国内競馬における機械学習及び
オッズの歪みを用いた購入法

長倉大輔研究会

田中奎帆*　出口雅也*　山田敬士*

2020 年 11 月 12 日

概要

本稿は、競馬予想の精度の向上を目的とする。競馬予想について 2 つの異なった視点から考察し、回収率による評価を行った。

（1）機械学習による予想

（2）オッズの歪みに着目した予想

（1）ではデータを距離・芝かダートかの 2 つで分類した後、この分類データごとに LightGBM とランダムフォレストで予想モデルを作成し、テストデータの回収率で評価を行った。特徴量や目的変数に工夫を加えたものの、予測結果に基づいて購入するのみでは回収率で 100%を超えることが出来なかった。予測結果に加え、オッズを考慮し割安な馬券を買うことにより回収率で 100%を超えることが出来た。（2）では単勝の値から馬単の理論値を算出し、実際の馬単オッズと比較することにより、割安だと思われる馬券を購入した。2019 年に JRA が主催した GI 全 24 レースのみのシミュレーション結果であるが、回収率で 100%を大きく超えることが出来た。[1]

*慶応義塾大学経済学部 3 年

[1] 今回の論文を執筆するにあたって、長倉大輔教授（慶應義塾大学経済学部）及び長倉大輔研究会の各員から大変有益な助言を頂いた。ここに記して感謝を申し上げたい。

目次

1. 序論

1.1 研究背景と先行研究

これまで競馬は様々な方法により予測されてきた。その予想方法はデータに基づくものからオカルト的なものまで様々である。近年、競馬データが整備され、豊富なデータに基づく予想が可能となっている。本論文ではそのデータを用いて、機械学習による予想とオッズの歪みに着目した予想という 2 つの異なった視点から考察し、競馬予想の精度の向上を目的とする。

機械学習を用いた競馬予想には大きく分けて 2 つの段階がある。1 つ目は競馬予想モデルの構築、2 つ目はそのモデルを元にどのように馬券を買うかである。競馬予想モデルでは順位やタイムを予測することが多いが、最終的に回収率を上げることが目的である。従って、競馬予想モデルが優れていても、買い方を誤っていれば結果を出すことが出来ない。本稿では、競馬予想モデルの構築と共に最適な買い方を論じる。

詳しく機械学習を用いた競馬予想について書かれた論文はないが、イベントや個人のブログなどでは盛んに行われている。論文として具体的な文献を明示することは出来ないが、全体的な傾向を述べる。競馬予想モデルでは目的変数として順位を用いた分類と走破タイムを用いた回帰の 2 つの手法がある。従来の研究では順位を用いた分類によるものが多い。走破タイムは馬場やレース展開に左右されるところが大きく、馬の実力として予想することが難しいためである。本論文でも分類の手法を用いる。使われるアルゴリズムとして多いのは LightGBM[2]であり、本論文では LightGBM に加えてランダムフォレストを用いた。

先行研究ではすべてのデータを分割せずに訓練データとして用いることが多かったが、レース条件に応じて傾向は変わると考えられる。例えば、1200m のレースと 3200m のレースや、芝とダートのレースでは傾向が違う。そのため、本論文では距離と芝・ダート別でデータを分類し、その分類データごとにモデルを作成した。予測をする際には、この分類の中であてはまるモデルによって予想を行う。例えば、1200m 芝のレースを予測する際には 1200m 芝の予測モデルを使用するということである。

一方、オッズの歪みに着目した予想の先行研究としては、小幡・太宰(2014)の研究がある。この研究では、一般の競馬ファンが過剰に大穴馬券を購入するために、その馬券が当たる本来の確率からすると割高なオッズになっており、逆に本命サイドの確率が相対的に高い馬券は割安になっており、そのバイアスは馬券が当たる確率が低いものほど大きくなっていることを実証している。具体的には、馬券の客観確率と主観確率を推計し、オッズが高い馬券ほど、客観確率に対して主観確率が大きくなっていることを実証している。こ

[2] 米マイクロソフト社よりリリースされている決定木アルゴリズムに基づいた勾配ブースティングのフレームワークである。様々なコンペティションでも用いられている。

こでは、配当額が高く、ギャンブル性も高いことから、こうした穴馬バイアスが大きいものとして三連単を挙げており、その支持率を主観確率としていた。一方で、単勝は、そうしたギャンブル性が低く、競馬の知識が豊富な投票者が購入層の中心であるため、穴馬バイアスは小さいとし、その支持率から求めた三連単の確率を客観確率としていた。

この結果を踏まえ、オッズの歪みに着目した予想では、客観確率よりも主観確率が小さい、割安な馬券を購入するというアプローチで、馬券の回収率を上げるような購入方法を導出した。

1.2 本稿の構成

本稿の構成を以下に示す。まず、第二章で機械学習を用いた競馬予想について述べる。分析手法や特徴量について述べた後、目的変数や購入方法を変更し、回収率によってそれぞれ評価を行う。次に、第三章でオッズの歪みに着目した予想について述べる。予測方法について述べた後、データを用いて実証を行う。最後に、まとめと今回の分析による課題や改善方法を述べる。

2. 機械学習による予想

2.1 使用したデータ

本章では JRDB よりダウンロードした 2015 年から 2019 年までの中央競馬のデータを用いる。新馬・未勝利のレースについては予想のためのデータが少ないため、除外した。また、取消や失格などで順位の情報が欠損している馬も除外している。

有料の競馬データの代表的なサービスとしては JRA-Van Data Lab と JRDB の 2 つが挙げられる。どちらもデータが充実しているが、JRDB は独自の指数のデータが充実しており、予想の幅を広げるために JRDB のデータを用いた[3]

2.2 分析手法

序論で述べたように走破タイムを用いた回帰の手法は、馬場やレース展開など馬の能力に関係ない影響が大きいため、順位を用いた分類を行う。しかし、順位をそのまま目的変数として用いるのは不適切であるといえる。私たちが予想したいのは馬券に絡む可能性がある 3 着以内の馬であり、13 着の予想等それ以外の予想が的中してもメリットがないからである。よって、単勝[4]を予想するために 1 着かどうかを予想することや、複勝[5]圏内を予想するために 3 着以内かどうかを予想すること等、1〜3 着のどこまでを予想するかが問題となるが、ここで問題が生じる。訓練データのうち True ラベルが False ラベルに比べて極端に少ないと、全てを False と判定してしまうことが多いため、多くのモデルではうまく学習することができないのである。つまり、1 着かどうかを予想するモデルでは True ラベルが少ないため、全てのデータを 1 着以外と予想するモデルが出来てしまうことが多い。そのため、先行研究では 3 着以内に入る馬を True、それ以外を False とした分類問題を解くことが多く、本論文でもその手法を用いる。

分類の手法として今回使うアルゴリズムはランダムフォレストと LightGBM の 2 種類である。どちらの手法においても、データセットを訓練データとテストデータの 2 つに分けて、訓練データで作成したモデルをテストデータに適用しモデルの精度を確認する。今回の分類でいうと、正しく分類が出来ているという正答率がモデルの精度の指標となるが、私たちが最終的に目的としているのは分類の正答率の向上ではなく、いくらお金を賭けていくら返ってきたかという回収率の向上である。そのため、出来上がったモデルにより回

[3] 競馬で 1 億 5000 万円を稼ぎ、馬券裁判の著者として有名な卍氏も当データを用いており、データの信頼性には問題がない

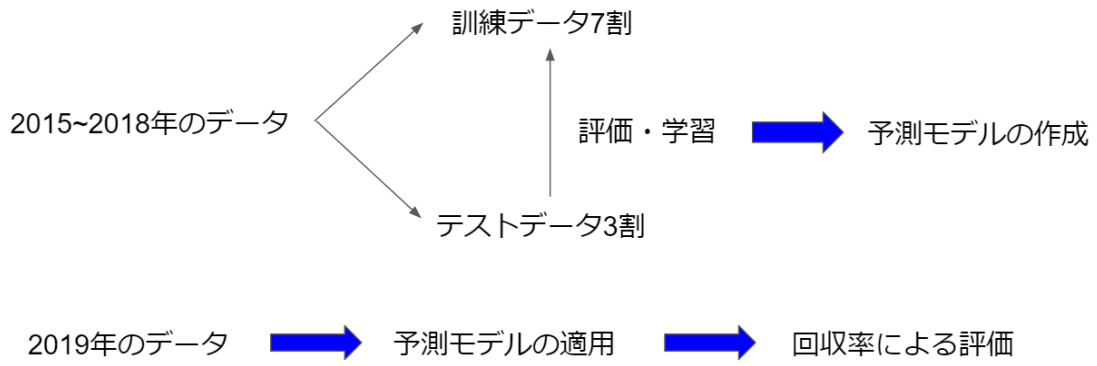
[4] レース内で 1 着になる馬を 1 頭予想する馬券

[5] レース内で 3 着以内に入る馬を 1 頭予想する馬券

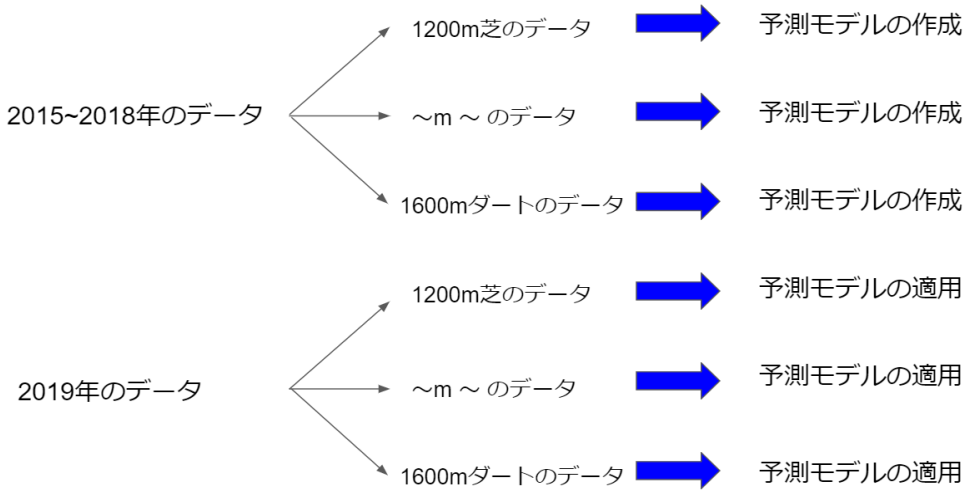
収率を算定することでモデルの精度を確認するが、回収率の算定をテストデータで行うとモデルの過剰評価となるのではないかと考えた。テストデータでの分類の正答率が良くなるようにパラメーターの調整を行い、学習させるからである。回収率が 100%を超えていても、この回収率がテストデータにより算出されたものだとすれば、実際に運用して馬券を買っても 100%を超えない可能性がある。よって、モデルの評価を行うためにモデルの構築に関わらない別のデータを用意することとした。これもテストデータであるが、上記のテストデータと混同するため実戦データと呼ぶ。以上のことを今回の分析に対してまとめると、2015 年～2018 年のデータを訓練データとテストデータに分割し学習を行い、作成したモデルを実戦データである 2019 年のデータに適用することで算出された回収率で評価する。

先行研究との大きな違いは条件ごとに分けてモデルを作成するということである。競馬は距離、芝・ダート、馬場状態、競馬場などによって求められる能力が違い、ある特定条件に限りとても強い馬というのが存在する。よって、全てのレースに一律のモデルを適用するというのは難しく、条件ごとに予想モデルを作成することで的中率が上がると考えられる。どこまで細かく分類するかというのは難しく、例えば有馬記念というようにレース名単位で分類を行うと各モデルの構築に使えるデータ数が少なくなってしまう。本論文では特に影響を及ぼすと考えられる距離と芝・ダートの 2 つについてのみ分類を行い、分類ごとに一定のデータ数を確保した。

図表 2.1　分析手法の流れ



図表 2.2　データの分類方法



2.3 特徴量の選定

競馬の結果を決めるものは大きく分けると、馬・騎手の能力とレース条件の 2 つである。馬・騎手の能力に関する特徴量として、IDM、騎手指数、馬体重、3走前までの順位、3走前までの上がり 3F のタイム、3 走前までの 1 着とのタイム差、脚質、性別、馬齢が取得できた。IDM とは JRDB が独自に算出しているスピード指数である。走破タイムを基準として、馬場・斤量・出遅れや不利等のレース内容・位置取り・レースペースの影響を反映させることによって算出される。これにより、比較が難しい前走以前のレースパフォーマンスを数値で比較できる。また、騎手指数とはオッズと騎手の連対率の関係を元に算出されたものであり、騎手の能力を数値化している。これらの特徴量の中で、性別と馬齢については意味のある特徴量となっていなかったため除外した。

レース条件に関する特徴量には同レースに出走する各馬それぞれで異なるものと同じものがある。例えば、馬番は各馬で異なるが、馬場や距離などは同レースに出走する各馬で同じである。今回の分析において後者はレース分類の時点である程度反映しており、また同レースに出走する各馬で同じならば特徴量に加えても意味を持たないと考え、除外した。特徴量に使ったのは、馬番、斤量、前走からの期間（何週間）、前走距離である。

オッズについては特徴量として用いていない。オッズと勝率は高い相関があり、オッズを加えることで的中率が上がる可能性が高い。しかし、いくら的中率が高くとも回収率が高くなければ予想モデルとして意味を持たない。オッズの高い馬を的中させ、回収率を向上させるために今回の分析では特徴量から除外した。

以上より、決定した特徴量について説明と理由を図表 2.3 にまとめた。IDM はそのまま特徴量として使うのではなく、同レースに出走する馬の平均値と標準偏差で標準化をしている。過去のパフォーマンスがどれくらい優れているのかは同レースに出走する馬との相対比較で判断すべきであるからだ。例えば、IDM50 の馬を考える。この馬が IDM 平均値

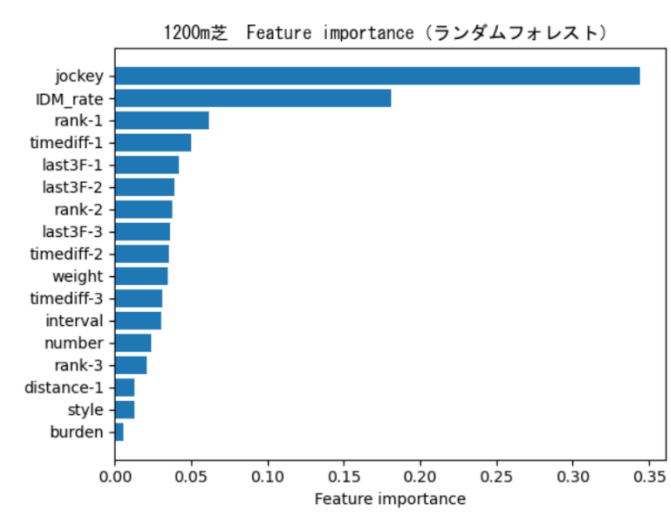
30 のレースに出走する場合は勝つ確率が高いが、IDM 平均値 70 のレースに出走する場合は勝つ確率が低い。このような状況に対応するために、IDM は標準化して用いている。

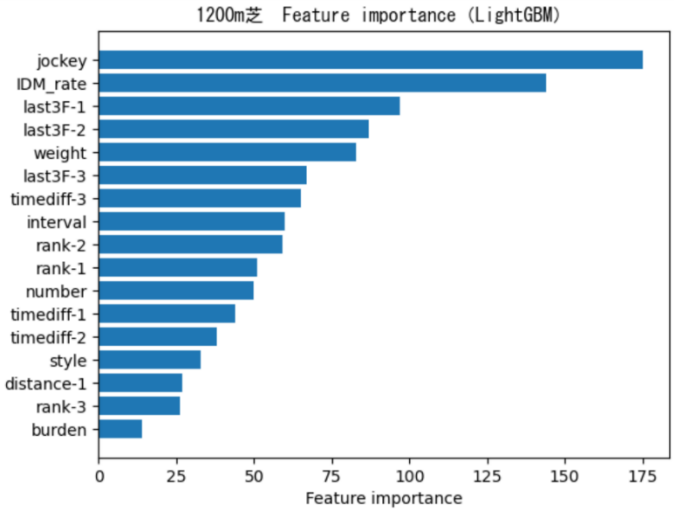
図表 2.3　特徴量の説明と使用理由

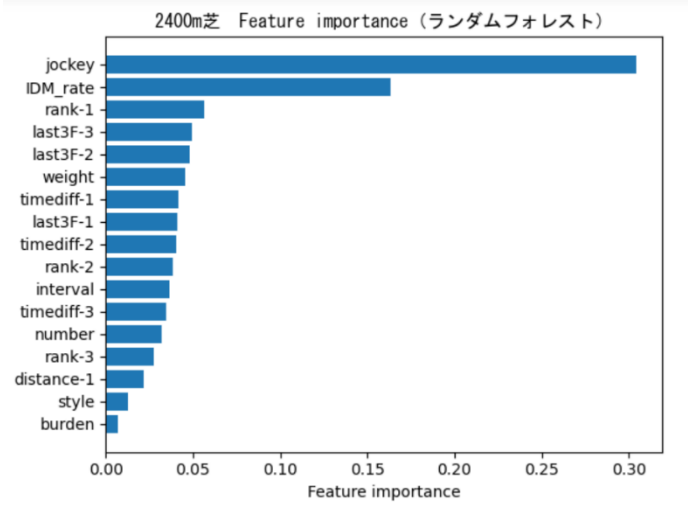
特徴量名	説明
IDM_rate	標準化したIDM 過去の馬のパフォーマンスを示しているため、高いほど3着内に入りやすいと考えられる
jockey	騎手の能力を数値化したもの この指数が高いほど、3着内に入りやすいと考えられる
number	馬番（スタート時の場所であり、内側から1、2、3・・)と数える 距離によっては内枠有利などの傾向が出る可能性がある
weight	馬体重 短距離は馬体重の重い馬が有利、長距離は軽い馬が有利という傾向が考えられる
interval	前走から何週間空いているか 前走から空きすぎているとレース感覚が鈍っている可能性がある
style	脚質、1:逃げ 2:先行 3:差し 4:追込 距離によって脚質の有利・不利などの傾向が出る可能性がある
burden	斤量　競走馬がレースに出走する時に背負う負担重量 斤量が軽いほど有利だと考えられる
distance-1	前走距離 前走の距離から短くなったか、長くなったかが結果に影響を及ぼす可能性がある
rank-1	前走順位 前走以前の順位が良いほど馬に力があると考えられる
last3F-1	前走上がり3Fタイム ラスト3F(600m)の脚のキレはレース結果に直結するため重要である このタイムが良いほど3着内に入りやすいと考えられる
timediff-1	前走の1着馬とのタイム差（1着の場合は2着とのタイム差） タイム差は馬の能力を示していると考えられる
rank-2	2走前順位
last3F-2	2走前上がり3Fタイム
timediff-2	2走前の1着馬とのタイム差（1着の場合は2着とのタイム差）
rank-3	3走前順位
last3F-3	3走前上がり3Fタイム
timediff-3	3走前の1着馬とのタイム差（1着の場合は2着とのタイム差）

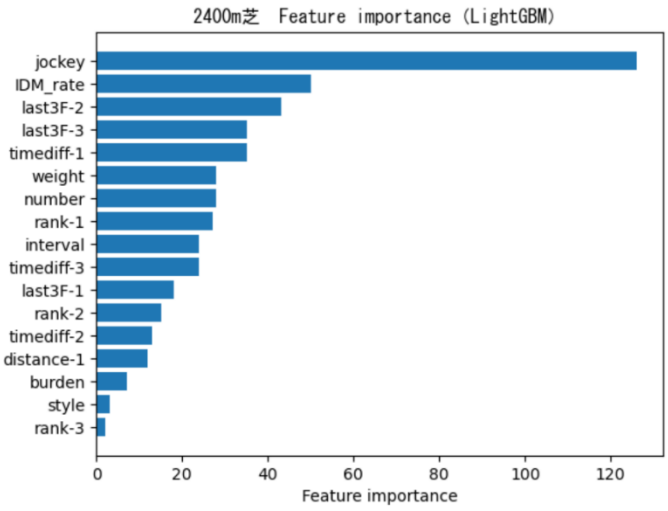
各分類のモデルで特徴量がどれだけ寄与しているかランダムフォレストと LightGBM のそれぞれについて次ページの図表 2.4 で示した。代表例として 1200m 芝・2400m 芝・1200m ダートについてのみ選んだ。

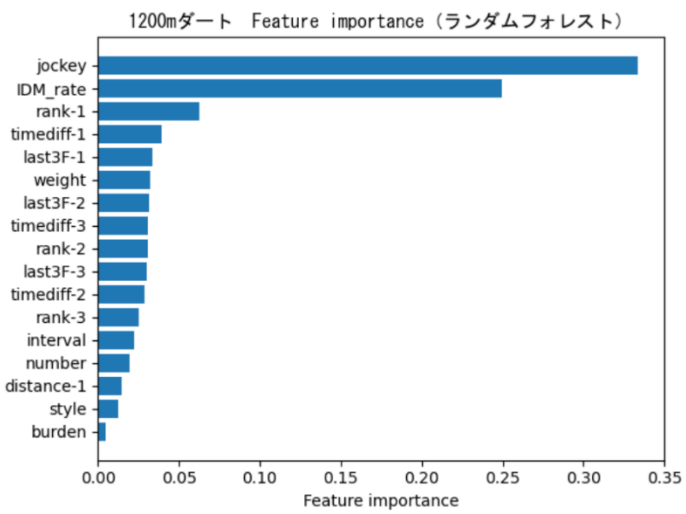
図表 2.4 特徴量の重要性

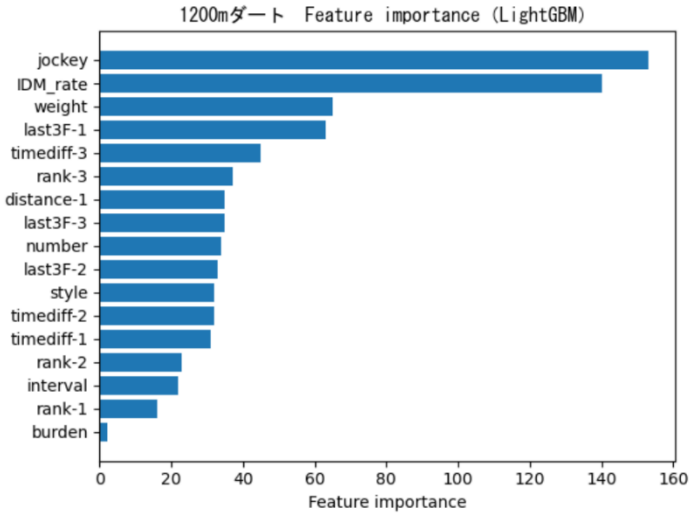












どのモデルでも jockey 指数と IDM_rate が大きく影響していた。騎手の能力と過去の馬のパフォーマンスが影響していると考えれば自然なことだといえる。個々のモデルについて見ていくと、全体的に LightGBM の方が特徴量を平均的に使っている。ランダムフォレストでは jockey 指数と IDM_rate の 2 変数に偏っている。実際にどの特徴量が競馬の結果に影響を及ぼしているか分からないため、どちらのモデルが良いとは言えず、次節以降で回収率を見ていく。また、分類による違いをみていくと分類によって特徴量の影響が異なっており、分類が意味のあるものになっているように思える。

2.4 二値分類の結果

ランダムフォレストによる予測モデルと LightGBM による予測モデルを 2019 年の予想に用いた結果は以下の通りである。条件についてはレース数が多く、G1 も行われている 1200m 芝・1600m 芝・2000m 芝・2400m 芝・1200m ダート・1600m ダートのみ予測した。距離、芝かダートかで分類を行ったが、マイナーな条件であるとレース数が少なく、精度の高い予測モデルが作れていないと考えられるからである。目的変数を 3 着以内に入目的変数として 3 着以内を 1, それ以外を 0 としているため、ランダムフォレストの場合でも LightGBM の場合でも、3 着以内に入る確率が結果として返ってくる。そのため何%以上の場合、購入するかが問題となる。50%以上、60%以上、70%以上の場合の購入、それぞれについて下記の表を作成した[6]。該当馬についてすべて同金額で購入している。3 着以内に入ることを目的変数としているため、理論的には複勝で買うべきであるが参考として単勝で買った場合の結果も出した。

図表 2.5 二値分類による的中率と回収率

		購入数	的中率	回収率
単勝	ランダムフォレスト(50%以上)	929	27.9%	92.7%
	LightGBM(50%以上)	970	27.3%	91.3%
	ランダムフォレスト(60%以上)	383	36.3%	101.2%
	LightGBM(60%以上)	429	33.6%	90.4%
	ランダムフォレスト(70%以上)	77	50.7%	127.3%
	LightGBM(70%以上)	83	50.6%	116.2%
複勝	ランダムフォレスト(50%以上)	929	61.3%	90.5%
	LightGBM(50%以上)	970	61.5%	91.0%
	ランダムフォレスト(60%以上)	383	68.2%	91.9%
	LightGBM(60%以上)	429	67.6%	91.1%
	ランダムフォレスト(70%以上)	77	76.6%	96.0%
	LightGBM(70%以上)	83	73.5%	89.7%

[6] 的中率と回収率について各分類の購入数で加重平均を取っている

表を見て分かるように単勝では 100%を超える場合があったが、複勝では 100%を超えることが出来なかった。単勝の場合、1 着を当てるため複勝よりオッズが高く、的中率が低いという傾向があり、オッズの高い馬一頭を当てるのみで回収率が跳ね上がる。そのため、複勝より回収率のぶれが大きいと考えられ、今回のデータでは上手くいっただけである可能性もある。今回の分析では 3 着以内に入ることを目的変数としているため、複勝の回収率の方が重要であるが、回収率で 100%を超えておらず何か工夫が必要であるといえる。3 着以内に入る確率が高いものに絞ることで的中率は上がるものの、このような買い方をするとオッズの低い人気馬を多く買うこととなり、回収率にはそこまで良い影響を及ぼしていない。

ランダムフォレストと LightGBM について比較すると全体的にランダムフォレストの回収率の方が高くなっている。このことから様々な変数を使うより、jockey 指数と IDM の影響を大きくしたモデルの方が精度は高いといえる。しかし、このデータでランダムフォレストが上手く行っただけの可能性もあり、更なる検討が必要である。

2.5 多分類の結果

前節では目的変数として 3 着以内を 1, それ以外を 0 とし、回収率を確認した。特徴量は十分に検討されており、このモデルについて改善できる点は目的変数であると考えられる。本節では目的変数として 3 着以内を 1, 4〜8 着を 2, それ以外を 3 とする多分類のモデルを考えた。精度が向上することを期待し、分類を細かくした。

二値分類と多分類の比較をすることが目的であり、アルゴリズムとして多分類の実装がしやすい LightGBM のみで行った。前節と買い方は同じであり、結果として出てくる 3 着以内の確率が 50%以上, 60%以上, 70%以上の各場合で検討を行った。比較結果が下の図表 2.6 である。図表から分かるように回収率が向上している部分もあるものの、悪化している部分もあり、全体としてはあまり効果がなかったといえる。

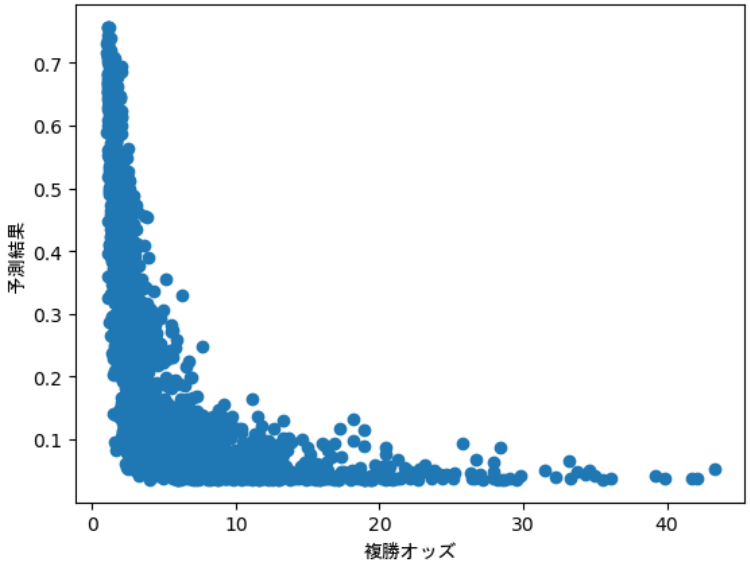
図表 2.6　二値分類と多分類の比較

		購入数	的中率	回収率
単勝	二値分類(50%以上)	970	27.3%	91.3%
	多分類(50%以上)	964	27.7%	94.9%
	二値分類(60%以上)	429	33.6%	90.4%
	多分類(60%以上)	427	34.4%	96.8%
	二値分類(70%以上)	83	50.6%	116.2%
	多分類(70%以上)	86	46.5%	107.7%
複勝	二値分類(50%以上)	970	61.5%	91.0%
	多分類(50%以上)	964	61.2%	90.9%
	二値分類(60%以上)	429	67.6%	91.1%
	多分類(60%以上)	427	67.0%	90.9%
	二値分類(70%以上)	83	73.5%	89.7%
	多分類(70%以上)	86	71.0%	86.5%

2.6 買い方の工夫

これまでの節において特徴量や目的変数に工夫を加えたが、回収率で 100%以上を出すことが出来なかった。これ以上モデルに工夫を加えることは難しく、買い方に工夫を加えることで 100%以上の回収率を目標とする。これまでは結果として出てくる 3 着以内に入る確率を元に購入を決めていたが、下の図表 2.7 でも分かるように予測結果が高いものはオッズが低いのである。そのため、的中率は上がっても回収率が上がらず、100%以上の回収率を出すことが出来なかった。本節では予測結果とオッズを比較し、オッズの要素を買い方に組み込むことで回収率の向上を目指す。

図表 2.7 1600m 芝の予測結果（3 着以内に入る確率）と実際の複勝オッズ



予測結果は 3 着以内に入る確率であるため、$\frac{1}{\text{予測結果}}$ = 理論複勝オッズ となる。例えば、予測結果が 0.25 であれば理論複勝オッズは 4 倍である。これと実際の複勝オッズを比較し、割安な場合、つまり理論複勝オッズ < 実際の複勝オッズ の場合に購入すれば、理論的には回収率で 100%を超えるといえる。このような買い方による結果が下の図表 2.8 である。

図表 2.8 オッズとの比較による買い方

		購入数	的中率	回収率
単勝	ランダムフォレスト(二値分類)	10845	7.3%	66.9%
	LightGBM(二値分類)	10946	7.3%	66.6%
	LightGBM(多分類)	10950	7.3%	67.0%
複勝	ランダムフォレスト(二値分類)	1655	24.0%	80.7%
	LightGBM(二値分類)	1559	24.4%	82.2%
	LightGBM(多分類)	1644	24.0%	78.6%

図表 2.8 から回収率が悪化していることが分かる。購入数を見ると多くの馬を買っており、予測結果は低いもののオッズが高いことで理論的には割安と判断された馬を多く買っていると考えられる。回収率が低くなっているのは、予測結果が低いものについては結果が少し変わるだけで理論複勝オッズが変化することが原因だと考えた。例えば、予測結果が 0.50 と 0.51 では理論複勝オッズは大きく変わらないが、0.01 と 0.02 では理論複勝オッズが 100 倍と 50 倍で大きく変化する。予測結果が低いものについてこのような細かい変化まで予測が正確であるとは考えられず、実際は割安ではない馬券を購入してしまっていると考えられる。

このように予測結果が低いものについて割安な馬券を判断できていないため、回収率が悪化している。そのため、予測結果が高いものについてのみこのような購入方法を適用することで回収率の向上が期待できる。予測結果が 0.5 以上でかつ理論複勝オッズ < 実際の複勝オッズとなる馬を購入すると結果は下の図表 2.9 のようになった。

図表 2.9 予測結果が 0.5 以上でかつ割安な馬の購入結果

		購入数	的中率	回収率
単勝	ランダムフォレスト	890	23.8%	92.6%
	LightGBM(二値分類)	560	21.4%	94.6%
	LightGBM(多分類)	533	19.8%	98.6%
複勝	ランダムフォレスト	253	50.5%	109.4%
	LightGBM(二値分類)	257	52.1%	106.5%
	LightGBM(多分類)	306	52.1%	104.0%

結果の通り、複勝で 100%をこえることが出来た。的中率は前節までの結果より悪化しているものの、オッズとの比較により割安な馬券を購入することにより回収率が向上した。予測結果が 0.5 以上でかつ割安と条件を絞っているため、購入数が少なくなっており、複勝では一週間で 5 頭程度しか購入しないこととなっている。そのため、このデータだけで偶然上手くいったのかを検討する必要があり、実際にこの手法で購入することで収益が出るのか実験していきたい。

3. オッズの歪みに着目した予想

3.1 対象としたデータ

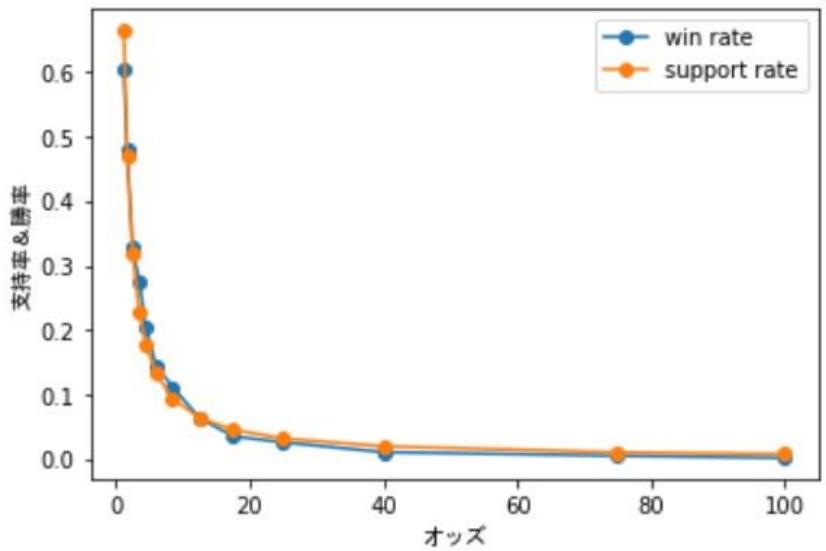
今回の分析対象は、2019 年に JRA が主催した GI 全２４レース(障害レースは除く）とし、24 レースの最終的な回収率が 100%を超えるような馬券の買い方を検討する。具体的には、それぞれのレースの単勝と馬単のオッズを用いる。三連単の場合、配当が高いため、穴馬バイアスも大きくなるが、その分当たる確率も低く、ある程度的中率を担保するためには、1 レースあたりに膨大な数の馬券数を購入しなくてはならないからだ。そこで、最終的に購入する馬券の種類を馬単とすることで、購入する必要がある馬券数を現実的な数に減らしても、ある程度的中できるようにした。

3.2 単勝の支持率と勝率の関係

本章では、先行研究と同様に、単勝の支持率が単勝の客観確率を表すと仮定し、馬単の客観確率を推計する。そこで、単勝の支持率がどの程度客観確率を反映できているのか、実際のデータを用いて検証した。

図は単勝オッズと単勝の支持率・勝率の関係を示している。青色のグラフが勝率、橙色のグラフが支持率である。データは2018年にJRAで施行された全レースの結果を用いた。見てわかる通り、両者のグラフはほぼ一致しており、単勝の支持率はかなりの精度で勝率を表していると言える。こうした結果からも、単勝の支持率を客観確率の推計に用いるのは妥当であると言えるだろう。

図表 3.1 単勝オッズと単勝の支持率・勝率の関係



3.3 予想方法

まず、単勝オッズから単勝馬券の支持率を割り出す。競馬では、全体の購入金額の 20%が JRA 側の収益となり、残り 80%が当選者で山分けされるので、0.8 をその馬券の支持率で割ったものがオッズとなる。したがって、支持率は0.8÷オッズによって求まる。次に、先行研究と同様に、この支持率を用いて馬単の客観確率を推計する。この際、Harville(1973)で提案されている推計方法を用いた。馬 A、馬 B が順番に 1 着、2 着となる確率をP_{AB}、それぞれが 1 着でゴールする確率をP_A,P_Bとして、

$$P_{AB} = \frac{P_A P_B}{1 - P_A}$$

とすることで、推計する。これは、馬 A が 1 着で馬 B が 2 着となる条件付き確率は、馬 A がいないレースでの馬 B の勝利確率であるという考え方に基づいている。

最後に、実際の馬単の支持率である主観確率をオッズから算出する。ただし、馬単の控除率は 25%であるため、0.75 をオッズで割る必要がある。この主観確率が客観確率よりも小さいもの、つまり割安な馬券のみを購入すれば良い。しかし、オッズが上がれば上がるほど、穴馬バイアスは大きくなるので、馬券の購入点数は減っていくことになる。そのため、この買い方でオッズの高い、低確率な馬券が当たることは少ない。つまり、割安な馬券を全て購入しても、いたずらに購入金額を増やし、回収率を下げてしまうのである。そこで、主観確率が客観確率よりも小さい、かつ客観確率が$\frac{1}{24} = 0.041666...$を超えるような馬券、つまり 24 回中 1 回以上は的中が見込める馬券のみを購入することとした。

3.4 実証

実際にこの理論を用いて、2019 年の GI 全 24 レースを購入した際のシミュレーションを行った。表では、購入する馬券につき 100 円ずつ賭けていった際の購入金額と払戻額を示している。

レース名	購入金額	払戻額
フェブラリーS	400	750
高松宮記念	400	0
大阪杯	200	0
桜花賞	200	0
皐月賞	600	1140
天皇賞（春）	200	0

NHK マイル C	400	0
ヴィクトリアマイル	0	0
オークス	200	0
日本ダービー	400	0
安田記念	300	0
宝塚記念	100	0
スプリンターズ S	300	2040
秋華賞	0	0
菊花賞	500	0
天皇賞（秋）	400	1170
エリザベス女王杯	400	0
マイル CS	500	2040
ジャパン C	200	0
チャンピオンズ C	400	1820
阪神 JF	300	0
朝日杯フューチュリティ S	200	950
ホープフル S	500	1170
有馬記念	500	0
合計	7600	11080

結果として回収率は、$\frac{11080}{7400} \times 100 = 145.78\ldots(\%)$となり、狙いとしていた回収率 100%超えを達成することができた。馬単の控除率は 25%であるため、期待できる回収率が 75%であることを踏まえると、この数値はかなり評価できるものと言える。平均購入点数は 3.166...であり、単勝人気の上位 2、3 頭での購入が多いと考えられる。そのため、的中した際の払戻額の平均は 1385 円となっており、1 回のレースで 50 倍を超えるような馬券の購入はなかった。

4. まとめと更なる課題

機械学習を用いた競馬予想では、予測結果とオッズを考慮することで最終的に複勝の回収率で 100%を超えることが出来た。しかし、このモデルには 2 点問題がある。一つ目は実際に馬券を購入する際には複勝のオッズが決まっていないということだ。レース前に複勝のオッズは決定せず、範囲で示される。複勝の払戻金額は

(当該勝馬に対する勝馬投票券の総券面金額 +

(出走した馬であって勝馬以外のものに対する勝馬投票券の総券面金額 ÷ 3))× 0.8

で計算される。つまり、はずれ馬券の総額を的中馬である 3 頭で山分けする形であり、他の 3 着以内に入った馬が人気馬であるとオッズが低くなる。一方、他の 3 着以内に入った馬が人気のない馬であるとオッズが高くなる。このようにレース前にオッズが確定しないため、予測結果に照らして正確に割安であるかどうかを判断できない。幅のあるオッズに対してどのように購入するか考える必要がある。

二つ目は予測結果が 0.5 以上でかつ割安な馬券を購入する方法であるが、0.5 という数字に理論的な裏付けがないということだ。この数字を上げると的中率が上がる代わりにオッズが低い馬を多く買うようになる。そのような中で回収率を最大にする値が理論的にいくつであるのかを見つける必要がある。また、該当馬についてすべて同金額で購入していたが、各馬の購入金額に差をつけて最適な金額で購入すれば、回収率を向上できると考えられる。

オッズに着目した予想では、かなり良い結果を出すことができた。しかし、今回は偶然うまくいったということもあり得る。24 レース中、上位決着になるようなレースが殆どない場合、この購入方法ではうまくいかないからだ。その場合、購入レース数を増やすか基準となる客観確率を下げる必要がある。特に基準となる客観確率の定め方には改善の余地があると思われる。ある程度の的中率がありながらも、購入点数が多くなりすぎないような基準値をより多くのレースデータを用いて検証することで、より確実性のある理論となるだろう。

参考文献

小幡績，太宰北斗(2014)「競馬とプロスペクト理論：微小確率の過大評価の実証分析」，行動経済学, 第 7 巻

Harville, D. A., 1973. Assigning probabilities to the outcomes of multi-entry competitions. Journal of the American Statistical Association 68, 312–316.

Sebastian Rachka、Vahid Mirjalili 著，福島真太郎訳，『第二版　Python 機械学習プログラミング』，株式会社インプレス，2019 年

NUKUI SHUN，「実践・競馬データサイエンス」, PyCon JP 2018, 2018 年（最終閲覧日 2020 年 11 月 1 日），https://alphaimpact.jp/downloads/pyconjp2018.pdf

JRDB, データルーム, 2020 年(最終閲覧日 2020 年 11 月 14 日), http://www.jrdb.com/

付録

ここでは機械学習を用いた競馬予想の章において、最終的に回収率で 100%を超えたモデルについて分類別の詳細データを示す。具体的には第二章六節の予測結果が 0.5 以上でかつ理論複勝オッズ < 実際の複勝オッズとなる馬を購入した場合のモデルについてである。

図表 1 ランダムフォレスト(二値分類)の詳細データ

		購入数	的中率	回収率
1200m芝	単勝	97	26.80%	96.30%
	複勝	22	54.50%	114.10%
1600m芝	単勝	209	23.90%	73.50%
	複勝	43	48.80%	94.00%
2000m芝	単勝	214	28.50%	100.00%
	複勝	38	57.90%	109.20%
2400m芝	単勝	101	24.80%	91.40%
	複勝	33	70.00%	136.70%
1200mダート	単勝	161	48.10%	101.30%
	複勝	54	27.30%	101.10%
1600mダート	単勝	108	24.10%	99.40%
	複勝	63	55.60%	111.10%

図表 2 LightGBM(二値分類)の詳細データ

		購入数	的中率	回収率
1200m芝	単勝	49	18.4%	96.7%
	複勝	28	53.6%	112.5%
1600m芝	単勝	135	11.9%	53.0%
	複勝	49	57.1%	109.6%
2000m芝	単勝	120	25.0%	126.0%
	複勝	41	48.8%	95.0%
2400m芝	単勝	73	19.2%	83.6%
	複勝	33	60.6%	116.1%
1200mダート	単勝	101	21.8%	98.1%
	複勝	51	45.1%	97.8%
1600mダート	単勝	82	22.0%	121.5%
	複勝	55	50.9%	111.6%

図表 3 LightGBM(多分類)の詳細データ

		購入数	的中率	回収率
1200m芝	単勝	48	18.80%	104.60%
	複勝	28	57.10%	125.00%
1600m芝	単勝	112	11.60%	51.60%
	複勝	42	54.80%	103.30%
2000m芝	単勝	120	24.40%	127.10%
	複勝	51	49.00%	93.90%
2400m芝	単勝	79	17.70%	74.30%
	複勝	46	56.50%	106.30%
1200mダート	単勝	104	25.00%	117.40%
	複勝	52	46.20%	97.10%
1600mダート	単勝	90	20.00%	115.60%
	複勝	67	52.20%	107.30%