Title: Online Credit Card Fraud Detection System

Team Members

B. Anil Kumar-700731647

G. Pranav-700741488

K. Satya Ishyanth-700735513

Abstract

Online transactions are fast rising, and credit cards will be used mostly. Loss of physical credit cards or loss of credit card information will result in a substantially higher payment. The hackers are there to commit fraud against people. As a result, there was a requirement to detect fraudulent transactions and safeguard online credit card transactions. To analyse this problem and combat credit card theft, we created a new system called "Online credit card fraud detection and prevention system" that employs machine learning. We tested the system's accuracy by experimenting with several algorithms. This system uses the random forest algorithm to determine if a transaction is legitimate or fraudulent.

Introduction

Many businesses are expanding rapidly over the world at the moment. Companies strive to deliver the greatest services to their clients. Companies process massive amounts of data on a daily basis for this reason. This data also includes the clients' personal and financial information. As a result, firms must store data in order to handle it, and data security is critical. If this data is not secured, it may be exploited by other companies or, in the worst-case situation, stolen. In rare circumstances, financial information is taken and utilized for fraudulent transactions, causing harm to the parties involved.

Today, online buying has become a widespread daily purchase habit. The perpetrators are engaging in malicious operations such as Trojan and spoofing. When criminals steal cardholder information, the increase in fraud incidents has become a severe issue. Credit card fraud detection is an important subject that has piqued the interest of the machine learning and computational intelligence fields, where a plethora of automation solutions have been presented. Data is available all around the world, and businesses of all sizes are loading information with great volume, variety, speed, and value. This data is derived from a variety of sources, including social media followers, likes, and comments, as well as user purchasing habits. All of this data is utilized to analyze and visualize hidden data patterns.

Dataset:

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

About Dataset

- Motive!

  The purpose of uploading this data collection is to gain insights into credit card defaulters based on the relevant parameters!
- Inside?

  In the application data set, there are roughly 122 columns and attributes like IncomeTotal, AMTAPPLICATION, and AMT CREDIT. The intriguing thing is that we can also leverage the PREVIOUS APPLICATION data set to gain more insights if we want to find patterns and variations.
- Inspiration

  We used this data set as our homework and did the best we could to complete the EDA!

Software And Hardware Requirements

Operating System: Window 10/Unix

Ide: Anaconda

Ram: 8GB

HDD: 100GB+

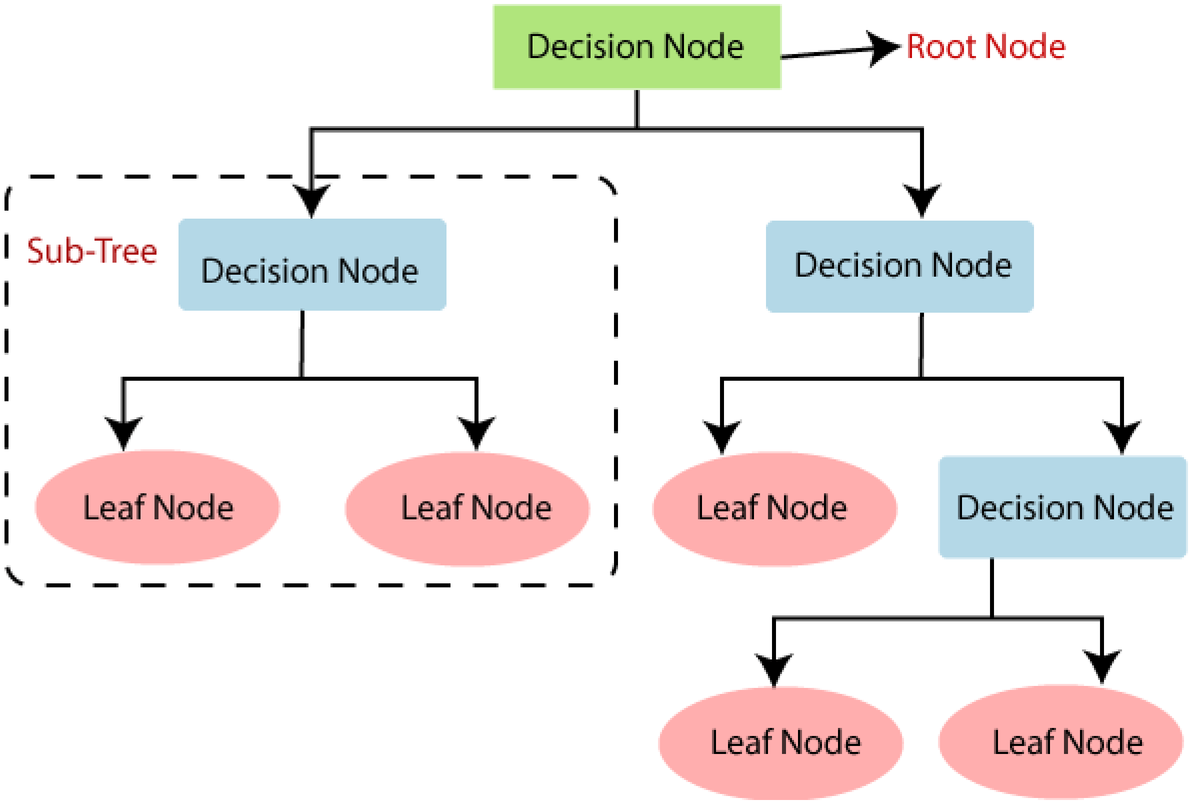
Programming Language: Python

Packages: Pandas, Numpy, SKLearn

Algorithm

Decision Tree Classification Algorithm

- Decision Tree is a Supervised learning technique that may be applied for both classification and regression issues, however it is most commonly employed for classification. It is a tree-structured classifier in which internal nodes contain dataset attributes, branches represent decision rules, and each leaf node represents the result.
- A Decision tree has two nodes: the Decision Node and the Leaf Node. Decision nodes are used to make decisions and have numerous branches, whereas Leaf nodes represent the results of those decisions and do not have any additional branches.
- The decisions or tests are based on the characteristics of the presented dataset.
- It is a graphical representation of all possible solutions to a problem/decision given certain conditions.
- It is named a decision tree because, like a tree, it begins with the root node and then branches out to form a tree-like structure.
- The CART algorithm, which stands for Classification and Regression Tree algorithm, is used to form a tree.
- A decision tree simply asks a question and divides the tree into subtrees based on the answer (Yes/No).



Process Flow

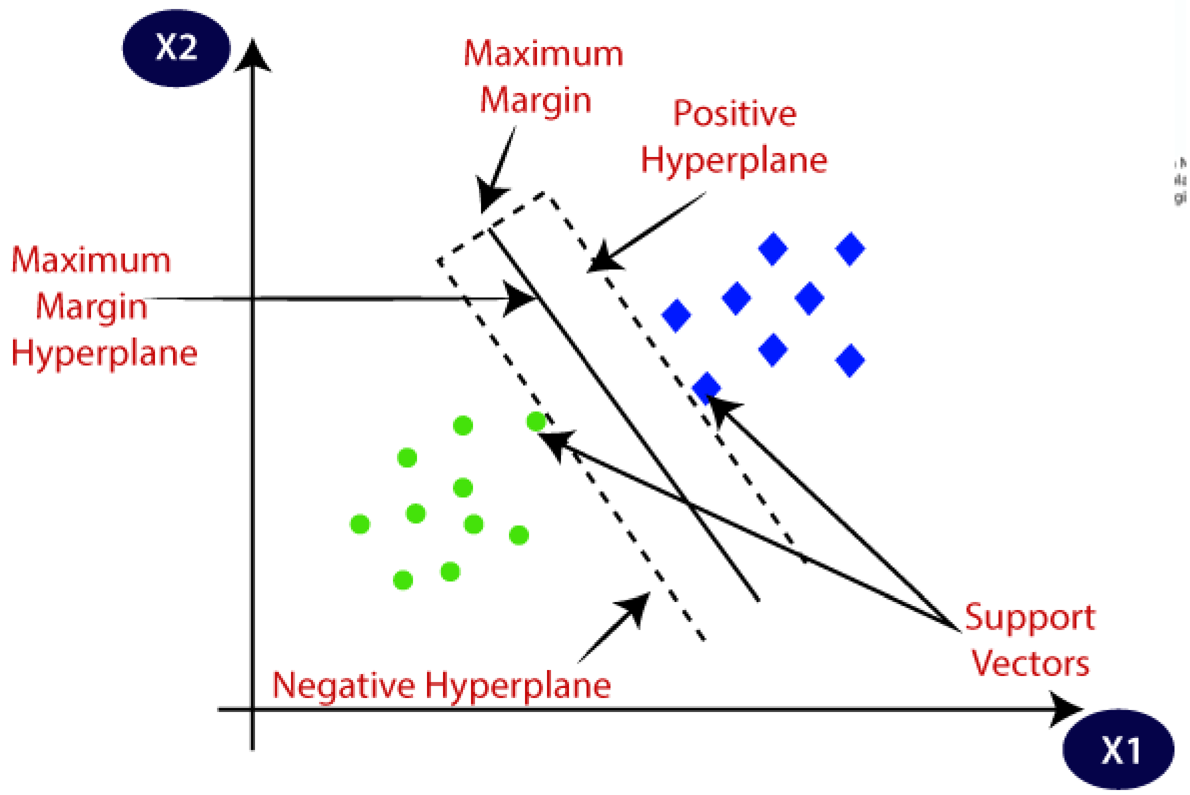
- **Step 1: Begin the tree with the root node, which contains the entire dataset, says S.**
- **Step 2: Using the Attribute Selection Measure, find the best attribute in the dataset (ASM).**
- **Step 3: Subdivide the S into subsets containing potential values for the best qualities.**
- **Step 4: Create the decision tree node with the best attribute.**
- **Step 5: Create new decision trees recursively using the subsets of the dataset obtained in step 3. Continue this process until you reach a point where you can no longer categorize the nodes and refer to the final node as a leaf node.**

Support Vector Machine

Support Vector Machine, or SVM, is a prominent Supervised Learning technique that is used for both classification and regression issues. However, it is mostly utilized in Machine Learning for Classification difficulties.

The SVM algorithm's purpose is to find the optimum line or decision boundary for categorizing n-dimensional space so that we may easily place fresh data points in the correct category in the future. A hyperplane is the optimal choice boundary.

SVM selects the extreme points/vectors that aid in the creation of the hyperplane. These extreme examples are referred to as support vectors, and the method is known as the Support Vector Machine. Consider the picture below, which shows two distinct categories that are classified.



Types of SVM

SVM can be of two types:

- Linear SVM: Direct SVM is utilized for straightly distinguishable information, and that implies if a dataset can be characterized into two classes by utilizing a solitary straight line, then such information is named as straightly distinct information, and classifier is utilized called as Direct SVM classifier.
- Non-direct SVM: Non-Direct SVM is utilized for non-directly isolated information, and that implies in the event that a dataset can't be grouped by utilizing a straight line, such information is named as non-direct information and classifier utilized is called as Non-straight SVM classifier.

MLP Classifier

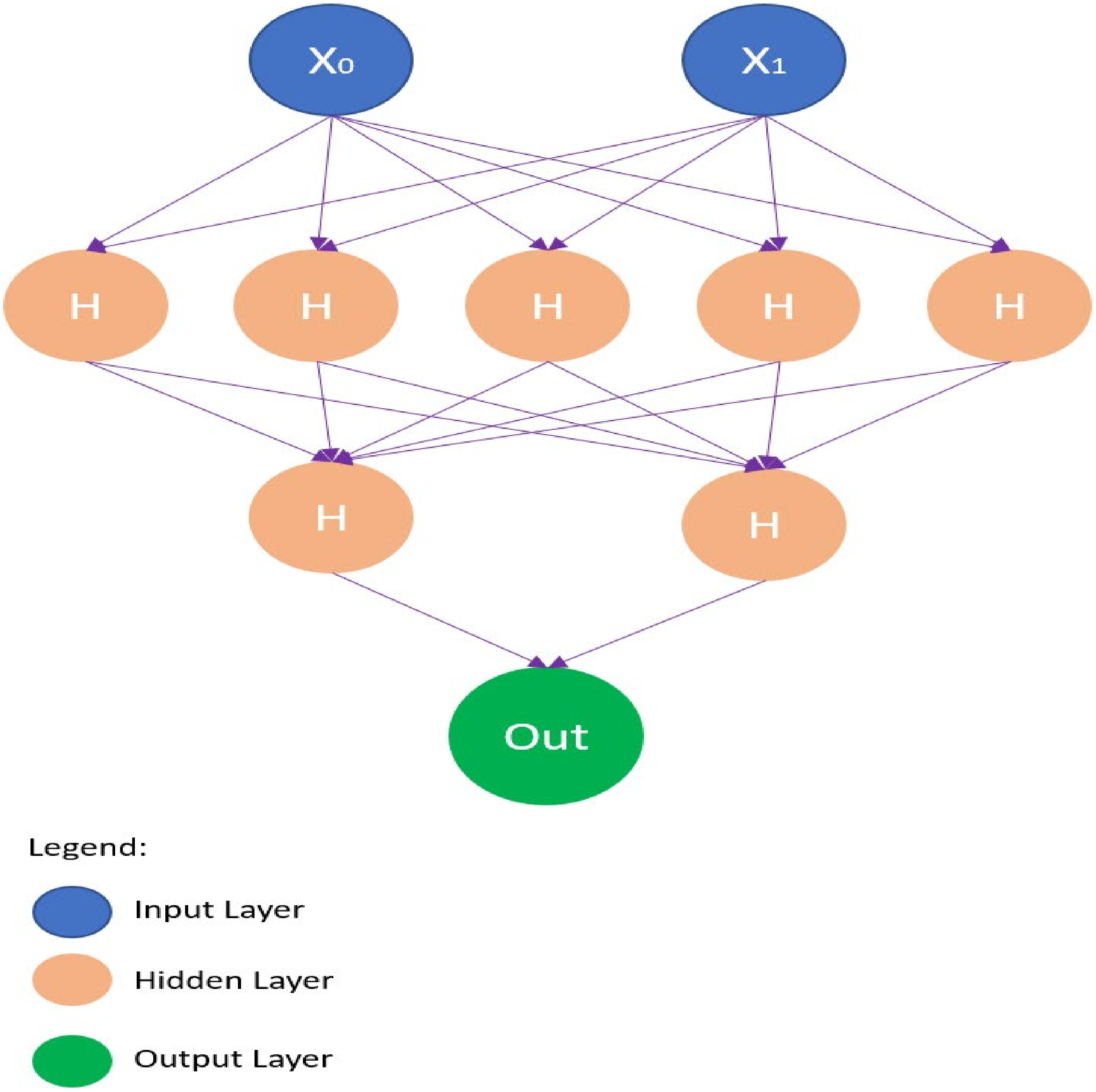
The multilayer perceptron (MLP) is a feedforward counterfeit brain network model that guides input informational indexes to a bunch of suitable results. A MLP comprises of different layers and each layer is completely associated with the accompanying one. The hubs of the layers are neurons with nonlinear enactment capabilities, aside from the hubs of the information layer. Between the information and the result layer there might be at least one nonlinear secret layers.

Suppose we have two predictor variables and want to do a binary classification. For this I can enter the following parameters at the model:

```
mlp_clf = MLPClassifier(hidden_layer_sizes=(5,2),
                        max_iter = 300,activation = 'relu',
                        solver = 'adam')
```

- hidden_layer_sizes: With this boundary we can determine the quantity of layers and the quantity of hubs we need to have in the Brain Organization Classifier. Every component in the tuple addresses the quantity of hubs at the ith position, where I is the file of the tuple. Subsequently, the length of the tuple demonstrates the absolute number of stowed away layers in the brain organization.

- max_iter: Indicates the number of epochs.
- activation: The activation function for the hidden layers.
- solver: This parameter specifies the algorithm for weight optimization over the nodes.



Expected Results

At the point when we get the information, after information cleaning, pre-handling, and fighting, the initial step we do is to take care of it to an extraordinary model and obviously, get yield in probabilities. Be that as it may, hang on! How on earth might we at any point method the viability of our model. Better the viability, better the presentation, and that is precisely exact thing we need. What is more, it is where the Disarray framework comes into the spotlight. Disarray Network is an exhibition estimation for AI grouping.

True Positive:

Interpretation: You predicted positive and it's true.

True Negative:

Interpretation: You predicted negative and it's true.

False Positive: (Type 1 Error)

Interpretation: You predicted positive and it's false.

False Negative: (Type 2 Error)

Interpretation: You predicted negative, and it is false.

Conclusion

Visa misrepresentation connotes an intense business issue. These cheats can prompt enormous misfortunes. web based paying cash by the Mastercard is expanded and furthermore Visa fakes so there is need to recognize this happened misrepresentation exchanges and give security to the clients about charge card. Thus, the principal reason for this paper is for distinguishing as well as forestalling the cheats during exchanges. There are still a few issues in past framework like precision.

References and Bibliography

[1] Global Facts (2019). Topic: Startups worldwide. [online] Available at: https://www.statista.com/topics/4733/startupsworldwide/ [Accessed 10 Jan. 2020].

[2] Legal Dictionary (2019). Fraud - Definition, Meaning, Types, Examples of fraudulent activity. [online] Available at: https://legaldictionary.net/fraud/ [Accessed 15 Jan. 2020].

[3] A. Mishra, C. Ghorpade, “Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques” 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) pp. 1-5. IEEE

[4] S. Bachmayer, “Artificial Immune Systems,” Artificial Immune Systems, vol. 5132, pp. 119–131, 2008.

[5] M. Krivko, “A hybrid model for plastic card fraud detection systems,” Expert Systems with Applications, vol. 37, no. 8, pp. 6070–6076, Aug. 2010.

[6] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, “Data mining for credit card fraud: A comparative study,” Decision Support Systems, vol. 50, no. 3, pp. 602–613, Feb. 2011.