Data Science @UL-FRI

Homework 2: Logistic regression

Matej Kalc

Relationship between Shot type and other variables

Data preprocessing

During data preprocessing the string features ("Competition", "PlayerType", and "Movement") have been one-hot encoded and one of each string feature has been removed. The numerical features were standardized. Binary features remain unchanged.

Correlation analysis

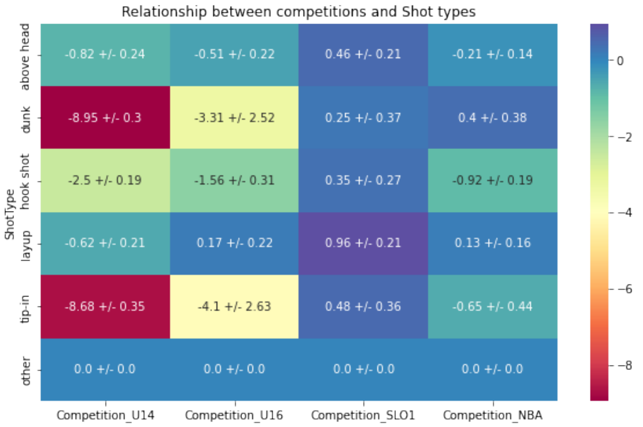

Figure 1. Relationship between competitions and shot types. 95% confidence intervals were calculated using bootstrapping.

In different competitions play a different levels of players. Since "Competition_EURO" has been removed, all the relationships are a comparison with this feature. As shown in Table 1, in "Competition_U14" the least common shots are dunks and tip-ins with other types of shots being the prevalent shot types, when compared to the "Competition_EURO". Also in "Competition_U16" dunks and tip-ins are the most uncommon shots, when compared to the "Competition_EURO", but they are more prevailing than in "Competition_U14". The uncertainty of dunks and tip-ins are large when compared to other 95% confidence intervals, which indicates that those two betas are not a good representation of a real-world relationship. Layup and other types of shots are the most common in "Competition_U16". In leagues of higher rank, "Competition_SLO1", and "Competition_NBA", the different shot frequencies are similar to the "Competition_EURO" frequency type of shots. Above head, hook shots, layups and tip-in are commonly performed in "Competition_SLO1", and dunks are most frequent in "Competition_NBA".

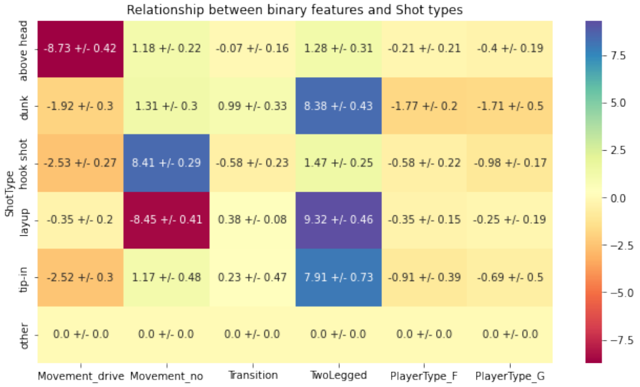

Figure 2. Relationship between binary features and shot types. 95% confidence intervals were calculated using bootstrapping.

Possible movements are "drive", "dribble or cut" and "no movement". Feature "dribble or cut movement" was removed. Above head with "drive movement" is the least likely shot type, when compared to "dribble or cut movements" (see Figure 2). For feature "no movement" layups are infrequent and hook shots are the prevalent shot type. Layups are most frequently performed with a "dribble or cut movements" when considering only the movement features. With "drive movement", the most common shots are other types of shots. There are three player types: F, C, and G. Tip-ins, layups, hook shots, dunks, and above head shots are most frequent with "C player types". We can observe this since all betas for those shots are negative for "PlayerType_F" and "PlayerType_G". Beta parameters are for a magnitude of 10 bigger in movement type features than in player type features. This indicates that movement features have a bigger weight during classification. Feature "Transition" indicates if there was a transition attack or not. During transition attack dunks are the prevailing shot type and during non-transition attacks hook shots are prevalent. Again beta parameters of feature "Transition" are smaller than betas of movement features, indicating a smaller weight for this feature. Feature "TwoLegged" indicates if the shot was performed on two legs or not. Dunks, layups, and tip-in are most common when the player is on two legs. This statement doesn't make any sense, since those shots are performed mid-air. After a quick look through the data, I discovered that the beta parameters are correct since I couldn't find a dunk with the feature "TwoLegged" set to 0. Probably is just the feature name wrong and should be renamed. Feature "TwoLegged" has a bigger weight than player-type features and "Transition" since its betas are bigger. Across all the shot types tip-in has on average the largest confidence interval, which could indicate a false perspective of the real-world correlation between feature and tip-in.

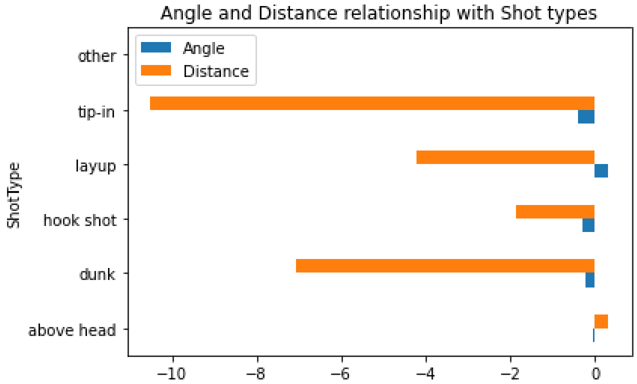

Figure 3. Relationship between distance and angle of the player and shot types.

ShotType	Angle	Distance	B_0
above head	-0.02 ±0.07	0.33 ±0.13	0.62 ±0.41
dunk	-0.2 ±0.07	-7.07 ±0.86	-14.74 ±1.07
hook shot	-0.29 ±0.08	-1.87 ±0.14	-8.47 ±0.5
layup	0.32 ±0.07	-4.2 ±0.2	-2.75 ±0.41
tip-in	-0.4 ±0.13	-10.52 ±1.13	-18.69 ±1.46
other	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0

Table 1. Relationship between distance, angle and intercept of the player and shot types. 95% confidence intervals were calculated using bootstrapping.

Features "Angle" and "Distance" have been normalized for a fair comparison. "Distance" has a bigger impact then angle (see Table 3). A smaller distance than average indicates a higher probability of dunks and tip-ins when compared to other types of shots. We can not state the same for "Angle", where the betas are small and could also be considered insignificant to the outcome. More precise values of betas are shown in Table 1. If all features are set to 0, then the most frequent shots are above head or other types of shots. The most uncommon are dunks and tip-ins. This is confirmed by the intercept (shown as B_0 in Table 1). Again we observe that tip-in has the largest 95% confidence interval.

Multinomial logistic regression accuracy

The accuracy of the model was calculated using the out-of-bag instances of 10 bootstraps. The mean accuracy is 73.9% ±0.02%.

Data generating process

Ordinal logistic regression works better than Multinomial logistic regression when the train size is small enough. The implemented generating process generates an independent variable X with a single feature and a dependent variable Y. X has n instances. Instance in row i is defined as Uniform([0, 1]) + i. Variable Y is defined as 0 for the first third of instances, 1 for the second third, and 2 for the last third. Multinomial and ordinal logistic regression were tested on the same generated data set (1050 instances). For the training set, 50 random instances were used and for the test set, 1000 random instances were used. Ordinal logistic regression performs better then multinomial (see Table 2). When increasing the train set, the multinomials log score decreases and performs like ordinal logistic regression if not better.

	Multinomial log. reg.	Ordinal log. reg.
Log score	2722.49	8.97e-07

Table 2. Log scores of multinomial and ordinal logistic regressions.