Junior Quant Test - Horse Racing

Instructions

- Dataset: test_dataset.csv (provided).
- Submit a concise report (PDF, RMarkdown/html, or Jupyter Notebook).
- Use **R or Python**.
- Include code, figures, and reasoning.
- Do **not** use any obs__* variables as direct predictors, since these are only known after the race starts. You can use past values though. You may also use them for evaluation where relevant.

Questions

Q0. Explore the dataset

Q1. Explore the peak age of horses

- For each race_type_simple, analyze horse performance by age.
- Identify and report the peak performance age.

Q2. Feature engineering: Build a rating for each horse, jockey and trainer

- Construct a rating system for horses, jockeys and trainers based only on past performance data (i.e. excluding future information and obs__* variables from the current race).

Q3. Build a predictive model

- Build a predictive model for race winners using your ratings and other covariates (but excluding current race obs__* variables). Your model could produce win probabilities for each runner that sum to 1 in each race for example.

Q4. Compare to the betting market

- Compare your model performance vs the Betfair Starting Price (obs__bsp).
- Does your model beat the Betfair Starting Price overall or in some subset of races?
- Can you find a profitable strategy assuming zero commission?
- Assess the **statistical significance** of any of the results reported.

Evaluation Rubric

Criterion	Weight	Notes
Reasoning & interpretation	50%	Quality of analysis, clarity of assumptions, detailed discussion of results, understanding of peak age, ratings, and model behavior.
Modeling & methodology	25%	Appropriateness of rating system and predictive model; correct exclusion of obs__* variables; thoughtful feature usage.
Data handling & exploration	15%	Correct handling of dataset, summaries, and plots.
Code clarity & reproducibility	10%	Readable, well-structured code; clear workflow; reproducible results.

Key point: Reasoning and discussion are **more important than perfect coding**. Focus on explaining *why* your choices make sense and what insights the analysis reveals.

Dataset Description

The dataset contains one row per **horse per race**. We removed races with late non-runners and also put Bumpers races into Flat Turf

Column	Description
date	The date the race took place.
racecourse_country	The country where the racecourse is located.
racecourse_name	The name of the specific racecourse.
race_time	The scheduled start time of the race.
race_id	A unique identifier for the race.
race_distance	The length of the race in meters.
race_type	The specific type of race (e.g., Bumpers, Flat, Hurdle, Chase).
race_type_simple	A simplified version of the race type (e.g., Flat Turf, Flat AW, Hurdle, Chase).
going_clean	The condition of the race track surface (e.g., Good, Soft, Standard).
n_runners	The total number of horses participating in the race.
horse_id	A unique identifier for each horse.
horse_name	The name of the horse.
age	The horse's age on the day of the race.
official_rating	A rating assigned to the horse by the official handicapper.
carried_weight	The total weight the horse was assigned to carry during the race.
draw	The horse's starting stall number.

Column	Description
jockey_id	A unique identifier for the jockey.
jockey_name	The name of the jockey.
trainer_id	A unique identifier for the trainer.
trainer_name	The name of the trainer.
ltp_5min	The last traded price of the horse 5 minutes before the race.
obs__bsp	The Betfair Starting Price (decimal odds) at the start of the race.
obs__racing_post_rating	A post-race performance rating assigned by the Racing Post.
obs__uposition	The horse's finishing position in the race.
obs__is_winner	A binary indicator (1 if the horse won, 0 otherwise).
obs__top_speed	A post-race rating indicating the horse's top speed.
obs__distance_to_winner	The distance in lengths the horse finished behind the winner.
obs__pos_prize	The prize money earned for the finishing position.
obs__completion_time	The time it took for the horse to complete the race in seconds.