Predictive model evaluation

Biljana Vitanova
MLDS1 2024/25 , FRI, UL
bv7063@student.uni-lj.si

I. Introduction

The purpose of this project is to assess how well different models generalize when predicting shot type in a basketball dataset, using different evaluation metrics. We also analyze the dependence of prediction error on shot distance and how model performance changes under the true relative frequencies of features.

II. Model Comparison and Evaluation

A. Methodology

In the first part, we compare three different models for predicting ShotType. The first two are a baseline classifier, which predicts according to the relative class frequencies, and logistic regression. As the third model, we use a Support Vector Machine (SVM) with a radial basis function (RBF) kernel, chosen due to its sensitivity to the cost hyperparameter. The cost parameter was tuned over a range from 10^{-3} to 10^{3}.

To estimate generalization performance, we applied cross-validation. For the baseline classifier and logistic regression, standard cross-validation was used. For the SVM, two approaches were implemented: (1) selecting the hyperparameter based on training fold performance, and (2) nested cross-validation, where hyperparameters were tuned based on performance on an independent validation set.

Model performance was evaluated using two metrics: classification accuracy and log-score. Due to class imbalance in the dataset, and the fact that the sample is representative of the data-generating process, we used stratified splits for all cross-validation procedures to preserve class distribution across folds.

B. Results and Discussion

A key consideration when using cross-validation is the number of folds, as it impacts both the stability of the performance estimates and the computational cost. To find a good trade-off, we tested three classifier: Baseline Classifier, Logistic Regression, and SVM with training fold optimization, across a range of splits from 2 to 25.

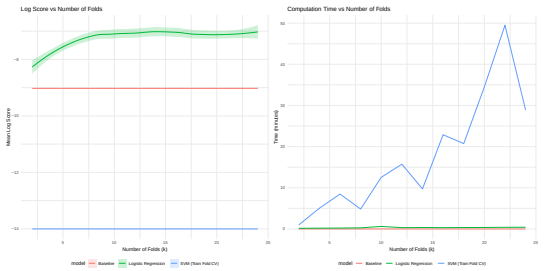

Fig. 1: Model performance and computation time with different number of folds in cross-validation.

From Figure 1, we observe that increasing the number of folds has little effect on performance, with a minor improvement only for logistic regression. However, computational cost increases noticeably for the SVM. Based on this, we chose 5-fold cross-validation for all models, as it provides a good balance between computation time and training set size.

Model	Accuracy	Log Score
Baseline	0.418 ± 0.0039	-9.680 ± 0.101
Logistic Regression	0.723 ± 0.0078	-6.884 ± 0.150
SVM (Train)	0.746 ± 0.0082	-14.002 ± 0.111
SVM (Nested)	0.752 ± 0.0088	-14.002 ± 0.224

TABLE I: Performance metrics for different models.

From Table I, we observe that the SVM model optimized on the training fold achieves nearly the same performance as the SVM evaluated with nested cross-validation. This is somewhat surprising, as one would typically expect the training-fold-optimized SVM to overfit and therefore generalize worse. Despite this, both SVM configurations and logistic regression perform similarly when evaluated using classification accuracy. However, when evaluated using log-score, their performance differs substantially. In particular, the SVM performs worse than the baseline in terms of log-score. This result is not surprising, as logistic regression is explicitly trained to minimize log-loss (maximize log-score), whereas SVM optimizes hinge loss. With this violation, we do not allow for fair evaluation and comparison of the models.

III. Dependency of Prediction Error on Shot Distance

A. Methodology

Second, we are asked to estimate whether the prediction error depends on the distance value. To do that, we fitted a Generalized Additive Model (GAM), using the prediction error as the dependent variable and distance as the predictor. We chose GAM because it can model non-linear and non-monotonic relationships, which is important since the error is not expected to change linearly with distance.

After fitting the model, we generated a sequence of distance values across the observed range and predicted the probability of error for each value. We also constructed 95% confidence intervals to reflect the uncertainty around these predictions. This process was repeated for each of the models.

B. Results and Discussion

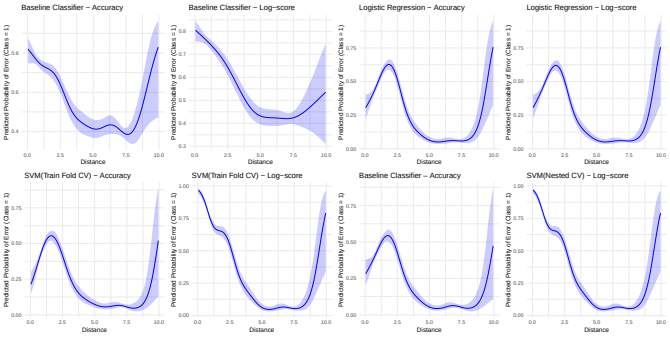

Fig. 2: The figures show the dependency of prediction error based on Shot Distance.

From the figure 2, we observe that the predicted error probability is not constant across distances, indicating a clear dependency. The relationship is non-linear and non-monotonic. For most models, error is lowest between approximately 5 to 7 meters. Beyond that range, especially after 7 meters, both the error probability and the uncertainty tend to increase. This suggests that prediction error does indeed depend on shot distance.

IV. Model Performance under true Competition Type distribution

A. Methodology

To estimate how the model would perform under the true distribution of competition types, we weighted the predictions for each competition type based on their actual frequencies. We computed classification accuracy and log loss separately for each competition type, then took a weighted average using these frequencies. This corrects for the fact that the dataset has roughly equal representation across types.

To estimate the standard error, we used stratified bootstrapping: resampling the data with replacement while preserving competition type structure.

B. Results and Discussion

Model	Accuracy	Log Score
Baseline	0.439 ± 0.0093	-10.045 ± 0.101
Logistic Regression	0.740 ± 0.0081	-6.685 ± 0.017
SVM (Train)	0.752 ± 0.0079	-12.780 ± 0.289
SVM (Nested)	0.767 ± 0.0077	-12.780 ± 0.308

TABLE II: Weighted performance metrics for different models.

From table II, we can see an improvement in the log score for SVM in both configurations, which indicates that the model correctly classified the instances belonging to the majority competition type (NBA). A slight improvement is also observed with logistic regression, and a decrease is observed with the baseline classifier.

V. Conclusion

We evaluated different models for predicting basketball shot types. SVM and logistic regression showed similar performance when using classification accuracy, but logistic regression was more reliable, outperforming in terms of log score. We also found that prediction error depends on shot distance, and that for some models the performance changes when accounting for the true distribution of competition types.