41ST ASIAN RACING CONFERENCE
RIYADH 2026

Machine Learning for Predictive Horse Performance

Jack Zuber

The Hong Kong Jockey Club

11/02/26

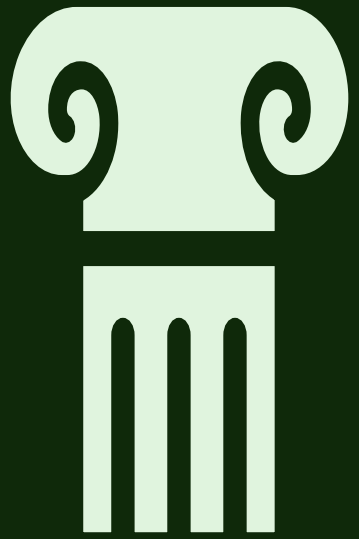
ASIAN RACING FEDERATION

نادي سباقات الخيل
JOCKEY CLUB OF SAUDI ARABIA

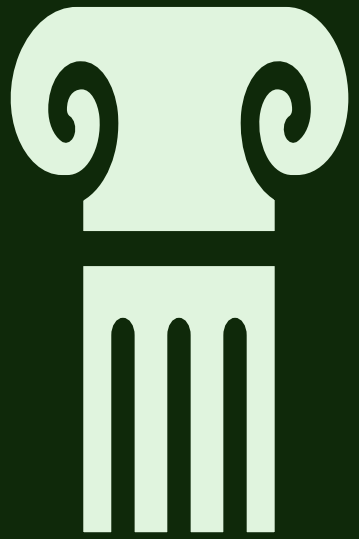
BACKGROUND

What does a betting analyst do?

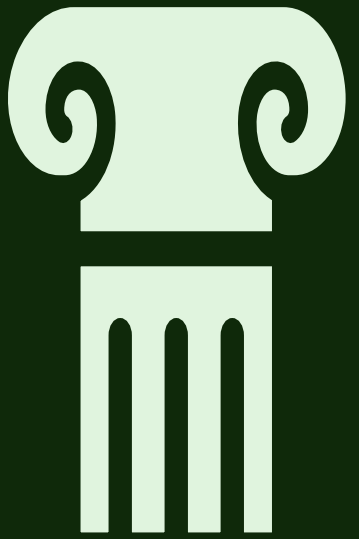
Monitor Betting Markets
Live + Historical



Analyse Betting Data
X
Analyse Performance Data



Identify Suspicious Trends



Past Performance can Define an Expectation for both Betting and Actual Performance

There usually exists a parity between expectation and actual outcomes

When the link between:

Past Performance and Betting

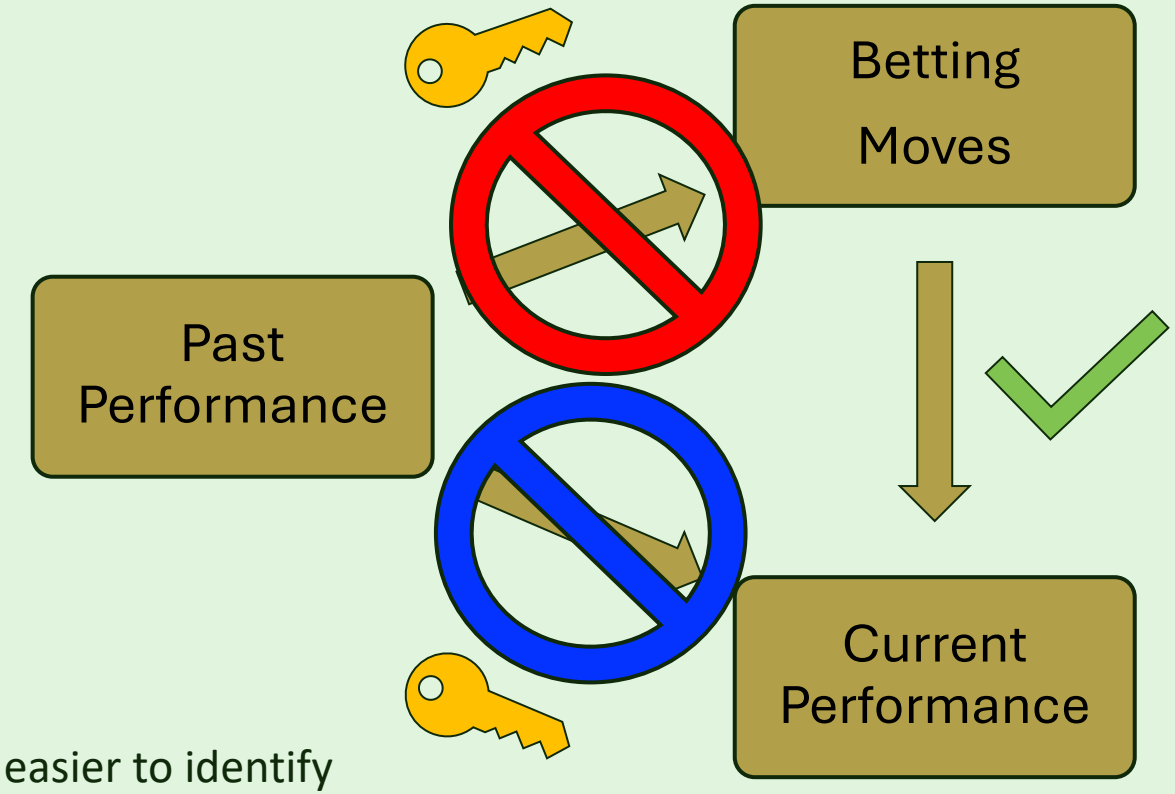
and/or

Past Performance and Current Performance

is broken there may be an integrity issue, particularly when

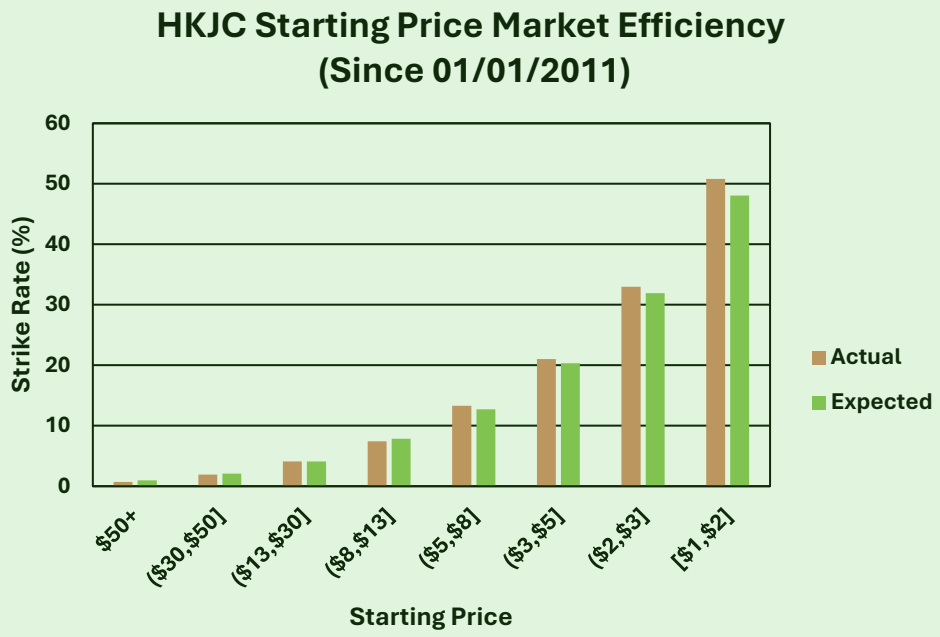
Betting Implied Current Performance

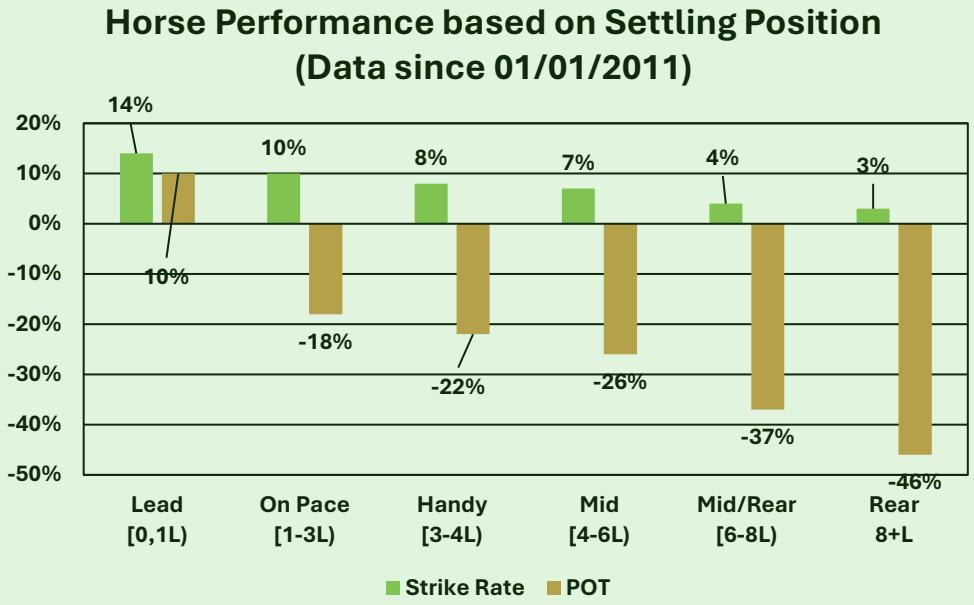
With an accurate expectation, anomalies become easier to identify



What Do We Want To Predict?

Starting Price and Settling Position are Highly-Correlated to Actual Performance





Machine Learning (ML) Basics

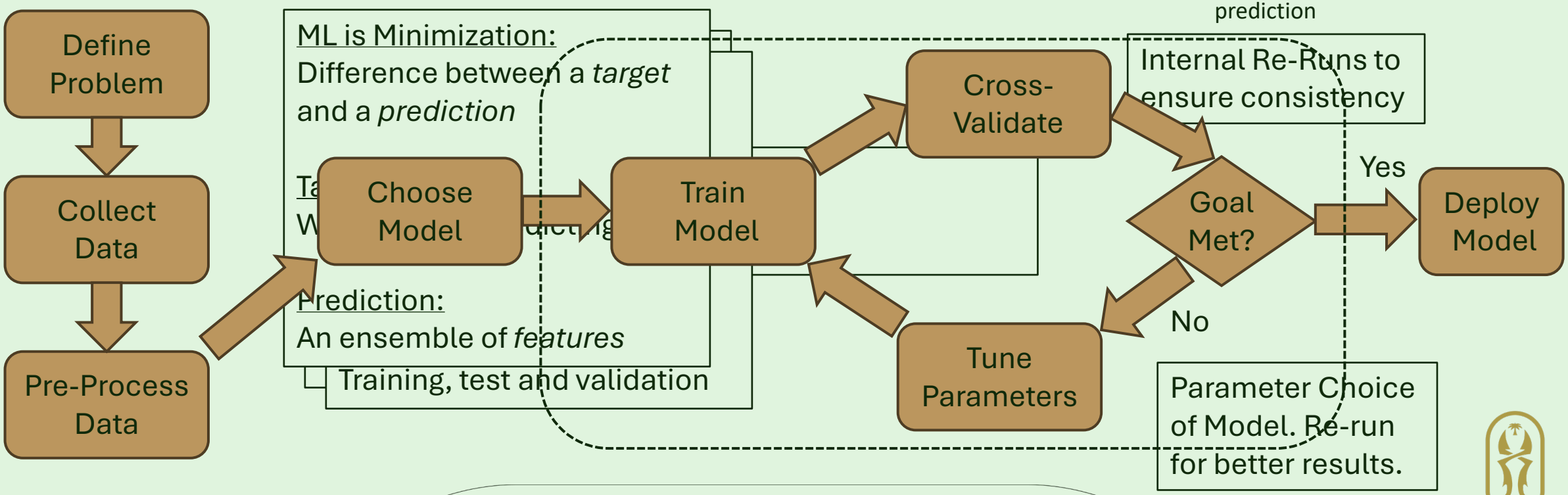
Racing is a data-rich sport, naturally lending toward the use of new and emerging technologies

A Simple 'ML' Model For Settling Position

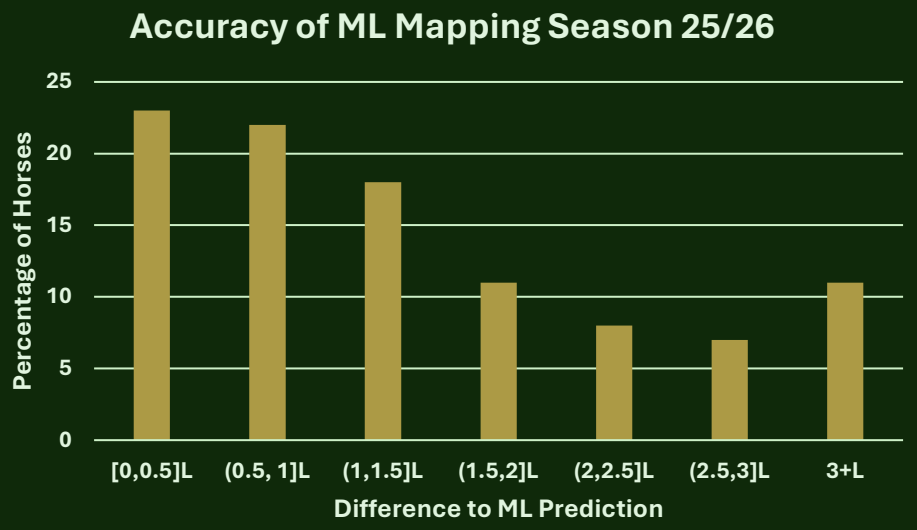
$$\text{MFL} = w_1 \times \text{Avg MFL}_{\text{Distance}} + w_2 \times \text{Avg MFL}_{\text{Class}}$$

Target — Weights — Features — Prediction

In this example the aim is to find the weights which minimize the difference between the target and the prediction



PREDICTING SETTLING POSITIONS OF HORSES



Settling Position is an easily interpretable variable highly correlated to actual performance and a driver of integrity questions

Use on race day and as a historical reference.

Integrity:

- Horse further forward to ensure pace
- Horse further back to decrease chances of winning

Poy and German Betting Case at Racing Victoria

5 ML Models are used in parallel:

- Cubist
- Deep Neural Network
- Generalised Linear
- Random Forest
- Gradient Boosted Tree

Each ML model uses 51 variables

Trained on over 100,000 horse runs (average error 1.49L)

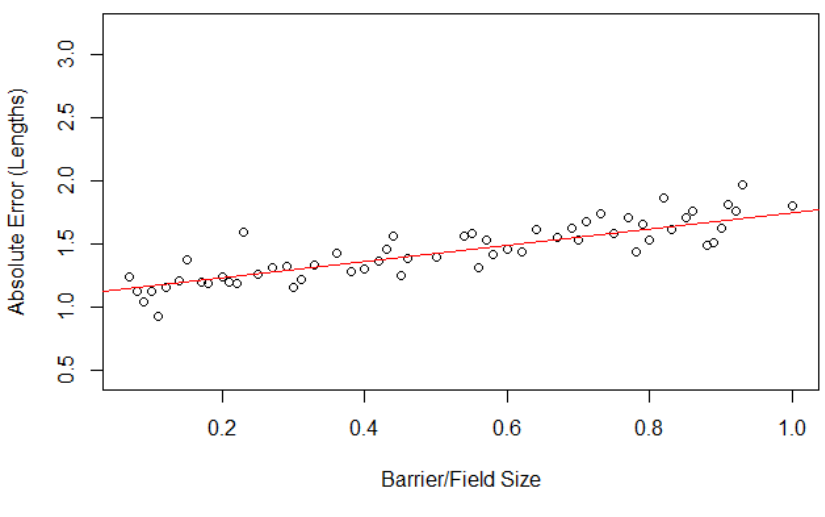
Tested on over 40,000 horse runs (average error 1.52L)

Average error in production this season is 1.57L.

Average deviation of a horse's own pattern in 1.2L

The ML Pricing Models – Results

Horses not on debut and not recorded as slow to begin have an average accuracy of 1.45L with 78% within 1L of prediction.



Key Features:

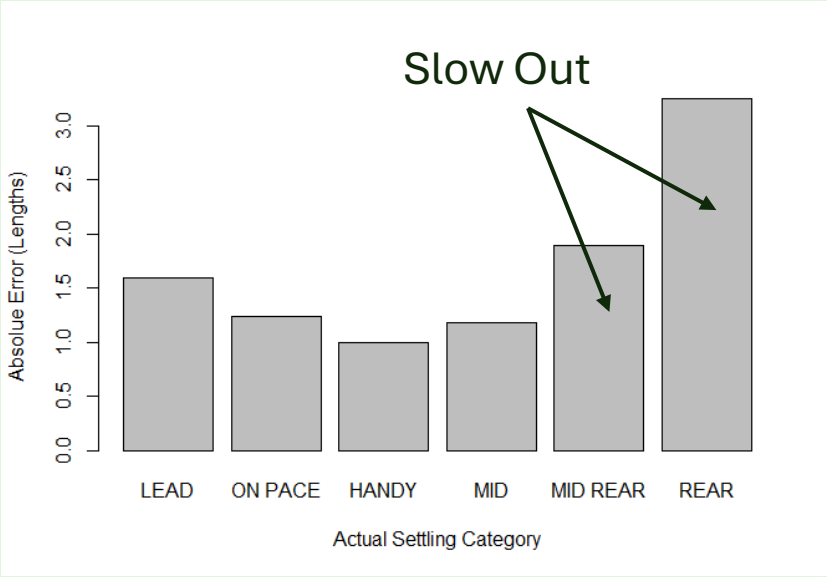
- *Previous MFL* – Past implies Future
- *Barrier* – Correlated to Track and Distance
- *Jockey* – Especially on this horse
- *Predicted pace*
- Which position has the horse *performed well* from previously.

Difficulties

- Lack of data (debutants)
- Wider Draws
- Slow out
- Extremal cases

Trainer and Jockey Patterns:

- Difference Between Least and Most Consistent Jockey is 1.2L
- Difference Between Least and Most Consist Trainer is 0.5L



PREDICTING STARTING PRICE AND TRUE ODDS OF WINNING

Discerning what the true probability of winning is, even post-race, is paramount to producing an accurate model.

Sha Tin	1200m	Good	AWT

Horse	Margin	True Odds
Bright Mortar	-0.1	$5.77
One Man Show	0.1	$5.86
So My Folks	0.5	$7.23
Vulcanus	0.75	$7.48
No Other Choice	0.75	$7.51
Swagger Bro	0.75	$7.62
Lean Master	4	$22.55
Cirrus Speed	6.25	$51.51
My Triumph	6.5	$52.79
Notthesillyone	7	$54.43
Goko Win	8	$82.52
Francis Maynell	9.75	$149

We follow the same recipe as before in producing two further models:

1) Predicting the Starting Price (Predicted Price – 121%)
2) Predicting the True Price (Assessed Price – 100%)

These models are far more complex and contain 329 features each.

When used for integrity, interpretability is key for feature-rich models.

Model:
Cubist ML model was used exclusively for both. Uses *decision trees* and *linear regression.*

Targets:
- Target the implied winning percentage of the horse (1/Odds)
- The *metric* used in training is the *Root Mean Squared Error (RMSE)*
- The Assessed Price Model Target is determined by *Re-Pricing*

The ML Pricing Models – Results

Starting prices have been within 3.6% of prediction whilst true odds have been within 5.2% and proven more efficient than Actual SP

Accuracy (MAE):
- Training Set:
  Predicted: 3.76%, Assessed: 5.10%
- Test Set:
  Predicted: 3.92%, Assessed: 5.14%
- This Season:
  Predicted: 3.63%, Assessed: 5.23%

Difficulties:
- Short-Priced Horses
- Lack of Data: Debutants/Griffin Races
- Horses coming off long injury-enforced breaks

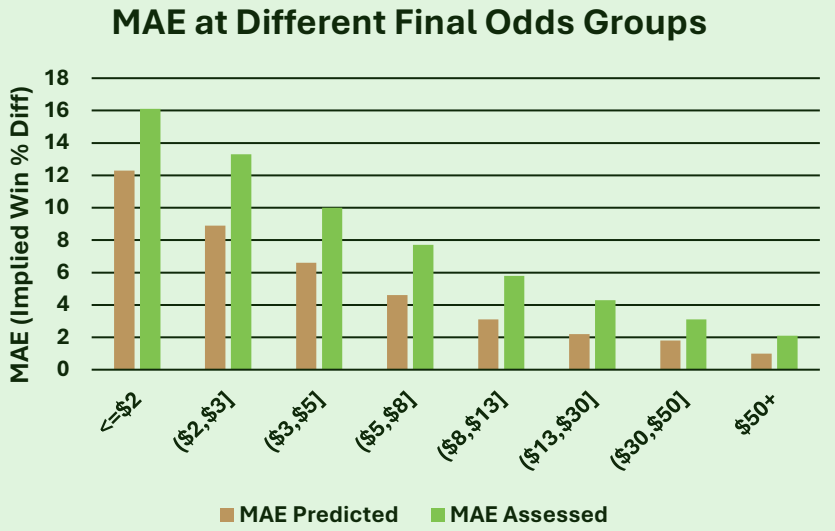
Favourite-Longshot Bias:
Apply an exponential distribution to the raw results: Wind longshots out and favourites in.

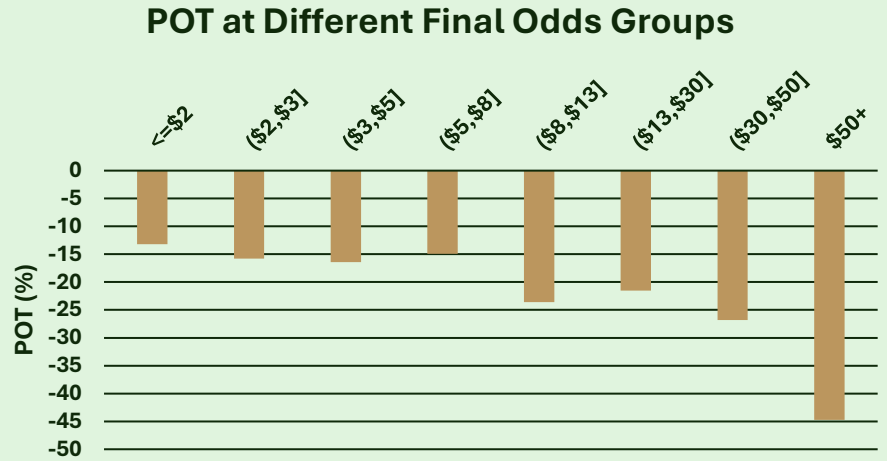
Re-Training:
Models should be re-trained 4-times per season (~220 races) to keep up with trends.

Late Price Movement Predictions (This Season):
- 42% of late market *firms* predicted correctly
- 68% of late market *drifts* predicted correctly

Over/Under Valued Horse Results (This Season):
- 14.1% of *Under-Valued* horses won vs market expectation of 11.9%
- 11.3% of *Over-Valued* horses won vs market expectation of 13.5%

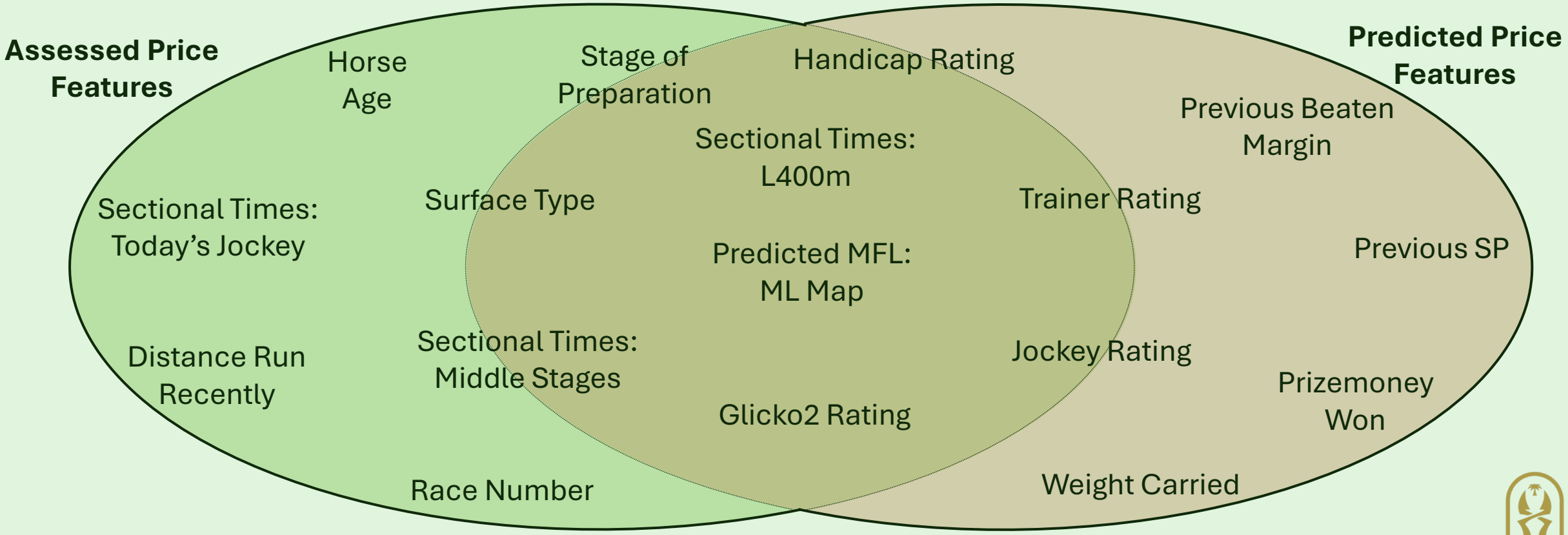
Modelling for Overvalued horses seem to be more robust. This is key for integrity when overlaying vs illegal market activity.





The ML Pricing Models – Feature Analysis

The ML algorithm has produced a Sectional Times-focused model for the Assessed Price and a Ratings-focused model for the Predicted Price



SYSTEM REQUIREMENTS

All run from a laptop on free software

Both models were trained and deployed using R by leveraging the Caret package which includes a host of different ML models ready to use out of the box.

Each model takes around 24 hours to train (approximately 150,000 training points) and 30 minutes to calculate predictions for a full race card (100 – 150 horses).

The main considerations involved in tailoring models to your own problem are:

- Variable Selection
- Hyperparameter tuning
- Data Collection, Retention and Cleaning
- Presentation (we use Shiny)

41ST ASIAN RACING CONFERENCE
RIYADH 2026

THANK YOU

ASIAN RACING FEDERATION

نادي سباقات الخيل
JOCKEY CLUB OF SAUDI ARABIA