National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Christy Davis Maliyekkal
Student ID: x22151648

School of Computing
National College of Ireland

Supervisor: Teerath Kumar Menghwar

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Christy Davis Maliyekkal
Student ID:	x22151648
Programme:	Data Analytics **Year:** 2024
Module:	MSc Research Project
Lecturer:	Teerath Kumar Menghwar
Submission Due Date:	31/01/2024
Project Title:	Predicting Customer Lifetime Value (CLV) in UK and Brazil using Machine Learning and Deep Learning: A Comparative Analysis
Word Count:	536 **Page Count:** 4

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:	Christy Davis Maliyekkal
Date:	31/01/2024

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):	

Configuration Manual

Christy Davis Maliyekkal
Student ID: x22151648

1 Data Description

For the Research named "Predicting Customer Lifetime Value (CLV) in UK and Brazil using Machine Learning and Deep Learning: A Comparative Analysis", used 2 different datasets, one from Kaggle and another one from UCI Repository website. The detailed description is given below;
[1] Name of the Dataset: Online Retail

Description: This dataset consisting of a wide collection of purchases that occurred for an online retail company based in UK. The company markets mainly all occasion gifts. The majority of the clients of this company is wholesalers. Data Consist of 8 attributes and 541909 entries.
URL of the location of the dataset: https://archive.ics.uci.edu/dataset/352/online+retail
Dataset size: 22.6 MB

[2] Name of the dataset: Brazilian E-commerce public dataset by Olist

Description: This is a Brazilian public E-commerce dataset of Olist store, consists the orders purchased at the store. Data includes information of orders made from multiple Brazilian markets. Its attributes consist of the price, order status, payment, freight performance customer location, reviews by customer and finally product attributes.
URL of the location of the dataset: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?resource=download
Dataset size: 42.6 MB

2 Implementation of the Data in Code

After selecting the relevant datasets, we moved to the implementation part. We downloaded the two datasets to the device. The datasets now located in downloads of the device. Online Retail dataset (UK) showed a size of 22.6 MB on device and Brazilian E-commerce public dataset by Olist showed a size of 42.6 MB on device. The Online Retail dataset is in XLSX worksheet format and the Brazilian dataset is a file with 9 separate datasets in XLX worksheet format. The system that we have used is 8 GB RAM. System OS: Windows 10.
For the further execution of the code, we used Jupiter Notebook (anaconda 3). Firstly, we created a folder named 'Final Project' in Jupiter Notebook. Then from the Upload option we uploaded the two datasets. As already mentioned, Brazilian dataset consists of 9 separate datasets, from this we have only uploaded 6 relevant datasets in CSV file format. Olist_customers_dataset.csv, olist_geolocation_dataset.csv, olist_order_items_dataset.csv, olist_order_payments_dataset.csv, olist_order_reviews_dataset.csv, olist_orders_dataset.csv

are the 6 datasets that we have uploaded. The Online Retail dataset we uploaded as xlsx file format named Online Retail.xlsx. For the code implementation we created a new python 3 notebook named 'CLV_Analysis' in the same folder 'Final Project'.

Figure .1. Folder created in Jupyter Notebook

3 Necessary Libraries and the Execution

There are certain libraries that are important for the implementation of the code, with the help of our datasets. Pandas, numpy, matplotlib.pyplot, seaborn, xgboost, sklearn.metrics, sklearn.model_selection, sklearn.preprocessing, sklearn.compose, sklearn.pipeline, sklearn.ensemble.RandomForestRegressor, sklearn.neural_network.MLPRegressor, joblib, tensorflow.keras.models.sequential, tensorflow.keras.layers.Dense. These are the libraries that we have used in this project.

3.1 Brazilian Data

importing the necessary libraries and preprocessing the data

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import xgboost as xgb
from sklearn.metrics import mean_squared_error, mean_absolute_error, mean_absolute_percentage_error
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.metrics import r2_score
import numpy as np
from sklearn.metrics import mean_squared_error, mean_squared_error, r2_score
import warnings
warnings.filterwarnings('ignore')

#Brazil data consist of 9 seperate datasets, from the loading the necessary datasets
# Load the datasets
customers = pd.read_csv('olist_customers_dataset.csv')
geolocation = pd.read_csv('olist_geolocation_dataset.csv')
order_items = pd.read_csv('olist_order_items_dataset.csv')
order_payments = pd.read_csv('olist_order_payments_dataset.csv')
order_reviews = pd.read_csv('olist_order_reviews_dataset.csv')
orders = pd.read_csv('olist_orders_dataset.csv')

# Data Cleaning
# Remove duplicate rows
customers.drop_duplicates(inplace=True)
geolocation.drop_duplicates(inplace=True)
order_items.drop_duplicates(inplace=True)
order_payments.drop_duplicates(inplace=True)
order_reviews.drop_duplicates(inplace=True)
orders.drop_duplicates(inplace=True)
```

Figure.2. Necessary Libraries and preprocessing

CLV Calculation

```
2]: # Merge relevant datasets to create a comprehensive dataset
    merged_data = pd.merge(orders, customers, on='customer_id')
    merged_data = pd.merge(merged_data, order_items, on='order_id')
    merged_data = pd.merge(merged_data, order_payments, on='order_id')

    # Select relevant features and target variable (CLV)
    features = merged_data[['customer_id', 'order_purchase_timestamp', 'price', 'payment_value']]
    target = merged_data[['customer_id', 'payment_value']]

    # Calculate CLV for each customer
    clv = target.groupby('customer_id')['payment_value'].sum().reset_index()
    clv.rename(columns={'payment_value': 'CLV'}, inplace=True)

    # Merge CLV back into merged_data
    merged_data = pd.merge(merged_data, clv, on='customer_id')
    merged_data.head()
```

Figure.3. CLV Calculation

Customer lifetime value is calculated after importing libraries and preprocessing. The CLV is merged back to the data.

- Then implemented the machine learning models (xgb regressor, SVM, Random Forest) and Deep learning model (MLP Regressor)
- Evaluated the models using plot (Actual CLV Vs Predictions)

3.2 UK Data

```
# importing necessary libraries and data preprocessing
```

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.neural_network import MLPRegressor
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error, mean_absolute_percentage_error
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
import seaborn as sns
import xgboost as xgb
import joblib
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')

# Load the dataset
data = pd.read_excel('Online Retail.xlsx')

# Filter data for the United Kingdom
data_uk = data[data['Country'] == 'United Kingdom']

# Data Preprocessing
# Feature Engineering
data_uk['TotalPurchase'] = data_uk['Quantity'] * data_uk['UnitPrice']

# Group by 'CustomerID' and calculate the sum of 'TotalPurchase' as CLV
features = data_uk.groupby('CustomerID')['TotalPurchase'].sum().reset_index()
features.rename(columns={'TotalPurchase': 'CLV'}, inplace=True)

# Merge CLV back into the original dataset
data_uk = pd.merge(data_uk, features, on='CustomerID')
```

```
data = pd.read_csv('data.csv')

# Check for missing values
missing_values = data.isnull().sum()

# Check for duplicate rows
duplicate_rows = data.duplicated().sum()
```

Figure.4. Necessary Libraries and Preprocessing, CLV Calculation

- Then implemented the machine learning models (xgb regressor, Random Forest) and Deep learning model (MLP Regressor)
- Evaluated the models using plot (Actual CLV Vs Predictions)