G-013

ランダフォレストを用いた J1 リーグにおけるホームチームの試合結果予測に関する研究
A study on Predicting Home Team Results in the J1 League Using the Random Forest Algorithm

山本 英輔† 樽松 理樹†
Eisuke Yamamoto Masaki Kurematsu

1. はじめに

日本国内には、J リーグというプロサッカーリーグがある。J リーグには、J1・J2・J3 の 3 つのリーグが存在しており、各リーグ内のチーム同士で対戦し、対戦の勝敗に基づく勝ち点数にて順位が決定される。対戦方法として、対戦チームごと、ホームゲームとアウェーゲームが行われる。ここで、ホームゲームとは、チーム A から見た場合、チーム A が属する地域にあるスタジアムで行われる試合であり、アウェーゲームとは、チーム A の対戦相手チームの地域にあるスタジアムで行われるゲームのことである。各リーグにおいては優勝に加え，昇格降格を争う構成になっており、降格が決定したチームは、次年度は 1 つ下のリーグへの参戦することとなる。一方、昇格したチームは、次年度は 1 つ上のリーグへの参戦できる。このように、J リーグでは各リーグ内で各チーム同士が対戦し、優勝や昇格・降格を争う構成となっている。

また、近年は各チームの戦術が多様化している。例えば、長い時間ボールを保持し得点を狙いに行くような戦術をするチームもあれば、硬い守備を構築し相手からボールを奪い、隙を狙いカウンターで得点を狙いに行く戦術をするチームもある。

このようにチーム毎に戦術が異なると、各チームによるシュート数やボール保持率などといったデータの数値が異なってくる。例えば、ボールを長い時間持とうとするチームは、ボール保持率が高くなり、守備を固め相手からボールを奪い取り得点を取りに行くチームでは、ボール保持率が低くなることが考えられる。このように，戦術が異なれば、試合データも大きく異なる。このようなデータは，例えば「Football lab」[1]という web サイト上に公開されている。しかし、どのような戦術をすれば次の試合に絶対に勝てるかは明確ではない。例えば、ボール保持率を高くしてチームでボールを持つ戦術で全試合処理するとは限らない。

一方、ホームゲームでは、生で応援している観客が多いことから、勝利がより求められる。チームとしても、多くのサポーターに喜んでもらいたいという想いから、ホームゲームでの勝利は絶対に必要であると考えられる。よって，直近の試合結果から，ホームゲームの勝敗が予測できれば，より効果的な戦術の検討が可能になると予想される。しかし，明確な予測手法がないのが現状である。

以上の背景から，本研究では，各チームのボール保持率や守備力・シュート数などのデータを基に、ホームゲームの勝敗の予測手法の提案を行う。本手法が確立することができれば、チームの戦術の検討にも貢献できると考える。

2. 関連研究

サッカーを含め，スポーツの試合結果の予測に関する研究は，すでに取り組まれている。

サッカーを題材としたものとしては、例えば Fátima Rodrigues ら[2]は、イングランドのプロサッカーリーグである「プレミアリーグ」を対象に勝敗予測を行っている。2013/2014 年から 2018/2019 年を対象に、「得点・シュート数・コーナー数・失点数・イエロー数・レッドカード数・各試合のオッズ・試合の最終結果・試合の審判」のデータを用いた。機械学習手法として、ナイーブベイズ、KNN、ランダムフォレスト、サポートベクターマシン、決定木、Xgboost、多項ロジスティック回帰、ANN を用いて予測を行い比較した。結果、SVM が最も優れており精度が 61.32%であった。

また、Joâo Gomes ら[3]は、データマイニングの分類モデルによるサッカーの試合のベッティングの利益を増加させるための試合予測知的意思決定支援システムのための第一歩となる研究を行っている。この研究では、イングランド・プレミアリーグの 14 シーズンのデータを使用し、データの中身として、「日付・ホームチーム・アウェーチーム・ホームチームのゴール数・アウェーチームのゴール数・試合結果・ハーフタイム時点でのホームとアウェーのゴール数・ハーフタイム時点での試合結果・観客動員数・レフェリー・ホームとアウェーチームのシュート数・コーナーキック数・ファール数・オフサイド数・イエローカード数・レッドカード数」を用いた。機械学習手法として、ナイーブベイズ・決定木・サポートベクターマシンの 3 つと 10 倍交差検証とパーセント分割の 2 種のサンプリング手法を用いた。結果、サポートベクターマシンが最も優れており、精度が 50.8%であった。今後の課題として、天候や選手の休息時間を表す変数の設定や、今回しようされていない他のデータを使用するといった点が挙げられた。

さらにサッカーではないが、櫻井ら[4]は、野球の勝敗予測に、試合内のデータである長打率や出塁率等を用いて取り組んでいる。システムの概要として、 549 試合分のデータを収集し、集めたデータを訓練用とテスト用に分割し、モデルの構築を行っている。使用したデータとして、各先発野手の昨年の出塁率と長打率、昨年の先発投手の防御率、昨年のチーム救援防御率、イニング毎の点差、ホーム/ビジター、勝敗を用いた。使用モデルは、時系列データを扱うのに適している LSTM を使用している。2019・2020・2021 年シーズンのデータを訓練データ、2022 年シーズンのデータをテストデータとし、過去 6 試合のデータを入力し、一試合先の試合を出力とした。実験の結果、テスト時は正し

†岩手県立大学 Iwate Prefectural University

く予測を行えなかった。その理由として、データ数の不足や、予測に用いるデータの複雑さ、説明変数に多重共線性が起こった可能性を挙げている。

3. 提案手法

3.1 手法の概略

本研究では、J1 リーグに所属しているチームを対象に、各チームのホームチームの試合結果を、スタッツをもとに予測を試みる。予測手法として、各チームの試合内のスタッツを収集し、それらのデータから構築した予測モデルにより、対象ホームゲームの「勝ち」「引き分け」「負け」の予測を行う。

ここでスタッツとは、1 試合のプレー成績をまとめたものであり、1 試合毎にホームチーム及びアウェーチームのスタッツがある。スタッツの種類として、シュート数やボール保持率、シュート成功率などがある。詳細は 3.2 節で説明する。

N 試合目の試合結果を予測する流れは、次の通りである。なお J リーグにおいては，節という呼び方をするが，厳密には試合目と合っていないため，本研究では試合目とする。

(1) N-1 試合目のアウェーチームのスタッツから、N-1 試合目のホームチームのスタッツを予測する。
(2) 予測した N-1 試合目のホームチームのスタッツを N 試合目のホームチームのスタッツとし、この値から、勝敗結果予測する。

以後、(1)の予測を行うモデルを、スタッツ予測モデル、(2)の予測を行うモデルを勝敗予測モデルと呼ぶ。各モデルの入力と出力の流れを図 1 に示す。

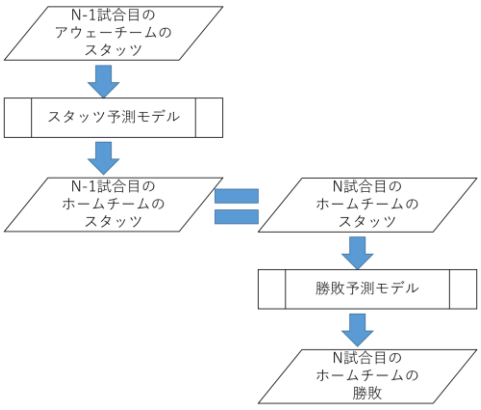

図 1　ホームゲーム試合結果予測方法

本来なら N 試合目の勝敗予測を行うには、N試合目のホームチームのスタッツが必要である。しかし、事前にその値を得ることはできない。そこで、N-1 試合目のアウェーチームのスタッツから予測する N-1 試合目のホームチームのスタッツを、N 試合目のホームチームのスタッツとして、本研究では利用する。

以後、3.2 節でスタッツについて説明した後、3.3 節、3.4 節で各モデルの構築方法を述べる。また 3.5 節において本手法で用いるランダムフォレストについて説明する。

3.2 スタッツ

スタッツとは、1 試合のプレー成績をまとめたものである。具体的には、表 1 に示す、「Football lab」[1]で公開されている値を用いる。これらのスタッツについては、各試合につき，ホームチーム、アウェーチームのスタッツがそれぞれ公開されている。これらのスタッツから各予測モデルの構築を行い、ホームゲームの勝敗を予測する。

表 1 本研究で用いるスタッツと説明

項目	説明
ゴール数	1 試合で決めた得点数
シュート数	1 試合でシュートを打った数
シュート成功率	1 試合のシュート数の内、ゴールが決まった数の割合
ボール保持率	1 試合 90 分間の内、ボールを持った時間の割
AGI	AGI とは Approach Goal Index の略であり、攻撃の際にどれくらいゴールに近づけたかを表す値
KAGI	KAGI とは、Keeo Away from Goal Index の略。守備でどれだけ相手をゴールに近づけなかったかを表す値
passCBP	パスによるチャンス構築率を表す値。CBP は Chance Building Point の略
奪取 P	ボール奪取力
守備 P	守備力

3.3 スタッツ予測モデル

スタッツ予測モデルを、回帰モデルを用いて構築する。今回は、回帰に対応したランダムフォレストを用いる。ランダムフォレストについての詳細は 3.5 節で述べる。ランダムフォレストを選択した理由は、回帰目的でも活用されることが多く、規模が大きいデータでもより早く学習と識別が可能だからである。説明変数としては、 i 試合目のアウェーチームのスタッツ、目的変数としては、同じ i 試合目のホームチームのスタッツを用いる。

3.4 勝敗予測モデル

勝敗予測モデルは、分類モデルを用いて構築する。分類モデルとしては、スタッツ予測モデルと同様にランダムフォレスト（分類に対応したもの）を用いる。ランダムフォレストを選択した理由は、分類（判別）の目的でも活用されており、規模が大きいデータでもより速く学習と識別が可能だからである。説明変数としては、 i 試合目のホームチームのスタッツ、目的変数としては、同じ i 試合目のホームチームの勝敗結果を用いる。勝敗としては、勝ち、負け、引き分けの 3 つのラベルとする。

3.5 ランダムフォレスト

ランダムフォレスト[5]とは、機械学習アルゴリズムの一つであり、複数の決定木の出力を組み合わせて 1 つの結果に到達させたものである。また、決定木は分類木や回帰木

を組み合わせたもので、ツリー(樹形図)によってデータを分析する手法である。ランダムフォレストでは分類と回帰の両方の問題を処理することが可能である。

ランダムフォレストにおける処理から予測の流れを図 2 に示す。

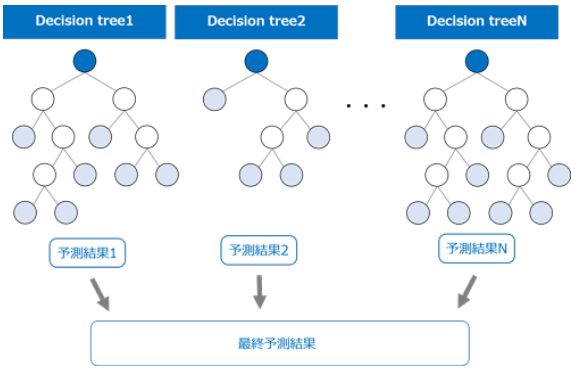

図 2　ランダムフォレストにおける処理と予測の流れ[6]

4. 評価実験

4.1 評価実験概要

本提案手法の有用性を検証するために、次の評価実験を行う。モデル構築には、「Football lab」より入手した 2023 年度の J 1 リーグのデータ、578 件（34 試合×17 チーム）を用いる。また評価においては、2024 年度の 1 試合目から 6 試合目までの J 1 リーグのデータを用いる。ここで，モデル構築に用いるデータと評価に用いるデータとではシーズンが異なる。そのため、選手や監督などに違いが生じると考えられる。しかし、変更の範囲は一部であり、その変化は小さいと仮定し、検証を行う。

本実験ではチーム単位で予測を試みる。具体的には，チーム A の 2023 年度シーズンのホームゲームのスタッツより，スタッツ予測モデル及び勝敗予測モデルを学習する。これを用いて，2024 年度シーズンの 1 試合目から 6 試合目の予測を行う。この処理を 2023 年度及び 2024 年度に J1 リーグに属していたチームに対して行う。チーム単位で行う理由は，チームごとの特徴が得られると考えたためである。各チームの学習及び評価用データ数を表 2 に示す。

本実験におけるスタッツ予測モデルと勝敗予測モデルの学習から予測の流れを図 3，図 4 に示す。

検証においては、スタッツ予測モデル、勝敗予測モデルそれぞれ評価を行う。スタッツ予測モデルについては、各スタッツの誤差から評価する。勝敗予測モデルでは、各チームの 2024 年度の 1 試合目から 6 試合目までのホームチームの勝敗を予測する。予測結果に対して、正解率、再現率、適合率、F 値の各値により評価する。これらの値は、式(1)から式(4)により求める。

$$\text{正解率} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{再現率} = \frac{TP}{TP + FP} \quad (2)$$

表 2　実験対象チームと学習データ・評価用データ数

チーム名	学習用データ数	評価用データ数
コンサドーレ札幌	17 試合	6 試合
鹿島アントラーズ	17 試合	6 試合
FC 東京	17 試合	6 試合
横浜 F マリノス	17 試合	6 試合
川崎フロンターレ	17 試合	6 試合
柏レイソル	17 試合	6 試合
湘南ベルマーレ	17 試合	6 試合
浦和レッズ	17 試合	6 試合
アルビレックス新潟	17 試合	6 試合
名古屋グランパス	17 試合	6 試合
京都サンガ FC	17 試合	6 試合
セレッソ大阪	17 試合	6 試合
ガンバ大阪	17 試合	6 試合
ヴィッセル神戸	17 試合	6 試合
サンフレッチェ広島	17 試合	6 試合
アビスパ福岡	17 試合	6 試合
サガン鳥栖	17 試合	6 試合

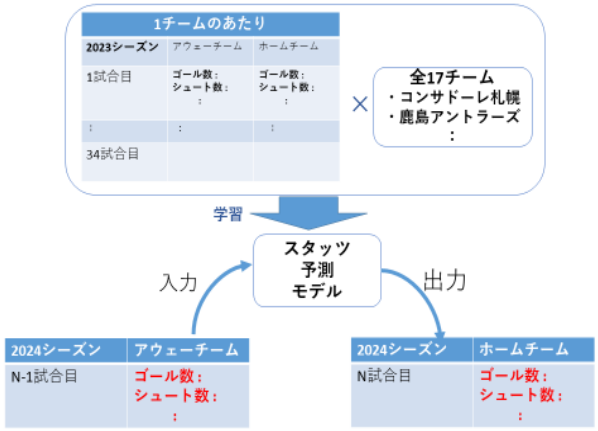

図 3　スタッツ予測モデルにおける学習から予測の流れ

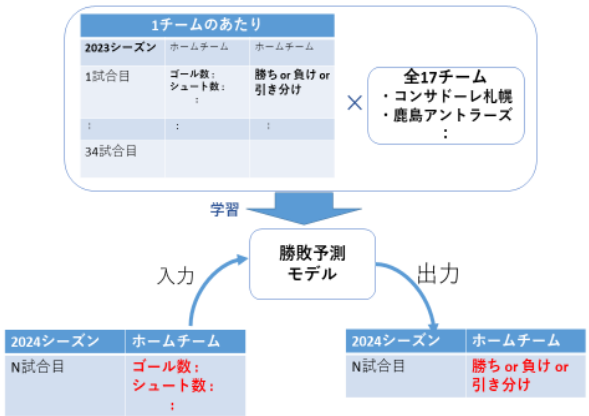

図 4　勝敗予測モデルにおける学習から予測の流れ

$$適合率 = \frac{TP}{TP + FN} \quad (3)$$

$$F\ 値 = \frac{2 \times 適合率 \times 再現率}{適合率 + 再現率} \quad (4)$$

	実際は勝ち	実際は勝ち以外
予測は勝ち	TP	FP
予測は勝ち以外	FN	TN

図 5　勝ちにおける TP・FP・FN・TN

式(1)から式(3)における TP、FP、FN、TN の各値の意味については図 5 に示す。これらの値を，勝ち・負け・引き分けのそれぞれの場合について求める。

本検証の実行においては、Python にてランダムフォレストを用いたモデルの構築を行う。モデル構築を行う際の実行環境として GoogleColablatory 上で行い、機械学習ライブラリ Scikit-learn のほか、NumPy、pandas、tqmd を使用する。

4.2 評価観点

実験結果をもとに、本提案の有用性について以下の観点から考察する。

スタッツ予測モデルの予測結果の評価基準値としては，ゴール数は実際の値から 1 以上、シュート数・シュート成功率・syubiP は 5 以上、ボール保持率・AGI・KAGI・passCBP は 10 以上、dassyuP は 20 以上とする。各要素および全組合せについて，誤差範囲内の件数から，本予測モデルの有用性を評価する。

勝敗予測モデルについては，2023 年度の勝敗数をもとにランダムで予測した結果との比較を行う。ランダムよりも高い予想結果が得られれば、本手法は有用である可能性が高いと判断する。

また本提案手法は 2 つの予測モデルを用いている。そのため、スタッツ予測モデルの誤差と，勝敗予測モデルの精度との比較も行う。

以上の評価方法に加え、次にあげる点についても行う。

(1) 構築した予測モデルの比較に基づく、チーム特性の分析
(2) 他のモデルで予測した結果との比較、または統計データとの比較
(3) 全チームのデータを用いた予測との比較
(4) 昇格により前年度のスタッツがないチームの予測とその評価

5. おわりに

本研究では、自チームのスタッツを基に、ホームゲームの勝敗の予測について提案した。提案手法は、アウェーチームのスタッツからホームチームのスタッツを予測するスタッツ予測モデルと、ホームチームのスタッツから勝敗を予測する勝敗予測モデルの 2 つのモデルから構成される。またそれぞれの予測モデルの構築には、回帰・分類に対応したランダムフォレストを用いる。

今後の課題としては、提案手法の有用性を確認するため、実データを用いた検証と評価、予測モデルの比較によるチームの特性分析、他のモデルとの比較などが挙げられる。

参考文献

[1] データによってサッカーはもっと輝く Football Lab : https://www.football-lab.jp（最終アクセス日 2024 年 6 月 13 日）
[2] Fátima Rodrigues, Ângelo Pinto, “Prediction of football match results with Machine Learning”, Procedia Computer Science, Vol.204, pp.463 – 470 (2022)
[3] Joâo Gomes, Filipe Portela, Manuel Filipe Santos : “Decision Support System for predicting Football Game result”, Recent Advance in Computer Science, ISBN : 978-1-61804-320-7 (2015)
[4] 櫻井 一成, 冨山 蓮 , “機械学習を用いた野球の勝敗予測に関する研究”, 南山大学(2022)
[5] ランダムフォレストとは ： https://www.ibm.com/jp-ja/topics/random-forest（最終アクセス日 2024 年 6 月 13 日）
[6] ランダムフォレストによる予測モデル作成 ：https://www.stats-guild.com/analytics/12543（最終アクセス日 2024 年 06 月 13 日）