Predicting MLB Runs Scored: Feature Significance and Feature Reduction

Miguel Corona

Dec 2020

1 Introduction

This document contains the efforts involved in identifying the significant features for predicting an MLB team's season Runs Scored as well as the effects of omitting the least significant features in a model's training. The experiments were performed with the use of the neural network developed to predict the Runs Scored based on the features: Slugging Percentage (SLG), Batting Average (BA), and On-Base Percentage (OBP). The model's architecture was optimized to yield the minimal loss on the validation set provided all features were utilized in the model's training. The assessments for feature significance and feature reduction were performed by providing a subset of features into the model's training and evaluating on the validation set.

2 Feature Significance

The objective of the feature significance experiment was to observe how each feature affects the model's performance by training the model with only the feature of interest. The architecture that yields the prediction of the Runs Scored based on the BA, OBP, and SLG was utilized for the experiment. Each feature's contribution to the model's learning was assessed via the resulting validation set loss. Each assessment involved training with a tensor input consisting of only the feature of interest. The initial weights were held constant across each single feature training to mitigate any potential deviations that may occur due to the initialization of the trainable parameters. ModelCheckpointing was utilized to capture the parameters that yield the minimal Mean Average Error on the validation set.

The resulting learning curves for each single feature training are contained within Figure 1. The learning curves indicate that the features are within the model's capacity and that the validation loss and training loss align in the training. The validation loss from the single feature training are contained within

Table 1 and depicted within Figure 2 respectively. The validation loss metrics indicate that a team's Batting Average is not a reliable statistic to assess a team's run scoring potential as it yields a greater loss than training with Slugging Percentage or training with On-Base Percentage in isolation. Slugging Percentage and On-Base Percentage serve as better indicators for a team's runs scoring potential as these yield comparable losses with Slugging Percentage outperforming On-Base Percentage. The results align with the correlations observed in the initial phase of the assessment.

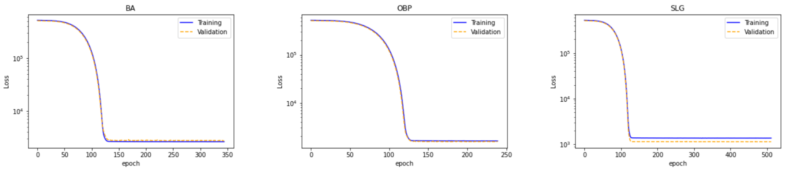

Figure 1: Single Feature Learning Curves

Interestingly, the difference between Batting Average and On-Base Percentage is much greater than the difference between On-Base Percentage and Slugging Percentage and this behavior may be explained by delving into the meaning of each metric. Batting Average serving as a poor indicator for Runs Scored aligns with the meaning of the statistic as it provides a probability of a hit occurring during the at-bat regardless of the type of hit. A homerun and a single are treated equally within the context of Batting Average although each hit yields a different outcome. The outcome of the hit is contained within Slugging Percentage as it represents the expected bases per at-bat and this feature plays the most prominent role in predicting the Runs Scored. The On-Base Percentage also serves as a strong indicator for Runs Scored even though it suffers from a similar setback as Batting Average as it treats all hits equally. The statistic may offset the all hits treated equally setback by including other means of getting on base to yield a probability of the batter becoming a runner. Knowing whether a batter will be a runner and the expected bases per at-bat are more reliable measurements for predicting the Runs Scored than knowing the probability of a hit occurring.

Feature	MAE
SLG	26.589
OBP	30.913
BA	42.748

Table 1: Single Feature Validation Loss

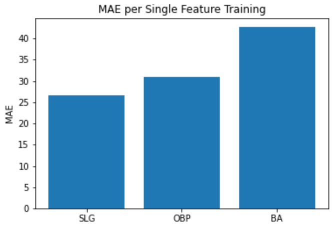

Figure 2: Single Feature Validation Loss

3 Feature Reduction

The objective of the feature reduction experiment was to observe the model's performance by omitting the least significant in the model's training. The significance of each feature's contribution to the learning is contained within the preceding section. The experiment consists of training the model with the defined architecture and the omission of features as the input. The assessments were performed similarly to those of feature significance by evaluating the model's performance on the validation set. ModelCheckpointing was utilized to acquire the model that yields the minimal loss on the validation set for each training. Only one feature was omitted as further reductions duplicate the efforts involved in the feature significance experiment.

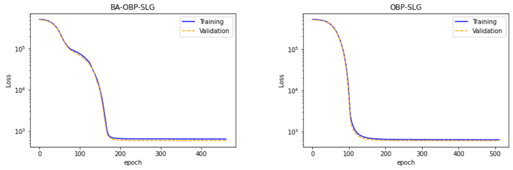

Figure 3: Feature Reduction Learning Curves

The resulting learning curves for each single feature training are contained within Figure 3. The learning curves indicate that the model architecture generalized the data as the validation loss is aligned with the training loss regardless of the reduced input. No overfitting was observed with the reduced input training for a model whose hyperparamters were optimized for all features. Training terminates sooner for the reduced input training than when all features are utilized. The quicker convergence can be attributed to the hyperparameters as the problem becomes less complex with the omission of features. The observed validation loss for each training in the feature reduction experiment are contained within Table 2 and depicted within Figure 4 respectively. The omission of the Batting Average minimally affects the model's performance on the validation set as both training cycles yield comparable outcomes. The On-Base Percentage and Slugging Percentage training performed slightly better than when trained with all features. The slight deviation may be attributed to the initialization of the weights. No loss of information occurs with the omission of Batting Average in training.

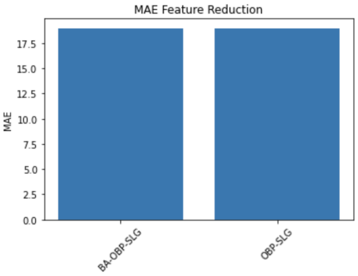

Figure 4: Feature Reduction Validation Loss

Features	MAE
BA-OBP-SLG	18.977
OBP-SLG	18.908

Table 2: Feature Reduction Validation Loss

4 Conclusion

It was observed that not all features contribute equally to the model's learning as it was determined that Slugging Percentage and On-Base Percentage serve as strong indicators for Runs Scored. The impact of each feature was assessed by training a model with a single feature at a time and evaluating the resulting validation loss. The validation loss between training with Slugging Percentage and training with On-Base Percentage were comparable unlike the loss of training with Batting Average with the loss of the preceding features. The minimal impact of Batting Average on the model's performance was further demonstrated through feature reduction as features were omitted in the model's training. A model that omits Batting Average has a comparable performance than that of a model that accounted for the statistic. Slugging Percentage and On-Base Percentage should be emphasised for understanding the manner in which runs are generated.