Team Analytica

Group 2
Shane Conner, Rosa Ortiz, Dianne Pham, Matias Stanham, Jacquelyn Tencza

Agenda

1. Business Case
2. Desired Outcome
3. Actual Outcome
4. Lessons Learned
5. Demo of the Visualizations

Business Case - The Project

A professional soccer team is deciding whether or not to make an investment in a new stadium. In order to help with this investment, our team is brought on as consultants.

Our team aims to evaluate the impact of fan presence on in-game performance to aid in the decision process regarding future stadium capacity changes. The project will explore the relationship between key variables and develop a statistical model to drive optimal investment decisions.

The outcome will be a report showing the impact of fan attendance on wins.

Business Case - History

- Every year Professional Soccer Teams all around the world make significant capital investments to renovate their stadium. The goal is to improve the fan experience which leads to higher ticket sales as well as increase stadium capacity which leads to increased revenue per game
- While increasing stadium capacity can increase potential revenue, these improvements are slated as being expensive and can risk a sports team's financial future.
- The decision on whether to invest a significant amount of capital is important and can be guided using data analytics and optimization to understand the optimal decision for a sports team
- These effects can now be studied due to the state of sports during the COVID-19 pandemic

Business Case - Approach

- Collect data: from http://fbref.com/ which is publically available. Dataset is provided by StatsPerform
  - Web Scraping required to pull data from the website. This was be done using the BeautifulSoup package in Python
  - Data was stored in a cloud based platform - Azure Data Lake Gen1
- Understand the data: Generalizations around individual league's home and away teams are easier due to a larger sample of games compared to individual teams' fan impact
- Exploratory Data Analysis: Dataset would need to be transformed and aggregated in Python
- Hypothesis testing: Complete testing in Python
- Win Percentage Modeling: Complete fit in R
- Final Dashboard developed using R-Shiny
- Produce a report and present recommendations and findings to our clients

Business Case - Benefits

- Minimize/eliminate impacts to fan attendance and its effect on in-game performance using a proven data-driven methodology

- Decrease the probability of an investment misjudgment resulting in an adverse financial impact

- A standardized report showcasing the effect of stadium capacity and fan attendance on in-game performance for the client, so they understand the relationship

Project Proposal Desired Outcomes

The proposed outcome of the project is to provide a concise report that summarizes the effect of fan attendance on wins for soccer teams. The vision of the project is that this report will be used to assist professional soccer teams in making decisions regarding stadium improvements and renovations.

Result 1: Define Scope & Hypothesis Tests of Games

Result 2: Test and Configure Web Scraper

Result 3: Load Dataset into Cloud base data storage

Result 4: Connect to cloud based data storage via Python

Result 5: Exploratory Data Analysis and Data Preparation

Result 6: Hypothesis Testing

Result 7: Perform OLS & Analyze Final Model

Result 8: Report For Client & Share Recommendations

Project Proposal Actual Outcomes

While the activities remained the same, the actual outcome is a report that predicts whether a soccer match is a win/loss/tie based on different variables (including fan attendance). The vision of this actual outcome will be to assist professional soccer teams in making decisions to affect wins during games throughout the season.

Result 1: Define Scope & Hypothesis Tests of Games

Result 2: Test and Configure Web Scraper

Result 3: Load Dataset into Cloud base data storage

Result 4: Connect to cloud based data storage via Python

Result 5: Exploratory Data Analysis and Data Preparation

Result 6: Hypothesis Testing on VAR and Fan Attendance

Result 7: OLS Regression to Predict Average Stadium Seasonal Win Rate

Result 8: Report For Client & Share Recommendations

Lessons Learned

What went well?

Communication & collaboration within the group went extremely well. Strong leadership and a willingness on every team member's part to contribute are equal reasons why the communication was strong. Each team member brought their own strengths into the project and the utilization of those strengths went well. The data itself was generally fairly clean once it was ingested, and there were very few issues in relation to the quality of the data.

What went wrong?

Ingesting the rich event level data took a lengthier time than originally planned, due in large part to the website viewing our traffic as a bot and then blocking access. To workaround this we had to limit requests to 100-125 records and wait a few seconds between runs. These delays then led to delays in other areas which eventually led to a change in the final outcome (from a report that looks at level of fan attendance and what their effect will have on wins to a report that looks at different variables during a game to predict whether or not a game is a win/loss/tie.

What we would do differently

With more time the proposed objective could have been completed. If we were to do this differently, we would look at a way to bring in the entire dataset via a web scraper that wasn't flagged as a bot, by leveraging FBref's new scraper policy
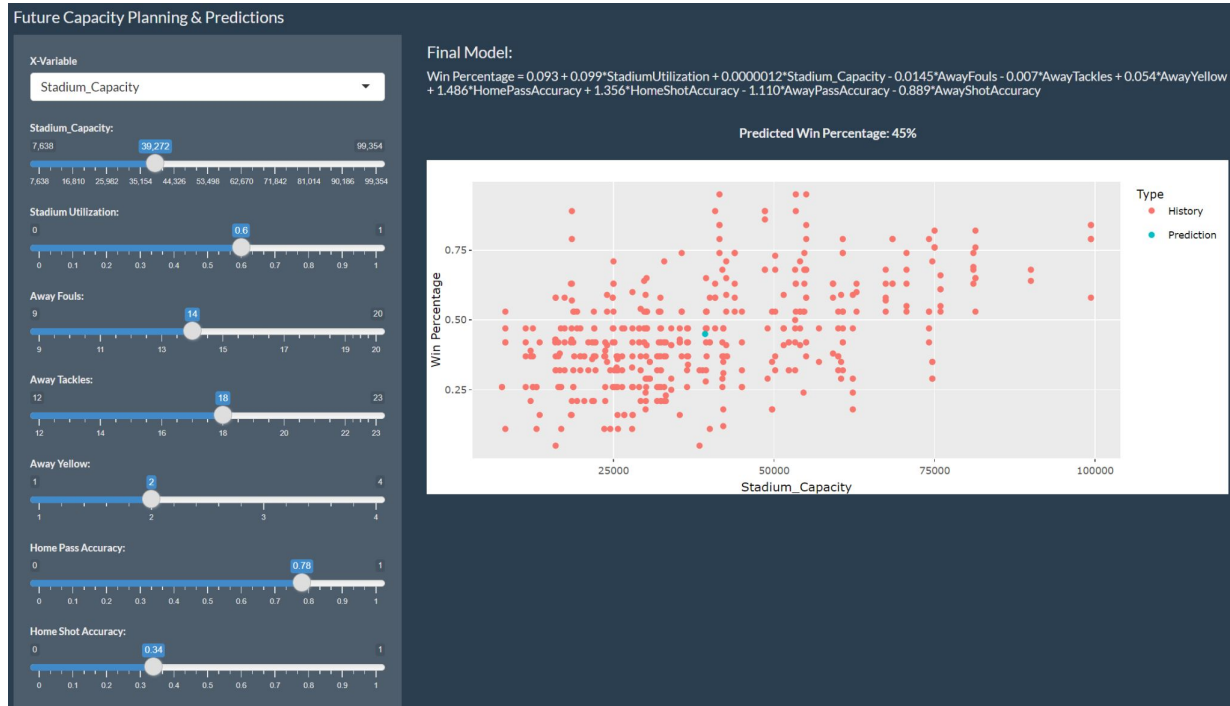
Demo of Visualization

https://team-analytica.shinyapps.io/Final_Report/
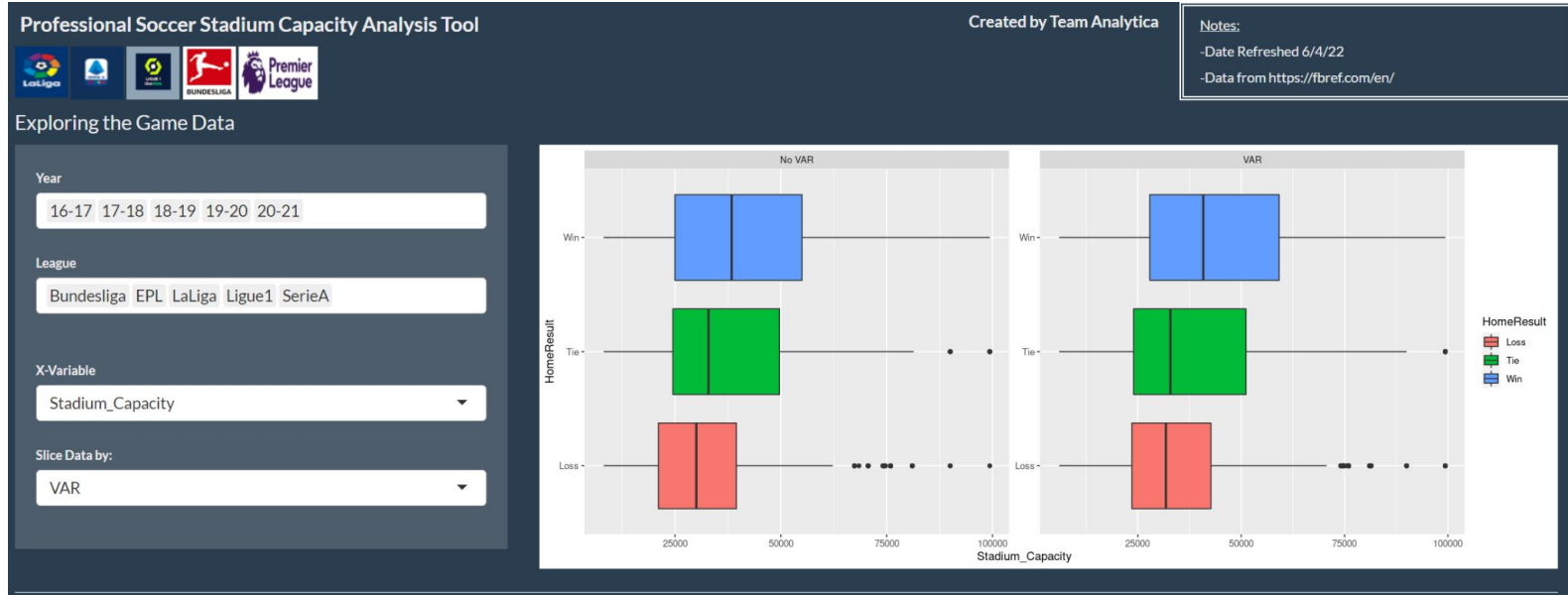
Demo of Visualization

https://team-analytica.shinyapps.io/Final_Report/

References

Albert, J. (2017). Referee Bias in Football. In *Handbook of statistical methods and analyses in sports*. Boca Raton: CRC Press, Taylor & Francis.

International football Association Board. (n.d.). Retrieved March 17, 2021, from https://www.theifab.com/laws/chapter/38/section/134/

Martin, C. B., PhD. (2019, February 18). EPL History, Part 1: Scraping FBref. [Scholarly project]. Retrieved March 17, 2021, from https://chmartin.github.io/2019/02/18/EPL-History-Scraping.html

Pratt, S. (2002). *The Cost of Capital: Estimation and Applications*. Hoboken, NJ: Wiley Press