DS 6600 Project Check In 1

*Project Check In 1 for DS 6600: Data Engineering

Hieu Vu
Department of Computer Science
University of Virginia
Charlottesville, VA, United States of America
brr2tu@virginia.edu

***Index Terms*—Data Engineering**

I. Introduction

Sports analytics has evolved into a data-driven discipline that enables sports teams, analysts, and fans to make informed decisions based on current and historical data. One major branch of this field involves understanding the relationship between sports betting markets and actual game outcomes. Betting odds represent the collective expectations of bookmakers and bettors, offering an implicit prediction of future events. By examining how accurately these odds forecast real-world results, we can evaluate the predictive efficiency of betting markets and uncover patterns of overestimation or underestimation in different sports.

The purpose of this project is to collect, clean, and integrate sports betting and game outcome data from multiple sources like public APIs and open datasets to create a unified and analyzable data pipeline. This pipeline will support the evaluation of prediction accuracy across different leagues and time periods. Beyond enabling statistical analysis, this project emphasizes the engineering challenges associated with real-world data acquisition, transformation, and integration. Ultimately, the project aims to demonstrate how data engineering techniques can be applied to build reliable systems for continuous data collection and analysis within the sports analytics domain.

II. Background

Sports betting markets are often studied as real-world laboratories for prediction and market efficiency. Classic work argues that bookmakers set prices to maximize profit while responding to bettor biases rather than to clear markets, which has implications for how well odds reflect true event probabilities [1], [2]. A long-standing empirical regularity is the *favorite–longshot bias*: longshots tend to be overpriced and favorites underpriced, producing odds that are systematically miscalibrated relative to realized outcomes [3], [4]. These findings motivate evaluating the degree to which pre-game odds provide unbiased and well-calibrated predictions across leagues and seasons.

From a data-engineering perspective, prior public work demonstrates the feasibility of building pipelines that join betting information with historical game results at scale. Open repositories from FiveThirtyEight provide league-wide, longitudinal results and ratings (e.g., NBA and NFL Elo datasets), which are widely reused for benchmarking predictive systems and for reproducible research [5]–[7]. Complementary, sportsbook-aggregated odds feeds expose machine-readable pre-game markets (moneyline, spreads, totals) through REST APIs suitable for automated ingestion and schema design [8]–[10].

To support cross-league integration of outcomes and context, several public sports APIs provide structured endpoints for schedules, teams, and final scores. For soccer, `football-data.org` offers match fixtures and results across major competitions [11], [12]. For basketball, `balldontlie` exposes historical and current NBA games, teams, and box scores [13], [14]. For college football, the `CollegeFootballData` API provides games, play-by-play, and advanced statistics and has a mature open-source ecosystem (e.g., SportsDataverse wrappers) that facilitates robust ETL and validation [15]–[17]. These sources collectively enable a pipeline that ingests pre-game odds, converts these odds to implied probabilities, joins to realized outcomes, and evaluates calibration and accuracy over time and by league.

In this project, we build on current existing literature, open infrastructure, and accessible data. Our data engineering goal is to design a reproducible pipeline that continuously acquires pre-game betting odds, integrates them with authoritative game outcomes, and produces analysis-ready tables for evaluating prediction accuracy and calibration across the NBA, NFL/college football, and international soccer.

III. Data

This project will draw from several public sports and betting data sources to build an integrated data engineering pipeline that links pre-game odds to actual game outcomes. The primary source of betting information will be *The Odds API* [8], [9]. This RESTful service provides JSON-formatted pre-game and live betting odds from major bookmakers across multiple sports and regions. Data can be retrieved through authenticated HTTP requests to endpoints such as `/v4/sports` and `/v4/odds`, which return metadata about leagues and associated betting markets. According to the provider's terms of service, free-tier access is permitted for research and non-commercial purposes, but redistribution or large-scale com-

mercial use requires prior authorization. The API's documentation specifies rate limits and encourages responsible use to avoid service termination. These odds datasets will serve as the project's measure of pre-game predictive expectations, later converted into implied probabilities for model calibration analyses.

To obtain verified game outcomes for comparison, this project will rely on publicly available sports result APIs. For international soccer, the *football-data.org* API provides detailed JSON endpoints containing fixtures, match results, and competition standings [11], [12]. The service allows users to query leagues such as the English Premier League, UEFA Champions League, and FIFA World Cup. While the API offers a free tier suitable for educational and non-commercial research, the site requests that commercial users contact the provider for explicit permission. The documentation and pricing pages clarify that its "Free Forever" plan covers a limited number of competitions with fair use constraints.

For basketball, the project will use the *balldontlie* API, which exposes historical and current NBA statistics, including game results, player box scores, and team-level information [13], [14]. Access is provided through a RESTful JSON interface, and the data can be retrieved programmatically via endpoints such as `/games` or `/teams`. The API is freely available for academic or hobbyist use, and its associated GitHub repository indicates that the client code is MIT-licensed. While no explicit data license is listed, public documentation states that it is "free to use for sports-related applications," which suggests minimal restrictions for educational purposes. These data will provide the actual NBA outcomes against which pre-game betting probabilities can be evaluated.

In addition to live APIs, this project can potentially incorporate historical data from *FiveThirtyEight*'s open datasets [5]–[7]. FiveThirtyEight also has downloadable CSV files containing league-level statistics, team Elo ratings, and historical game results for major U.S. sports leagues. These datasets are hosted on GitHub and Kaggle under Creative Commons Attribution (CC BY 4.0) licenses, which permit reuse and redistribution with attribution. Because these files are accessible in tabular format, they provide a convenient mechanism for cross-validation, archival storage, and reproducibility testing in this project's pipeline.

Collectively, these datasets enable the design of a unified system that collects betting odds and game outcomes from multiple APIs, transforms them into standardized relational schemas, and evaluates the predictive calibration of betting markets across sports. All of the chosen sources provide either open licenses or explicit free-use tiers, making them appropriate for academic analysis under fair-use and attribution guidelines. Prior to large-scale ingestion, all data will be verified for license compliance, and the final data pipeline will document the provenance and terms of each source in its metadata.

IV. Potential Analyses

The integrated dataset will include key variables such as league, event date, team names, bookmaker odds, implied probabilities, final scores, and match results. Additional fields from FiveThirtyEight, such as Elo ratings and model win probabilities, will serve as benchmarks for comparison. These variables enable both descriptive and comparative analyses of betting accuracy across sports.

The planned analyses will primarily focus on evaluating how well pre-game odds predict actual outcomes. This will involve calculating and visualizing calibration curves that compare implied probabilities to observed win rates. Additional analyses will include cross-tabulations of favorite versus underdog outcomes, histograms of implied probabilities across leagues, and correlation tables between bookmaker odds, Elo ratings, and game results. A public-facing dashboard will present these insights interactively, allowing users to filter by league, season, or bookmaker and explore trends in betting market accuracy over time.

V. Challenges

One of the main challenges in this project will be ensuring consistent data integration across multiple APIs and formats. Team names, league identifiers, and date formats often differ between data sources, which can lead to mismatches during joins. To address this, the project will include a data-cleaning step that standardizes identifiers and timestamps before merging. Another challenge involves handling incomplete or missing odds data, especially for smaller leagues or historical matches. If these gaps significantly affect analysis quality, an alternative approach will be to focus on a single league such as the NBA, where both odds and outcomes are more consistently available.

API rate limits and access restrictions may also pose obstacles during large-scale data collection. If API limits certain stats or metrics that are planned to be used, then I plan on using only one league's data instead to lower the amount of API usage. Finally, aligning real-time odds with finalized game results will require careful timestamp management. If real-time integration proves unreliable, the fallback plan will use historical CSV data from FiveThirtyEight to demonstrate the same pipeline design on archived datasets.

VI. Gen AI Statement

Generative AI tools, including ChatGPT, were used to assist in refining the written sections of this project check-in, specifically the introduction, background, and methodology descriptions. All ideas, project design decisions, and data source selections were developed and verified by the author. The final submission has been reviewed for accuracy, originality, and academic integrity.

References

[1] S. D. Levitt, "Why are gambling markets organised so differently from financial markets?" *The Economic Journal*, vol. 114, no. 495, pp. 223–246, 2004. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0297.2004.00207.x

[2] ——, "Why are gambling markets organised so differently from financial markets?" University of Chicago Price Theory, Tech. Rep., 2004, working paper. [Online]. Available: https://pricetheory.uchicago.edu/levitt/Papers/LevittWhyAreGamblingMarkets2004.pdf
[3] E. Snowberg and J. Wolfers, "Explaining the favorite-longshot bias: Is it risk-love or misperceptions?" IZA Institute of Labor Economics, Tech. Rep. DP4884, 2010. [Online]. Available: https://www.iza.org/publications/dp/4884/explaining-the-favorite-longshot-bias-is-it-risk-love-or-misperceptions
[4] M. Ottaviani and P. Sørensen, "The favorite-longshot bias: An overview of the main explanations," NBER, Tech. Rep. w15923, 2010. [Online]. Available: https://www.nber.org/system/files/working_papers/w15923/w15923.pdf
[5] FiveThirtyEight, "Fivethirtyeight data repository," https://data.fivethirtyeight.com/, 2017, cC-BY-4.0. Accessed Oct. 23, 2025.
[6] ——, "Nba elo dataset," https://www.kaggle.com/datasets/fivethirtyeight/fivethirtyeight-nba-elo-dataset, 2019, accessed Oct. 23, 2025.
[7] ——, "Nfl historical scores and elo (forecasting game data)," https://github.com/fivethirtyeight/nfl-elo-game, 2018, accessed Oct. 23, 2025.
[8] The Odds API, "Sports odds api," https://the-odds-api.com/, 2025, jSON feed for bookmaker odds. Accessed Oct. 23, 2025.
[9] ——, "Odds api documentation (v4)," https://the-odds-api.com/liveapi/guides/v4/, 2025, endpoints for moneyline, spreads, totals. Accessed Oct. 23, 2025.
[10] Odds API IO, "Odds api: Real-time sports betting odds (250+ bookmakers)," https://odds-api.io/, 2025, accessed Oct. 23, 2025.
[11] football-data.org, "Api reference," https://www.football-data.org/documentation/api, 2025, rEST endpoints for fixtures and results. Accessed Oct. 23, 2025.
[12] ——, "Api quickstart (v4)," https://www.football-data.org/documentation/quickstart, 2022, accessed Oct. 23, 2025.
[13] BALLDONTLIE, "Balldontlie sports apis," https://balldontlie.io/, 2025, multi-league coverage. Accessed Oct. 23, 2025.
[14] ——, "Nba api (documentation)," https://nba.balldontlie.io/, 2025, historical and current NBA data. Accessed Oct. 23, 2025.
[15] CollegeFootballData.com, "College football data api," https://api.collegefootballdata.com/, 2025, games, play-by-play, advanced stats. Accessed Oct. 23, 2025.
[16] ——, "College football data api: Documentation," https://www.postman.com/api-evangelist/college-football-data/documentation/el6ukhf/college-football-data-api, 2025, postman docs. Accessed Oct. 23, 2025.
[17] S. Gilani and SportsDataverse, "cfbfastr: College football data in r (sportsdataverse)," https://cfbfastr.sportsdataverse.org/reference/index.html, 2025, open-source wrappers around CollegeFootballData API. Accessed Oct. 23, 2025.