Optimizing Horse Racing Predictions through Ensemble Learning and Automated Betting Systems

Nguyen Quoc Khanh
khanhnq@ventusgroup.ai

Nguyen Phuoc Sang
sangnp@ventusgroup.ai

Tran Thi My Quyen
quyenttm@thinkprompt.com

Vu Anh Tran
andy@ventusgroup.ai

Abstract

Horse racing's burgeoning popularity has spurred extensive research into predictive analysis and strategic decision-making within the sport. The digital age has ushered in an era of unprecedented data complexity and volume related to horse racing, demanding sophisticated analytical approaches. Machine learning, with its broad reach across various sectors, is making significant inroads into the sports betting industry, particularly in horse racing, a domain characterized by substantial financial stakes and widespread interest. Predicting horse racing outcomes is inherently challenging due to the multitude of factors at play. This study endeavors to advance the field by leveraging machine learning algorithms to forecast horse racing results using a comprehensive dataset encompassing race data from Australia (2015-2024) sourced from Punters.com.au, Australia's premier horse racing platform, and data from the United Kingdom (including Great Britain and Ireland) spanning 2021 to 2023. A stacked ensemble model incorporating six algorithms – LightGBM, XGBoost, CatBoostRegressor, HistGradientBoostingRegressor, AdaBoostRegressor, and TabNetRegressor – was employed to predict the victor of each race. This research can be valuable to stakeholders and researchers alike, offering insights for further analysis and experimentation in this field.

Keywords: machine learning, prediction, classification model, sports betting, horse racing.

1 INTRODUCTION

Enjoyed by people from all backgrounds, contemporary horse racing is an open-air sport. It represents a multifaceted industry encompassing sports, entertainment, wagering, and commercial aspects. With 3.2 million races held annually, the global prize pool for horse racing reaches an impressive 360 billion yuan. [1]. Currently, over 70 nations and territories worldwide are involved in horse racing. Horse racing thrives in countries like the United States, Britain, France, Australia, Japan, Hong Kong, and others, where it has become a significant industry [2]. Horse racing serves as a catalyst for economic activities, a cornerstone for public welfare and charitable endeavors, and a substantial source of tax revenue for nations. It fosters social interaction and provides a significant form of entertainment. [3]. Participating in horse racing events, joining horse clubs, and owning horses have become fashionable pursuits, often symbolizing an individual's identity and social standing.

As living standards have significantly improved, horse-racing-related activities have experienced rapid and consistent growth. The horse racing lottery, a component of the horse-racing industry, has garnered both public interest and national backing. Fueled by the horse racing lottery, there's a surge in the number of everyday individuals engaging in horse racing. The horse racing industry in the UK exemplifies this trend, generating over a billion pounds in revenue. This positions it as the second largest sports industry, trailing only football. The horse racing industry's growing popularity is evident in the estimated 130 million individuals globally who participate in horse racing lottery predictions. With the increasing prevalence of horse racing, enthusiasts have begun identifying patterns and trends across various race formats. Their goal is to enhance the accuracy of their win-loss predictions [**?**]. They aim to discover methods for improving the accuracy of their predictions. However, in today's era of rapid technological advancement and data analysis, conventional horse racing prediction models are proving to be inadequate. They often suffer from low prediction accuracy and have structural flaws in their system design. These limitations render them incapable of meeting the demands of modern horse racing management and competition outcome forecasting. Consequently, there's a pressing need to develop a robust horse racing decision-making and management model. This model should leverage scientific methodologies to address forecasting challenges effectively. This challenge has become a focal point for both researchers and the general public. This issue has captured the attention of both academics and the public.

The evolution of horse racing has been significantly influenced by information technology. This progress, driven by scientific and technological advancements, has not only elevated horse racing but also raised the bar for decision-making, management, and performance forecasting in the field [**?**]. This paper leverages extensive data analysis to introduce a horse racing system based on machine learning ensemble algorithms. This system aims to create a robust decision-making and management optimization model for horse racing competitions. It empowers decision-makers with advanced tools and resources, enabling well-informed choices. Furthermore, it offers a concrete quantitative foundation for devising effective training management strategies and training schedules. Lastly, it delivers scientifically-grounded predictions to guide public engagement in horse racing. This approach facilitates well-informed training management decisions and the development of tailored training regimens. It also equips the public with scientifically sound predictions to enhance their participation in horse racing. These insights provide a data-driven approach to understanding and engaging with the sport.

2 RELATED WORK

Horse racing, a test of human control and equestrian mastery, centers around speed as its core competitive element. A cornerstone of equestrian sports, it encompasses various formats while consistently emphasizing the pursuit of speed. Despite variations in format, the fundamental principle of horse racing remains consistent: a contest of speed. Beyond its sporting nature, horse racing holds a global presence as a traditional competitive event, as documented in source [**?**]. Historical evidence, including textual analysis, reveals the existence of horse riders in Neolithic Age stone carvings. Notably, the ancient Olympic Games in Greece featured four-horse chariot racing as early as the seventh century BC. This event, initially a chariot race, later transitioned to a competition of human driving, as detailed in source [7]. Originally, horse racing served as a selection process for identifying superior horses. Only the most exceptional horses, those demonstrating outstanding performance on the racetrack, were deemed suitable for breeding purposes. In contemporary times, the evolution and widespread popularity of horse racing have led to significant advancements. Organization, management, and competition have become increasingly sophisticated compared to historical horse racing events. Modern horse racing, as highlighted in source [8], embraces a more scientific and advanced approach, reflecting its ongoing development. The modern era has witnessed remarkable advancements in horse racing, particularly in its organization, management, and competitive aspects. Compared to their historical counterparts, contemporary horse racing events exhibit a heightened emphasis on organization, management, and competition. In contrast to ancient horse racing events, modern competitions have evolved significantly, incorporating more sophisticated organizational structures, management practices, and competitive formats. This evolution is characterized by the adoption of more scientific and advanced methods, as noted in source [8]. Source [8] emphasizes the significant advancements in modern horse racing, particularly the implementation of more scientific and advanced methodologies.

The enduring fascination with horse racing has spurred extensive academic inquiry. Ghe zelsefloo, for instance, advocated for applying service operation management principles to optimize horse racing event management [9]. This approach suggests integrating the established theoretical frameworks of service operations management into the practical domain of horse racing event management. Fenner et al. further emphasized the value of model-driven approaches, using them to comparatively analyze service operation management models in Hong Kong and Japanese horse racing events [9]. This analysis focused on the service aspects of these events, highlighting the importance of customer experience. From a service perspective, they examined the operational models of horse racing events in Hong Kong and Japan. Zhang and Liu highlighted the unique symbiosis between horse and rider in equestrian events, stressing that equine health is paramount and necessitates meticulous event organization and management [11]. They argued that the well-being of horses directly impacts the event's success. Quintana et al.'s research revealed a correlation between high-level equestrian competition and elevated injury rates, underscoring the need for robust emergency medical services at such events [12]. Their findings indicated that high-level equestrian competitions have the highest injury rates among participants. They emphasized the critical need for competent emergency medical services in equestrian competitions. Fenner et al. stressed the importance of long-term strategic planning for equestrian events, advocating for the cultivation of unique brand identities and the development of distinctive event features [13]. They recommended establishing a system for comprehensive long-term event planning. They also proposed the creation of a specialized talent development pipeline to support the equestrian sector. Sun and Li's research demonstrated the economic ripple effect of equestrian sports, noting that competitions stimulate related consumption [14]. Padalino et al. delved into the evolution of large-scale sports events in Hong Kong, examining the mechanisms governing sports development and management [15]. Their study analyzed the development and management frameworks within the sports industry. They specifically focused on how horse racing has fueled the growth of Hong Kong's gaming industry and sports culture. This research highlighted the role of horse racing in driving the expansion of Hong Kong's gaming sector and sports culture. Hoseini and Amani introduced a B/S model-based equestrian event management system, aiming to streamline event operations through digitalization [16].

Information technology serves as both the underpinning and the hallmark of contemporary society. As artificial intelligence and big data advance, numerous experts leverage information technology for the management and forecasting of sporting competitions. In the context of research concerning the comprehensive assessment and prediction of national gold medal standings in the Olympic Games, Xiao et al. employed a multi-faceted approach. This approach encompassed tracking statistics, comprehensive evaluation, comparative analysis, and expert consultation. Their objective was to comprehensively evaluate Olympic gold medal rankings. This evaluation focused on Olympic events within major global competitions. The timeframe for their analysis spanned the Olympic cycle from 1989 to 1992. Their prediction methodology relied on a comprehensive and up-to-date understanding of the strength levels and developmental trajectories of Olympic events across various nations and regions. This approach enabled them to predict strength patterns. Furthermore, they forecasted the number of gold medals and the overall team rankings for the Olympic Games. These predictions were detailed in their publication [17]. In a separate study, Chen examined the strength of diving powerhouses during the 2000 Sydney Olympics. This analysis also included gold medal predictions. Chen employed the sports expert method to assess the relative strengths of the diving powers participating in the 2000 Sydney Olympics. Based on this analysis, predictions were made regarding the gold medal outcomes for the Chinese diving team. These findings were presented in [18]. 2000 Sydney Olympics and predicted the gold medal situation of the Chinese diving team [18].

Predictive management research and applications are currently being explored across numerous research domains. Notably, the field of horse racing research is no exception to this trend. Within horse racing research, predictions are relevant to several areas. These areas include strategizing, decision-making, coaching, and the formation of athlete teams and training regimens. In recent years, various prediction algorithms have been widely applied to nonlinear prediction problems across diverse fields. Some examples of these algorithms include BP neural networks, wavelet neural networks, and support vector machine regression. In the context of horse racing, association rules and neural networks have emerged as the most effective prediction methods. This success stems from their ability to easily implement complex nonlinear mapping functions. Utilizing scientific forecasting methods enables the identification of the direction, trend, and laws governing the development and change of things. This understanding allows for the implementation of effective measures to control co-development. Ultimately, this leads to enhanced optimization of horse racing management decisions. This optimization is achieved through forecasting using scientific methods. Through these methods, we can understand the direction, trends, and laws driving the development and changes in horse racing. This understanding allows us to implement effective measures to manage co-development. This, in turn, leads to better optimization of horse racing management decisions. This optimization is crucial for improving the management of horse racing decisions. decisions.

3 DATASET AND PROCESSING

3.1 Data Source

This research paper presents experimental findings derived from two primary geographic regions: Australia (AUS) and the United Kingdom (UK), encompassing Great Britain and Ireland (GB).

Primary Data Sources and Concise Descriptions: The following sources were selected based on their relevance and reliability in the field of horse racing data: (1) punters.au.com: A comprehensive Australian repository offering historical race results, horse profiles, and form guides; (2) attheraces.com: A UK-based platform providing in-depth coverage of horse racing events, including live streaming, race cards, and detailed race results; (3) racingbetdata.com: A specialized provider focused on betting-related data, historical odds, and market trend analysis.

The selection of these data sources was informed by several key factors: (1) Reliability: Each platform is recognized for its accuracy and trustworthiness within the horse racing industry; (2) Data Diversity: The inclusion of these platforms ensures access to a broad spectrum of data points, enabling a more comprehensive analysis; (3) Timeliness: The sources provide real-time or near-real-time updates, ensuring that the research is based on the most current information; (4) Geographical Scope: By incorporating data from both Australian and UK sources, the research benefits from a wider geographical perspective, enhancing the depth of the analysis.

3.2 Data Selection and Preprocessing

Criteria for selecting data: The data selection process was guided by the following criteria: (1) Time period: The most recent five years (2019–2024) of racing data were considered to ensure both relevance and a sufficient sample size; (2) Race categories: Only flat races were included in the analysis, while jump races were excluded to maintain consistency in the dataset; (3) Geographical locations: The study focused on major racecourses in Australia and the UK, targeting high-profile events for detailed analysis.

Steps for preprocessing data: The preprocessing phase involved the following steps: (1) Eliminating duplicate or incomplete records: Redundant data points and races missing essential information were identified and removed; (2) Standardizing data formats: Consistent formats for dates, naming conventions, and units of measurement were enforced across all data sources; (3) Addressing missing data or outliers: Techniques such as imputation were applied to handle missing data, while extreme outliers were either removed or capped, depending on their influence on the analysis.

3.3 Feature Engineering and Dataset Preparation

3.3.1 Feature Engineering

Derivation of New Features

In this study, new features were derived to enhance the predictive model. Rolling performance metrics, such as the average finishing position in the last five races, were computed to track recent trends in horse performance. Additionally, speed figures were calculated based on race times and adjusted for track conditions to provide a more accurate measure of performance. Interaction terms, such as combinations of trainer and jockey pairs, were also introduced to capture important collaborative effects that might influence race outcomes.

Development of Complex Indicators

Several complex indicators were developed to further refine the analysis. One such indicator was a proprietary "form index," which incorporated various factors, including recent performance, class level, and consistency, to give a comprehensive view of a horse's form. Furthermore, speed ratings were adjusted to account for track bias and race conditions, providing a more contextualized assessment of performance.

Feature Selection

For feature selection, statistical tests, such as chi-square and ANOVA, were employed to identify features with significant predictive power. Dimensionality reduction techniques, particularly Principal Component Analysis (PCA), were utilized to reduce the feature space while retaining the most important components. Additionally, regularization methods, including Lasso and Ridge, were applied to determine the relative importance of each feature, helping to avoid overfitting and improve the model's generalizability.

Dataset Preparation

The dataset was divided into training and testing sets using a chronological split, where the first 80% of the data by date was used for training, and the remaining 20% was reserved for testing. This approach ensured that temporal integrity was maintained, thereby preventing data leakage. A time series cross-validation strategy was adopted to account for the temporal nature of the racing data, ensuring that the model was trained and validated in a way that reflected real-world conditions.

Data Scaling and Normalization

Numerical features were standardized to ensure consistency across variables, which is essential for models sensitive to the scale of input data. For categorical variables, encoding techniques such as one-hot encoding and label encoding were applied, ensuring that these variables could be used effectively in the modeling process.

4 METHODOLOGY

4.1 Model Selection

After extensive research and experimentation, the following models were selected for the horse racing prediction system:

- **LightGBM**: A gradient boosting framework that uses tree-based learning algorithms. It was chosen for its high efficiency and
- ability to handle large-scale data with lower memory usage. **XGBoost**: An optimized distributed gradient boosting library. It was selected for its proven performance in various machine learning competitions and its ability to handle complex feature interactions.
- **CatBoost**: A gradient boosting library with advanced handling of categorical features. It was included for its ability to automatically deal with categorical variables without extensive preprocessing.
- **AdaBoost**: An ensemble learning method that combines weak learners to create a strong predictor. It was chosen for its ability to focus on difficult-to-predict samples.
- **HistGradientBoosting**: A scikit-learn implementation of gradient boosting that builds histograms for faster training. It was selected for its speed and memory efficiency.

- **TabNet Regressor**: A deep learning model designed for tabular data. It was included to capture complex non-linear relationships that tree-based models might miss.

These models were selected based on the following criteria:

- **Performance**: All chosen models have demonstrated strong predictive power in similar tasks and competitions.
- **Diversity**: The selection includes both traditional boosting methods and newer approaches like TabNet to capture different aspects of the data.
- **Handling of Feature Types**: Given the mix of numerical and categorical features in horse racing data, models like CatBoost and LightGBM, which handle categorical features well, were prioritized.
- **Scalability**: Models that can efficiently handle large datasets were preferred, considering the volume of historical racing data.
- **Interpretability**: While some models (like TabNet) are less interpretable, the ensemble includes models that provide feature importance, allowing for some level of explanation in predictions.

4.2 Ensemble Stacking

Stacking is a strong ensemble learning strategy in machine learning that combines the predictions of numerous base models to get a final prediction with better performance. It is also known as stacked ensembles or stacked generalization.

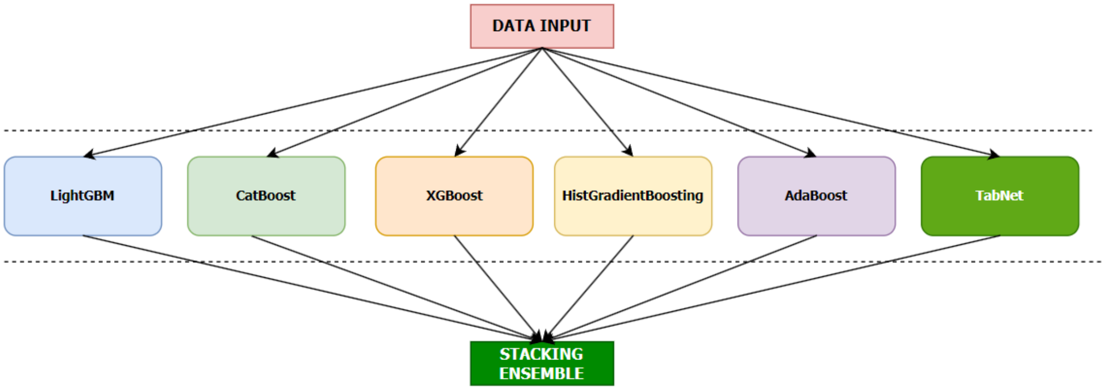

Figure 1: Architure of Stacking Ensemble

The diagram illustrates the architecture of a stacking ensemble in machine learning, a powerful ensemble strategy that combines predictions from several base models to achieve improved performance. This method, also referred to as stacked generalization, leverages the complementary strengths of various models.

Data Input: The process begins by feeding input data into several base models, which are designed to capture different aspects of the data. In this particular ensemble, the models used are:

- **LightGBM**: A gradient boosting model focused on speed and efficiency.
- **CatBoost**: A boosting method that handles categorical data efficiently.
- **XGBoost**: A widely-used, scalable gradient boosting model.
- **HistGradientBoosting**: Another efficient gradient boosting algorithm for tabular data.

- **AdaBoost**: A boosting algorithm that combines weak learners to form a strong learner.
- **TabNet**: A deep learning model using attention mechanisms, specifically designed for tabular data.

The predictions from these base models are passed to a meta-model, which in this case is **XGBoost**. The meta-model aggregates the outputs of the base models and makes the final prediction. **XGBoost**, known for its robustness and performance in gradient boosting, helps refine the final output by learning how to best combine the individual model predictions.

4.3 System design

The system pipeline is designed to process vast amounts of real-time and historical data to provide accurate predictions for horse racing events. The architecture integrates data retrieval, aggregation, processing, and prediction modeling, all the way through to betting automation.

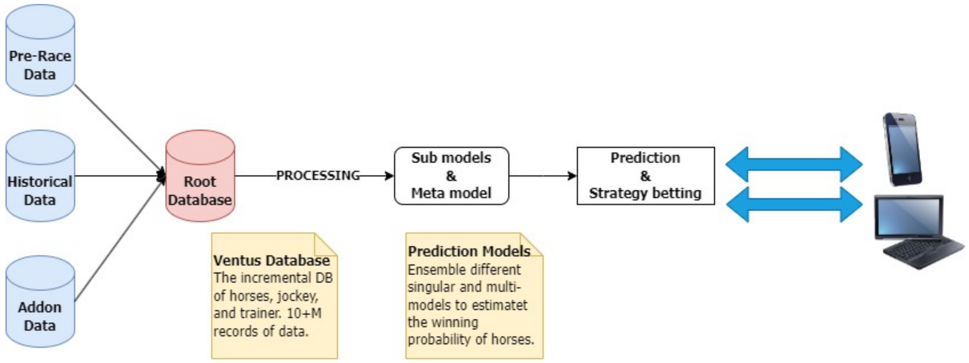
Pre-Race Data
Historical Data
Addon Data
Root Database
PROCESSING
Sub models & Meta model
Prediction & Strategy betting
Ventus Database
The incremental DB of horses, jockey, and trainer. 10+M records of data.
Prediction Models
Ensemble different singular and multi-models to estimatet the winning probability of horses.

Figure 2: System Architecture

4.3.1 Data Retrieval

System retrieves data from multiple sources, including: 3rd party data sources for daily updates. Historical data that spans several years. The Root Database, which is continuously updated with horse, jockey, and trainer information. This system works in real-time to gather relevant data about upcoming races, ensuring that the predictions are based on the most recent and historical performance metrics.

4.3.2 Data Aggregation and Processing

Once data is collected, it is passed through a Data Aggregator, where information from different sources is combined. This is followed by Data Processing, where the aggregated data is prepared for use in prediction models. Here, the data is cleaned and structured, ready to be used by the prediction system.

4.3.3 Root Database

The Root Database serves as the central repository for all racing data, containing over 10 million records on horses, jockeys, and trainers. This database is incrementally updated, allowing for continuous improvements in the accuracy of predictions. In addition to the main database, there are specialized sub-databases for:

- **Horses**
- **Jockeys**
- **Trainers**

These sub-databases ensure that detailed profiles of each entity are maintained and accessible for model predictions.

4.3.4 Prediction Models

The system utilizes an ensemble of singular and multi-models to predict the likelihood of a horse's success in an upcoming race. Each horse is individually evaluated by the prediction models, which analyze factors such as past performance, jockey, and trainer influence. A Master Model then consolidates the predictions from these individual models to generate the final betting strategy. The use of an ensemble approach helps to minimize prediction errors by combining the strengths of multiple models. 5. Betting Automation Once the predictions are finalized, the system moves into the betting phase:

- **Ranking Model:** The predictions are ranked to prioritize the best betting opportunities.
- **Betting Model:** The betting strategy is applied, determining how much to bet on each horse based on the predicted probability of success.
- **Automated Betting (AmTote)**: Bets are automatically placed through the AmTote gateway, ensuring timely execution of the strategy.

4.3.5 Result Display and Reporting

After the predictions and betting strategies have been processed, the results are automatically displayed on the system's website interface. Users can view the predictions in real-time, along with the recommended betting strategies.

- **Daily Reporting**: In addition to live results, the system generates daily performance reports. These reports provide an overview of the day's betting outcomes, highlighting successes and areas for improvement.
- **End-of-Day Updates**: The reports are updated at the end of each day, offering detailed insights on performance metrics for the day's races. These reports are tailored based on specific regions, ensuring that the predictions and outcomes are localized for different time zones and racing events.
- **Regional Customization**: Depending on the user's location, the reports can be customized to display information relevant to their region, offering insights into local betting trends and performance statistics.

These continuous updates help users track the effectiveness of their betting strategies, while also allowing the system to further optimize and refine its prediction models based on real-world outcomes.

5 EVALUATION METRICS

To assess the performance of prediction system, following:

1. **Accuracy Score (Accu)**: It is a measure of performance of the model given in percentage.

$$\text{Accuracy} = \frac{T_rP + T_rN}{(T_rP + F_sP) + (T_rN + F_sN)} \tag{1}$$

where, T_rP = number of True Positives, T_rN = number of True Negatives, F_sP = number of False Positives, F_sN = number of False Negatives.

2. **F1 Score**: It is a performance metric for classification systems, especially useful for imbalanced class structures, and is calculated as follows:

$$\text{F1} = 2 \cdot \frac{1}{\left(\frac{1}{\text{precision}} + \frac{1}{\text{recall}}\right)} \tag{2}$$

where:

$$\text{precision} = \frac{T_r P}{T_r P + F_s P} \tag{3}$$

$$\text{recall} = \frac{T_r P}{T_r P + F_s N} \tag{4}$$

Precision is the ratio of correctly predicted positive observations to the total predicted positives, and recall is the ratio of correctly predicted positives to the total actual positives. A high F1 score indicates that both precision and recall are high, making it a desirable metric for the performance of the model.

3. **Receiver Operator Characteristic (ROC)**: It is a probability curve that ranges from 0 to 1. The True Positive Rate (TPR) is plotted against the False Positive Rate (FPR).

   The True Positive Rate (TPR) is given by:

   $$\text{TPR} = \frac{T_r P}{T_r P + F_s N} \tag{5}$$

   The False Positive Rate (FPR) is given by:

   $$\text{FPR} = \frac{F_s P}{F_s P + T_r N} \tag{6}$$

4. **Area Under the Curve (AUC)**: It is used to determine the performance of various machine learning models by measuring the area under the ROC curve. A higher value of AUC signifies that the classifier distinguishes well between positive and negative classes.

For this study, Python 3.12.3 is used to apply machine learning methodology. After loading and data preprocessing, the dataset is split into train and test data (70-30 split) using a random number so as to ensure random sample representation of the original problem dataset. Then different ML classifier models are applied and their performance efficiency is tested by using metrics as discussed earlier. The results are discussed in the following next section.

6 RESULTS

The results are presented in two ways. The first is the favorite horse method, considered naive, as it involves betting solely on the horse with the lowest odds in each race. Unsurprisingly, favorite horses (with odds between 1 and 2) have a 30% higher chance of winning compared to outsiders. However, betting on the horse with the lowest odds often leads to a negative net gain.

6.1 Favorite Horse and Base Strategy

In horse racing terms, the "**favorite**" is the horse with the most money bet on it to win. This lowers its parimutuel odds. Determine the horse race favorite by finding the horse with the lowest odds. While a favorite has popularity, it also often has a low payoff; most often a \$3-\$6 range for a \$2 bet.

Favourite horse results (independent of model output) across Australia (AUS) and United Kingdom (UK) regions were evaluated using four strategies. Each strategy involves betting on the favorite horse(s) under different conditions to optimize accuracy or profit:

1. **Bet only on the favorite horse**: The horse with the lowest odds (typically between 1 and 2) is selected, often leading to low returns.
2. **Bet on the top 3 favorite horses:** The three horses with the lowest odds are selected, increasing the chances of winning but reducing profit due to low odds.

Base strategy refers to simple and straightforward strategies based on the output of the model. Two such strategies are proposed below:

3. **Top 1 MODEL OUTPUT**
   This strategy involves selecting the horse predicted to have the highest probability of winning, according to the model output. Essentially, you place a bet on the horse ranked first by the model in terms of predicted performance. This is a straightforward approach that assumes the model's top-ranked horse has the highest chance of winning.
4. **Top 3 MODEL OUTPUT**
   In this strategy, you select the top three horses according to the model output. Instead of only betting on the top-ranked horse, you place bets on the first three horses predicted by the model. This approach spreads the risk across more potential winners and increases the chance of selecting the actual winner from a small pool of contenders.

Note: In all strategies, the bet type is WN (Win or Not), focusing on whether the selected horse(s) win or not. Results provide insight into how betting on favorites affects profitability in AUS and UK regions.

Approach	No. Bet	Precision	F1	ROC-AUC	ROI 1st
1	5149	32.96%	32.97	62.98	-3.03%
2	15447	22.32%	33.49	72.25	-4.05%
3	5149	25.66%	25.67	58.86	2.84
4	15447	18.96%	28.45	66.17	-0.45%

Table 1: Favorite Horse Performance in the Australian Region from 01/06/2024 to 30/09/2024

Approach	No. Bet	Precision	F1	ROC-AUC	ROI 1st
1	4504	16.85%	16.83	54.51	-51.41%
2	13511	17.1%	25.64	65.42	-36.34%
3	4504	25.84%	25.82	59.31	-14.78%
4	13511	19.07%	28.58	67.78	-19.36%

Table 2: Favorite Horse Performance in the United Kingdom Region from 01/06/2024 to 30/09/2024

The results presented in Tables 1 and 2 provide an insightful comparison of the performance of base strategies in both the Australian and United Kingdom horse racing markets. The Top 3 MODEL OUTPUT strategy consistently offers a better balance between recall and precision across both regions, leading to more consistent model performance. However, the Top 1 MODEL OUTPUT strategy can provide positive returns in specific market conditions, as observed in the Australian region. The analysis highlights that while precision is important, a diversified betting strategy (betting on the top 3 horses) can yield more reliable results, especially in terms of model robustness and predictability. These results suggest that future efforts could focus on refining the Top 3 approach to improve ROI, particularly in more competitive and unpredictable markets such as the UK.

6.2 Proposed Strategy

To improve upon the favorite horse method, a more sophisticated strategy is proposed. Instead of betting solely on the lowest odds, this strategy analyzes various factors such as the horse's recent performance, jockey statistics, track conditions, and historical race data. By combining these factors with machine learning models, this approach aims to identify not only the favorites but also potential value bets with higher returns. The goal is to strike a balance between higher winning probabilities and maximizing profits, avoiding the pitfalls of simply betting on the lowest-odds horse.

The proposed strategy is based on the model output discussed in the previous sections. By leveraging the predictions generated from our model, this approach dynamically adjusts bets according to the horse's predicted performance, allowing for more informed decisions and optimizing potential returns.

Two main strategies are presented, focusing on different approaches.

1. **Optimize Precision (optimize_pre)**: This strategy aims to maximize the model's accuracy in predicting winning horses, excluding the favorites (horses with odds between 1 and 2).
2. **Optimize ROI for Underdogs (optimize_roi_underdog)**: This strategy focuses on predicting underdogs (horses with higher odds and lower expectations of winning), with the goal of maximizing potential profit.

Approach	No. Bet	Precision	F1	ROC-AUC	ROI 1st
1	3055	9.03%	6.73	49.72	-1.98%
2	736	4.48%	1.12	49.57	23.98%

Table 3: Proposed Strategy Performance in the AUS Region from 01/06/2024 to 30/09/2024

Approach	No. Bet	Precision	F1	ROC-AUC	ROI 1st
1	847	5.43%	1.71	49.59	-11.6%
2	288	17.7%	0.18	50.3	20.94%

Table 4: Proposed Strategy Performance in the UK Region from 01/06/2024 to 30/09/2024

The proposed strategies highlight the importance of balancing precision and profitability in horse racing predictions. While precision-focused strategies may seem attractive, they tend to result in lower returns due to the competitive nature of the markets. Conversely, targeting underdogs, though riskier, can yield substantial profits if managed correctly. These results suggest that the Optimize ROI for Underdogs strategy is particularly effective in generating positive ROI in both Australian and UK markets, making it a compelling approach for bettors looking to maximize returns while taking on higher risks.

6.3 Compare Results

Approach	No. Bet	Precision	F1	ROC-AUC	ROI 1st
FAVORITE HORSE AND BASE STRATEGY					
1	5149	32.96%	32.97	62.98	-3.03%
2	15447	22.32%	33.49	72.25	-4.05%
3	5149	25.66%	25.67	58.86	2.84
4	15447	18.96%	28.45	66.17	-0.45%
PROPOSED STRATEGY					
1	3055	9.03%	6.73	49.72	-1.98%
2	736	4.48%	1.12	49.57	23.98%

Table 5: All Strategy Performance in the AUS and UK Region from 01/06/2024 to 30/09/2024

The comparative analysis of the strategies shows that betting on favorites (Approach 1 and 2 of the base strategy) leads to higher precision but generally results in negative ROI, making these strategies less attractive from a profitability standpoint. On the other hand, the proposed strategy, particularly Approach 2 (Optimize ROI for Underdogs), demonstrates the highest potential for profitability despite low precision. This suggests that a riskier approach focusing on underdogs could be more rewarding in the long term, especially when the goal is maximizing ROI rather than achieving high prediction accuracy.

The Receiver Operating Characteristic (ROC) curves - The Area Under the Curve (AUC) value is used to measure the quality of each strategy in predicting successful outcomes.

- Across both regions (AUS and UK), the strategies that optimize Top 3 Live Odds and Top 3 Model Outputs generally show better predictive performance compared to others, with AUC values above 0.65 in both regions. This suggests that selecting a group of top-ranked horses either based on live odds or model output leads to better discriminative power.

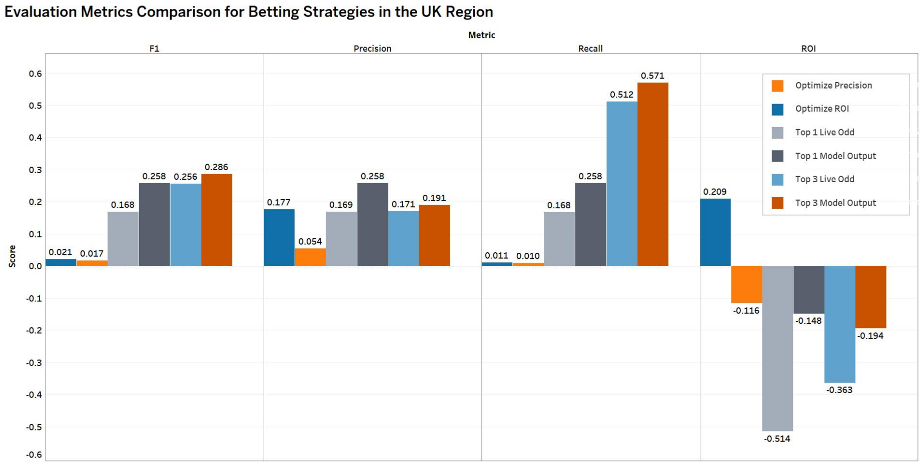

Figure 3: Evaluation Metrics Comparison for Betting Strategies in the UK Region

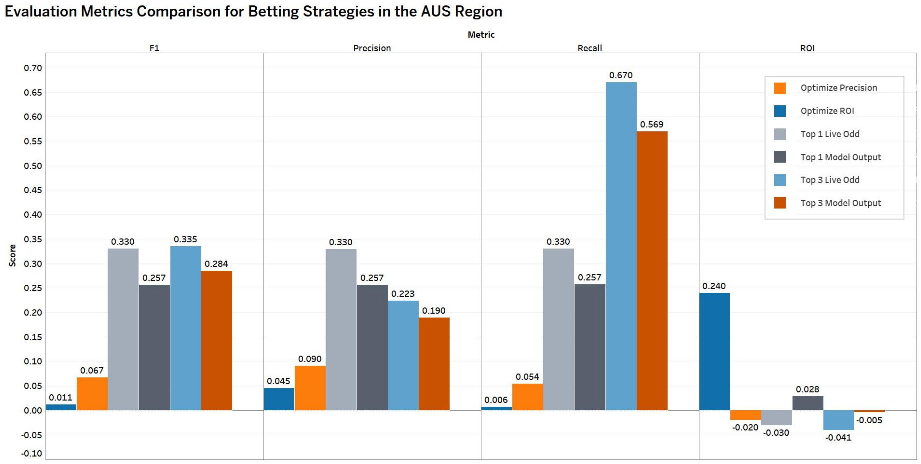

Figure 4: Evaluation Metrics Comparison for Betting Strategies in the AUS Region

- Strategies that focus on Optimize ROI and Optimize Precision consistently perform worse than others, with AUC values around 0.49-0.50, which is close to random guess (AUC = 0.5). Solely focusing on maximizing return on investment or precision might not be effective for horse racing prediction.

- Interestingly, the Top 1 Live Odd and Top 1 Model Output strategies show a moderate level of performance, with slightly different results between the AUS and UK regions. The reliability of the top single prediction can vary depending on the specific characteristics of each region's horse racing environment.

While the base strategy offers higher accuracy, the proposed strategy focusing on underdogs is more effective in terms of ROI, making it a better option for bettors who are willing t take higher risks for potentially greater financial rewards. Strategies focusing on ROI or a single top selection are comparatively less effective. These insights can be used to enhance betting models by focusing on the inclusion of multiple promising candidates to improve predictive accuracy in both AUS and UK racing environments.

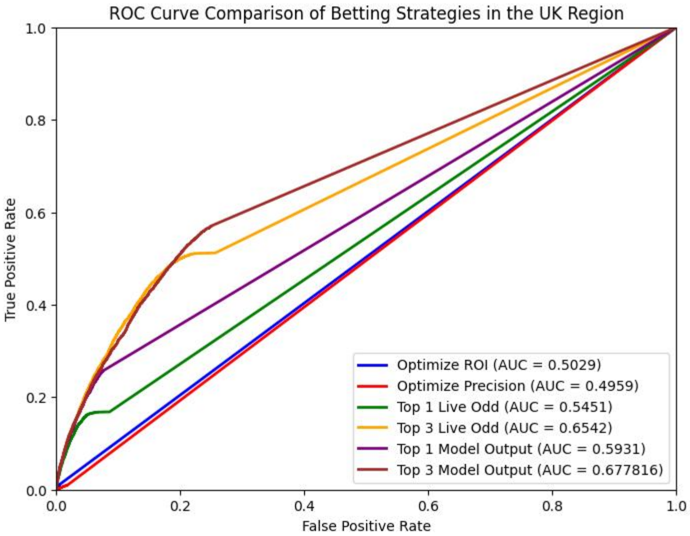

Figure 5: ROC Curve Comparison of Betting Strategies in the UK Region

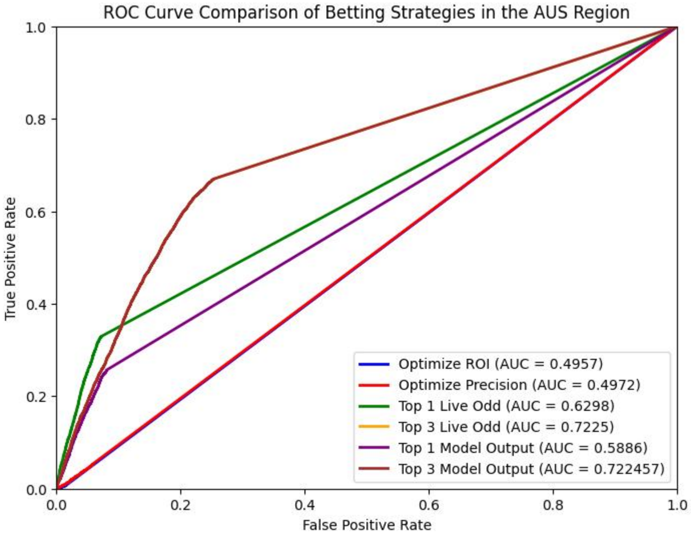

Figure 6: ROC Curve Comparison of Betting Strategies in the UK Region

7 CONCLUSION

Through this study, the aim is to analyze Machine Learning modeling techniques to predict horse race winners. The proposed ML framework demonstrates effective prediction performance and has outperformed previously reported predictive models. This paper also reviews the available literature on ML modeling, focusing on algorithms for predicting sports results. The insights gained from this research can be beneficial to both fellow researchers and stakeholders such as punters, horse race managers, horse owners, and club owners, who can use the findings to improve decision-making. Furthermore, this research can be extended to other domains where prediction is valuable, such as healthcare, finance, and education etc.

Acknowledgements We are grateful to My Quyen for her fruitful comments, corrections and inspiration.

References

[1] P. L. Hitchens, K. Ryan, and S. I. Koch, “A sustainable structure for jockey injury data management for the North American horse racing industry,” Injury, vol. 50, no. 8, pp. 1418–1422, 2019.

[2] J. M. Fiedler and P. D. McGreevy, “Reconciling horse welfare, worker safety, and public expectations: horse event incident management systems in Australia,” Animals, vol. 6, no. 3, pp. 16–35, 2016.

[3] E. Juckes, J. M. Williams, and C. Challinor, “Racing to a staffing solution: an investigation into the current staffing crisis within the UK horseracing industry,” Comparative Exercise Physiology, vol. 17, no. 1, pp. 73–89, 2021.

[4] S. H. Kang and G. S. Park, “Overcoming ethical issues through symbolic management, cultivating proponents and storytelling: the institutionalization of Korea’s horseracing industry,” Asia Pacific Business Review, vol. 22, no. 3, pp. 439–451, 2016.

[5] E. Davies, W. McConn-Palfreyman, and J. M. Williams, “*e impact of COVID-19 on staff working practices in UK horseracing,” Animals, vol. 10, no. 11, pp. 2003–2034, 2020.

[6] S. Zhang and M. Liu, “Design of horse race registration system based on wireless network and simulation system,” Techniques and Applications, vol. 34, no. 17, pp. 1661–1669, 2021.

[7] K. Clayton-Hathway and U. Fasbender, “Women as leaders and managers in sports: Understanding key career enablers and constraints in the British horseracing industry,” Women, Business and Leadership, vol. 7, no. 3, pp. 309–322, 2019.

[8] Z. Ma, “Research on system and developing path of talents collaborative cultivation in horse racing industry between hubei and xinjiang,” Education Science and Economic Management, vol. 16, no. 23, pp. 612–615, 2017.

[9] H. Ghezelsefloo, “Designing structural model of elemental, behavioral and motivational traits in horse racing betting with mix method,” Applied Research in Sport Management, vol. 9, no. 2, pp. 99–110, 2020.

[10] K. Fenner, K. Dashper, and J. Serpell, “*e development of a novel questionnaire approach to the investigation of horse training, management, and behaviour,” Animals, vol. 10, no. 11, pp. 56–68, 2020.

[11] S. Zhang and M. Liu, “Computer aided management system of sports horse registration based on distributed storage system and deep Fusion learning,” Microprocessors and Microsystems , vol. 56, no. 26, pp. 3144–3153, 2021.

[12] C. Quintana, B. Grimshaw, and H. E. Rockwood, “Differences in head accelerations and physiological demand between live and simulated professional horse racing,” Comparative Exercise Physiology, vol. 15, no. 4, pp. 259–268, 2019

[13] K. Fenner, M. Hyde, and A. Crean, “Identifying sources of potential bias when using online survey data to explore horse training, management, and behaviour: a systematic literature review,” Veterinary Sciences, vol. 7, no. 3, pp. 18–33, 2020.

[14] Z. Sun and Y. Li, “Research on Chinese speed horse racing guessing lottery issuance based on internet big data,” Design Engineering, vol. 183, no. 59, pp. 405–476, 2020.

[15] B. Padalino, S. L. Raidal, and E. Hall, “Survey of horse transportation in Australia: issues and practice,” Australian Veterinary Journal, vol. 94, no. 10, pp. 349–357, 2016.

[16] S. V. R. Hoseini and M. Amani, “Impact of constraints and behavioral motivations on loyalty of horse racing spectators,” Journal of History Culture and Art Research, vol. 7, no. 2, pp. 14–27, 2018.

[17] Y. Xiao, C. Xing, and T. Zhang, “An intrusion detection model based on feature reduction and convolutional neural networks,” IEEE Access, vol. 117, no. 2, pp. 103–187, 2019

[18] R. Y. Chen, “A traceability chain algorithm for artificial neural networks using T–S fuzzy cognitive maps in blockchain,” Future Generation Computer Systems, vol. 183, no. 23, pp. 109–119, 2018.

[19] M. Hahsler and R. Karpienko, “Visualizing association rules in hierarchical groups,” Journal of Business Economics, vol. 87, no. 3, pp. 317–335, 2017.

[20] T. Osadchiy, I. Poliakov, and P. Olivier, “Recommender system based on pairwise association rule,” Expert Systems with Applications, vol. 115, no. 12, pp. 1871–1888, 2019.

[21] R. Rekik, I. Kallel, and J. Casillas, “Assessing web sites quality: A systematic literature review by text and association rules mining,” International Journal of Information Management , vol. 38, no. 1, pp. 201–216, 2018.