Can Post-Race Commentary Improve the Predictive Power of Horse Racing Models?

Alex Nikic

Department of Statistical Science

STAT0034

Introduction

Source: British Horseracing Authority (BHA)

From a Gambling Perspective

- Horse racing is a widely popular sport for betting.
- Bettors like to wager their money on horse races in the hopes of winning their bets and turning a profit.
- Bookmakers need to set betting odds in such a way that they manage risk exposure and maintain profitability.
- Predictive models for horse races inform these decisions.

The Need for Novel Data Sources

- In order to gain an edge in the modern era of sports betting, one must look toward new sources of information that are not currently being used by competitors.
- The Efficient Market Hypothesis states that all publicly available information is incorporated into betting prices.
- In a fully efficient market, it is impossible to consistently make a profit from public information, as any advantage would have been arbitraged away by other participants in the market.
- If we can find data sources not reflected in betting prices, it may be possible to create a profitable model.

The Post-Race Card

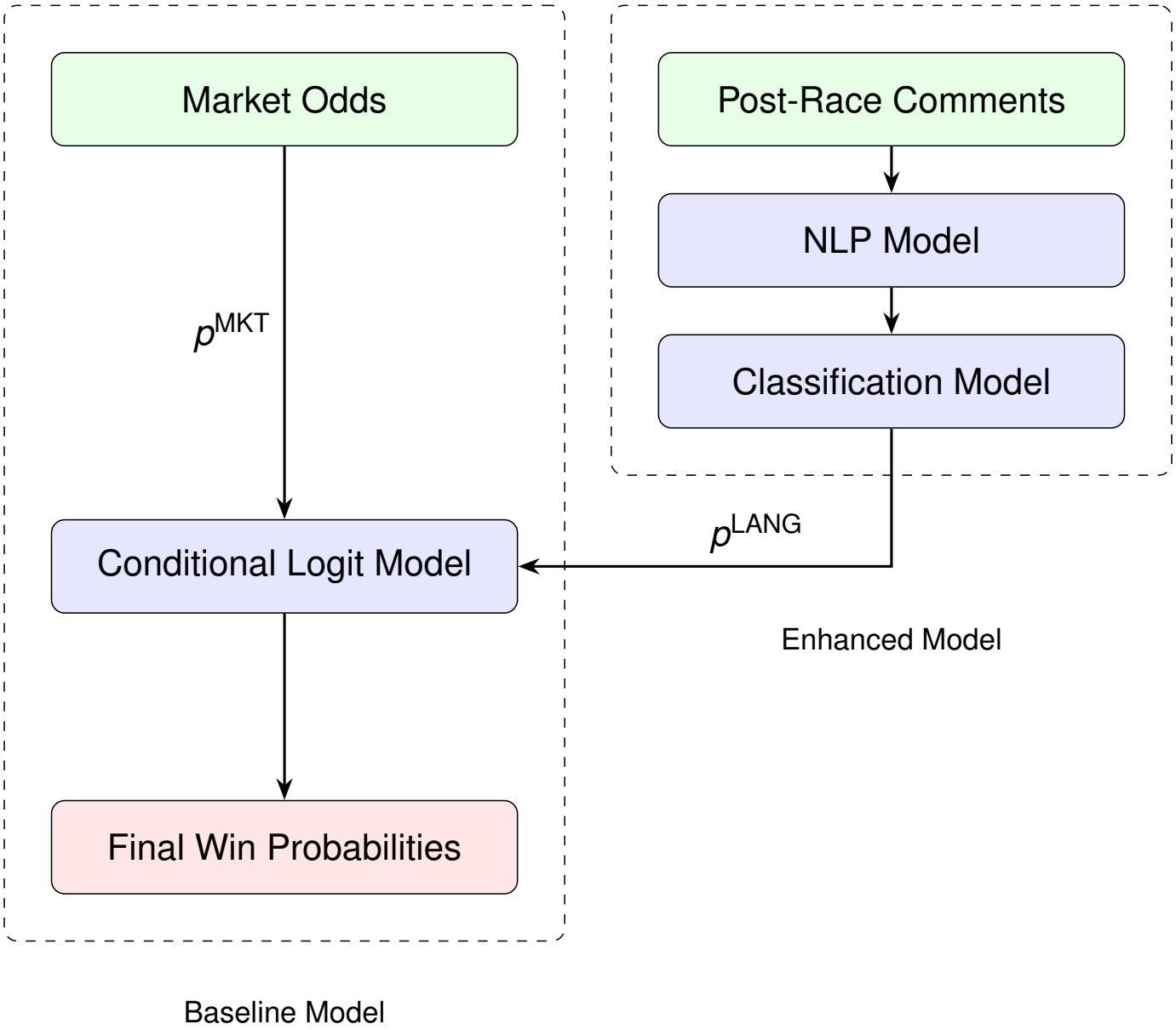
Pos	Horse	SP	Comment
1	Eium Mac (GB)	16/1	Tracked leaders on inner - chased leader halfway - slight lead over 2f out - ridden clear and edged right over 1f out - kept on (op 12/1)
2	General Tufto (GB)	8/1	Soon pushed along - ridden and outpaced after 3f - soon behind - headway and wide straight - ridden to chase leaders when rider dropped whip 2f out - chased winner and edged left entering final furlong - no impression (tchd 13/2)
3	Miami Gator (IRE)	6/1	Led - joined halfway - ridden along over 3f out - headed narrowly well over 2f out - ridden and edged left over 1f out - soon driven and kept on same pace (tchd 11/2)
4	Major Rowan (GB)	3/2	Dwelt and in rear - headway on outer halfway - chased leaders and ridden along over 3f out - driven over 2f out - kept on approaching final furlong - nearest finish (op 15/8 tchd 2/1)

Project Aims

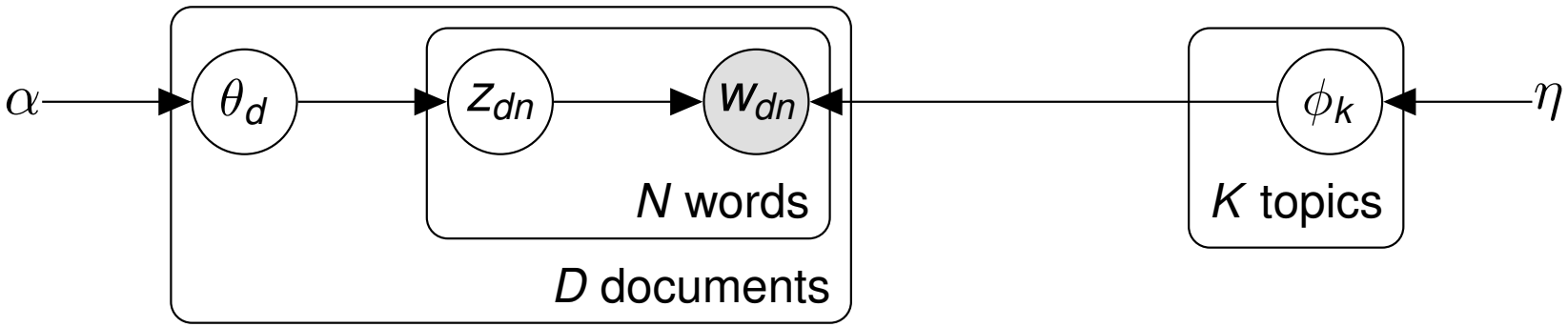
This project attempts to answer two keys questions.

- Is post-race commentary predictive of future horse races?
- If so, do post-race comments provide additional value beyond existing models and market odds?

The Pipeline



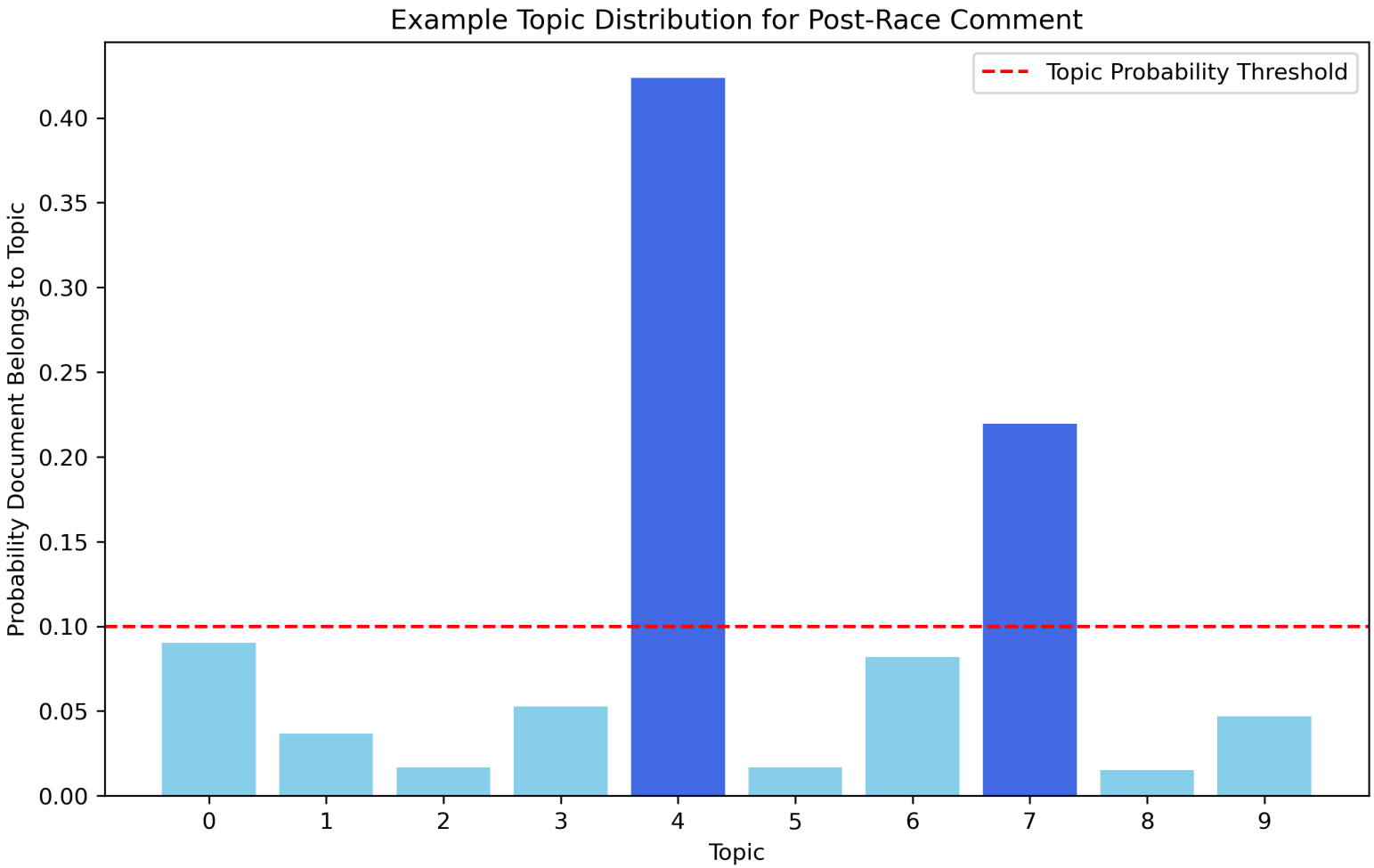
Latent Dirichlet Allocation



- For each topic $k = 1, \ldots, K$, sample $\phi_k \sim \mathit{Dir}(\eta)$.
- For each document $d = 1, \ldots, D$, sample $\theta_d \sim \mathit{Dir}(\alpha)$.
  - For each word $n = 1, \ldots, N_d$ in document d:
    - Choose a topic assignment $z_{dn} \sim \mathit{Categorial}(\theta_d)$.
    - Choose a word from the assigned topic $w_{dn} \sim \mathit{Categorial}(\phi_{z_{dn}})$.

Topic Assignment

Chased leaders - ridden along well over 2f out - soon weakened(op 11/1)

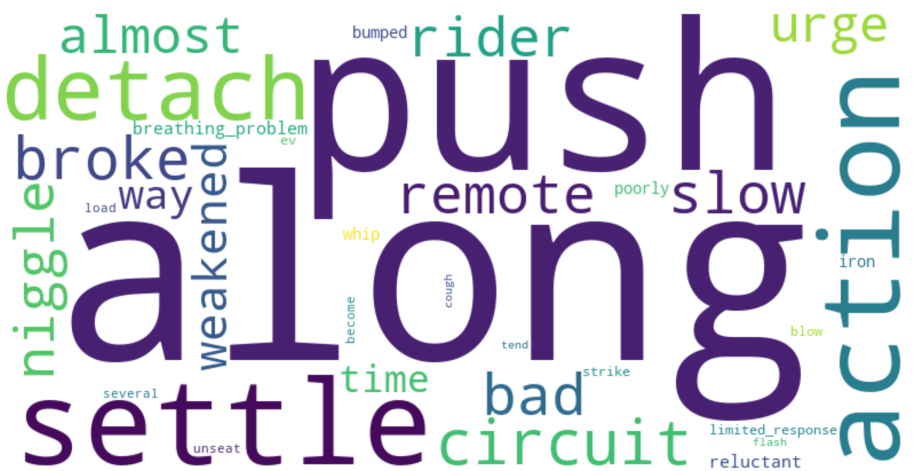


topic_features = [0,0,0,0,1,0,0,1,0,0]

Predictive Model

Topic 4

Topic 0



Naive-Bayes/XGBoost Classifier → p^{LANG}

Transformer Embeddings

Chased leaders - ridden along well over 2f out - soonweakened(op 11/1)

$$\Downarrow$$

$$\underbrace{[-1.300532, -0.219023, -0.688411, \ldots, -0.46522, 1.586039]}_{\text{128 dimensions}}$$

Richer text representation, but at the cost of interpretability.

The Conditional Logit Model

Let Y_{ij} be a binary random variable, where $Y_{ij} = 1$ represents horse j winning race i and $Y_{ij} = 0$ otherwise. Let $p_{ij}^{MKT}, p_{ij}^{LANG}$ be the probability horse j wins race i according to the market odds and language model respectively. Then according to the Conditional Logit Model,

$$\mathbb{P}(Y_{ij} = 1 | p_{ij}^{LANG}, p_{ij}^{MKT}) = \frac{\exp(\beta logit(p_{ij}^{LANG}) + \gamma logit(p_{ij}^{MKT}))}{\sum_{j=1}^{j=J_i} \exp(\beta logit(p_{ij}^{LANG}) + \gamma logit(p_{ij}^{MKT}))},$$

where $logit(p) := \log\left(\frac{p}{1-p}\right)$.
Here β is the effect of the language model, and γ is the effect of the market odds. To retrieve the baseline model, we simply set $\beta = 0$.

Betting Strategy

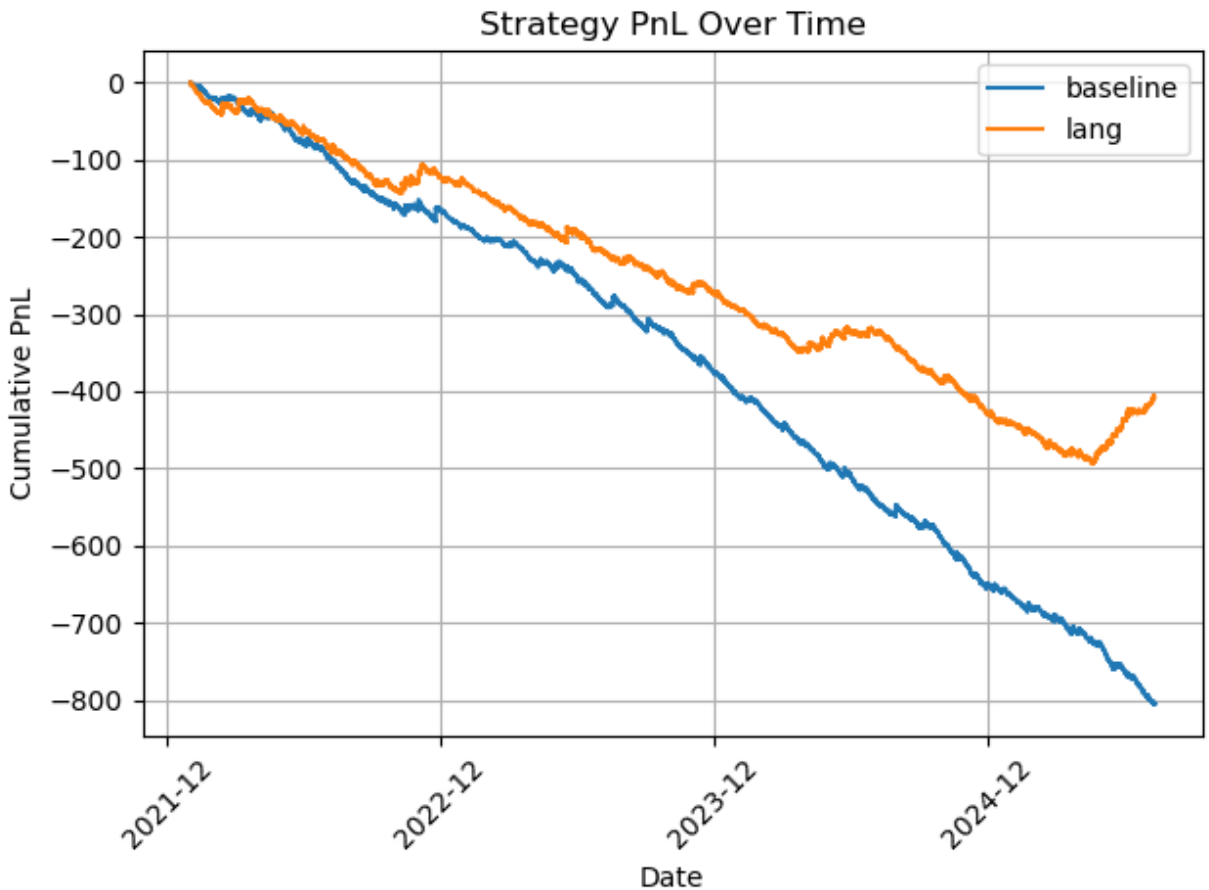
- Kelly's criterion is a formula that determines the optimal amount of money to wager on a bet in order to maximise growth and minimise risk of ruin.
- For each race, we will bet a fraction of £1 on the most likely horse to win according to our model.
- However if according to our model the bet does not meet a specified minimum Expected Value threshold, we do not bet anything at all.
- For race i and horse $j^* = \arg\max_{j=1,\ldots,I_j} p_{ij}$, the fraction to bet is

$$f_{ij^*} = \frac{o_{ij^*} p_{ij^*} - 1}{o_{ij^*} - 1},$$

where o_{ij} is the market odds (in decimal form) of horse j^*, and p_{ij^*} is our model's predicted win probability of horse j^*.

Language Model vs Random Guessing

Model	p-value	Mean Brier Diff
LDA-10-Topics + CalibratedNB	6.893×10^{-4}	0.3210



Market Odds vs Full Model

Model	p-val	Mean Brier Diff
Embeddings + CalibratedXGB	1.50×10^{-6}	2.09×10^{-5}

Discussion

- **Strengths**
  - The commentary can predict horse races better than random choice.
  - We can analyse the content of the commentary and extract qualitative information about horse performance.
  - The commentary can provide useful insights for bettors in an easily-digestible, language-based format.

- **Limitations**
  - However we conclude that the commentary is a redundant form of already available information.
  - In the future, it is likely that analysts will use GenAI to create the comments.