Homework 4: Logistic Regression

Stefanela Stevanović
MLDS1 23/24, FRI, UL
63220492

I. Introduction

We were given a dataset that contains data for 5024 basketball shots in real-world basketball game. Our target variable is shot type, which is a categorical variable with 6 categories: above head, hook shot, layup, tip-in, dunk and other. In this homework we implemented multinomial and ordinal logistic regression and used them to get insights into the relationship between the shot type and other variables from the given dataset. At the end, we came up with a data generating process where ordinal logistic regression has a better log score than multinomial logistic regression.

II. Implementation of Multinomial and Ordinal Logistic Regression

A. Data Pre-processing

Firstly, target variable was encoded so that names of the categories were replaced with numerical values: above head - 0, hook shot - 1, layup - 2, tip-in - 3, dunk - 4, other - 5. Categorical independent variables (Competition, Player Type and Movement) were one-hot encoded - converted into a set of binary variables where each binary variable corresponds to a unique category in the original variable. First of the dummy variables for each dummy set was left out as a reference category to avoid the issue of multicollinearity. Independent variables were scaled with MinMaxScaler from Python's scikit-learn library, so that they are all in the range between 0 and 1.

B. Multinomial and Ordinal Logistic Regression

For the multinomial logistic regression, target variable was one-hot encoded. This encoding is necessary because multinomial logistic regression assumes that the target variable follows a multinomial distribution, which means that each observation belongs to one and only one category. This allows multinomial logistic regression model estimates separate coefficients for each category of the target variable using maximum likelihood estimation (MLE). The coefficients are estimated for each category of the target variable, except for one category which is chosen as the reference category, which in our case is the last category - other. The main advantage of using the reference category is that it is used as a baseline for comparison with the other categories, simplifying the interpretation of the model coefficients. Multinomial logistic regression model was implemented without intercept. The coefficients multinomial logistic regression are estimated using the L-BFGS-B optimization algorithm (fmin_l_bfgs_b from scipy) with numerical approximation of gradients. One of the encountered errors was that the optimizer was internally flattening the initial guess which was provided as 2D matrix. We solved this by reshaping the optimizer output in the defined negative log-likelihood function we want to minimize. The first version of the multinomial logistic regression algorithm was too slow (training was taking around 6 minutes), because the for loop was used to find the value of the target variable between the dummies for each observation. This was solved by masking the computed probabilities with the target variable dummies, so that only probabilities which correspond to the true class (value of 1 in the target variable dummies) were saved. This helped to significantly speed up the algorithm and the training is now taking only 8-15 seconds.

For the ordinal logistic regression implementation target variable was kept in an ordinal numeric form, as prepared in data-preprocessing. L-BFGS-B optimization was used to optimize coefficients which measure the impact of each feature on the target variable and differences between the threshold values for adjacent classes in the ordinal logistic regression model.

III. Computing the coefficients

Coefficients of the multinomial and ordinal logistic regression models and their uncertainty was computed with bootstrap technique. The original data set was resampled 100 times with replacement in order to create the bootstrap samples with same number of observations as the original data set. Both multinomial and ordinal model were trained on these subsamples in order to estimate the coefficients and 95% confidence intervals. The results are shown on Fig 1. and Fig 2.

From the range of coefficients in the Fig.1 we can conclude that distance, movement and if shot was performed one-legged or two-legged are the most significant features for predicting the shot type with multinomial logistic regression model. With distance all types of shots become less likely compared to other, except the above head shot, which is is the most often used shot-type when shooting from distance. This is more pronounced for the dunk and tip-in shot types than can be attempted only very close to the basket. Inconsistencies

were found between the results of two-legged and the actual distribution of shot types in reality. According to the results, two-legged shots were more likely to be layups, tip-ins, or dunks, which does not align with what is commonly observed in basketball. Additionally, it is possible that the two-legged shot feature may have been mislabeled and could actually represent a one-legged shot. In transition, dunk, layup and tip-in shots become more likely, while hook shot becomes less likely compared to other. These results align with the real-life basketball situations. At higher angles, layups become more likely, while tip-in, hook shot and dunk become less likely. Compared to EURO competition, layups are more likely to be attempted in SLO1 and NBA competitions. Also, tip-in and dunk shots are the least likely to be attempted in U14 and U16 competitions. As for the player types, compared to centers, both forwards (Player F) and guards (Player G) are more likely to attempt layup and less likely to attempt dunk, hook shot and tip-in. During the drive movement, all shot types except the layup become less likely compared to dribble or cut. This is especially true for the above head shot, which is the least likely shot to be attempted during the drive movement. When there is no movement, hook shot and above head become more likely, while tip-in, dunk and especially layup become less likely to be. Furthermore, we observed that the above head shot is the most similar shot type to the reference category in transition, angle, distance, competition and player type that attempted it. Thus, the type of the movement is the key variable that distinguishes these two shot types.

IV. Multinomial bad ordinal good

For the purpose of generating a data set on which ordinal logistic regression has a better log score than multinomial logistic regression, we generated 100 independent variables with mean 0 and standard deviation 1 using the random.gauss function from the random module. Target variable was generated by randomly selecting the integer between 0 and 5 with equal probability of each integer being selected. We trained both models on the training set of 500 and 2000 observations and tested on the test set of 1000 observations. This was repreated 100 times and the the results are shown in table 1.

TABLE I
Log Loss of the Multinomial and Ordinal Logistic Regression trained on different train set size

Model	Train set size	Mean Log-loss	Standard Error
Multinomial	500	3.054	0.017
LR	2000	1.940	0.002
Ordinal	500	2.207	0.005
LR	2000	2.061	0.002

From the table 1 we can see that ordinal logistic regression has lower trained on 500 observations has lower log-loss than multinomial logistic regression trained on the same number of observations. However, when the size of the training set increased to 2000 observations, multinomial model had lower log-loss, as with more data available it has been able to better capture the underlying patterns and relationships in the data.

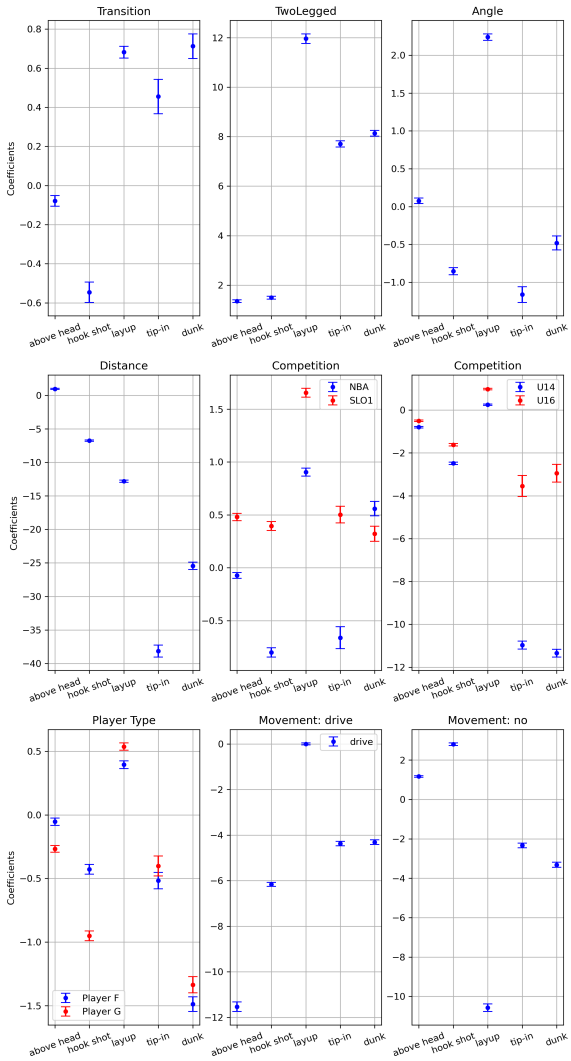

Fig. 1. Estimated coefficients for multinomial logistic regression with shot type 'other' used as a reference category. The higher the coefficient, the more likely is that shot type to be attempted compared to the shot type 'other' with the higher value of numerical variable (Angle and Distance) or with the value of 1 for all other binary variables.

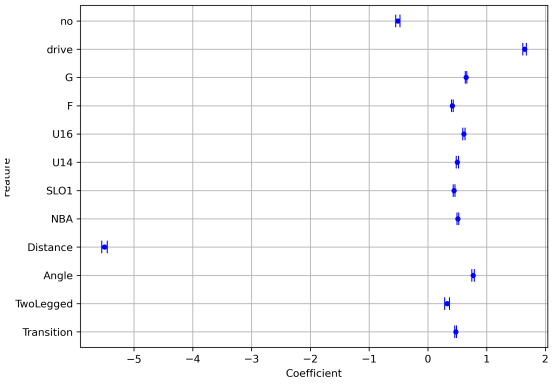

Fig. 2. Estimated coefficients for ordinal logistic regression