7C-06

競馬における荒れるレースの予測に関する研究

靱勝彦 †

クォンツ・リサーチ株式会社 †

1 はじめに

日本中央競馬の出馬表，成績表等のデータを用い，機械学習を用いた競馬予想の手法の開発及びその有効性の検証を行った。単勝オッズに着目し，勝ち馬のオッズが 10.10 倍以上で決着するかどうかの 2 値分類モデルの構築を行い，そのモデルによる予測の有用性を確認した。

2 関連研究

Sung ら [1] はファンダメンタル変数を入力とし，ステップ１でファンダメンタル変数をロジットモデルでパラメタ計算，ステップ２でオッズから逆算された勝率を合わせて入力としパラメタ計算を行う，２段階による条件つきロジットモデル分析で勝率を予測している。Silverman[2] は条件つきロジットモデルで勝ち馬の予測を行うが，Frailty スコアという項を組み合わせている点が特徴である。この Fraity 項はオッズの逆数に基づく項でありこの項によりオッズの割には勝率の高い馬を予測することが可能である。

本研究ではオッズに着目しオッズの高くなりそうなレースの選別を機械学習により行うことを目的とする。

3 分析方法

3.1 データ及び手法

2010 年 1 月 1 日から 2014 年 12 月 31 日までの 16520 レースを学習用，2015 年 1 月 1 日から 2016 年 12 月 31 日までの 6908 レースを検証用データとして用いた。

オッズの分析に当たっては，単勝オッズのみについて行い，「荒れる」の定義及び，「荒れる」際の条件パラメタを RandomForest を用いて機械学習し検証した。

3.2 荒れの定義

「荒れ」を定義するため，学習用データの 1 着馬の単勝オッズのヒストグラムを**図 1** に示す。1 着馬のオッズのピークは 0.5 あたりにあるが右方向に長いグラフとなっている。対数軸であることを考慮すると，1.0 以上はオッズで 10 倍以上を示している。また 1 着馬の単勝オッズの統計量は Median が 5.0，Mean が 10.42，第３四分位数は 10.10 である。そこで，単勝オッズ 10.10 を荒れる荒れないの閾値とにした。

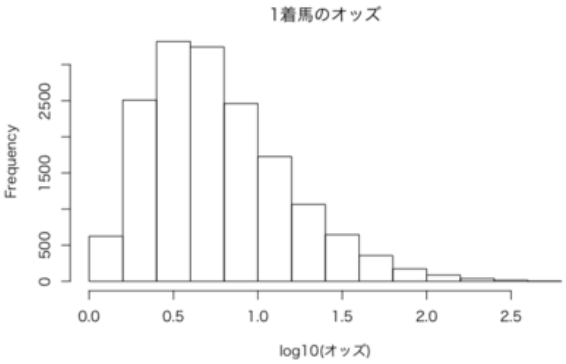

図 1: 1 着馬のオッズ

さらに、単勝人気と 1 着になった時の期待値の関係を**図 2** に示す。単勝の期待値は控除率 20%から, 80%[3] が期待値となる。1 から 8 番人気までほぼ 80%近辺であるが, 9 番人気以降は極端に期待値が落ちている。これは穴馬への過剰な投資が行われており，1 着になる確率から計算されるオッズが歪められているため（穴馬バイアス）である [4]。よって，10 倍以上となる荒れるレースを予測することにより，期待値の低くなるレースを選別することには意味があると言える。

4 予測モデル

4.1 分類器の選定

1 着のオッズが 10.10 以上か未満かという 2 値分類を RandomForest を用い行い分類した。パラメタは，競馬場，時間帯，開催月，出走頭数，レース番号，トラック条件，距離，競争条件，重量条件，当日の馬場状態の中から個別にデータの確認を行った結果，1 着のオッズが異なる傾向が現れたパラメタであるレース番号、競馬場、重量条件、出走頭数、トラック条件を

A study of prediction of a big upset at horse racing
Katsuhiko UTSUBO †
Quants Research Inc. †
105-7209, Tokyo, Japan
utsubo@qri.jp

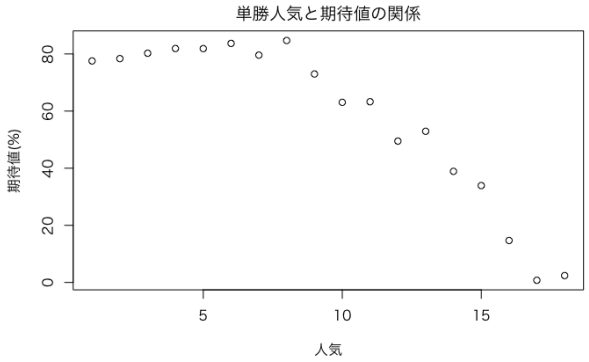

図 2: 単勝人気と期待値の関係

表 1: RandromForest による 2 値分類

	予測 10.10 以上	予測 10.10 未満
10.10 以上	515	1124
10.10 未満	1292	3725

正解率	適合率	再現率	F 値
63.7%	31.4%	28.5 %	29.9%

用いた。

正例とする 10.10 倍以上のレースはデータ全体の 1/4 であるため，SMOTE アルゴリズム (Synthetic Minority Over-sampling Technique) により，不均衡データを解消した。訓練データにより学習させ，テストデータを評価した結果を**表 ??**及び**表 1** に示す。正解率 63.7% となり，F 値に関しては 29.9%となった。

4.2 信頼度による分別

分類の際には分類器から最終的に出力される 0〜1 までの確率を 0.5 を基準として 2 値に分類している。ここで，確率が 0.5 近辺のものと，0 もしくは 1 近辺の値で分類されるものでは信頼度が異なると考え，RandomForest で出力される確率別に分類してみた結果を**表 2** に示す。

下限 0.5 は正例（10.10 倍以上）と負例（10.10 未満）がほぼ均衡した予測となることを意味している。また上限 1.0 は正例となる確率が高いと予測されることを意味しており，下限 0.5 上限 1.0 は全データを用いた場合を示し，正例，負例の全体の正解率は 63.7%である。0.1 刻みに 0.5 から 1.0 まで正例の信頼度を変化させた場合に，正解率はほぼ増加していると言える。

また，正例の予測も 10.10 倍以上となる確率は，10.10 を分けた理由が**図 1** からもとまる 1 着馬の第 3 四分位数の点であることを考えると，25%である。このことから適当に予測した場合 25%であり，これに対し，本

表 2: RandromForest による正例の信頼度による正解率

下限	上限	データ数	全体正解率
0.5	1.0	6640	63.7%
0.5	0.6	1081	50.2%
0.6	0.7	1227	61.5%
0.7	0.8	1441	64.6%
0.8	0.9	1325	67.1%
0.9	1.0	1445	70.7%

下限	上限	予測 10.10 以上	予測 10.10 未満
0.5%	1.0%	28.5%	76.8%
0.5%	0.6%	27.5%	77.4%
0.6%	0.7%	26.9%	76.0%
0.7%	0.8%	30.0%	76.1%
0.8%	0.9%	28.5%	77.3%
0.9%	1.0%	31.3%	77.6%

モデルでは 27.5〜31.3%という結果が得られた。

5 まとめ

レースの荒れの定義として単勝オッズ 10.10 倍以上が出やすいかどうかを確率で示せるかどうかの確認をした。レース番号、競馬場、条件・重量、頭数、トラックを入力とし，RandomForest で単勝オッズ 10.10 倍以上かどうかの 2 値分類を RandomForest 行い，予想モデルの優位性を確認した。さらに，分類の際の確率を考慮することで，統計的に 10.10 倍以上が出る確率が 25%に対し，本モデルの予測では 31.3%の確率で予測できた。

参考文献

[1] M. Sung , J.E.V. Johnson: *Comparing the effectiveness of one-and two-step conditional logit models for predicting outcomes in a speculative market*, The Journal of Prediction Markets , (2007).

[2] Silverman, Noah:*Optimal Decisions with Multiple Agents of Varying Performance*,UCLA Electronic Theses and Dissertations,(2013).

[3] JRA: 馬券のルール，`https://www.jra.go.jp/kouza/baken/index.html`,2020/12/8 取得.

[4] 芦谷政浩: 「穴馬への過剰な選好 *(longshot bias)*」に関するサーベイ, 国民経済雑誌,(2010).