Data Science @UL-FRI

Evaluating Prediction Models: Performance, Error Analysis, and Distribution Shifts

Teodora Taleska

Introduction

This report evaluates models for predicting basketball shot types. We compare a *Baseline Classifier*, *Logistic Regression*, and *Random Forest* optimized through standard and nested cross-validation. Random Forest achieved the highest accuracy and lowest log-loss.

We also examined the relationship between shot distance and prediction errors, finding a weak negative correlation using the *Spearman rank coefficient*, indicating errors slightly decrease with distance but are not strongly dependent on it.

Finally, adjusting for the true competition type distribution increased log-loss significantly while accuracy remained unchanged, suggesting that while predictions were correct at the same rate, their confidence degraded. This underscores the importance of evaluating models under real-world data distributions.

Part 1: Model Evaluation and Comparison

Methodology

Dataset and Problem Statement

The dataset consists of basketball shot data with the goal of predicting the *ShotType* (6 categories) using all other variables. The dataset is assumed to be a representative sample of the data-generating process.

Chosen Models

Three models were compared: a *Baseline Classifier* predicting the most frequent class, *Logistic Regression* for multi-class classification, and *Random Forest*, an ensemble model sensitive to hyperparameters such as the number of trees. Random Forest was chosen for its strong performance and sensitivity to hyperparameter tuning.

Evaluation Method

The models were evaluated using *5-fold cross-validation* with a 70-30 train-test split. The dataset contains 5024 instances, resulting in approximately 3500 training instances. With 5 folds, each fold contained around 700 instances, ensuring a balance between robust performance estimation and computational efficiency. For Random Forest, two hyperparameter tuning strategies were used: *Standard Cross-Validation*, where hyperparameters were tuned on the training folds, and *Nested Cross-Validation*, where an outer loop evaluated performance and an inner loop tuned hyperparameters.

Metrics

The models were evaluated using *Accuracy*, the proportion of correctly predicted instances, defined as $\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(y_i = \hat{y}_i)$, where N is the total number of instances, y_i is the true label, and $\hat{y}_i$ is the predicted label. Additionally, *Log-Loss*, which measures the confidence of predicted probabilities for multi-class classification, is defined as $\text{Log-Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}$ where $y_{i,c}$ is 1 if the true label of instance i is class c, and $p_{i,c}$ is the predicted probability of class c.

Results

Part 2: Error Analysis and Adjusting for True Distribution

Assessing the Influence of Shot Distance on Prediction Errors

Results

Adjusting for True Distribution

Results

References

[1] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[2] Trevor Hastie, Robert Tibshirani, Jerome Friedma. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*