CS6890: Fraud Analytics Using Predictive and Social Network Techniques
Assignment 5: Synthetic data generation using Variational Autoencoder

Archit Vivek Ganvir (CS21BTECH11005)
Maharshi Kadeval (CS21BTECH11027)
Abhinav Yadav (CS21BTECH11002)
Harsh Goyal (CS21BTECH11020)
Anshul Sangrame (CS21BTECH11004)

May 2024

1 Problem Statement

Given a set of credit card transactions, generate synthetic data that is similar to the given data.

2 Introduction

In this assignment, we use a variational autoencoder to generate synthetic data using the given dataset of credit card transactions.

An autoencoder is a type of artificial neural network used to learn efficient codings of unlabeled data (unsupervised learning). An autoencoder learns two functions: an encoding function that transforms the input data, and a decoding function that recreates the input data from the encoded representation.

A variational autoencoder maps an input point to a distribution within the latent space, rather than to a single point in that space.

3 Description of Dataset

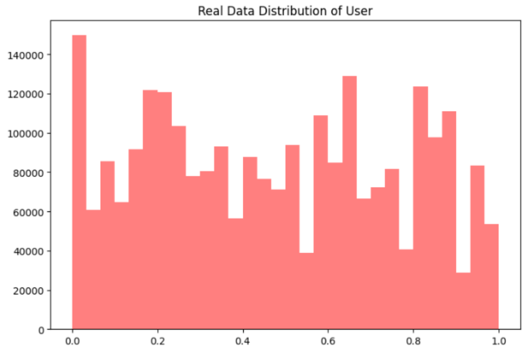
The transactions have 15 features: User, Card, Year, Month, Day, Time, Amount, Use Chip, Merchant Name, Merchant City, Merchant State, Zip, MCC, Errors?, Is Fraud?.

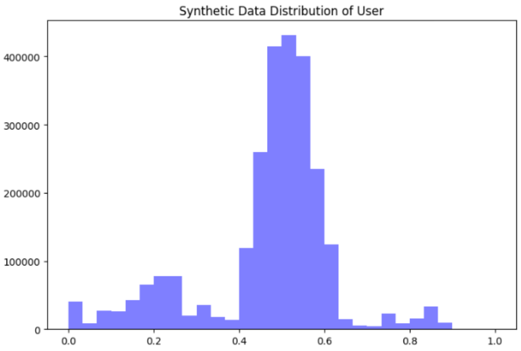
4 Procedure

- Label encoding is used for all the columns except Amount.
- The values in each column are normalized (by MinMax scaling).
- The variational autoencoder is designed with 15-7-4-7-15 units in each layer (input layer, 2 hidden layers for encoder, 1 hidden layer for decoder and 1 output layer) with ReLU as the activation function.
- The variational autoencoder is trained on the dataset with the adam optimizer, MAE (Mean Absolute Error) as the loss function and a batch size of 64 for 50 epochs.
- All the transactions in the dataset are reconstructed.
- The distribution of each column of the real as well as synthetic data is plotted.
- Some course-grained metrics are calculated.

5 Results

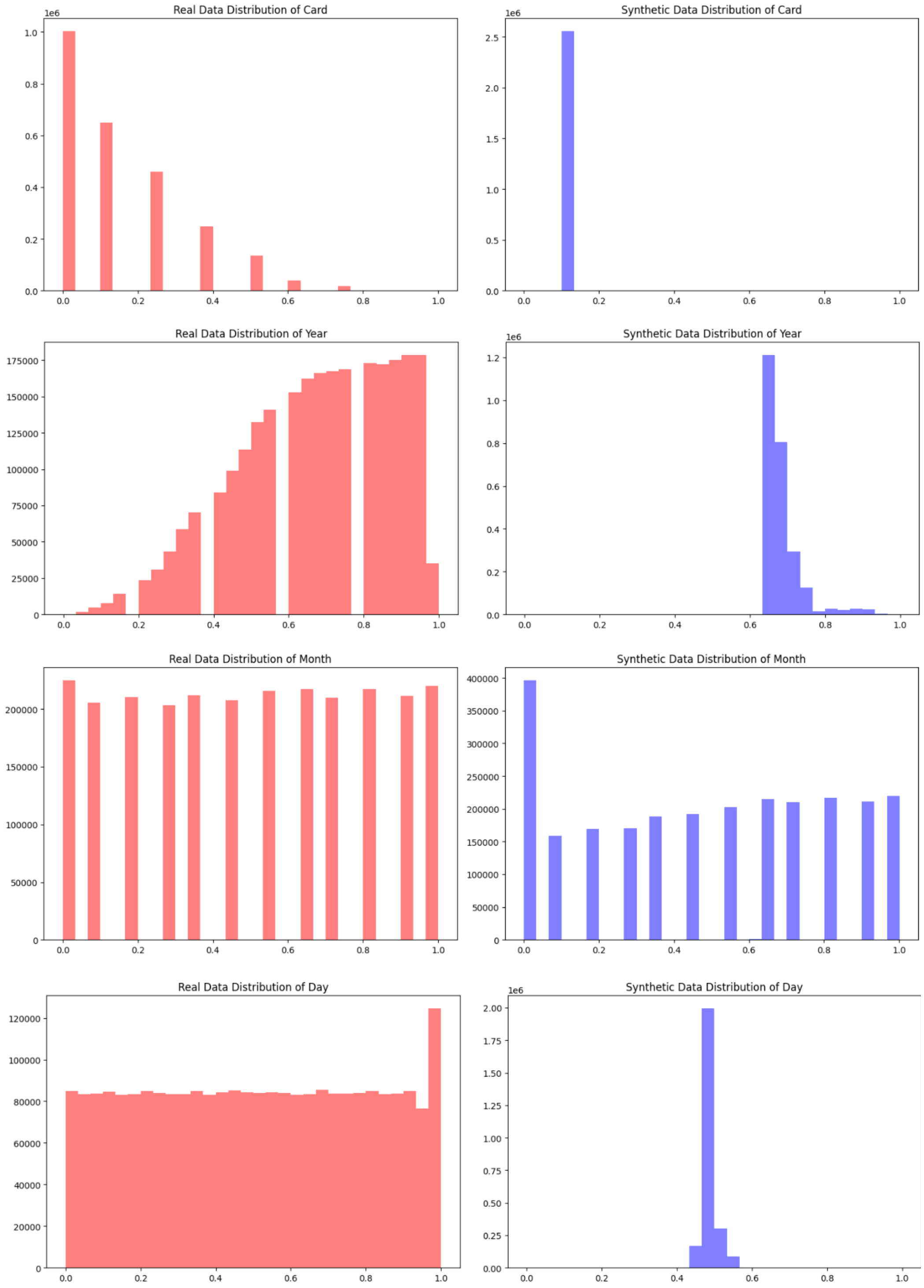
Following are the distributions of each column of the real as well as synthetic data:

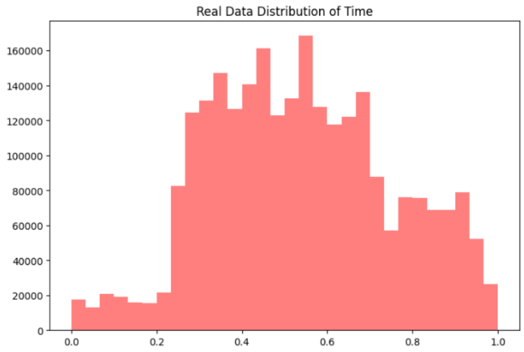


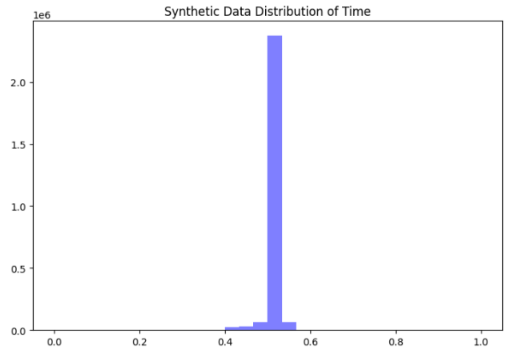


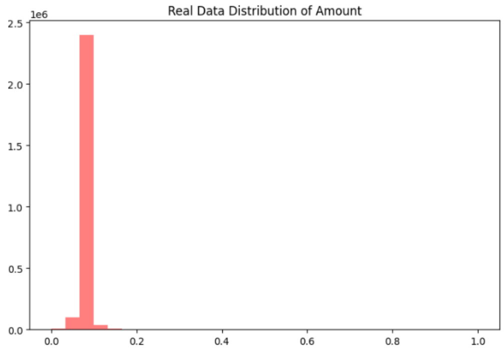
Following are some of the course-grained metrics that were evaluated:

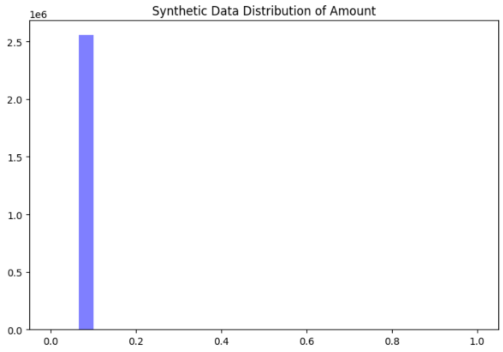
- % of data that is a direct copy of the real data: 0.00%
- % of data that is a self copy: 0.07%

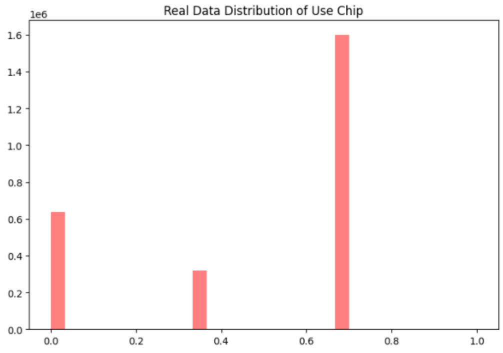

Real Data Distribution of Card
1e6
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Synthetic Data Distribution of Card
1e6
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Real Data Distribution of Year
175000
150000
125000
100000
75000
50000
25000
0
0.0
0.2
0.4
0.6
0.8
1.0
Synthetic Data Distribution of Year
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Real Data Distribution of Month
200000
150000
100000
50000
0
0.0
0.2
0.4
0.6
0.8
1.0
Synthetic Data Distribution of Month
400000
350000
300000
250000
200000
150000
100000
50000
0
0.0
0.2
0.4
0.6
0.8
1.0
Real Data Distribution of Day
120000
100000
80000
60000
40000
20000
0
0.0
0.2
0.4
0.6
0.8
1.0
Synthetic Data Distribution of Day
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0

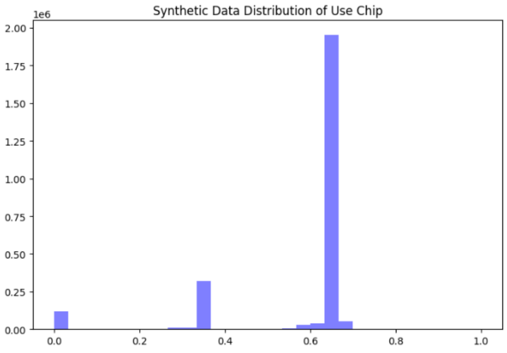

Real Data Distribution of Time
160000
140000
120000
100000
80000
60000
40000
20000
0
0.0
0.2
0.4
0.6
0.8
1.0

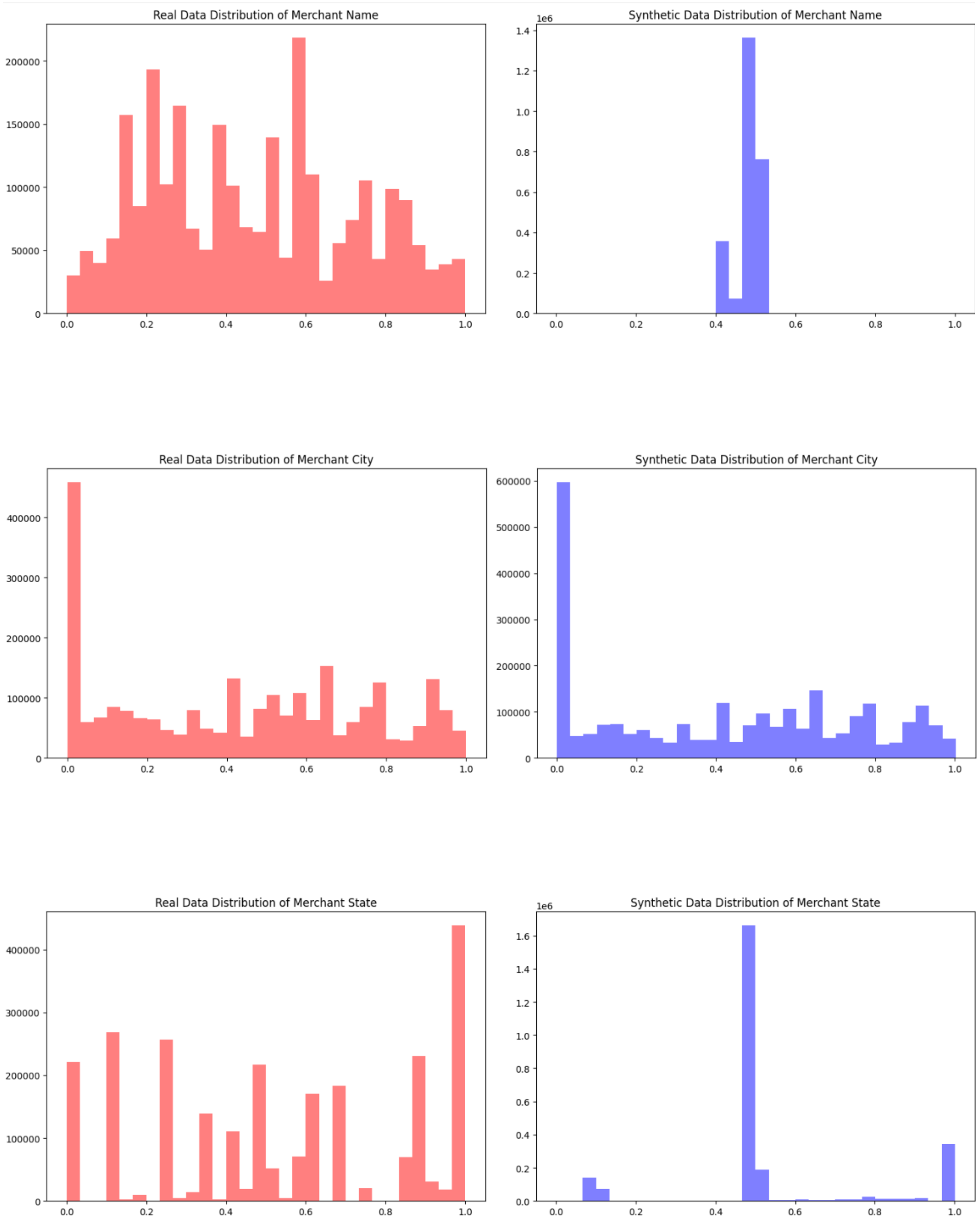

Synthetic Data Distribution of Time
1e6
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0

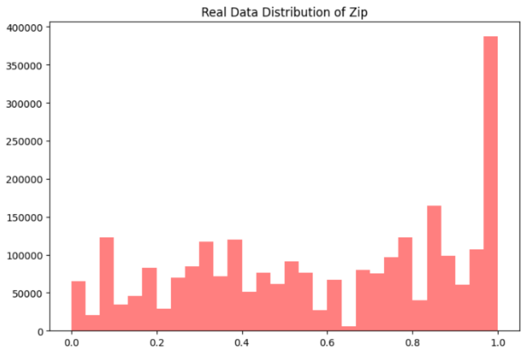

Real Data Distribution of Amount
1e6
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0

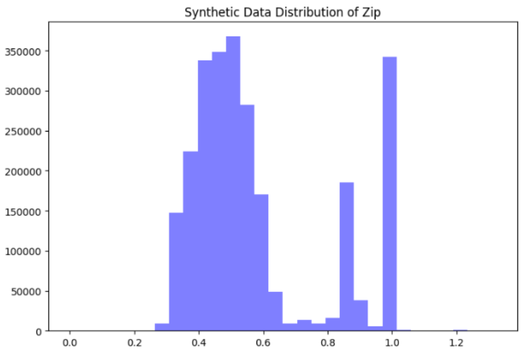

Synthetic Data Distribution of Amount
1e6
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0

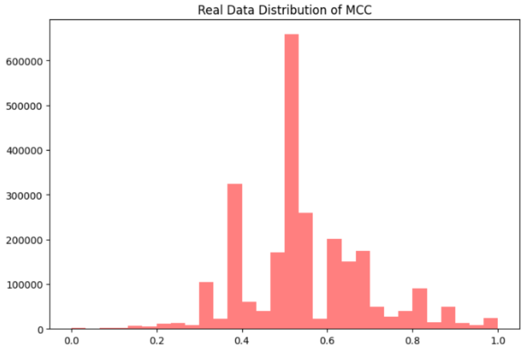

Real Data Distribution of Use Chip
1e6
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

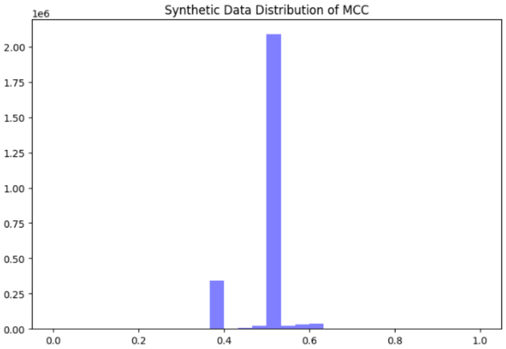

Synthetic Data Distribution of Use Chip
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0

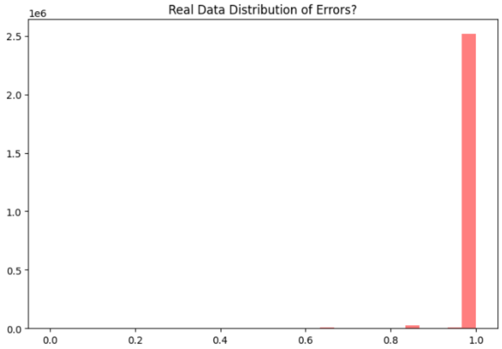

Real Data Distribution of Merchant Name
0
50000
100000
150000
200000
0.0
0.2
0.4
0.6
0.8
1.0
Synthetic Data Distribution of Merchant Name
1e6
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.0
0.2
0.4
0.6
0.8
1.0
Real Data Distribution of Merchant City
0
100000
200000
300000
400000
0.0
0.2
0.4
0.6
0.8
1.0
Synthetic Data Distribution of Merchant City
0
100000
200000
300000
400000
500000
600000
0.0
0.2
0.4
0.6
0.8
1.0
Real Data Distribution of Merchant State
0
100000
200000
300000
400000
0.0
0.2
0.4
0.6
0.8
1.0
Synthetic Data Distribution of Merchant State
1e6
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0.0
0.2
0.4
0.6
0.8
1.0

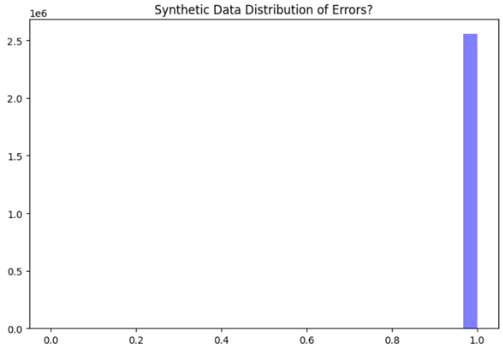

Real Data Distribution of Zip

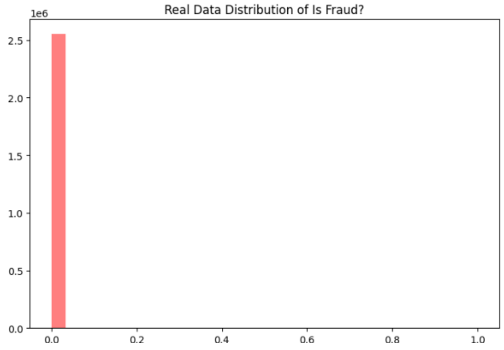

Synthetic Data Distribution of Zip

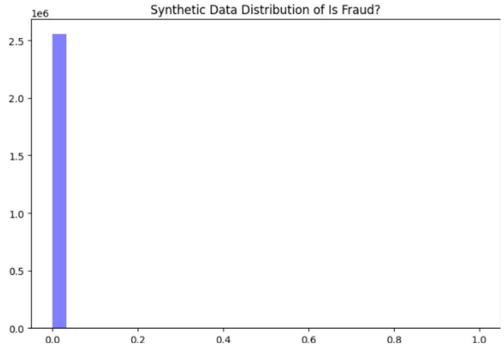

Real Data Distribution of MCC


Synthetic Data Distribution of MCC


Real Data Distribution of Errors?


Synthetic Data Distribution of Errors?


Real Data Distribution of Is Fraud?


Synthetic Data Distribution of Is Fraud?