

大阪電気通信大学
総合情報学部 デジタルゲーム学科
2021 年度 卒業論文

機械学習を用いた競馬予想

HW18A094 中谷 翼
デジタルアーキテクチャ研究室
指導教員：上善 恒雄

2022 年 01 月 25 日

概要

競馬は馬に騎手が乗り馬たちを競い合わせる競争競技である。そして競馬は競争競技でありながら公営賭博でもあり、競馬ファンは競争の着順を予想しお金をかけることによって楽しんでいる。この着順の予想は人それぞれによってやり方が異なり、競馬新聞を見て予想する人や、オッズを見て予想する人もいる。本論文ではこの競馬予想を機械学習を用いて予想的中率を 100 % に近づけることを目指す。

目次

1	はじめに	1
1.1	研究背景	1
1.2	研究目的	1
1.3	本論文の構成	1
2	研究の背景知識	3
2.1	関連研究	3
2.2	中央競馬について	3
2.3	機械学習について	6
2.4	ロジスティック回帰分析について	6
3	研究内容	8
3.1	研究に使用したもの	8
3.2	環境説明	11
3.3	研究方法	12
3.4	実験手順	12
4	結果と考察	15
4.1	結果	15
4.2	考察	17
5	まとめと今後の課題	21
5.1	まとめ	21
5.2	今後の課題	21
5.3	最後に	21

1 はじめに

1.1 研究背景

競馬は馬に騎手を乗せ馬同士を競い合わせる競争競技である。そして競馬は国や各地方自治体によって経営される公営賭博でもある。競馬ファンは競馬の着順を予想しお金を賭けることにより楽しんでいる。この予想とは勝馬投票券の種類によって異なるのだが、基本1着馬を決める予想のことである。

競馬のレースに出走する各馬たちにはそれぞれ人気が存在する。この人気は各馬の単勝馬券の総売上数によって決められる。例に挙げると、A馬とB馬という馬がいて、A馬の単勝馬券の総売上数が100万円、B馬の単勝馬券の総売上数が50万円だとA馬の方が人気になる。この人気度がファンたちによって強いと思われている馬が人気することが多く、言うなればこの人気はファンたちによって決められた強さ投票のようなものである。実際にこの人気で決まるレースもあり、2020年のジャパンカップ[1]は1着は1番人気の「アーモンドアイ」、2着に2番人気の「コントレイル」、3着に3番人気の「デアリングタクト」で決着した。このように人気な馬だけで勝馬投票券を買えば的中すると思いきやそのようなことではなく、2021年エリザベス女王杯[2]のように1着10番人気の「アカイト」、2着7番人気の「ステラリア」、3着9番人気「クラヴェル」のように人気の馬は全て馬券外になるレースもある。このようなレースもあるので人気通りで買えば良いと言うわけではない。ここで重要となるのが予想である。

予想はファンそれぞれやり方が異なり、競馬新聞や競馬情報サイト「netkeiba.com」[3]に載っている各馬ごとの過去の成績などを見て予想する人もいれば、過去のレース映像や調教映像を見て予想する人もいる。また馬の情報を見ずにオッズ理論と呼ばれる、馬券ごとに決められているオッズを用いて予想する人もいる。今回の研究では競馬の過去のレース結果の情報を使用し、機械学習を用いて競馬予想をしていく。

1.2 研究目的

本研究の目標としては競馬予想を機械学習で行うことにより、的中率を100%に近づけることである。またこの競馬予想とは、中央競馬のレースを対象とし、1着馬、2着馬、3着馬、4着以下の馬を予想することである。

1.3 本論文の構成

本論文の構成は以下の通りである。第2章は競馬予想に関連する研究や、中央競馬の基本知識、機械学習についてなど本論文を進めていくために必要な知識について紹介する。第3章は研究方法について説明している。ここでは使用言語や使用環境について、使用データの入手法やデータの加工について、実験手順が書かれている。第4章は予想結果とその的中率、実験からの考察について

書かれている。第 5 章はまとめと課題についてである。

2 研究の背景知識

2.1 関連研究

以下は競馬のレース予想に関する研究について示す。

2.1.1 機械学習を使用した競馬予想

本研究テーマである機械学習を使った競馬予想は実際に行われており、日本中央競馬会 (JRA) が提供している「JRA-VAN」[4] や競馬情報サイトの netkeiba.com にて提供されている「的中型人工知能」[5] などがある。またイギリスでも研究が進められており、勝者と敗者を区別するデータに依存した分類モデルが提案されており、イギリス競馬のデータを使用し実証実験も行われている [6]。

2.1.2 Harville の公式と Gini 係数を用いた競馬予想

競馬の予想の基本は過去のレース結果や、各馬の特徴や当日の状態などを見て行う。しかしこの研究は本命バイアスが顕著に見られる 1 番人気の馬に注目し、1 番人気の馬が勝つレースと負けるレースを、投票率や人気の偏りの投票率を表す Gini 係数を用いて行われている。本命バイアスとは人気の高い馬が実際の結果に見合うほどの得票を必ずしも集めない傾向のことを言う [7]。

2.1.3 穴馬への過剰な選好に関する研究

世界各国の競馬で「各馬の勝つ確率」に「その馬のオッズ」に応じた値 (= 期待払戻率) を比較すると、本命馬は値は大きくなり、穴馬は値が小さくなる傾向がある。この穴馬の過剰な選好を「Favorite-Longshot-Bias (穴馬バイアス)」と呼ばれており、数多くの研究がされている [8]。本命馬とはレースに出走される馬の中で最も強いと思われる馬のことである。穴馬は能力がわからず人気していない馬のことで、この馬が 3 着以内に入ると高配当が期待できる。

2.2 中央競馬について

2.2.1 概要

中央競馬は農林水産省監督下の元、日本中央競馬会 (JRA) によって開催されている [9]。1 年間で約 3400 回レースが行われており、基本毎週土日日本の 10 箇所の競馬場中の 2 箇所か 3 箇所、各競馬場 1 日 12 レース行われている。

2.2.2 競争について

中央競馬にて行われている競争には「一般競争」と「特別競争」の 2 種類がある。特別競争は、出走するために特別登録が必要なレースで現在中央競馬の特別競争には全てレース名が付けられている。例で言うと「若駒ステークス」や「壇ノ浦ステークス」がある。また特別競争には重賞競走

と呼ばれるものがあり、一般的な特別競争よりも賞金が高く重要な意義を持ったレースだ [10]。重賞レースには位置づけを明確にするために、「グレート制」があり 1984 年に格付けされた。表記としては「G1」「G2」「G3」と表記され G1 が一番格が上である。[11] この G1 は「日本ダービー」や「有馬記念」が該当する。

次に一般競争についてだが、一般競争は特別競争以外の競争のことで、「3 歳未勝利」や「4 歳以上 1 勝クラス」などレース条件がレース名となっている。

2.2.3 勝馬投票券について

競馬はスポーツでもあるが賭博の面もあるのでお金の駆け引きがある。その駆け引きをするためにあるのが「勝馬投票券」またの名を「馬券」だ。表 1 は勝馬投票券の種類についてである。この勝馬投票券は競馬場や JRA が経営する場外馬券所で購入することができる。また 2002 年からはインターネット投票が導入され、電話やパソコンなどで手軽に購入できるようになった。

表 1 勝馬投票券の種類

単勝	1 着馬を当てる馬券
複勝	3 着までに入る馬を当てる馬券
枠連	1 着と 2 着になる馬の枠番号の組み合わせを当てる馬券
馬連	1 着と 2 着になる馬の馬番号の組み合わせを当てる馬券
馬単	1 着と 2 着になる馬の馬番号を着順通り当てる馬券
ワイド	3 着までに入る 2 頭の組み合わせを当てる馬券
3 連複	1 着、2 着、3 着となる馬の組み合わせを馬番号で当てる馬券
3 連単	1 着、2 着、3 着となる馬の馬番号を着順通りで当てる馬券
WIN5	JRA が指定する 5 レース全ての 1 着馬を当てる馬券

勝馬投票券の販売開始時間は、土曜日のレースは前日の 18 時 30 分から各レース発走 1 分前、日曜日のレースは前日の 19 時 30 分から各レース発走 1 分前と決まっている。しかし G1 レースや日曜日にある重賞レースは前日販売があり、G1 レースだと金曜日の 18 時 30 分から販売されており、日曜日の重賞レースは土曜日の 7 時 00 分から販売が開始されている。

表 2 当日販売の販売時間

対象レース	販売開始時間	販売締め切り時間
土曜日全レース	金曜日の 18 時 30 分	各レースの発走 1 分前
日曜日全レース	土曜日の 19 時 30 分	各レースの発走 1 分前

表 3 前日販売の販売時間

対象レース	販売開始時間	販売締め切り時間
G1 レース・目黒記念 (G2)	金曜日の 18 時 30 分	土曜日の 17 時 30 分
日曜日の重賞レース	土曜日の 7 時 00 分	土曜日の 17 時 30 分

2.2.4 払戻について

日本競馬の払戻には「パリミュチュエル方式」[12] が採用されている。このパリミュチュエル方式は、売り上げから一定割合を差し引いて、残りの金額を当選者に配当する方式である。売り上げから差し引く方式なので、勝馬投票券購入時はまだ払戻配当は決定していないのがパリミュチュエル方式の特徴である。パリミュチュエル方式の一定割合は、競馬でいうと各勝馬投票券ごとの払戻率のことである。各勝馬投票券の払戻率は表 4 の通りである。

表 4 各勝馬投票券の払戻率

投票法	単勝	複勝	枠連	馬連	ワイド	馬単	3 連腹	3 連単	WIN5
払戻率	80 %	80 %	77.5 %	77.5 %	77.5 %	75 %	75 %	72.5 %	70 %

そしてパリミュチュエル方式が採用された払戻対象総額の計算式 [13] は以下の式 1 の通りである。

$$(W + D/P) \times R \quad (1)$$

また WIN5 の場合、的中者がいないとキャリーオーバーになり次回の WIN5 の払戻対象総額の計算式が以下の式 2 のようになる。

$$(W + D/P) \times R + A/P \quad (2)$$

これらの式の「W」は勝馬に対する勝馬投票券の総券面金額を示す。「D」は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を示し、「P」は勝馬の数、「R」は各勝馬投票券の払戻率のことを示している。式 2 の「A」は WIN5 のキャリーオーバーの金額を示している。またこの払戻対象金額からその勝馬投票券の販売金額を割ることにより、その勝馬投票券のオッズを計算している。計算式は式 5 の通りである。

$$\text{払戻対象総額} / \text{その勝馬投票券の販売金額} \quad (3)$$

オッズは各勝馬投票券ごとにある払戻時のおおよその倍率のことで、もしオッズが 5 倍でその勝馬投票券が的中すると、100 円で買っていた場合 500 円が払戻される。このオッズは各勝馬投票の組み合わせごとに存在しており、もし 10 頭のレースだと、単勝は 10 通りあるので 10 通りごとにオッズがあり、馬連だと 45 通りあるので 45 通りごとにオッズがある。

2.3 機械学習について

機械学習 (MachineLearning) は 1960 年頃から人工知能における課題研究として研究されており、データから規則性や判断基準を機械に学習させ、それに基づき未知のものを予想する技術である。機械学習は主に 3 種類に分類することができる。以下の表 5 は機械学習の分類についてのものである [14]。表の「特徴量」とは対象の特徴が数値として表されたものである。教師データとは機械学習で学習するデータのうち例題と答えに関するデータのことである。また教師データは訓練データと呼ばれることもある。

表 5 機械学習の分類について

	入力に関するデータ (特徴量)	出力に関するデータ (教師データ) [正しい答え]		活用法
機械あり学習	有り	○	有り	株価予測
教師なし学習	有り	×	無し	自動運転 AI
強化学習	有り (試行)	△ (間接的)	答えはないが、評価は有り	将棋 AI

教師あり学習とは、既知となった過去の入力に関するデータと出力に関するデータを、機械学習アルゴリズムにあらかじめ与えることで、それらを正解データとして計算する手法である。また教師あり学習には「分類モデル」と「回帰モデル」がある。分類モデルは過去のデータから抽出された工学的特徴の特定のセットに基づいて、項目をクラス分けするものである [15]。回帰モデルはある 2 つの変数の関係を表す式のうち、統計的手法で推測された式により、一方の大きさが他方の大きさをどの程度説明できるかを分析するものである [16]。本研究では分類モデルの「ロジスティック回帰分析」を用いて競馬予想していく。

教師なし学習は教師あり学習とは異なり、特徴量のみのデータを用いて学習する手法である。データの特徴を学習しデータのグループ分けを得意とする。しかし教師なし学習は正解のデータがないことから、教師あり学習よりかは精度は低くなる傾向があり本研究には不向きだと判断した。

強化学習は、教師あり学習や教師なし学習に使われる静的データには依存せず、動的な環境で動作し収集した経験から学習する手法である。強化学習ははじめに評価を設定し、あることをすることにより評価を与えられ、それを繰り返すことにより評価を与えられる行動を学習する。この強化学習も動的な環境で動作する点で、競馬の過去のレース結果のデータから学習させたい本研究とは不向きと考え、使用しない。

2.4 ロジスティック回帰分析について

ロジスティック回帰分析はいくつかの要因 (説明関数) から 2 値の結果 (目的変数) が起こる確率を説明予測する統計手法で一般化線形モデルの一種である。また 3 値以上の目的変数で予測する手法を、多項ロジスティック回帰分析という。説明関数とはある現象や値を説明する変数のことで

ある。目的変数は説明変数を受けて発生した結果を表す変数である。一般化線形モデルは、正規分布を仮定とした解析をする一般線形モデルとは違い、正規分布以外の確率分布に従う場合にも使えるようにした線形モデルである。また一般化線形モデルは以下の式 4 で表することができる [17]。

$$g(y) = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_n x_n \quad (4)$$

ただし β は偏回帰係数、 x は説明関数、 n は説明変数の数を示す。そしてここでの関数 $g()$ はリンク関数を示す。リンク関数はモデルの推定を行えるよう累積確率を変換するためのものである [18]。累積確率はある値を x としたとき、 x 以下の確率のことをいう。またリンク関数は使用モデルによって異なり、ロジスティック回帰分析の場合はロジットを使用する。ロジットはロジット関数とも呼ばれ、対数オッズを関数にしたものである。ここで出てきたオッズとは競馬に出てくるオッズではなく、統計学用語のオッズである。このオッズはある事象が起こる確率を p 、事象が起こらない確率を $1-p$ とするとき、この二つの確率の比をオッズと呼ぶ。オッズの式は以下の通りである。

$$\frac{p}{1-p} \quad (5)$$

そしてオッズの対数を取ったものを対数オッズと呼ぶ。式 6 が対数オッズである [19]。

$$\log \frac{p}{1-p} = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_n x_n \quad (6)$$

この式 6 の左辺を関数化した式 7 がロジット関数である。

$$f(p) = \log \frac{p}{1-p} \quad (7)$$

次にロジット関数を y としたときその逆関数にしたのが式 8 である [20]。

$$p = \frac{1}{1 + e^{-y}} \quad (8)$$

式 8 はロジスティック関数と呼ばれ、ロジスティック回帰分析はこの関数を用いて確率を求めたり、式を変形させ説明変数の偏回帰変数を求めることもできる。偏回帰係数は 1 つの説明変数に注目しその他の説明変数を 0 としたとき、注目した説明変数が 1 単位変化した時の目的変数の変化量を表したものである。この係数が高いほどその説明変数はプラスに捉えるに重要になり、係数が低いほどその説明変数はマイナスと捉えるにために重要になる。

3 研究内容

以下は本研究にて使用したもの、研究方法・実験手順について述べる。

3.1 研究に使用したもの

本研究では機械学習を行うために MacBookPro を用いて行った。この MacBookPro の macOS のバージョンは 10.15.7 を使用した。

また以下は使用した言語やデータについて説明する。

3.1.1 使用言語について

本研究では機械学習を用いて競馬予想をしていくため、プログラミングを用いて機械学習を行う必要がある。そのために使用したプログラミング言語は、機械学習のライブラリが豊富な「Python」である。Python には機械学習を行う際に頻繁に使われている sklearn や、スクレイピングしたデータを機械学習するためにデータを加工できる pandas がある。これらのライブラリを使用することにより Python で機械学習を行うことができる。

3.1.2 JupyterLab について

Python を実行するために導入したものは、「JupyterLab」と呼ばれるブラウザ上で動作する対話型実行環境である。JupyterLab は Python パッケージ管理ツールの 'pip' を用いてインストールができる。JupyterLab はデータ分析の結果をプログラミング時に確認できるのと、ローカルサーバーで実行完結するので実行速度やセキュリティ面で安心ができるのがポイントである [21]。

3.1.3 使用データについて

本研究で使用するデータは、2019 年、2020 年、2021 年の中央競馬で行われたレース結果のデータを使用し、この 3 年のデータを結合し 3 年間分の 1 つのデータにする。スクレイピングするページは netkeba.com のデータベース [22] からである。以下の図 1 がスクレイピングし結合したデータである。

	着順	枠番	馬番	馬名	性別	斤量	騎手	タイム	着差	単勝	人気	馬体重	調教師
	0	1	1	1	ゴルコンダ	牡2	54.0	ルメール	1:48.3	NaN	1.4	1.0	518(-16) [東] 木村哲也
	1	2	3	3	ブントファイヤー	牡2	54.0	岩田康誠	1:50.1	大	3.5	2.0	496(-8) [東] 手塚貴久
	2	3	4	4	ラグリマスネグラス	牡2	51.0	団野大成	1:50.9	5	46.6	6.0	546(+6) [東] 藤沢和雄
	3	4	8	9	キタノコドウ	牡2	51.0	菅原明良	1:51.5	3.1/2	56.8	7.0	458(-8) [東] 高木登
	4	5	5	5	ネモフィラブルー	牡2	54.0	川島信二	1:51.7	1.1/2	140.3	9.0	436(0) [西] 矢作芳人
	...	...	...	...	...	...	...	...	...	...	...	...	...
139852	7	2	2	2	ジオルティ	牡3	51.0	角田大和	2:42.0	1.1/4	49.6	10.0	502(+8) [西] 杉山晴紀
139853	8	7	9	9	ロックグラス	牡4	57.0	幸英明	2:42.2	1.1/4	43.7	8.0	464(0) [西] 谷潔
139854	9	7	8	8	サマーカナロア	牡3	53.0	富田暁	2:42.2	ハナ	46.5	9.0	502(-10) [西] 見真
139855	10	8	10	10	レッドレイル	セ5	57.0	浜中俊	2:42.4	1.1/4	14.0	6.0	436(+2) [西] 辻野泰之
139856	11	6	7	7	ニシノトランザム	セ6	57.0	西村淳也	2:42.6	1.1/2	82.5	11.0	486(-2) [西] 森秀行

139857 rows × 13 columns

図 1 スクレイピングして結合したデータ

しかしこのままだと欠損値や今回使用しないデータも含まれるため、これらのデータを加工する必要がある。またタイムのデータも使用するため、分と秒で表記されているのを秒で表記するようにする。その他にも馬体重のデータは当日の馬体重と前走との馬体重の変化量の2つのデータが1つとなっているので、体重と前走との体重変化量に分けたい。これらの行為を「データの前処理」と呼ばれている。そして以下の図2がスクレイピングしてきたデータを前処理したデータである。

	着順	枠番	馬番	馬名	斤量	騎手	タイム	単勝	人気	性	年齢	体重	体重変化	
	0	1	1	1	ゴルコンダ	54.0	ルメール	108.3	1.4	1.0	牡	2	518	-16
	1	2	3	3	ブントファイヤー	54.0	岩田康誠	110.1	3.5	2.0	牡	2	496	-8
	2	3	4	4	ラグリマスネグラス	51.0	団野大成	110.9	46.6	6.0	牡	2	546	6
	3	4	8	9	キタノコドウ	51.0	菅原明良	111.5	56.8	7.0	牡	2	458	-8
	4	5	5	5	ネモフィラブルー	54.0	川島信二	111.7	140.3	9.0	牡	2	436	0
	...	...	...	...	...	...	...	...	...	...	...	...	...	...
139852	7	2	2	2	ジオルティ	51.0	角田大和	162.0	49.6	10.0	牡	3	502	8
139853	8	7	9	9	ロックグラス	57.0	幸英明	162.2	43.7	8.0	牡	4	464	0
139854	9	7	8	8	サマーカナロア	53.0	富田暁	162.2	46.5	9.0	牡	3	502	-10
139855	10	8	10	10	レッドレイル	57.0	浜中俊	162.4	14.0	6.0	セ	5	436	2
139856	11	6	7	7	ニシノトランザム	57.0	西村淳也	162.6	82.5	11.0	セ	6	486	-2

138668 rows × 13 columns

図 2 前処理したデータ

また本研究で使用する目的変数と説明変数は以下の表6の通りである。

表 6 本研究で使用した目的変数と説明変数

目的変数	
着順	レースの順位。1 着だと 1、2 着だと 2 と順位の数字が入っている。
説明変数	
枠番	スタートする際に入るゲートの番号。枠番は 1 から 8 までである。
馬番	各レースの各馬ごとに付けられる馬の番号。馬番は 1 から 18 までである。
馬名	走る馬の名前。強い馬弱い馬がいるため重要なデータと考えている。
斤量	レースの際に馬が背負う重さ。基本的に牡馬 57kg、牝馬 55kg と決まっている。
騎手	レースで馬に乗る人。騎手は免許試験に合格しレースで乗ることができる。
タイム	レースでスタートからゴールまでに掛かった時間。
単勝	単勝オッズの数値。数値が低いと人気しており、高いと人気がない。
人気	単勝オッズの数値を元にした人気。
性	馬の性別。牡馬、牝馬、セン馬がいる。
年齢	馬の年齢。
体重	レースに出る前に馬体重が図られその時の体重。
体重変化	前走との馬体重の差。

だがこのままでは馬名、騎手、性のデータは使えない。そのために使用したのが「ダミー変数」である。ダミー変数は数値以外のデータを数値に変換する変数のことで、0 か 1 のどちらかで表される。ダミー変数化することにより性のデータなどを扱えるようになる。また今回目的変数の着順だが 4 着以下は全て 4 とし、勝馬投票券に関係する 3 着までをそのまま使用する。そして図 2 のデータをダミー変数化し、着順を加工したデータが図 3 である。このデータを使用し機械学習での予想を行っていく。

	枠番	馬番	斤量	タイム	単勝	人気	年齢	体重	体重変化	rank	...	騎手_鰐島克駿	騎手_鰐島良太	騎手_鰐島祥行	騎手_黒岩悠	騎手_黒岩弘人	騎手_C.デム	騎手_M.デム	性_セ	性_牝	性_牡
0	1	1	54.0	108.3	1.4	1.0	2	518	-16	1	...	0	0	0	0	0	0	0	0	0	1
1	3	3	54.0	110.1	3.5	2.0	2	496	-8	2	...	0	0	0	0	0	0	0	0	0	1
2	4	4	51.0	110.9	46.6	6.0	2	546	6	3	...	0	0	0	0	0	0	0	0	0	1
3	8	9	51.0	111.5	56.8	7.0	2	458	-8	4	...	0	0	0	0	0	0	0	0	0	1
4	5	5	54.0	111.7	140.3	9.0	2	436	0	4	...	0	0	0	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
139852	2	2	51.0	162.0	49.6	10.0	3	502	8	4	...	0	0	0	0	0	0	0	0	0	1
139853	7	9	57.0	162.2	43.7	8.0	4	464	0	4	...	0	0	0	0	0	0	0	0	0	1
139854	7	8	53.0	162.2	46.5	9.0	3	502	-10	4	...	0	0	0	0	0	0	0	0	0	1
139855	8	10	57.0	162.4	14.0	6.0	5	436	2	4	...	0	0	0	0	0	0	0	1	0	0
139856	6	7	57.0	162.6	82.5	11.0	6	486	-2	4	...	0	0	0	0	0	0	0	1	0	0

138668 rows × 21059 columns

図3 実際に学習させるデータ

3.2 環境説明

本研究を行うため Python の実行環境として「JupyterLab」を使用した。また JupyterLab をインストールする際に「Anaconda」のインストールも行なった。Anaconda は Python で使用できるデータサイエンスに必要なライブラリを提供する、ディストリビューションである。この Anaconda をインストールしないと JupyterLab は正しく動作しなかったのでインストールした。また本研究で使った Python のライブラリとモジュールは以下の図 4 のものである。

```
import pandas as pd
import time
from tqdm.notebook import tqdm
from sklearn.model_selection import train_test_split
from imblearn.under_sampling import RandomUnderSampler
from sklearn.linear_model import LogisticRegression
```

図4 研究のためにインポートしたライブラリとモジュール

ここからは本研究でインポートしたものに関して説明していく。はじめに「pandas」についてである。pandas はデータ解析を容易にするために提供されている python ライブラリである。この pandas の特徴としては独自のデータ型の「DataFrame 型」があることである。この DataFrame 型は二次元の表形式のデータで、本研究の競馬の過去レースのデータはこの DataFrame 型を使用した。また pandas はデータの前処理も得意としている。ロジスティック回帰分析をするためには、スクレイピングしてきたデータそのまま使うことはできず、必ずデータの前処理が必要である。それを容易にできるのが pandas である。

次に「time」である。time は経過時間や日時の差分を測定するモジュールである。本研究では for 文を使用して Web スクレイピングを行なったのだが、そのまま行ってしまうとスクレイピングする Web サイトのサーバーに負荷をかけてしまう。そこで負荷をかけないように一時的に処理を停

止するために `time` を使う。`time.sleep(1)` とすることにより、1 秒間処理を一時停止することができる。これによって一度処理を終えると 1 秒間の停止が入りサーバー負荷を防ぐことができる。

次は「`tqdm`」についてである。`tqdm` はプログラムの進捗状況をプログレスバーとして表示するライブラリである。今回の場合だと 1 年分の過去のレース結果をスクレイピングするとき 2 時間半ほどかかった。そのときスクレイピングの進捗状況を確認するために `tqdm` を用いて進捗状況の可視化するためのプログレスバーを表示させた。

次に「`sklearn`」である。これは機械学習を行う際に使用するライブラリで、主に回帰・分類・クラスタリング・次元削減を行える。本研究で使用するロジスティック回帰分析は分類にあたるので `sklearn` で実装することができる。また `sklearn` ではデータを二分割にすることができ、それを使用することで訓練データとテストデータを作っている。

最後に「`imblearn`」についてである。`imblearn` は先ほど紹介した `sklearn` とともに使用できる。`imblearn` の特徴はアンダーサンプリングを行える点である。アンダーサンプリングは多数派データを少数派データの数に合わせて削除することである。これをしないと機械が多数派のデータばかりを予想してしまい、的中率は高いがモデルとしては悪くなる。なので数を合わせることでよりバランスの良いデータになり、学習結果としても偏りのない学習ができる。

3.3 研究方法

本研究は、中央競馬の 2019 年、2020 年、2021 年の 3 年間のレース結果を対象として競馬予想を行う。競馬予想と言っても勝馬を予想するものや、入着タイムを予想するなどあるが、本研究では 1 着になる馬、2 着になる馬、3 着になる馬、4 着以降になる馬を予想する。それを行うために使用する予想方法として、ロジスティック回帰分析を用いた。

研究方法は、使用する 3 年間分のレース結果データを学習できるよう前処理し、訓練データとテストデータでランダムに分割する。訓練データとは機械に学習させるデータのことでモデルを形成する重要なデータである。テストデータとは訓練データを学習し作成したモデルの検証をするためのデータである。データを分割すると訓練データを学習させ、ロジスティック回帰分析を元に予想していく。的中率はテストデータでの予想結果から見ていく。また的中率の他に各説明変数がどれほど予想に影響しているのかを確認するため、各説明変数の回帰変数も求める。

3.4 実験手順

本研究の実験手順は以下の通りである。

1. `netkeiba.com` から 2019 年、2020 年、2021 年全ての中央競馬のレース結果をスクレイピングする。スクレイピングは `python` を使用して行っており、`pandas` の `DataFrame` 型にして使用する。
2. スクレイピングしているデータは 2019 年、2020 年、2021 年と各年度ごとに分かれているので、3 つのデータを結合し 1 つのデータにする。

3. 結合した3年間分のデータを機械学習できるよう前処理していく。加工したものは、着順、性齢、馬体重、タイム、単勝だ。着順は数値のデータ以外にも除外や取消と文字データも入っている。これらの文字データは今回使用しないので削除した。また勝馬投票券とは関係ない4着以下の着順は全て'4'とし、勝馬投票券に関係のある1着、2着、3着はそのままにし、着中から'rank'とデータ名を変更した。性齢は性別と年齢が1つのデータとなっているので、これを性別と年齢に分けて取り扱いたいため、性別と年齢で分けた。馬体重は性齢と同じようなデータで、当日の体重と前走との体重変化量の2つのデータが1つになっている。このデータも当日の体重と前走との体重変化量の2つに分けて使いたいため、当日の体重と前走との体重変化量とで分けた。タイムはスクレイピングしたままのデータだと'1:31.3'のように60秒以上かかっている場合だと、'60'が'1:'と表記されている。このままだと'.'が入っていることにより学習できないので、'1:'を秒に戻して使用していく。例にあげると'1:31.3'だと'91.3'となる。単勝はデータの型がobject型となっており、それによって機械学習ができないので、float型に変更した。また前処理でされた着順と性齢と馬体重は加工前のデータも列に残っているので、この3種は削除した。それと同時に今回使わない調教師と着差のデータも削除した。
4. データの前処理が終わるとそのデータを、訓練データとテストデータに分ける。訓練データとテストデータの分ける割合は訓練データが8でテストデータは2とした。
5. 今の状態のデータだと目的変数の'rank'が'4'の数が多く、このまま機械に学習してしまうと、'4'を予想しとけば的中するモデルとなってしまうため、それを阻止するために訓練データをアンダーサンプリングする。'4'のデータ数を'1'のデータ数と揃えることにより学習した際'4'のみを予想するモデルではなくなる。またアンダーサンプリングをする前とした後の各データ数は以下の図の通りである。アンダーサンプリング前が図5で、アンダーサンプリング後が図6である。

```
4    86707
1     8088
3     8077
2     8062
Name: rank, dtype: int64
```

図5 アンダーサンプリング前の各データ数

```
1     8088
4     8088
3     8077
2     8062
dtype: int64
```

図6 アンダーサンプリング後の各データ数

図5を見ていただきたい。4着以下のデータは86707もの数があるが、1着のデータは8088、2着のデータは8077、3着のデータは8062と4着以下のデータはその他のデータよ

り 10 倍近くあることがわかる。続いて図 6 を見ていただきたい。1 着、2 着、3 着のデータ数は図 5 と変わらないが、4 着以下のデータ数は 1 着のデータ数と同じ 8088 となっている。これは 4 着以下のデータ全体から 1 着のデータ数と等しくなるようランダムで選ばれている。

6. アンダーサンプリングされた訓練データを用いて機械学習を行う。使用するモデルはロジスティック回帰分析だ。学習後、テストデータの的中率、各特徴量がどれほど学習に影響しているかを確認することによって、精度を確認する。

4 結果と考察

4.1 結果

実験の結果は図 7 のようになった。

	pred	actual
109041	2	3
66721	4	4
74963	3	4
119073	1	3
23198	4	4
...	...	...
120390	4	4
121874	3	4
90589	4	4
57254	4	4
49868	2	4

27734 rows × 2 columns

図 7 実験の結果

図 7 の pred は予想した着順、actual は実際の着順を示している。そして以下の図 8 は的中した割合を示している。的中した割合は 56 %で的中しなかった割合は 44 %だった。

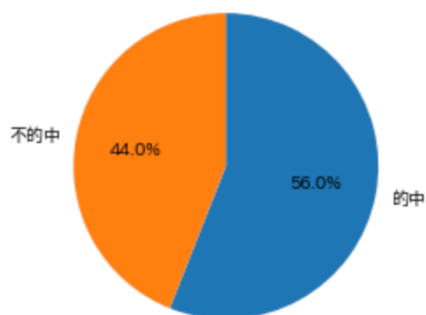


図 8 的中率を表したグラフ

また 1 着になると予想された馬、2 着になると予想された馬、3 着になると予想された馬、4 着以下になると予想された馬それぞれの実際の結果の数と的中率を表した図が図 9 である。

	forecast_1	forecast_2	forecast_3	forecast_4
1place	1331	274	216	201
2place	1095	298	315	307
3place	876	335	338	471
outofreturns	3228	1704	3160	13585
Hit rate	20%	11%	8%	93%

図9 予想された各着順の実際の着順と的中率

図7の「forecast_1」の列は1着と予想された馬の実際の着順について、「forecast_2」の列は2着と予想された馬の実際の着順について、「forecast_3」の列は3着と予想された馬の実際の着順について、「forecast_4」の列は4着以下と予想された馬の実際の着順についてを示している。「1place」の行は実際の着順が1着だった馬、「2place」の行は実際の着順が2着だった馬、「3place」の行は実際の着順が3着だった馬、「outofreturns」の行は実際の着順が4着以下だった馬を示している。例を挙げると、forecast_1と1placeが重なる1331の数字は、1着と予想された馬が実際に1着だった数を示している。また「Hit rate」行はそれぞれの的中率を表している。forecast_1を例に挙げると、これは1着と予想された馬の中からの的中率を述べている。Hit rateを見ると4着以下になると予想した的中率は93%とかなりの確率で的中しているが、勝馬投票券に関係する1着、2着、3着と予想された馬に関しては、的中率は50%も超えておらず、勝馬投票券を的中させるために使用するの是不向きである。

各特徴量がどれほど学習に影響されているか示しているのが図10である。図10の左側に特徴量名、右側にその特徴量の回帰係数が書かれている。

```

人気          -0.065020
単勝          -0.019125
斤量          -0.013714
年齢          -0.007102
馬番          -0.006922
...
性_牡         0.001722
体重         0.001768
騎手_川田将雅 0.002208
騎手_ルメール 0.002491
体重変化      0.006940
Length: 21058, dtype: float64

```

図10 各特徴量の回帰係数の下位5種と上位5種

4.2 考察

研究結果から考察すると、的中率は 56 %と 2 回に 1 度は的中しており、的中率が 100 %には程遠いが競馬予想としてはプロの予想家並みの的中率がある。しかし図 7 からわかるように、的中しているのは大半が 4 着以下になると予想された馬で、勝馬投票券に関係する 1 着、2 着、3 着と予想された馬の的中率は 1 着馬の的中率は 20 %あるものの、その他はどちらも 10 %程しかなく、的中率としては低い。

次に図 8 を見ると体重変化は着順上位に入るための予想には影響のあるデータで、逆に人気のデータは着順が下位になるために影響のあるデータである。回帰係数上位の騎手は昨年の騎手リーディングの 1 位 2 位 [23] の騎手なので着順上位を予想するために必要なデータだと推測できる。回帰係数下位の年齢は数字が増えていくことに馬が老いていっているという意味となり、それが実際にレースの着順がマイナスに影響していると推測できる。斤量は数字が増えるほどその馬への負担ハンデが重たくなるので、斤量が増えるほど着順に響いていると推測できる。また人気と単勝だが、これらは数字が増えると人気がないと判断できる。人気がない馬とはファンから弱いと判断され期待されていない馬ということだ。なので数字が増えると着順は 4 着以下と考えることはできるが、それだと 2021 年のエリザベス女王杯のようなレースは的中できない。しかし回帰係数最底辺であるように、人気のデータはこのモデルでの予想には大きく影響している。ここから推測できるのは 4 着以下の的中率だけが異様に高いのは人気に影響していることが原因ということだ。もちろんテストデータの目的変数に 4 着以下のデータが多いことも関係しているだろうが、人気がないほど弱いと判断されているため、4 着以下になることが多いと推測できる。全体の高い的中率を維持するためには、この人気のデータも必要なデータと言える。

だがこのままだと本研究の目標である的中率 100 %には到底及んでいない。なので新たに特徴量を追加し再度学習させ、的中率が上がるのか実験した。

4.2.1 2 度目の実験

考察を踏まえ新たに特徴量を追加し再度中央競馬の 2019 年から 2021 年の 3 年間のレース結果を用いて予想した。実験手順は前回の実験と同様に行い、訓練データとテストデータの割合も 8:2 と前回と同様にし、特徴量のみを新たに追加する。そして新たに追加した特徴量は以下の表 7 の通りである。

表 7 追加した特徴量

course_len	レースの距離。m 換算で 1800.0 だと、1800m
weather	当日の天気。晴、曇、雨、小雨、雪、小雪がある。
race_type	コースの種類。芝コースとダートコースがある。
ground_state	馬場状態。良、稍重、重、不良がある。

また前回の実験で使った説明変数はそのまま使用する。

4.2.2 2 度目の実験結果

2 度目の実験は図 11 のようになった。

	pred	actual
109041	2	3
66721	3	4
74963	3	4
119073	1	3
23198	4	4
...	...	...
120390	4	4
121874	2	4
90589	4	4
57254	4	4
49868	1	4

27734 rows × 2 columns

図 11 2 度目の実験結果

そして図 12 は的中した割合を表したグラフだ。的中した割合は 48 %で的中しなかった割合は 52 %だった。

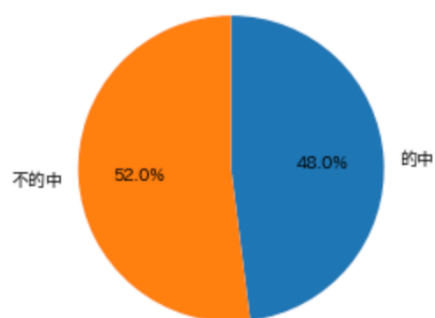


図 12 2 度目の的中率を表したグラフ

また予想された各着順の実際の結果と的中率は以下の図 13 の通りだ。

	forecast_1	forecast_2	forecast_3	forecast_4
1place	1595	158	182	87
2place	1373	222	250	170
3place	1159	249	331	281
outofreturns	4618	2029	3958	11072
Hit rate	18%	8%	7%	95%

図 13 2 度目の予想された各着順の実験の実際の着順と的中率

前回の実験と比べ 4 着以下と予想された馬の的中率は上がったが、1 着、2 着、3 着になると予想された馬に関しては的中率が下がってしまった。

また図 14 は各特徴量の影響度を示している。

```

単勝          -0.051157
人気          -0.022015
タイム       -0.015300
斤量         -0.006046
馬番         -0.003107
...
騎手_川田将雅    0.000502
騎手_ルメール    0.000586
体重          0.000619
course_len    0.001129
体重変化       0.003423
Length: 21072, dtype: float64

```

図 14 2 度目の実験の各特徴量の回帰係数の下位 5 種と上位 5 種

4.2.3 2 度目の実験を経ての考察

特徴量を増やし再度学習させた結果、的中率が下がってしまった。またそれぞれの着順の的中率は 1 着、2 着、3 着と予想された馬に関しても的中率が下がってしまった。しかし 4 着以下と予想した馬に関しては的中率が上がっていた。図 9 と図 13 を比べると特徴量を追加したことにより、1 着と予想された馬の数は増えているが、的中率は前回と比べ下がっていることがわかる。

図 14 を見ると図 10 と比べ下位 5 種と上位 5 種の内容が変わっている。下位 5 種から注目すると、タイムのデータが新たに出てきた。これは新たに特徴量にレースの距離が含まれたことにより、タイムが増えると 4 着以下だと判定していると推測できる。次に上位 5 種に注目すると course_len が上位着順と予想するために必要となっている。だが競馬はレースの距離が増えると上位着順に加わるとは限らない。的中率が下がってしまったのはこのデータが原因だと推測できる。また前回の実験で計算された回帰係数と比べると、今回の上位の回帰係数は前回よりも下がっていて、特徴量を追加したことは逆に予想モデルを悪くしてしまったと推測できる。

2 度の実験を行った結果、ロジスティック回帰分析による競馬予想は、的中率が高くて 56 %

と2回に一度は当たることがわかった。しかし馬券を購入するものからすると、1着、2着、3着と予想した馬の的中率が低く、実際にこのモデルを使用して馬券の買い目を考えるのはリスクではない。

5 まとめと今後の課題

5.1 まとめ

本論文は機械学習手法のロジスティック回帰分析を用いて、競馬予想の的中率を 100 % に近づけることが目標だった。しかし実験をした結果、1 度目の実験が的中率が 56 %、2 度目の実験が 48 % と 100 % には程遠かった。これはロジスティック回帰分析は機械学習手法の中では簡易な手法なのが原因だと推測できる。しかし 4 着以下と予想された馬は 1 度目は 93 %、2 度目は 95 % と高い数値が獲れた。馬券を購入する際も、4 着以下と予想された馬に関しては、買い目には入れずに買うと良い。だが、馬券の買い目を考えるために必要な 1 着、2 着、3 着と予想された馬の的中率が低いのが問題で、この予想モデルを元に馬券を購入するのは避けたほうが良い。

5.2 今後の課題

残っている課題として、全体の的中率を上げることが言える。いまのままだとこのモデルを元に馬券を購入するのはリスクでしかない。また全体的に上げることもそうだが、1 着、2 着、3 着の的中率を上げるのが最大の目的と言える。

5.3 最後に

本研究の目標の的中率 100 % に近づけることだったが、実際には程遠い的中率だった。これを機に改めて競馬予想の難しさを知った。しかし今回使ったモデルに使用した説明変数の他にも競馬には数多くのデータが存在する。競馬にはレース展開の予想やコースの特徴、コースと騎手や馬との相性など、予想に使えるデータは数多く存在する。次に行う際は各競馬場の特徴やレース展開のデータを入れると的中率が変わってくると推測できるので、実際に行っていきたい。

参考文献

- [1] ジャパンカップ (G1) 結果・払い戻し | 2020 年 11 月 29 日東京 12R レース情報, https://race.netkeiba.com/race/result.html?race_id=202005050912,
(参照 2022 年 1 月 11 日)
- [2] エリザベス女王杯 (G1) 結果・払い戻し | 2021 年 11 月 14 日阪神 11R レース情報, https://race.netkeiba.com/race/result.html?race_id=202109050411,
(参照 2022 年 1 月 11 日)
- [3] netkeiba.com, <https://www.netkeiba.com/>,
(参照 2022 年 1 月 20 日)
- [4] 競馬予想・競馬情報なら JRA-VAN, <https://jra-van.jp/>,
(参照 2022 年 1 月 11 日)
- [5] [的中型人工知能], https://yoso.netkeiba.com/?pid=yosoka_profile&id=101,
(参照 2022 年 1 月 12 日)
- [6] Stefan Lessmann: "Identifying winners of competitive events: A SVM-based classification model for horserace prediction", p.1, (2009)
- [7] 佐藤 裕也 華山 宣胤: "Harville の公式と Gini 係数を用いた競馬予想", pp1-2, (2015)
- [8] 芦谷 政浩: "「穴馬への過剰な選好 (longshot bias)」に関するサーベイ", pp1-18, (2010)
- [9] JRA の概要, <https://www.jra.go.jp/company/about/outline/>,
(参照 2022 年 1 月 16 日)
- [10] 重賞競争 (競馬用語辞典) JRA, <https://www.jra.go.jp/kouza/yougo/w326.html>,
(参照 2022 年 1 月 17 日)
- [11] グレード制 (競馬用語辞典) JRA, <https://www.jra.go.jp/kouza/yougo/w323.html>,
(参照 2022 年 1 月 17 日)
- [12] パリミュチュエル方式とは?, <https://baken-no-kaikata.com/parismutuel-method-1651>,
(参照 2022 年 1 月 21 日)
- [13] 馬券のルール JRA, <https://www.jra.go.jp/kouza/baken/index.html>,
(参照 2022 年 1 月 21 日)
- [14] 人工知能と機械学習, https://www.soumu.go.jp/ict_skill/pdf/ict_skill_3_5.pdf,
(参照 2022 年 1 月 26 日)
- [15] 機械学習の 2 つの壁「分類モデルの選定」と「過学習」への対処法, <https://monoist.itmedia.co.jp/mn/spv/2004/07/news017.html>, (参照: 2022 年 2 月 20 日)
- [16] 回帰モデル、回帰分析-Riskpedia, <https://www.riskdatabank.co.jp/rdb/riskpedia/314/>, (参照: 2022 年 2 月 20 日)

- [17] 一般化線形モデル, <https://bellcurve.jp/statistics/glossary/638.html>,
(参照 2022 年 1 月 26 日)
- [18] 順序回帰分析のリンク関数の使い分けについて, <https://www.ibm.com/support/pages/順序回帰分析のリンク関数の使い方について>, (参照日 2022 年 2 月 7 日)
- [19] Scikit-learn で ロジスティック回帰, https://qiita.com/ONE_shoT_/items/b702ab482466df6e5569, (参照 2022 年 2 月 7 日)
- [20] ロジット関数とロジスティック回帰, <https://mathwords.net/logitkansu>, (参照 2022 年 2 月 7 日)
- [21] JupyterLab とは?, <https://su-gi-rx.com/archives/5981>,
(参照 2022 年 1 月 30 日)
- [22] netkeiba.com データベース, <https://db.netkeiba.com/race/>,
(参照 2022 年 1 月 27 日)
- [23] 2021 年リーディングジョッキー (全国), <https://www.jra.go.jp/datafile/leading/j2021.html>,
(参照 2022 年 1 月 31 日)