Dean Oliver: Four Factor Analysis

Pioneer Valley Tip-Off: Data Analytics

01/07/2023

Introduction

Using the data that we collected from the Pioneer Valley Tip-Off, the Data Analytics team was interested in taking a look into Dean Oliver's Four Factors of basketball. Dean Oliver was a famous statistician who has made several contributions to advancing statistics in sports, mainly in basketball. He is well known for coming up with four factors of winning a basketball game in the early 2000s, which are making your shots (EFG% and OppEFG%), protecting the basketball (TOV% & OppTOV%), grabbing rebounds (ORB% & DRB%), and getting to the free throw line (FT/FGA & Opp_FT/FGA). Using these measurements on the offensive and defensive side (for defensive side, it's opposing team stats) of the ball for each team, we wanted to look further into detail of how this would be applied using the team data that we have gathered.

Data Cleaning

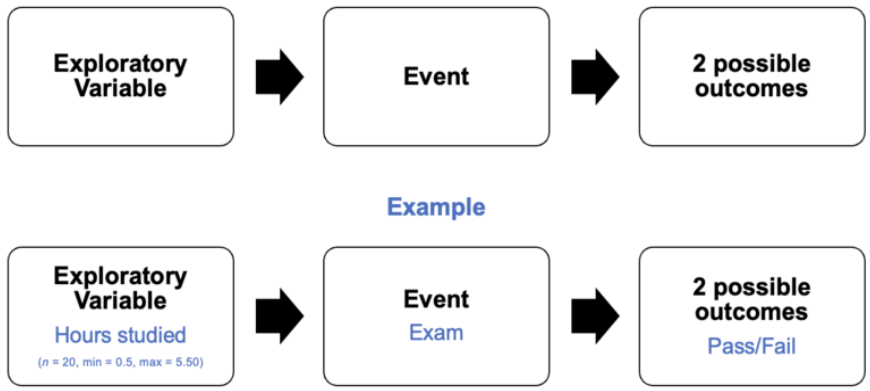
Before we can start doing any analysis, the first and usually the most important step of this process is organizing the data to what is required, since that is necessary to investigate for further analysis. To gather this data, we had to scrape it from Turbostats using a programming language called Python and gathered the team data. This allows us to add the four factors columns that were necessary on offense and defense for the modeling. Once we had the clean data set, we extracted the data points which consisted of the four factors for each of the teams who participated in the Pioneer Valley Tip-Off. Down below is an image of the four factors data for reference.

Name	Efg%	Opp_Efg%	Ts%	Ft/Fga	Opp_Ft/Fg	RebO%	RebD%	To%	Opp_To%
Granby G	0.333	0.371	0.344	0.091	0.113	0.353	0.613	0.505	0.193
South Hadley G	0.371	0.333	0.397	0.113	0.091	0.387	0.647	0.193	0.505
Amherst G	0.359	0.279	0.47	0.692	0.209	0.414	0.786	0.374	0.395
Northampton G	0.279	0.359	0.317	0.209	0.692	0.214	0.586	0.395	0.374
Amherst B	0.362	0.294	0.397	0.191	0.255	0.353	0.588	0.228	0.178
Northampton B	0.294	0.362	0.358	0.255	0.191	0.412	0.647	0.178	0.228
Frontier G	0.305	0.287	0.344	0.141	0.111	0.472	0.606	0.167	0.237
Mahar G	0.287	0.305	0.32	0.111	0.141	0.394	0.528	0.237	0.167
Putnam B	0.58	0.345	0.575	0.107	0.138	0.423	0.811	0.206	0.164
Agawam B	0.345	0.58	0.363	0.138	0.107	0.189	0.577	0.164	0.206
Chicopee G	0.346	0.295	0.331	0.077	0.114	0.591	0.786	0.21	0.399
Chic. Comp G	0.295	0.346	0.321	0.114	0.077	0.214	0.409	0.399	0.21
Chic. Comp B	0.543	0.489	0.578	0.234	0.295	0.348	0.767	0.218	0.145
Chicopee B	0.489	0.543	0.528	0.295	0.234	0.233	0.652	0.145	0.218
Easthampton G	0.343	0.289	0.387	0.275	0.193	0.452	0.6	0.307	0.308
Hampshire G	0.289	0.343	0.338	0.193	0.275	0.4	0.548	0.308	0.307
Easthampton B	0.447	0.392	0.508	0.298	0.098	0.345	0.9	0.203	0.185
Hampshire B	0.392	0.447	0.394	0.098	0.298	0.1	0.655	0.185	0.203

Along with the four factors, we added the numbers 0 and 1, in which 0 indicated the team lost the game and 1 indicated that the team won the game. The reasoning behind adding this column is because we wanted to use the four factors data that we collected from the Pioneer Valley Tip-Off to predict if a team would win or lose a game based on the percentages of the four factor stats in the game for offensive & defensive side of the ball.

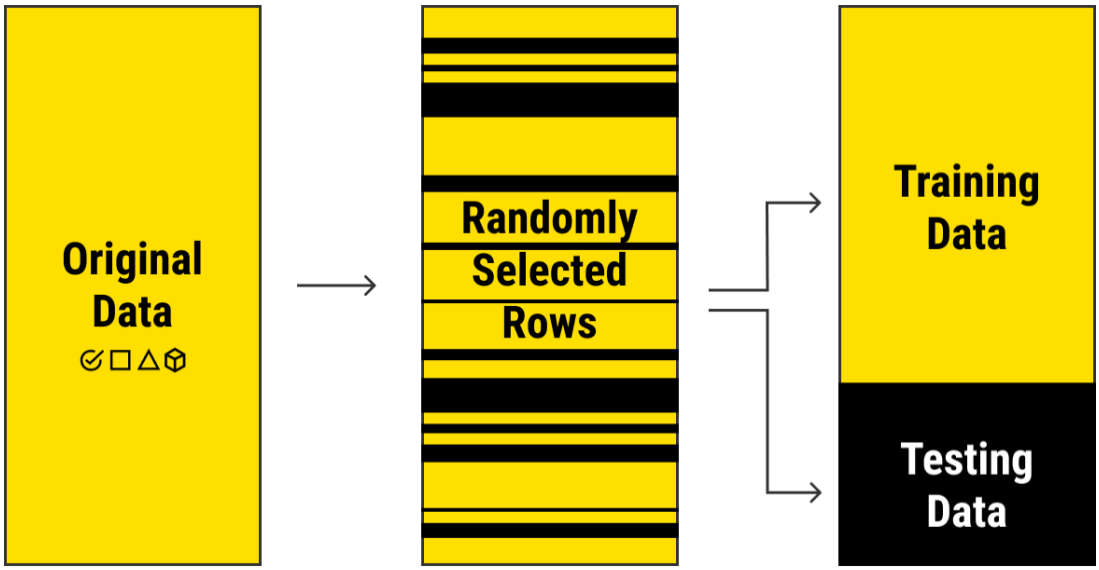
Modeling Process

After cleaning up the data as needed, we decided for this analysis to use a regression model called Logistic Regression. To explain briefly, what regression means in this case is that we are studying the relationship between the different four factor variables and the result of the game. Using the Logistic Regression model works here because we are predicting whether a team would win or lose a game from the given data of the four factors. Down below is an image of a basic overview for better reference.



After going through and choosing the Logistic Regression model as our choice, we then had to use Python to enable us to run this model. When running the model, the first step is to train the data. What this essentially means is that we are teaching the data set to recognize different patterns that have occurred. One simple example is that having a higher EFG% would mean that there's a higher chance that you win the game. In essence, we need to teach the data how to recognize certain patterns that we may already know. In this case, anywhere from 60-80% of the data will be needed to see what patterns can be found from the data set. These patterns are then used to give a prediction of what the result could possibly be using the remaining 20-40% of the data. What we are looking for in this case is whether the predictions made from training the data match the results from these games.

To fully grasp the results and other analysis, we need to be able to run it multiple times. The reason behind running it so many times is due to the randomness of the model. Since the specific teams split between training and testing the data is completely random, running it can produce many different results. In this case, running the model a large amount of times allows further analysis on the patterns that occurred from the model.



Important Variables

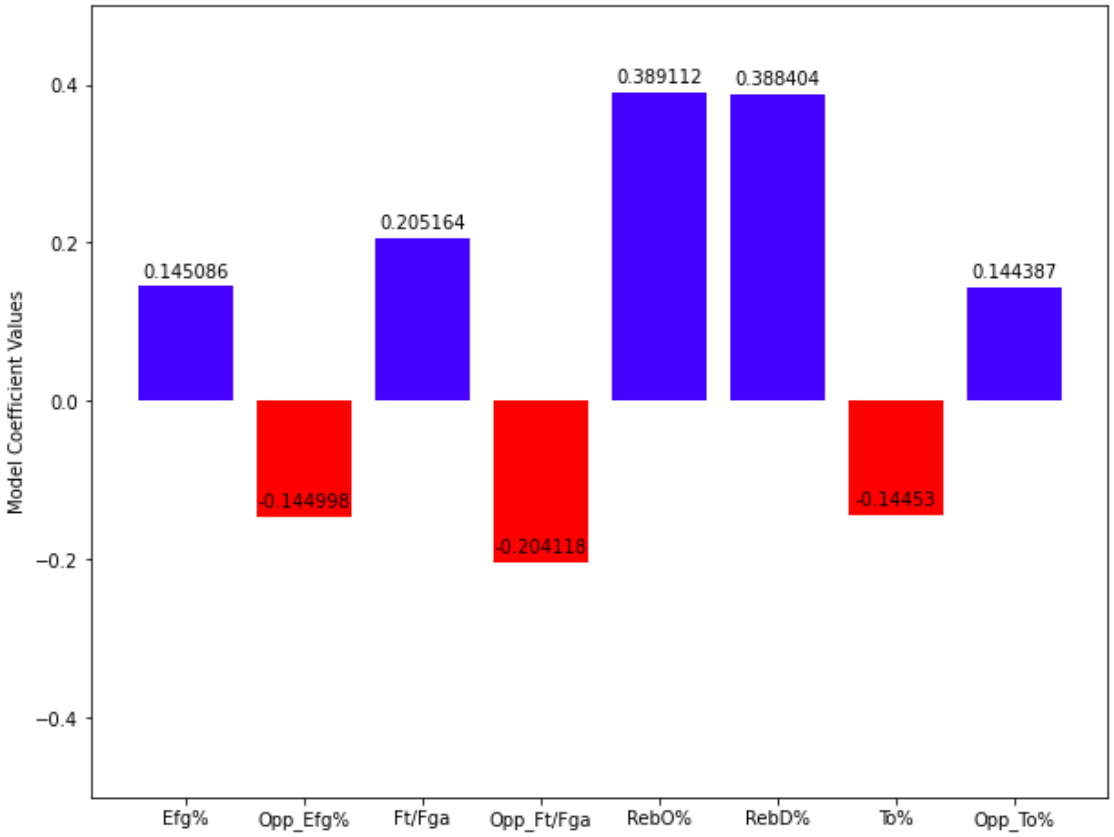
In this section, we are going to investigate the Logistic Regression equation and the important analysis that we can deduce from it. The equation is down below for reference.

Our equation: w = -0.391283 + (0.145086 * Efg%) + (-0.144998 * Opp_Efg%) + (0.205164 * Ft/Fga) + (-0.204118 * Opp_Ft/Fga) + (0.389112 * RebO%) + (0.388404 * RebD%) + (-0.144453 * To%) + (0.144387 * Opp_To%)

$$P(Y = 1) = \frac{e^w}{1 + e^w}$$

$$P(Y = 0) = 1 - P(Y = 1)$$

For us to visualize it better, we will put a graph above to show the importance of each variable on the overall model.



In the Dean Oliver's 4 Factors, he had allocated the following percentages: Shooting (40%), Turnovers (25%), Rebounding (20%), and Free Throws (15%). To our surprise, we have that the best indicator for increasing your chance of winning the game for the teams in this data set was having a high offensive and defensive rebounding percentage. Referring back to the data set, we can see that teams here did not shoot the ball as well, which led to having more opportunities to grab rebounds. Over the course of the game, the ability to keep crashing the boards not only shows that the team is willing to win the game but also it shows that the team is able to keep up the energy over the course of the game. On the offensive side of the ball, having more offensive rebounds can help give your team possessions, which can crucial in key moments of the game while on the defensive side, it can prevent teams from getting more opportunities. Due to the fact that there are a lot more missed shots in this level of play, grabbing more rebounds can help get more opportunities, offsetting shooting percentages.

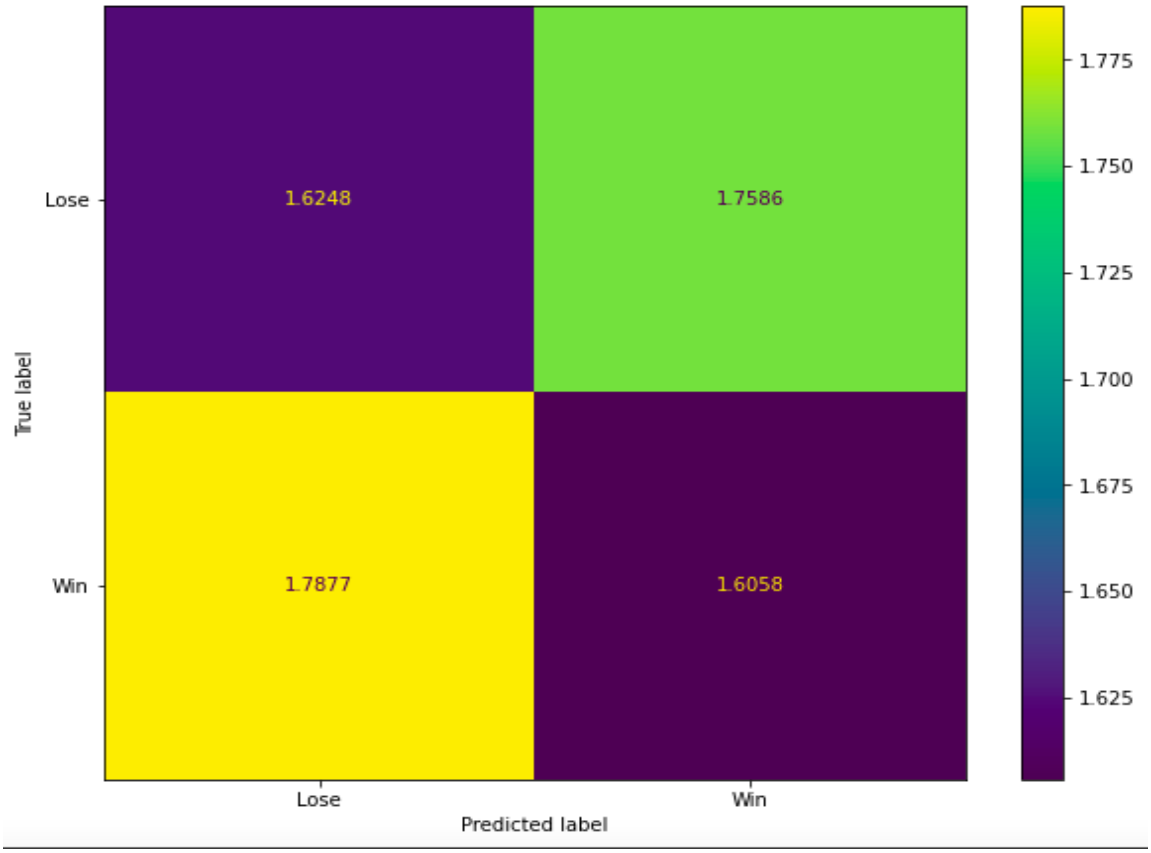
What was also interesting was the fact that the free throw rate was a better indicator of increasing your chance of winning than effective field goal percentage. Looking at the data set, we see that there is an outlier and that Amherst Girls had a .692 free throw rate. They were the only team with a free throw rate above .3. Looking at this, what we can deduce from this is that if we were to remove Amherst Girls from the data set, there would be significant changes in the model due to the amount of teams in the data along with the

significance of each of the variables on the chance of your team winning the game. Since teams at this level are not shooting too great from the field, getting to the free throw line at a high rate and making them would also increase your chance of winning. Since there is the one and one after 6 team fouls and double bonus after 10 in each half at the high school level, drawing more shooting fouls can help your team put up points since shooting percentages aren't as high in the high school level and drawing shooting fouls can potentially be an easy way to generate points if you make your free throws.

Comparing to the Dean Oliver's 4 Factors, his percentages may not be as accurate as it would be in a different level of play. Due to the fact that teams are not shooting the ball well, it is more important at this level of play that they are crashing the boards and getting to the free throw line at a high rate and having good ball control. This isn't taking away from the importance of having a good shooting split but focusing on other factors such as going after loose balls or drawing fouls will lead to better success over the long term.

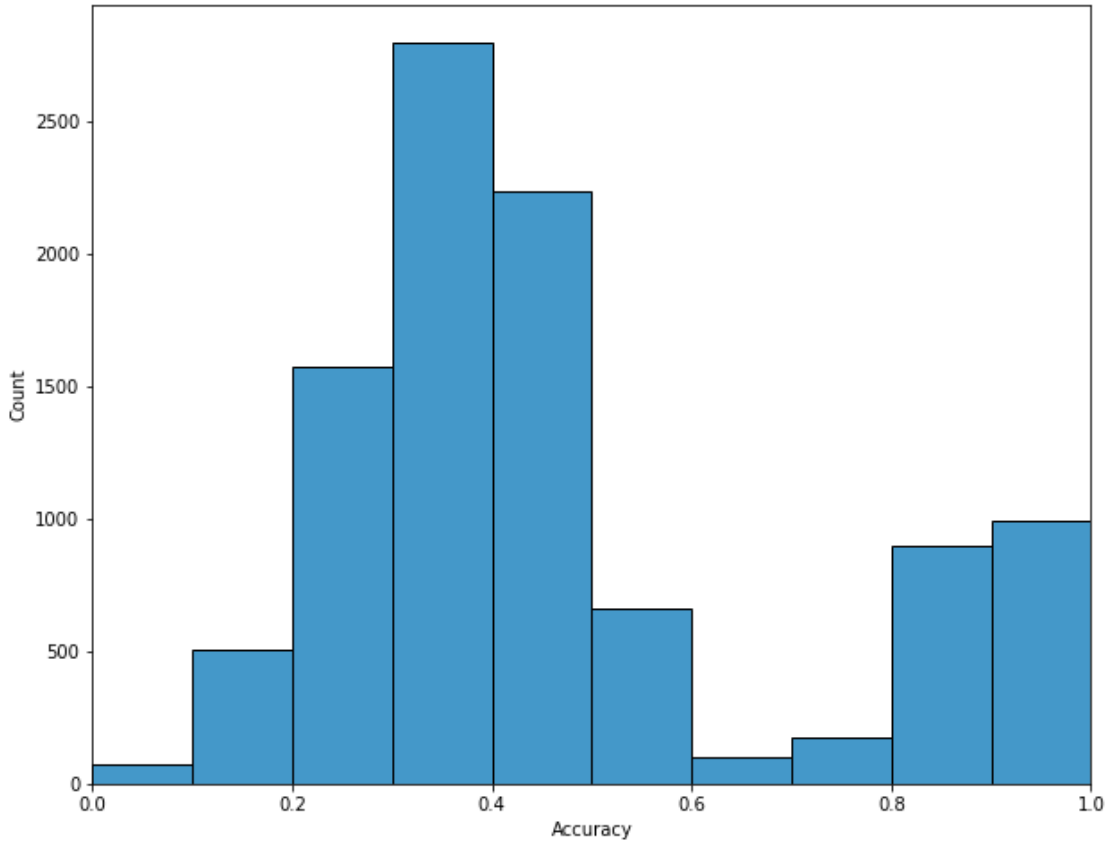
Model Accuracy

This section contains the actual performance of the model itself. What this means is that we are taking a look at how accurate it is in using the four factors to indicate if a team won or lost the game. As mentioned earlier, we needed to run the model 10,000 times to fully get the picture of what we are working with. Attached below is diagram which will show us the average amount of correct and incorrect predictions for each time the model ran.



The purpose of including this diagram is to access and analyze in further detail of how accurate the model is in predicting whether a team wins or loses the game, and which one is the model predicting better with the reasoning behind it. One interesting thing to point out is that it seems that the model is predicting that a team would lose more often than it would win. We see here that the model would predict that a team would lose 50.4% and it can range from 50-51% while for winning, it would be 49.6% and it can range from 49-50% of the time. What this can possibly mean is that there could be some instances where the model is possibly underestimating the team's performance. This can possibly happen if there are lower scoring games and teams wouldn't be shooting as well.

An important measure to take is the overall of the model, since there is not too much of an average. To do that, we take the correct predictions divided by the sum of all of the correct and incorrect predictions. In this case, we get that the accuracy is 47.6% and it can range anywhere from 47-48% in which it is inaccurate 52.4% and it can range anywhere from 52-53%. Although this is not that great accuracy, we have to take into account that the data set is extremely small, so this is a misleading value. What we need to do instead is take a look at the accuracy and how frequent it predicts a certain percent of teams results correct.



What we observe here is that there are a lot of instances where the accuracy is near perfect or exactly perfect. Since there was an extremely small amount of data to work with, we had to somewhat randomize how much data we have to train and test the model as described earlier. Although that is definitely not the best procedure for this issue, the purpose of this was to allow us to see how the accuracy of the model would affected along having more

meaningful analysis of how often the accuracy was between a certain range of values.What we can see here is that there is an absurd amount of times that all of the predictions would be correct and almost 20% of the time, we are getting anywhere from 80-100% accurate. The small sample size allows much higher variance since each data point has a much larger influence on the behavior of the model as a whole. Examples such as Amherst Girls having a .692 free throw rate or Putnam Boys having an EFG% of 58% in this case have a much bigger influence, thus letting us conclude that the small changes in the data can affect the model as a whole.

Conclusion

To wrap it up, we can say that there was a lot left to be desired as there are still a lot of questions we can consider. One issue that was left out was the fact that the gender was unable to be accounted for. It would be more worth it to take a look at if there was more data since the gender can influence stats such as shooting percentages, free throw rates, and more. Another possible issue was the technical side of things as we ran into some issues due to small sample size. We can also improve upon this model as having some sort of cross checking with the predictions would go a long way to having a better model given the scope of the data set. Overall, this analysis was a good start to find a way to predict who would win or lose a game but it needs to be expanded upon. Collecting more data and doing further research on the four factors and it's impact would definitely go a long way to improving upon this model as that would open up more doors to further analysis.