# NBA Team Performance Prediction using PML

##### **Mikhail Karimov\***

\* FIT CTU

This project targets it to develop a machine learning model to predict the outcome of sport events, especially NBA match score, by analyzing historical data of two competing teams. The system will use features such as player statistics, team performance trends, and game conditions to forecast the final team score

Collaborative filtering, Content-based filtering, Hybrid, Recommendation system.

1. Introduction

T

he prediction of sports outcomes, such as NBA game results,

has garnered significant attention due to its applications in analytics, betting, and strategic planning. This paper explores the development of a machine learning model leveraging recommendation system techniques to predict NBA match outcomes. The model is trained, tested, and evaluated using historical NBA datasets covering the 2010 to 2019 seasons.

The task of predicting NBA game results was approached by decomposing it into two subproblems: predicting the final score for each team independently. This decomposition allows for greater flexibility and granularity in analyzing team performance. To achieve this, the datasets were preprocessed separately for home and away team score predictions, ensuring a consistent methodology across both tasks.

Three distinct recommendation techniques were employed to address the problem:

* Collaborative Filtering, which utilizes patterns in historical data to make predictions based on similarities between teams and outcomes.
* Content-Based Filtering, which relies on team-specific features to predict scores.
* Hybrid Filtering, which combines elements of both collaborative and content-based methods to enhance predictive accuracy.

This work aims to demonstrate the efficacy of these techniques in the domain of sports analytics while providing insights into the relative performance of each approach.

1. Related works

Existing paper *Survey on Collaborative Filtering, Content-based*

*Filtering and Hybrid Recommendation System* reviews recommendation system techniques, focusing on collaborative, content-based, and hybrid filtering. It highlights several limitations, including:

* Data Sparsity: Recommender systems often face sparse user-item interaction data, leading to challenges in accurately predicting preferences.
* Scalability Issues: Traditional algorithms struggle with large-scale datasets.
* Diversity Concerns: Some systems favor popular items, reducing recommendation diversity.
* Cold Start Problem: New users or items lack sufficient data for effective recommendations.
* Vulnerability to Attacks: Systems are susceptible to manipulative user activity.

Other works also summarize typical outperformance of hybrid algorithm over CF and CBF. But while most of them is used to propose film to watch, music to listen or some other value, that can be only relevant or irrelevant to current user, this study is focused on something that can be objectively measured in discrete values.

1. Method

The proposed methodology integrates Collaborative Filtering (CF) and Content-Based Filtering (CBF) into a hybrid framework to predict NBA game scores. This section outlines the data collection process, chosen features, and the implementation of CF, CBF, and the hybrid models.

The dataset comprises historical NBA game data, player statistics, and team-level features, acquired from publicly available sports data repositories and APIs.

* Features Extracted:
  + Historical Interactions (CF): Team matchup results (game scores) over multiple seasons 2010-2018.
  + Team-Level Features (CBF): Offensive and defensive ratings, average points scored, and team rankings from season 2018.
  + Player Features (CBF): Player statistics (e.g., points per game, assists, rebounds) from season 2018.
  + Game Context: Home/away status.
* Preprocessing:
  + Most features were selected automatically (CBF). Were implemented weighs for history result to increase dependency on recent matches in match history (CF): rows with recent matches were duplicated several times by formula *season year - 2010*.
* Collaborative Filtering (CF)
  + A Matrix Factorization model was implemented to predict game scores based solely on the historical interactions between teams.
  + Algorithm: Singular Value Decomposition (SVD) was used to reconstruct missing values in matchup history, representing historical relationships between teams.
  + Output: Predicted game scores based on learned latent relationships from the interactions matrix.
* Content-Based Filtering (CBF)
  + Approach: A supervised machine learning model was deployed, leveraging player and team-specific features to predict game outcomes.
  + Model: A Random Forest Regressor was selected for its ability to model non-linear relationships and handle mixed data types.
  + Feature Engineering: Player, team, and contextual features were aggregated for each game, translated into input vectors for prediction.
  + Output: Predicted home team score based on the specific attributes of teams and players in each matchup.
* Hybrid Model (Dynamic weights)
  + Methodology: Predictions from the CF and CBF models were combined using a Weighted Averaging strategy, wherein weights were dynamically adjusted for each game.
  + Higher CF Weight: Games with significant historical interaction data.
  + Higher CBF Weight: Cold-start event, games without significant historical data:
* Hybrid Model (Selectively Use Models)
  + Methodology: Selection between predictions from the CF and CBF models based on amount of historical data.
  + CF prediction is preferred if there is historical data from at least constant limit of matchups.
* Hybrid Model (Model Stacking)
  + Meta-Learner: A simple Linear Regression model was trained as a meta-learner to optimize weights for CF and CBF predictions, using a separate validation dataset.
* Hybrid Model (Mean Prediction)
  + Simple CF and CBF results combination model, that return mean from both predictions.
* Evaluation
  + To assess model performance, predictions were evaluated using the Relative Root Mean Squared Error (Relative RMSE) as the primary metric:

* Performance comparisons were made across:
  + Collaborative Filtering (CF): RMSE attributable to historical matchups.
  + Content-Based Filtering (CBF): RMSE explained by feature-driven predictions.
  + Hybrid Model: RMSE reflecting the integration of CF and CBF.

1. Experiment

Historical NBA game data obtained from Kaggle public datasets. The dataset included records of several thousand games spanning through 1947-2024, from which were used games that occurrent in 2010-2019 and team statistics from 2018.

Features for CF: Historical game scores between teams.

Features for CBF: Team and player performance statistics, including offensive rating, defensive rating, average points, rest days, injuries, and game conditions (e.g., home/away), from which were automatically selected subset of data using SelectKBest.

Features For Hybrid Model: A combination of the above features.

The dataset was divided into training (80%) and testing (20%) sets. To benchmark the hybrid model's effectiveness, we used a baseline CF model that predicted scores based on the league-wide history of team scores. Relative RMSE (Root Mean Squared Error) was chosen for its ability to compare the prediction accuracy of each method relative to the baseline model.

All experiments were implemented in Python (v3.12.2) using libraries such as NumPy, pandas, scikit-learn, and matplotlib.

The experiments were performed on a system equipped with AMD Ryzen 7 7700 CPU and 32 GB of RAM.

A matrix factorization model using Singular Value Decomposition (SVD) was implemented. For Content-Based Filtering a Random Forest Regressor was chosen, trained on team stats and game-specific features. Feature importance was also evaluated to understand the significant contributors to predictions.

For model stacking version of Hybrid Model weights were found via a meta-learner (linear regression) trained on the validation set.

Scenarios with high historical match data favored CF, while injury-driven, dynamic contexts weighed CBF more heavily.

The composition of NBA teams can change significantly over just a few seasons. With a minimum age requirement of 18 for players to join the league and the majority of veteran players retiring between the ages of 32 and 36, the lifespan of an NBA career is often relatively short. Consequently, teams frequently undergo roster changes, driven by trades, retirements, and the influx of younger talent. These dynamics introduce variability that complicates predictive modeling, especially for algorithms heavily reliant on historical data.

Moreover, teams typically face each other only two to three times during a season. This limited interaction results in a slower accumulation of historical matchup data, further exacerbating the challenge for Collaborative Filtering (CF) models. Additionally, off-season changes—such as roster overhauls and shifts in coaching strategies—can render historical data less reliable, significantly impacting the accuracy of CF and CF-centric hybrid models.

Collaborative Filtering (CF) relies heavily on historical interactions between teams to make predictions. This approach is effective when there is abundant historical data, such as frequent matchups or stable team compositions over time. However, in the context of NBA game score prediction, CF faces several challenges:

Limited Interaction Frequency: NBA teams typically play against each other only 2–3 times per season. This results in sparse data for most team pairs, limiting the ability of CF models to accurately learn patterns.

Dynamic Team Compositions: NBA teams frequently undergo significant changes in their rosters due to trades, injuries, retirements, and new player drafts. These changes mean that the relationships learned from historical data may not reflect the current state of the teams, making CF less effective.

Impact of Off-Season Changes: Between seasons, teams can experience shifts in coaching strategies, player form, and overall dynamics. CF struggles to adapt to such changes since it relies on static historical data rather than dynamically updated features.

Cold-Start Problem: When teams or players have limited historical data, CF performs poorly because it lacks sufficient context for accurate predictions. This issue is especially pronounced for newer teams or teams with recently overhauled rosters.

In contrast, Content-Based Filtering (CBF) leverages detailed features such as player statistics, team performance metrics, and game context (e.g., home/away status). These features are updated regularly and reflect the current state of teams, making CBF better suited to handle the dynamic and evolving nature of NBA teams. Additionally, CBF can incorporate external information, such as injuries or recent trends, which CF inherently lacks.

In summary, while CF can be effective in settings with stable and frequent interactions, the dynamic, sparse, and evolving nature of NBA team data makes it less suitable for game score prediction compared to CBF.

The analysis revealed that pure CF and CF-centered hybrid models yielded higher Relative RMSE values in comparison to other approaches, particularly for teams with insufficient historical data. This limitation underscores the need for adaptive techniques that can effectively handle data sparsity and evolving team compositions.

Interestingly, simpler hybrid models, such as those employing mean prediction, demonstrated notable effectiveness. Despite their straightforward nature, these models provided competitive performance, balancing the strengths of both CF and Content-Based Filtering (CBF). While the meta-model—a stacking-based hybrid approach—achieved the highest overall accuracy, it introduced additional computational complexity by requiring another layer of machine learning. For practical applications, the simpler mean-prediction hybrid approach offers a compelling balance of efficiency and effectiveness, particularly in scenarios with limited resources.

Despite efforts to enhance the Collaborative Filtering (CF) model by assigning greater weights to recent matchups, the improvement in Relative RMSE was negligible. The results showed only a minor difference, with the metric changing from 0.283 to 0.2828.

Obsah obrázku text, snímek obrazovky, diagram, řada/pruh

Popis byl vytvořen automaticky

Obsah obrázku text, diagram, řada/pruh, snímek obrazovky

Popis byl vytvořen automaticky

1. Conclusion

This project developed and evaluated a machine learning framework to predict NBA team performance using a hybrid recommendation system approach. By leveraging historical game data, team statistics, and player features, we demonstrated the efficacy of combining Collaborative Filtering (CF) and Content-Based Filtering (CBF) to enhance prediction accuracy.

The CF model effectively utilized historical matchup data to predict outcomes in scenarios with abundant historical records, while the CBF model excelled in cold-start situations by incorporating team-specific and player-specific attributes. The hybrid methodologies—dynamic weighting, selective model use, and model stacking—showcased adaptability by balancing the strengths of CF and CBF according to the data context. Among these, the stacking approach emerged as a robust solution, with the meta-learner optimizing prediction weights to achieve the lowest Relative RMSE.

These findings highlight the versatility of hybrid recommendation techniques in addressing challenges like data sparsity and dynamic game contexts, providing actionable insights for sports analytics. Future work could explore integrating advanced deep learning methods, additional features like real-time game dynamics, and extending the approach to other sports or competitive domains.

References

1. Geetha, G., Safa, M., Fancy, C., & Saranya, D. (2018). A Hybrid Approach using Collaborative filtering and Content based Filtering for Recommender System. *Journal of Physics Conference Series*, *1000*, 012101. https://doi.org/10.1088/1742-6596/1000/1/012101
2. Widayanti, R. (2023). Improving Recommender Systems using Hybrid Techniques of Collaborative Filtering and Content-Based Filtering. *Journal of Applied Data Sciences*, *4*(3), 289–302. https://doi.org/10.47738/jads.v4i3.115
3. Thorat, P.B., Goudar, R.M. and Barve, S., 2015. Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications*, *110*(4), pp.31-36.
4. *NBA stats (1947-present)*. (2024b, December 1). Kaggle. https://www.kaggle.com/datasets/sumitrodatta/nba-aba-baa-stats?resource=download&select=Advanced.csv
5. *NBA games data*. (2022b, December 23). Kaggle. https://www.kaggle.com/datasets/nathanlauga/nba-games?select=players.csv

Author

**Mikhail Karimov** – Bachelor of Informatics and management, FIT CTU, karimmik@fit.cvut.cz