|  |
| --- |
| AI ALGORITHMS I  Credit Card Fraud Detection (V1)  D |
|  |
| November 6,2020   |  |  | | --- | --- | | Submitted To: | Submitted By: | | Marcos Bittencourt | Aman Verma | | AI ALGORITHMS I | 100799391 | |

Contents

1. Executive Summary……………………………………………………………………………………….….3
2. Problem Statement……………………………………………………………………………….………….3
3. Rational Statement………………………………………………………………………………….……….4
4. Data Requirement …………………………………………………………………………………………...4
5. Data Analysis Approach………………...………………………………………………………………….4
6. Outline to follow and Test Process…………………………………………………………………….5

## Executive Summary:

What is Credit Card Fraud?

Credit card fraud is when someone uses another person's credit card or account information to make unauthorized purchases or access funds through cash advances. Credit card fraud doesn’t just happen online; it happens in brick-and-mortar stores, too. As a business owner, you can avoid serious headaches – and unwanted publicity – by recognizing potentially fraudulent use of credit cards in your payment environment.



## Problem Statement

The Credit Card Fraud Detection Problem includes modeling past credit card transactions with the knowledge of the ones that turned out to be a fraud. This model is then used to identify whether a new transaction is fraudulent or not. Our aim here is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications.

## Rational Statement

Build a machine learning models, in order to make real-time fraud detection decisions tailored to the individual customer. Applying supervised learning and classification algorithms to build models which provide better insights, accurate decision and find interesting pattern in customer data. Improve decision making process by correctly predicting probability of default for detecting fraud.

## Data Requirements

We will use a dataset from the KAGLE to identify the fraud on Credit Card. Application and previous history are required for model building. The labels are included in the training data and train a model to predict the labels from the features. Details of the data required are as follow:

In the dataset, the features are scaled and the names of the features are not shown due to privacy reasons. Nevertheless, we can still analyze some important aspects of the dataset.

Assumption, Limitations and Constraints with the data:

* Understand the little distribution of the "little" data that was provided to us.
* Create a 50/50 sub-data frame ratio of "Fraud" and "Non-Fraud" transactions. (Near miss Algorithm)
* Determine the Classifiers we are going to use and decide which one has a higher accuracy.
* Create a Neural Network and compare the accuracy to our best classifier.
* Understand common mistake made with imbalanced datasets.

This dataset contains total of 284707 records which further composed of 31 columns and 284707 rows.

## Data Analysis Approach

This is a Classification problem and it is a scenario of Supervised Learning. This problem requires classification algorithms like Logistic Regression, Clustering, smote for sampling, Classifier etc., and try to understand the accuracy. This would be determined in the modeling phase of the project.

The software tools used will be: -

* Python – for EDA (Exploratory Data Analysis), Data Cleaning, Model Building, and Testing
* Jupyter Notebook – IDE for development
* Tableau/Power BI – for detailed visualizations

## Outline to Follow and Test Process:

I. Understanding our data

a) Gather Sense of our data

II. Preprocessing

a) Scaling and Distributing

b) Splitting the Data

III. Random Under Sampling and Oversampling

a) Distributing and Correlating

b) Anomaly Detection

c) Dimensionality Reduction and Clustering (t-SNE)

d) Classifiers

e) A Deeper Look into Logistic Regression

f) Oversampling with SMOTE

IV. Test Process

a) Testing with Logistic Regression

b) Neural Networks Testing (Under sampling vs Oversampling)