GroupS Analysis Component

James Street, Sarah Fatihi, Sophie Diop

04-03-2024

Project Aim:

The purpose of our study is to see if we can predict actual player transfer value from in-game FIFA stats for premier league transfers in the summer transfer window for the 23/24 season. If we are able to find a relationship between actual transfer value and in-game stats, we could attempt to transfer these findings and predict unsold players transfer price by their in-game statistics.

Research Question:

Can in-game FIFA statistics accurately predict the actual transfer value of Premier League players during the 23/24 summer transfer window?

Variables:

We have both categorical and quantitative variables as our explanatory variables. These are purchasing club (categorical), in-game Short Passing (quantitative), in-game Shot Power (quantitative), in-game Dribbling (quantitative), in-game Tackling (quantitative), in-game Physicality (quantitative), in-game Sprint Speed (quantitative). All of our quantitative in-game stats are on a scale of 0-100 determined by the people working at FIFA.

Step 1: Univariate Analysis

In order to use these variables within our project we first need to examine the data to discover the shape, spread, and distribution of each explanatory variable. Visualizations to assess normality of distributions of our quantitative variables can be created using the R command gf_histogram. This operation allows us to see the distribution of our variable values presented in a histogram. From this point we can analyze if they have a normal distribution, and if not, what we can do to solve this issue. We will then use the favstats command to determine valuable information like mean, median, IQR's, standard deviation and normality. For our one categorical variable (Club), we will use a density plot to see how many players each club signed as well as the average price of those players. Additionally, we will perform a favstats command and a gf_histogram command of our explanatory variable in transfer price. This will allow us to analyze the mean, median, IQR's, standard deviation and normality.

Step 2: Bivaraite Analysis

In this second step, we want to individually compare each of the six explanatory variables to our response variable of Transfer Price. For transfer_price (quantitative) vs our quantitative variables: short_pass, shot_power, dribbling, stand_tackle, and sprint_speed, we plan to use various scatterplots for bivariate

analysis. For transfer_price (Quantitative) vs club, which is a qualitative variable, we plan to use a boxplot for bivariate analysis. We will also create a GGpairs plot to identify which explanatory variables have the strongest relationship with transfer_price using the provided correlation coefficients and scatterplots. In this plot we are looking for the explanatory variables with clear linear relationships and high correlation coefficients with transfer price. Multicollinearity is also a concern we would like to address, and the correlation coefficients between the explanatory variables are a good starting point. However, we also want to conduct VIFs for all the explanatory variables to get a definitive check on multicollinearity. We will remove at least one of the variables if the VIF test shows a value of at least 5. The variables we choose to keep in our model will be used in multiple linear regression. Here, we would check the conditions of MLR to ensure our model is suitable for analysis.

Step 3: Model Building

Our analysis involves several steps to determine the best predictive model for soccer players market value. Initially, we will create various models incorporating both quantitative and categorical variables. We'll assess these models based on their significance, significant predictors, adjusted R-squared values, standard residuals, and whether they meet mandatory conditions. Following this, we'll employ the best subsets method for variable selection to identify the model and optimal number of variables, prioritizing low Mallow's Cp. We will use this method because the forward or backward methods have blinders preventing the new models from re-evaluating old predictors. Afterward, we'll use added variable plots to see if there are any additional variables that will strengthen predictive power and that we should include in our final model. To select the final model, we will consider various contenders: the "kitchen sink" model, the one recommended by best subsets, any models suggested by added variable plots, and the ones we found significant through individual Single Regression Models and Multiple Regression models. The model we choose will be simple, have a high adjusted R-squared, low residual standard error, low Mallow's Cp, and meet all the conditions. If multiple models seem usable, we may prioritize specific characteristics or conduct nested F-tests to determine the most appropriate option. Finally, we'll validate our chosen model by checking for all conditions and ensuring equal variance and normality. Ultimately, our final model will exhibit the highest adjusted R-squared, lowest Mallow's Cp, lowest standard error, absence of multicollinearity, and optimal conditions.