SHAP を用いた MLB の配球分析

石橋　京也† 　唐　 惠東†† 　蒋　 帥†† 　兪　 樺†† 　亀井　清華††
森本　康彦††

† 広島大学情報科学部情報科学科　〒 739–8527 広島県東広島市鏡山一丁目４番１号
†† 広島大学大学院先進理工系科学研究科　〒 739–8527 広島県東広島市鏡山一丁目４番１号
E-mail: †{b195083,d216083,m21111,huachongyu,s10kamei,morimo}@hiroshima-u.ac.jp

あらまし　野球において,「配球」を分析することは重要である. 本研究では, eXplainable AI（以下, XAI）技術を利用して配球の分析を試みる. 特定の捕手に注目し, 配球のブラックボックスの予測モデルを作成し, XAI でそのモデルを解釈することでプロ野球選手が行う「配球の意図」を推定する. 本研究では, Major League Baseball（以下, MLB）独自の解析ツールである “statcast” のデータを対象にランダムフォレストを用いて予測モデルを作成し, XAI 技術の一つである SHAP でモデルの予測を解釈した. その結果から，各捕手がどのような意図をもって配球していたかを推定することができた.

キーワード　ランダムフォレスト, 機械学習, SHAP, アンサンブル学習

1　は じ め に

野球において「配球」は, 試合結果を大きく左右する要素である. 一般に打者が予測していた球を打つ時とそうでない球を打つ時ではヒットを打てる確率が異なるため, バッテリー（投手と捕手）は打者を抑えるために球種やコースの組み合わせの仕方を工夫する. それが「配球」である. 配球は重要な戦術であり, 現役のプロ野球選手の配球の考え方を知ることができる機会はほとんど無い. そこで, 外から分析する方法として, 機械学習で「配球」をモデリングし, SHAP を適用した結果から, 対象となった選手の「配球の意図」を推定するという方法を考えた.

モデリングの対象となる「配球を考える人」について, 野球の試合の中で配球を決めるのは, 投手, 捕手, 監督, コーチ（データ分析班など）と様々である [1]. 誰を対象としても興味深い研究となると考えられるが, 稲福らの文献 [2] において, 日本野球機構（以下, NPB) では捕手ごとに配球の特徴がみられると言われていることから, 本研究では捕手を対象とした. したがって, MLB から 5 名の捕手をピックアップし, それぞれ予測モデルを作成することで比較しながら分析を行った.

2　関連研究と基礎事項の準備

2.1　配球予測に関する研究

Sidle らは文献 [3] において, Linear Discriminant Analysis（以下, LDA)) [4], Support Vector Machine（以下, SVM) [5], ランダムフォレスト [6] の 3 つの手法で 7 種類の球種を予測した. その結果, ランダムフォレストが最も精度が高く, その精度は 66.62 %であった. 本研究では, 球種だけでなく投球コースも予測している点や, 作成した予測モデルに対して SHAP を適用し, 分析を行っている点が異なる.

稲福らは文献 [2] において,「配球」を球種, 球速, 座標の推移ベクトルを構築することで表し, 主成分分析を行うことで捕手ごとに配球が異なることを示した. 本研究は, SHAP を使うことで特徴量をそのまま分析するという点で異なる.

菊地らは文献 [7] において, 大学野球においてカウント 0-0 からどのような投球がストライクを取りやすいのかに着目した. 球種, コースの割合および投球結果の割合について単純集計を行い, スイングの有無とストライクの取り方については球種別, コース別のクロス集計により割合の算出を行った. そして独立性の検定に χ^2 検定を用いることで分析を行った. 本研究は, 配球を機械学習を用いてモデリングして分析を行うという点で異なる.

2.2　SHAP

機械学習モデルの複雑な動作の全容を解釈したいというニーズから, その予測根拠を説明する XAI 技術の研究が急速に進んでいる [8]. その中でも今回使用する SHapley Additive exPlanation（以下, SHAP) [9] は代表的な XAI 手法の 1 つである. SHAP は経済学の Shapley 値 [10] が基になっている. Shapley 値は協調ゲームにおいて, 複数のプレーヤーが存在した場合に, 全体の連携に対して与えられた報酬金を協調した複数のプレーヤーでいかに分配すべきかを求める. それをベースに SHAP では, 機械学習モデルのプレーヤーを特徴量に置き換え, 報酬をモデルの予測値ととらえる [11]. その結果, 複雑で解釈が難しいとされる機械学習モデルについて, どの特徴量が予測にどれくらい貢献しているかを可視化できる.

3　方　　法

3.1　「配球」の定義

図 1 のように予測の対象となる「配球」を球種（速い球 or 遅

い球）×高さ（高め or 低め）×コース（内 or 外）=8 分類と定義する．右打者側のコース高めの速い球をクラス 0, 遅い球をクラス 1, 左打者側のコース高めの速い球をクラス 2, 遅い球をクラス 3, 右打者側のコース低めの速い球をクラス 4, 遅い球をクラス 5, 左打者側のコース低めの速い球をクラス 6, 遅い球をクラス 7 とする．

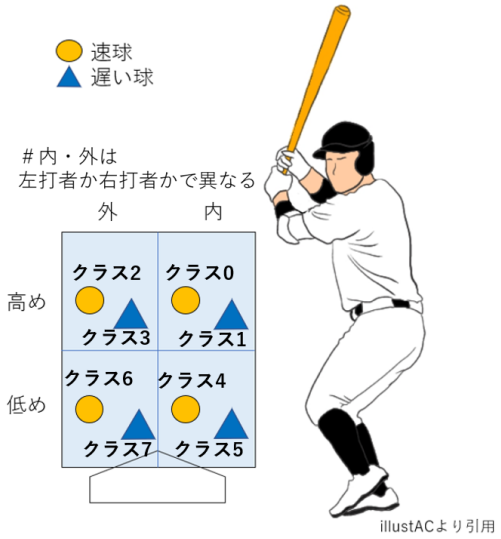

図 1　配球の分類

3.2　データセット

本研究では, MLB 独自の解析ツールである “statcast” のデータを利用する．MLB のドジャース, パドレス, ジャイアンツ, ダイヤモンドバックス, ロッキーズの 5 チームの正捕手について, 2022 年シーズンの 1 球ごとのデータ（表 1）を用いる．さらにそのデータは, それぞれのチームで年間を通して先発ローテーションに入った 4 名の投手（表 2）が投げた球に絞る．このように投手を絞った理由は, この研究では球種のみでなく投球コースを予測するため, 制球力があるという前提が必要だからである．これらのデータについて, シーズン前半の 80 %を訓練データ, シーズン後半の 20 %をテストデータとして, モデル作成, 精度評価を行った．

表 1　捕手ごとのデータ数

捕手名 (チーム名)	データ数
ウィル・スミス (ドジャース)	5232
オースティン・ノラ (パドレス)	6287
ジョーイ・バート (ジャイアンツ)	4735
カーソン・ケリー（ダイヤモンドバックス）	5389
エリアス・ディアス (ロッキーズ)	5735

3.3　特　徴　量

特徴量について表 3 に示す．statcast のデータの中で, 配球決定に関わるであろう 14 個を使用した．「*tensa*」は, 守備側のチームが勝っていればプラスの値をとり, 負けていればマイナスの値をとる．

表 2　モデルに使用した投手

チーム名	投手名（利き腕）
ドジャース	タイラー・アンダーソン（左）
	フリオ・ウリアス（左）
	トニー・ゴンソリン（右）
	クレイトン・カーショー（左）
パドレス	ダルビッシュ・有（右）
	ジョー・マスグローブ（右）
	ショーン・マナイア（左）
	フレーク・スネル（左）
ジャイアンツ	ローガン・ウェブ（右）
	カルロス・ロドン（左）
	アレックス・カッブ（右）
	アレックス・ウッド（左）
ダイヤモンドバックス	メリル・ケリー（右）
	ザック・ゲーレン（右）
	マディソン・バムガーナー（左）
	ザック・デービーズ（右）
ロッキーズ	ヘルマン・マルケス（右）
	カイル・フリーランド（左）
	チャド・クール（右）
	オースティン・ゴンバー（左）

表 3　特　徴　量

属性値	説明
balls	ボールの数（0〜3）
brl_pa	打者のバレル%
inning	イニング数
mae_pitch_name	1 球前の球種 {0（速球）or 1（遅い球）}
mae_X	1 球前のコース（内 or 外）
mae_Z	1 球前の高さ（高め or 低め）
on_1b	1 塁ランナーがいるか（0 or 1）
on_2b	2 塁ランナーがいるか（0 or 1）
on_3b	3 塁ランナーがいるか（0 or 1）
outs_when_up	アウトの数（0〜2）
pitch_sum	その試合での総投球数
pitch_number	打席内での投球数
p_throws	右投手ならば 0，左投手ならば 1
stand	右打者ならば 0，左打者ならば 1
strikes	ストライクの数（0〜2）
tensa	両チームの点差

3.4　予測モデルの作成

本研究では, 機械学習の手法としてランダムフォレスト [6] を使用する．ランダムフォレストは決定木を用いたアンサンブル学習の 1 つであり, 高い予測性能を得ることが出来る．また, 効率が良く, 計算速度が高速であることが知られている．Sidle らの文献 [3] において球種を予測するモデルを作成した時に最も精度がよかった手法がランダムフォレストであったことから, 本研究でもランダムフォレストを使用した．モデルの精度は, 表 4 のとおりである．次節からは, このモデルに SHAP を適用し, 分析していく．

表 4　各予測モデルの精度

捕手名 (チーム名)	正解率
ウィル・スミス (ドジャース)	21.01 %
オースティン・ノラ (パドレス)	22.21 %
ジョーイ・バート (ジャイアンツ)	28.61 %
カーソン・ケリー（ダイヤモンドバックス）	22.73 %
エリアス・ディアス (ロッキーズ)	23.35 %

4　SHAP による分析

本研究では分析対象のモデルが 5 つあるが, そのまま分析結果を載せると冗長になってしまうため, ドジャースのモデルの分析結果を軸として詳しく説明し, 他のモデルについてはドジャースのモデルと同じ特徴, 違う特徴があった場合に記載する.

4.1　重要な特徴量分析

図 2 が, ドジャースの予測モデルに SHAP Feature Importance メソッドを適用したものである. まず, 横軸が各データについての SHAP 値の絶対値の平均を取ったものであり, 「特徴量が予測にどの程度影響を与えたか」という値である. そして縦軸の特徴量は, 上から順に貢献度の高い順に並んでいるため, 例えば一番上の p_throws（左投げ・右投げ）は, このモデルが配球を予測するのに最も貢献した特徴量だと言える.

次に色の違いは, クラスごとの SHAP 値の値を表している. クラスと配球の対応は, 3 章の図 3.1 の通りである. SHAP を分類クラスモデルに適用する場合は, 各クラスごとに SHAP 値を出す. 例を挙げると, p_throws（左投げ・右投げ）の水色のバー（Class7）とは, クラス 7 のゾーンに投手が投げると予測するか投げないと予測するかという予測に対して p_throws（左投げ・右投げ）が貢献した量となる. 横向きのバーの色の割合を見て, p_throws（左投げ・右投げ）のクラス 7 に対する貢献が, 全体的に見た p_throws（左投げ・右投げ）の貢献度を押し上げていることが分かる.

分析を進めていく. 図 2 を見て, p_throws（右投げ・左投げ）, strikes（ストライク数）, stand（右打ち・左打ち）の 3 つの特徴量が配球予測に重要であることが分かる. p_throws（右投げ・左投げ）については, 水色で表される Class7 の重要度の大きさが特徴的である. 他に分かることとして, on_1b（1 塁ランナーの有無）, on_2b（2 塁ランナーの有無）, on_3b（3 塁ランナーの有無）の走者に関する特徴量は重要度が低いことが分かる. ドジャース以外のモデルについても, p_throws（右投げ・左投げ）, strikes（ストライク数）, stand（右打ち・左打ち）の 3 つの特徴量が重要であることが確認できた. その他に pitch_sum（総投球数）も重要度が高いことが分かった. また, on_1b（1 塁ランナーの有無）, on_2b（2 塁ランナーの有無）, on_3b（3 塁ランナーの有無）の走者に関する特徴量の重要度が低いことは全モデル共通であった. 表 5 に重要な特徴量について 5 つのモデルを調べた結果をまとめる.

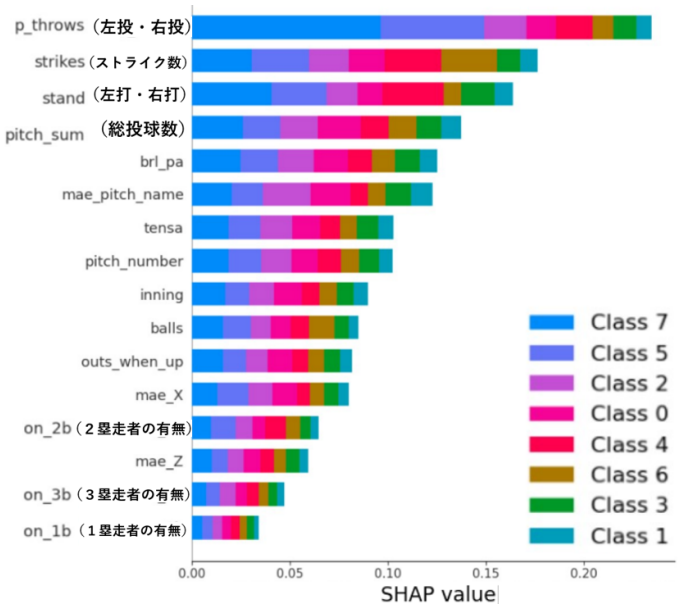

図 2　SHAP Feature Importance（ドジャース）

表 5　配球決定に重要な特徴量

特徴量	重要度 Top3 に入ったモデル数
strikes（ストライク数）	5
stand（右打ち・左打ち）	4
p_throws（右投げ・左投げ）	3
pitch_sum(総投球数)	3

4.2　特徴量の寄与の仕方分析

次に SHAP Dependence Plot メソッドを適用して, SHAP 値と特徴量の関係を分析した. SHAP Dependence Plot メソッドはクラスごとの SHAP 値と特徴量の値の関係を散布図に表す. 図 3 のクラス 4 を例にとって説明する. まず, 特徴量 stand（左打ち・右打ち）は 0（右打者）と 1（左打者）の 2 値を取るのだが, 横軸でそれが表されている. そして, 例えば 1（左打者）について, 縦軸 SHAP 値の分布を見てみると, SHAP 値が正に集まっていることが分かる. ここから特徴量 stand（左打ち・右打ち）の値が 1（左打者）であることは, クラス 4 を予測することに対して, 正に寄与することが分かる. 色は, SHAP が特徴量と最も相互作用の強い特徴量との関係を表すが, 本研究では特に言及しない.

分析を進めていく. まず, 図 3 から図 6 は, ドジャースのモデルの低めのコース（クラス 4, 5, 6, 7）について, SHAP 値と stand（右打ち・左打ち）の関係を表した図である. 色は, それぞれのクラスで stand（右打ち・左打ち）との相互作用が最も大きい特徴量との関係を表す. これらの散布図から分かることは, 右打者側のコース（クラス 4,5）では, 左打者（stand の値が 1）の SHAP 値がプラスに集まっており, 左打者側のコース（クラス 6,7）では, 右打者（stand の値が 0) の SHAP 値がプラスに集まっていることが読み取れる. この様子は, パドレス以外のモデルに共通して見られた. パドレスの stand（右打ち・左打ち）は, クラスごとに SHAP 値との関係が様々で規則性は見つけられなかった.

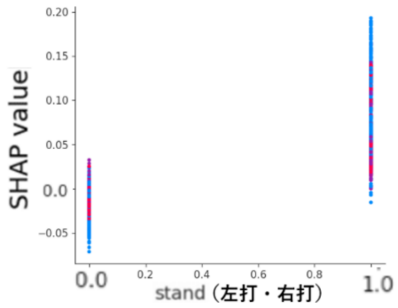

図 3　クラス 4

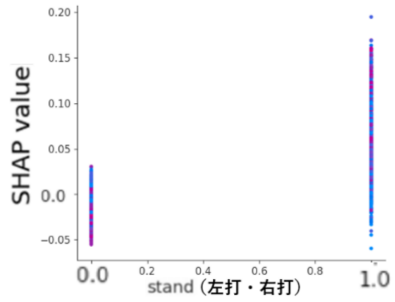

図 4　クラス 5

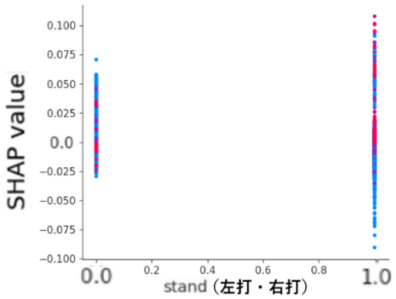

図 5　クラス 6

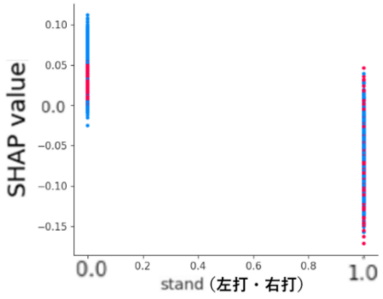

図 6　クラス 7

次に, 図 7 から 図 10 は, ドジャースのモデルの低めのコース（クラス 4, 5, 6, 7）について, SHAP 値と strikes（ストライク数）の関係を表した図である. これらの散布図から, 速球（クラス 4, クラス 6）では反比例の関係が, 遅い球（クラス 5, クラス 7）では比例の関係が見て取れる. つまり, 速球はカウントを稼ぐために使い, 遅い球（主に変化球）は追い込んでから投げるという方針であると分かる. この様子は全モデルで見られ, パドレスのモデルでは高めのコースにも見られた.

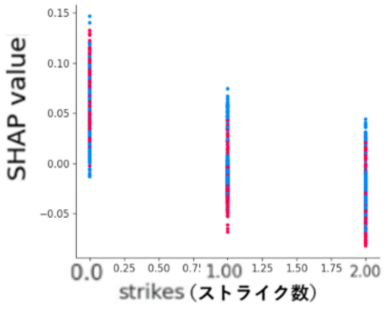

図 7　クラス 4

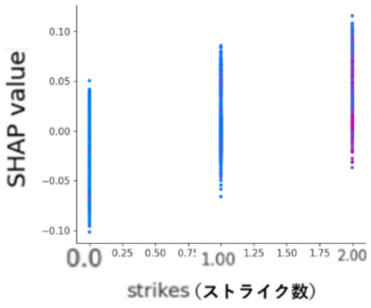

図 8　クラス 5

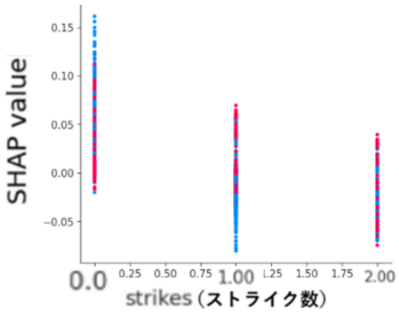

図 9　クラス 6

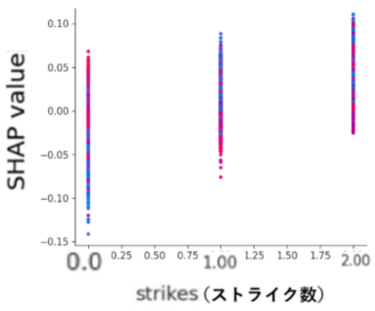

図 10　クラス 7

次に, 図 11 から図 17 は, ドジャースのモデルの低めのコース（クラス 4, 5, 6, 7）について, SHAP 値と p_throws（右投げ・左投げ）の関係を表した図である. 利き腕側のコースの SHAP

値はマイナス, 反対側のコースの SHAP 値はプラスに集まっていることが分かる．この様子は, パドレスのモデルの右打者側のコースにも見られた．

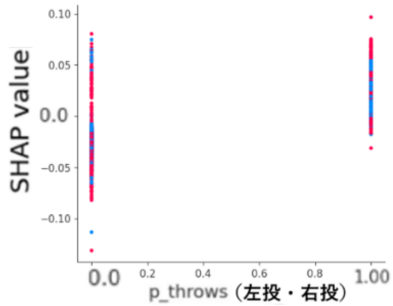

図 11　クラス 4

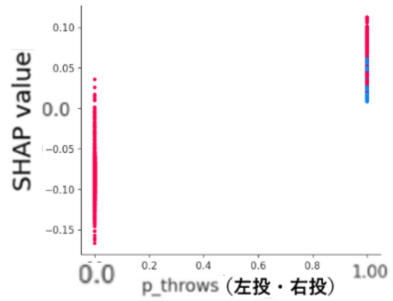

図 12　クラス 5

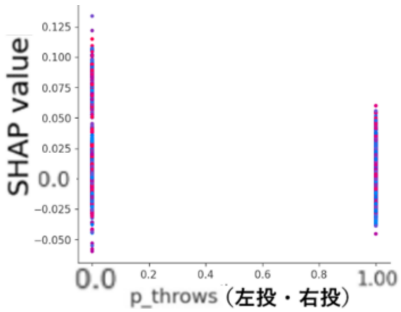

図 13　クラス 6

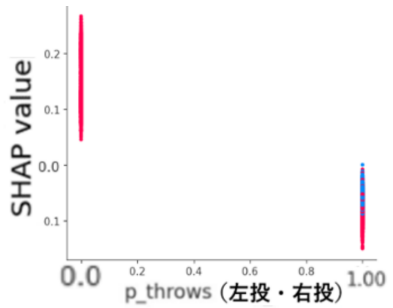

図 14　クラス 7

次に, 図 15 から図 18 は, ドジャースのモデルの高めのコース（クラス 0, 1, 2, 3）について, SHAP 値と mae_pitch_name(1 球前の球種) の関係を表した図である．これらの図から, 速い球（クラス 0,2）は 1 球前が遅い球（1.0）であるとの予測に正に寄与し, 遅い球（クラス 1,3）はその逆であることが分かる．このような特徴は, ダイヤモンドバックスのモデルでも見られた．

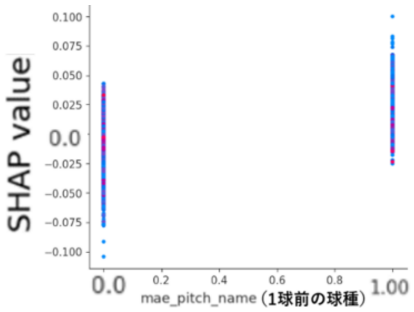

図 15　クラス 0

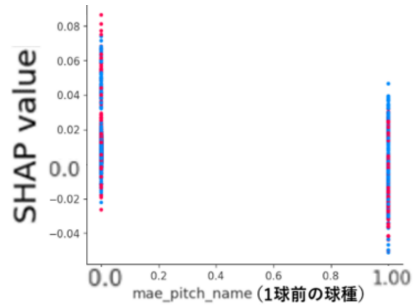

図 16　クラス 1

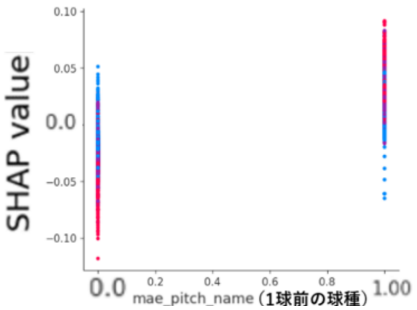

図 17　クラス 2

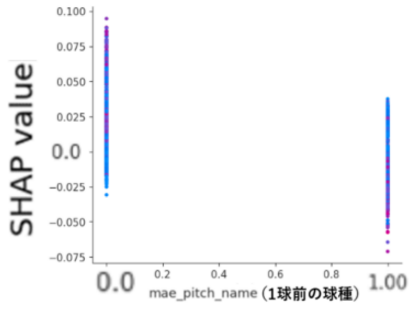

図 18　クラス 3

SHAP Dependence Plot で分析した結果について表 6 にまとめる．

表 6　SHAP 適用で分かったことまとめ

特徴量	分かったこと（当てはまったモデル数）
stand （右打ち・左打ち）	・全クラスでアウトコースの寄与度が高い（4） ・寄与の仕方はクラスごとに様々（1）
mae_pitch_name (1 球前の球種)	・1 球前が遅い球だと速い球の予測に寄与 速い球だと遅い球の予測に寄与（2）
strikes （ストライク数）	・低めのコースについて, 速球はストライク数と SHAP 値が反比例, 遅球は比例（4） ・クラス 4 を除いて, 速球はストライク数と SHAP 値が反比例, 遅球は比例（1）
p_throws （右投げ・左投げ）	・低めのコースについて, 利き腕と同じ側のコースの 寄与度が高く, 反対は低い（1） ・右のインコースについて, 左投手が寄与（1） ・クラス 0，2，3 は左投手の寄与 クラス 1，5，6，7，は右投手の寄与（1）

4.3 考　　察

4.4 球種に関する考察

まず, mae_pitch_name(1 球前の球種) から, 2 つのモデルで速球と遅い球を交互に投げる配球が行われていることが分かったことについて, これは緩急を使うことで打者のタイミングをずらすことを目的としていると考えられる. 第 1 節で, 打者は予測していた球は高確率で打つことができると述べたが, 1 球前との間に球速差がある場合, 体が前の球に「慣れてしまう」ことで, 予測していた球でも上手く対応できなくなる. よって, 緩急を使うことは有効とされる [12]. 他の 3 つのモデルで緩急を使っている様子が見られなかった理由としては, 緩急を使うには, 投手の持ち球に速い球と遅い球があることが前提なので, データの対象となった投手の違いが出たと考えられる.

次に strikes（ストライク数）から, 各チームとも共通してストレート系の速い球でカウントを取り, 追い込んだら変化球で抑えるという方針であると分かったことについて, この特徴が全チームに共通していることは意外であった. 甲子園, NPB を観ていると「変化球でカウントを取る」, 「ストレートで空振り三振を取る」という配球も頻繁に行われている. よって, どれかのモデルで「変化球でカウントを取り, 速い球で抑える」という逆の配球結果があっても不思議でない. それでもこのような結果になった理由として考えられることは 2 つある. 1 つは日本とアメリカの配球の考え方の違いである. 日本で一般的な配球がアメリカではされていないということはあり得る. もう 1 つは「流行」である. 配球にも年ごとにトレンドがあるならば, 現在の流行りが 「ストレート系の速い球でカウントを取り, 追い込んだら変化球で抑える」なのかもしれない.

4.5 投球コースに関する考察

まず, stand（右打ち・左打ち）が配球に及ぼす影響として, 4 つのモデルで, 打者のアウトコースの予測に正に寄与した. このことについて, 一般論として打者にとって体から遠いアウトコースの球は, 物理的に目からの距離が離れるため捉えることが難しく, また, 体から離れることで力も入りにくいとされ, 特に低めのアウトコースは投手の生命線といわれる [13]. 本研究でアウトコースを基本とする配球が MLB でも行われていることが分かり, 野球のレベルに関係なく通用する基本の配球であることが確かめられた.

次に p_throws（右投げ・左投げ）が配球に与える影響について, ドジャースとパドレスの 2 つのモデルでは利き腕と反対のコースの予測に寄与した. つまり利き腕と同じ側のコースに投げるより, 利き腕と反対側のコースに投げる方が打者を抑える可能性が高いと考えていることになる. 考えられる理由としては, いろいろなコースに順番に投げるよりも 1 つのコースに投げ続ける方がコントロールしやすいということが挙げられる. インコースとアウトコースを十分でない制球力で投げ分けるよりもどちらかに絞って基本はそのコースに投げた方がよい結果になると考えていると推測する.

5 まとめと今後の課題

本研究では, SHAP を使うことで配球予測モデルごとに重要とする特徴量は何か, 特に寄与した特徴量に関しては特徴量がどういう値であれば予測に影響するのかを分析した. その結果, 捕手によって配球を決める時に重要とする要因を発見することができたと考える. 「配球の意図」というモデリングしないと定式化できないものを分析しようと思った場合, このような分析の仕方は一定の意義を得られるのではないだろうか. 予測モデルの精度や特徴量の選び方などを工夫し, 打者の得意コースや苦手コースなどのデータを入れることでさらに興味深い研究になると考える.

6 謝　　辞

本研究は科研費 20K11830 の助成を受けたものです.

文　　献

[1] 小宮山 悟, 勝亦 陽一, 福永 哲夫. 「プロ野球球団におけるゲーム分析データの活用事例」. スポーツパフォーマンス研究, Vol.7, pp.346–355, 2015.

[2] 稲福 和史, 伏見 卓恭, 佐藤 哲司. 「最終球への配球推移に基づくキャッチャー成績分析」. DEIM Forum, I1-2, 2019.

[3] Sidle Glenn, Tran Hien. “Using multi-class classification methods to predict baseball pitch types”. *Journal of Sports Analytics*, Vol.4, No.1, pp85-93, 2018.

[4] John Lafferty. “Latent Dirichlet Allocation”. *Journal of Machine Learning Research*, Vol.3, pp.993-1022, 2003.

[5] Lorenza Saitta. “Support-Vector Networks”, *Machine Leaming*, Vol.20, pp.273-297, 1995.

[6] Leo Breiman. “Random Forest”. *Machine Learning*, Vol.45, pp5-32, 2001.

[7] 菊地 啓太, 中島 宣行, 綿田 博人. 「大学野球における配球について: カウント 0-0 における投球の分析」. 体育研究所紀要, Vol.49, No.1, 2010.

[8] 恵木 正史. 「XAI(eXplainable AI) 技術の研究動向」, 日本セキュリティ・マネジメント学会誌, Vol.34, No.1, 2020.

[9] Scott M Lundberg, Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. *Advances in Neural Information Processing Systems*, Vol.30, 2017.

[10] Lloyd Stowell Shapley. “A value for n-person games”. *Contributions to the Theory of Games*, Vol.2, No.28, pp.307-

317, 1953.
[11] 大坪 直樹, 中江 俊博, 深沢 裕太, 豊岡 祥, 坂元 哲平, 佐藤 誠, 五十嵐 健太, 市原 大暉, 堀内 新吾. 「XAI（説明可能な AI）そのとき人工知能はどう考えたのか？」. 4 章, 3 節 , リックテレコム, 2022.
[12] 平井 成二. 「これで完璧！野球ピッチング・守備」. 4 章, ベースボールマガジン社, 2014.
[13] 吉見 一起. 「コントロールの極意」. 1 章, 竹書房, 2022.