NBA Player Analytics Inferential Statistics/EDA Report

Brief Background of the Data

Given that the goal of the Capstone Project was to find out the differences between US and EU NBA centers and create predictive models for (Points per Game, Assists per Game, Rebounds per Game etc) we begin with a general data overview. It was important to note that since we scraped the data of ALL the seasons played by NBA centers since 1980 (3pt shot introduced in 1980) the dataset, by definition, contained the data from the whole population.

Data Overview

Having imported the wrangled dataset and corrected a few variable types in the dataframe, we began with a simple count of the number of seasons played by US and EU players. We found that 3804 seasons were played by players originating in the US, and 592 by their EU counterparts. It was important to get an idea of the size of each group, since if the size of either group had been much lower than the other (say 10:1) then it would have been difficult to make reliable statistical inferences, however in this case that wasn't a problem.

In the true EDA spirit, we then computed a time series plot of the number of players from each category over the years, and as expected the ratio of EU players in the NBA has grown since the year 2000. This was an important plot, as not only were we able to test the hypothesis that the ratio of EU players has grown, but given that the ratio has been increasing (as a whole) allows us to better compare the 2 groups (from the same 'proportions perspective' discussed above).

The Differences

I chose a violin plot to visualise US and EU players in the most common descriptive statistics (Points, Assists, Rebounds etc) as not only are we able to compare summary statistics between groups (box plot inside the violin), but we are also able to see the distribution of the data (edge of the violin). As mentioned above, given that this dataset represents the populations of EU NBA centers and US NBA centers, by having the violin plots side to side, we made direct comparisons between the 2 groups for each statistic.

The exact details can be found in the Jupyter Notebook, however the main takeaway (perhaps even surprisingly so) is that with regards to the statistical measures compared in the notebook, no significant differences were found between the groups. This meant that from a statistical basis US and EU players (as populations) are the same, even the distributions of values were almost identical. It's important to note, however, that while numerically the populations are the same, the ages at which EU and US players reach peak performance for example cannot be inferred. However, this does mean that we can disregard the 'US or EU' categorical variable (which classifies is the season was played by a player from the US or EU).

Features Analysis/Selection

Having found the answer (perhaps uninterestingly) to the first part of the capstone (the differences between US and EU centers in the NBA) we could move on to the second half - building the models. Before building the models though, we needed to analyse the relationships between the independent variables and the predictor variable. We began with PTS (points per game) as the dependant variable.

We first covered the relationship with the 'Age' column. There was a clear relationship, however given that we had aggregate data instead of discrete player career data, we could not easily track the progression of each individual player. To get a better idea of the relationship we viewed 'Age' as a series of box plots standardized by playing time (per 36 minutes) and found a decline in player efficiency with age.

We also dealt with 'Draft Placing' and 'eFG%' (Effective Field Goal %) columns and found relationships. However, we recognised the downside of heatmaps as a form of 'semi-automatic' feature selection given that heatmaps are only indicative of linear relationships, and we found that most of the features that we (rightfully) assumed would have some relationship with PTS had a non-linear relationship, which the heatmap could not make visible.

The conclusion then, in terms of feature selection, was that plotting scatter plots for all dependant variables and all independent variables is an inefficient time-consuming process (which is why we did not repeat it for the other dependant variables). Instead, we opted to begin training and testing of our linear regression models before spending an endless amount of time on feature selection (especially so given that there are functions that perform feature selection).

Hypothesis Testing

Previously we had concluded that there is no difference between US and EU NBA centers, through our analysis of the violin plots. However, to confirm such a bold (and uninteresting) outcome, we needed a more rigorous approach, such as a hypothesis test. However, we ended up not conducting a hypothesis test (null hypothesis being no difference in the 2 groups, alternate being a difference in the 2 groups) for 2 reasons.

Firstly, the assumption of a Z hypothesis test is that the data is normally distributed, and in fact the data was not normally distributed. Secondly, hypothesis tests are used to infer aspects of a population from a sample with a certain confidence interval. However, since are dealing with population data directly, there is no need to make inferences as all aspects of the population can already be observed directly, rendering the hypothesis test useless.

Conclusion

The possibilities of EDA and statistical inference are endless. The most important thing is using these methods as tools and not exercises in their own right. So, from the perspective of gaining insight into our data to answer our initial question ("what are the differences between US and EU centers playing in the NBA") and helping us build our predictive models, the analysis conducted in the EDA notebook (and described in this document) has served its purpose.