Hi, I’m Martin and let me talk about the explorative data analysis

Our dataset consists of 8582 observations and 34 characteristics, 4 are strings and the rest are floats

It includes some variate NBA players specific statistics such as “usg%” representing**…**

In those specified variables, we observed nice relevant correlations between them helping to understand their roles in the dataset.

For example, the first scatter plot shows that bpm and per have a positive correlation meaning that the more efficient is a player, the more he contributes to the team. Or on the opposite, on the second, the less a player attempts 3-points, the more he secures offensive rebounds available to his team.

For more details, we will now look at the correlation matrix on the following slide

This tells many information such as the fact that we had to handle a lot of variables.

Those seem to be consistent since height and weight are naturally related, as are the years of experience the age of a player.

Another thing is that while some stats seem to highly correlate positively and negatively, most of the features show small or even no correlation between each other.

Now that we talked about the correlation between the features, let’s talk about their distributions.

Unfortunately, none of the variables could be considered as gaussian distributed. Nevertheless, some of them were quite close as for usg% and tm.drtg which are the percentage of plays utilized by a player while he is in the game, as said before, and the team defensive rating.

As you can see, the histogram and the normal distribution of both variables have been plotted as for the qq-plot showing that they are closely gaussian distributed.

We also made a Shapiro-wilk test with an alpha of 0.05 and both p-values were very close to zero but not null, meaning they are far from the significant level but they are better than the one being null.

While on the next slide, features ws and vorp.gm, being the win shares and the value over replacement player, are pretty far from a normal distribution and therefore highly skewed.

So, we can conclude that, the features do correlate, but on another side, mostly differ from the gaussian distribution.

Correlation between variables to check if they are significant for the prediction models

What I did:

Research and planning

Features data understanding and cleaning

Features importance + First insight of models on current data

Explorative processed data analysis