CREDIT CARD FRAUD DETECTION USING DATA SCIENCE TECHNIQUE

T. Ahathyan[1], V. Deepak[2], Mrs. A. Sengodi[3]

Student, B.E. Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India[1,2]

Assistant Professor, CSE, Anand Institute of Higher Technology, Chennai, India[3]

Abstract: A credit card is issued by a bank or financial services company that allows cardholders to borrow funds with which to pay for goods and services with merchants that accept cards for payment. Nowadays as everything is made cyber so there is a chance of misuse of cards and the account holder can lose the money so it is vital that credit card companies are able to identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. This type of problems can be solved through data science by applying machine learning techniques. It deals with modelling of the dataset using machine learning with Credit Card Fraud Detection. In machine learning the main key is the data so modelling the past credit card transactions with the data of the ones that turned out to be fraud. The built model is then used to recognize whether a new transaction is fraudulent or not. The objective is to classify whether the fraud had happened or not. The first step involves analyzing and pre-processing data and then applying machine learning algorithm on the credit card dataset and find the parameters of the algorithm and calculate their performance metrics.

I. INTRODUCTION

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domainsThe term "data science" has been traced back to 1974, when Peter Naur proposed it as an alternative name for computer science. In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic. However, the definition was still in flux.
The term "data science" was first coined in 2008 by D.J. Patil, and Jeff Hammerbacher, the pioneer leads of data and analytics efforts at LinkedIn and Facebook. In less than a decade, it has become one of the hottest and most trending professions in the market.Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.

Data science can be defined as a blend of mathematics, business acumen, tools, algorithms and machine learning techniques, all of which help us in finding out the hidden insights or patterns from raw data which can be of major use in the formation of big business decisions.Data scientists examine which questions need answering and where to find the related data. They have business acumen and analytical skills as well as the ability to mine, clean, and present data. Businesses use data scientists to source, manage, and analyze large amounts of unstructured data

1.1 OBJECTIVE
The goal is to develop a machine learning model for Credit Card Fraud Prediction.
To potentially replace the updatable supervised machine learning classification models by predicting results in the form of best accuracy by comparing supervised algorithm.

1.2 SCOPE
The main Scope is to detect the Fraud Prediction, which is a classic text classification problem with a help of machine learning algorithm.
It is needed to build a model that can differentiate between Fraud OR not

II. ANALYSIS

2.1 SYSTEM ANALYSIS
System Analysis is a combined process dissection the system responsibilities that are based on problem domain characteristics and user requirement.

2.1.1 Problem Definition

Prices of real estate properties are sophisticatedly linked with our economy. Despite this, we do not have accurate measures of housing prices based on the vast amount of data available. Theref\ore, the goal of this project is to use machine learning to predict the selling prices of houses based on many economic factors. A systematic method can be built to derive a layered knowledge graph and design a structured Deep Neural Network (DNN) based on it. Neurons in a structured DNN are structurally connected, which makes the network time and space efficient; and thus, it requires fewer data points for training. The structured DNN model has been designed to learn from the most recently captured data points which allows the model to adapt to the latest market trends. To demonstrate the effectiveness of the proposed approach, we can use a case study of assessing real properties in small towns.

2.1.2 Existing System

They proposed a method and named it as Information-Utilization-Method INUM it was first designed and the accuracy and convergence of an information vector generated by INUM are analyzed. The novelty of INUM is illustrated by comparing it with other methods. Two D-vectors (i.e., feature subsets) a and b, where Ai is the ith feature in a data set, are dissimilar in decision space, but correspond to the same O-vector y in objective space. Assume that only a is provided to decision-makers, but a becomes inapplicable due to an accident or other reasons (e.g., difficulty to extract from the data set). Then, decision-makers are in trouble. On the other hand, if all two feature subsets are provided to them, they can have other choices to serve their best interest. In other words, obtaining more equivalent D-vectors in the decision space can provide more chances for decision-makers to ensure that their interests are best served. Therefore, it is of great significance and importance to solve MMOPs with a good Pareto front approximation and also the largest number of D-vectors given each O-vector.

2.1.3 Proposed system

Multiple datasets from different kaggle would be combined to form a generalized dataset. In data wrangling section the loaded data will check for cleanliness, and then trim and clean given dataset for analysis. Different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy. The data set collected for predicting given data is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using machine learning algorithms are applied on the Training set and based on the test result MAE, Test set process is done.

III. MODULES

4.1 Data Pre-processing
4.2 Data Analysis of Visualization
4.3 Comparing Algorithm with prediction in the result
4.4 Deployment Using Flask

4.1 Data pre-processing

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset.

4.2 Data analysis of visualization

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding.

4.3 Comparing algorithm with prediction

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn.

4.4 Deployment

Flask is a micro web framework written in Python. It is classified as a micro-framework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where preexisting third-party libraries provide common functions.

IV. RESULTS AND DISCUSSION

The result is generated by comparing the accuracy of the different algorithm with the existing system to that of the proposed system. Various experiments have performed to evaluate the accuracy of the result. The comparison is based on the analyzed value of the existing. By comparing the following four algorithms the result is evaluated, the below algorithms are used to predict the result,

V. CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out. This application can help to find the Prediction of credit card fraud or not.

REFERENCES

1. Paruchuri, Harish. "Credit Card Fraud Detection using Machine Learning: A Systematic Literature Review." ABC Journal of Advanced Research 6.2 (2017): 113-120.
2. Chandini, S. B., and S. P. Devika. "A research on credit card fraudulent detection system." Int J Recent Technol Eng (IJRTE) 8 (2019): 5029-5032.
3. Shirgave, Suresh, et al. "A Review On Credit Card Fraud Detection Using Machine Learning." International Journal of Scientific & technology research 8.10 (2019): 1217-1220.
4. Abinayaa, S., et al. "Credit card fraud detection and prevention using machine learning." International Journal of Engineering and Advanced Technology (IJEAT) 9.4 (2020).
5. Bansal, Akanksha, and Hitendra Garg. "An Efficient Techniques for Fraudulent detection in Credit Card Dataset: A Comprehensive study." IOP Conference Series: Materials Science and Engineering. Vol. 1116. No. 1. IOP Publishing, 2021.