Ekaterine Jorjoliani

MDA 620

Professor Gurpreet Singh

December 14, 2023

Final Capstone Project: Report

Men's Tennis Grand Slam Data Analysis

https://www.kaggle.com/datasets/wonduk/mens-tennis-grand-slam-winner-dataset

Table of Contents

Background

For the final capstone project, I created data analysis about men's tennis grand slam winners and tried to predict their prize money. As a tennis player myself, I found this data to be interesting and relevant to work with. In the tennis and athletics world in general, it can be beneficial and serve many good purposes such as financial planning for the players, players' career strategy, tournaments attracting top players by offering the rewards, understanding player success in sports analytics, and last but not least it can also help with sponsorships and endorsements.

The analysis focuses on men's tennis grand slam tournaments from the years 2000 to 2020. Player attributes considered include age, ATP ranking of winners and runners-up, match statistics, tournament surface, age, year, whether the players were right or left-handed, the country they were from, etc.

Problem Scenario/Business Issue

The world of tennis is a large and complex ecosystem with financial implications. The amount of money that circulates in it is highly vast. It is crucial to understand the determinations of prize money for both the players and the tournament staff, however, the distribution of prize money amongst tournament winners is complex.

Tennis players need to plan their careers, find sponsors, and decide which tournaments to participate in. Understanding prize money, how it works, and seeing predictions regarding it, can help the players with their goals.

Tournament organizers must attract top players, and offering a competitive amount of prize money is the key to their business. An understanding of the relationship between player performance and prize money can guide tournament organizers in structuring attractive financial prizes.

Interpreting the relationship between player characteristics and tournament results can help us understand player success more deeply from the standpoint of sports analytics. Beyond just financial concerns, this information might reveal trends that point to a player's competitive advantage.

Objective/Goals of the Project

The primary goal/objective for this project was to analyze men's tennis grand slam winners' historical data to develop a predictive model for their prize money by exploring the relationship between variables of player attributes and tournament outcomes and building predictive models to estimate prize money of player performance and other factors. The main research questions to be answered are: do specific player attributes impact player's prize money and can historical data of men's grand slam winners predict prize money for future tournaments?

Data Exploration/Data Visualization

The dataset contains 292 entries and 10 columns. The Columns include information such as the year, tournament details, winner and runner-up players, their nationalities, ATP rankings, right or left-handedness, tournament surface, and winner's prize. The columns “Winner ATP Ranking', ‘Runner Up ATP Ranking and 'Winner Prize’ have some missing values. Therefore, it is important to perform functions that will remove the missing values. Our target variable will be the winner's prize money and we will use it for prediction models.

I ran a few visuals to interpret the data more clearly. I created a correlation matrix, which showed that not a lot of the variables have a high correlation between one another. The variables with the highest positive correlation are winner prize and year, while the highest negative correlation is between the winner and runner-up ATP rankings. However, these correlations are still not very high, which might make our predictions more complex in accuracy.

Another plot created was the distribution of winner prize money. The X-axis represented the range of winner prize values from 0 to 4,500,000 and Y-axis was the frequency within the prize values from 0 to 16. The plot shows the distribution of winner prize values across different intervals, the bar is on the highest on 2.0 which means the highest frequency is 2,000,000 and the most common prize money is $2 million, after that it is 1.5, 1.0, 2.5, 3.5, 3.0, and 4.0 being the lowest.

Another plot for interpretation was the boxplot of winner prize money distribution amongst the surface types. There were no outliers in the plot, the highest prizes were on plexicushion prestige and DecoTurf Outdoor surfaces. Clay and DecoTurf had the best median

which means the data was normally spread out. We do not have much information about the Rebound Ace on the plot.

The last plot I did was the pair plot which shows us the distribution of different variables using bar charts and scatterplots. The plots are easily understandable. The main purpose of the plots was to observe the correlations and patterns between variables. They showed us that there is not a big correlation, however, it gave us more insight into the variables.

I also used plots like a scatterplot, bar plot, and confusion matrix after creating the models to better see the accuracy of the predictions, which will be discussed later in the report.

Data Manipulation

During the preprocessing stages, the dataset was thoroughly analyzed to determine its structure and identify any areas that needed improvement. Managing the dataset's missing values was an important component. It was crucial to handle missing values properly since they can negatively impact machine learning models' performance. The dataset was examined, and it was found that some variables had missing values. To guarantee the model's integrity and dependability, it was decided to eliminate cases in which the relevant variables had missing values. This method was selected as the initial step because of the small number of missing variables and the possible effect on the flexibility of the model.

The dataset also had categorical variables, which could have made our model building more complex. To avoid this, it was important to deal with this issue in the preprocessing stage. Therefore, I decided to encode the categorical variables. Label encoding is a technique used to convert categorical data into numerical form, making it suitable for machine learning algorithms that require numerical input. Each unique category is assigned a unique integer.

After removing the missing values and encoding the categorical data, the dataset was ready for further analysis.

Methodology/Model Building

To build prediction models, regression, decision tree classification, regular decision tree, and random forest were used.

Regression is a type of supervised learning algorithm used for predicting a continuous outcome variable based on one or more predictor variables. The goal of a regression model is to establish a relationship between the input features and the continuous target variable. In this project, I used linear regression for the model building. Linear Regression has a linear relationship between the input and target variables. It finds the line that minimizes the sum of the squared differences between the observed and predicted values. After running the linear regression model using winner and runner-up ATP rankings, and year as independent variables I got certain winner prize money predictions, however, the mean squared error was 486742065541.7877, which is very high and indicates that the prediction accuracy is very low. To interpret the results better, I created a scatterplot, which showed that the accuracy was indeed low, as the dots on the plot were not close to the middle line, which is where they should be if the accuracy is high.

In the next step, a decision tree classification model was employed. Following the division of the dataset into training and testing sets and the execution of predictions, the model exhibited a good accuracy of approximately 92%. While this high accuracy suggests the efficacy of the classification model, it's noteworthy that the target variable in this case was not the primary variable of interest. Instead, the 'winner' variable served as the target, with 'winner nationality' and 'left or right-handed' as the independent variables. As our primary goal is to predict prize money, this classification model, focused on player attributes, won't be the primary choice. However, it serves as a valuable demonstration of a tree classification model. A visual

representation of the model's predictions is shown in a bar plot, indicating Novak Djokovic as the most successful player in terms of Grand Slam victories. In addition, a confusion matrix heatmap illustrates the accuracy of the model by showing the alignment between actual and predicted values, confirming the model's effectiveness in correctly identifying outcomes.

The next evaluation of two distinct models, a regular decision tree, and a random forest, involved considering key variables such as 'winner nationality,' 'winner ATP ranking,' 'tournament surface,' and 'year,' with the primary target variable being winner prize money this time. Despite employing various independent variables, models, and approaches, the accuracy of the predictions remained unsatisfactory. The mean square error for the decision tree model was 614,071,572,649.5726, and for the random forest, it was 424,048,950,023.0769. These high errors indicate a significant lack of accuracy in the models. Scatter plots were used to visually represent the results, revealing a pronounced misalignment of points from their expected positionings. This inconsistency further highlights the low accuracy of the predictive models.

Model Selection

When it comes to model selection, it is important to select the model which gives us the highest accuracy score. However, all our models with the primary target variable gave us very low accuracies. The explanation for this will be provided further in the report. As of now, the highest accuracy given by the models that involved our main target was the random forest model. It gave us a mean squared error of 424048950023.0769, which is still significantly high, but is nonetheless lower than other models. Therefore, if we had to choose a model for predictions based on this project, the random forest model would be selected.

Conclusions/Recommendations

Unfortunately, the pursuit of accurate predictions for the primary target variable, winner prize money, proved challenging even with the usage of different independent variables and models. Several factors could contribute to the observed low accuracy. It is possible that the dataset lacks key features that strongly correlate with winner prize money. Additionally, it can be difficult to predict because of the natural complexity of factors impacting prize money in tennis, economic conditions in sports change quite often, making it possible for the predictions to get complex. In recommendation, it could be necessary to conduct more research by adding more variables or using more advanced modeling strategies. Alternatively, there is still a chance that the winner prize money, with its complex factors, will always be difficult to accurately estimate within the boundaries of this dataset.

Bibliography/References

Men's Tennis Grand Slam Winner Dataset by "Wonduk", 2023

https://www.kaggle.com/datasets/wonduk/mens-tennis-grand-slam-winner-dataset