**Predicting Horse Race Outcomes: A Data-Driven Modeling Approach – Gaoxiang Chen**

**1. Introduction**

The prediction of horse race outcomes presents a challenging yet highly relevant application of machine learning in real-world decision-making. By leveraging structured and semi-structured data from historical races, this project seeks to develop a robust predictive model capable of estimating finishing positions or performance scores of horses in upcoming races. This project integrates XML data parsing, data cleaning, feature engineering, and probabilistic modeling using Bayesian Additive Regression Trees (BART).

**2. Data Sources and Structure**

The raw data comes from two XML datasets:

* **Past Performance Data**: Contains detailed pre-race information about horses, including weight carried, odds, jockey, trainer, and origin.
* **Results Data**: Captures post-race outcomes such as finish time, official placement, speed ratings, and final odds.

By parsing both sources and merging them on key identifiers (RaceNumber, HorseName), we construct a supervised learning dataset suitable for regression modeling.

**3. Data Processing and Feature Construction**

Key steps in data processing include:

* **XML Parsing**: Python's xml.etree.ElementTree is used to extract structured information from the unstructured XML format.
* **Cleaning and Merging**: Fields are stripped of whitespace, standardized to lowercase, and joined into a unified DataFrame.
* **Feature Engineering**:
  + Odds strings (e.g., '5/2') are converted to floating point decimal values.
  + Numerical columns (FinishTime, WeightCarried, SpeedRating) are coerced to float.
  + Categorical fields are retained for potential downstream encoding.

The final dataset is exported as a clean .csv file for modeling and reproducibility.

**4. Modeling Methodology**

We adopt a probabilistic approach using **Bayesian Additive Regression Trees (BART)** implemented in PyMC-BART for the following reasons:

* BART provides a nonparametric way to model nonlinear relationships and interactions without explicit specification.
* Bayesian methods enable us to quantify uncertainty in predictions—a key advantage in high-stakes applications like betting or decision support.

Steps involved:

* **Preprocessing**: We scale the feature space using MinMaxScaler.
* **Training**: A posterior distribution over predictions is sampled using Markov Chain Monte Carlo (MCMC) with pm.sample().
* **Prediction**: Posterior means are computed for both training and test sets.

**5. Evaluation and Visualization**

The model is assessed via:

* **Quantitative Metrics**: RMSE, MAE, and R² for both training and testing sets.
* **Visual Tools**:
  + Scatter plot of actual vs predicted results.
  + Histogram of residuals for test data.

These tools give a holistic view of model performance and potential overfitting.

**6. Key Results**

* The model demonstrates reasonable predictive performance, with consistent behavior across training and test splits.
* Residuals are roughly normally distributed around zero, suggesting low systemic bias.

**7. Reflections and Further Considerations**

While the project successfully builds a pipeline from raw XML to probabilistic forecasting, several deep considerations arise:

**a. Data Bias and Representativeness**

The dataset includes only completed races, meaning horses that were scratched or disqualified are not part of the training data. This selection bias could affect generalizability, especially in forecasting rare events.

**b. Temporal Drift**

No temporal component was incorporated. Performance of jockeys, horses, or trainers may drift over time. Incorporating a time-indexed feature or using time-series models (e.g., BART + GP) could improve forecasting fidelity.

**c. Explainability**

While BART offers uncertainty estimates, its black-box nature makes it hard to extract clear decision rules. Future work could include:

* Feature importance via BART split frequencies
* Model-agnostic tools like SHAP for post-hoc interpretability

**d. Practical Use and Ethics**

In betting contexts, model misuse or overfitting to historical data could lead to financial loss. Additionally, if used for animal performance prediction in real races, ethical considerations (e.g., pressure on jockeys or horse health) must be acknowledged.

**e. Comparative Modeling**

It would be valuable to benchmark BART against traditional regressors (e.g., Ridge, Random Forest, XGBoost) using the same feature set. This helps validate the Bayesian model’s benefits and understand trade-offs in interpretability and performance.

**8. Conclusion**

This project demonstrates a full-cycle machine learning application—from data ingestion and preparation, through modeling, to rigorous evaluation—applied to a novel and complex domain. While BART provides flexibility and probabilistic insight, it invites further exploration in interpretability, temporal modeling, and real-world deployment constraints. Future work will benefit from richer feature sets, time-awareness, and ethical reflection on model impact.