Eish Kapoor & Kunal Bhandarkar
CSE 163 Final Project

Analyzing the Effects of COVID-19 on the Financials of the Premier League

Summary of Research Questions & Results

1. What will be the effect of COVID-19 be on the financials of the British Premier League?
    a. COVID-19 has impacted almost all financially tracked categories negatively. The effect of the pandemic is most pronounced in particular areas that are most impacted by the fact that the league has halted all games. Each club is also suffering a tremendous average loss in revenue, so it is imperative that actions are taken to allow the host of clubs to minimizes losses.
2. How much will Premier League clubs pay for a player of the same quality after the coronavirus pandemic relative to the fees paid before the onset?
    a. Premier League clubs paid an average of 289,831 euros for every overall point in a player's FIFA rating before the pandemic, which after the financial effects of the coronavirus pandemic fell to 210,301 euros for every overall point. So, for the average transferred Premier League player in 2020 whose rating was 74.18, the cost for transferring the average PL player fell from 21,499,663 euros to 15,600,128 euros.
3. What combination of qualitative variables best determines the market value of a player?
    a. From the qualitative features described in the dataset we used, it is clear that the some like "Power-Free Kick" and "Leadership" are of most importance in the overall value of a player. However, the accuracy of the decision tree regression model that we used to predict market value tells us that qualitative variables are not a strong indication of market value.

Motivation and Background

The motivation for this project lies in our passion for soccer but also our interest in how the coronavirus pandemic affects the financial situations of the sports industry, especially in a sport so crucial to the fabric of our global society. The Premier League should be the most financially stable out of all the leagues due to its global reach and massive TV contracts, but there have been whispers that the Premier League is not in the greatest condition either. With this project, we want to take a closer look at a big part of club football- transfers. Transfers are continually happening to improve clubs, but with transfers also being the "least essential" part- because clubs aren't trying to improve, they're trying to stay afloat- we wanted to predict the spending power of the Premier League clubs. As for the last question, our motivation is that market value is very hard to evaluate- just like in stocks, there's no guarantee the price you're buying for is the right price, so the right evaluation of a player's value is invaluable. We want to shed some light on the process and better understand what intangible traits dictate market value.

Datasets

Scraped Data

- TransferMarkt- Transfer Data for the Premier League

    - 2017 - https://www.transfermarkt.us/premier-league/transfers/wettbewerb/GB1/plus/?saison_id=2017&s_w=&leihe=0&intern=0&intern=1
    - 2019 - https://www.transfermarkt.us/premier-league/transfers/wettbewerb/GB1/plus/?saison_id=2020&s_w=&leihe=0&intern=0&intern=1
- TransferMarkt- Player Market Values for the Premier League
    - Numerous Club URLS in the form of https://www.transfermarkt.us/manchester-united/startseite/verein/985/saison_id/2018 , with the club name and year being altered.

Manually Created Data

- Guardian Data- Financial Reports of PL Teams
    - 2016-17 Season - https://www.theguardian.com/football/2018/jun/06/premier-league-finances-club-guide-2016-17
    - 2017-18 Season - https://www.theguardian.com/football/2019/may/22/premier-league-finances-club-guide-2017-18-accounts-manchester-united-city
- COVID-19's Impact on the Financials of the Premier League
    - https://www.lexology.com/library/detail.aspx?g=61487a90-7cf4-4f0e-86d7-df10796b5e94

Pre-Processed Datasets

- Kaggle- Complete Dataset of all players and their attributes from the EA FIFA Series
    - FIFA 15-20 iterations - https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset

Methodology

1. What will be the effect of COVID-19 be on the total revenue of the British Premier League?

This problem's methodology was straightforward on paper, but posed challenges in implementation. The first hurdle was to find a way to predict data in 2020 based on previous years' data. The first thought that popped into my head was to use machine learning, however, a simpler solution entered my mind: we could use linear regression to predict values for each financial category. To do this, I used the two years of data from the Guardian (years 2017 and 2018), and created a 'delta' DataFrame that was the change in each category between the years. I then multiplied each value by 2 to cover the two years between 2020 and 2018 and added the resulting data to the original 2018 data.

The above methodology gave us data on what a predicted version of 2020 club financials would look like, but what about accounting for the coronavirus? That's where the dataset curated from lexology comes in. The site contained information about the loss for particular clubs. Extrapolating that information to all clubs, I was able to get the losses for each club. I then applied those losses and obtained the predicted financials for the premier league in 2020 accounting for coronavirus.

To fully paint the picture and to allow the findings to be used and accessed, I exported CSVs for both 2020 data and plotted bar plots using matplotlib's pyplot library on the three most prominent features of each club: turnover, wages, and profit.

2. How much will Premier League clubs pay for a player of the same quality after the coronavirus pandemic relative to the fees paid before the onset?

The methodology for this problem was tedious and took a great amount of data processing to reach a conclusion. The foundation of our approach lied in us relating the transfer market values not to some natural occurring growth of players, but rather the revenues(turnovers) of the clubs. The clubs' spending power, which is tied to their revenue, is what set the market, not the players' quality. We reached this conclusion from a real life situation occurring with star striker Timo Werner, who at once usually valued at around 80 million euros, after the coronavirus pandemic is now being sold for around 50-60. This, along with other reading we'd did, made us sure that it was the correct move to make our logical foundation that market values and overall revenue was tied inextricably.

To tackle the same quality issue, we realized that the quality of player transferred was not always the same for each transfer window, as the average FIFA overalls for transfers changed by 1-3 points every window, so we couldn't just compare the average costs of the players transferred, because they were at different levels. Thus, we decided to break down the spending by Premier League clubs into cost that the clubs were spending for each overall point, thus to break down the per unit cost the clubs were paying and thus adjust for player quality in answering the research question that stated the same player. So, after deciding to use the cost per overall as our standard metric in seeing the effect of coronavirus on the transfer market, we looked to our financials predictions from question 1 to understand the financial trends from the COVID pandemic. Using the average predicted revenue for 2020 without the pandemic and with the actual average revenue for 2017, we calculated the relation of the cost per overall growth to revenue growth, which gave us a direct relation. Then, using the drop off on the average predicted revenue due to coronavirus, we multiplied that change by the relation of the cost per overall growth, to get the change in the cost per overall. Then we multiplied that new cost per overall by the average quality of the player transferred in 2020 to get the average market value of a transferred Premier League player in the new market.

3. What combination of qualitative variables best determines the market value of a player?

This question was explored through the depth of machine learning. The first big hurdle was deciding what model to use. Going through sklearn's various models for predicting a numerical value, we decided to use a decision tree regression model to answer our problem. The data we used came from Kaggle and provided a list of qualitative features that describe each player, though they were formatted strangely.

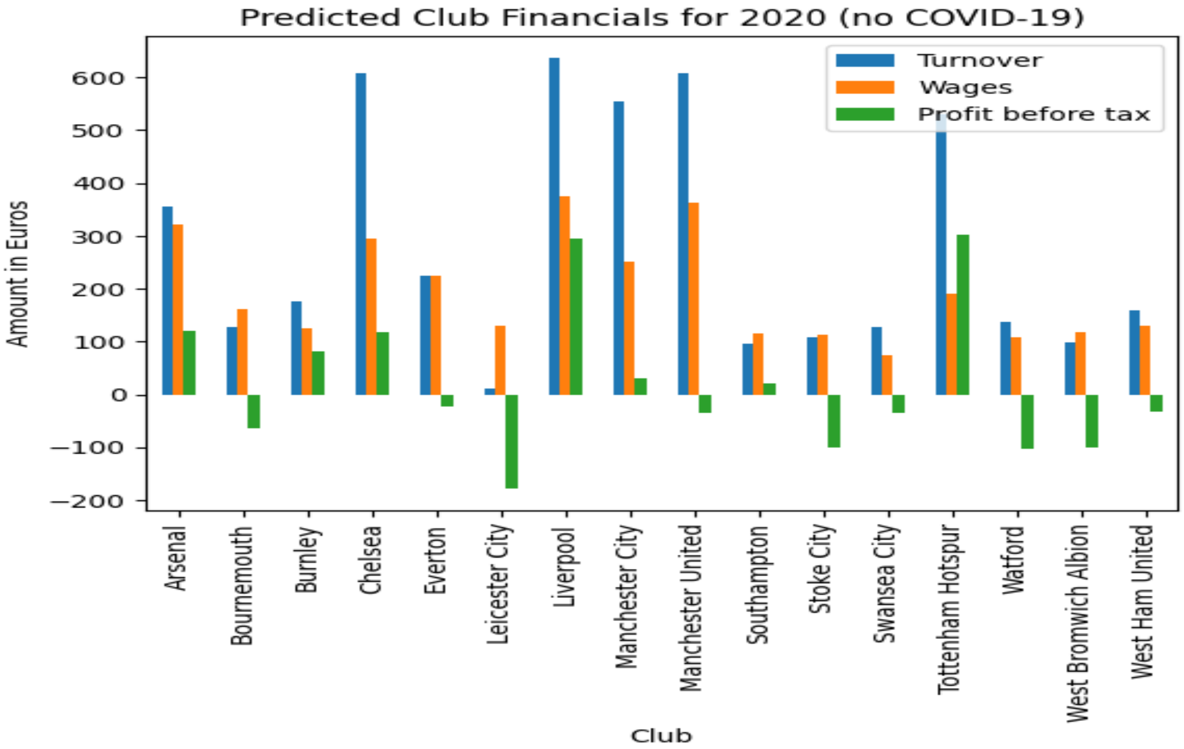
Strangely, the most difficult part of answering this question came in preparing and cleaning the data. This was an extensive process that took a lot of time to complete. To do this, I filtered out columns that we did not need till we just had two: player_traits and value_eur. From there, I had to go through the player_traits feature and split up the multiple descriptions of each player. After adding those two a DataFrame, we performed our version of one-hot encoding where each description is a new column with a value of 1 if the player had that tag and 0 if he did not.

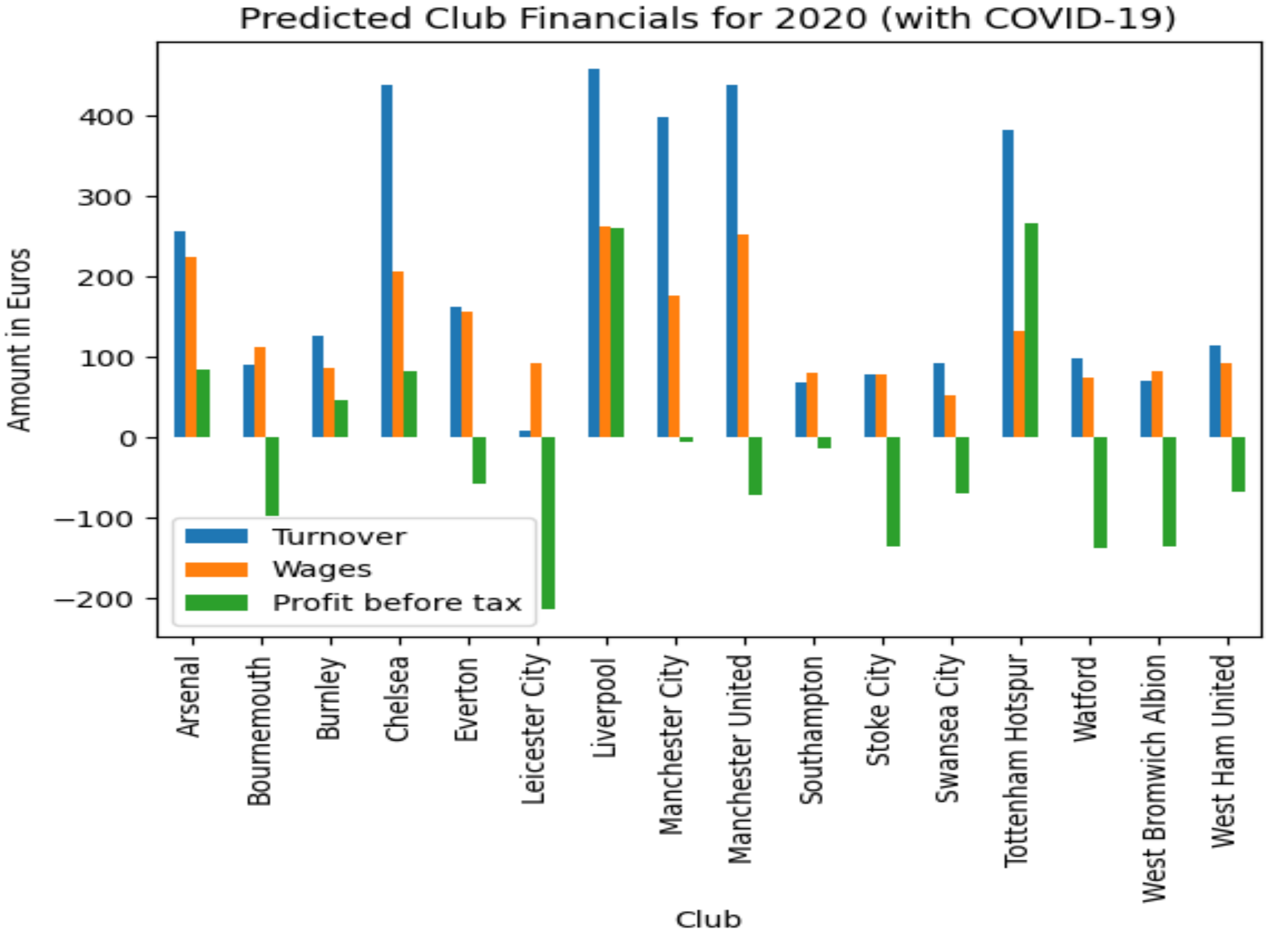
After making sure the data was clean enough, I used various sklearn libraries to split the data into labels and features (labels being the player value and features being all the qualitative features) and then further split the data into 80% and 20% for training and testing. From there, we created a DecisionTreeRegressor model, feed the data into it, and then get the accuracy score from the model. So that the model and it's fields can be accessed I packaged the model, labels, features, and accuracy scores into a dictionary that's returned. In addition to the model information, we saved PNG plots of the most impactful features in determining market value for a player. Moreover, we included a visualization of the model itself just for normal people to be able to visualize how the model looks – this was done using a library in sklearn called tree that allows us to print the tree.

Results

1. What will be the effect of COVID-19 be on the total revenue of the British Premier League?

The effect of the coronavirus on the Premier League was staggering. From the data extrapolated to 2020, there was a steady average pace of growth, so mostly all clubs were well off. However, due to COVID-19, almost all financial statistics took a major hit, most tanking by as much as 24%. Additionally, average profits were down 35 million across the league which is terrifying for the future of the soccer. Below are the major financial statistics (turnover, wages, profits) not accounting for and accounting for coronavirus. The values on the y-axis are in millions.

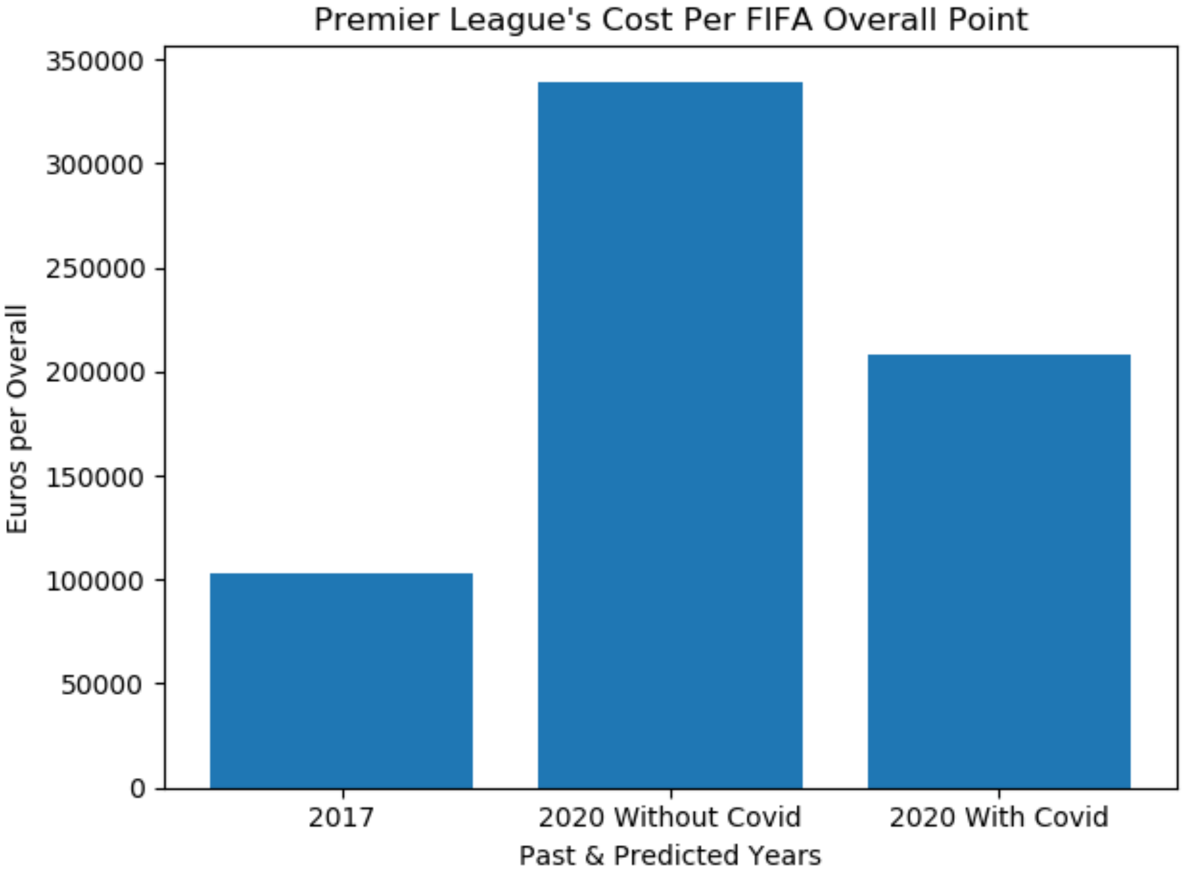

Predicted Club Financials for 2020 (no COVID-19)
Turnover
Wages
Profit before tax
Amount in Euros
600
500
400
300
200
100
0
−100
−200
Arsenal
Bournemouth
Burnley
Chelsea
Everton
Leicester City
Liverpool
Manchester City
Manchester United
Southampton
Stoke City
Swansea City
Tottenham Hotspur
Watford
West Bromwich Albion
West Ham United
Club


Predicted Club Financials for 2020 (with COVID-19)
Amount in Euros
400
300
200
100
0
−100
−200
Turnover
Wages
Profit before tax
Arsenal
Bournemouth
Burnley
Chelsea
Everton
Leicester City
Liverpool
Manchester City
Manchester United
Southampton
Stoke City
Swansea City
Tottenham Hotspur
Watford
West Bromwich Albion
West Ham United
Club

As you can see, the toll of the corona virus is staggering. It was really interesting to see the devastating results of the coronavirus on soccer since we usually think about how we personally are affected by such circumstances. The financial hit of 2020 has strong implications on the future of the league if unchecked. Many companies and organizations cannot survive these losses and it is important that we as data scientists help present issues plaguing all manner of areas.

2. How much will Premier League clubs pay for a player of the same quality after the coronavirus pandemic relative to the fees paid before the onset?

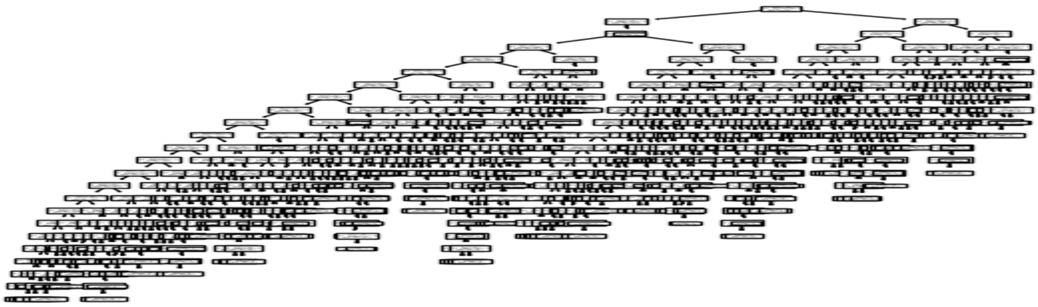
The results were fascinating. We found that in the two transfer datasets that we had 2017 & 2019- that the average overall of players were different and so was the volume. In 2017 the average overall was 80.33, while the average overall in 2020 was 72.75. This overall difference made us decide that we should find a different metric than the one that uses sums. Thus we went for average. To reduce the discrepancy in quality of players moved, we used the cost per overall metric. To get the cost per overall, we took the average fee of the dataset and divided that by the average rating, which gave us two starkly different numbers for the 2017 season and the 2020 season. The cost per overall for the 2017 season was approximately 103,029 euros while the average fee paid was 8,276,666 euros. As for the 2020 season, the average fee paid was 24,681,250 euros and the cost per overall was 339,261 euros. When we made our relation to the financials that were presented in #1, we found the estimated growth rate for the revenue from 2017 to 2020 without coronavirus, which was about 5.95%. This occurred in the same time that the cost per overall rose by 76.43%. From that, we extrapolated that for every percent change in revenue, there is a 12.85% in the cost per overall that the Premier League on average spends in the transfer market. Now, after taking in the effect of the coronavirus pandemic on the revenues of the club and getting that as a percentage, we found that the pandemic will result in an approximately 28% loss in revenue. That, however, is being buoyed by a infusion by the League into each club by a line of credit that for each club, is 25% of their expected revenue before coronavirus. Thus, we offset the -28 by 20, to get approximately -3% loss of revenue. This is manifested in the transfer market by the 3.34% cost multiplier, which got us the result that the cost per overall would fall by about 38.5% to 208,494 euros. To see how this would play itself out for the same quality of player, we calculated the average FIFA rating of the transferred British Premier League player in 2020 which was 74.18, and multiplied that by the pre-COVID cost per overall and post-COVID cost per overall. The pre CPO was 339,261 euros, and post-COVID was 210,301 euros. Thus, for the average player, the cost fell from 25,166,380 euros to 15,466,084 euros.



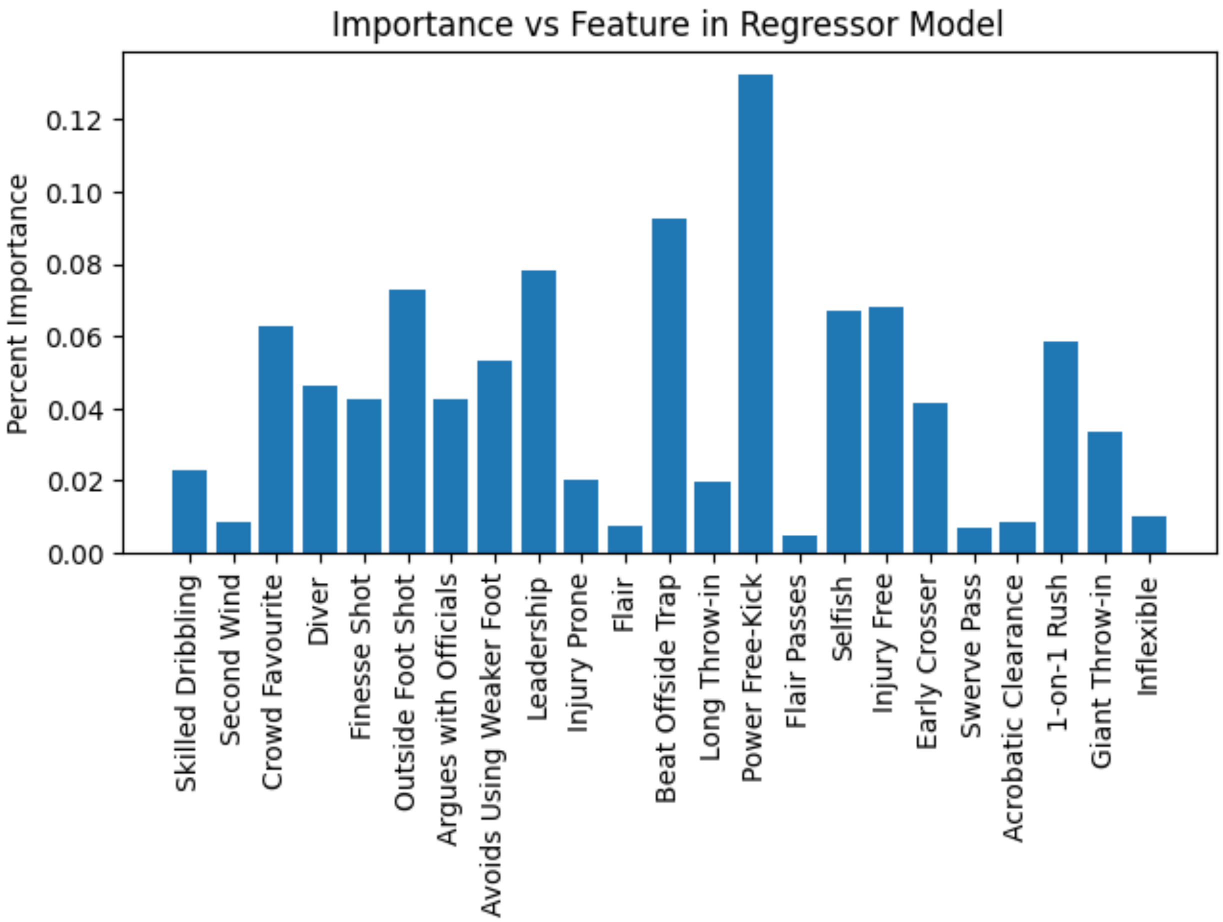
This graph was created by the cost_per_ovr_plot function in q2_data_analysis

3. What combination of qualitative variables best determines the market value of a player?

The results of this question were undoubtedly interesting. Player traits included as many as eight unique descriptions for each player. After feeding the information into our model, we were able predict market value of a player based on the qualitative features. Below is the tree diagram that represents the model. Notice how many branches there are – so many that they overlap in the image. This shows the complexity in building a prediction model on this data.



Additionally, feeding the information into the model allowed us to learn which qualitative variables were most indicative of market value. Below is a breakdown of the most impactful variables.



Unsurprisingly, features like ‘Leadership’ and ‘Power Free-Kick’ are among the top in importance to the model. However, it was interesting to learn that the accuracy score of predicting market value from these features was only around 0.35. This value, while really low is actually telling of how market value is not significantly determined by any qualitative feature. Rather, it is likely made up more of quantitative statistics that are more easily tracked and that make it easier to distinguish a good player from a great one.

This question was really important because we so often hear in sports media that someone is better than another person because they are more “clutch” or that they are a better “leader” or are “selfish”. While these descriptions may help describe someone overall, this question helped prove that they are not really what matters in terms of monetary value of a player. We can now do away with the general statements and make way for more accurate portrayals of players.

Challenge Goals

- Multiple Datasets- We used multiple different datasets, as we used one from Kaggle and scraped the other key datasets using BeautifulSoup. Additionally, we then had to deal with joining the datasets and running data analysis on all of them.
- Messy Data- For our financial data and our player data in the context of the transfer market, we had to use web scraping in order to ascertain the data. Additionally, because the conventions in the data in terms of names were different, we had to also deal with the messy cleaning and re-ordering of the data for successful joins of datasets (i.e. FIFA and TransferMarkt transfers)
- Machine Learning- For our third question, we used the scikit-learn ML module and used the machine learning models along with machine learning practices such as splitting the test data and training data. Additionally, we also checked our data by using our own accuracy metrics.

Work Plan Evaluation

Work Plan

- Web Scraping Script: For this part of the project, we plan on collaborating on the web scraping tool by using GitHub to share code. Our goal is to get done with the script by 5/26 so we can begin the next part of the project. We expect to spend around 3-5 hours each on this part of the project.
- Data Cleaning: We expect data cleaning to take till the end of that week, 5/29. Again, we will use GitHub to share code and verify that we have engineered an appropriate and accurate solution. We expect to spend 2-5 hours each on this part of the project.
- Machine Learning Model: We expect training the machine learning model and reasoning if we are doing so correctly will take the most amount of time as there will be more of a learning curve. We hope to be done with this part of the project by 6/2 and spend anywhere from 4-7 hours each on ensuring we are getting accurate results.
- Visualizations and Result Presentation: We do not expect this part of the project to take too long as we have already got the solutions from the previous steps. We hope to be done with the final project on 6/4 and spend 3-4 hours each making sure we have everything ready to submit.
- Additional Comments: Throughout this group project, we expect to be collaborating frequently through video calls and on platforms such as GitHub to ensure that we can be as streamline in delivering a finished project as possible.

Evaluation

Overall, the plan worked fairly well, but it was in a much shorter amount of time. Due to the protests and political climate during this project, we started work later than expected, which crunched our plan and accelerated our work. For the web scraping, it took about as much time as expected, and due to us having to manually create the csv for the Guardian financial data due to the inability to reach the Guardian’s server, it took about the time we expected. The actual web scraping from TransferMarkt took less time than anticipated. As for the data cleaning, that was an extremely painstaking process, especially for the Machine Learning Model, and it took much more time than anticipated and probably took the longest time to do overall. Just cleaning the data and getting it into the correct format was painstaking. That was the main time crunch for the Machine Learning model, and due to us having less time to work on it than anticipated, took a great deal more time. And as for the visualizations and result presentations, this didn’t take much time at all and we are still currently doing it, but we don’t foresee it being an issue.

Testing

- For our second question, since we are making a prediction, there was no way to completely feasibly test our results, as the effect of the coronavirus pandemic has not made itself clear on this year’s transfer market, with the season being delayed and now restarting. Thus the tests written for this question were to ensure the data parsing was done correctly and so were the results.

Collaboration

In our process to completing this project, we used a plethora of online resources including but not limited to: Stack Overflow, tutorialspoint, YouTube tutorials, and the documentation for the libraries we used. Other than that, all our work was done by us with no help from others.