DSE 309: Advanced Programming in Python

Project Report

Moneyball: Uncovering The Process

Submitted by: Hritik Bana

Roll No.: 19141

Department: DSE

Introduction

This project draws inspiration from the book 'Moneyball: The Art of Winning an Unfair Game', written by Michael Lewis, based on the story of the Major League Baseball team, The Oakland Athletics, which consistently performed well in the tournament despite having one of the lowest payrolls among its competitors.

The Major League Baseball is an American professional baseball league. A total of 30 teams participate in this tournament, one of which is the Oakland Athletics, also known as the Oakland A's.

At the time of the 2001 season, Billy Beane was the manager of the Oakland A's. He was a baseball player who was tagged as an up-and-coming player by the recruitment scouts at his debut. Although he could not become a superstar as a player, having experienced the recruitment and judging criteria in the MLB teams, he came to believe that these methods were not efficient. After his retirement, he became the manager of the Oakland A's. The Oakland A's invariably had the lowest to second lowest payrolls among all its competitors. All the players who suited the traditional judging criteria were signed by the teams willing to pay more, leaving the groups like Oakland A's to compete with whatever was left. But when Billy became the manager, he was ready to use the scientific approach of Data Science for creating the team, and to help him with that he hired, Harvard graduate Paul DePodesta. The rest of the story is how Oakland A's succeeded by using their efficient methods.

I read Moneyball in the summer of 2021, and I found it very exciting. Since I did not know the terminologies and rules of baseball, I had to spend some time reading about them before starting the book.

In the story, there are mentions of a few strategies by Paul, like winning 95 games to qualify for the playoffs and the importance of On-base percentage, but the process behind arriving at that conclusion is not there in the book. Through this project, I tried to apply the data analysis aspect to its dataset and try to work out these strategies using ML.

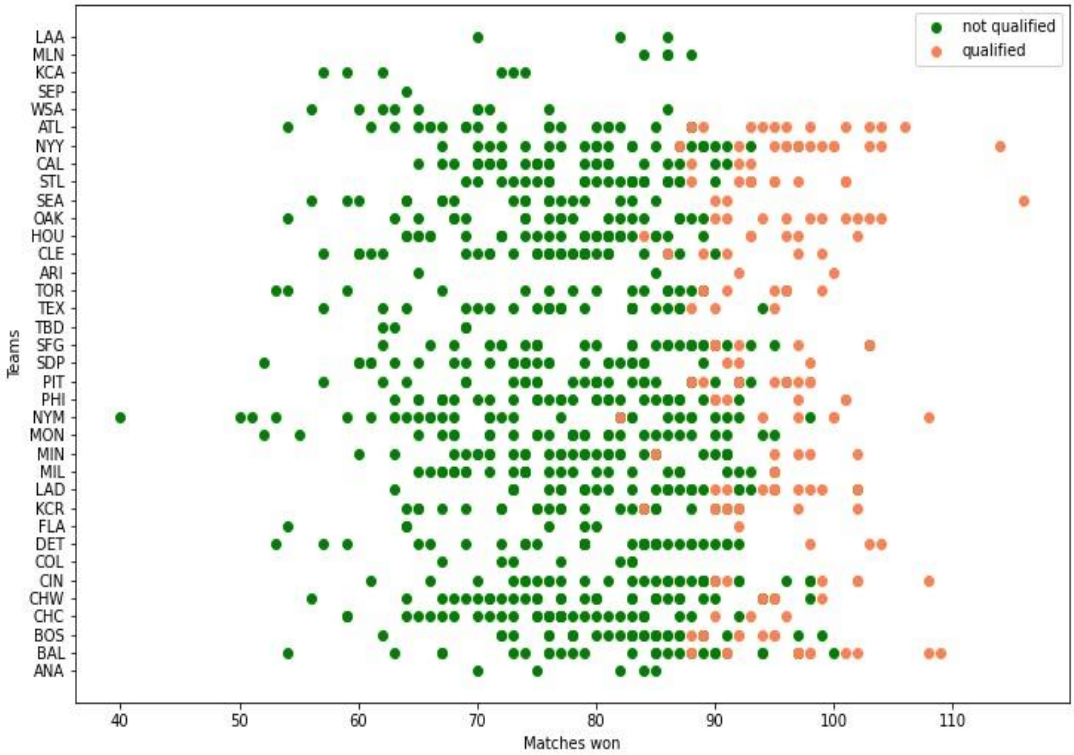
Data

I got the dataset for this project from Kaggle.

Approach and Methodology

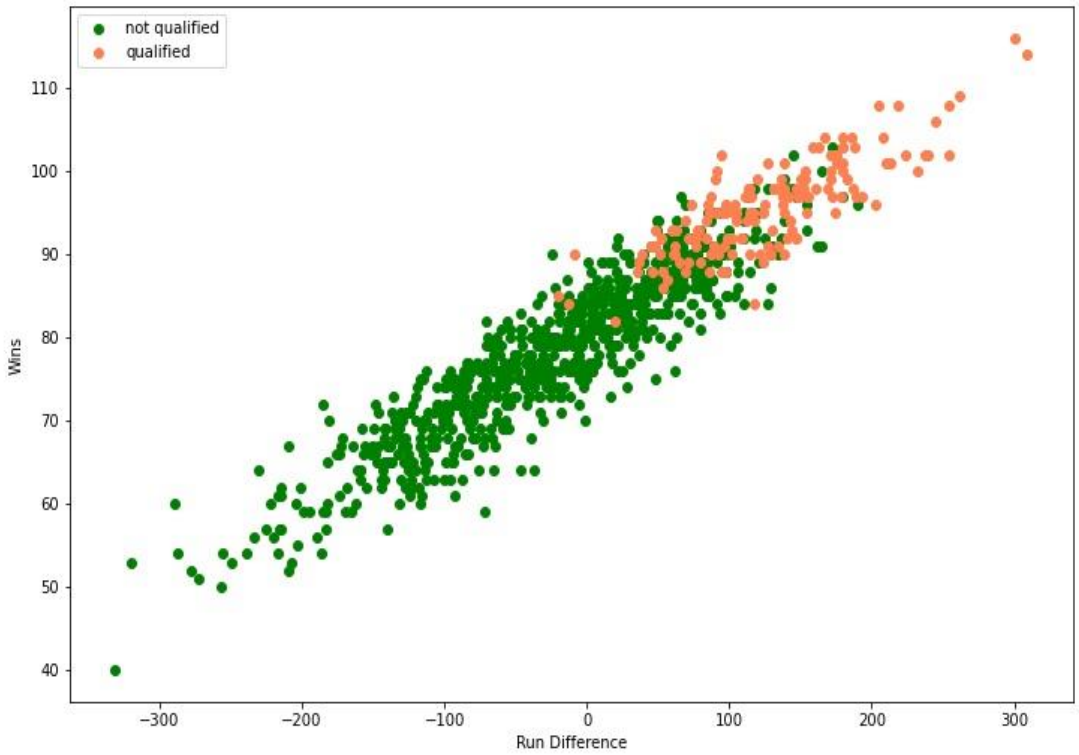
The dataset from Kaggle contains data till 2012, but the Moneyball analysis performed by Billy Beane and Paul DePodesta of the Oakland A's for the first time was before the 2002 season. So we first stored the original dataset in a different data frame and kept a separate data frame with the data till 2001 only.

The first step for the Oakland A's for deciding their strategy was to determine what target they wanted to achieve and how their new techniques of Data Science and ML could help them do it? Data Science and ML requires a large amount of data for accurate predictions; for instance, it can't say if the result of a coin toss will be a head or a tail, but it can say that for a large number of tosses, the number of heads and tails would be nearly equal. Similarly, the ultimate prize of a league or competition is for the winner, but to win, a team has to perform well in the playoffs, in which there is typically a minimal number of matches for each team. But if we look at the league stage of a regular season, there are 162 games. The target for all the teams is to win more games and qualify for the playoffs. It provides a comparatively better sample for answering some questions related to that, for example, how many games are needed to qualify for the playoffs. To answer this, I used a scatter plot of teams with their matches won, with different colour codes for depicting the qualified teams and the teams that did not. Since we have panel data, there will be multiple points for each team corresponding to different seasons. And we get a scatter plot like this:



Through this plot we get an idea that winning at least some 95 matches provides a good chance of qualifying for the playoffs. We used the quantile method to establish the fact further and calculated that winning 95 games provides a 98% chance of qualifying for the playoffs.

Now the target was set, win at least 95 games. Baseball has many analogies to cricket. One of them is that, in a match, the team which scores more runs wins. In short, the run difference (runs scored - runs conceded) for the winning team is larger. On calculating the correlation matrix, the correlation coefficient between run difference and wins was 0.9385. We can say that the number of wins is also more significant for teams with a higher run difference. I also made a scatter plot between run difference and wins and got



To connect this with the earlier finding, we need to predict the run difference needed to win 95 games. I used a linear regression model, as the plot above also represents a linear relationship between run difference and wins.

I got the OLS regression equation,
$\hat{RD}$ = -673.5757054841915 + 8.32794581 * W
On plugging in the target for wins = 95, we solve the equation to get the predic ted Run difference as 117.57914646580844

Through this, the Oakland A's strategy was to select a team that would score nearly 118 runs more than they concede. They needed to know the players they had to target for signing to play for them. Jason Giambi, a perfect batter that they had, was leaving the Oakland A's, and now they had to bring in batters whose stats would ensure that the team achieved a run difference of at least 118 runs and one component was to figure out how to score more runs.

Traditionally, the way a batter (analogous to a batter in cricket (Sidenote: The terminology has changed from batsman to batter in cricket as well)) or pitcher (equivalent to the bowler in cricket) was judged was just by the conviction of the scouts, while they saw them play. The scouts themselves were retired

baseball players, and a lot of times, they judged players just by their appearance. The only stats that were somewhat considered were superficial of the actual abilities of the players. This included the number of home runs, stolen bases and batting average. They were oblivious to the stats that spoke for the so-called 'unfit' players. However, it was Oakland A's Paul DePodesta and Billy Beane who argued that batting average is overvalued, while On-base percentage and Slugging percentage were more important indicators for the batting ability. By definition OBP refers to how frequently a batter reaches base per plate appearance, this can be thought of as the ability of a player to take singles and identify and leave wide balls in cricket.

I regressed runs scored on batting average, on-base percentage and slugging percentage, and noticed that the OLS coefficient of batting average was negative, while the same for OBP and SLG was highly positive, which proved that batting average was indeed overvalued.

I obtained my final regression equation for runs scored as

$\hat{RS}$ = -804.6270610622403 + 2737.76802227 * OBP + 1584.90860546 * SLG

Similarly I predicted the runs allowed i.e. RA using Opposition On-base percentage (OOBP) and Opposition slugging percentage (OSLG) and got

R = -837.3778886133363 + 2913.59948582 * OOBP + 1514.28595842 * OSLG

Considering that much focus was given to recruiting batters, I assumed that the OOBP and OSLG of Oakland A's will remain the same as it was in the previous year, and predicted how many runs will Oakland A's concede in the 2002 season. It came out to be 635.4394172160479.
Given that they will allow nearly 636 runs in the season, to keep a run difference of 118, they needed to score 754 runs in the season.
This was their player recruitment strategy; they looked for players who would bring the OBP and SLG to a level where the runs scored, as predicted by our OLS equation for RS, will be 754.
So they focussed on OBP and SLG while recruiting the players. Since they knew that all the other teams would consider batting average, home runs and stolen bases for their recruitment, they were confident that they would get the players they wanted at a low price, which they eventually did!

After this I collected the 2002 season data for Oakland A's and compared the predictions of my model with the actual data.

Conclusions

In 2002, given the OOBP and OSLG my model predicted the runs allowed by Oakland A's is 661.8917574504227, close to actual runs allowed i.e 654.
Given the OBP and SLG, the model predicts the runs scored by Oakland A's is 808.1568160466265, again close to the actual 800.
Then I checked the predicted number of wins for this run difference and it was 96.35118968785162, which is near to the actual games won i.e 103.
Oakland A's qualified for the playoffs with a league ranking of 1, but could not win. Billy Beane and Paul DePodesta believed that in a 162-game season, luck would be compensated for and it won't really affect the final outcome, while skills will be the ultimate factor. So, they focussed on preparing a team that could qualify for playoffs, now what happens in playoffs can't be predicted. It provides a very small sample size and through the course of those knockout stages, that luck factor; just a single moment of brilliance can turn any game towards any team. So they believed that their method will work and take them to the playoffs, and it did.

Well, not everyone enjoyed seeing Oakland A's spending less and winning more, especially the teams with very big payrolls, and the scouts whose traditional ways of recruitment were being challenged. They would rather vent their frustration by commenting that what the Oakland A's did, was just a fluke, and their method does not work. On hearing this, Paul DePodesta would say to himself:-

“I hope they continue to believe that our way doesn’t work. It buys us a few more years.”

References

[1] Moneyball: The Art of Winning an Unfair Game, by Michael Lewis
[2] The Bill James Historical Baseball Abstract, by Bill James
[3] Dimitris Bertsimas. *15.071 The Analytics Edge.* Spring 2017. Massachusetts I nstitute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Cr eative Commons BY-NC-SA.
[4] Link for data Source: https://www.kaggle.com/wduckett/moneyball-mlb-stats-19622012