ML2 Project Report

Nick Chandler, Tharun Johny Mekala, Ashritha Shreedhar, Luke Richard

2025-12-31

Contents

1 Abstract

In this report, we examine the application of two statistical learning methods to baseball data from the years 2020-2024. Specifically we are interested in predicting the next season's batting average of arbitrary players given the standard and advanced metrics of the previous season from this source. The report begins with a quick exploratory data analysis (EDA), it then moves into a

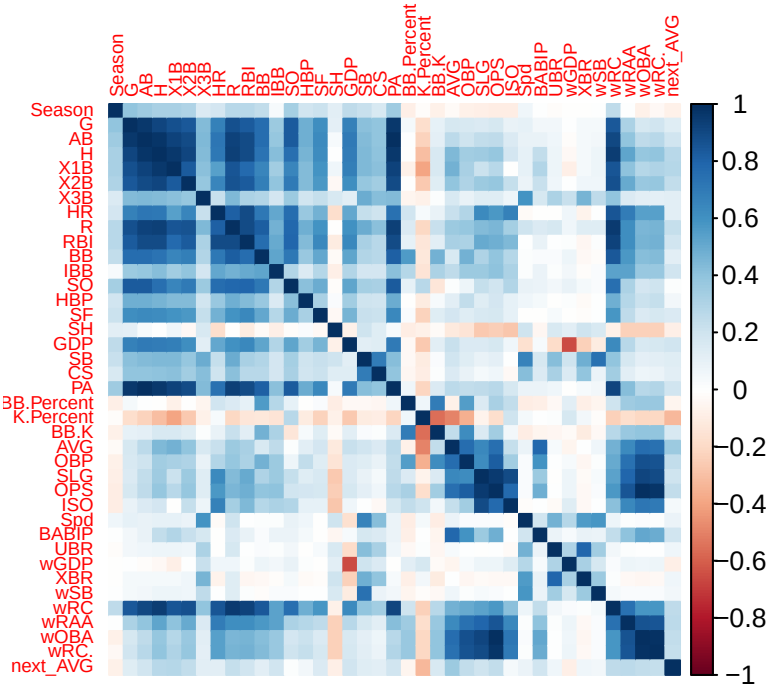
mathematical description of the statistical learning methods applied, then the application of each method, and finally a performance evaluation and description. This report illustrates how classical, interpretable methods can be applied, and then how more modern (and less interpretable) methods can be applied for better performance, and with the help of XAI, also be interpretable.

2 EDA

Here we do a short exploratory analysis of the data to understand some of the basic characteristics with a focus primarily on the correlation structures between variables since the modeling techniques we utilize rely on this. For more information on each feature, fangraphs provides a glossary, here.

The dataset has 1455 rows, 41 columns, and contains 559 unique players.

Here are correlation plots using Pearson's correlation coefficient (Pearson 1895). Pay special attention to the *next_AVG* column so see which features have a high (or low) correlation with the target.



3 Mathematical Description of Methods

This section briefly describes the mathematical background of each of the methods used. All equations herein can be traced back to the literature of (Hastie, Tibshirani, and Friedman 2009; Wagner 2025).

3.1 Mathematical Description of Regularized Regression

This section examines the mathematical background for the regularized regression model types of: Ridge, Lasso, and Elastic-Net regression. All assume a model of the form (here, without interaction

terms):

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon.$$

Where the β_is are found by solving (though specific instances introduce additional terms):

$$\min_{\vec{\beta}} \left\| X\vec{\beta} - \vec{y} \right\|_2^2, \text{ with } X \in \mathbb{R}^{n \times p} \text{ the design matrix}$$

The general solution is given (analytically) by:

$$\hat{\vec{\beta}} = \left(X^T X\right)^{-1} X^T \vec{y}$$

3.1.1 Elastic-Net Regression

The elastic-net regression uses a convex combination of the lasso penalty (L1) and ridge penalty (L2). Specifically, the optimization problem is given by:

$$\min_{\vec{\beta}} \left\| X\vec{\beta} - \vec{y} \right\|_2^2 + \lambda \left((1-\alpha) \left\| \vec{\beta} \right\|_1 + \alpha \|\vec{\beta}\|_2^2 \right)$$

Since the L1 norm is not differentiable at 0, there is no closed-form, analytical solution to the above problem. Solvers for convex problems capable of dealing with non-smooth objectives must be applied. The optimality condition is given by:

$$0 \in 2X^\top (X\hat{\beta} - y) + \lambda \left(2\alpha \hat{\vec{\beta}} + (1-\alpha) \partial \|\hat{\beta}\|_1 \right), \quad (\partial \|\hat{\beta}\|_1)_i = \begin{cases} \{1\}, & \hat{\beta}_i > 0, \\ \{-1\}, & \hat{\beta}_i < 0, \\ [-1, 1], & \hat{\beta}_i = 0 \end{cases}$$

The parameters λ and α are hyperparameters to be tuned via cross validation.

3.2 Mathematical Description of Gradient Boosting Decesion Trees

In this section we give a short mathematical description of Gradient Boosting Decision Trees (focusing on the gradient boosting aspect) (Wagner 2025).

Suppose our training data is given by: $\{(\vec{x_i}, y_i)\}_{i=1}^n$

Gradient boosting builds an additive sequential model interatively, aiming to minimize a loss function: $L(y, F(x))$. Specifically, gradient boosting builds a model (defined recursively):

$$F_m(x) = F_{m-1}(x) + \eta \cdot \gamma_m \cdot h_m(x)$$

with,

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$

The two step process (for each iteration) to fit the model consists of: 1. Building weak learners $h_m(x)$. 2. Finding the optimal step size at that iteration γ_m.

For step 1. we must compute the pseudo residual:

$$\tilde{y}_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

and then fit the weak learner $h_m(x)$ (in our case, a tree model) to the given pseudo residual.

We then do step 2. to find the optimal step size γ_m using an optimizer to solve:

$$\arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$$

Finally, the η given in the original equation is a hyperparameter and must be tuned.

4 Train-Val-Test Split

We split the data into a train set (60%), validation (val) set (20%), and a test set (20%). Specifically, we adhere to the following condition when deciding on the splits: A player may occur in no more than one set. Additionally, the newest data (that from the 2024 season) is saved for the test set.

In numbers:

- The train set has 872 rows and 352 unique players.
- The validation set has 291 rows and 117 unique players.
- The test set has 290 rows and 88 unique players.

5 Regularized Regression Application

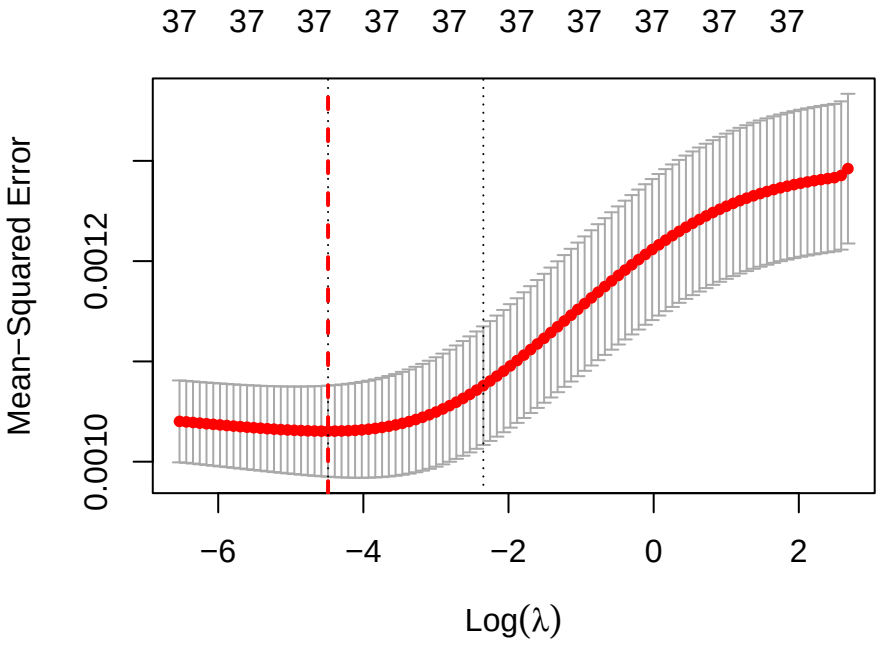
In this section, we apply the regularized regression techniques doing cross validated grid-search hyperparameter optimization over values for alpha.

Through our analysis, we find that ridge attains the minimum validation RMSE of all models tested at (rounded): 0.031623. Which means that the model mispredicts the next season's batting average by approximately 3.1623% on average. We examine the best model further, using classical diagnostics.

The features with larger magnitude coefficients can be seen as contributing more to the regression and so we plot them here.

Feature	Coefficient
AVG	0.0980
K.Percent	-0.0764
BB.Percent	-0.0575

The following is a plot of the lambda value in the ridge regression.

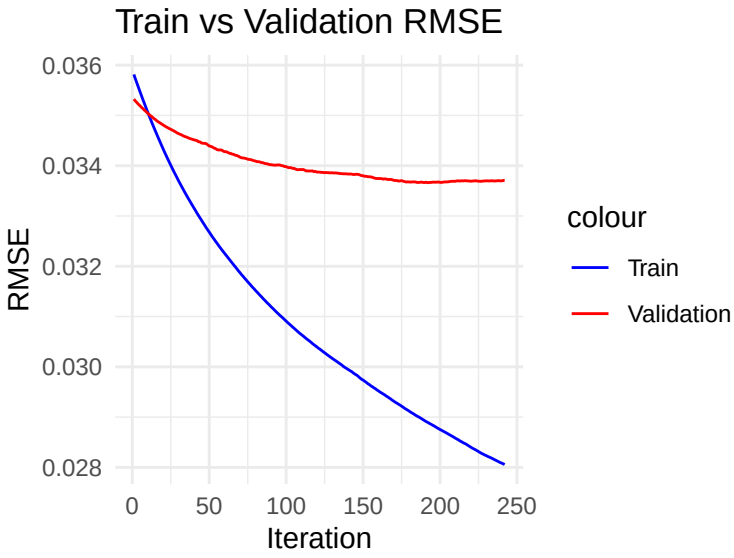


6 Gradient Boosted Decision Trees Application

In this section, we apply LightGBM (Ke et al. 2017).

6.1 LightGBM

Here we do grid-search hyperparameter optimization to find the best LightGBM model.




```
## ====================================
##           Best LightGBM Model
## ====================================
```

```
## Best Validation RMSE : 0.0337
## Best Training RMSE   : 0.0281
## Best Iteration       : 192
##
## Selected Hyperparameters:
##   learning_rate        : 0.01
##   max_depth            : 6
##   num_leaves           : 8
##   feature_fraction     : 1
##   objective            : regression
##   metric               : rmse
## ====================================
```

6.2 LightGBM Feature Importance

This section examines the LightBGM-Specific feature importances.

Table 2: LightGBM - Top 10 Features by Gain

Feature	Gain	Cover	Frequency	Descriptions
AVG	0.4030	0.2655	0.1466	Batting average – hits per at-bat
K.Percent	0.1751	0.2221	0.1912	Strikeout percentage – proportion of plate appearances ending in strikeout
IBB	0.0452	0.0514	0.0387	Intentional walks – plate appearances ending in intentional walk
SO	0.0395	0.0356	0.0439	Strikeouts – total number of strikeouts
BABIP	0.0359	0.0447	0.0580	Batting average on balls in play – luck-independent hitting measure
wRAA	0.0317	0.0448	0.0409	Weighted runs above average – overall offensive contribution
ISO	0.0279	0.0294	0.0402	Isolated power – extra base hits per at-bat
SF	0.0238	0.0222	0.0439	Sacrifice flies – fly balls scored for a run
XBR	0.0203	0.0217	0.0387	Grounded into double play – number of double plays hit into
GDP	0.0197	0.0169	0.0320	At-bats – total plate appearances minus walks, hits, etc.

7 Comparison of Model Classes

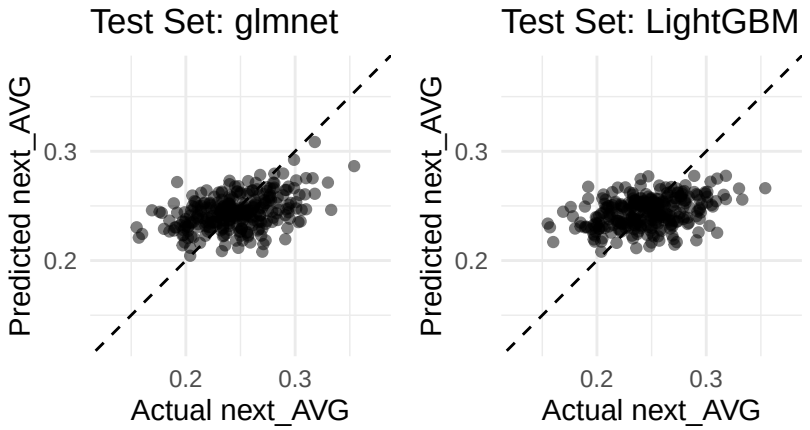
Here we evaluate the best model from each of the previous sections using performance diagnostics and XAI techniques.

7.1 Test Set Evaluation

We evaluate on the test data and calculate RMSE, R-Squared, and plot one-to-one plots.

Table 3: Test Set Performance (RMSE and R-Squared)

Model	RMSE	R_Squared
Regularized Regression (glmnet)	0.0302	0.17674
LightGBM	0.0312	0.12131



7.2 XAI / IML Analysis

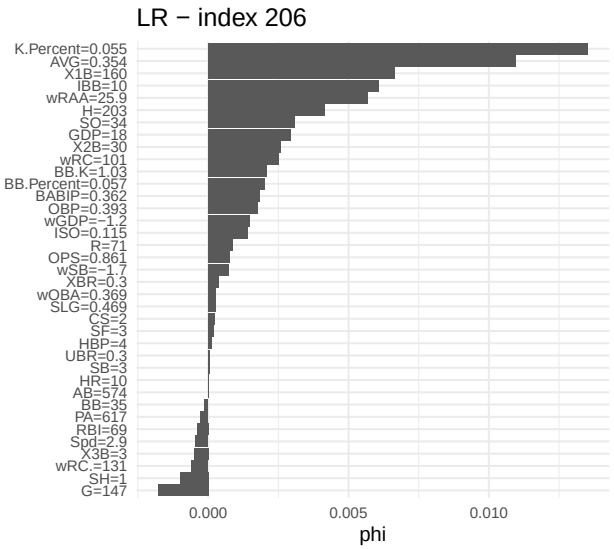
We examine model-agnostic explainability results on the *test set*. This includes global permutation feature importance, partial dependence feature effects, and microscopic SHAP.

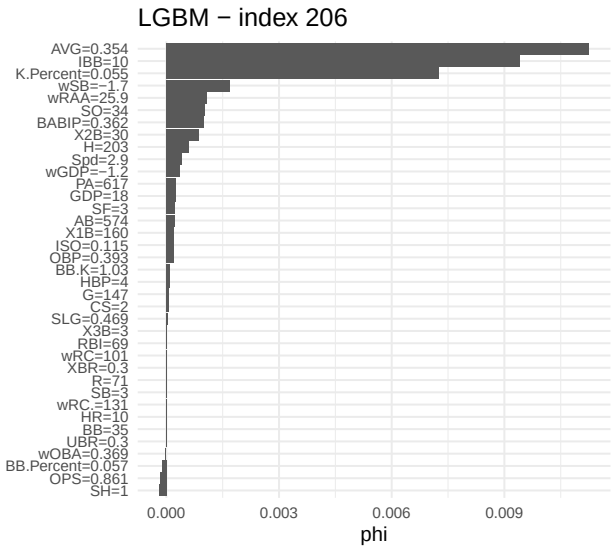
7.2.1 Microscopic

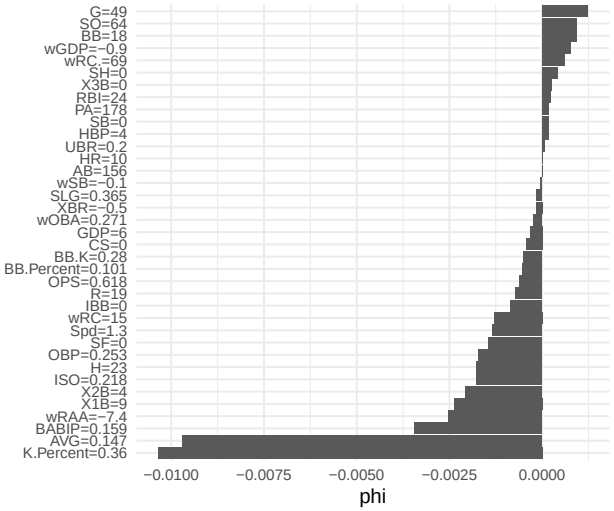
Here we examine the predictions, errors, and actual values for a set of selected observations before examining the SHAP values (Lundberg and Lee 2017) of those instances.

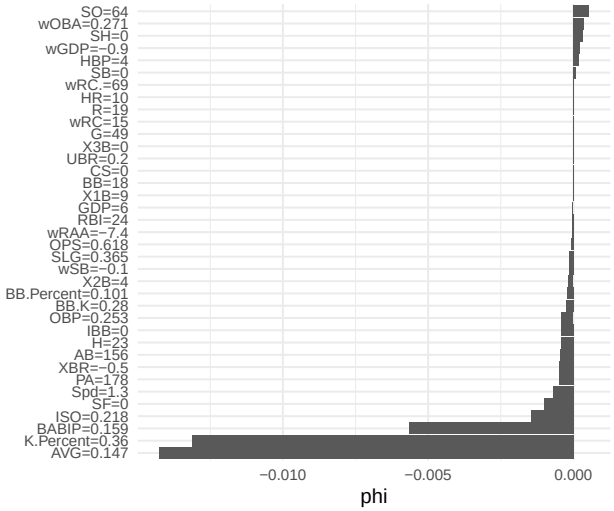
Table 4: Chosen test-set cases for SHAP explanations

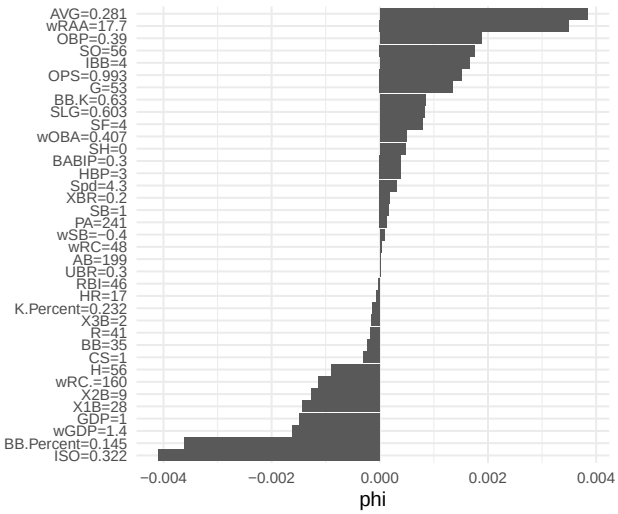
Case	Index	Actual	Pred_LR	Pred_LGB	AbsErr_LR	AbsErr_LGB
Highest predicted (LGBM)	206	0.318	0.3085	0.2776	0.0095	0.0404
Lowest predicted (LGBM)	87	0.204	0.2044	0.2081	0.0004	0.0041
Largest error (LR)	225	0.333	0.2464	0.2558	0.0866	0.0772

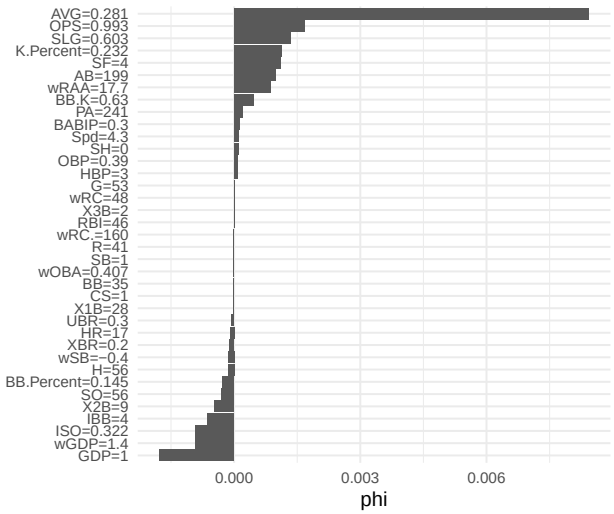

LR − index 206
K.Percent=0.055
AVG=0.354
X1B=160
IBB=10
wRAA=25.9
H=203
SO=34
GDP=18
X2B=30
wRC=101
BB.K=1.03
BB.Percent=0.057
BABIP=0.362
OBP=0.393
wGDP=−1.2
ISO=0.115
R=71
OPS=0.861
wSB=−1.7
XBR=0.3
wOBA=0.369
SLG=0.469
CS=2
SF=3
HBP=4
UBR=0.3
SB=3
HR=10
AB=574
BB=35
PA=617
RBI=69
Spd=2.9
X3B=3
wRC.=131
SH=1
G=147
0.000
0.005
0.010
phi


LGBM − index 206
AVG=0.354
IBB=10
K.Percent=0.055
wSB=−1.7
wRAA=25.9
SO=34
BABIP=0.362
X2B=30
H=203
Spd=2.9
wGDP=−1.2
PA=617
GDP=18
SF=3
AB=574
X1B=160
ISO=0.115
OBP=0.393
BB.K=1.03
HBP=4
G=147
CS=2
SLG=0.469
X3B=3
RBI=69
wRC=101
XBR=0.3
R=71
SB=3
wRC.=131
HR=10
BB=35
UBR=0.3
wOBA=0.369
BB.Percent=0.057
OPS=0.861
SH=1
0.000
0.003
0.006
0.009
phi


LR − index 87
G=49
SO=64
BB=18
wGDP=−0.9
wRC.=69
SH=0
X3B=0
RBI=24
PA=178
SB=0
HBP=4
UBR=0.2
HR=10
AB=156
wSB=−0.1
SLG=0.365
XBR=−0.5
wOBA=0.271
GDP=6
CS=0
BB.K=0.28
BB.Percent=0.101
OPS=0.618
R=19
IBB=0
wRC=15
Spd=1.3
SF=0
OBP=0.253
H=23
ISO=0.218
X2B=4
X1B=9
wRAA=−7.4
BABIP=0.159
AVG=0.147
K.Percent=0.36
−0.0100
−0.0075
−0.0050
−0.0025
0.0000
phi

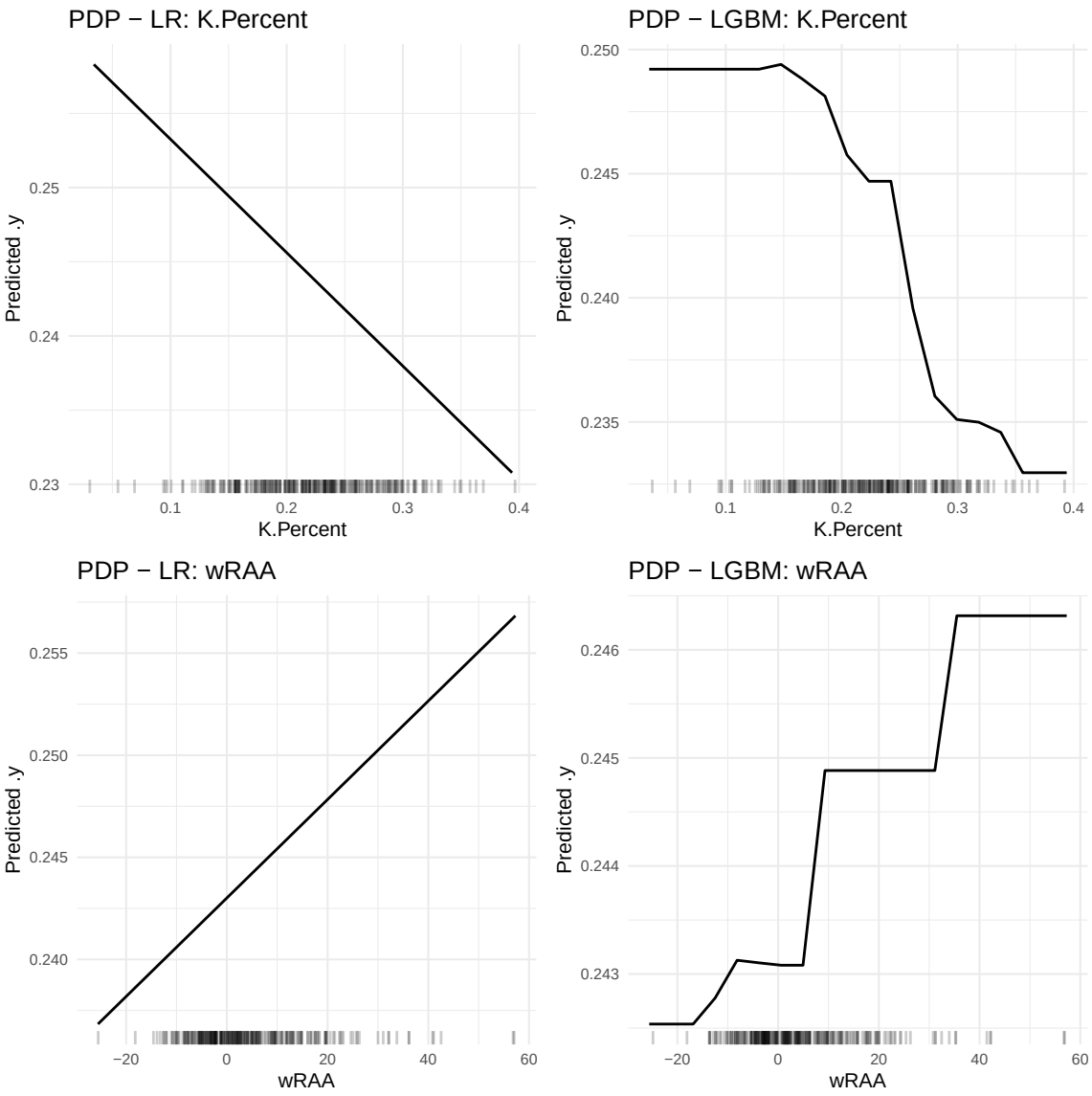

LGBM − index 87
SO=64
wOBA=0.271
SH=0
wGDP=−0.9
HBP=4
SB=0
wRC.=69
HR=10
R=19
wRC=15
G=49
X3B=0
UBR=0.2
CS=0
BB=18
X1B=9
GDP=6
RBI=24
wRAA=−7.4
OPS=0.618
SLG=0.365
wSB=−0.1
X2B=4
BB.Percent=0.101
BB.K=0.28
OBP=0.253
IBB=0
H=23
AB=156
XBR=−0.5
PA=178
Spd=1.3
SF=0
ISO=0.218
BABIP=0.159
K.Percent=0.36
AVG=0.147
−0.010
−0.005
0.000
phi


LR − index 225
AVG=0.281
wRAA=17.7
OBP=0.39
SO=56
IBB=4
OPS=0.993
G=53
BB.K=0.63
SLG=0.603
SF=4
wOBA=0.407
SH=0
BABIP=0.3
HBP=3
Spd=4.3
XBR=0.2
SB=1
PA=241
wSB=−0.4
wRC=48
AB=199
UBR=0.3
RBI=46
HR=17
K.Percent=0.232
X3B=2
R=41
BB=35
CS=1
H=56
wRC.=160
X2B=9
X1B=28
GDP=1
wGDP=1.4
BB.Percent=0.145
ISO=0.322
−0.004
−0.002
0.000
0.002
0.004
phi

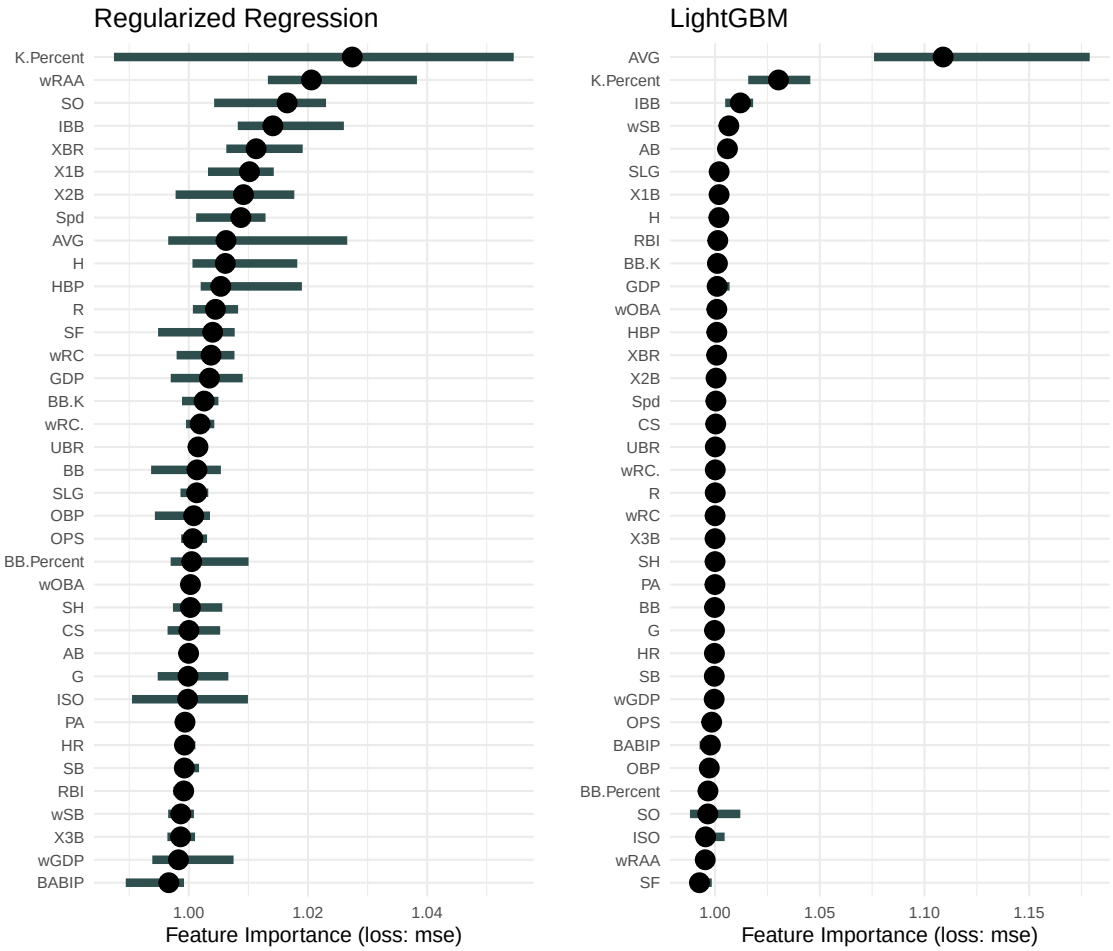

LGBM − index 225
AVG=0.281
OPS=0.993
SLG=0.603
K.Percent=0.232
SF=4
AB=199
wRAA=17.7
BB.K=0.63
PA=241
BABIP=0.3
Spd=4.3
SH=0
OBP=0.39
HBP=3
G=53
wRC=48
X3B=2
RBI=46
wRC.=160
R=41
SB=1
wOBA=0.407
BB=35
CS=1
X1B=28
UBR=0.3
HR=17
XBR=0.2
wSB=−0.4
H=56
BB.Percent=0.145
SO=56
X2B=9
IBB=4
ISO=0.322
wGDP=1.4
GDP=1
0.000
0.003
0.006
phi

7.2.2 Macroscopic

Here we plot the Partial Dependence Plots (Hastie, Tibshirani, and Friedman 2009) for two chosen features.



Here we plot the permutation feature importances for the two models. Note that the LightGBM importances agree with the model specific ones for the first few features. The regularized regression explanation technique disagrees aside from the K percent. Also note that K percent is ranked high in both models' feature importance plots.



References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." *Advances in Neural Information Processing Systems*.

Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems (NeurIPS)* 30.

Pearson, Karl. 1895. "Note on Regression and Inheritance in the Case of Two Parents." *Proceedings of the Royal Society of London* 58: 240–42.

Wagner, Steffen. 2025. *Machine Learning 2 Course Notes*.