Feature Importance Analysis: Predicting Tennis Match Outcomes

Liz Garcia Ovalles, Computer Science Department, Class 2025.

MOTIVATION

- Investigate discrepancy in the findings of scholarly papers
- Provide valuable insights to bettors and tennis training programs
  - Facilitate understanding of the key factors that influence men's single tennis match outcomes on the ATP circuit through data-driven analysis

GOAL

- Train Random Forest and XGBoost models to perform feature importance analysis and determine the predictive power of rank vs. serve-related features
- Draw conclusions in support or against the findings of existing research

RELATED WORK

- **What has been done before?**
  - There is vast research on the use of machine learning models to predict tennis match outcomes
    - Logistic regression, random forest, gradient boosting models, and many other models, have been deployed to predict and analyze tennis match outcomes.
- **What is my contribution?**
  - Application of XGBoost model with rank and serve-related features as predictors
    - Rarely any studies have done this
  - Verify the findings of previous studies which arrived at contradictory conclusions regarding the importance of rank versus serve strength in accurately predicting tennis match outcomes

RF = Random Forest
XGB = XGBoost

RESULTS

Predict with all features

Model	Accuracy	F1 Score
RF	0.93	0.93
XGB	0.94	0.93

Predict with rank

Model	Accuracy	F1 Score
RF	0.59	0.58
XGB	0.63	0.63

Predict with serve-related features

Model	Accuracy	F1 Score
RF	0.81	0.80
XGB	0.84	0.84

Predict with rank and serve-related features

Model	Accuracy	F1 Score
RF	0.82	0.82
XGB	0.85	0.85

APPROACH

- The inclusion of an XGBoost model is of particular interest as there is a scarcity of research employing this tool for predicting tennis match outcomes based on players' characteristics and match statistics, including rank and serve strength data.

IMPLEMENTATION

- **Models**
  - **Random Forest**
    - Employed by studies which inspired this research (included for reproducibility and comparison).
  - **XGBoost Regression**
    - Can capture nonlinear relationships.
    - Regularization and adjustable parameters (helps reduce overfitting).
- **Other Tools Employed**
  - **Python**: project implementation.
  - **Pandas**: data manipulation.
  - **Matplotlib and Seaborn:** visual data analysis.

DATA AND VARIABLES

- **Data:** ATP data on men's singles matches from 2000 to 2024.
- **Source:** Jeff Sackmann's repository on GitHub.
- **Features**: year, tournament name, tournament level, minutes, surface type, draw size, round, winners' and losers' handedness, height, age, rank, ratio of aces over double faults, percentage of first/second serves in, percentage of first/second serves in and won, and percentage of break points saved.

CONCLUSION AND FUTURE WORK

- Training models with only serve-related features retained strong performance relative to when the models were trained with all features, with accuracies of around 80-85%, unlike with rank which achieved accuracies of around 60-65%.
  - Indicates serve strength is a better predictor over rank of tennis match outcomes
- Future work could perform a comparative analysis of the impact of serve-related features on doubles versus singles matches to evaluate the generalizability of these findings and offer more targeted insights for players based on their specialization.