



Creating a Tennis Betting Pipeline

Emily Wang | August 2022



Agenda

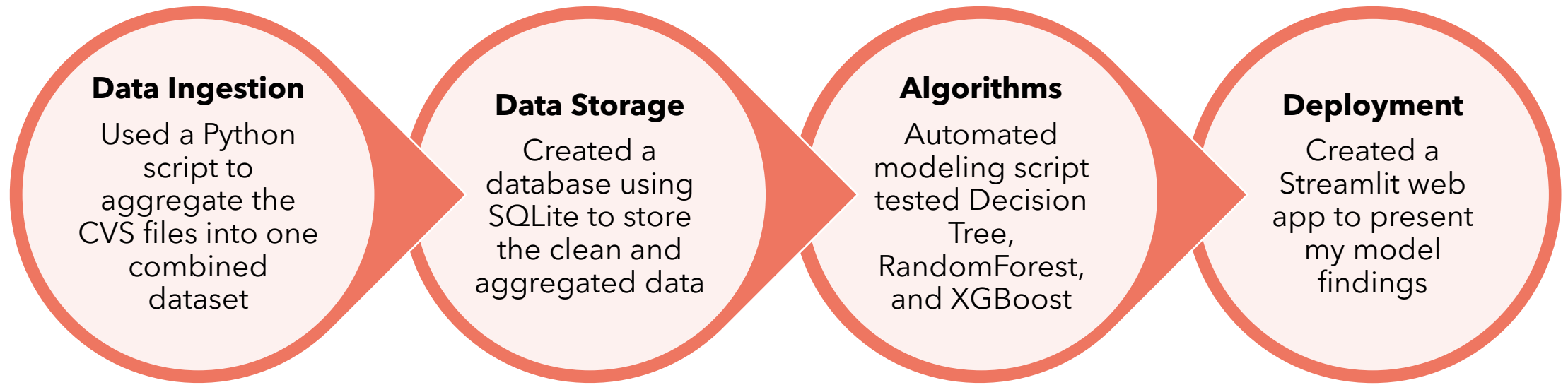
Pipeline Design

Dataset

Streamlit Demo

Future Considerations

Pipeline Design



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

A large blue circle with the text "Streamlit Demo" in white. To the left of the circle is a dashed teal line, and at the bottom right is a small purple circle.


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