Creating a Tennis Betting Pipeline

Emily Wang | August 2022

Agenda

Pipeline Design

Dataset

Streamlit Demo

Future Considerations

Pipeline Design

Data Ingestion

Used a Python script to aggregate the CVS files into one combined dataset

Data Storage

Created a database using SQLite to store the clean and aggregated data

Algorithms

Automated modeling script tested Decision Tree, RandomForest, and XGBoost

Deployment

Created a Streamlit web app to present my model findings

Dataset

Processed Variables

- *Elo_variable*: The Elo Rating is a well known probability calculation that considers player ranking to determine a match outcome. For example, if a player has an Elo rating of 1,800 and his opponent has a rating of 2,000, the probability of the lower Elo rating player winning becomes is 24.1%. Learn more about it here
    - *elo_loser*= The Elo Model ranking of the loser calculated *before* the match based on the playing history of the two players
    - *elo_winner*= The Elo Model ranking of the winner

Tournament Variables

- *Series* = Name of ATP tennis series (Grand Slam, Masters, International or International Gold)
- *Match Date*=converted into year
- ATP = Tournament number (men bracket)
- Date = Date of match
- Series = Name of ATP tennis series (Grand Slam, Masters, International or International Gold)
- Court = Type of court (outdoors or indoors)
- Surface = Type of surface (clay, hard, carpet or grass)
- Round = Round of match
- Best of = Maximum number of sets playable in match

Streamlit Demo

Future Considerations

Some future improvements to further automate the data pipeline

- Scraping website for updated CSV files using beautifulsoup package
- Implementing Pyspark to handle larger datasets
- Deploying Streamlit app using cloud based data storage

The way to get started is to quit talking and begin doing.

Walt Disney