countries, and the rates of fraud of the bank account used.
    4. *Network:* It checks for the number of mobile numbers and emails used within a network for the transaction.

- **Training the Algorithm**: Once you have created a fraud detection algorithm, you need to train it by providing customers data so that the fraud detection algorithm learns how to distinguish between 'fraud' and 'genuine' transactions.

Training differents models to choose the best out of them

Importing and Creating models objects:

In [36]:
```
from sklearn import tree
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier

model_descTree = tree.DecisionTreeClassifier()
model_logic=LogisticRegression(solver='lbfgs', max_iter=600)
model_rf=RandomForestClassifier()
```

Using cross_val_score function

In [37]:
```
from sklearn.model_selection import cross_val_score
```

In [38]:
```
cross_val_score(model_descTree, x, y, cv=3)
```

Out[38]:
```
array([0.99925498, 0.99940333, 0.99935569])
```

In [39]:
```
cross_val_score(model_logic, x, y , cv=3)
```

Out[39]:
```
array([0.99429648, 0.99817427, 0.99816452])
```

In [63]:
```
cross_val_score(model_rf, x, y, cv=3)
```

Out[63]:
```
array([0.99915536, 0.99930696, 0.99927339])
```

we got 99.9% accuracy from both random forest and from DecisionTreeClassifier but RandomForestClassifier take loo much time so the our best model is "DecisionTreeClassifier".

- **Creating a Model:** I have trained fraud detection model on a specific dataset.

- The advantage of Machine Learning in fraud detection Model is that it keeps on improving as it is exposed to more data.

In [41]:
```
from sklearn.model_selection import train_test_split
```

In [53]:
```
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.3,stratify=y)
```

In [54]:
```
len(x_train)
```

Out[54]:
```
1939275
```

In [55]:
```
len(y_train)
```

Out[55]:
```
1939275
```

In [56]:
```
len(x_test)
```

Out[56]:
```
831118
```

In [57]:
```
len(y_test)
```

Out[57]:
```
831118
```

In [58]:
```
model_descTree.fit(x_train, y_train)
```

Out[58]:
```
DecisionTreeClassifier()
```

In [59]:
```
y_pred = model_descTree.predict(x_test)
```

In [60]:
```
model_descTree.score(x_test, y_pred)
```

Out[60]:
```
1.0
```

- The model works for detecting ‘fraudulent’ and ‘non-fraudulent’ transactions with the accuracy of approx. 99.9%.

```
In [61]: from sklearn.metrics import classification_report

In [62]: print(classification_report(y_test,y_pred))
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	828659
1	0.91	0.88	0.90	2459
accuracy			1.00	831118
macro avg	0.95	0.94	0.95	831118
weighted avg	1.00	1.00	1.00	831118

finally we have achieved the accuracy of approximately 99.9%.

–

11. Concept Development

– The concept can be developed by using the appropriate API (flask in this case).

–

12. Final Report Prototype

– The product takes the following functions to perfect and provide a good result.

Back-end

– Model Development: This must be done before releasing the service. A lot of manual supervised machine learning must be performed to optimize the automated tasks.

1. Performing EDA to realize the dependent and independent features.
2. Algorithm training and optimization must be done to minimize overfitting of the model and hyperparameter tuning.

Front End

1. Different user interface: The user must be given many options to choose form in terms of parameters. This can only be optimized after a lot of testing and analysis all the edge cases.
2. Interactive visualization the data extracted from the trained models will return raw and inscrutable data. This must be present in an aesthetic and an “easy to read” style.
3. Feedback system: A valuable feedback system must be developed to understand the customer’s needs that have not been met. This will help us train the models constantly.

13. Product details - How does it work?

– To detect fraud, a machine learning model first needs to collect data. The model analyzes all the data gathered, segments, and extracts the required features from it. Next, the machine learning model receives training sets that teach it to

predict the probability of fraud. Finally, it creates fraud detection machine learning models.

- The first step, data input, differs for ML and humans. Whereas humans struggle to comprehend massive amounts of data, such a task is a piece of cake for ML. The more data an ML model receives, the better it can learn and polish its fraud detection skills.

- Feature extraction is the next step. At this point, features describing good customer behavior and fraudulent behavior are added. These features usually include (but are not limited to) the customer's location, identity, orders, network, and chosen payment method. Based on the complexity of the fraud detection system, the list of investigated features can differ.

- Next, a training algorithm is launched. In a nutshell, this algorithm is a set of rules that an ML model has to follow when deciding whether an operation is legitimate or fraudulent. The more data a business can provide for a training set, the better the ML model will be.

- Finally, when the training is over, the company receives a fraud detection machine learning model suitable for their business. This model can detect fraud in next to no time with high accuracy. To be effective in credit card fraud detection, a machine learning model needs to be constantly improved and updated. Payment fraud detection can be eliminated for a while using ML. But sooner or later, fraudsters will come up with new tricks to game the system unless you keep it updated.

14. References/Source of Information

- https://intellias.com/how-to-use-machine-learning-in-fraud-detection/
- https://www.kaggle.com/datasets/ealaxi/paysim1
- https://www.monster.com/career-advice/article/business-opportunity-fraud
- https://www.sas.com/en_in/insights/articles/risk-fraud/fraud-detection-machine-learning.html
- https://intellipaat.com/blog/fraud-detection-machine-learning-algorithms/
- https://www.monster.com/career-advice/article/business-opportunity-fraud

Thank you