競馬への機械学習の応用

競馬とは？

- 騎手の乗った馬が着順を競い合う**スポーツ**
- その着順を予測する**ギャンブル**

競馬で勝つ! → 着順を予測したい！

競馬における機械学習で解くべき問題

機械学習と競馬予想が相性の良い理由

- データ数が多い
  毎週土日に, 各レース平均10頭, 100レース
  毎週1000 頭分のデータ数

- 明確に順位という教師が与えられる
  レースが終わるごとに明確に, 順位が数字で付けられ, 予測に用いやすい

目的

実際の競馬では, レース内の**3着以内**の, 勝敗やその組み合わせを予想
それ以外は, 払い戻しに全く関係がない

⇩

その馬が, 3着以内に入るか, 入らないか(1, 0)の予測

データの確保

データは全て, netkeiba.com「https://www.netkeiba.com/?rf=navi」から
2015～2020年の6年分スクレイピング

当日レース情報, 過去レース情報, 各馬情報, 各騎手情報

11R 天皇賞(秋) GI WIN5

15:40発走 / 芝2000m (左) / 天候:曇 / 馬場:良
4回 東京 8日目 サラ系３歳以上 オープン (国際)(指) 定量 12頭
本賞金:15000,6000,3800,2300,1500万円

コース詳細 | IPAT 連携

特集へ ▸ 俺プロへ ▸

出馬表 | オッズ・購入 | 予想 | 調教 | 厩舎コメント | NEW データ分析 | 結果・払戻

出走馬 | 競馬新聞 | 馬柱(5走) | 馬柱(9走) | タイム指数 | 血統 | 対戦表 | 調子偏差値

枠	馬番	印 切替	馬名	性齢	斤量	騎手	厩舎	馬体重 (増減)	オッズ	人気
1	1		ブラストワンピース	牡5	58.0	池添	美浦 大竹	550(+8)	32.3	7
2	2		カデナ	牡6	58.0	田辺	栗東 中竹	478(+2)	146.1	11
3	3		ダイワキャグニー	セ6	58.0	内田博	美浦 菊沢	496(-2)	89.8	10
4	4		ダノンキングリー	牡4	58.0	戸崎圭	美浦 萩原	450(-6)	13.3	3
5	5		ウインブライト	牡6	58.0	松岡	美浦 畠山	484(0)	180.6	12
5	6		フィエールマン	牡5	58.0	福永	美浦 手塚	478(-12)	17.4	5
6	7		クロノジェネシス	牝4	56.0	北村友	栗東 斉藤崇	464(0)	4.4	2
6	8		キセキ	牡6	58.0	武豊	栗東 角居	508(+6)	16.7	4
7	9		アーモンドアイ	牝5	56.0	ルメール	美浦 国枝	490(+2)	1.4	1
7	10		スカーレットカラー	牝5	56.0	岩田康	栗東 高橋亮	488(+14)	42.3	8
8	11		ダノンプレミアム	牡5	58.0	川田	栗東 中内田	496(-10)	21.6	6
8	12		ジナンボー	牡5	58.0	Mデムーロ	美浦 堀	492(+6)	89.0	9

- レース場所
- 天候
- コースの距離
- コースの向き
- 何頭出場するか
- 馬番
- 枠番
- 負担重量
- 馬体重
- 馬体重増減
- 性別
- 人気
- 血統

etc...

レース結果

11R 天皇賞(秋) GI WIN5

15:40発走 / 芝2000m (左) / 天候:曇
4回 東京 8日目 サラ系３歳以上 オープン
本賞金:15000,6000,3800,2300,1500万

着順	枠	馬番	馬名
1	7	9	アーモンドアイ
2	5	6	フィエールマン
3	6	7	クロノジェネシス
4	8	11	ダノンプレミアム
5	6	8	キセキ
6	3	3	ダイワキャグニー
7	8	12	ジナンボー
8	2	2	カデナ
9	7	10	スカーレットカラー
10	5	5	ウインブライト
11	1	1	ブラストワンピース
12	4	4	ダノンキングリー

- 順位
- タイム
- 途中結果

データの前処理

レース内の3着以外は予測しても意味がない

レース場所
天候
コースの距離
コースの向き
何頭出場するか
馬番
枠番
負担重量
馬体重
馬体重増減
性別
人気

etc...

⇨

データの整形
型変更・統一
欠損値埋め
ダミー変数化
ラベルエンコーディング
特徴量追加

11R 天皇賞(秋) GI WIN5

15:40発走 / 芝2000m (左) / 天候:曇

4回 東京 8日目 サラ系３歳以上 オープン

本賞金:15000,6000,3800,2300,1500万

着順	枠	馬番	馬名
1	7	9	アーモンドアイ
2	5	6	フィエールマン
3	6	7	クロノジェネシス
4	8	11	ダノンプレミアム
5	6	8	キセキ
6	3	3	ダイワキャグニー
7	8	12	ジナンボー
8	2	2	カデナ
9	7	10	スカーレットカラー
10	5	5	ウインブライト
11	1	1	ブラストワンピース
12	4	4	ダノンキングリー

着順

1
2
3
4
5
6
7
8
.
.
.

⇨

2値化

1
1
1
0
0
0
0
0
.
.
.

特徴量エンジニアリング例：レース内正規化

ある馬がレースで勝てるかどうかは，その馬の絶対的な能力ではなく，
同じレースの他の馬とのの相対的な能力で決まる

同じレースの馬のデータだけを集めて正規化

	単勝	体重	体重変化	着順平均_5R	獲得賞金平均_5R	勝率_1Y	連対率_1Y	複勝率_1Y
201503010101	2.0	484	8	4.800000	91.000000	0.043	0.093	0.143
201503010101	11.6	494	-8	5.000000	83.333333	0.029	0.062	0.094
201503010101	250.2	472	-6	12.500000	0.000000	0.020	0.040	0.067
201503010101	35.4	508	0	0.000000	NaN	0.023	0.048	0.089
201503010102	108.6	482	-2	8.000000	0.000000	0.017	0.047	0.081
201503010102	7.1	436	-2	6.000000	0.000000	0.076	0.149	0.235
201503010102	75.2	456	-1	14.000000	0.000000	0.020	0.040	0.067
201503010102	42.4	484	-2	12.000000	0.000000	0.034	0.092	0.185

同じレース

レース内正規化

	正規化_単勝	正規化_体重	正規化_体重変化	正規化_着順平均_5R	正規化_獲得賞金平均_5R	正規化_勝率_1Y	正規化_連対率_1Y	正規化_複勝率_1Y
201503010101	0.000000	0.693878	0.900	0.300000	1.000000	0.500000	0.513812	0.572000
201503010101	0.038678	0.795918	0.100	0.312500	0.915751	0.337209	0.342541	0.376000
201503010101	1.000000	0.571429	0.200	0.781250	0.000000	0.232558	0.220994	0.268000
201503010101	0.134569	0.938776	0.500	0.000000	NaN	0.267442	0.265193	0.356000
201503010102	0.813364	0.857143	0.500	0.571429	0.000000	0.000000	0.049645	0.076503
201503010102	0.033794	0.387755	0.500	0.428571	0.000000	0.855072	0.773050	0.918033
201503010102	0.556836	0.591837	0.625	1.000000	0.000000	0.043478	0.000000	0.000000
201503010102	0.304916	0.877551	0.500	0.857143	0.000000	0.246377	0.368794	0.644809

	features	importance
36	オッズ_0	8088
254	season_cos	8014
201	正規化_前走着差	7136
56	前走着差	6757
249	days_interval_0	6669
224	正規化_獲得賞金平均_allR	6352
3	horse_id	6347
186	正規化_オッズ_0	5647
61	前走上り	5322

モデル作成・パラメータ最適化

- 機械学習アルゴリズムは, **LightGBM**を選択

  決定木アルゴリズムに基づいた勾配ブースティングの機械学習フレームワーク

  1, 欠損値をそのまま扱える
  2, 特徴量のスケーリングが不要
  3, feature importanceが確認できる
  4, 精度が出やすくKaggleでもよく用いられている

- ハイパーパラメータは, Optunaを使って最適化

- いろいろなモデルを模索

  3着以内に入るか入らないか（分類）を予測するモデル
  2着以内に入るか入らないか（分類）を予測するモデル
  タイム（回帰）を予測するモデル
  オッズを予測して（回帰）能力と人気の乖離を予測するモデル
  レース内の馬の相対的な強さ（ LambdaRank ）を予測するモデル
  ↑のアンサンブルモデル

  etc …

結果・検証①

2020/11/01天皇賞(秋)

予測

⑨ – ⑦ - ⑥

馬番	馬名	人気	予測確率
9	アーモンドアイ	1	1.0
7	クロノジェネシス	2	0.691
6	フィエールマン	5	0.541
8	キセキ	4	0.452
4	ダノンキングリー	3	0.399
11	ダノンプレミアム	6	0.365
12	ジナンボー	9	0.276
1	ブラストワンピース	7	0.27

予測が当たっている！

結果

⑨ – ⑥ - ⑦

11R 天皇賞(秋) GI WIN5

15:40発走 / 芝2000m (左) / 天候:曇

4回 東京 8日目 サラ系３歳以上 オープン

本賞金:15000,6000,3800,2300,1500万

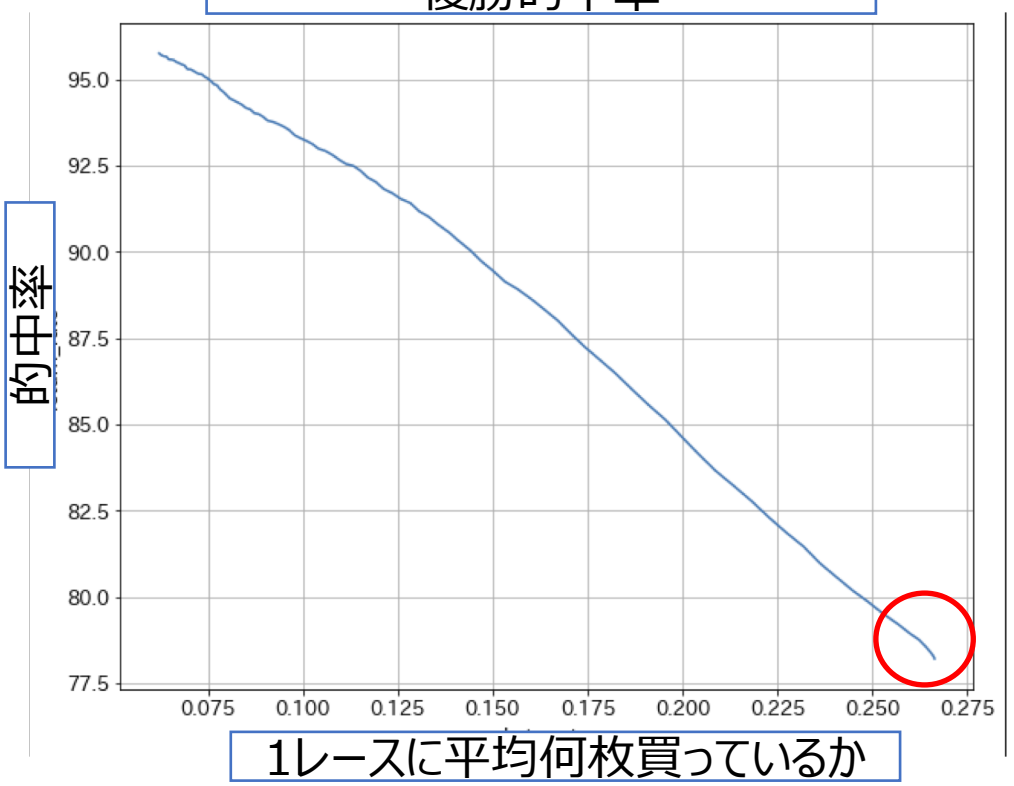
着順	枠	馬番	馬名
1	7	9	アーモンドアイ
2	5	6	フィエールマン
3	6	7	クロノジェネシス
4	8	11	ダノンプレミアム
5	6	8	キセキ
6	3	3	ダイワキャグニー
7	8	12	ジナンボー
8	2	2	カデナ
9	7	10	スカーレットカラー
10	5	5	ウインブライト
11	1	1	ブラストワンピース
12	4	4	ダノンキングリー

結果・検証②

テストデータ(2018年~2020年途中, 約5000レース)の単勝の回収率と複勝の的中率

1着を当てる
単勝回収率

3着に入るかどうかを当てる
複勝的中率



枚数を単勝回収率85%, 複勝的中率75%達成

今後やりたいこと

単勝回収率85%, 複勝的中率75%達成

このままでは勝てない

- 完璧に予測が可能であるわけではない

  どのぐらいのモデルの予測で買うべきか?

- 人気がある馬では的中した際の払戻金は少ない

  どれぐらいのリターンが見込めたら買うべきか？

- **馬券の買い方は, 自分次第(予測はできても, 買い方を間違える)**

  「複勝」は一番当てやすく, 一番配当の低い賭け方

  複勝, 単勝, ワイド, 3連複, 3連単, 馬連, 馬単

期待値の高い勝負を続ける必要がある!

数理最適化？

GAなどによって, 近似解を探索したい

期待値を考慮した, 馬券の買い方や組み合わせを最適化

概要とアピールポイント

概要

- 機械学習を勉強したい + 好きな競馬に機械学習を応用したい
- 趣味で, 競馬の順位を予測するAIを自作
- テストデータ(2018年~2020年)で
  単勝回収率85%, 複勝的中率75%達成
- 数理最適化を使うことで,
  期待値を考慮した買い目や賭け金の最適化に挑戦したい

アピールポイント

- 独学でPython文法・機械学習を学んだ
- データのスクレイプ, 特徴量加工, モデル作成, パラメータ最適化,
  検証など全体を通して行った
- 2020年5月から8月までの短期間で開発した
- とりあえず最後予測できるまで作り上げることを意識した