Ben Katz
benkatz21@gmail.com
Brain Station Data Science Bootcamp
4/4/2022

Capstone Read Me

For my capstone project, I attempted to model transfer market values for soccer players in the big 5 European soccer leagues. These leagues – the *English Premier League, French Ligue 1, German Bundesliga, Italian Seria A,* and *Spanish La Liga* –represent the 5 best soccer leagues in the world.

In soccer, a *transfer* is when one team buys another team's player. The buying team pays a *transfer fee* to the selling team for the rights to this player. A player's transfer market value is an estimation of how much a team should pay to buy a this player.

I chose this project because I love sports, and I wanted to do a project in an area where I had some domain knowledge. Though not my favorite sport, I have found myself getting more and more into European Soccer each year. Soccer has two distinct factors that were not available in other sports-related contexts. Transfer values provided a unique cross-section of sports and business. These valuations are generated similarly to how businesses are valuated, by looking at important statistics, putting production in the context of their environment, and trying to understand and project potential. This gave me a chance to create business adjacent models in a space in which I was more comfortable. Also, because there are 5 major soccer leagues in Europe, there is a larger set of athletes to gather data from than any of the American sports leagues. These two factors, along with the readily available data that comes from all sports, led me to choose this specific problem.

Soccer is a big business. Teams spend millions of dollars every year purchasing players from other clubs. Purchasing these players comes with a huge risk. Spending your whole budget on a failed signing can be disastrous, as can selling a player for far less than their true market value. In professional sports, winning in the margins is crucial, especially for teams that do not have as many financial resources as some of their competitors. It is in these margins that I hope my project will provide some value. The goal is to create models that can help teams better understand the value of their players, make more informed decisions on who to purchase, and ultimately win more games. This is not just for the glory that comes with victory. With wins comes increased revenues, which means more and better players to target for *transfers*, which leads to more wins. In this cycle where every decision matters, I aim to make these decisions easier.

In my submission folder are the following documents.

1. **This Read Me Document**
2. **Capstone Enviroment**
    a. All the packages used in my capstone notebook.
    b. *Plotly* is in here but used sparingly. It is not necessary for running these notebooks
    c. **Again, *Beautiful Soup* is only need if you plan on using the scraping technique for future analysis.**
3. **Capstone Data Scrape Jupyter Notebook**
    a. In this notebook, I go through the process of scraping data from transfermrkt.us. It covers what data is being scraped, and how it is gathered. I used *Beautiful Soup* for my data scraping package. All data in this project is accurate as of 2/27/2022, but the scrape will work regardless of date. The statistics outlined below will not match if you choose to use this scraping formula, so you will have to gather your own statistics.
    b. **YOU DO NOT HAVE TO RUN THIS NOTEBOOK, THE FINAL DATA SET IS ALSO INCLUDED.**
4. **Capstone Excel Data Gathering and Manipulation**
    a. In this excel document, I gather and combine all of the statistics I downloaded from fbref.com. This was done in sequential order to best display the process of combining the statistics from the 2020-21 and 2021-2022 seasons. It contains the following sheets
        i. **2020-21 Stats:** All of the statistical information from the 2020-2021 season for all players across all big 5 European Leagues
        ii. **2021-22 Stats:** All of the statistical information from the 2021-2022 season, as of 2/27/2022 for all players across all big 5 European Leagues.
        iii. **All Stats:** All of the stats describe above from both seasons. Created by joining **2021-22 Stats** with **2020-21 Stats** using *xlookup*. I used player names as my joining criterial
        iv. **Combined Stats:** The sum of all like categories across **All Stats.** These represent the raw totals of all statistics from the start of the 2020-21 season through 2/27/2022.
        v. **Market Values Data Scrape:** The scraped data from **Capstone Data Scrape Jupyter Notebook.**
        vi. **Market Values and Stats:** The combination of all the data gathered thus. I joined **Market Values Data Scrape** with **Combined Stats**, again using *xlookup* and again using player name as my joining criteria
5. **Capstone EDA and Statistical Testing Jupyter Notebook**
    a. In this notebook, I explore the market values of players across positions and leagues.
    b. I conduct statistical tests to determine if there is a statistically significant difference between the market values of the groups outlined above.
    c. I go into an explanation of the statistics in the data set.
    d. I also explore the difference in distributions of selected soccer statistics across positions groups and subgroups. This allows the reader to see the difference in roles between the different players on the field.
        i. The creation and explanation for these subgroups are also explained in this notebook.
6. **Capstone Modeling Notebook**
    a. I engineer features and explain why I think they are important.
    b. I attempt to model player market values across three distinct position groups, defender, midfielder, and attacker, and attempt to determine which statistics are most important. The modeling process is described below.

        i. Make a baseline linear regression
        ii. Introduce engineered features
        iii. Transform target variable into log space
        iv. Use lasso regressions to reduce the number of features
        v. Examine similar statistics with contradicting coefficients
        vi. Find optimal model
        vii. Convert predictions back to linear space and compare them to actual market values
        viii. Note the most important statistics
        ix. Examine other types of machine learning models (KNN, decision tree regressors, and support vector regressors) to see if they worth examining in the future.
    c. I run statistical linear regressions. The reason for this step was to gain a better understanding of which statistics are important.
    d. Discuss the successes and limitations of the models created

7. **Capstone Boosting Notebook**
    a. Here I attempt to use more powerful models to model the entire data set.
    b. This notebook is unfinished, due to time constraints and technical difficulties
        i. I completed on Gradient Boosting Model, which did not serve much better than the models in **Capstone Modeling**
        ii. I started to use XGBoost, which showed promising Preliminary Results
8. **Data Folder**
    a. Final Capstone Data Sheet.xslx
        i. This is the actual data used for my Capstone.
    b. Population.csv/Population Market Values.CSV
        i. Needed for the boosting notebook
9. **Capstone Final Report**
    a. A report on my findings and future steps.