CSC 522 Fall 2024 Project

Soccer Match Outcome Prediction

Team-5

Apurv Choudhari
Chinmay Singhania
Madhur Dixit
Parth Kulkarni

NC STATE UNIVERSITY

Introduction

- Soccer match outcome prediction is challenging due to complex factors
- Traditional methods often focus on historical team performance but miss the impact of individual players and recent trends.
- Our project aims to address this limitation by developing a more comprehensive prediction model that integrates:
  - Player-specific metrics
  - Team dynamics
  - Advanced feature engineering techniques (Feature Tools, SMOTE, Manual feature engineering)

Related Work

- Classifier Comparisons: Research on a 10-year dataset (2002-2012) explored classifiers like Random Forest and SVM, achieving error rates of 0.50. This highlighted challenges in classifying balanced outcomes like draws.
- Deep Learning Approaches: The SoccerNet model, using GRU-based deep learning on Qatar Stars League data, achieved over 80% accuracy. It revealed the critical role of defenders and the last 15-30 minutes of gameplay in determining outcomes.
- Hybrid Models: A study combining machine learning with statistical analysis on a dataset of over 205,000 matches achieved 46.6% accuracy. Interestingly, league-specific data only marginally improved predictions.

Method

Datasets:

- SQLite Database: Seven interconnected tables capturing relational data.
- Premier League Match Results: Match data spanning 10 seasons.

Machine learning techniques applied to the problem:

Machine learning techniques applied include feature engineering, SMOTE for class imbalance, Random Forest, XGBoost, Neural Networks, ensemble learning, hyperparameter tuning, and evaluation with cross-validation and performance metrics.

Method

Novel Approach:

- Innovative Transformation: Converted relational data into a machine-learning-ready format using Feature Tools and manual feature engineering.
- Advanced Preprocessing: Applied techniques like SMOTE, feature scaling, and encoding to enhance model accuracy.
- Scalable Design: Developed an adaptable pipeline for both datasets, leveraging cloud infrastructure (AWS) for efficient processing and training.

Method

Rationale:

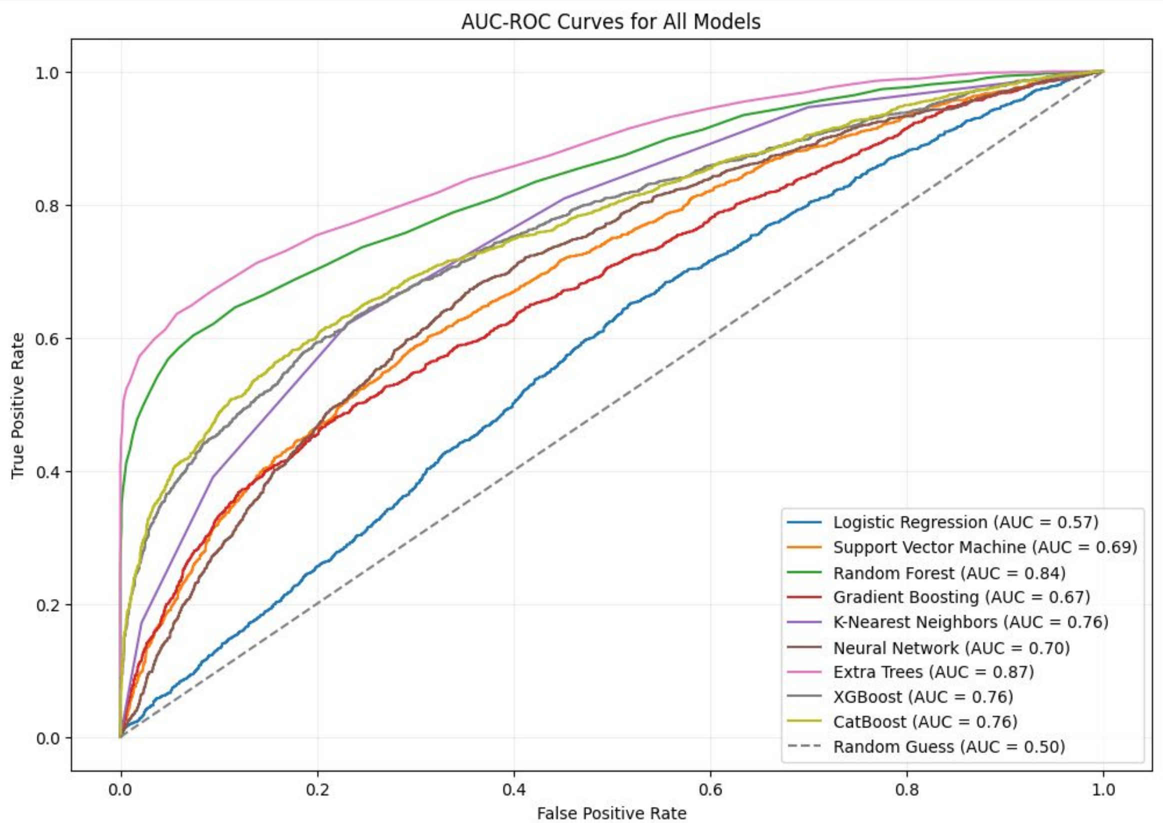
- Streamlined complex data structures for predictive analysis.
- Designed features capturing domain-specific dynamics:
    - Player abilities
    - Team performance metrics
    - Match outcomes and trends
- Addressed class imbalance for unbiased predictions across datasets.

Method

Approach:

- Integrated Feature Tools for data transformation and domain knowledge for feature engineering.
- Leveraged advanced preprocessing techniques for consistency.
- Validated pipeline on both datasets, demonstrating adaptability and achieving significant performance improvements.

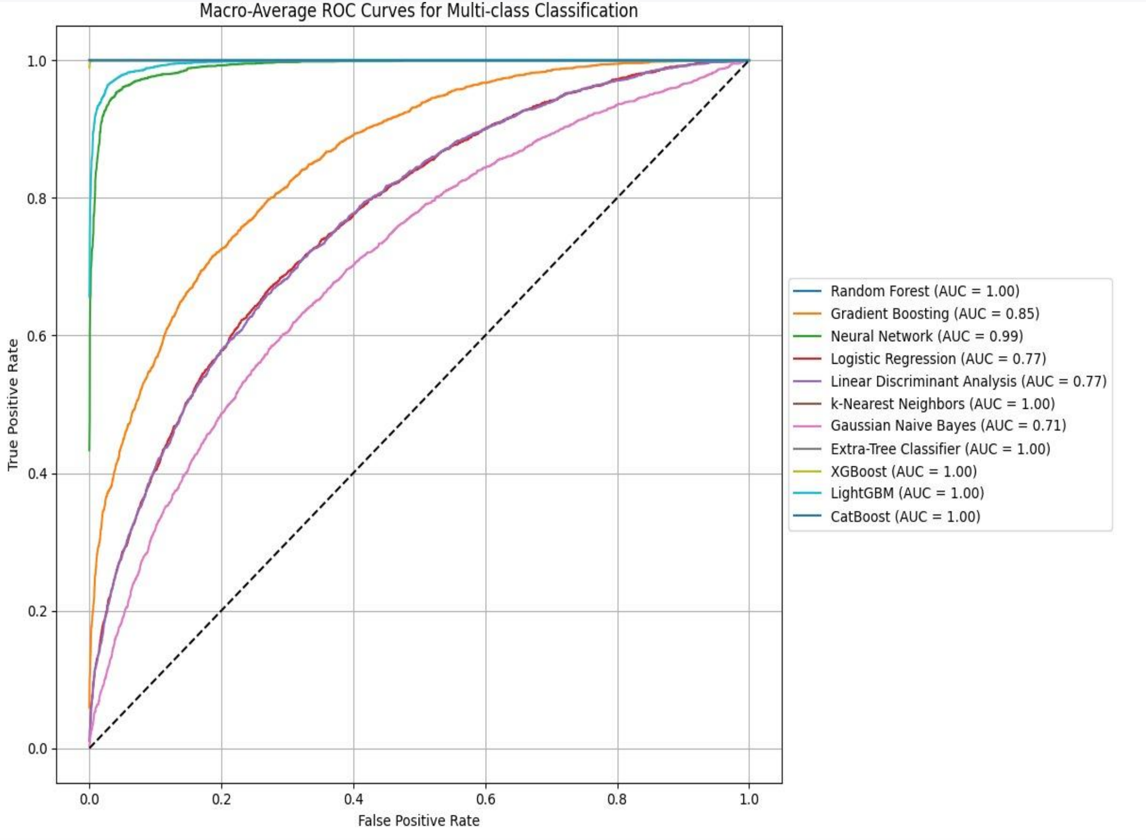
Results



These are the AUC-ROC curves that we obtained from the models trained on manual feature engineering dataset.

As observed from the plot the models do not particularly perform well on this dataset and has poor results.

Results



The new dataset when implemented, had considerably better accuracies and AUC-ROC scores for similar models as evident from the diagram.

Results

- The previous dataset was overly complex; even after feature engineering the features of the dataset were inadequate for sophisticated machine learning algorithm pipelines.
- The complexity arises from the fact that the first dataset did not adequately represent the underlying patterns needed for accurate prediction.
- The second dataset contained features that aligned better with the target variable, allowing the models to leverage these patterns effectively for better predictions.
- The second dataset provided clearer, more actionable signals for the models to learn from, unlike the first dataset, which was clouded by noise and irrelevant variance.

Discussion

- Football match prediction by Fatima, Angelo (2022) - https://doi.org/10.1016/j.procs.2022.08.057

Dataset	Variables	Matches	Accuracy
Paper	31	1580	65.2
Old Dataset	61	25000	69
New Dataset	16	45000	92

- Highlights
  - Joblib, Dask, and Cloud infrastructure
  - Transformed raw SQLite data effectively
  - Pipeline of 9 Models, GridSearchCV
  - Complex models with domain informed feature engineering

thank you