

勾配ブースティング決定木を用いた イングランドプレミアリーグの試合結果予測

Forecasting the results of English Premier League matches using Gradient Boosting Decision Trees

2041060 志多野 健太

Kenta SHITANO

指導教員 秋葉 知昭

In recent years, the popularity of football has surged worldwide and broadcasting rights are on the rise. The English Premier League (EPL) has seen particularly strong growth. According to the data in 2021, 3.2 billion viewers a year had paid attention to matches [1]. The EPL has become the most watched league in the world. With the growing popularity of football, match predictions, such as the 'Sports Lottery WINNER', are attracting attention as a new form of entertainment. In this study, three gradient boosting decision trees (it is called GBDT) are used to create a model with a high percentage of correct answers. The results showed a 61.5% correct prediction rate, but the inability to correctly classify the matches into draws was a challenge.

1. 諸言

近年、世界中でサッカーの人気の上昇し、放映権も増加傾向にある。特に、イングランドプレミアリーグ（以下 EPL）はその成長が著しく、2021 年のデータによれば、年間 32 億人もの視聴者が試合に注目し、世界で最も視聴者数の多いリーグとなっている[1]。このサッカーの人気の拡大に伴い、「スポーツくじ WINNER」など、試合の予測が新たなエンターテインメントとして注目を集めている。そこで本研究では、3 つの勾配ブースティング決定木（以下 GBDT）を用いて正解率の高い試合結果予測モデルを作成する。

2. 勾配ブースティング決定木 (GBDT)

GBDT はアンサンブル学習の一種であり、弱学習器である決定木を順番に訓練していき、前の決定木の誤差を次の決定木で補正していく手法である。誤差を繰り返し補正することで高い精度を期待できる機械学習手法である[2]。

本研究では Xgboost, Lightgbm, Catboost 計 3 種の GBDT を用いて学習する。

3. 先行研究

田中・細井の研究[3]では、2017 年から 2021 年のイロレーティング、ゲーム会社が出している選手能力値、ホーム&アウェイ、審判、を特徴量として収集し、Light gbm を用いて、勝、引き分け、負けの 3 値に分類するモデルを構築した。結果は正解率 58% のモデルが完成している。

このモデルの課題として引き分けへの分類がされていないこととイロレーティングがモデルに与える影響度の低さが挙げられた。

4. 研究内容

4.1 本研究について

本研究では、2018 年から 2023 年 EPL 全 1900 試合のデータを用いて、ホームチームの勝、引き分け、負けの予測を行う。先行研究から、2 つの GBDT と 7 種の特徴量を追加し、正解率 58% より高いモデルを作ることとする。

4.2 使用するデータ

一般利用可能なサイトから 2018 年から 2023 年を収集対象とし、7 種の特徴量を追加した。追加した特徴量をまとめたものを表 1 に示す。ppg は先行研究で使用した、選手の能力値に掛け合わせている。特徴量は全てホーム、アウェイ 2 チームずつのデータを収集している。

表 1 追加した特徴量

データ名	解説
WDL	直近 5 試合の勝、分、敗戦数
5point	直近 5 試合の勝ち点数
Money	その年の移籍金予算
Manager	監督がそのチームで獲得した平均勝ち点
Income	その年に得た移籍金
ppg	前節の選手の出来高（機械採点で決まる）
Rank	前節までのチームの順位

4.3 実験方法

本研究はGBDTを実装する段階で交差検証とoptunaを用いる。交差検証は5年分のデータを使うため5分割にし、optunaはまとめサイト[4]を参考にしてパラメータ範囲を設定する。分類先は勝,分,負の三値に分類し,正解率が最も高いモデルを最適モデルとする。

各GBDTを実装後,正解率が最大になるモデルの,特徴量の重要度,混合行列を可視化し考察する。図1にXgboostの特徴量の重要度を表したグラフを示す。

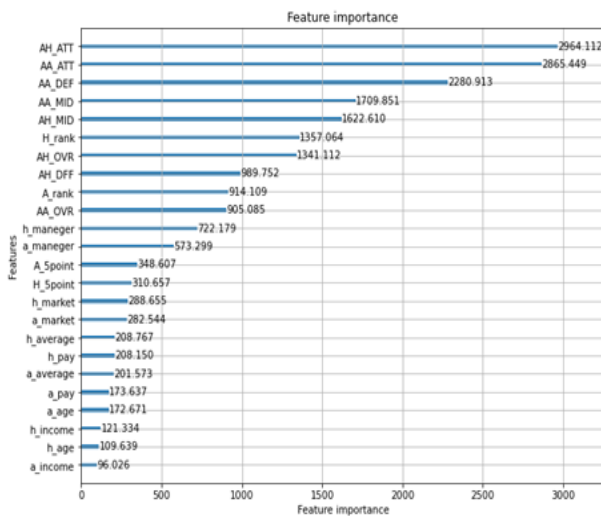


図1 Xgboost 特徴量の重要度

4.4 結果及び考察

表2に各GBDTの結果を示す。

正解率はXgboostが最も高くその次に,Light gbm,Catboostの順に高かった。

Light gbmの正解率がXgboostを下回った要因は,一部をサンプリングする性質上,正解率を上げる特徴量を早い段階で切り捨てられた可能性があると考ええる。実際,Light gbm以外が一番重要度が高い特徴量はいずれも前節の選手の調子(ppg)を含むデータであったのに対し,Light gbmは前節までの順位(rank)であった。

Cat boostの正解率が低かった要因は,質的変数の処理に特化した学習器であるCat boostに対して,質的変数を使わなかったことだと考える。

今回最大61.5%の正解率で,田中・細井の正解率58%より高い数値を出すことができたが大幅な改善はみられなかった。

正解率が上昇した要因は影響度の高い特徴量を扱集できたことだと考えられる。図1から特徴量の

重要度上位5つはいずれも選手能力値とppgを掛け合わせた値であった。次に,前節のチーム順位となっている。このことから,試合結果予測において,時系列上で試合に近いデータが特に有益であると考えられる。本研究では収集できなかったが,怪我をした選手やスターティングメンバーなどより試合に近い時系列データも有効であると考えられる。

正解率が伸び悩んだ要因として,引き分けへの分類の改善ができなかったことが挙げられる。負けの分類は70%,勝は80%になるが引分の分類は2%とほとんど正解へ分類されなかった。また,改善できなかった要因として,期待値を大幅に下回る又は,上回ることが容易に起こりうる,サッカーの不確実性に考慮できなかった可能性がある。改善するには,引き分けの試合についてより細かい分析を行い,適切なデータ,学習手法を探す必要がある。

表2 GBDT 正解率

GBDT	Xgboost	Light gbm	Cat boost
accuracy	61.5%	60.5%	57%

5. 結言

本研究は,田中・細井[3]の研究に対し,2つのGBDTと7種の特徴量を追加することで正解率の改善を計った。時系列データの特徴量が機能し,正解率は上昇したものの大幅な改善には至らなかった。今後の課題は,引き分けへの分類の改善である。本研究では引き分けの試合に対しての正解率は2%だった。そのため,引き分けの試合を分析し,引き分けの分類を正確に行える手法やデータを見つけることで大幅な改善に繋がると考える。

文 献

- [1] allstars : プレミアリーグのイギリス経済への貢献は約1兆1900億円
<https://www.all-stars.jp/news/premier-league-united-kingdom-economy/> (2024年時点)
- [2] RAPIDMINER : Gradient Boosted Trees
https://support.rapidminer.jp/rapidminerstudio/operatorslist/9.5/modeling/predictive/trees/gradient_boosted_trees (2024年時点)
- [3] 田中聡太朗 細井祐希 : 機械学習手法を用いたサッカーの試合結果予測モデルの構築とその活用
日本大学文理学部情報科学科谷聖一研究室(2021)
- [4] @m_yama123(M Yamaguchi) : Xgboostのハイパーパラメーターまとめてみた
https://qiita.com/m_yama123/items/133e5b73b2e2b0d6bd6a (2024年時点)