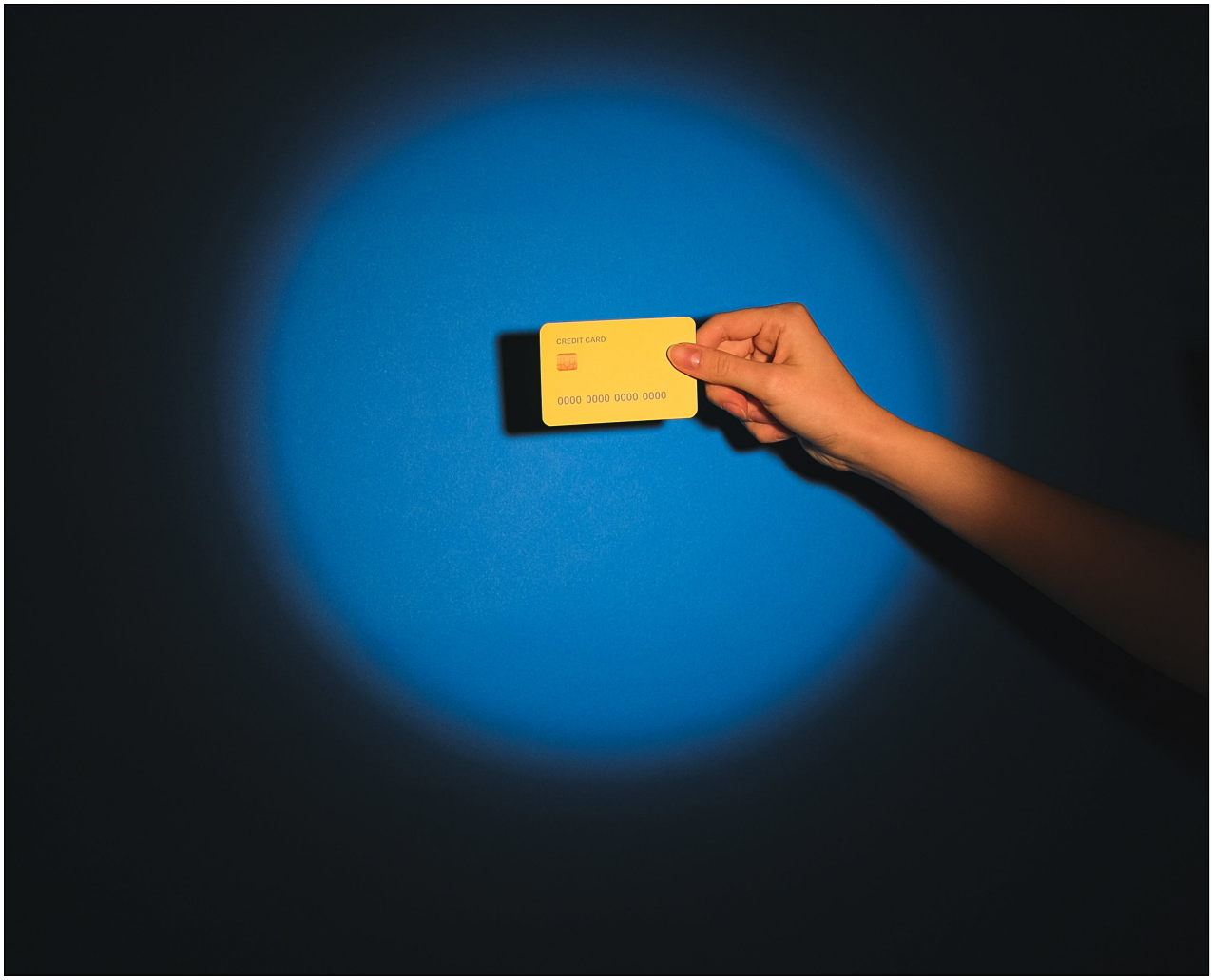

Project Proposal

Prepared for: Data Mining [CS634101] Project Midway Report

Prepared by: Venkata Sai Rahul Sunkari, Vinay Sai Teja Gamidi

10 November 2023

PROJECT

Credit card fraud detection

TEAM MEMBERS

1. Venkata Sai Rahul Sunkari
2. Vinay Sai Teja Gamidi

DATASET

The dataset selected to conduct the analysis and create a model is:
https://www.kaggle.com/datasets/kartik2112/fraud-detection

This dataset doesn't contain data from real time transactions, but it was simulated by the Kaggle user through a python simulator with a python library called 'faker'. The tool is Sparkov Data Generation which can be found here: https://github.com/namebrandon/Sparkov_Data_Generation .

What made us choose this dataset?
Although there were many other datasets, one very popular dataset on kaggle with 10k+ votes, this dataset resembled real time information rather than feature vectors as the data. This demanded understanding how to preprocess data suitable for model creation and analysis which will be evident from the notebook file.
Another important aspect is that the dataset is huge, imbalanced and the training data and test data came in two different files, which makes model testing and accuracy measurement more pragmatic.

WORK

The notebook file is attached along with this PDF file. Please note that it is not structurally organized yet as we're in the process of discovering various methods, processes and techniques to prep the data ready for model generation.

The following preprocessing tasks were applied for the fraudTrain.csv

- Checked for null values in the data frame - there weren't any.
- Removed duplicates if any.
- Identified from the class categorization that there are huge number of genuine transactions but only a few fraudulent transactions in comparison. This is a case of under-sampling - our goal is to use the rare events and balance the dataset to create a more accurate model.
- There are 7506 datasets that are fraudulent transactions, so we've sampled an equal number of datasets that are genuine.
- We converted the transaction date and time column to date format and created new columns - year and month, convert date of birth to date format and subsequently the age column. In addition, we've calculated the distance between the transaction location (latitude and longitude) and the merchant location.
- We're evaluating if normalization or standardization of the remaining data. The data that is irrelevant to generate the model will be dropped from the dataset.

CHALLENGES

- Initially it was unclear how to correlate the columns and predict a model based on data that wouldn't directly attribute to the frauds.
- Selection of the main algorithm for creating the model.
- Which columns to drop from the data frame, whether a dropped column would still contribute towards the model generation is unclear yet.
- Upon reading, we came across logistic regression, EDA, SMOTE etc. that are most appropriate for generating the model. But we lack to knowledge to implement them immediately.