🐎 Horse Racing Probabilistic Modeling Assignment

Overview

You are provided with a dataset containing information about horse races and individual runners.

Your task is to build a model that **outputs the probability of each horse winning its race**. The **final output must be probabilities**, such that for each race, the **sum of predicted probabilities across all runners is exactly 1**.

You may choose any appropriate **target variable and modeling approach**, as long as the final output satisfies the probability constraint.

You are not required to use any specific algorithm. We are more interested in your reasoning, feature engineering, and ability to produce a valid probabilistic output.

Deliverables

You Must Submit:

1. **Predicted probabilities** for the provided **test set**, formatted as a CSV with the following columns:
    - Race_ID, Horse, Predicted_Probability
    - These probabilities must **sum to 1 for each race**.
2. Your **code** (in a reproducible script or notebook)
    - You are free to use any modeling approach or target variable, as long as your **final output is a set of probabilities** for each horse winning.
3. A brief **write-up** (1–2 pages max) covering:
    - Your **feature selection and modeling choices**
    - Any **feature engineering**, especially race-relative features (e.g., how a horse's previous odds, speed, or rating compares to the rest of the field)
    - How you **transformed model outputs into valid probabilities** that sum to 1 per race
    - **Evaluation results** using at least **Log Loss** and/or **Brier Score**
    - Any key **assumptions, limitations, or challenges** you encountered

Dataset Format

You have been provided with:

- train.csv: historical races to train your model
- test.csv: races for which you will predict win probabilities

Do **not** train or tune your model using the test data.

Variable Descriptions

Variable Name	Description
Race_Time	The official start time of the race and date (e.g., "16:02").
Race_ID	Unique identifier for the race.
Course	Name of the racecourse (e.g., "Ascot", "Cheltenham").
Distance	Distance of the race in standard format (e.g., "1m4f").
distanceYards	Distance of the race in yards.
Prize	Total prize money for the race (in local currency, e.g., GBP).
Going	Track condition (e.g., "Good", "Soft", "Heavy").
Horse	Name of the horse competing in the race.
Trainer	Name of the trainer
Jockey	Name of the jockey
betfairSP	Betfair Starting Price (decimal odds) at race start.
Position	Finishing position of the horse (1 = winner, 2 = second, etc.).
timeSecs	Time the horse took to complete the race, in seconds.
pdsBeaten	Pounds the horse was beaten by (based on distance and standard scale).
NMFP	Normalized finishing position defined as 1-(Position/Runners)
Runners	Total number of horses that started the race.
Age	Age of the horse (in years).
Speed_PreviousRun	Speed rating from the horse's previous race.
Speed_2ndPreviousRun	Speed rating from the horse's second previous race.
NMFPLTO	Normalized finishing position from the horse's last race
MarketOdds_PreviousRun	Market odds from the horse's last race.
MarketOdds_2ndPreviousRun	Market odds from the horse's second last race.
TrainerRating	Numerical performance indicator of the trainer
JockeyRating	Numerical performance indicator of the jockey.
daysSinceLastRun	Number of days since the horse's last race
SireRating	Average performance indicator of the horse's sire (father).
DamsireRating	Average performance indicator of the horse's damsire (mother's sire).
meanRunners	Average number of runners in last four runs

⚠️ Important Note on Data Leakage

The following variables are **only known after the race has started or finished**, and **must not be used as input features** in your predictive model:

- betfairSP – the Betfair Starting Price
- Position – the finishing position of the horse
- timeSecs – the horse's final race time
- pdsBeaten – how far the horse was beaten by
- NMFP – normalized finishing position

Including these variables in your model would constitute **target leakage** and invalidate your results. You may use them **only for evaluation purposes**, such as computing log loss or Brier score on the test set.

📊 Evaluation Criteria

Your submission will be evaluated on the following:

1. Probabilistic Accuracy

- We will use **Log Loss** and **Brier Score** to assess the quality of your predicted probabilities.
- Your predictions should reflect **realistic uncertainty**, not just classification.

2. Validity of Output

- Predicted probabilities must **sum to 1 per race**.
- Submissions that fail this constraint will not be considered valid.

3. Methodological Soundness

- We'll look at how well you:
    - Justified your feature and target choices
    - Engineered features thoughtfully, especially **race-relative** ones

4. Market Comparison (Optional)

- While the market (betfairSP) is not allowed as a feature, you may compare your predicted probabilities against **market-implied odds** for insight.

Submission Instructions

Please submit all deliverables within one week of receiving this email to toby@parametricai.co.uk

Late submissions may not be considered.