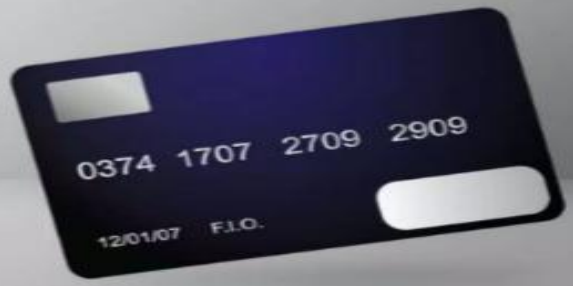
Credit Card Fraud Detection



Problem Defination:

This Project is centered on credit card fraud detection in actual international scenarios. Nowadays credit score card frauds are extensively growing in variety compared to in advance times. Criminals are the usage of faux identification and numerous technology to entice the customers and get the cash out of them. Therefore, it's far very important to discover a way to those sorts of frauds. In this

proposed undertaking we designed a version to discover the fraud hobby in credit score card transactions. This device can offer maximum of the vital capabilities required to discover unlawful and illicit transactions. As generation modifications constantly, it's far becoming tough to song the conduct and sample of crook transactions. To provide you with the answer you will make use of technology with the boom of device learning, synthetic intelligence and different applicable fields of statistics generation;

EXISTING SYSTEM

With growing advancement in the electronic commerce field, fraud is spreading all over the world, causing major financial losses. In the current scenario, Major cause of financial losses is credit card fraud; it not only affects tradesperson but also individual clients. Decision tree, Genetic algorithm, Metalearning strategy, neural network, HMM are the presented methods used to detect credit card frauds. In contemplating system for fraudulent detection, artificial intelligence concept of Support Vector Machine (SVM) & decision tree is being used to solve the problem. Thus by the implementation of this hybrid approach, financial losses can be reduced to greater extent.

DEEP LEARNING ALGORITHM :

LSTMs are a type of Recurrent Neural Network (RNN) that can learn and memorize long-term dependencies. Recalling past information for long periods is the default behavior.

LSTMs retain information over time. They are useful in time-series prediction because they remember previous inputs. LSTMs have a chain-like structure where four interacting layers communicate in a unique way. Besides time-series predictions, LSTMs are typically used for speech recognition, music composition, and pharmaceutical development.

MODULES:

- 1.Data collection
- 2.Data pre-processing
- 3.Feature extraction
- 4.Evaluation model

1.DATA COLLECTION:

Data used in this paper is a set of product reviews collected from credit card transactions records. This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called labelled data.

2.DATA PRE -PROCESSING:

Pre-processing is the process of three important and common steps as follows:

- ✓ Formatting: It is the process of putting the data in a legitimate way that it would be suitable to work with. Format of the data files should be formatted according to the need. Most recommended format is csv files.
- ✓ Cleaning: Data cleaning is a very important procedure in the path of data science as it constitutes the major part of the work. It includes removing missing data and complexity with naming category and so on. For most of the data scientists, Data Cleaning continues of 80% of work.
- ✓ Sampling: This is the technique of analyzing the subsets from whole large datasets, which could provide a better result and help in understanding

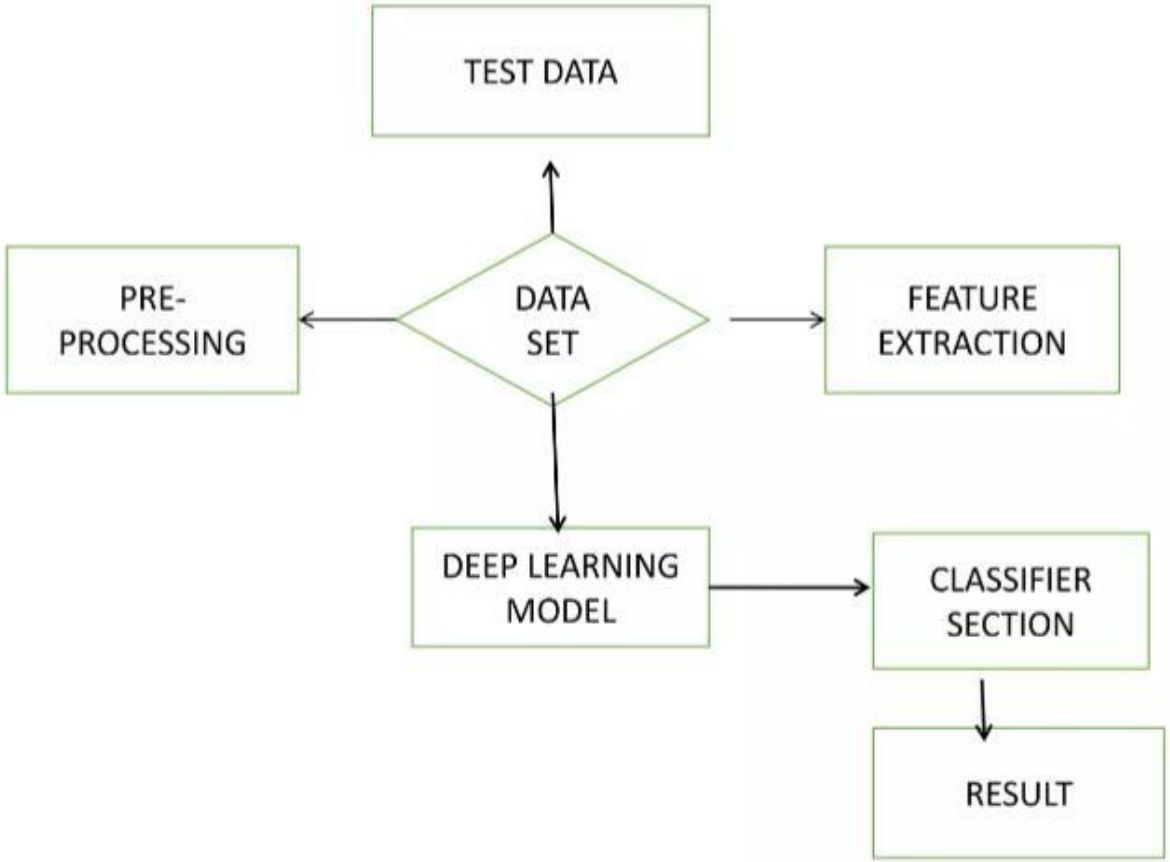
3.Feature extraction:

Feature extraction is the process of studying the behavior and pattern of the analyzed data and draw the features for further testing and training. Finally, our models are trained using the Classifier algorithm. We use classify module on Natural Language Toolkit library on Python. We use the labelled dataset gathered. The rest of our labelled data will be used to evaluate the models. Some Deep learning algorithms were used to classify pre-processed data. The chosen classifiers were Random forest. These algorithms are very popular in text classification tasks.

4.Evaluation model:

Model Evaluation is an essential part of the model development process. It helps to find the best model that represents our data and how well the selected model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can effortlessly generate overoptimistically and over fitted models. To avoid overfitting, evaluation methods such as hold out and cross-validations are used to test to evaluate model performance. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is well-defined as the proportion of precise predictions for the test data. It can be calculated easily by mathematical calculation dividing the number of correct predictions by the number of total predictions.

ER-DIAGRAM:


TEST DATA
PRE-
PROCESSING
DATA
SET
FEATURE
EXTRACTION
DEEP LEARNING
MODEL
CLASSIFIER
SECTION
RESULT

Contd...

- **Logistic Regression Classifier:**
  - Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable, where there are only two possible outcomes.
  - The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest, and a set of independent variables.
  - Logistic Regression generates the coefficients of a formula to predict a Logit Transformation of the probability of presence of the characteristic of interest.

- **Logistic Regression Classifier**

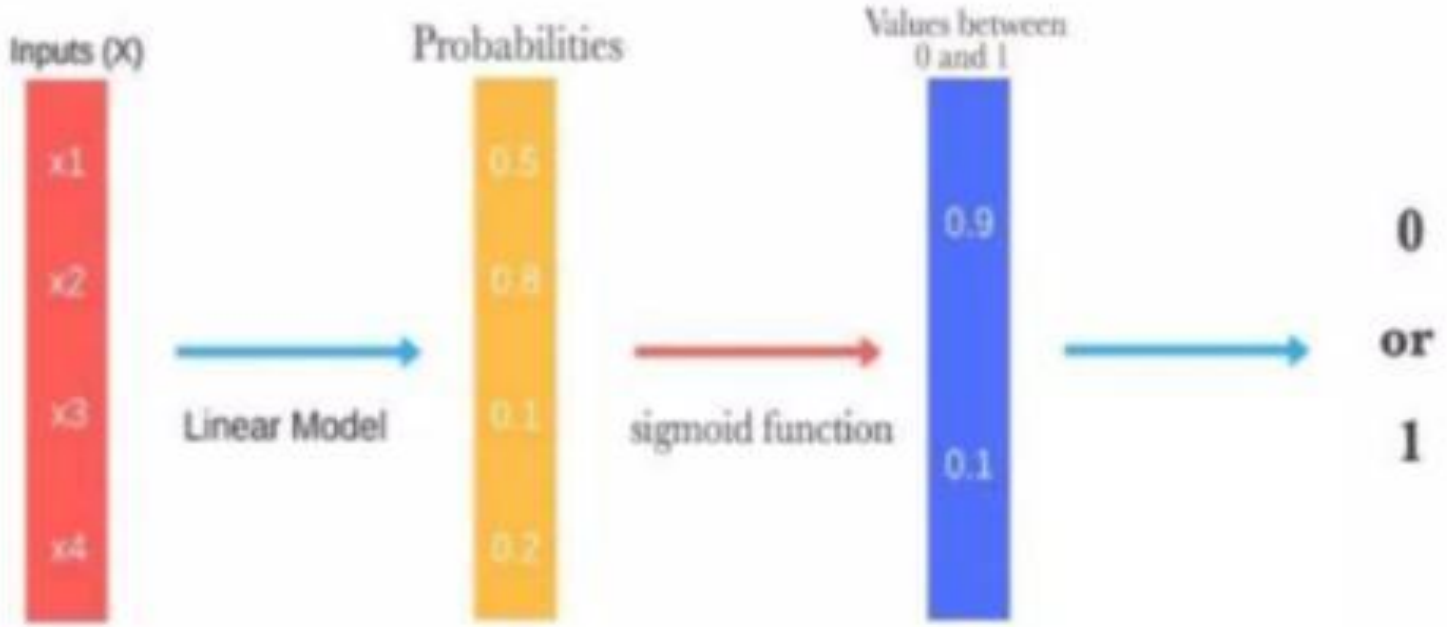


Fig.3 Working of Logistic Regression Model

Contd...

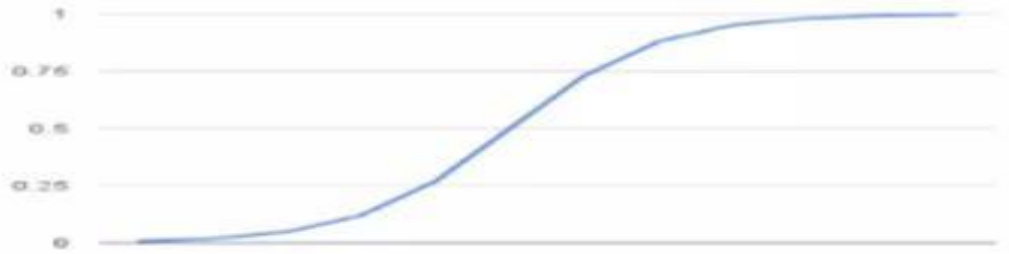


Fig.4 Sigmoid Function

Sigmoid Function:

➢A sigmoid function is a mathematical function having a characteristic "S"-shaped curve or sigmoid curve.

➢A sigmoid function is a bounded, differentiable, real function that is defined for all real input values and has a non-negative derivative at each point.

$$\sigma(x) = \frac{1}{(1+\ell^{-x})}$$

$$x = w_0 z_0 + w_1 z_1 + \ldots + w_n z_n$$

Performance Evaluation and Results

- Four metrics used in evaluation
  - True Positive Rate(TPR)
  - True Negative Rate (TNR)
  - False Positive Rate (FPR)
  - False Negative Rate (FNR)

$$TPR = \frac{TP}{P}$$

$$TNR = \frac{TN}{N}$$

$$FPR = \frac{FP}{N}$$

$$FNR = \frac{FN}{P}$$

Contd ...

- **Performance of naïve bayes, k-nearest neighbour and logistic regression classifiers are evaluated based on :**
  - Accuracy
  - Sensitivity
  - Specificity
  - Precision
  - Matthews Correlation Coefficient (MCC)
  - Balanced Classification Rate(BCR)

CONCLUSION:

Hence, we have acquired the result of an accurate value of credit card fraud detection i.e. 0.9994802867383512 (99.93%) using a random forest algorithm with new enhancements. In comparison to existing modules, this proposed module is applicable for the larger dataset and provides more accurate results. The Random forest algorithm will provide better performance with many training data, but speed during testing and application will still suffer. Usage of more pre-processing techniques would also assist Our future work will try to represent this into a software application and provide a solution for credit card fraud using the new technologies like Machine Learning, Artificial Intelligence and Deep Learning