P10: Team Performance and Formation Project Proposal

CSE 575: Statistical Machine Learning

Archies Bhandary

Sreekar Manuka

Zaina Mushtaq

Katelyn Vanderwolde

Parva Sheta

Mingqi Wang

September 30, 2025

Team Members and Their Roles:
All team members will be collaborating on all steps of this project. During model development, each member may work on a separate model.

Katelyn Vanderwolde:
Build in-team synergy features, such as, how lineups and pairs perform together, and on-off court impact. Build Team A vs Team B matchup features by position and likely lineups to improve Team A's chances of winning. Overall, this should contribute to a more accurate lineup recommendation.

Zaina Mushtaq:
Build models like logistic regression and gradient boosting, focus on model training, validation, and testing splits that respect seasons and time. Run hyperparameter tuning and keep a clear leaderboard of results to find the best player attributes.

Parva Vasudevbhai Sheta:
Focus on model training and evaluation. Calibrate the model's win probabilities well and check with reliability plots and track accuracy with clear metrics like margin errors. Implement features for role coverage, bench depth, and rotation consistency.

Archies Bhandary:
Focus on chemistry metrics (pair synergy, lineup net rating), player replacement module (graph kernels + pruning), and recency features (Optimal Time Window), with some work on constraint-based optimization, and document results into the final report.

Mingqi Wang:
Extract game, player, lineup, injury, and schedule data and create a consistent, clean schema. Handle missing or duplicate data and automate preprocessing. Add data checks to prevent leaks from future games and maintain a data dictionary.

Sreekar Manuka:
Build reusable feature code like the paces, possessions, opponent-adjusted ratings, rolling, decay recent form, clutch, and time split stats. Set up experiment tracking so results are reproducible. Create a prediction script/API that takes upcoming games and returns model predictions.

Description of the Problem:
The primary challenge in this project is understanding and predicting what makes a successful team. For the scope of this project, we will be focusing on NBA (National Basketball Association) teams as it represents a high-level competitive sport with abundant publicly available data, allowing us to analyze complex team dynamics and chemistry. In basketball, success is not only determined by each player's talent alone, but also the combination of skills, chemistry, tournament schedule and recent performance trends that each individual brings to the court. This problem involves identifying which combinations of player metrics, such as scoring, assists, rebounding, and defense, are most likely linked to winning outcomes. It also requires building a predictive model that estimates the likelihood of a team winning a game by

considering both its own performance statistics and how those compare against its opponent. Finally, this challenge extends to also recommending strong team formations or replacements that maximize skill coverage while also preserving player chemistry.

The problem can be divided into two parts. First, we aim to predict the outcome of NBA games using both player and team statistics. This will be treated as a classification task, where the model learns patterns from features such as points, rebounds, assists, net rating and recent performance trends like winning streaks or lineup changes. Second, is recommending effective team formations. This is more of an optimization task, where the goal is to suggest balanced lineups that not only look good statistically but also make sense in real basketball terms. For example, a strong lineup needs the right mix of scorers, defenders while also respecting natural positions and salary limitations. Similarly, replacing injured players is not just finding players with similar stats, but someone who can blend into the existing team with minimal disruption.

Overall, the problem our project team aims to address is the design of a framework that combines skill metrics, chemistry measures, and contextual trends to predict NBA team performance more accurately and suggest balanced high-synergy team formations that reflect the realities of NBA basketball.

Preliminary Plan (Milestones):

Milestone 1: Background Research

- Target Date: 10/02/25
- Find and review relevant research papers and articles regarding team performance, play metrics, and predictive modeling in sports

Milestone 2: Finding Datasets

- Target Date: 10/06/25
- Find reliable NBA datasets - game scores, advanced stats, etc.
- Datasets should include both individual player performance and team-level performance
- Document data sources and confirm licenses/availability for academic use.

Milestone 3: Data Cleaning and Preprocessing

- Target Date: 10/10/25
- Explore initial data and find features like: total amount, loss rate, outliers, and duplicates.
- Try feature engineering like Age = curYr - birYr to improve data quality.
- Standardize and normalize data to address the long-tail distribution problem and data domination issue, if necessary.

Milestone 4: Model Development and Training

- Target Date: 10/31/25
- This will be the longest part of this project; therefore, we need to allocate enough time.
- Implement logistic regression, comparative features, gradient boosting, outcome prediction, and optimization outside of skillset.

Milestone 5: Model Validation & Testing

- Target Date: 11/10/25
- Evaluate the model using historical game data (split into training, validation, and testing sets during the model development stage)
- Validation will be done to ensure there is no overfitting or underfitting
- Testing will be done after the model is finalized and after validation on data completely new to the model in order to obtain unbiased results

Milestone 6: Develop Frontend and Gather Visual Results

- Target Date: 11/17/25
- Create a simple dashboard or interface to demonstrate the model results.
- Show predicted win probabilities, team chemistry scores, and recommended replacements.
- Allow basic inputs (choose two teams, select lineup, test replacement).
- Visualize synergy using graphs (nodes = players, edges = chemistry links).
- Prepare visual outputs (charts/graphs) for the project presentation.

Milestone 7: Project Presentation

- Target Date: 11/19/25
- Presentation Date: TBD
- Create slides to go along with our presentation
- Outline the problem description and the work completed thus far.
- Practice team presentation during meetings.

Milestone 8: Final Project Report

- Due Date: 12/08/25
- Target Date: 12/01/25 (to leave 1 week for edits and finalizing)
- Write the report: introduction, problem description, methodology, results, conclusion/future work, and references.
- For results, include graphs and stats.

References:

L. Li, H. Tong, N. Cao, K. Ehrlich, Y.-R. Lin, and N. Buchler, "Replacing the irreplaceable: Fast algorithms for team member recommendation," in *Proc. 24th Int. World Wide Web Conf. (WWW '15)*, Florence, Italy, May 18–22, 2015, pp. 636–646, doi: 10.1145/2736277.2741132.

S. Kusmakar, S. Shelyag, Y. Zhu, D. Dwyer, P. Gastin, and M. Angelova, "Machine learning enabled team performance analysis in the dynamical environment of soccer," *IEEE Access*, vol. 8, pp. 90266–90284, May 2020, doi: 10.1109/ACCESS.2020.2992025

S. Siddique, L. Li, and Y. Wang, "Finding optimal teams: An analysis of NBA statistics and constraints," in *Proc. IEEE 11th Int. Conf. Big Data Computing Service and Machine Learning Applications (BigDataService)*, Newark, CA, USA, Apr. 2025, pp. 20–28, doi: 10.1109/BigDataService65758.2025.00010

T. Horvat, J. Job, R. Logozar, and Č. Livada, "A data-driven machine learning algorithm for predicting the outcomes of NBA games," *Symmetry*, vol. 15, no. 4, p. 798, Mar. 2023, doi: 10.3390/sym15040798

Theodoros Lappas, Kun Liu, Evimaria Terzi, "Finding a Team of Experts in Social Networks," KDD'09, June 28–July 1, 2009, Paris, France, doi: 10.1145/1557019.1557074