In tennis, there are 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. 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.

The goal of this worksheet is to build a regression model to predict tennis players match win percentages.

|  |  |
| --- | --- |
| 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 1: Tennis Terminology

(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 quantitative variables. Discuss what type of association each one of these variables might have on a player’s win percentage. All of the variables below are for each individual surface.

Table 2: Available Variables

|  |  |
| --- | --- |
| Variable | Description |
| Matches | The number of matches the player has played |
| EloRank | The ELO ranking of this player |
| WinPct | The win percentage of this player |
| DoubleFP | The percentage of the players’ services that result in a double fault |
| ReturnPWP | The percentage of return points that the player wins |
| AcesPDF | The ratio of how many aces the player serves to double faults |

EloRank: As it increases Win Percentage goes down.

DoubleFP: As it increases Win Percentages goes down.

ReturnPWP: As it increases Win Percentage goes up.

AcesPDF: As it increases Win Percentage goes up.

1. Below is a model fitted with only the quantitative variables. Interpret the EloRank and ReturnPWP coefficients and the R-sq value.

A close-up of a graph

Description automatically generatedA screenshot of a computer

Description automatically generated

EloRank: A 1 unit increase in EloRank is associated with a 0.002182 decrease in Win Percentage.

AcesPDF: A 1 unit increase in AcesPDF is associated with a 0.0651 increase in Win Percentage.

R-sq: About 54.51% of the variability of win percentage is explained by the multiple linear regression model.

1. Below is a model fitted including the categorical surface variable. Interpret the grass and hard coefficients and compare the R-Sq value to the R-sq(adj) value.

A close-up of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

Grass: Compared to playing on Clay courts, we would expect the win percentage for playing on grass courts to be 0.1421 percent higher, on average, holding all else constant.

Hard: Compared to playing on Clay courts, we would expect the win percentage for playing on grass courts to be 0.1050 percent lower, on average, holding all else constant.

R-sq vs. R-sq(adj): The R-sq(adj) value is slightly smaller meaning that there could be some unnecessary predictors.

1. A close-up of a chart

   Description automatically generatedNotice, the number of matches increasing typically means that a player is getting deeper into a tournament, as well as their comfort on the surface increases. This makes it important to use an interaction term between matches and surface. Below is this model, interpret the hard surface interaction coefficient assuming a player has played 10 matches on hard courts. Comment on the R-sq value.  
     
   A table of numbers and symbols

   Description automatically generated with medium confidence

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 0.0719 lower, on average, holding all else constant.

R-sq: About 75.71% of the variability of win percentage is explained by the multiple linear regression model. This is a pretty good value, and the best one we have seen so far.

1. Given the normal probability plot of residuals and the residuals vs. fitted values plots, does this model seem to meet the necessary conditions to be valid to use?

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, and there isn’t any curvature so it is linear.

1. Use the ANOVA table below to comment on the overall quality of the model. A black text with numbers and numbers

   Description automatically generated with medium confidence

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, Aces Per Double Fault, and the interaction between matches and surface is useful for predicting win percentage. F=39.10, P-Value = 0.

1. Use the model to predict the Win Percentage of a player on Grass surface with an EloRank of 45, DoubleFP of 0.051, ReturnPWP of .45, 12 matches played, AcesPDF of 1.01.

WinPct = -.385 + (0.01927 \* 12) + (-0.001474 \* 45) + (0.082 \* 0.051) + (1.639 \* 0.45) + (0.0376 \* 1.01) + (0.0290 \* 1) + (0.01885 \* 12) 0.81