Fraud Detection in Financial Transactions Using Machine Learning

SURTECH, DUM DUM

A Presentation By Group No :-09

Name	University Roll No	Registration No
Vaibhaw Sinha	25500319030	004575
Koustav Dhara	25500319024	009718
Juber Ali	25500319037	000928
Tanmoy Dey	25500319029	004884
Hemant Tiwary	25500319035	001033

Under the supervision of
Mrs. Manjari Bharti,
Department of ECE, DSCSDEC KOLKATA

CONTENTS

OVERVIEW

As the market of e-commerce is growing day by day, cash flow, digital payments and transfers etc. is expanding at huge extents. With the expansion of these financial transactions, lot of cases involving several fraud/false activities are growing.

- These cases involve false payments, fraud transfers, fraud cash flows.
- Cybercriminals are getting smarter and, ironically, are advanced in technology for their own benefit.

According to a report published by Nilson, in 2007 alone, the worldwide losses in card fraud related cases reaches 22.8 billion dollars. Banks and financial institutions have no choice but to tighten their defenses and develop their own capabilities faster.

LITERATURE SURVEY

- Traditionally, banks and financial institutions have approached fraud detection with manual procedures, or rule-based solutions. A rule-based approach means that a complex set of criteria for flagging suspicious transaction needs to be established and reviewed manually. Since the criteria involves manual reviewing of all the data, it is quite inefficient.

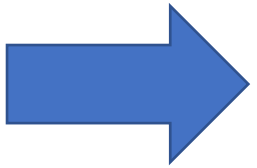
- With this project, we aim to use machine learning to determine a model which can most accurately predict if a transaction is fraud or not.

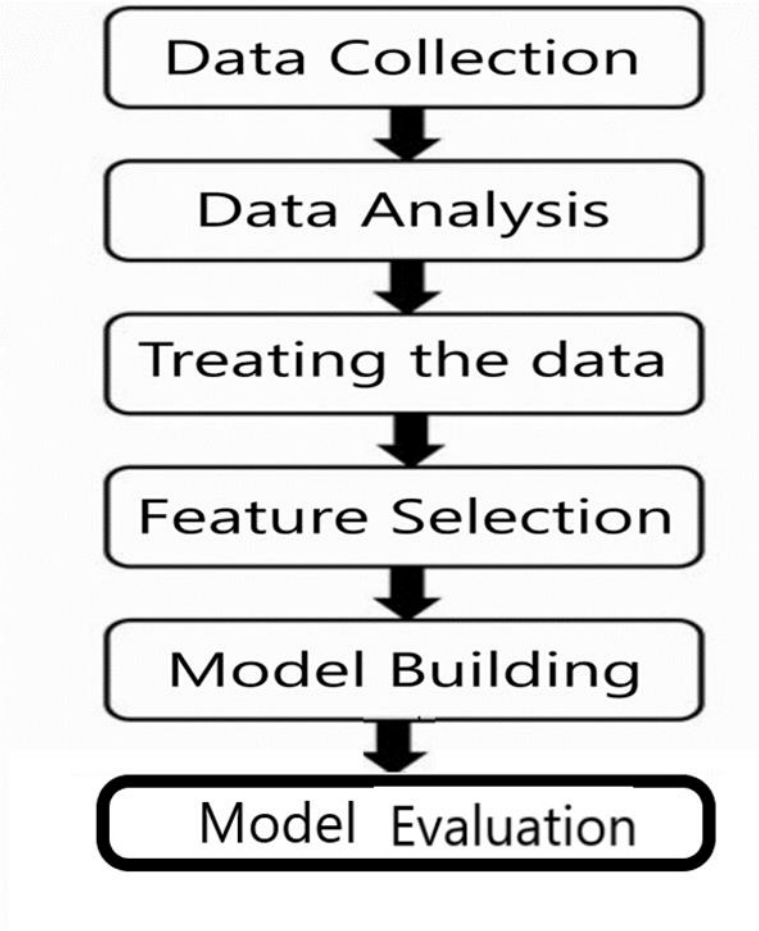
- We train various models using past tabular data containing details about the transactions. The models would differ in terms of training algorithms and the number of attributes fed to them while training.

- These models will find out how different attributes affect a fraud transaction from the past data. So, when a new transaction occurs, the trained model can know if the transaction is fraud or not based on the value of attributes of the transaction

PROCESS FLOW DIAGRAM

The various steps involved in determining the best model are shown in the flowchart.



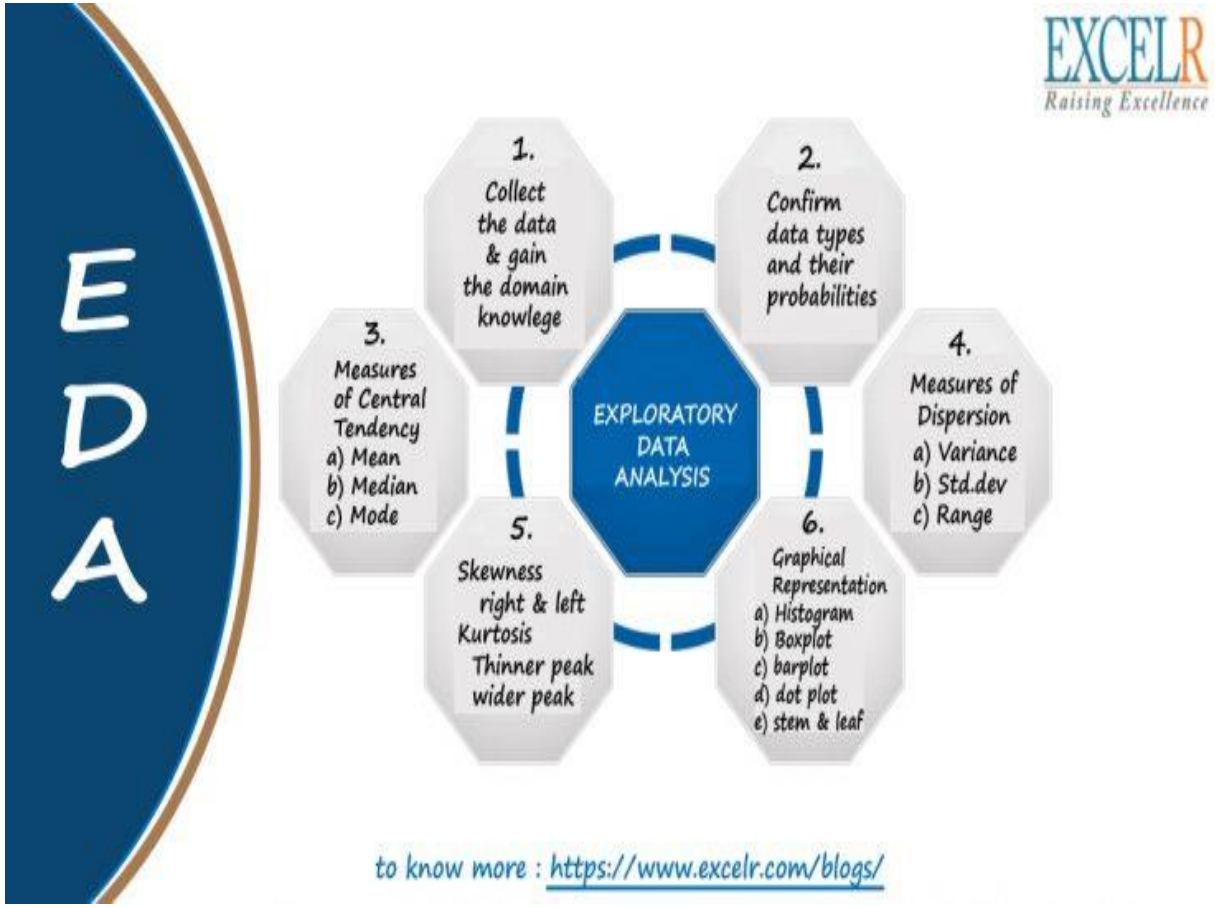


DATA COLLECTION

- The dataset that we will use in our project to train our ML models is obtained from Kaggle community. The presented dataset is a synthetic dataset generated using the simulator called Pay Sim.
- This dataset resembles the normal operation of transactions and injects malicious behavior for evaluation of the performance of fraud detection methods.
- The database is based on a sample of real transactions extracted from one month of financial logs from a mobile money service implemented in an African country which is scaled down to 1/4th of original dataset. This dataset was created on 31-03-2017 and evaluated over a month. It has 6362620 cases with 11 features.

DATA ANALYSIS

Exploratory Data Analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task.



SAMPLE OF DATASET

	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

TREATING THE DATA

- In order to apply the algorithms on the datasets, the data provided for training and testing models should be treated and prepared accordingly. So that we get precise outcomes from the various models built.
- Several statistical algorithms are applied to scale and normalize the data. Missing values are taken care of in this step.

FEATURE SELECTION

Since the features present in the dataset is scaled and standardized, it is needed to check which data is influencing the output feature with what magnitude. So that the features can be compared in order to give priorities. Based on the type of data present in the input features(continuous or categorical), the feature selection can be done by several methods.

- Co-relation method, being one of them, takes one input feature of continuous type and output feature of continuous type.
- Chi-square method, takes one input feature of categorical type and output feature of categorical type.
- Anova method, takes one input feature of continuous type and output feature of categorical type.

Information gain method, takes one input feature of categorical type and output feature of continuous type.

MODEL BUILDING

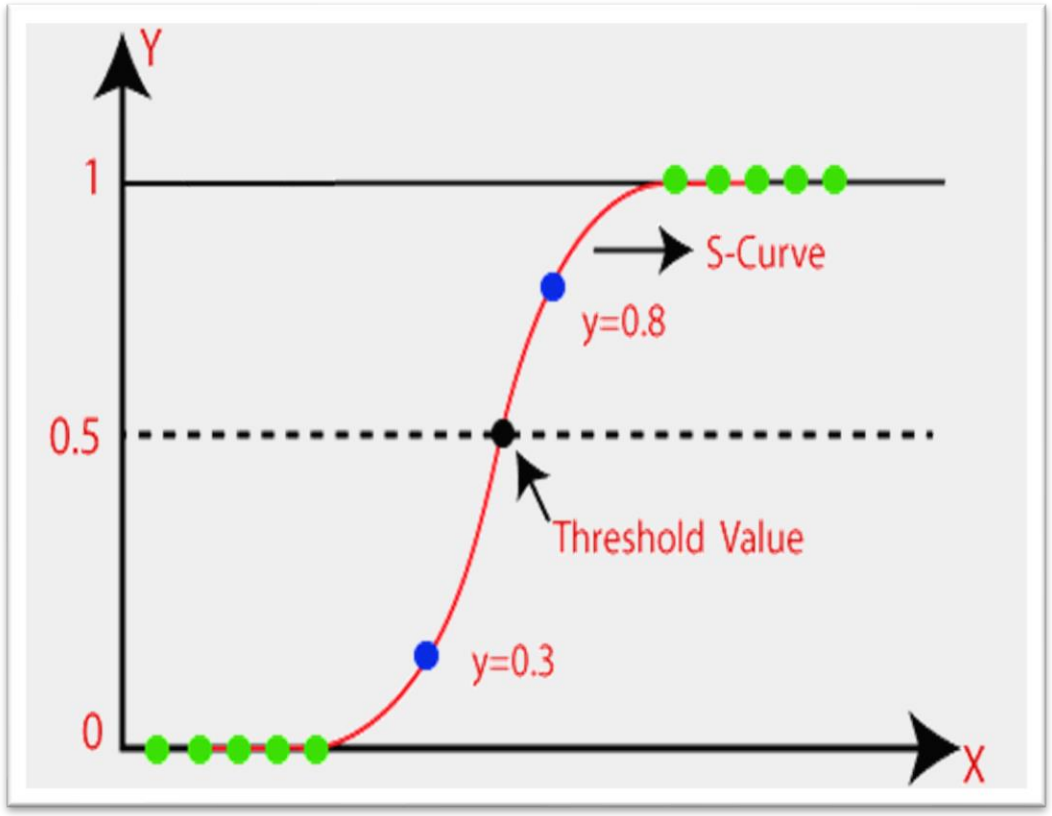
A machine learning model is a mathematical model that generates predictions by finding patterns from organized datasets fed during the training of model..

Machine learning model selection depends on the type of output feature of the dataset that if it is of categorical of continuous type.

In this project, the output feature is Fraud is a classifier type output having values 0 or 1. So, various type of classifier models are used like

- Logistic Regression model
- Decision Tree Classifier model
- KNN classifier model

LOGISTIC REGRESSION

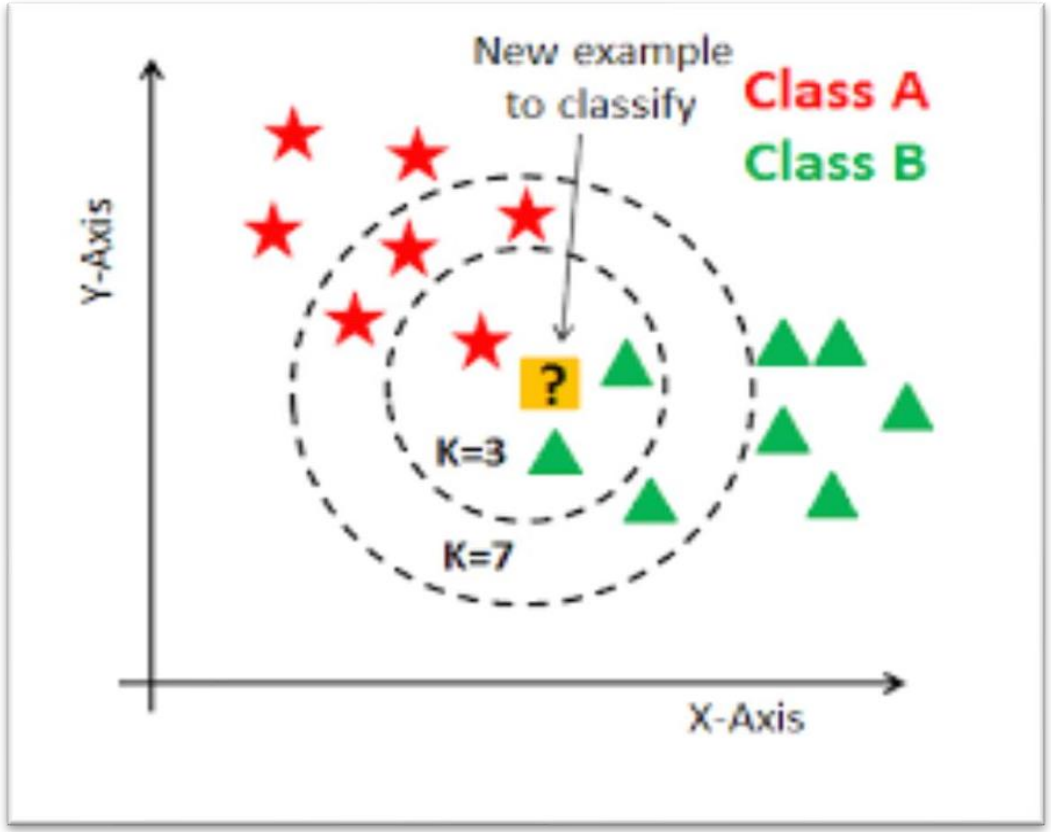


$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

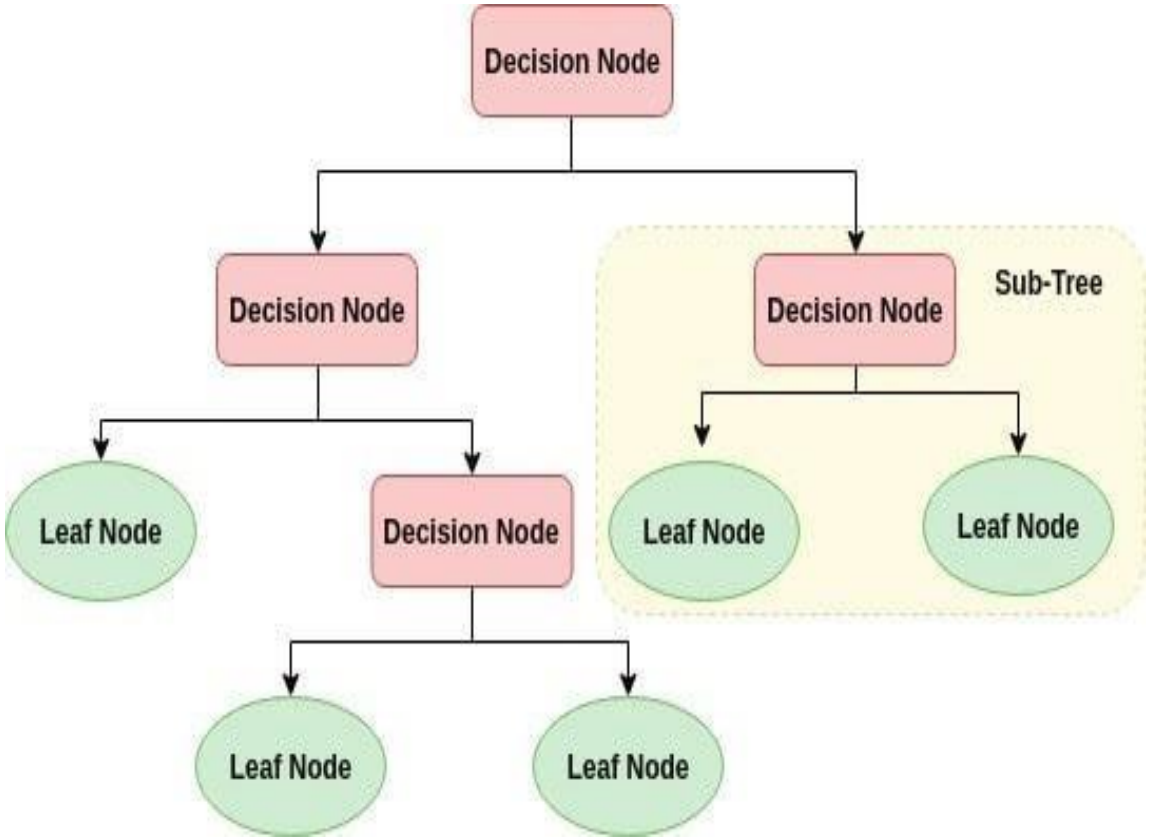
It uses a sigmoid function to map the value of x(dependent attribute) and y(output attribute) between 0 to 1. 0.5 is called the threshold value. If the value of y(Is Fraud) is less then 0.5 we consider the transaction is genuine if it is greater then 0.5 we consider that the transaction is fraudulent.

K-NEARSET-NEIGHBORS CLASSIFIER MODEL



- A k-nearest-neighbor is a data classification algorithm that attempts to determine what group a data point is in by looking at the data points around it.
- In the diagram, looking at one point on a grid, trying to determine if a point is in group A or B, looks at the states of the points that are near it. The range is arbitrarily determined, but the point is to take a sample of the data. If the majority of the points are in group A, then it is likely that the data point in question will be A rather than B, and vice versa.
- The k-nearest-neighbor is an example of a "lazy learner" algorithm because it does not generate a model of the data set beforehand.

DECESION TREE CLASSIFIER



- Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects(s), each belonging to one of the classes.
- A decision tree classifier model is based on the concept of decision tree building in which several decision conditions are placed as nodes in a tree which are processed further to achieve the solution.

MODEL EVALUATION

- Model evaluation involves testing of the trained models with suitable datasets. This evaluation is achieved by building and analyzing the confusion matrix and some other scoring parameters.
- Before model building and feeding data into different model or algorithm we split the data set into training set and testing set in the ratio 7:3. 70% data is used for training of models and 30% data for testing the model.
- The model is then tested and evaluated in order to extract the most precise/suitable model that should be preferred for the prediction of the output feature when a new set of input features/data is tested/fed to the model.

CONFUSION MATRIX

n = total predictions	Actual: No	Actual: Yes
Predicted: No	True Negative	False Positive
Predicted: Yes	False Negative	True Positive

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\textit{Precision} = \frac{TP}{TP + FP}$$

$$\textit{Recall} = \frac{TP}{TP + FN}$$

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.

- Accuracy:

Ratio of number of correct prediction by classifier by to all number of prediction made by classifier.

- Precision:

It is number of correct output predicted by the model by total number of prediction in the positive class.

- Recall:

It is defined as the out of total positive classes, how our model predicted correctly. The value of recall must be as high as possible.

The value of above terms will be between 0 an 1.
The more the value approaches close to 1 the more accurate the model is.

MODEL SELECTION

- The best model is then selected based on the accuracy and precision scores.
- The model can be implemented by banking systems after they are fed with the data of previous transactions of that system so that the models are trained with similar data they would be working on.

COMPONENT AND SOFTWARE USED

- Software Used: Jupyter notebook
- Programming Language used: Python 3

Python Libraries Used:

- Panda
- NUMPY
- Sci Kit Learn
- Matplotlib

REFERENCES

- www.Kaggle.com
- Docs.python.org
- Numpy.org
- Javatpoint.org
- Data obtained from: www.Kaggle.com

THANK YOU