COMP4801 Final year project

Comparative Analysis of Machine Learning Algorithms for Predicting Horse Racing Outcomes

Detailed Project Plan

Hong Yuk Sing, 3035927257
Group: FYP24058

Supervisor:
Prof. Difan Zou

1. Background

Horse racing, a widely favoured form of gambling, enjoys substantial popularity in various regions worldwide, including Hong Kong, Japan, and Great Britain. Data from the Hong Kong Jockey Club [1] indicate that the total wagering on horse racing reached approximately HKD 136.1 billion in the fiscal year 2023-24. Prior to each race, a "racing booklet" is published, containing essential information about the forthcoming event. This booklet includes details about the participating horses, jockeys, the track conditions, and historical race outcomes, all crucial for formulating predictions about race results and guiding more informed betting decisions. The informed and data-rich nature of horse racing favours the use of scientific approaches for systematic prediction. In academia, horse racing serves as a testbed for machine learning scientists to present and compare the latest machine learning algorithms.

In this research, several machine learning techniques will be employed to develop different machine learning models that predicts horse racing outcomes. The objective of this project is to construct a model that can accurately predict these outcomes, thereby assisting in making informed betting choices. Additionally, the study will investigate the factors that significantly influence race outcomes and evaluate the performance of the model when various deep machine algorithms are implemented.

2. Objective

The structure of this project will be composed as follows. In stage 1, a single learning model will be trained. This stage aims at constructing a dataset, which involves data collection and cleaning work. The outcome of this stage will be used to support stage 2 study, which is focus on a comparative analysis of the performance of a selected set of machine learning algorithm.

2.1 Understand the historic racing data

For every machine learning project, it is essential to delve in to the data, recognizing fields which are significant to the prediction . This understanding will help in refining the data, eliminating irrelevant or noisy information, and setting a solid foundation for predictive accuracy.

2.2 Optimization of the machine learning model

To improve the performance of the machine learning model, extensive work on optimization, such as fine-tuning and weight initialization, has to be carried out. Additionally, the use of cloud computing can be considered to meet the requirements for computational resources during training. Cloud platforms offer advantages in terms of scalability and often provide GPUs specialized for model training, which can reduce training times and offer a cost-effective solution for project management.

2.3 Explore the performance of different machine learning algorithm

The second stage of the project entails a systematic comparative analysis of various machine learning algorithms. A select group of algorithms will be studied to determine which performs best under different conditions relevant to horse racing competitions. Factors like model complexity and computational resource demands will be considered. This comparative analysis will not only identify the model with the best overall performance but also provide deep insights into each model's strengths and weaknesses. The outcomes of this study will enhance our understanding of how to optimize the use of different machine algorithms for prediction tasks, applicable not only to horse racing but also to daily practical scenarios.

3. Methodology

3.1 Data collection and understanding

Horse racing data are available publicly for access via the Hong Kong Jockey Club website or several third-party data providers such as hkHorseDb. These race record contains features such as date, list of horses, jockey, track, distance and odds. The data will be collected will be saved to a database for the model training.

3.2 Data preparation

Before passing the data for model training, data cleaning has to be carried out to integrate and standardised to create a dataset, this process will involve process such as adding id to label horses and jockey, using encoding

techniques to convert non numerical data to number value and standardized the data format.

3.3 Modelling and Evaluation

Machine learning model should be built using cloud GPUs or GPU farm and trained using the prepared dataset, which consisted the train dataset and testing dataset for validating the model's performance. The model's performance will be evaluated through standard metrics such as accuracy and loss measurements. In addition, the profit of gambling will be evaluated based on the betting odds.

3.4 Comparative analysis

The comparative analysis is a critical component of this research project, aiming to assess and compare the effectiveness of different learning algorithms in predicting horse racing outcomes. This analysis not only helps in identifying the most suitable algorithm but also provides insights into the strengths and weaknesses of each approach under various conditions.

4. Schedule and Milestones

Period	Work Description
Sep – Oct, 2024	• Development of a detailed project plan • Establishment of the project web page
Oct 1, 2024	*Deliverables of Phase 1* • *Detailed project plan* • *Project web page set up*
Oct - Nov, 2024	**Stage 1: Data Development** - Collecting data from racing database - Data cleaning and standardization - Constructing training and testing datasets
Nov – Dec, 2024	**Stage 1: Model Development** - Building and training the machine learning model using a selected machine learning method - Hyperparameter tuning to optimize model performance - Evaluating the model accuracy and performance metrics
Jan, 2024	- Preparing interim report and presentation
Mid-Jan, 2025	*First presentation*
Jan 26, 2025	*Deliverables of Phase 2* • *Preliminary implementation* • *Detailed interim report*
Feb, 2025	**Stage 2: Model Enhancement and Comparative analysis** - Selecting a set of machine learning algorithms for comparative analysis - Training of the model using the shortlisted algorithms - Collecting and evaluating of model performance data

Mar - Apr, 2025	- Source code clean-up to ensure clarity and maintainability
Late-Apr, 2025	*Final presentation*
Late-Apr, 2025	*Project exhibition* • *3-min video*

5. References

[1] The Hong Kong Jockey Club, “Annual report for the year ended 30 June 2024,” 2024. [Online]. Available: https://corporate.hkjc.com/corporate/english/history-and-reports/2024/index.aspx#1. [Accessed: Sep. 23, 2024]