Champions League Prediction Model Report

Team CudaNotAvailable

June 2025

Problem Statement

The task was to predict UEFA Champions League knockout stage outcomes (from the Round of 16 to the final) utilizing historical match data, team performance metrics, and advanced analytics. Predictions spanned seasons from 2017/18 to 2023/24, with the training data cutoff set as April 31, 2017.

Data Collection

- Historical match data from European leagues and Champions League matches.
- Elo ratings were scraped from `clubelo.com` to incorporate real-time team form indicators.

Data Preparation and Feature Engineering

- Data was cleaned, normalized, and structured chronologically.
- Distinct weighting was applied: 70% for UEFA Champions League data and 30% for domestic league data.
- Features used from the dataset included:
  - Attack and defense strength ratios (overall, UCL-specific, and league-specific)
  - Overall team form and goal difference ratios
  - UCL-specific form and goal difference
  - League-specific form and goal difference
  - Experience differentials in UCL matches
  - Historical win counts
  - Elo rating differences
  - Home advantage indicators
- Elo ratings were integrated to reflect current form.

- Team statistics were calculated for each match up to the match date, capturing historical performance including goals scored/conceded, points per game, and goal differences.

Preprocessing

- Missing numerical values from the 2023 season were imputed using an Iterative Imputer with Random Forest Regressor.
- Numerical data underwent standard scaling for normalization.
- Teams were label-encoded, significantly reducing data dimensionality and improving computational efficiency.

Model Development

- Employed XGBoost for its robustness and performance, particularly effective for binary-like classification tasks as needed here.
- A predictive pipeline included numeric data scaling followed by an XGBoost regressor/classifier. Predictions were binarized into win/loss outcomes using a custom wrapper.
- Hyperparameter tuning using Optuna and feature selection optimized the model's predictive capability.

Simulation of Tournament

- Actual Round of 16 matchups served as the baseline for predicting subsequent stages.
- The model iteratively predicted match outcomes and advanced winners through each round, progressively updating the data used for predictions.

Results

- Achieved a competitive prediction score, with best public score of **41.66666** and second highest score of **41.0742**.
- **Highest public score attempt:** Training data `best_output_training_data.csv` was generated using `notebook_to_generate_2nd_best_output_train_data.ipynb` and its output is uploaded to `notebook_to_generate_best_output_train_data.ipynb`. This is passed through `best_output_pipeline.ipynb`, which generates ELO data for test teams as well (requires uploading `ELO_ratings_scraped.zip`).
- **Second highest score attempt:** Training data `2nd_best_output_training_data.csv` was generated using `notebook_to_generate_2nd_best_output_train_data.ipynb`. This was passed through `2nd_best_output_pipeline.ipynb` which simulates the tournament accordingly.

Recommendations and Future Work

- Enhance the pipeline by exploring further sophisticated metrics like Expected Goals (xG).
- Consider integrating advanced NLP models to automate literature review and discover additional influential predictive metrics and strategies.

Conclusion

This structured approach yielded reliable predictions, effectively balancing sophisticated modeling with critical sports-specific feature engineering.