Codebook for Final Project

Prashit Parikh Stats 242

Background Information:

The data used in this analysis was sourced from a data set built by Jordan Goblet on Kaggle.com. Kaggle retrieved the data from the large pool of statistics, leaderboards, and rankings on atpworldtour.org. The original data set was comprised of 8,637 match observations with 27 variables and data on betting odds. However, this was condensed to include a total of 1,593 observations of 21 variables. Incomplete cases and those matches which were in a series other than the Grand Slam were discarded. Match Data ranges from as early as 2000 to as recent as 2016. The 21 variables included in the data set are: Location, Tournament, Surface, Round, Best.of, Winner, Loser, Winner’s Rank preceding play, Loser’s Rank preceding play, Number of Sets won by match winner, the number of games won by both players in each match, and our binary outcome of Retired (0=Match Complete, 1=Match Retired). Categorical variables were presented as names of a tournament, rounds, Surface etc. However, numerical variables were presented in strings of numeric values such as 5, representing 5 games won or 5 sets (for variables Wsets and Best.of). Scores are self-reported by officials who proctor the matches on the ATP World Tour and are then kept in a database run by the Association of Tennis Professionals.

The Data is updated constantly on Kaggle by various collaborators and usually occurs at the end of a tournament year. At this point, Kaggle users and the data set creator then construct a new data set with updated data and put that out to the public for use.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable Name | Variable Description | Type of Variable | Values | Re-Coded Values | Source File |
| ATP\* | This is the Tournament number for the Men’s ATP circuit | Numerical | Numbers assigned by tournament draw | N/A | ATP Men’s Tour |
| Location | Location of the match being played | Categorical | Name of Location | N/A | ATP Men’s Tour |
| Tournament | Name of the tournament | Categorical | Name of tournament | N/A | ATP Men’s Tour |
| Date\* | Date the matches are being played | Numerical | Date in the form 00/00/0000 | N/A | ATP Men’s Tour |
| Series\* | Which series the matches are being played under | Categorical | All Matches are played under the Grand Slam Series for this Analysis | N/A | ATP Men’s Tour |
| Court\* | Court type (i.e. Outdoor or Indoor) | Categorical | Court values unassigned | N/A | ATP Men’s Tour |
| Surface | Surface the court is made out of (Clay, Hard, Carpet, Grass) | Categorical | Surface Values Unassigned | N/A | ATP Men’s Tour |
| Round | Which round of play the match being examined is in | Categorical | No Round Values assigned | N/A | ATP Men’s Tour |
| Best of | Maximum number of sets playable in specified match | Numerical | N/A | N/A | ATP Men’s Tour |
| Winner | Match Winner | This variable is the name of the winner | N/A | Made into a factor variable | ATP Men’s Tour |
| Loser | Match Loser | This variable is the name of the loser | N/A | Made into a factor variable | ATP Men’s Tour |
| WRank | The current ATP rank of the match winner before beginning play in the specified tournament | Numerical | N/A | Made into a numerical variable | ATP Men’s Tour |
| LRank | The current ATP rank of the match loser before beginning play in the specified tournament | Numerical | N/A | Made into a numerical variable | ATP Men’s Tour |
| WPts\* | The current ATP points amount of the match winner before beginning play in the specified tournament | Numerical | N/A | N/A | ATP Men’s Tour |
| LPts\* | The current ATP points amount of the match loser before beginning play in the specified tournament | Numerical | N/A | N/A | ATP Men’s Tour |
| W1 | Number of games won in 1st set by match winner | Numerical | N/A | N/A | ATP Men’s Tour |
| L1 | Number of games won in 1st set by match loser | Numerical | N/A | N/A | ATP Men’s Tour |
| W2 | Number of games won in 2nd set by match winner | Numerical | N/A | N/A | ATP Men’s Tour |
| L2 | Number of games won in 2nd set by match loser | Numerical | N/A | N/A | ATP Men’s Tour |
| W3 | Number of games won in 3rd set by match winner | Numerical | N/A | N/A | ATP Men’s Tour |
| L3 | Number of games won in 3rd set by match loser | Numerical | N/A | N/A | ATP Men’s Tour |
| W4 | Number of games won in 4th set by match winner | Numerical | N/A | N/A | ATP Men’s Tour |
| L4 | Number of games won in 4th set by match loser | Numerical | N/A | N/A | ATP Men’s Tour |
| W5 | Number of games won in 5th set by match winner | Numerical | N/A | N/A | ATP Men’s Tour |
| L5 | Number of games won in 5th set by match loser | Numerical | N/A | N/A | ATP Men’s Tour |
| Wsets | Number of sets won by match winner within the match | Numerical | N/A | N/A | ATP Men’s Tour |
| Lsets\* | Number of sets won by match loser within the match | Numerical | N/A | N/A | ATP Men’s Tour |
| Comment\* | This is just a space for defining what happened in a game, such as if a walkover or retirement of match occurred. | This variable is just to inform the data set user about the status of the match being analyzed and is not intended for use in building a model | 1=Complete  2=Retired | 0=Complete Match  1=Retired | ATP Men’s Tour |
| Season | This is a variable based on 4 quarters of the year to determine how the time of year impacts the binary outcome Retired | Categorical | 1 = Winter (December, January, February)  2 = Spring (March, April, May) 3 = Summer (June, July, August) and 4 = Fall (September, October, November) | None | ATP Men’s Tour |
| Retired | This is a binary predictor of whether a match will be completed or not, used as the outcome variable in this analysis | Binary | Comment-1 | 1=Match is Retired  0=Match Completed |  |

\*Indicates Variables Not Used in Analysis