Evaluating Performance and the Role of Coaching in Women’s Tennis

Methodology and Data Extraction

In the realm of professional tennis, the intricacies of player performance and the impact of coaching have long been subjects of fascination and study. Understanding how players evolve and improve over time, especially under the guidance of different coaches, is not only of academic interest but also holds practical implications for players, coaches, and the tennis community.

In recent years, the availability of comprehensive data sets from the Women's Tennis Association (WTA) has opened up new avenues for analyzing player trajectories and assessing the efficacy of coaching strategies. Through meticulous data extraction and analysis, we aim to uncover patterns, trends, and insights that shed light on the dynamics of player-coach relationships and the factors influencing success on the professional circuit.

In this section, we delve into the process of WTA data extraction and outline the methodology employed in our analysis. Drawing on specific case studies, we aim to demonstrate how this data-driven approach deepened our understanding of player development and performance enhancement in the competitive world of women's tennis.

WTA Data Extraction: Methodology and Approach

Our methodology encompasses several key steps:

1. Data Collection and Preprocessing: The foundation of our analysis lies in extracting and organizing pertinent data from the vast repositories maintained by the WTA, journals, and news articles. This process involves accessing and collating information on player rankings, coaching affiliations, and professional trajectories. Preprocessing involved tasks such as removing duplicates, standardizing formats, resolving discrepancies or missing values, streamlining the data set, and enhancing its reliability and usability for subsequent analysis.

1a. Creation of Grade Slabs

Player earnings are categorized into different earning brackets based on their global ranking. For instance, players ranked within the top 10 earn approximately $2,811,484, while those ranked between 11-20 earn $1,614,396, and so forth. Each earning bracket corresponds to a specific grade, ranging from Grade 1 for the top 10 players to Grade 15 for players ranked between 501-550th globally. These earnings serve as a representation of the financial rewards associated with different levels of performance in women's tennis.

1b. Creation of Economic Coaching Zones

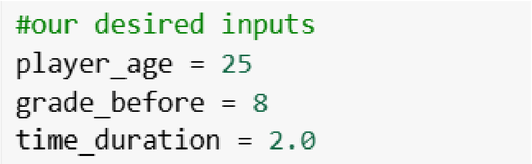
The 33 coaches for whom the dataset was complete were categorized based on the economic zones of their countries, determined by per capita GDP statistics of their countries. The countries were divided into four income brackets: Grade 1 for GDP per capita ranging from $40k and above, Grade 2 for $20-40k, Grade 3 for $10-20k and Grade 4 $0 to $10k and below. Determining the income grade was essential for understanding the economic background of each coach, offering insights into their potential adaptability and effectiveness in coaching players across various age groups and skill levels and geographic regions.

2. Feature Encoding: We identified and constructed relevant features encompassing various player and coach dynamics aspects to extract meaningful insights. These features included metrics related to player rankings, coaching tenure, and geographical regions. We carefully selected features to capture nuanced patterns and correlations within the data.
(See Appendix A)

3. Model Development: Leveraging machine learning techniques and statistical modeling to comprehensively analyze the impact of coaching on player performance. We harnessed historical data and uncovered underlying trends and relationships between coaching interventions and player outcomes.
Subsequently, a Linear Regression model was employed to predict future coach selection based on the player's current career point which included age and current grade to attain maximum grade improvement in set time. Model performance evaluation was conducted using Mean Squared Error (MSE), providing insights into the predictive capabilities of our framework. Predictions were rounded to the nearest integer grade to understand potential grade advancements.
This comprehensive mathematical framework ensured that our model effectively captures the dynamic nuances of player performance and offers actionable insights into the role of coaching in shaping player rankings and earnings trajectories.
(See Appendix B)

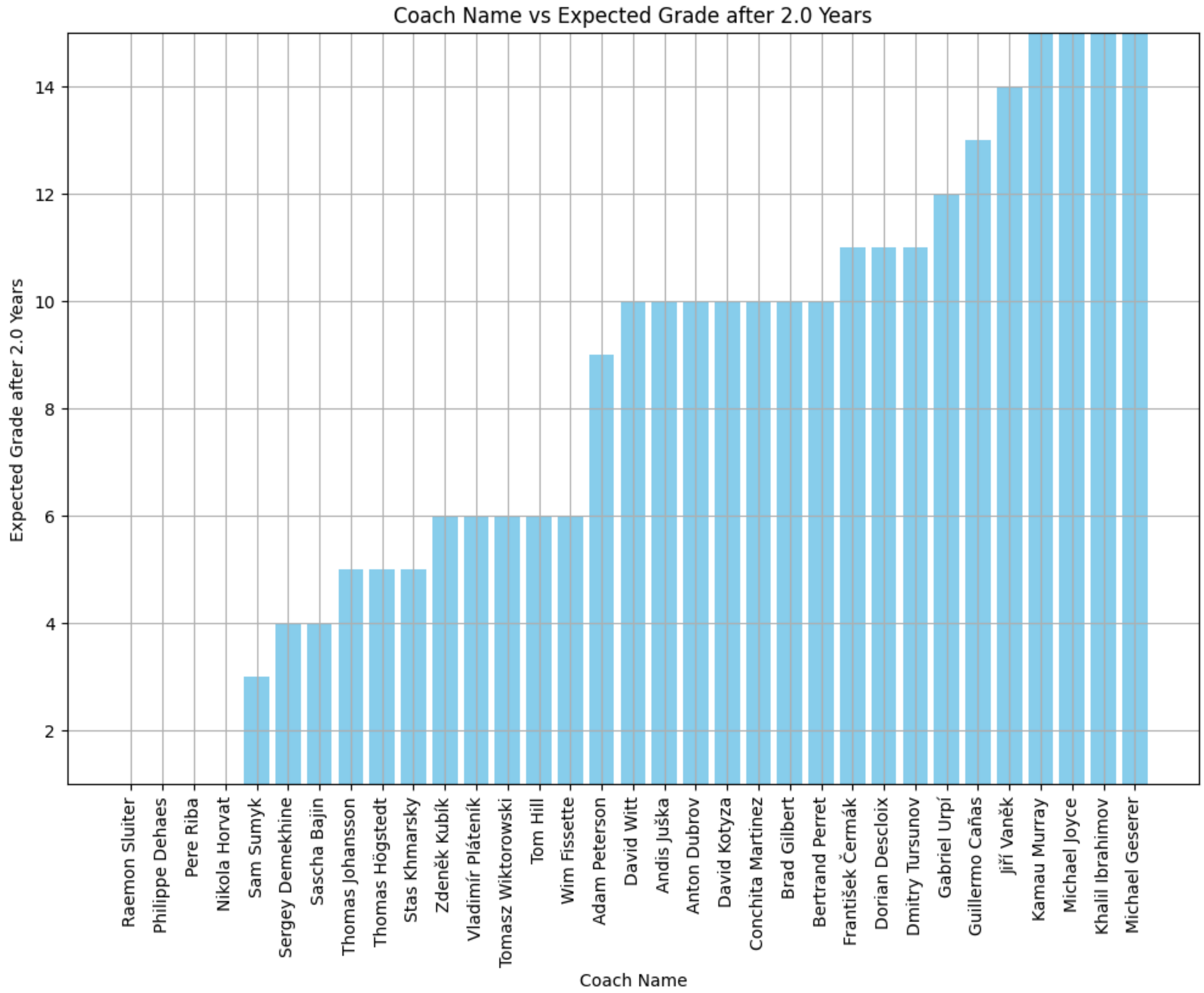
Practical Example
To demonstrate the efficacy of our approach, we present illustrative examples featuring notable player-coach partnerships:
In a hypothetical scenario, Player A, from Grade 8, 25 in age, aims to gain maximum grade improvement in 2 years time. Our model steps in here, leveraging historical data and predictive analytics to streamline the coach selection process for Player A. By evaluating coaching success metrics and financial implications, the model provides tailored recommendations, empowering Player A to make informed decisions aligned with her goals and resources.
Further, the model also outputs The Mean Squared Error (MSE), which measures the regression model's average squared difference between the actual and predicted values. Lower MSE values indicate a better accuracy of predictions, implying that the model's performance in estimating the maximum potential to advance from one grade to another in set time is relatively acceptable.

For our practical scenario, the inputs are as follows:



```
#our desired inputs
player_age = 25
grade_before = 8
time_duration = 2.0
```

Output graph:



Analysis of Graph

- □ The graph illustrates the anticipated Grade improvement under the coaching of various coaches, with the x-axis denoting individual coaches and the y-axis scaled to reflect predicted Grades, reaching up to 15 for clarity.
- □ Notably, Raemon Sluiter, Philippe Dehaes, and Nikola Horvat emerge as highly recommended choices for player-coach partnerships. Projections suggest a remarkable advancement to Grade 1 within the set 2-year timeframe, aligning with substantial

financial rewards. Grade 1 attainment typically correlates with an average earning of $2,811,484, highlighting the lucrative potential of these partnerships.

- □ In contrast, coaches spanning from Adam Peterson to Michael Gesserer are predicted to be the least recommended, as their guidance may lead to counterproductive outcomes, potentially resulting in a decline in Grade, even dropping to Grade 15. This downward trajectory not only signifies a setback in sporting performance but also threatens players' financial prospects, with Grade 15 earners averaging only $11,738.
- □ Between these extremes, coaches such as Sascha Bajin to Wim Fissette fall into a moderately recommended category. While their guidance is projected to yield Grade improvements from 8 to 6, the associated financial benefits, averaging around $614,339 for Grade 6 players, still present a reasonable pathway to prosperity.

In essence, this analysis underscores the intertwining of coaching effectiveness, Grade progression, and financial outcomes, emphasizing the critical role of coach-player partnerships in shaping both athletic partnerships and economic success.

Analysis of Coach’s Salaries based on Economic Zones

- □ To analyze how the salary of a coach affects their standard of living based on their country's economic standing, the coaches' zones were analyzed according to per capita GDP figures. Countries were classified into four income zones: Zone 4 for $0-$10k, Zone 3 for $10k-$20k, Zone 2 for $20k-$40k, and Zone 1 for $40k+.
- □ We examined the financial impact of two tennis coaches, Zdeněk Kubík from Czechia (Zone 2) and Sam Smyuk from France (Zone 1), on Player A, whose starting Grade is 8 with a corresponding average income of $521,465 annually, aiming for the maximum possible grade jump in 2 years.
- □ Under Kubík's coaching, Player A is assured to reach Grade 6, resulting in a substantial income boost to $614,339. Kubík's salary, 20% of Player A's earnings, would be $122,867.80 annually. For each grade improvement, Kubík's compensation is $61,433.90 .This places Kubík in the upper strata of Czechia's income distribution, affording him a luxurious lifestyle with ample opportunities for savings and investments.
- □ In contrast, Sam Smyuk's coaching strategy boosts Player A to Grade 3 within the same timeframe, resulting in an income of $1,614,396. Smyuk's salary, also 20% of Player A's earnings, would amount to $322,879.20 annually. His compensation per grade improvement is $64,575.84 . While Smyuk's earnings are substantial, they align more closely with the economic norms of France. This provides him with a comfortable lifestyle but does not elevate him to the same level of affluence as Kubík in Czechia.
- □ In conclusion, the analysis highlights the profound impact of coaching salaries on the standard of living, particularly when considered within the context of their respective countries' economic standings. Kubík's earnings significantly enhance his living standards, placing him comfortably within Czechia's high-income bracket. Smyuk, while

enjoying a respectable income reflective of France's prosperity, doesn't reach the same level of affluence. His earnings are notable but in line with the economic standards of his affluent nation, resulting in a less dramatic increase in his living standards compared to Kubík.

The analysis underscores the significant impact of coaching on a player's progression and financial success in the tennis industry. Further, with special focus on coaches and their inte
The varying outcomes under different coaches highlight the importance of selecting the right mentorship to maximize player development and achieve desired career milestones.

- □ A deep analysis also indicates that the coaches predominantly come from countries in the global north, with the United States providing the highest percentage at 21.6%, followed by Czechia (13.5%), France, and Spain (both 10.8%) from Europe. In contrast, other northern countries like Poland, Croatia, Germany, and Sweden contribute a smaller percentage. Only a few coaches are from the global south countries like Belarus, Russia, Tunisia, Ukraine, Argentina, and Latvia.
- □ An examination of our dataset furthermore brings to light a notable lack of female representation among professional tennis coaches. Of 148 coaches, only 14 are women, constituting approximately 9.46% of the total. This finding underscores a broader systemic gender disparity within the sport. The underrepresentation of women in professional tennis coaching is consistent with existing research, which indicates that only 8% of WTA Top 100 players currently employ a female coach. High-profile examples, such as Conchita Martinez coaching Karolina Pliskova, are exceptions rather than the norm [1]

[1] https://www.wtatennis.com/news/1439899/why-aren-t-there-more-female-coaches-on-tour-coaches-and-players-weigh-in-

Appendix A: Data Collection, Preprocessing and Description

In our comprehensive analysis of player-coach partnerships in women's tennis, we compiled detailed datasets using authoritative sources such as the official WTA website,[2] renowned tennis news platforms like Tennis.com[3] and ATP Tour[4], and respected tennis journals including Tennis Magazine and The Tennis Times. Our data focused on the top 100 women's tennis players from 2013 to 2023, capturing the most impactful player-coach relationships.

We employed a rigorous manual web scraping process to gather the necessary information, meticulously combing through player profiles, tournament reports, and coaching announcements. We cross-referenced information from multiple sources to mitigate discrepancies and gaps, ensuring the completeness and accuracy of our dataset. We documented 148 unique coaches and created a detailed mapping of each coach to the players they worked with during the specified period. This dataset included crucial details such as starting and ending ranks, ages, years of partnership, peak rankings, and financial metrics like changes in player income and projected coach earnings.

During the data cleaning and preprocessing phase, we resolved inconsistencies, handled missing data points, and standardized the data format. We performed deduplication to eliminate redundant entries and used feature engineering to derive meaningful insights, incorporating age groups (novice: 0-25 years, middle: 26-30 years, experienced: 31+ years) to better understand each coach's impact on players at different career stages.

For model training, we utilized the comprehensive dataset capturing player-coach relationships and coaching histories, focusing on 33 carefully selected coaches with proven track records and consistent data across multiple seasons. By concentrating on these highly successful coaches, we aimed to establish a robust baseline for our model's predictions and recommendations, aligning to optimize coaching partnerships for players from diverse economic backgrounds and regions.

Dataset 1: Player-Coach Relationship Dataset

The first dataset, 'WTA Database(Players-Coach).csv' [5], provides a comprehensive overview of WTA players and their coaches over a decade, from 2013 to 2023. This dataset captures essential details such as the annual rankings of players and the corresponding coaches they were associated with each year.

For each player, the dataset includes:

- Their rank at the end of the year.
- Prior rank at the beginning of the year.

[2] https://www.wtatennis.com/

[3] https://www.tennis.com/

[4] https://www.atptour.com/en

[5] WTA Database(Players-Coach)

- First name.
- Last name.
- Nationality.
- The specific year of data entry.

The coach information is a critical component, recording the name of the coach working with the player during that specific year. This detailed data allows for an analysis of the coach's influence on the player's performance.

Dataset 2: Coach-Player Performance Dataset

The second dataset, ‘WTA Coach- Player Mapping .csv ‘[6], maps coaches to the players they worked with from 2013 to 2023, capturing essential performance metrics, player demographics, and financial outcomes of each coaching partnership. This dataset includes the names of coaches and corresponding players, player ages before and after the coaching period, years of partnership, rankings before and after the coaching period, peak rankings, and financial earnings before and after coaching. It also records the projected coach earnings, typically 20% of the player's income, offering a perspective on the financial rewards for coaches based on their players' success.It also includes the defined economic zones of the coach’s countries.

Dataset 3: Player Financial Earnings and Rankings Dataset

The third dataset, ‘WTA Earning Database.csv’ [7], categorizes WTA players based on their annual rankings and corresponding financial earnings, providing a detailed overview of the economic landscape of women's tennis. Each player is assigned a grade based on their rank, with specific earnings associated with each grade. The rankings are divided into several tiers, ranging from the top 10 players to those between 501 and 550.

[6] 🟩 WTA Coach- Player Mapping

[7] 🟩 WTA Earning Database

Appendix B: Model Description and Mathematical Framework

Data Description

The datasets `WTA Coach- Player Mapping .csv` and `WTA earning Database.csv` contain historical performance data for women tennis players. The first dataset includes their current ranks, earnings, starting and ending age with a coach, grade, and rank change with a coach. The latter dataset includes the grade-wise earnings of the top 100 globally ranked players based on stark differences in the pay slab. The ranking system assigns grades to players, with corresponding earnings for each grade. The goal of the model is to predict the future grade of a player based on their current grade, age, and coach and subsequently estimate the financial implications of the predicted grade.

Data Preprocessing

1. Label Encoding: Coach names are categorical variables and, thus, need to be encoded numerically. We employed Label Encoding to convert the coach names into numerical values.

2. Feature and Target Variables: The features (X) include `CoachName,` `PlayerAge,` and `GradeBefore.` The target variables (y) are `GradeAfter` and the number of years (`year`).

3. Custom Weight Function: A custom weight function is used to assign weights to the samples based on their initial grades. This approach allows us to account for the varying significance of different grades in the model training process.

4. Train-Test Split: The data is split into training and testing sets, with 80% used for training and 20% for testing.

Model Training

A Linear Regression model is used to predict the future grade and best suited coach in desired time. The model is trained using the training set, with sample weights applied to account for the significance of different grades.

Model Evaluation

Predictions are made on the test set, and the model performance is evaluated using Mean Squared Error (MSE).

Prediction and Analysis

The model predicts the best suited coaches for a player for whom we have her age, current grade and the set time the player wants maximum results in. The predictions are rounded to the nearest integer grade, and the change in income is calculated based on the predicted grade.

Visualization

The visualization presents a line graph where coaches are arranged from left to right based on their predicted compatibility. The y-axis shows the predicted grade attainment after the desired duration, ranging from 1 to 15. This scaling is employed for clarity and to accommodate the variation in the dataset's grade levels ranging from 1 to 15.

Mathematical Framework

The linear regression model aims to fit a linear equation to the data points that best predict the target variables. The mathematical representation is given by:

$Y = X\beta + \epsilon$

where:

- y is the vector of target variables.
- X is the matrix of input features.
- β (Beta) is the vector of coefficients (weights).
- ϵ (epsilon) is the error term.

The model parameters B are estimated by minimizing the weighted sum of squared residuals:

Minimize: $\Sigma (wi (yi - X_i \beta)^2)$ for i = 1 to n

where w_i are the sample weights derived from the custom weight function.

Evaluation Metric

The Mean Squared Error (MSE) is used to evaluate the model's performance:

$MSE = 1 / n * \Sigma (yi - \hat{y}i)^2$ for i = 1 to n

Where:

- MSE represents the Mean Squared Error
- n represents the total number of data points
- yi represents the total number of target values for data point i
- $\hat{y}i$ represents the predicted value of data point i obtained from the model

This comprehensive framework ensures that the model captures the sensitivities of the player performance data and provides meaningful predictions regarding the impact of different coaches on player rankings and earnings.