Patrick Quinn

ADAN 8888

28 October 2024

Week 8 Assignment

For this week's task, I utilized the LightGBM model, a decision tree-based ensemble that leverages gradient boosting for high performance on structured data. LightGBM was selected due to its computational efficiency and capacity to handle large datasets with complex patterns, which aligns well with the requirements of predicting fantasy football points based on varied and intricate features. This model's strengths in managing categorical variables and interactions without extensive preprocessing made it a suitable choice for optimizing predictions in a high-dimensional dataset like this one.

The LightGBM approach is complex, as it uses gradient boosting to iteratively build on the residuals of previous models. Its structure relies on multiple hyperparameters to manage tree depth, learning rate, subsampling, and feature fraction—each contributing to its ability to fit complex, non-linear patterns. However, this complexity can also introduce challenges, such as overfitting, which is managed through carefully tuned hyperparameters and early stopping. Despite its power, LightGBM's complexity means that the model's performance heavily depends on finding the right balance in parameter settings, as observed in the model variations.

In this analysis, three key hyperparameters were adjusted across variations:

1. Learning Rate: Controls the step size of each iteration, balancing the model's fit speed and accuracy. A lower learning rate often improves model generalization.

2. Max Depth: Limits the depth of each tree, which can help prevent overfitting. Deeper trees capture more detail but may also capture noise in the data.

3. Num Leaves: Determines the maximum number of leaves per tree, which affects the model's ability to fit to complex structures. A higher value increases complexity but requires careful tuning to avoid overfitting.

These parameters were selected because they directly influence the trade-off between model accuracy and generalization, which is essential in building a robust model capable of predicting fantasy football points reliably.

The performance metrics selected were Root Mean Square Error (RMSE) and R^2 (coefficient of determination).RMSE reflects the average deviation of predictions from actual values, providing a straightforward interpretation of model error in the context of fantasy points. R^2 indicates the proportion of variance explained by the model, offering insight into how well the

model captures the underlying relationships within the data. These metrics are well-suited for regression tasks like this, where the goal is to minimize error and maximize explanatory power in predictions.

Here is a summary table showcasing the performance of each variation on training and validation datasets:

Model Variation	Validation RMSE	Validation R^2	Observations
Variation 1	168.71	0.665	Lowest performance in both RMSE and R^2, indicating insufficient complexity or suboptimal parameter settings
Variation 2	156.57	0.712	Best performance across variations, suggesting an effective balance of parameters for model generalization.
Variation 3	161.25	0.694	Moderate performance with acceptable error but slightly lower explanatory power than Variation 2.

Variation 1 performed with the highest RMSE (168.71) and the lowest R^2 (0.665). This result indicates that the parameters were likely too restrictive, limiting the model's complexity and reducing its ability to generalize effectively on the validation set. Variation 2 achieved the best results, with the lowest RMSE (156.57) and the highest R^2 (0.712). This suggests that the parameter tuning in Variation 2 struck an optimal balance, allowing the model to effectively capture the underlying data patterns without overfitting. Variation 3 performed moderately, with an RMSE of 161.25 and an R^2 of 0.694. While its performance is competitive, it falls short of Variation 2 in both metrics. The settings in Variation 3 may have introduced a degree of complexity that slightly impacted generalization, as indicated by its higher RMSE relative to Variation 2.

The training and validation metrics demonstrate the impact of tuning hyperparameters on model generalization. Variation 1's results show underfitting, likely due to restrictive depth and leaf settings, limiting the model's ability to capture data complexity. Variation 2's improvement on both metrics suggests a more appropriate complexity level, while Variation 3's results, though close, reveal a slight overfit, evidenced by its relatively higher validation RMSE.

Based on the analysis, Variation 2 is the best model for this week's prediction task. It achieved the lowest Validation RMSE and the highest Validation R^2, indicating a successful balance of accuracy and generalization. This variation demonstrates the model's optimal capacity to predict fantasy football points with minimal error, supporting its reliability for practical application in the prediction task.

This LightGBM model approach, particularly with the optimized parameters of Variation 2, serves as a robust foundation for fantasy football predictions. Future iterations could focus on additional feature engineering and exploring ensemble techniques to potentially improve performance further.