National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Sean Dwyer
Student ID: 22124314

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Sean Dwyer
Student ID:	22124314
Programme:	Data Analytics
Year:	202024
Module:	MSc Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	12/08/2024
Project Title:	Configuration Manual
Word Count:	712
Page Count:	3

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:	15th September 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

Sean Dwyer
22124314

1 Introduction

This document contains detailed instructions as to how to install and run the ICT/code aspect of the research project.

2 Prerequisites

This manual assumes that there is an installation of Docker available to use. PostgreSQL will be installed using Docker.

Python 3.9+ - preferably 3.12.2

Use will be made of Python virtual environments.

3 Database install

To install and run PostgreSQL (in Docker), please use this command

docker pull postgres

docker run –name postgres-default -d -p 2022:5432 -e POSTGRES_PASSWORD=postgres postgres

note, it is important that this username and password is used.

4 Project folder - activate

The main folder is Project folder:

Open Terminal/Command prompt here and type the following (not in brackets)

(if you don't already have virtualenv installed) pip install virtualenv

to create your new environment (called 'venv' here) virtualenv venv

(activate the virtual environment) source venv/bin/activate

5 Create database and python requirements

In the project folder (command line), after environment is activated

navigate to the **src** folder

type

project_master.cmd this install all python required packages (via requirements.txt).

this connects to the default docker installation of PostgreSQL and creates the database (horseracedata).

6 Project/src and Project/data folders

6.1 src folder

This folder contains all of the python code, and notebooks required to run the solution.

the four python files contain functions pertaining to their names.

They are:

1. config.py - this is a configuration object that is initailized with database settings from the config.yaml file.

2. dbfunc.py - this contains connection string and connection object functions that are used by the python notebooks.

3. createDatabase.py - this is the module that creates the horseracedata database on PostgreSQL database server (only of not already existing.

4. racingbetdata.py - this contains functions used in python notebooks that read from excel and csv files, and write to the databases.

6.2 data folder

This is the 'working' data folder of the solution. All interim data files created will be done so in here.

7 Main project files - Project/src

Because of ease-of-use and secondly because there are some visualizations necessary, the python notebooks are chosen. The methodology aligning to the adapted CRISP-DM process is followed by the number of these notebooks. Each with its own part of the process.

They are designed to be run through in order from 00 - Data Load to 05 - Evaluation, each contributing to the rlevant part of the research.

7.1 00 - Data load

This contains the database load functionality that loads from raw data (in Project/data/01_raw into the database tables.

7.2 01 - Data Exploration

this takes data from the database tables into the pandas dataframe structure saving to the data directory for the subsequent step. There is some prelininary feature engineering in this step. It also provides visual analysis of the structures as input into the document.

7.3 02-Data Preparation and Feature engineering

IN this step, there are further engineered features, and the data is prepared for modelling. That is, the data is split into training, testing, and validation sets.

7.4 03-Feature-Selection

The sole function of this step is to provide feature importance models to the overall process so the feature selection process can be optimized.

7.5 04-Implementation

This is the step that creates, trains, and tests the accuracy of the models. The output of this step is a series of predicted dataframes from the validation dataset. This will be evaluated in the next, and final step.

7.6 05-Evaluation

This final step of the process contains the crux of the research. It takes each of the models' predictions and evaluates them against real or actual values. Then, using those results, for each model it applies a betting strategy to assess the capability of each model in a financial aspect. And finally, it uses the tipster dataset to compare the performance of the human against a machine learning model counterpart.

References