National College of Ireland

Configuration Manual

AI-Powered Improvement in Inventory Management for E-Commerce Supply Chains

MSc in AI for Business

Rana Shehbaz Khan
Student ID: 23163054

School of Computing National College of Ireland

Supervisor: Anderson Simiscuka

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Rana Shehbaz Khan
Student ID:	23163054
Programme:	MSc in AI for Business **Year:** 2023-24
Module:	Practicum Thesis
Lecturer:	Anderson Simiscuka
Submission Due Date:	12 August 2024
Project Title:	AI-Powered Improvement in Inventory Management for E-Commerce Supply Chains
Word Count:	1000 **Page Count:** 03

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: Rana Shehbaz Khan

Date: August 12 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

Rana Shehbaz Khan
Student ID: 23163054

1 Introduction

This configuration manual aims to offer specific instructions on how to set up the environment, configuration used to process the Olist dataset together with the signals used for analysis and the execution of the code used for the analysis and prediction of outcomes. The project was carried out on Google Colab platform using Python which offers extensive libraries to carry out the data acquisition, processing, visualization as well as modeling. This manual includes the list of required software and hardware, data pre-processing, and application of different kinds of machine learning algorithms.

2 System Configuration

I then discuss the hardware and the software that should be used in order to duplicate the project.

2.1 Hardware Configuration

The project was implemented on Google Colab and, therefore, does not require special hardware characteristics expect for the stable Internet connection and a contemporary Web browser like Google Chrome.

2.2 Software Configuration

The following software components were utilized:The following software components were utilized:

2.2.1 Python

Python 3. x was the dominant language travelled in programing. The code used several dependencies to import libraries that carry data manipulation, visualization, and machine learning algorithms functions.

2.2.2 Google Colab

For development and executing the model, Google Colab environment was taken which is a cloud-based Jupyter notebook with provision of GPUs and TPUs.

2.2.3 Pandas, NumPy IE Matplotlib

These libraries were employed for manipulating and visualizing data derived from the study's quantitative data analysis. Pandas deals with the data frame structures, NumPy is for the numerical computations and Matplotlib along with Seaborn is used to plot the graphs.

2.2.4 Scikit-Learn

This library was employed in the preprocessing step, in building the model as well as in model assessment. It contains a scaler for scaling data, a method for grid search, and RandomForestClassifier and LogisticRegression for a machine learning algorithm.

2.2.5 LightGBM and XGBoost

These libraries used in the gradient boosting models, and these models used for the classification problem.

2.2.6 Imbalanced-learn

Imbalanced datasets were managed using the SMOTE (Synthetic Minority Over-sampling Technique) from this very library.

3 Data Preparation

The Olist dataset is several CSV files with distinct contextual features of the Olist e-commerce firm, including customers, orders, products, and reviews. The data was then copied from the Google Drive and pasted in the Colab environment were all the processing was to be done.

3.1 Loading the Dataset

The datasets were loaded into Pandas DataFrames:The datasets were loaded into Pandas DataFrames:

```
from google. colab import drive
drive. mount('/content/drive')
raw_path = '/content/drive/MyDrive/datasets/archive (1)/'
olist_customer = pd. read_csv(raw_path +'olist_customers_dataset. csv')
# Similar data were loaded to other datasets as well. ..
```

3.2 Preprocessing

- **Handling Missing Values:** It has functions to check and manage the null values especially in the review comments section.
- **Datetime Conversion:** The columns relevant to time were transformed into datetime type for better time analysis.
- **Merging Datasets:** Several dataframes were combined to generate a consolidative customer order product and seller Data frame.

4 Implementation

This section outlines procedures by which machine learning models can be deployed so that results for the Olist dataset can be predicted.

4.1 Feature Engineering

- **Distance Calculation:** Haversine formula was used for Hypothesis 4 to determine the distance of customers and the sellers.
- **Time Features:** Some of these variables included shipping time and order approval time on which computations and analysis were made.

4.2 Model Training

- **Pipelines:** They were named by the models that they were intended to support (LightGBM, RandomForest, LogisticRegression, XGBoost) to achieve greater data preprocessing and model fitting uniformity.

- **Grid Search:** Hyperparametes used for each of the models used was tuned from from the best value based on a Grid Search with Stratified K-Fold Cross-Validation.

4.2.1 Example of Model Training with Grid Search

```
grid_search = GridSearchCV(pipeline, param_grids[model_name], cv=skf,
scoring='f1_weighted')
grid_search. fit(X, y)
```

5 Evaluation

With respect to this table, the models were trained and tested using validation set in terms of accuracy, precision, recall and F1-score. T-Test and Formula 9 were also applied to test the hypothesis and confusion matrices were also plotted for visual representation of the classifiers.

5.1 Evaluation Metrics

Therefore, in order to predict the efficiency of the models, Accuracy, Precision, Recall, and F1-Score were computed for each of the models.

CROSS TABULATIONS were done to get confusion matrices which were plotted for further analysis.

6 Conclusion

The following configuration manual contains all that is needed in order to achieve the Olist dataset analysis and prediction project Using Google Colab. The worked example illustrates the use of different machine learning techniques to make predictions out of e-commerce data.

7 References

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.
2. Keras. (n.d.). *Keras Documentation*. Retrieved from https://keras.io/
3. TensorFlow. (n.d.). *TensorFlow Documentation*. Retrieved from https://www.tensorflow.org/
4. Google. (n.d.). *Google Colab*. Retrieved from https://colab.research.google.com/