Introduction

I do not have a particular business problem for this project but I want to work with spatiotemporal data as I am very interested in working with such datasets in team sports analytics. The role of deep learning in sports analytics has been massively expanded by the technological breakthrough which allows organizations to efficiently (cost-wise) and accurately collect near continuous data on players' locations during a game. There are two primary end users for this sort of data: 1) Gambling Books & 2) Sport Organizations.

The rise in online gambling companies has obviously made sports betting more accessible and popular. But perhaps the bigger change from this deregulation trend is the rise in in-game or live betting. Judging by the amount of advertising showcasing these live betting opportunities, it is safe to say that it is a big part of gambling companies' business plans[i]. However, setting the odds for these bets can be very difficult for sports like soccer, basketball and hockey because of the dynamic/fluid pace of play. Having a model that can take a sequence of game states, each defined by the location of all players (& ball/puck), to predict future game states in the short-term can be funneled into less reactive models for predicting medium- and/or long-term outcomes.

For sport organizations, the potential opportunities are much more robust:

- Identifying general strategies that are successful
- Identifying/Predicting opponent strategies
- Optimal personnel for different approaches
- Optimal training and practice regiments
- Post-game analysis of player and team efficiencies/deficiencies

Specific to Horse Racing

My primary goal is to build ever-more-complex models to better predict outcomes of horse races:

- Simple regression model(s) based on horse features, track/race features and (ideally) past performance.
- Some form of time-adjusted clustering model(s) to start to identify different race strategies (like start slow-and steady w/ a big final push).
- Neural network using geospatial data; and, as an extension, a graph-based model to capture horse interactions.

There are some interesting features of horse racing that can make this project interesting for wider applications:

- The state of the track (weather included) should probably be considered when trying to model outcomes.
- The state of the horse & jockey (specifically weight) is likely to be an important feature.
- There are also unique track features (like distance) that can be included in the feature space.
- Multiple Modeling Targets & Approaches:
    - Trajectory/Path
    - Results (Time to Finish)
    - Results (Placement)
- Adjusting for different race length and even track size, there is a single "goal" for every sequence (the finish line). However, there is still secondary strategy considerations/goals that can be deployed/modeled.
- ***Interaction Between Horses (although not as direct as team sport examples)*:** **This is the most interesting aspect—especially with horse gambling have such distinct/specific betting options—with enough data, one should be able to model the race results by considering how each horse-jockey pair is most likely to react to the other participants movements.**
- Interesting extension with race descriptions. As you can see in the example below, Equibase provides a race description (that almost reads like live-analysis/commentary for television or radio) with very unique vocabulary. I am not sure of the exact application but it may be an interesting extension of this project's efforts to see if a generative AI model can be built based on horse's race path.

JC'S SHOOTING STAR away without apparent mishap, was unhurried three to four wide during the initial stages, saw the pilot resort to a light hand ride after departing the backstretch, got switched over to a drive coming up to the five-sixteenth marker, angled towards path four for entrance into the stretch and took off in earnest after the one to catch, whittled away at the deficit relentlessly, getting the job done in the shadow of the finish line. SOUNDS DELICIOUS established the top after being on the receiving end of an abbreviated brisk hand ride, cut the pace rated along afterwards while saving ground, with a pair of opponents in attendance parked out in paths two and three respectively, turned for home however with neither of them latched on, got set down right around the three-sixteenths marker, emerged a sixteenth down the road in possession of her largest advantage, had the winner close in a steadfast manner, came out a few paths from some let side stick work in deep stretch, got collared. CRIMSON FROST guided onto the rail approaching the conclusion of the backstretch, took it to the end of the bend, was ridden out in the direction of path six heading for the three-sixteenths marker, went on to finish a non-threatening third. FRIEND OF LIBERTY prompted SOUNDS DELICIOUS from path two for a half and folded. BOBBY'S SONG pursued the front runners three wide, altered course into path five in midlane, lacked a further response and backed away.

Cons:

- ***Will require lots of feature engineering/wrangling***. Even for the non-geospatial data.
- Only one result / race

- Enough Observations and are sequences long enough?
- Not a team sport & no puck/ball to track—a lot of the research I have seen makes use of team sports having a ball that participants must pay at least some attention. In fact, to the extent that teams are drawn away from the ball/puck is often a good way to measure individual's off-ball contribution to overall success.
- Complicated limitations for the individual path both in time and space.
- Live-betting angle does not apply

Datasets

The biggest problem for this project is finding enough data to successful build accurate models; especially, if it is necessary to consider individual historical data (for the horse, trainer or jockey). A lot of this data is behind paywalls, at best (as I found out with trying to build a hockey-related project).

Kaggle-NYRA Data

Kaggle-Kentucky Derby Data

Equibase Race Charts (Provided At Request for 2023)

Other Resources

Spatiotemporal Machine Learning

Generating Long-Term Trajectories Using Deep Hierarchical Networks

A Graph Attention Based-Approach for Trajectory Predictions in Multi-Agent Sports Games

Spatiotemporal, ML & Sports

A Spatial Perspective on Sports Analytics

Horseracing Analytics/Strategy

https://royalsocietypublishing.org/doi/10.1098/rsbl.2011.1120

https://www.biorxiv.org/content/10.1101/2020.06.11.145797v1

https://pubmed.ncbi.nlm.nih.gov/33264298/

Gambling

https://help.draftkings.com/hc/en-us/articles/4405230615699-What-is-a-live-bet-US https://www.risk.inc/blog/how-sportsbooks-make-money---a-look-inside-the-online-betting-business

New Timeline

- Saturday-Continue organizing code & collect/explore new equibase data; needs to be translated from thousands of XML files to pandas dataframes that, ideally, can also be combined with older data.
- Sunday—Data exploration coded; build and fit all basic models (regression, time series and clustering).
- Monday—Begin exploring neural network options using geopspatial data.
- Tuesday—Finalize Deep neural network model (s) and consider implications/conclusions.

[i] Per Google: recent quotes from DraftKings CEO