サッカーの勝敗要因に関するデータ分析

伊藤 光毅(19X4005)　鹿野 紘樹(19X4022)　北林 良太(19X4024)　指導教員 劉 慶豊

1. はじめに

スマートフォンの普及やSNSの発達により，スポーツをいつでも好きな時間に観ることが出来るようになった．こうしてスポーツは，人々にとってより身近になった．またスポーツ観戦の伴いあらゆる企業がスポーツと関連するビジネスに参入し収益を獲得している．こうしてスポーツは世界経済に貢献する１つの産業として目まぐるしく成長してきた．昨今のワールドカップの盛り上がりを考えると，その中で最も注目されるスポーツはサッカーであると考える．

本研究は「機械学習」という手法を用いて「サッカーの勝敗要因」に関して分析する．そして分析結果を試合戦術に取り込みチームの勝率に貢献し最終的にはサッカーの経済発展に貢献できる研究になると期待している．

2. 利用するモデル

2.1 ロジスティック回帰

ロジスティック回帰分析は，統計学において分類問題を解くために使用する手法である．ロジスティック回帰の利点は以下の 4 点である．

① 解釈性が高い

ロジスティック回帰は回帰係数を推定し，各特徴量（説明変数）の重要度を算出することができ，予測結果がどの特徴量によって決まったかを理解することができる．

② 高い汎化性能

ロジスティック回帰は離散変数と連続変数の両方を取り扱うことができるため，汎化性能が高い．

③ 不均衡データに対しても有効

クラス分類においてのクラスのサンプル数が不均衡であっても有効．

④ 多クラス分類にも対応

多項ロジットを利用すれば，多クラス分類に対応可能．

2.2 ニューラルネットワーク

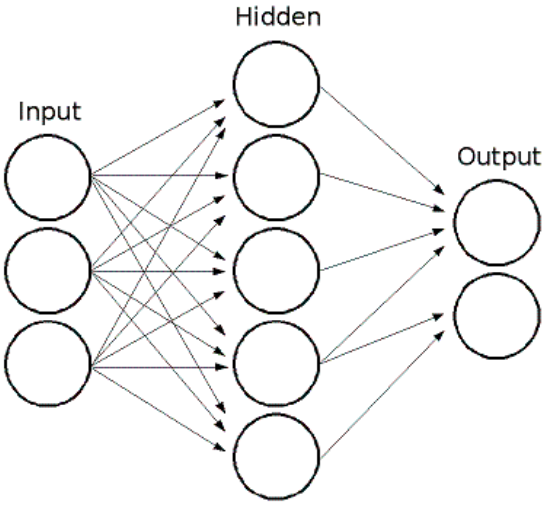

図 1. 階層型ニューラルネットワーク

ニューラルネットワークは複雑な非線形関係をモデル化するために，複数の層に分かれて，それぞれの層が異なるタスクに対応している人間の神経細胞の原理を真似した予測用ネットワークである．

ニューラルネットワークには，入力層，隠れ層，出力層の 3 種類の層からなる．各層は，入力層における入力信号，隠れ層におけるニューロン，出力層における出力信号がある．

入力層　x_i　(1)

隠れ層　$z_j = \sum_{i=1}^{m} w_{ij} x_i + b_j$　(2)

出力層　$y_k = \sum_{j=1}^{s} w_{jk} z_j + b_k$　(3)

w_{ij} は入力層iから隠れ層jへの重み，b_jは隠れ層jのバイアス，w_{jk}は隠れ層jから出力層kへの重み，b_kは出力層kへのバイアスである．この重みとバイアスは学習過程でニューラルネットワークが自身の改善のために更新される．

2.3 勾配ブースティング決定木(GBDT)

GBDT (Mason, et al, 1999)はアンサンブル学習の一つであるブースティングと勾配降下法，決定木の三つ組み合わせた手法である．テーブルデータに対しては，他の機械学習モデルに比べて精度・計算速度が優れているので，データ解析コンペティションの回帰・分類問題ではよく用いられる手法である．

3. 実験概要

本研究は欧州 5 大リーグ（イングランド・スペイン・ドイツ・フランス・イタリア）の過去 5 年分の試合データをもとに分析を行った．目的変数を「ホームチームの試合の勝敗」とし，説明変数は，各チームのシュート本数やコーナーキック回数など計 23 項目を説明変数とした．

また，ロジスティック回帰，ニューラルネットワークと GBDT の 3 つの分析手法を用いて勝敗要因を分析した．

4. 分析結果

表 1:各モデルの交差エントロピー誤差

モデル＼指標	交差エントロピー誤差
ロジスティック回帰	0.5206
NN	0.5279
GBDT	0.5102

表 1 は各モデルの分析結果の評価指標を比較したものである．表 1 より，3 つの手法の中で GDBT によるモデルが優れていることがわかる．

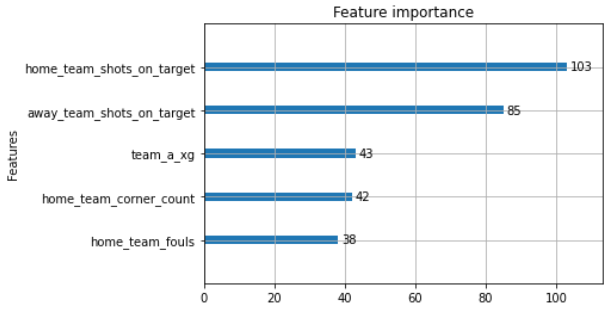

図 2: LightGBM による各特徴量の重要度

図 2 は LightGBM において重要な特徴量上位 5 項目を示したものである．

LightGBM による分析では，ホームチームの枠内シュート数が最も大きい重要な特徴量であること，次いで，アウェイチームの枠内シュート数が重要な特徴量であることが示された．また，ホームチームのゴール期待値は，アウェイチームのゴール期待値よりも重要度が高いことが示された．さらに，ホームチームのコーナーキック・ファウルの数もホームチームの勝利に対して重要な特徴量であることが読み取れた．

5. 考察

各モデルを比較して重要度が共通して高かった説明変数は，ホームチーム，アウェイチームの枠内シュート本数である．シュートを多く放っているチームが勝利の確率を上げていることが分かる．重要だと予想していたボール保持率の重要度は高い数値は出ず，大きく関係はしないということが分かる．ポゼッションサッカー，カウンターサッカーどちらが良いかという点より，いかにシュートまで持っていけるかが重要であると考える．ホームチームのコーナーキック数の重要度が高いという点から，サイド攻撃からコーナーキックを得るようなプレーが推奨できる．また，ホームチームのファウル数に関しては，守備において，セーフティな守備をし，ファウルにならないような守り方をするべきである．

6. おわりに

本研究では，サッカー試合の結果に大きな影響を与える要因を検証し,重要度の高い特徴量を示した.その結果をもとに，試合の有効な戦略を検証し，提案した．本研究では，簡単なスタッツデータしか取得することができず，もし工夫して様々なデータを取得することが可能であれば，より詳細な分析を行い，モデルの精度を上げることができると考えられる．今後の研究テーマにしたい．

参考文献

[1] Mason, Llew, Jonathan Baxter, Peter Bartlett, and Marcus Frean. "Boosting algorithms as gradient descent." Advances in neural information processing systems 12 (1999).

[2] Footy Stats
https://footystats.org/jp/download-stats-csv