In professional tennis the rankings are typically based on set parameters. Throughout the year there are tournaments that are worth different numbers of points. There are four tiers of events, Grand Slams, Masters 1000, ATP 500 and ATP 250. You receive the most points in your Association of Tennis Professionals (ATP) rankings for winning a Grand Slam, and the least from winning an ATP 250 tournament. In the ATP, Grand Slam tournaments are played in a best of 5 format, and non-Grand Slam tournaments in a best of 3. In the WTA, all tournaments are played in a best of 3 format. The source of this data has calculated an ELO ranking, this ranking considers not only what tournament you are playing in, but who you are playing. For instance, if 4th ranked Coco Gauff beats 1st ranked Iga Swiatek, it will be worth more to her ELO ranking than if she beat 14th ranked Emma Navarro in the same tournament.

|  |  |
| --- | --- |
| Ace | A winning serve in which the opposing player doesn’t touch the ball |
| Double Fault | When two serves are missed in a row causing the serving player to lose the point |
| Return Point | Winning a point when the opposing player is serving |

Table : Tennis Terminology

In Tennis, there are also three different types of surfaces that are played on. The options are Grass, Hard, and Clay. The surfaces are important to keep track of as the speed of tennis changes, e.g., clay generally slows the ball down whereas grass speeds it up. Certain players perform better on certain surfaces. This dataset contains information for each player on each surface.

(In order to be included in the data set, players must have played a minimum of 10 matches overall or 5 matches on a particular surface. This data was filtered so only players who have recorded data on all three surfaces are present)

1. Below is a table of the available variables. Discuss how they might impact predicting win percentage and which ones might be helpful to use.

Table : Available Variables

|  |  |
| --- | --- |
| Variable | Description |
| Player | The name of the player |
| Surface | The surface of the tennis court |
| Matches | The number of matches the player has played |
| EloRank | The ELO ranking of this player |
| WinPercentage | The win percentage of this player |
| DoubleFaultPercentage | The percentage of the players’ services that result in a double fault |
| ReturnPointsWonPercentage | The percentage of return points that the player wins |
| AcesPerDoubleFault | The ratio of how many aces the player serves to double faults |

A player’s ranking will likely be a good determining factor in their win percentage so we should use that. Since we know the players play on multiple surfaces, we should use surface. The number of matches shows how much they play so those who play more are likely to be better. Both return points won percentage and aces per double fault are good metrics to see how good and how likely a player is to win a game as well so those should be included. Aces per double fault also will include data on double faults, so including the double fault percentage is redundant.

1. Below is a table of the characteristics of different court surfaces and how they benefit certain types of players. Think about which variables might benefit from an interaction term with surface.

Table : Characteristics of Court Surfaces

|  |  |  |
| --- | --- | --- |
| Surface | Characteristics | Best for |
| Clay | Slow, High bounce | Baseline players, Drop shots |
| Grass | Fast, Low bounce | Serve and volley, Big Servers |
| Hard | Medium speed, Highest bounce | Baseline players, Longer rallies |

Based on logic, if a player plays more matches on a surface, they will likely be more accustomed to that surface and therefore better. So, our number one variable that should include an interaction with surface is matches. As seen in the table above, we can see that all the surfaces will impact a players serve, and players with better serves will perform better. For this reason, we should also include an interaction between aces per double fault and surface in order to capture the difference the surface makes on serves.

1. Write the equation for our multiple linear regression model.

A screenshot of a computer

Description automatically generated

1. Based off the following plots, do you believe that this model meets the assumptions of a multiple linear regression model?

A graph with a line

Description automatically generated A graph with blue dots

Description automatically generated

The normal probability plot looks to be normal as all points fit closely to the given line. As well, the variance in the residuals looks to be consistent but there might be a couple outliers.

1. Comment on the overall quality of the model.

A close-up of a chart

Description automatically generated A black and white text with numbers and numbers

Description automatically generated

About 79.69% of the variability of win percentage is explained by the multiple linear regression model. This is quite a goo R-Squared value.

In our Analysis of Variance test, there is very strong evidence that the model containing Return Points Won Percentage, Elo Rank, Number of Matches, Surface, and Aces Per Double Fault is useful for predicting win percentage. F=29.27, P-Value = 0.

1. Interpret the coefficients for return points won percentage, grass, and hard with 10 matches played.

A screenshot of a computer

Description automatically generated

Return Points: A 1 unit increase in the return points won percentage is associated with a 1.632 increase in win percentage holding all else constant.

Grass: Compared to playing on Clay courts, we would expect the win percentage for playing on grass courts to be 0.0255 higher, on average, holding all else constant.

Hard: Compared to playing on clay courts, we would expect the win percentage for a player having played 10 games on hard courts to be 0.716 lower, on average, holding all else constant.