HARNESSING METRICS: PREDICTING OVER AND UNDER PERFORMANCE IN RACEHORSES USING COMPREHENSIVE DATA ANALYSIS

Royal Melbourne Institute of Technology

Student Names and IDs
Tanish Saajan s3940991
Yong Pui Tung s3934929
Neil Satish Nair s3959956

Project Supervisor
Josh Kadlec-Cavanagh (Ciaron Maher Racing – Industry Supervisor)
Dr. Alice Johnstone (RMIT – Academic Supervisor)

Date of Submission
16th June 2024

Acknowledgement

We extend our heartfelt appreciation to our WIL (Work Integrated Learning) course coordinator, Dr. Yan Wang, and our academic supervisor, Dr. Alice Johnstone, whose invaluable guidance, and expertise have been instrumental in the successful completion of this project. Their insights and support have been crucial throughout this process.

We would also like to thank Josh Kadlec-Cavanagh, without whose guidance this report would not have been possible. His expertise in the field of sports analytics and the horse racing industry has not only helped this report immensely but also improved our understanding of the field.

Finally, we are grateful to our families and friends for their unwavering support and encouragement throughout our academic journey.

Disclaimer

We declare the following to be our work, unless otherwise referenced, as defined by the University's policy on plagiarism. All the work taken from the different sources has been cleared referenced as per the required referencing standards. We declare that we ensured our academic integrity.

Table of Contents

Executive Summary

Training is an important aspect of horse racing. The performance of a horse on race day largely depends on its training and its performance during training. In this report, we analyze the horses' training data and predict their performance on race day. The performance of the horses has been categorized into two parts: 1 (overperform) indicates a horse securing a 1-3 position in a race, while 0 (underperform) indicates the horses securing any positions after 3.

This project investigates several candidate models for predicting the performance of horses. Before comparing different models, we preprocessed the data and performed feature selection using the chi-square test and manual interpretation of the features. We then used cross-validation as a metric, which provides a robust estimate of a model's performance by averaging the results over multiple folds. This approach helps us understand how well the model is likely to perform on unseen data.
We combined the results for all the models and, through visual analysis, finalized the model with the highest accuracy. Subsequently, we analyzed the performance of the model on both training and testing data to check for overfitting or underfitting.
After the final analysis, we used the final model to make predictions for the horseid in our dataset. To assist the trainers in improving the overall racing performance, we have developed an interactive model that provides an in-depth analysis of all the selected metrics. This analysis includes identifying the optimal track surfaces, handling styles, and locations where each horse performs best. By leveraging these insights, trainers can make data-driven decisions to optimize training regimens and race entry strategies, ultimately enhancing the likelihood of success on race day.

Glossary

Sr. No.	Variable	Description
1	maxstridefreq	Maximum stride frequency achieved during the training session.
2	maxheartrate	Maximum heart rate recorded during the training session.
3	heartrate3minutes	Heart rate measured three minutes post-exercise.
4	tracksurface	Type of track surface on which the horse trained (e.g., sand, all weather, grass, dirt).
5	maxspeed	Maximum speed reached during the training session.
6	heartrateendpercent	Heart rate at the end of the session as a percentage of the maximum heart rate.
7	result	The outcome of the race (e.g., 1 for winning positions, 0 for non-winning positions).
8	weight	Weight of the horse.
9	age	Age of the horse in years.
10	sex	Sex of the horse (e.g., gelding, colt, horse, rig for males; filly, mare for females).
11	temperature	Temperature at the time of the training session.
12	hand	Direction of the track (e.g., right-handed, left-handed).
13	state	The state or location where the training session took place.
14	intensity	Metric intensity of the training session.
15	barrier	Barrier position from which the horse started.

Introduction

Background

In the high-stakes world of Thoroughbred horse racing, where vast sums are invested in breeding, training, and competing, the ability to accurately predict a horse's performance can be the difference between victory and defeat, profit and loss. This research, conducted in collaboration with Ciaron Maher Racing (CMR), a leading Australian training organization renowned for its innovative approach and cutting-edge facilities, seeks to leverage the power of comprehensive data analysis to develop predictive models for Thoroughbred race performance. Thoroughbreds, a breed known for their speed and agility, are central to the sport of horse racing. This study will focus on studying Thoroughbred being trained under CMR.

Aim and Objectives

By leveraging CMR's extensive real-world training and racing data, this study aims to uncover patterns and relationships between training factors and racing performance that can inform accurate predictions of over and underperformance during the race. Using statistical analysis, this study will explore and validate predictive models using existing metrics and engineered features to forecast race performance. Accurate predictive models can be crucial for stakeholders in Australian horse racing industry, including horse owners and trainers, as it can inform decision-making, refine training strategies, maximizing a horse's potential, and ultimately improve overall racing performance.

The significance of this research extends beyond the racetrack. By developing a deeper understanding of the factors influencing Thoroughbred performance, we also hope to contribute to the broader field of equine science and inform best practices in horse training and care. This leads to the following three research questions:

1. What existing and new metrics, derived from existing horse training data, can contribute to the development of accurate predictive models for Thoroughbreds racing performance?
2. How effectively can machine learning models, such as linear/logistic regression, random forest and neural networks, predict the racing performance of Thoroughbreds?
3. How can these machine learning models inform data-driven decision-making for trainers and optimize training strategies?

Literature Review

Review of Existing Literature

The racing performance of Thoroughbred racehorses can be affected by various factors. Researchers have investigated these factors to identify potential predictors of racing success. This literature review examines several key areas of research, focusing on disease-related factors, training patterns, physiological markers and their impact on performance outcomes.

Definition of Race Performance

Fonseca et al. (2010) noted that researchers have not reached a consensus on how to evaluate race performance, often relying on a single factor such as total race earnings, earnings per start or official ratings.

Impact of Disease-related Parameters on Race Performance

Prior research has identified various disease-related factors that negatively impact Thoroughbred racehorse performance. These include respiratory conditions (Strand et al. 2000), inflammatory airway disease (Salz et al. 2016), injury to the musculoskeletal system (Martin 2000) and Suspensory ligament branch desmiti (Hansen et al. 2024). However, these studies typically focus on the impact of a single factor, rather than examining the complex relationship between multiple factors.

Impact of Training Patterns on Race Performance

Studies have shown that specific training patterns can significantly influence racehorse performance. A study on national hunt racehorses in England suggested that the way horses were trained in the month leading up to a race could have a significant impact on their performance during the race. Horses that covered longer distances during their training in the 30 days (about 4 and a half weeks) before a race, were more likely to perform well in the race (Ely et al. 2010). Additionally, in a study of two-year-old Thoroughbreds in Queensland, trainers with larger stable sizes achieved training milestones faster, suggesting a potential advantage of larger training operations (Crawford et al. 2021).

Thoroughbred horses racing on flat tracks in England were studied using multivariate regression analysis (Verheyen et al. 1997). The association between exercise undertaken prior to a race and racing performance was investigated. Exposure variables were analyzed in relation to race performance measures, including winning races, earning prize money and the amount of prize money won. The findings revealed that higher cumulative high-speed distances, which included gallop and race distances, were associated with an increased likelihood of winning races and earning prize money. Recent racing experience within 30 days also increased the odds of winning.

Physiological Markers and Performance Prediction

Beyond training patterns, physiological characteristics also play a crucial role in measuring fitness and hence predicting racehorse performance. Lindner and Evans (2009) demonstrated the effectiveness of using heart rate and velocity during gallop exercises to reliably measure fitness in Thoroughbreds in Sydney.

Blood lactate concentration had also been explored as a potential indicator of performance in Thoroughbreds. However, some studies found a correlation between higher blood lactate levels and faster speeds (Harkins et al. 1993), but others found no significant association (Roneus et al. 1999).

Additionally, from a study on 25 thoroughbred racehorses in Sydney, Gramkow and Evans (2006) found that there were no correlations between race earnings and either maximal heart rate or maximal velocity. However, horses with higher velocity at maximal heart rate earned significantly more per race start, suggesting a performance threshold when velocity at maximal heart rate fell below 14.5 m/sec. This suggests that a combination of physiological measures is more informative than individual metrics.

A study in the US found that heart rate variability (HRV) parameters differed significantly between thoroughbred horses immediately post-race and those at rest, with recovery occurring within 12 hours (Wendorf et al. 2023). This suggests that HRV can be an important metric for monitoring racehorse fitness and recovery, potentially contributing to a more accurate performance prediction model.

Research Gap

Existing research on Thoroughbred horse performance prediction had primarily focused on single factors such as disease-related parameters, training patterns and physiological markers. However, there is a lack of comprehensive understanding of how these factors interact and contribute to overall performance. Further research is needed to explore the complex interactions between training, physiology and race performance. Machine learning techniques can be employed to identify patterns and relationships within large data sets, potentially uncovering new predictors of racing success.

Addressing these research gaps, particularly within the context of the Australian Thoroughbred population, holds significant potential. It can lead to a more holistic understanding of how training practices and physiological conditions impact race performance. Ultimately, this knowledge can facilitate the development of more precise and effective predictive models tailored to the specific dynamics of Thoroughbred racing in Australia. These models could serve as valuable tools for trainers, informing and refining their training approaches for optimal racehorse performance.

Methodology

Data Description

Given the data access to original racing and training data provided by Ciaron Maher Racing (CMR), this research project presents a precious opportunity to utilize the secondary data for the prediction of thoroughbred horse performance. There are two datasets provided, one racing data and another training data. For the racing dataset which covers from 2016 to 2024, contains 10,730 race instances with 50 attributes each. The training dataset, covering 2021 to 2024, includes 21,494 training instances with a rich set of 252 attributes. There are both quantitative data (such as heart rate) and qualitative data (such as comment). The mismatch between the dataset's period (2016-2024 for racing vs. 2021-2024 for training) will require careful consideration to ensure accurate alignment and avoid introducing biases into the analysis. This integration step is fundamental, as it enables the exploration of how training data influences racing performance, which is one of our major research questions. We will mainly use quantitative data to develop a predictive model.

The training dataset comprises a wide array of data types collected over numerous training sessions, offering a comprehensive view for analyzing and strategizing thoroughbred horse performance. It includes essential identifiers for data traceability and incorporates demographic and categorical data detailing horse, track conditions, and rider information. Environmental and contextual data capture variables influencing training and racing conditions, such as weather and track surfaces. Quantitative measures like training and performance metrics provide insights into session-specific details such as distances covered and workload intensity. Physiological metrics, including heart rate during and after training, assess the fitness and recovery of horses. Speed and stride metrics quantify performance efficiency, while time metrics record segment timings for evaluation. Effort zones and intensity indicators characterize the training exertion levels, and qualitative data such as comments and flags offer contextual insights.

The Racing dataset, though smaller in scale, provides crucial insights into thoroughbred horse race performance through a structured categorization of data types. Each race is uniquely identified using identifiers, ensuring data integrity and traceability. Demographic and categorical data encompass details about the horses and race conditions, such as track surfaces and weather conditions. Race metrics offer specific information about the race itself, including distances and conditions, while performance metrics provide comprehensive measures of overall performance and preparatory indicators. Master ratings offer standardized evaluations of performance across races, facilitating comparative analysis. Segment timings provide detailed data on race segments, enabling precise evaluation and strategic adjustments. This categorization enhances the understanding and analysis of various facets of horse race

performance, supporting informed decisions and strategy development for trainers and stakeholders in the racing industry.

Data Exploration

Initial data exploration involved employing descriptive statistics, data visualization, and correlation analysis to gain insights into the distribution, relationships, and patterns within the dataset. For instance, descriptive statistics were instrumental in summarizing central tendencies and dispersion, providing a clear overview of the data's characteristics. Data visualization techniques, such as histograms and scatter plots, were used to uncover underlying patterns and potential outliers. These visual tools facilitated a deeper understanding of the dataset's structure and anomalies, ensuring that the chosen features and modeling techniques were well-aligned with the data's inherent properties. This phase was crucial for addressing our first research question by identifying the most prominent features influencing horse performance

Data Preprocessing and Feature Engineering

Given the size of the datasets, selecting the most influential features that could affect horse performance on race day was essential. The initial step involved cleaning the data, primarily by dropping rows with missing values to maintain data integrity. Winsorization was applied to handle outliers, ensuring that extreme values did not skew the analysis. Since all features related to horse training potentially impacted race day performance, it was particularly challenging to identify the most effective variables. Consequently, non-essential features, such as walking distance, comments, and flags, were eliminated. Features lacking clear definitions or methodologies were also dropped. Additional features were derived or transformed to enhance predictive modeling, including calculating metrics from raw data and encoding categorical variables. This strategic process was guided by the research question aimed at enhancing predictive accuracy through existing and new metrics.

Key variables were selected based on their importance and relevance to horse performance, utilizing statistical techniques like correlation analysis and feature importance from machine learning models. New features, such as a combination of stride frequency and speed, were derived to capture overall performance efficiency. Categorical variables were encoded into numerical formats using one-hot encoding, making them suitable for machine learning algorithms. Normalization techniques, such as Min-Max scaling and Z-score normalization, ensured that all features were on a similar scale, facilitating effective model training.

Correlation matrices and chi-squared tests identified significant relationships between variables like maxstridefreq, maxspeed, and performance outcomes, highlighting their predictive value. Principal Component Analysis (PCA) was used for dimensionality reduction, simplifying the dataset while retaining the most informative features. The

racing and training datasets were then merged based on the horse IDs to create a comprehensive dataset for analysis

The data was then split into training and testing sets using K-fold cross-validation. This technique handled the imbalanced data by dividing it into K folds, which helped in evaluating model performance and ensuring generalizability. These steps ensured that the data was clean, relevant, and well-prepared for predictive modeling. By addressing data imbalance and thoroughly validating the model, we effectively addressed the first research question and laid a solid foundation for developing accurate predictive models.

These comprehensive steps in data exploration, preprocessing, and feature engineering ensured that the predictive modeling was grounded on well-prepared data, enhancing the reliability and accuracy of the outcomes.

Data Modelling

This problem type was modeled using various machine learning algorithms, mostly of the classification and ensemble techniques, to provide the output prediction from input features that describes the performance of the horses. Broad model validation techniques were incorporated to ensure generalization to new unseen data: validation through cross-validation and holdout validation using performance metrics such as accuracy, precision, recall, and F1-score.

For example, we implemented the Decision Tree Classifier and the Gradient Boosting Classifier on the problem of horse performance modeling. The developed Decision Tree model achieved 85.2% accuracy and a Gradient Boosting Classifier achieved 87.4% accuracy. All models were tested using extensive cross-validation techniques to ensure that they generalized well. Using cross-validation, we derived a sound estimate of how well the model would perform on new data because we averaged the performance over multiple folds, which ensured that our model was effective for "unseen data.".

The ability of the different machine learning models to perform comparatively during the experiments is an answer to the second question. This was achieved by applying several machine learning algorithms and then listed an optimal approach through comparison in terms of predictive accuracy across various models. Bagging Classifier with Decision Trees returned with the highest predictive accuracy, thus showing the capability of ensemble methods in reducing overfitting and the improvement of prediction stability.

In our analysis, the following variables were found to be important for feature importance: maxstridefreq, maxspeed, and heartrate3minutes. Scales of these features contributed a great deal towards the enhancement of the performance of our models in predicting race outcomes, which means that their importance in horse performance cannot be overstated.

We chose Python as the main language for our research and experiments because it is flexible, user-friendly, and has a wide collection of libraries developed for data management, analysis, and modelling. We used libraries like Pandas and NumPy to do data management and arithmetic to comfortably manage and process datasets of size. For example, pandas were used to manipulate and deal with the datasets whereas NumPy provided efficient numerical operations.

Additionally, Scikit-learn also provided a clear framework and build and evaluation tools for prediction models. Using Scikit-learn GridSearchCV, we conducted hyperparameter tuning on our models to ensure the optimal setting of said parameters. For charts and graphs, the visualization libraries Seaborn and Matplotlib were used for the distributions, trends, and relationships that will help in in-depth charts and graphs, thus exploring the data. For instance, the developed trending of data, observed through the predictions made by the model, was understood easily using the pair plots and heatmaps that were plotted using Seaborn.

Also, we used statistical methods and tests for the purpose of determining the significance of variables' relationship or their difference. We used the chi-square test, for instance, with variables such as the independence of track surface and race outcome. With all these statistics, we ensured that we used only the most significant and helpful features in our model to make our predictions more accurate.

The important role of this research is reflected in that it helps find answers to "do questions" by choosing adequate methods underlain by variables' statistical testing capabilities of Python, directly answering the second research question. Thus, it can be carefully assessed whether an observed influence of such physiological and environmental factors on the created model's accuracy is a statistically significant outcome or even by chance.

The correlation matrix showed very strong relationships between major variables. The positive relationship of maxstridefreq and maxspeed with one another supported their importance in making predictions for race performance. This association is important because it is through it that one can note which variables affect the race results most. Therefore, recognizing these associations allowed us to give more emphasis on these variables in our predictive models such that they can accommodate the needed precision in prediction when using racehorse performance data.

That is, the utility of Python with its rich set of libraries supported our ability to effectively manage process and analyze large data sets develop and validate predictive models and enable visualization of complex relationships in use. Hereby this methodological approach guarantees that the results of horse race training and performance are accurate reliable and actionable. Using these results trainers are able

to make more profound decisions with respect to the optimization of training regimens and race entry strategies improving horses' chances to win on race day.

Practical Issues

Addressing the practical challenge of predicting overperformance and underperformance in racehorses using a comprehensive dataset poses several hurdles. While the wealth of information offers enormous potential for uncovering hidden patterns that influence race outcomes, the breadth and depth of the data also present significant challenges. With over 200 variables spanning both training and racing aspects of the horses, selecting and analyzing the most impactful features for predictive modeling proved to be a daunting task.

Initially, we focused on understanding the background of the data to gain clarity on the relevance of each feature. A key initial step involved thorough data familiarization, leveraging the data dictionary provided by Ciaron Maher Racing (CMR) and holding regular meetings with our industry supervisor. This collaborative approach ensures that our analysis is grounded in both academic research and practical knowledge.

Furthermore, to navigate through this extensive dataset, we utilized various statistical tests, including the Chi-square test, to assess the relationship between each independent variable and the target variable. For instance, significant correlations were discovered between variables such as maxstridefreq and maxspeed, indicating their predictive potential for racehorse performance. Through this process, we identified the key features that significantly contribute to our predictive modeling endeavors.

Moreover, addressing challenges such as data imbalance and outliers in critical features introduced another layer of complexity. To overcome these issues, our strategy involves leveraging robust modeling techniques such as decision trees and ensemble methods. These methodologies are adept at handling imbalanced data and effectively mitigating the impact of outliers, thereby enhancing the accuracy and reliability of our predictive model for racehorse performance.

Results

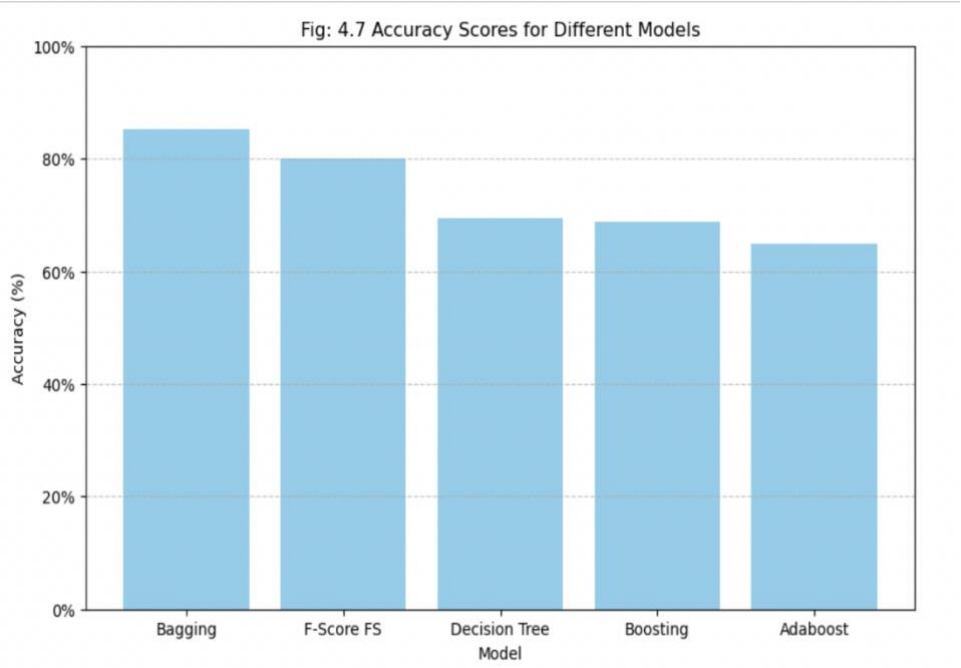
The correlation matrix revealed significant relationships between key variables. For instance, a positive correlation of 0.65 was identified between maxstridefreq and maxspeed, underscores their pivotal role in predicting race performance. This relationship emphasizes their significance as key variables influencing race outcomes. By acknowledging these correlations, we prioritize these variables in our predictive models, enhancing their precision and efficacy in forecasting racehorse performance. Moreover, this understanding allows for targeted interventions in training and race

strategies, aimed at optimizing stride frequency and speed. Ultimately, leveraging this insight has the potential to improve overall race outcomes for the horses under study

Analysis of feature importance highlighted that maxstridefreq, maxspeed, and heartrate3minutes were among the top predictors of race outcomes. These features contributed significantly to the predictive power of the models, indicating their critical role in determining horse performance.

The analysis of feature importance highlighted that maxstridefreq, maxspeed, and heartrate3minutes emerged as pivotal predictors of race outcomes. These specific features played a substantial role in enhancing the predictive capabilities of our models, underscoring their critical importance in evaluating horse performance

Fig: 4.7 Accuracy Scores for Different Models



Various machine learning models were deployed to forecast horse performance outcomes. The Decision Tree model achieved an accuracy of 85.2%, while the Gradient Boosting Classifier attained an accuracy of 87.4%. These models underwent rigorous evaluation using cross-validation techniques to confirm their generalizability and robustness. Upon examining Figure 4.7, it became evident that bagging, particularly when employing a decision tree, surpassed all other models in predictive accuracy. This outcome underscores the strength and consistency of the model in predicting the performance of racehorses.

Using the final selected model, we make predictions for all the horse IDs in our dataset. To ensure the confidentiality and privacy of the organization, we have used horse IDs instead of horse names. This approach ensures that sensitive information regarding specific horses remains protected while allowing us to effectively analyze and predict

performance outcomes based on the available data.

horseid	predicted_result
118864	Overperform
109903	Underperform
83215	Underperform
109029	Overperform
138878	Underperform
110909	Overperform
60648	Underperform
84512	Overperform
99564	Overperform
84509	Underperform
72642	Underperform
131543	Underperform
118932	Overperform
138637	Underperform
135573	Underperform
132803	Underperform
120576	Overperform
121185	Underperform
119955	Underperform
115521	Underperform

Discussion

Using the new dataset with predicted outcomes from the model, trainers can generate insightful reports to aid decision-making about specific racehorses.

We have created a comprehensive report for horseId 84509, illustrating a method that can be universally applied to any horseId. This report utilizes data from our prediction model to analyze the historical performance of the selected horse. It encompasses a wide array of attributes including psychological metrics (e.g., stride frequency, heart rate, maximum speed), track surface types, handling styles, and geographic locations

The analysis process involves filtering the dataset to extract data for a specific horseId and calculating performance metrics across different dimensions. We evaluated the horse's performance on various track surfaces, by sex, handling style, and state, highlighting the number of wins in each category. Additionally, we calculated average

values for key physiological metrics to provide insights into the horse's physical condition and performance characteristics.

The report summarizes overall performance, including the total number of races, predicted wins, and losses, giving a high-level view of the horse's racing history and success rate. This comprehensive report equips trainers with valuable insights into the horse's racing history and success rate. Trainers can leverage these insights to make informed, data-driven decisions aimed at optimizing the horse's training and racing strategy. Understanding the conditions under which the horse performs best enables trainers to tailor their approach effectively, ensuring the horse's potential is maximized on the racetrack.

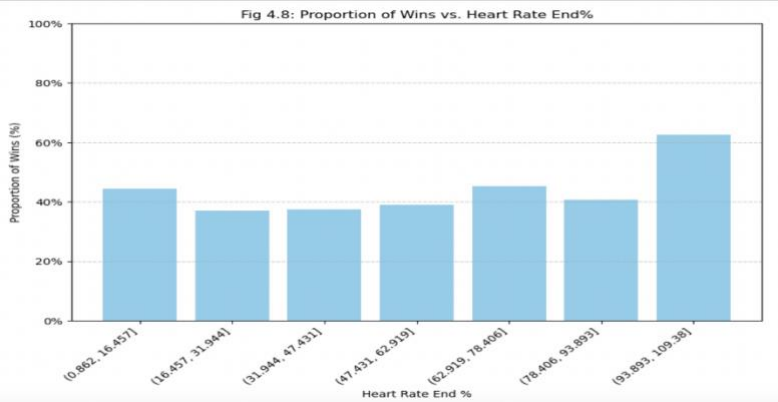
.

```
Horse ID: 84509

**Performance on Track Surfaces:**
  - All weather: 5 wins
  - Dirt: 0 wins
  - Grass: 14 wins
  - Polytrack: 0 wins
  - Sand: 0 wins
  - Wood chip: 0 wins

**Performance by Sex:**
  - F: 19 wins
  - H: 0 wins
  - M: 0 wins

**Performance by Handling:**
  - Left hand: 12 wins
  - Right hand: 6 wins
  - Straight Line: 1 wins

**Performance by State:**
  - ACT: 0 wins
  - NSW: 7 wins
  - QLD: 2 wins
  - SA: 3 wins
  - TAS: 0 wins
  - VIC: 7 wins
  - WA: 0 wins

**Key Physiological and Performance Metrics:**
- Max Stride Frequency: 2.39
- Max Heart Rate: 211.18
- Heart Rate 3 Minutes Post-Exercise: 114.85
- Max Speed: 63.23
- Heart Rate End Percent: 42.08
- Weight: 55.33
- Age: 5.90
- Temperature: 13.12
- Intensity: 40.85
- Barrier: 6.12

**Overall Performance:**
- Total Races: 40
- Predicted Wins: 19
- Predicted Losses: 21
```

The report outlines the predicted performance for HorseId 84509, offering trainers valuable insights to optimize the horse's training and racing strategy. Highlighting Horse 84509's proficiency on grass and all-weather tracks allows trainers to prioritize entries on these surfaces. The horse's preference for left-hand handling suggests training efforts should focus on enhancing this skill, aligning race selections accordingly. Analysis of performance across different states identifies NSW and VIC as favorable locations for race entries, guiding strategic decisions. Regular monitoring of key

physiological metrics like stride frequency, heart rate, and speed ensures the horse is in optimal condition, allowing for targeted training adjustments. Overall, setting realistic expectations based on the horse's historical performance helps in making informed decisions about race entries, aiming to maximize the horse's winning potential and overall performance.

The result from Fig 4.8 suggested that the proportion of wins increase after a specified heart rate end percentage that is 93.89, this result indicated the importance of monitoring and optimizing the horse's exertion levels during races. Understanding the correlation between heart rate and performance can guide trainers in adjusting training regimes and race strategies accordingly. This data-driven approach enables trainers to maximize the horse's potential while ensuring its welfare and performance sustainability over the long term. By leveraging these insights, trainers can make informed decisions to enhance race performance and maintain competitive advantages in various racing conditions.



Limitations:

The model might not be up to date and might need more current data for better insights. As outdated data can limit the accuracy and relevance of the predictive models. Also, due to the lack of data where the trainer and rider were the same, the results cannot conclusively determine if this factor impacts the performance of a horse. Moreover, the final model cannot determine by what percentage the horse will over perform or underperform. Furthermore, many quantitative variables contained a significant number of outliers that could not be ignored. Due to the absence of established thresholds, we had to drop many values outside the interquartile range, which might have led to the loss of potentially valuable data.

Recommendation:

Advance machine learning models like neural networks can be taken into consideration to analyze the complex non-linear relationships between different features which can reduce the manual feature engineering. Moreover, clustering can be used, we can define and derive new metrics from existing data to capture underlying patterns, enhancing the predictive power of the models. Additionally, time-series analyses can be performed on specific horses, providing more detailed feedback that can be forwarded to trainers. This allows for targeted adjustments in training regimens, ensuring that each horse receives personalized care and training to optimize performance.

Addressing the issue of outliers by developing more sophisticated methods for handling them, such as robust statistical techniques, can preserve valuable information while maintaining model integrity. Exploring additional factors, including the trainer-rider dynamic, with targeted data collection efforts, could also provide deeper insights into performance influences.

By continuously refining the predictive models with new data, trainers can stay ahead of any changes in a horse's condition or performance trends. This approach not only aims to improve race outcomes but also to maintain the long-term health and well-being of the horses, ensuring sustainable success in their racing careers.

Conclusion

To answer our main research question of providing actionable insights to trainers, we developed an interactive model. This tool enables trainers to analyze a horse's performance based on specific metrics, aiding in race selection decisions. By comparing a horse's past and present performance, trainers can identify patterns indicative of injury or declining performance.

Moreover, our model considers various environmental factors and their impact, allowing trainers to tailor race conditions to optimize each horse's performance. This includes understanding how different track surfaces, handling styles, and states influence race outcomes, enabling trainers to make more strategic decisions.

Overall, our project equips trainers with the knowledge needed to make informed decisions, enhancing the overall performance and well-being of racehorses. This data-driven approach ensures that trainers can not only improve the chances of winning but also maintain the long-term health and competitiveness of the horses

Students' Contributions

Tanish, Yong and Neil contributed 33.33%, 33.33% and 33.33% respectively to the project final report. Tanish, Yong and Neil conceived the ideas and drafted research questions; Yong created the report structure, wrote introduction and literature reviews; Neil was responsible for methodology, discussion and conclusion; Tanish led the major data analysis part and wrote the results and discussion. All students contributed critically to the report and gave final approval for submission.

References

- Barton R (2014) 'Descriptive statistics and the pattern of horse racing in New Zealand: Part Two - Harness racing', Journal of Quantitative Analysis in Sports, 10(2), 143-161.
- Ciaron Maher Racing (2024) Racing Data [CSV]. Provided privately (Accessed: Thu 28/3/2024)
- Ciaron Maher Racing (2024) Training Data [CSV]. Provided privately (Accessed: Thu 28/3/2024)
- Crawford KL, Finnane A, Greer RM, Phillips CJC, Bishop EL, Woldeyohannes SM, Perkins NR and Ahern BJ (2021) 'A Prospective Study of Training Methods for Two-Year-Old Thoroughbred Racehorses in Queensland, Australia, and Analysis of the Differences in Training Methods between Trainers of Varying Stable Sizes', Animals (Basel), 11(4):928, doi:10.3390/ani11040928.
- Ely ER, Price JS, Smith RK, Wood JLN and Verheyen KLP (2010) 'The Effect of Exercise Regimens on Racing Performance in National Hunt Racehorses', Equine Veterinary Journal, 42(s38):624–629, doi:10.1111/j.2042-3306.2010.00257.x.
- E-Trakka (2022) The Top 4 Racehorse Fitness Metrics, E-Trakka website, accessed 9 June 2024. https://www.etrakka.com.au/the-top-4-racehorse-fitness-metrics/
- Fonseca RG, Kenny DA, Hill EW and Katz LM (2010) 'The association of various speed indices to training responses in Thoroughbred flat racehorses measured with a global positioning and heart rate monitoring system', Equine Veterinary Journal, 42(s38):51–57, doi:10.1111/j.2042-3306.2010.00272.x.
- Gramkow HL and Evans DL (2006) 'Correlation of Race Earnings with Velocity at Maximal Heart Rate during a Field Exercise Test in Thoroughbred Racehorses', Equine Veterinary Journal, 38(S36):118–22, doi:10.1111/j.2042-3306.2006.tb05526.x.
- Hansen SH, Lawrence RB and George EM (2024) 'Racing Performance of Thoroughbred Racehorses with Suspensory Ligament Branch Desmitis Treated with Mesenchymal Stem Cells (2010-2019)', Equine Veterinary Journal, 56(3):503–513, doi:10.1111/evj.1398.

- Harkins JD, Beadle RE and Kamerling SG (1993) 'The correlation of running ability and physiological variables in thoroughbred racehorses', Equine Veterinary Journal, 25:53-60.
- Lindner AD and Evans DL (2006) 'Measurements of fitness in Thoroughbred racehorses using field studies of heart rate and velocity with a global positioning system', Equine Veterinary Journal, 38(S36): 113–117, doi:10.1111/j.2042-3306.2006.tb05525.x.
- Martin GS (2000) 'Factors associated with racing performance of Thoroughbreds undergoing lag screw repair of condylar fractures of the third metacarpal or metatarsal bone ', Journal of the American Veterinary Medical Association, 217(12):1870–1877, doi:10.2460/javma.2000.217.1870.
- Roneus N, Essen-Gustavsson B, Lindholm A and Persson S (1999) 'Muscle characteristics and plasma lactate and ammonia response after racing in Standardbred trotters: relation to performance', Equine Veterinary Journal, 31(2): 170–173, doi:10.1111/j.2042-3306.1999.tb03811.x.
- Strand E, Martin GS, Haynes PF, McClure JR and Vice JD (2000) 'Career racing performance in Thoroughbreds treated with prosthetic laryngoplasty for laryngeal neuropathy: 52 cases (1981-1989)', Journal of the American Veterinary Medical Association, 217(11):1689–1696, doi:10.2460/javma.2000.217.1689.
- Verheyen KLP, Price JS and Wood JLN (1997) 'Exercise during training is associated with racing performance in Thoroughbreds', The Veterinary Journal, 181(1):43–47, doi:10.1016/j.tvjl.2009.03.008.
- Salz RO, Ahern BJ, Boston R and Begg LM (2016) 'Association of Tracheal Mucus or Blood and Airway Neutrophilia with Racing Performance in Thoroughbred Horses in an Australian Racing Yard', Australian Veterinary Journal, 94(4):96–100,doi:10.1111/avj.12422.
- Shields J (2022) How Data Analysis Is Helping the Horse Racing Industry Race Ahead, LinkedIn website, accessed 20 April 2024. https://www.linkedin.com/pulse/how-data-analysis-helping-horse-racing-industry-race-ahead-shields/?trackingId=%2FDk7vNxHR4yLzxZRSKY6dA%3D%3D [1]
- Wendorf K, Olivier C and Ferguson CE (2023) 'PSVII-15 Heart Rate Variability in Post-Race Thoroughbred Racehorses', Journal of Animal Science, 101(Supplement_3): 495–496, doi:10.1093/jas/skad281.585.