Vol.2021-IS-155 No.6
2021/3/8

機械学習を用いた所属メンバー入れ替え数による組織のパフォーマンス変化の分析

足利 太嘉[1,a)]　松井 藤五郎[2,b)]　武藤 敦子[1]　森山 甲一[1]　犬塚 信博[1]

概要：本稿では，近年盛んに行われているデータマイニングの中でも所属先の変化に伴う組織の成績に着目し, 分析を行った. 所属先の変化に関係がある, 加入者数, 継続者数, 脱退者数を用いることで組織の成績にどのくらいの影響があるのか, 有効であるかをロジスティック回帰を用いて MLB の前年度のチーム成績から次年度の成績の予測を行い, 加入者数, 継続者数, 脱退者数の有効性を示すことに成功した. また所属先の変化に個人成績のを考慮した変数を用いることでの予測精度の向上を試みた.

Analysis of changes in organizational performance due to the number of members replaced using machine learning

1. はじめに

近年, 大量のデータに対して統計学, パターン認識, 人工知能等のデータ解析手法を用いて様々な知識を獲得するデータマイニングが盛んに行われている. データマイニングは様々な分野で利用されており, その例としてスポーツや投資信託, マーケティングなどがある. このデータマイニング技術を用いることでスポーツにおいては相手の行動パターンや戦略を把握し, 自身に活かすことができる. 投資信託, マーケティングでは利益を捻出するための傾向分析, 予測に使用されている. これらのようにデータマイニングを用いることでより利便性の改善や分析が盛んになっている. 例えば橋本公雄ら [1] はスポーツ競技におけるパフォーマンスを予測するための分析的枠組みとして重回帰分析を用いることで心理的要素がもたらす影響について調査を行った.

他にもデータマイニングが対象としているデータは様々なものがあり, そのうちの一つに所属関係の変化による影響がある. これは例えば会社の場合, ある部署から別のある部署への異動に伴う変化や, スポーツの場合, チームを加入すること, 脱退することによる変化等様々な事象が身近に存在している. そこで永柄真澄ら [2] は異動した個人の職業性ストレスと満足要因について関連性を明らかにするために要因分析を行った.

これらのようにデータを用いた予測や要因分析等盛んに行われている. しかし, 異動に伴う個人の評価は行われているが組織としての評価は行われていない. この組織の評価をするに当たって所属先の変化における適切な人数の割り当て, 入れ替えが重要であると考えている. 例えばスポーツの場合, 加入してくる人数が多すぎるとチーム全体へ方針や指導が行き渡らないという問題が発生する. 反対に加入してくる人が少ないとチームメンバーがほとんど同じになり, 新たな発見やチームの大きな変化は見込みにくいという問題がある. 本論文では, パフォーマンスを評価しやすいスポーツチームに焦点を当て, MLB データを対象として分析を行った.

2. 所属先の変化について

2.1 所属関係

ここでは, 所属関係について説明する.

所属とは, あるメンバーがあるグループに属していることである.

所属関係とは, あるメンバーがあるグループに所属しているというその関係のことを指している.

例えば, ある生徒 m があるクラス c に属していることを m が c に所属しているといい, そのときの m と c の関係を

[1] 名古屋工業大学
Nagoya Institute of Technology Uniersity
[2] 中部大学
Chubu Uniersity
[a)] t.ashikaga.072@stn.nitech.ac.jp
[b)] TohgorohMatsui@tohgoroh.jp

所属関係という. ここで所属関係は次のように定義される.

所属関係の定義

メンバーの集合を $M = \{m_1, m_2, \ldots, m_n\}$, グループの集合を $G = \{g_1, g_2, \ldots, g_k\}$ とする. このとき,

$$b(m,g) = \begin{cases} \mathrm{T} & (m \text{ が } g \text{ に所属しているとき}) \\ \mathrm{F} & (\text{そうでないとき}) \end{cases}$$

となる $m \in M$ と $g \in G$ の関係 $b : M \times G \to \{\mathrm{T}, \mathrm{F}\}$ を所属関係という. ここで, n はメンバーの数,k はグループの数を表す.

2.2 動的所属関係

動的所属関係とは所属関係に時刻を加えたものである.

時刻 $t-1$ においてあるメンバー m があるグループ g に所属関係があるとき, 時刻 t においてあるメンバー m があるグループ g に所属関係があるこの時刻 $t-1$ と時刻 t の所属関係を動的所属関係という.

ここで動的所属関係は次のように定義する.

動的所属関係の定義

時刻 t において m が g に所属することを
$b_t(m,g)$ と表す.
b_t を動的所属関係という.

この動的所属関係には実生活にも多くの例が存在する. 例えば大学生の編入や小学校のクラス内での班分け, 野球チームの加入などがある.

2.3 加入, 継続, 脱退

移籍とは動的所属関係においてメンバーが所属しているグループから別グループに所属することである. 移籍先のグループから見た移籍を加入, 移籍元のグループから見た移籍を脱退という. 継続とは動的所属関係においてメンバーが所属しているグループが変わらないことである. 加入, 継続, 脱退は次のように定義される.

加入, 継続, 脱退の定義

$b_{t-\Delta}(m,g) \wedge b_t(m,g')$
$g \neq g'$ のとき m は時刻 t に g' に加入した
$g = g'$ のとき m は時刻 t に継続した
$g \neq g'$ のとき m は時刻 t に g から脱退した
という. ただし,$t-\Delta$ は,t より前で,m がいずれかのグループに所属していたことがわかる最後の時刻である.

例えば上記の動的所属関係の例から大学生の編入の場合一般的な大学生と比べると編入の確率はかなり低いため加入した時点で希少性の高いメンバーとなる. 他にも小学校のクラス内での班分けがある. こちらの例の場合, 班を加入することは当たり前であるため一回の加入では希少性は高くない. この二つに比べ MLB の加入は常に同じチームに所属するメンバーもいれば年々別のチームに加入するメンバーもいる. このようにメンバーの加入はグループ作成のシステムに大きく依存する.

3. 提案手法

本研究では, 所属グループが高々 1 個の動的所属関係において, 統計的な手法を用いて加入者数, 継続者数, 脱退者数の説明変数としての有効性を示す手法とメンバー個人の成績を考慮した説明変数を用いた組織のパフォーマンス向上における手法を提案する.

3.1 加入者数, 継続者数, 脱退者数を用いた有効性

ここでは, 加入者数, 継続者数, 脱退者数の説明変数としての有効性の確認方法の説明をする. 有効性の確認には, ロジスティック回帰分析という統計的な分析方法を用いる.

ロジスティック回帰分析とは, 目的変数が二値の場合に用いられる分析手法で, 主に確率やダミー変数について分析, 予想したいときに用いられる.

そこで, 組織の成績の良し悪しを 1 or 0 とする目的変数とし, その成績に影響がありそうなものを説明変数としてロジスティック回帰分析を行う. その後, 説明変数に加入者数, 継続者数, 脱退者数を加えることで結果にどのような変化があるかを確認していく.

3.1.1 ロジスティック回帰分析

ロジスティック回帰分析とは目的変数が二値であるデータの分析に用いられる統計学的手法である. この手法は目的変数が二値であるデータを扱う場合に説明変数から目的変数の確率を求める式を求め, 予測, 説明をするのに用いられる. 以下の式を用いることで偏回帰係数を求められる.

$$p = \frac{1}{1+\exp(-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n))}$$

ここで p は目的変数が 1 となる確率,β は偏回帰係数,x は説明変数を表す. $p \ldots$ 確率, $\beta \ldots$ 偏回帰係数, $x \ldots$ 説明変数の値

ロジスティック回帰分析では患者が病気を発症する確率, 顧客がローンを返済できる確率, 顧客が DM に反応する確率等を求める場合に用いられることが多い.

3.1.2 アルゴリズム

組織ごとの加入者数, 継続者数, 脱退者数に対してロジスティック回帰分析を行う手法を提案する. ここで提案手法のアルゴリズムを示す.

提案手法のアルゴリズム

入力：組織ごとの加入回数のベクトル V_1, 継続回数のベクトル V_2, 脱退回数のベクトル V_3, その他パフォーマンス向上に関係があるとされる変数 $A, B \ldots$
出力：成績の良し悪しの予測結果 X
(1) 組織の成績の良し悪しを 1, 0 とする Z を作成
(2) Z を目的変数, 成績の良し悪しに関係があるとされる $A, B \ldots$ と所属先の変化について V_1,V_2,V_3 を説明変数とし, ロジスティック回帰分析を行い, モデルを作成
(3) モデルに予測したいデータを代入し, 確率が 0.5 以上を 1, 0.5 未満を 0 と予測する

3.2 個人の成績を考慮した加入者数, 継続者数, 脱退者数によるパフォーマンス向上確率分析

ここではメンバー個人の成績を加味した説明変数を用いた組織のパフォーマンス向上における手法を提案する. まず, パフォーマンス向上について定義を行う. 本研究では組織のパフォーマンスのことを指す. パフォーマンス向上は対象とするデータによって大きく内容が異なるがパフォーマンスの変化という点においては同様である. 例えばスポーツの分野におけるパフォーマンス向上とは, あるチームの得点力が向上したということである. 会社の部門におけるパフォーマンス向上とは, ある部門の売上額が前年と比較し, 増加したことである. このように対象によって内容は異なるがパフォーマンスの向上という点においては同様である.

3.2.1 パフォーマンス向上

本研究ではパフォーマンスの向上をあるときのデータとその一つ前のデータの成績の比較によって成績が向上したか, そうでないかを判別する. そこで次のように定義する.

パフォーマンスの定義

時刻 t においてグループ g の成績を
$pf(t, g)$ と表す.
pf をパフォーマンスの値とする.

パフォーマンス向上の定義

$pf(t, g) - pf(t - \Delta, g) > 0$ のときグループ g は時刻 $t - \Delta$ から t にかけてパフォーマンス pf が向上したという.

3.2.2 個人の成績を考慮した変数の取得方法

ここでは個人の成績を考慮した手法について説明する. 上記の有効性の確認実験において組織の成績からその後どうなるか予測と, 加入者数, 継続者数, 脱退者数の必要性について実験を行った. その結果から加入者数, 継続者数, 脱退者数を説明変数として追加した方が良いことが分かったが, この加入者数, 継続者数, 脱退者数は定義上単純に組織を異動したメンバーの度数のことを指している. しかし, 現実問題においてどのような成績（優劣）のメンバーが異動してくるか, 組織に残るかということが重要である. そこで個人の成績を考慮した加入者数, 継続者数, 脱退者数を用いた精度改善を図る手法の提案を行う.

個人の成績を考慮した加入者数, 継続者数, 脱退者数

(1) データの中の考慮する成績の最大値 max_grade を取得
(2) for p in 全てのメンバー
(3) p の成績を max_grade で割った結果 p_grade を取得
(4) 複数年データがあるメンバーのみ for (row in 1 to メンバー p の列数-1)
メンバー p のデータのうち row 行における teamID と row+1 行の teamID を比較する
(5) if(teamID が同じ) メンバー p の成績を考慮した継続者数に p_grade 加える
else メンバー p の成績を考慮した加入者数, 脱退者数に p_grade 加える
(6) endif
(7) endfor
(8) endfor

3.2.3 アルゴリズム

ここでは加入者数, 継続者数, 脱退者数に個人の成績を考慮した変数を用いる手法を提案する. まずパフォーマンスの変化を見るために前年度の成績と比較し, パフォーマンスの変化を分類する. その後, 目的変数をパフォーマンスの変化, 説明変数を各組織の個人の成績を考慮した加入者数, 継続者数, 脱退者数とパフォーマンスの変化に関係があるとされる属性とし, ロジスティック回帰分析を行う. 分析結果の偏回帰係数からモデルを作成し, テストデータを用いて予測分析を行う. 提案手法のアルゴリズムを以下に示す.

提案手法のアルゴリズム

入力：組織ごとの個人の成績を考慮した加入回数のベクトル V_1, 滞在回数のベクトル V_2, 脱退回数のベクトル V_3 その他パフォーマンス向上に関係があるとされる変数 $A, B \ldots$
出力：パフォーマンスの変化の良し悪しの予測精度確認表 X, Y
(1) パフォーマンスの良し悪しを 1, 0 とする Z を作成
(2) Z を目的変数, 各組織の個人の成績を考慮した加入者数 V_1, 継続者数 V_2, 脱退者数 V_3 を説明変数とし, ロジスティック回帰分析を行い, 式を求める
(3) 求めた式にテストデータを代入し, 確率が 0.5 以上のとき 1（パフォーマンスが向上した）, 0.5 未満のとき 0（パフォーマンスが向上しなかった）と予測する

表 1　Batting リストデータ例

	playerID	yearID	stint	teamID	lgID	G	AB	R	H	X2B	X3B	HR	...
1	abercda01	1871	1	TRO	NA	1	4	0	0	0	0	0	...
2	addybo01	1871	1	RC1	NA	25	118	30	32	6	0	0	...

表 2　Teams リストデータ例

	yearID	lgID	teamID	franchID	divID	Rank	G	Ghome	W	L	DivWin	WCWin	...
1	1984	NL	ATL	ATL	W	2	162	81	80	82	N	NA	...
2	1984	AL	BAL	BAL	E	5	162	81	85	77	N	NA	...

4. 実験

まず実験に使用したデータについて説明する.

4.1 使用データ

近年, 組織のパフォーマンスの向上に向けた編成法等の研究が盛んに行われている. その上で本研究では加入数, 継続数という属性がパフォーマンスの向上に与える影響が分析において重要であると考えその有効性を示す. 組織のパフォーマンスの向上には, スポーツにおける勝利数, 会社における売上金額, 学校における進学実績等様々なものがあるが本研究ではパフォーマンスの向上が分かりやすく, 成績を表す属性が複数存在しているスポーツの分野の Major League Baseball について分析を行った.Major League Baseball についての基本的な情報を取得するため,R のパッケージである Larman の中から打者についてのデータである Batting リスト, チームについてのデータである Teams リストを抽出した.Batting リスト,Teams リストの一部を表 1,2 に示す.Batting リストは 1871 年から 2019 年までの Major League Baseball に登録されている選手すべての成績が入っておりメンバーの総数は 19689 人, データの行数は 107429, 列数は 22 であった.Teams リストは 1871 年から 2019 年までの Major League Baseball に登録されているチームすべての成績が入っておりチームの総数は 149 チーム, データの行数は 2925, 列数は 48 であった.

また実験には以下の条件を満たすデータを採用する.

- 1975 年〜2010 年の期間
- 空白の年がある場合はプレー年数には含めない
- プレー年数が複数年ある選手データのみ
- 加入はシーズンオフ時のときのみ

一つ目の条件は加入に関しての制度や選手が古すぎると評価が難しいためである. 二つ目の条件は選手によってはマイナーに落ちていて MLB に登録されていないことや他国でプレーしている期間は空白になっており選手ごとによって異なるためである. 三つ目の条件はプレー年数が 1 年の場合加入ができないためである. 四つ目の条件はシーズン中に加入する場合他の選手とタイムスタンプが異なり適切な判定ができないためである. 前処理を行った結果今回対象になったメンバーの総数は 7073 人, 総チーム数は 35 種類だった.

データの列情報について説明する.

- transferred は加入者数
- stayed は継続者数
- transferring は脱退者数
- H は安打数
- HR は本塁打数
- HA は被安打数
- HRA は被本塁打数
- win は成績が良いか悪いか 1, 0 で表す変数
- strong は強くなったかを 1, 0 で表す変数

4.2 加入者数, 継続者数, 脱退者数の有効性確認実験

4.2.1 実験方法

今回は加入者数, 継続者数, 脱退者数の有効性確認実験を行う.MLB における成績の良し悪しを今年度の順位成績で判別する.

実験方法

(1) リーグ上位 2 チームを成績が良い（win=1）, リーグ下位 2 チームを成績が悪い（lose=0）と分類, 3 位のチームを成績が普通 (middle=1) と分類する
(2) 前処理したデータの中で成績が普通 (middle=1) に分類されたデータを排除
(3) 目的変数をリーグの順位成績 (win), 説明変数を成績の良し悪しに関係があると考えられている前年度のチーム成績とし, ロジスティック回帰分析を行う (モデル 1)
(4) 目的変数をリーグの順位成績 (win), 説明変数を成績の良し悪しに関係があると考えられている前年度のチーム成績に加えてチーム間の加入者数, 継続者数, 脱退者数とし, ロジスティック回帰分析を行う (モデル 2)
(5) (3) と (4) の結果からモデルを作成する
(6) (5) のモデルに対して学習データに使用していない 2011 年から 2015 年をテストデータとし, 有効性の確認を調査する

4.2.2 実験結果

表 4 にモデル 1, 表 6 にモデル 2 のテストデータに対する分割表を示す. 表 3 にモデル 1, 表 5 にモデル 2 における説明変数の係数と切片の推定値 (estimate), その標準語差 (std.error), 有意確率を求めるための z 値 (z value), 有意確率 (pr(>|z|) を示す. また, モデル 1 では説明変数として使用した H, HR, HA, HRA の全てにおいて統計的に有意であるという結果になった. 偏回帰係数の符号を見ても H, HR が多いほど強く, HA, HRA が多いほど弱いというイメージとあった結果になった. 今回行った実験ではモデル 1 で 76/118 チームを正しく判別した. また表 3,5 から正解率について,「2 つの比率の差の検定」で統計的有意な差が

あるか調査したが, p 値=0.406 となり, 統計的に有意な差は見られなかった.

モデル 2 では 82/118 チームを正しく判別した. モデル 2 の真陽性率が 36/48=0.75, モデル 1 の真陽性率が 35/54=0.65, モデル 2 の真陰性率が 46/70=0.66, モデル 1 の真陰性率が 41/64=0.65 とモデル 2 の方が良い精度になっている.

表 3 モデル 1 :所属先の変化を考慮していないモデル

	estimate	std.error	z value	pr(>|z|)
intercept	0.578	0.974	0.593	5.533e-1
H	0.005	0.001	4.904	9.392e-7
HR	0.015	0.003	4.963	6.951e-7
HA	−0.006	0.001	−5.313	1.080e-7
HRA	−0.010	0.004	−2.864	4.183e-3

表 4 モデル 1 :予測結果

	強いチーム数	弱いチーム数	合計
強いと予測した数	41	23	64
弱いと予測した数	19	35	54
合計	60	58	118

表 5 モデル 2:所属先の変化を考慮したモデル

	estimate	std.error	z value	pr(>|z|)
intercept	3.608	1.368	2.638	8.327e-3
transferred	−0.164	0.036	−4.615	3.937e-6
stayed	−0.100	0.037	−2.727	6.383e-3
transferring	0.027	0.033	0.825	4.092e-1
H	0.005	0.001	4.521	6.151e-6
HR	0.016	0.003	5.292	1.212e-7
HA	−0.006	0.001	−5.204	1.952e-7
HRA	−0.007	0.004	−1.938	5.265e-2

表 6 モデル 2 予測結果

	強いチーム数	弱いチーム数	合計
強いと予測した数	36	12	48
弱いと予測した数	24	46	70
合計	60	58	118

4.2.3 モデルの当てはまり具合による考察

今回行った実験では所属先の変化数の説明変数を加えたことによる結果への影響を調べるため実験を行った. モデル 1 では野球の成績に関係がある属性 (H, HR, HA, HRA) 4 種類を使用した. モデル 2 では野球の成績に関係がある属性に, 加入者数, 継続者数, 脱退者数を加えた 7 種類を使用した. モデル 1 では 76/118 チームを正しく判別した. モデル 2 では 82/118 チームを正しく判別した. 所属先の変化を説明変数として加えたことで見つけられなかった 6 チームを見つけることができた. 次に適合度の検定として Hosmer-Lemeshow 検定を用いて評価した. 一般的に Hosmer-Lemeshow 検定では p 値が 0.05 より大きければ適合しているとみなす. 結果, モデル 1 の p 値=0.811, モデル 2 の p 値=0.853 となり, モデル 1,2 どちらについてもより適合しているとわかった.

またさらに目視による当てはまり具合の確認ができる ROC 曲線で評価をした. その結果を図 1 に示す. 青線がモデル 1, 黒線がモデル 2 を指している.ROC 曲線は次のように評価する.

(1) 当てはまりがよいモデルの ROC 曲線
点 (0, 0) から点 (0, 1) の近くに進み, そこから点 (1, 1) に向かって進む曲線

(2) 当てはまりの悪いモデルの ROC 曲線
ROC 曲線が 45 度線の近くを通過する

図 1 の結果からモデル 2 の方が当てはまり具合が良いことが見てわかる. また ROC 曲線下面積 (以下 AUC) について調査した. AUC(モデル 1)=0.711, AUC(モデル 2)=0.733, AUC の差の検定を行ったところ, p 値=0.031 となり, 有意水準 5 %で統計的に有意な差でモデル 2 の方が良いことが確認できた.

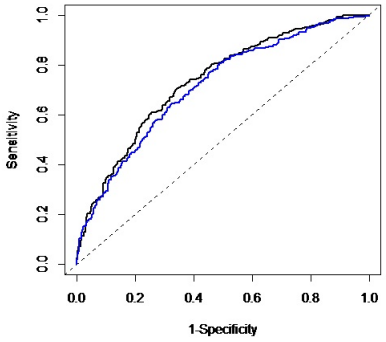

図 1 ROC 曲線結果. 青線がモデル 1, 黒線がモデル 2

さらに実験に使用したデータへの当てはまり具合についてもクロス集計表で確認してみる. 表 7, 8 の真陽性数, 真陰性数からモデル 2 の方が当てはまり具合が良い.

表 7 モデル 1 当てはまり具合結果

	強いチーム数	弱いチーム数	合計
強いと予測した数	233	118	351
弱いと予測した数	101	187	288
合計	334	305	639

表 8 モデル 2 当てはまり具合結果

	強いチーム数	弱いチーム数	合計
強いと予測した数	241	114	355
弱いと予測した数	93	191	284
合計	334	305	639

4.2.4 説明変数の数による考察

モデル 1 では野球の成績に関係がある属性 (H, HR, HA, HRA) 4 種類を使用した. 説明変数の個数による結果への影響を調べるため同じ個数で実験を行った. 説明変数は所属先の変化に関する加入者数, 継続者数, 脱退者数に野球の成績に関係がある属性 (H, HR, HA, HRA) 4 種類を順番に加え, 合計説明変数が 4 種類になるようにモデルを作成し, 予測を行った. その結果が表 9 から表 12 である. 予測精度は, 説明変数に H を加えたモデルでは 80/118=0.678, HR を加えたモデルでは 79/118=0.670, HA を加えたモデルでは 80/118=0.670, HRA を加えたモデルでは 81/118=0.686 だった. 全てのモデルにおいてモデル 1 の予測精度よりわずかながら良い結果になった. この結果から今回, 加入者数, 継続者数, 脱退者数はモデル 1 との差である 4 種類の属性と同程度の価値があることがわかった. また前年度のチーム成績 4 種類のモデルの精度を前年度の成績をわずか 1 種類に野球の指標として用いられていない所属先の変化の 3 種類のモデルの精度が上回るということから野球の成績に関係がない他の属性に関しても成績を決める要因とすることを検討する必要性がある.

表 9 予測結果 (transferred, stayed, transfering, H)

	強いチーム数	弱いチーム数	合計
強いと予測した数	31	9	40
弱いと予測した数	29	49	78
合計	60	58	118

表 10 予測結果 (transferred, stayed, transfering, HR)

	強いチーム数	弱いチーム数	合計
強いと予測した数	29	8	37
弱いと予測した数	31	50	81
合計	60	58	118

表 11 予測結果 (transferred, stayed, transfering, HA)

	強いチーム数	弱いチーム数	合計
強いと予測した数	37	15	52
弱いと予測した数	23	43	66
合計	60	58	118

表 12 予測結果 (transferred, stayed, transfering, HRA)

	強いチーム数	弱いチーム数	合計
強いと予測した数	33	10	43
弱いと予測した数	27	48	75
合計	60	58	118

表 13 モデル 2 における標準回帰係数

	intercept	transferred	stayed	transfering
標準回帰係数	0.101	−0.625	−0.329	0.098

表 14 モデル 2 における標準回帰係数

H	HR	HA	HRA
0.677	0.649	−0.838	−0.260

4.2.5 加入者数, 継続者数, 脱退者数の影響度

加入者数, 継続者数, 脱退者数の影響度を調べるのに今回実験に使用したモデル 2 から読み取る. 加入者数はオッズ比 $= \exp(-0.16) = 0.85$ となる. つまり 1 人増えると強いチームになる確率が 0.85 倍になる. 継続者数の場合はオッズ比 $= \exp(-0.1) = 0.91$, 脱退者数の場合はオッズ比 $= \exp(0.03) = 1.03$ となる. 仮に, 加入者数の値が 10 増加すると, $\exp(-0.16 \times 10) = \exp(-1.6) = 0.2$ 倍になる. 標準回帰係数から加入者数については安打数, 本塁打数と目的変数に与える影響がおおよそ近い値になった. 表 5 から加入者数, 継続者数が増加すると強いチームになる確率が下がることから強いチームは他チームから選手を獲得せず, 弱いチームが選手をより多く獲得しているということが考えられる. また脱退者数の係数だけが正であり, 強いチームは脱退者数が多いということから, 強いチームは加入者数が少ないことと合わせて考えると, 強いチームではチーム内で若手が成長して, 選手が他のチームに移籍していく傾向があるということが考えられる.

今回の所属先の変化に関する属性（加入者数, 継続者数, 脱退者数）の影響について加入者数, 継続者数, は 1 人増えると確率が下がる, 脱退者数が増えると確率が上がるという結果になった. この結果は単純に加入者数, 継続者数を減らして脱退者数増やせば成績が良くなるという解釈ではなく, 新しい加入者を呼ぶことで戦力増強を測るのではなく, 今現在の状態からいかにして能力のある個人を脱退させないかが重要という意味だと考える.

4.3 個人成績を考慮したパフォーマンスの変化分析の実験

また前年度の成績との比較の場合, 成績が極端に高すぎる場合や低すぎる場合はそれ以上の変化が見られにくくなる. そのため前年度との得点数の差が 10 未満の場合は変化なしとした. つまりパフォーマンスの変化には前年と比較し, 得点数の増加, 変化なし, 減少の 3 パターンに分類したデータを使用した.

4.3.1 実験方法

以下に実験方法を示す.

実験方法

(1) 前年度との得点数の差が 10 未満のデータをパフォーマンスの変化がなしと分類
(2) 前処理したデータの中でパフォーマンスの変化がなしに分類されたデータを排除
(3) 目的変数を得点数の増減, 説明変数を得点数の増減に関係があると考えられているチーム成績 H(安打数), HR(本塁打数) とし, ロジスティック回帰分析を行う
(4) 目的変数を得点数の増減, 説明変数を加入者数, 継続者数, 脱退者数に得点数の増減に関係があると考えられている前年度のチーム成績 H(安打数), HR(本塁打数) を重みとして加えたものとし, ロジスティック回帰分析を行う
(5) 3.4. の結果からモデルを作成する
(6) 5. のモデルに対して学習データに使用していない 2015 年から 2019 年をテストデータとし, 有効性の確認を調査する

4.3.2 個人の成績を考慮することによる精度向上実験結果

表 16 にモデル 3, 表 18 にモデル 4 のテストデータに対する分割表を示す. 表 15 にモデル 3, 表 17 にモデル 4 における説明変数の係数と切片の推定値 (estimate), その標準語差 (std.error), 有意確率を求めるための z 値 (z value), 有意確率 (pr(>|z|) を示す. モデル 3 では 82/133 チームを正しく判別した. モデル 4 では 89/133 チームを正しく判別した. 表 16 からモデル 4 の結果では実験に用いた説明変数のうち stayed_HR を除いた 5 つにおいて統計的有意がみられた. この中でも transfering_H, transfering_HR の偏回帰係数がマイナスなのは脱退者の安打数, 本塁打数が減ることは得点の減少につながるためイメージ通りの結果である. また transferred_HR が増えると得点が増加するのでこちらも目的に合っていると考える. stayed_H に関してはチームに残ったメンバーの安打数が多いことが前年度の得点の高さを表しており, stayed_H が大きいほど今年度のハードルが上がるためと考えられる. ここで問題は transferred_H についてである. 一般的には加入者の安打数が増えると前年度より成績が良くなると考えられる. つまり偏回帰係数がマイナスなのはおかしい. このことについては推測であるがこのような理由が考えられる. 前年度安打数が多いメンバーが加入するチームは前年度得点数が高い, つまり前年度強かったチームに加入することが多いのではないかと推測する. その場合, 前年度得点数が高いチームはそれ以上の得点数を取ることが前年度得点数が低いチームと比べて難しい. つまり得点数が下がる可能性が高いのである. 加えて前年度得点数高いチームに所属していたメンバーは成績が良いと考えられている. そのため前年度安打数が多いメンバーが加入してもポジション被り等の理由から出場機会が減り, その結果がこの数字ではないかと考えた.

表 15 モデル 3:所属先の変化を考慮していないモデル

	estimate	std.error	z value	pr(>|z|)
intercept	16.689	1.591	10.488	9.773e-26
H	−0.011	0.001	−9.702	2.962e-22
HR	−0.003	0.002	−1.109	2.676e-01

表 16 モデル 3:予測結果

	強くなった	弱くなった	合計
強くなると予測した数	50	37	87
弱くなると予測した数	14	32	46
合計	64	69	133

表 17 モデル 4:所属先の変化に個人の成績を考慮したモデル

	estimate	std.error	z value	pr(>|z|)
intercept	9.047	0.890	10.161	2.956e-24
transferred_H	−0.656	0.246	−2.662	7.763e-3
stayed_H	−1.730	0.193	−8.956	3.361e-19
transferring_H	−1.070	0.280	−3.822	1.326e-4
transferred_HR	1.451	0.518	2.802	5.077e-3
stayed_HR	−0.259	0.193	−1.340	1.802e-1
transferring_HR	−1.488	0.508	−2.931	3.377e-3

表 18 モデル 4:予測結果

	強くなった	弱くなった	合計
強くなると予測した数	49	29	78
弱くなると予測した数	15	40	55
合計	64	69	133

4.3.3 多重共線性について

変数間の相関関係が強いことによる多重共線性についてを VIF の値を見ることで確認する. まず表 18 に各変数ごとの VIF を示す. VIF の値が 10 未満であれば多重共線性の問題は生じない. 表 18 から強い相関関係は見られず多重共線性の問題は生じていなかった.

表 19 モデル 4vif の値

	transferred_H	stayed_H	transfering_H	transferred_HR	stayed_HR	transfering_HR
transferred_H	Inf	1.288	1.351	3.810	1.025	1.275
stayed_H	1.288	Inf	1.866	1.172	1.626	1.413
transfering_H	1.351	1.866	Inf	1.269	1.111	3.544
transferred_HR	3.810	1.172	1.269	Inf	1.013	1.269
stayed_HR	1.025	1.626	1.111	1.013	Inf	1.033
transfering_HR	1.275	1.413	3.544	1.269	1.033	Inf

4.3.4 モデルの当てはまり具合による考察

今回行った実験では所属先の変化数の説明変数に個人の成績を重みとして加えたことによる結果への影響を調べるため実験を行った. モデル 3 では野球の得点に関係がある属性 (H, HR) 2 種類を使用した. モデル 4 では加入者数, 継続者数, 脱退者数の説明変数に対して野球の得点に関係がある属性 (H, HR) を重みとして加えたものを使用した. モデル 3 では 82/133 チームを正しく判別した. モデル 4 では 89/133 チームを正しく判別した. 所属先の変化を説明変数として加えたことで見つけられなかった 7 チームを見つけることができた. またさらに目視による当てはまり具合の確認ができる ROC 曲線で評価をしていく. 青線がモデル 3, 黒線がモデル 4 を指している.

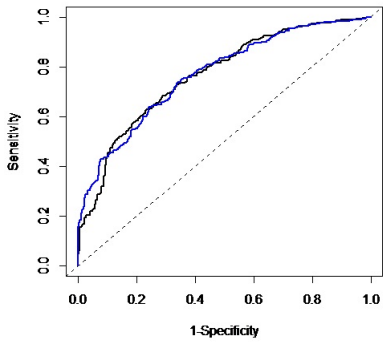

図 2 ROC 曲線結果. 青線がモデル 3, 黒線がモデル 4

図 2 の ROC 曲線結果からは目視ではどちらの方が良いか確認が出来なかった. そこで AUC で数値を分析してみた. モデル 3 は AUC=0.767, モデル 4 は AUC=0.764 とほとんど差はないという結果になった.

次に実験に使用したデータへの当てはまり具合についてもクロス集計表で確認してみる.

表 20 モデル 3 当てはまり具合結果

	強くなった	弱くなった	合計
強くなると予測した数	282	130	412
弱くなると予測した数	139	292	431
合計	421	422	843

表 21 モデル 4 当てはまり具合結果

	強くなった	弱くなった	合計
強くなると予測した数	270	109	379
弱くなると予測した数	151	313	464
合計	421	422	843

表 19 からモデル 3 の当てはまり具合は 574/843=0.681 だった. 表 20 からモデル 4 の当てはまり具合は 583/843=0.692 だった. この結果から実験データへの当てはまり具合はモデル 4 の方が良いと分かった.

今回はチームの成績から翌年の得点数が増加するか, 減少するか実験を行ったが, 重みづけしたことでものすごく精度がよくなったわけではない. この改善点としてチームの成績に関しても前年度からの変化した値にすることで改善できる可能性があると考えられる.

5. まとめ

本稿ではメンバーの所属先の変化の有効性の調査, パフォーマンスの変化の分析を目的として MLB データへの統計的分析を試みた. 所属先の変化が関係しているデータは身近にも多く存在している. グループ間の所属先の変化は様々な種類があり, その中でも本稿では移籍してくるメンバー（加入者）, 残るメンバー（継続者）, 移籍していくメンバー（脱退者）に注目し, 人数の変化がもたらす, 組織の成績への影響をロジスティック回帰を用いた分析を試みた.

加入者, 継続者が増加することで強いチームになる確率が下がる, 脱退者が増加すると翌年強いチームになる確率が上がる, という結果は単純に加入者数, 継続者数を減らして脱退者数増やせば成績が良くなるという解釈ではなく, 新しい加入者を呼ぶことで戦力増強をせず, 今現在の状態からいかにして能力のある個人を脱退させないかが重要という意味だと考えられる. 加入者数, 継続者数, 脱退者数を説明変数に加えたモデル 2 は, 加えないモデル 1 と比べて AUC が有意水準 5 %で統計的に有意な差があり, 適合度も統計的に有意だった. モデル 2 に追加した所属先の変化に関する説明変数の係数のうち, 加入者数 (transferred) と残留数 (stayed) は統計的に有意だった. したがって, 脱退者数 (transferring) はチームの強さとは関係がないが, 加入者数と残留数はチームの強さと関係があることが分かった.

また所属先の変化数に個人の成績を重みづけすることによりモデル改善を試みた. 安打数で重み付けした加入者数について一般的には加入者の安打数が増えると前年度より成績が良くなると考えられるが, 実験結果は加入者の安打数の偏回帰係数がマイナスとなり, 加入者の安打数が増えると前年度より成績が悪くなる. この結果に対しての考察として前年度安打数が多いメンバーが加入するチームは前年度得点数が高い, つまり前年度強かったチームに加入することが多いのではないかと考えられる.

今後の展望として MLB データと似た他のデータでも似た結果が得られるか実験を行う必要がある.

参考文献

[1] 橋本公雄, スポーツ競技におけるパフォーマンスを予測するための分析的枠組みの検討, 九州大学健康科学センター (2000)
[2] 永柄真澄, 人事異動の満足要因と職業性ストレス, 日本心理学会 (2011)
[3] Michael J.Crawley,「統計学:R を用いた入門書」, 共立出版 (2009)