ML–DS I: Project III Report

Sebastijan Trojer
ML-DS I 2024/25 , FRI, UL
st5804@student.uni-lj.si

I. Introduction

In this assignment we implemented the multinomial and ordinal logistic regression and used them to explore the relationships in the basketball dataset. We also evaluated the multinomial logistic regression on that same data with repeated 5-fold cross validation. We leveraged the model coefficients to analyze the relationships between the target variable and the features. Additionally we also performed a permutation-based feature importance experiment.

II. Methodology and results

A. Multinomial Logistic Regression

We implemented multinomial logistic regression using log-likelihood. We implemented it as class MultinomialLogReg, where we implemented methods for log-likelihood and gradient computation, and minimized it using the fmin_l_bfgs_b function from scipy. For effectiveness we fixed the latent strength of one of the classes to 0, so we only had $(m-1) \times k$ parameters to optimize, where m is the number of classes and k number of features. We tested the implementation with our own unit tests and on the basketball dataset. To get more robust estimate of model performance we utilized 10 repetitions of 5-fold cross validation on the basketball dataset. The model achieved a mean accuracy and F1 score of $73.51\% \pm 2.02\%$ and $37.95\% \pm 3.52\%$ respectively. The data was shuffled on each repetition to ensure the variance wasn't underestimated. Furthermore, Figure 1 shows the visualized decision boundary on binary mock data.

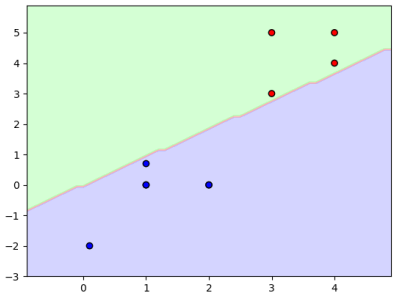

Fig. 1. Visualization of the decision boundary on mock data.

B. Ordinal Logistic Regression

We also implemented the ordinal logistic regression as class OrdinalLogReg, with the same methods as the multinomial logistic regression. Ordinal logistic regression is a model designed to handle ordinal target variables, where the classes have a natural order. Unlike multinomial logistic regression, which requires fitting separate parameters for each class, ordinal logistic regression optimizes a single set of β parameters for all classes and $m-1$ thresholds. To avoid overfitting, one threshold is fixed to 0. This approach makes the model more computationally efficient, as it only optimizes $(m-1)+k$ parameters, compared to the $(m-1) \times k$ parameters needed in multinomial logistic regression. Since we did not have a suitable dataset to evaluate the model on, we do not provide model evaluation results.

C. Application of the multinomial logistic regression

We wanted to utilize the model to provide insights into the relationships between the ShotType and the other variables. First, we computed the odds ratio of the coefficients relative to the tip-in shot type shown in Figure 2. To quantify the uncertainty we bootstrapped the data and recorded the coefficient variability 100 times. The vertical error bars on the plot indicate the 95% confidence intervals. The odds ratios are computed as e^{β_i} and give information about how a change in feature value influences the probability of other classes relative to the reference class.

We can observe a clear distinction when distance increases - the tip-in is the least likely class, except for potentially the dunk class, while the other classes are significantly more likely to occur as distance increases. Another clear division of classes is visible at the two legged feature, which is a binary variable. When a shot was made with a single leg, that increases the probability that it was a layup, hook shot, above head shot or other, relative to the tip-in, but when it was a two legged shot, it is more probable that it was a dunk, rather than a tip-in.

The other features provide some similar insights, although some do have a less distinct influence on the class probabilities, especially when we consider the variablity of the coefficients. Some other features that provide clear distinction are the categorical values of some one-hot encoded features like whether the player was moving, player type and competition type. While plotting odds-ratios

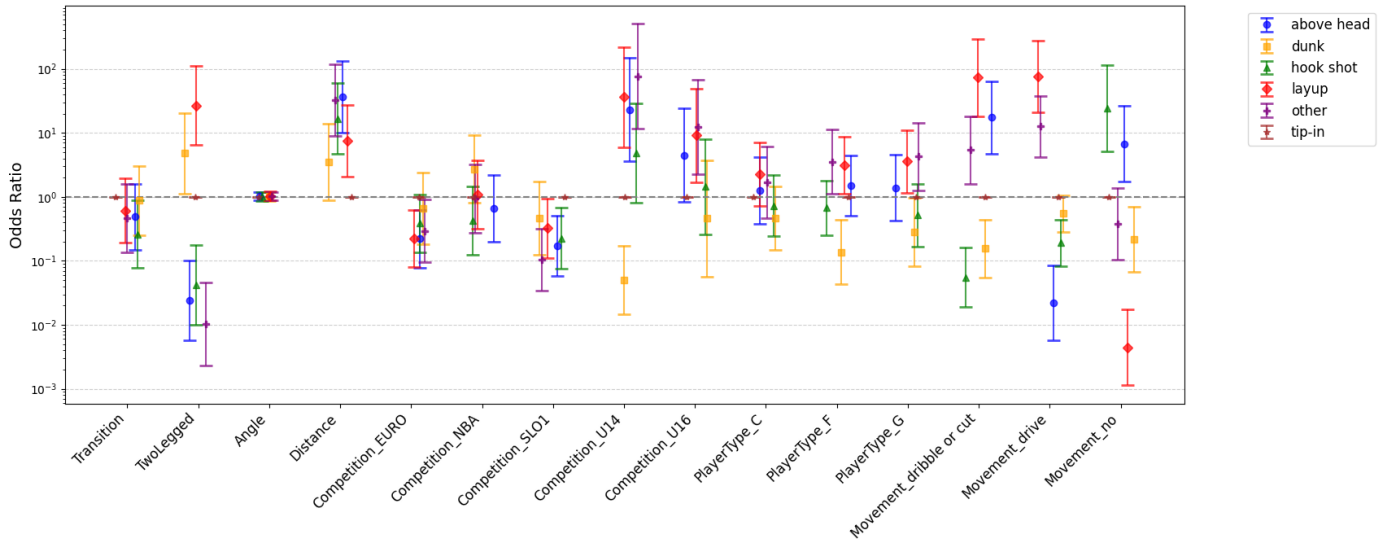

Fig. 2. Odds ratios relative to the tip-in class.

against other reference classes might reveal some additional insights, most of the information is already captured in this plot.

Additionally we performed a permutation-based feature importance analysis, where we logged the difference in performance on original data vs. on a permuted feature column. This way we gain insight on how much a feature helps the model – if the score with permuted feature is a lot worse, that means the feature holds a lot of information about the label. The scores are shown in Figure 3. Notice that the distance has the biggest impact on the score, which makes sense, since the strategy of throwing the ball largely depends on the player's distance from the hoop.

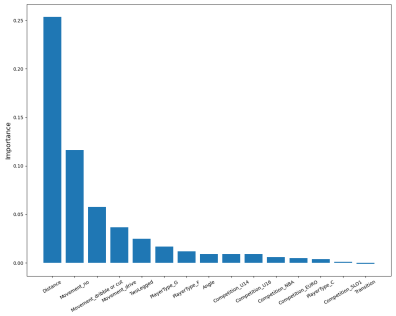

Fig. 3. Permutation based feature importances.

We also explored the predicted class probabilities relative to different distances, since it proved to be the most important feature. Figure 4 shows the probability of the above head shot being made growing with distance (which is also indicated by above head being the highest on distance category on Figure 2, but less obviously), and the probabilities of other classes mostly decreasing with it. This serves as an example on how different visualizations can complement each other to provide a more complete insight into the relationships between the variables.

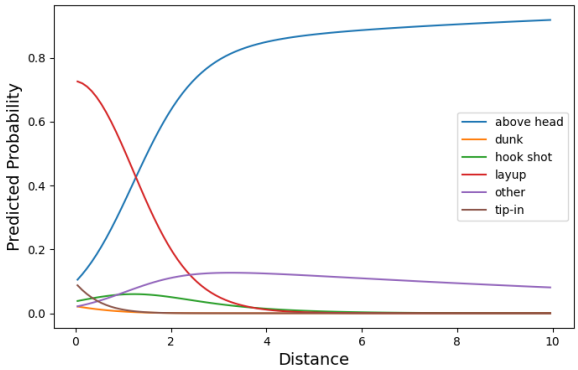

Fig. 4. Class probability predictions vs. distance.

D. Application of ordinal regression

Finally we were tasked with finding a DGP where the ordinal regression would outperform the multinomial regression. We designed a (DGP) via a linear combination of features X with coefficients β and some gaussian noise and threshold-based discretization of the latent variable into ordered classes. The discretization ensures that higher values of $X\beta$ correspond to higher ordinal classes. That allows ordinal regression to outperform multinomial logistic regression because of the explicit ordering we imposed on the data. We also improved that relationship by choosing thresholds based on percentiles. To validate the approach we sampled 100 datasets from this DGP and compared the performance

of both models. The data had 1000 samples, 5 features and 10 classes. The multinomial logistic regression was unable to levarage the ordinal relationships and achieved an average accuracy of $26.65\% \pm 2.01\%$ while the ordinal regression achieved a significantly better result of $93.83\% \pm 1.93\%$, thus validating our approach and DGP.

III. Conclusion

In this project we implemented multinomial and ordinal logistic regression. We considered the use cases of each, utilized the multinomial regression to provide insights in the relationships between the variables, and considered a DGP that fits the use of ordinal regression.