National College of Ireland

Implementing Ensemble Method with stacking approach for Machine Learning and Deep Learning Algorithms for Credit Card Fraud Detection

MSc Research Project
Data Analytics

Charan Teja Marlabeedu
Student ID: X22161163

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Charan Teja Marlabeedu
Student ID:	X22161163
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Vladimir Milosavljevic
Submission Due Date:	31/02/2024
Project Title:	Implementing Ensemble Method with stacking approach for Machine Learning and Deep Learning Algorithms for Credit Card Fraud Detection
Word Count:	958
Page Count:	6

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Implementing Ensemble Method with stacking approach for Machine Learning and Deep Learning Algorithms for Credit Card Fraud Detection

Charan Teja Marlabeedu
X22161163

1 Introduction

The main purpose of this document is to explain regarding how the configuration setup has been done for the implementation of the research question. It also explains what tools have been setup and have been used for the implementation in this study. Furthermore going to explain in-detailed what python libraries have been utilised for the data visualisation, data imbalance, training and building the models and also what models have been utilised for the research study purpose.

All the configuration setups and the implementations has been explained section wise accordingly.

2 The System Configuration Used for the Implementation

The system configurations are as shown in Figure 1 where the implementation has been done.

Figure 1: System Configurations

3 Software and the Tools

The entire implementation has been done by using the python programming language which has been executed in Jupyter notebook which is a Integrated Development Environment(IDE) on the platform called Anaconda. The web browser used for the process is the "Google Chrome". The versions of the particular tool and software is as shown in Figure 2

```
!python --version
Python 3.9.13

!conda --version
conda 23.1.0

!jupyter --version
Selected Jupyter core packages...
IPython          : 7.31.1
ipykernel        : 6.15.2
ipywidgets       : 7.6.5
jupyter_client   : 7.3.4
jupyter_core     : 4.11.1
jupyter_server   : 1.18.1
jupyterlab       : 3.4.4
nbclient         : 0.5.13
nbconvert        : 6.4.4
nbformat         : 5.5.0
notebook         : 6.4.12
qtconsole        : 5.3.2
traitlets        : 5.1.1
```

Figure 2: Software and tools used and their versions

4 The Python Packages Utilised

In order to process the data, to analyse the data to run the models we have utilised the python packages.

Package	Purpose
Pandas	Data Manipulation and Analysis
NumPy	Numerical Computing
Matplotlib	Data Visualization
Seaborn	Advanced Data Visualization
Scikit-learn	Machine Learning Algorithms and Tools
Imbalanced-learn	Handling Imbalanced Data
TensorFlow	Deep Learning and Neural Networks
Keras	Neural Network Design and Training

Table 1: Python Packages and Their Purposes in Data Analysis

5 Data Collection

The datasets have been taken from the kaggle. The datasets taken are published in public domain so anyone can use for the research purpose as per the guidelines. The dataset consists of two csv files. One is fraud_train data and second one is fraud_test data file.

6 Data Loading and Data Preprocessing

6.1 Data Upload and Loading Process:

6.1.1 Data Upload:

The two csv datasets fraudTrain.csv and fraudTest.csv have made availabe in the working directory of the Jupyter notebook for the analysis as shown in Figure 3. It has been uploaded in the jupyter notebook working directory through manual upload.

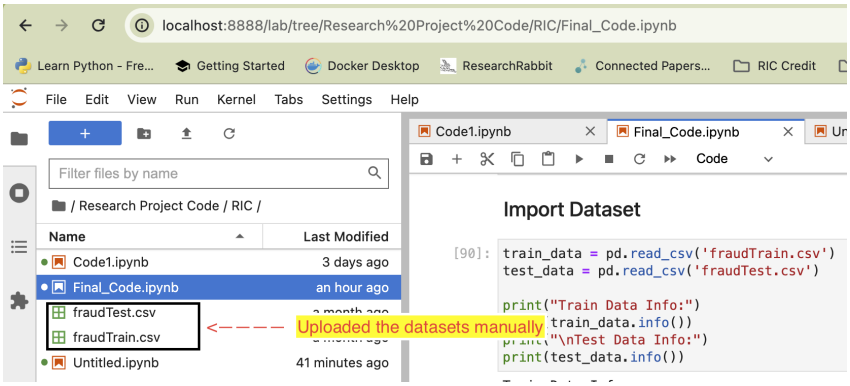

Figure 3: Uploaded manually into working directory the datasets

6.2 Data Loading:

6.2.1 Pandas:

In order to fetch or load the data for running the program pandas are the primary tool as shown in Figure 4.

6.3 Data Pre-processing Tools and Their Purposes:

6.3.1 Pandas:

Used for initial data exploration, cleaning, and transformation. Functions like info(), describe(), and head() are used for inspecting the data structure and content. Pandas also provides capabilities for handling missing values and duplicates, which are essential for data quality.

6.3.2 NumPy:

Employed for numerical operations that are more complex or not available directly in Pandas. It's particularly useful for operations on array data.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Import Dataset

```
train_data = pd.read_csv('fraudTrain.csv')
test_data = pd.read_csv('fraudTest.csv')
```

Figure 4: Data Loading/Import

6.3.3 Scikit-learn:

In order to split the datasets for the training and testing purpose we are going to use the Scikit-learn library.

Based on all these we can make a clear analysis for the model training.

7 Explorative Data Analysis

7.1 Data Visualization

In order to do the visualisation we are going to use the Matplotlib and Seaborn libraries which are very critical for doing the analysis as shown in Figure 5. Which helps in understanding the outliers and as well the parameters that needed to be considered for the model.

```
import matplotlib.pyplot as plt
import seaborn as sns
```

Figure 5: Libraries imported for the Visualisation

7.2 Data Pre-processing Tools and Their Purposes:

These libraries have been utilised in order to make some of the critical visualisations for the final analysis like the correlation matrix as shown in Figure 7, no of fraud transactions as shown in Figure 6, how the fraud transactions has been happening over certain period of time as shown in Figure 8, fraud transactions as per the area as shown in Figure 9 all these visual analysis have been helpful for selecting the parameters and training the models.

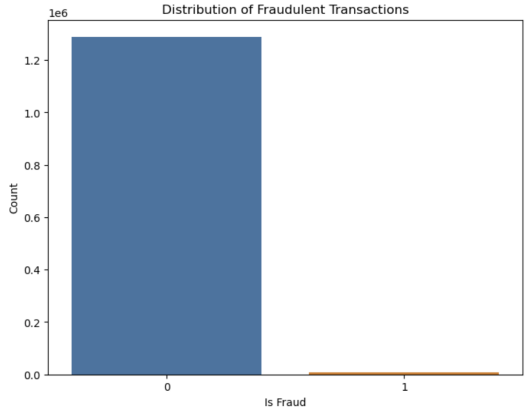

Figure 6: The visualisation of total fraud transactions

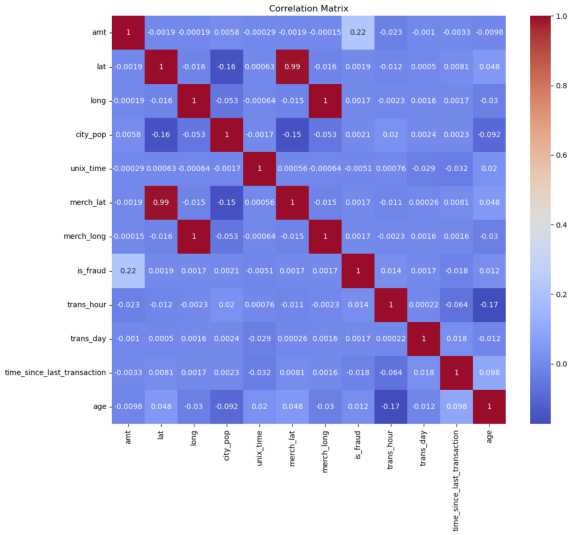

Figure 7: The visualisation of the correlation matrix between the variables

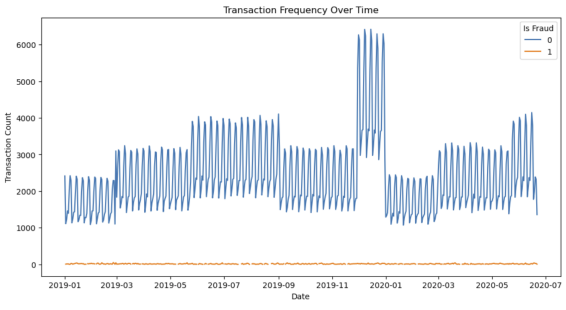

Figure 8: The visualisation fraud transactions over period of time

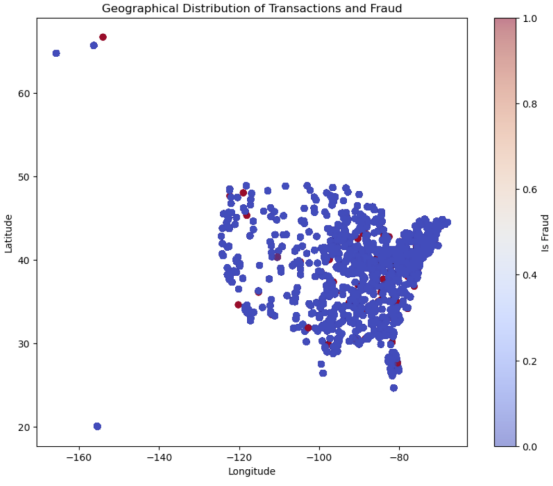

Figure 9: The visualisation of the fraud over areas

8 Data Training and Model Implementation & Evaluation

In this section we will be discussing how we have configured for training the data what all the libraries and packages that has been utilised in order to do for data training and also implementing various models for detecting the credit card fraud transactions.

The libraries that has been imported for data training and model implementation are as shown in Figure 10

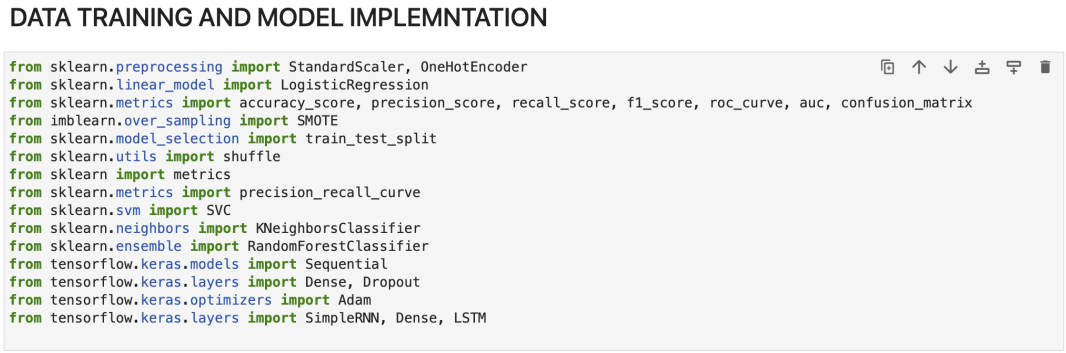

Figure 10: Libraries imported for data training and model implementation

Here as shown in Figure 10:

- We have imported the SMOTE from imblearn package in order to perform the data imbalance.
- We have imported the StandardScaler for adjusting the features with standard range and also imported OneHotEncoder for converting the features which are categorical into numerical values.
- We have imported all the machine learning model libraries from respective packages and also the deep learning models libraries from respective packages for implementing the models.
- We have imported the evaluation metric libraries in order to calcualte the implemented model performance based on those metric values.

9 Conclusion

As per the research question as discussed above those are all the system configurations and setup that has been made initially to excute the implementation. And all those software tools and the programming language, packages and libraries have been utilised for implementing the models as per my research question. Finally I would like to conclude saying that by making all the above configuration I had made a very significant progress for my research question implementation.