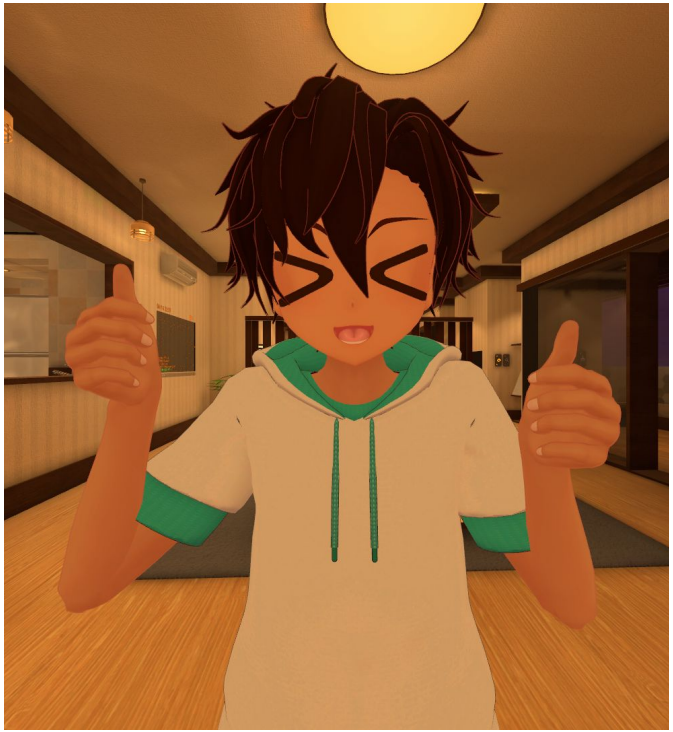
楽して競馬に勝ちたい！

機械学習初心者が
競馬予測モデルを作った話

#個人開発集会 2024.6.20 ラクティ

自己紹介

名前：ラクティ(rakti)
職業：Sier(Javaを使った業務システムの開発)
VRC：ラクティ(2024年1月～)
個人開発：2013年3月～(やったりやらなかったり…)
使用言語：Java(仕事),Python(仕事/個人),Ruby(個人)
ひとこと：機械学習もLTも初心者なので
生暖かく見守っていただけると幸いです！



目次

1. 作ろうと思ったきっかけ
2. 作成した内容
3. 作ってみて思ったこと
4. 困っていることや課題
5. 今後の展望
6. まとめ
7. おまけ（時間があれば）

作ろうと思ったきっかけ

競馬の予想システムを
作って楽して馬券で勝ちたい！

（後、機械学習の勉強のため）

競馬の魅力

競馬の魅力はギャンブルだけじゃない！

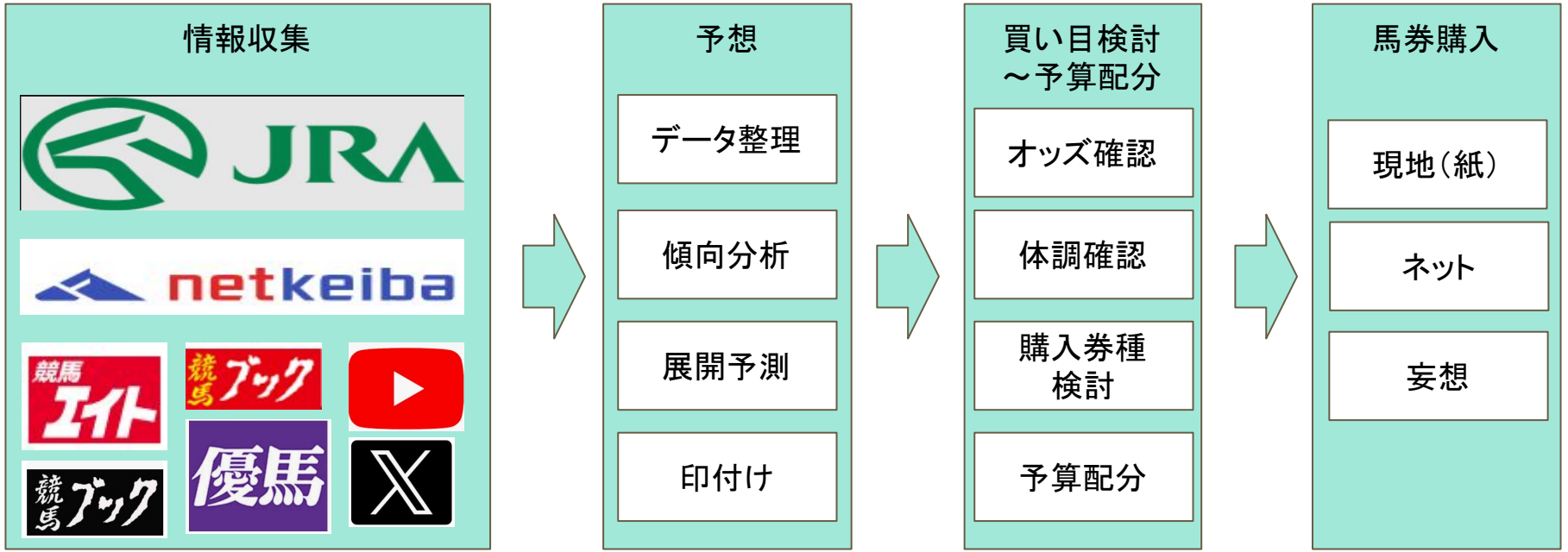
- 競馬は馬が主役のスポーツ
  - 馬同士が着順を競う競技
  - レースでは人馬の一瞬の駆け引きが行われ、展開次第で結果が大きく変わる
- レースの迫力がすごい
  - 実際にレースを見ると躍動感がすごい！
  - G1にもなると熱狂がすごい！興奮する！
- 推し活ができる
  - 推しの馬が勝った！
  - 推しの騎手がG1勝った！
  - 推しの馬の子供が勝った！

じゃあ競馬で一番嬉しい瞬間は？

自分の馬券が当たった時！！！

馬券で勝つには？

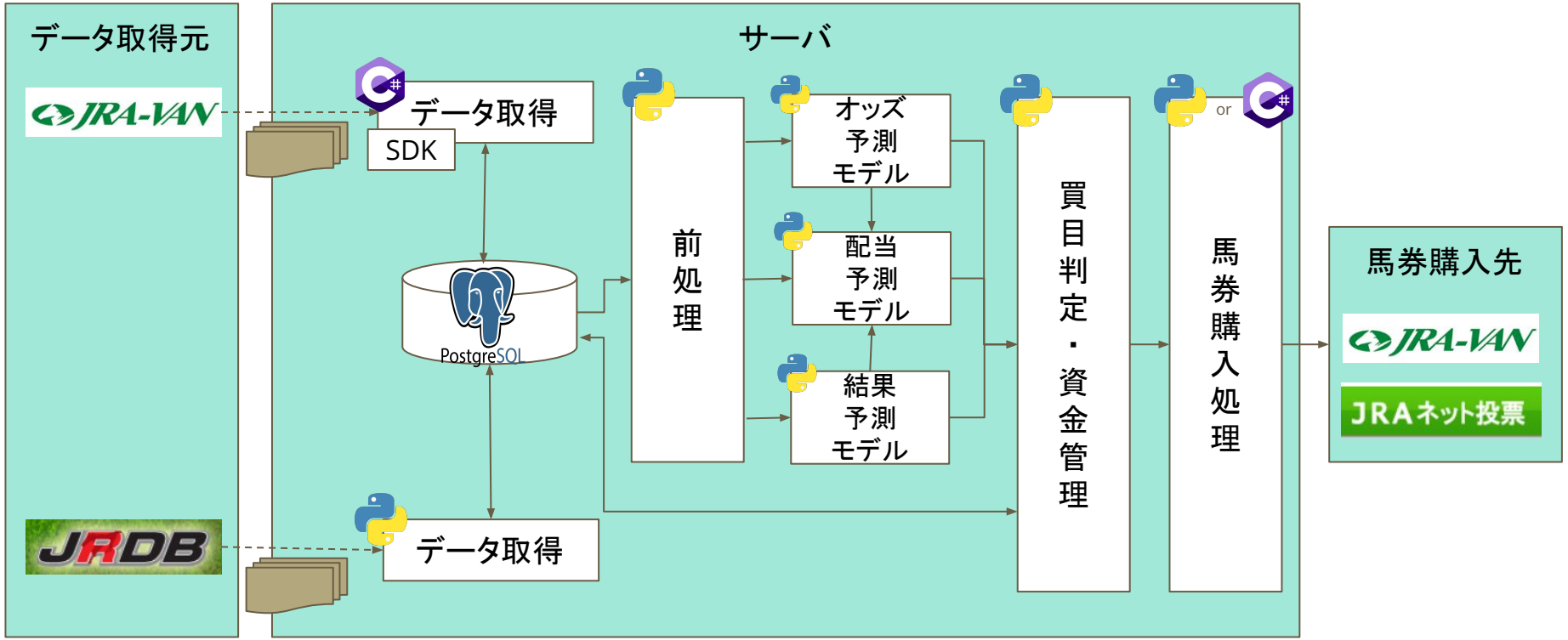
予想から買うまでは工程がいっぱい必要になる



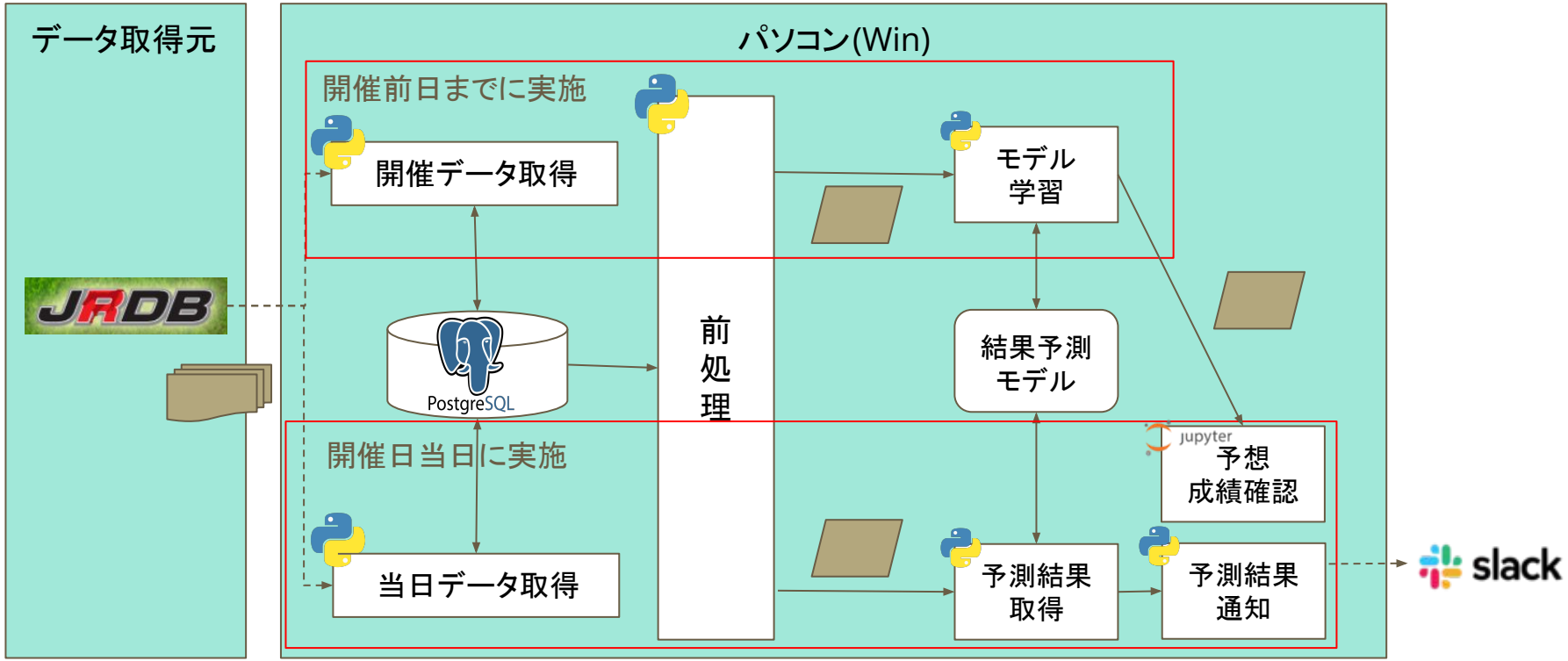
代表的なデータ取得元の比較

	JRA-VAN DataLab.	netkeiba.com	JRDB
月額料金	2,090円	0～1027円	1,980～2,480円
開発環境	Windows・Visual Studio	なんでもOK	なんでもOK
信頼度	公式なので◎	○	○
データ種類	○	△（無料だと指数値無し）	◎
手軽さ	△	◎	○
所感	自動購入やオッズ解析がしたい人は避けては通れないが、SDKがWin環境でしか使えないのが弱点。	スクレイピング用のソースがGitHubなどで公開されているので気軽に始められる。	公式にない独自のデータが多い。情報の取得もURLを叩いてDLする方式なのでどの環境でも使える。

作る予定だったシステムの概略図

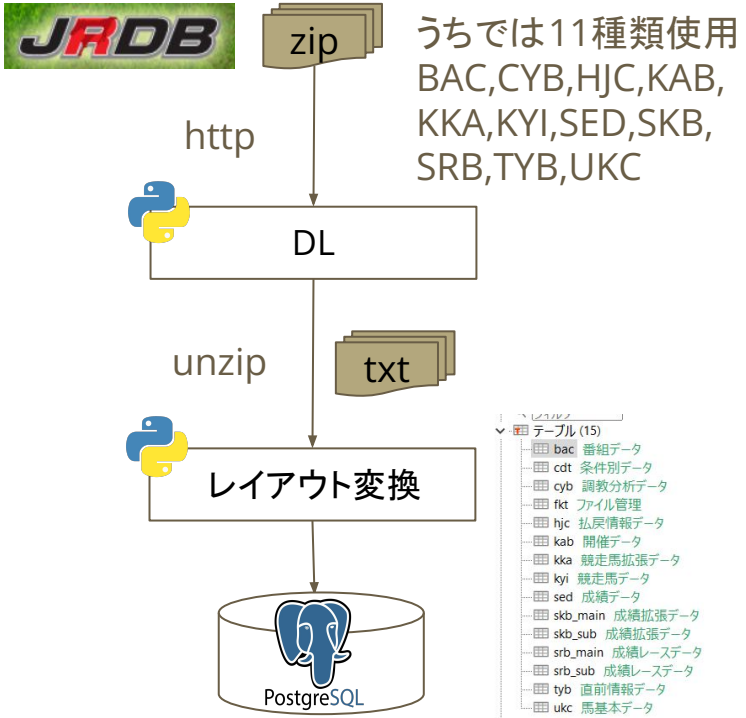


現在実装済みの機能


データ取得元
パソコン(Win)
開催前日までに実施
開催データ取得
モデル
学習
JRDB
PostgreSQL
前
処
理
結果予測
モデル
開催日当日に実施
jupyter
予想
成績確認
当日データ取得
予測結果
取得
予測結果
通知
slack

データの取得

1. JRDBから各データファイル(zip)を
   ダウンロードして解凍（固定長、詳細）
   zipファイルは週ごとだけでなく、
   1年分をまとめたファイルもある。
2. txtファイルを項目ごとに分割する。
   （当時はパース処理を公開してくれている人が
   いなくてすごく大変だった）
   分割した内容を後で分析や加工が
   やりやすいようにDBに入れる



ファイルのレイアウトの例

項目名	OCC	BYTE	TYPE	相対	備考
レコード長：292BYTE					
血統登録番号		8	X	1	
馬名		36	X	9	全角１８文字
性別コード		1	9	45	1:牡,2:牝,3,セン
毛色コード		2	9	46	コード表参照
馬記号コード		2	9	48	コード表参照
血統情報					
父馬名		36	X	50	全角１８文字
母馬名		36	X	86	全角１８文字
母父馬名		36	X	122	全角１８文字
生年月日		8	9	158	YYYYMMDD
					===以下第２版にて追加===
父馬生年		4	9	166	YYYY 血統キー用
母馬生年		4	9	170	YYYY 血統キー用
母父馬生年		4	9	174	YYYY 血統キー用
馬主名		40	X	178	全角２０文字
馬主会コード		2	99	218	競馬場毎にある。場コードと同じ
生産者名		40	X	220	全角２０文字
産地名		8	X	260	全角４文字
登録抹消フラグ		1	9	268	0:現役,1:抹消
データ年月日		8	9	269	YYYYMMDD
					===以上第２版にて追加===
					===以下第３版にて追加===
父系統コード		4	9	277	
母父系統コード		4	9	281	
					===以上第３版にて追加===
予備		6	X	285	スペース
改行		2	X	291	ＣＲ・ＬＦ

注記

・馬主会コードについて

参考データです。設定されていないデータが有ります。

・系統コードについて

前２桁：血統を１２系統に分けた大系統コードです。

後２桁：小系統コードです。

取得データの確認

- 実際に使用するデータの範囲を決める（うちでは2009年以降を使用）
- 取得したデータから分析の邪魔になるデータを取り除く

  （雪や台風などによる開催中止や、競争中止、発走除外etc・・・）

  時々ファイルが破損してて変なデータがいるので注意

- データを眺めたり（馬名を眺めたり）、ヒートマップを眺めたり、

  Pytorchにかけてみたりして使えそうな項目を絞り込んでいく

  →一番わくわくして楽しかった時期

学習データの前処理

2024年5月26日（日曜） 2回東京12日　発走時刻：15時40分

印刷用ページ

11R 第91回 東京優駿 GI

3歳 オープン（国際）牡・牝（指定） 馬齢　コース：2,400メートル（芝・左）

本賞金（万円） 1着 30,000　2着 12,000　3着 7,500　4着 4,500　5着 3,000
付加賞（万円） 1着 2,886.8　2着 824.8　3着 412.4

枠	馬番	馬名 戦績 / 総賞金 馬主名 / 調教師名 / 血統	性齢/毛色 負担重量 騎手名 プレレーティング	前走	前々走	3走前	4走前
1	1	サンライズアース (2.0.0.1) 2743.1万 (株)ライフハウス 石坂 公一（栗東） 父：レイデオロ 母：シャンドランジュ (母の父：マンハッタンカフェ)	牡3/栗 57.0kg 池添 謙一 108 L	2024年4月14日 中山 皐月賞 GI 12着 17頭 15番 12番人気 M.デムーロ 57.0kg 2000芝 1:58.5 良 103 532kg 11 11 11 13 3F 35.2 ジャスティンミラノ (1.4)	2024年2月24日 阪神 すみれS L 1着 10頭 9番 7番人気 M.デムーロ 57.0kg 2200芝 2:12.0 良 108 542kg 9 9 1 1 3F 35.0 ジューンテイク (0.3)	2023年12月28日 中山 ホープフルS GI 取消 16頭 17番 M.デムーロ 56.0kg 2000芝 良 レガレイラ	2023年10月22日 京都 新馬 1着 8頭 8番 7番人気 M.デムーロ 56.0kg 2000芝 2:01.4 良 538kg 1 1 1 1 3F 35.2 ヴィスマール (0.1)
1	2	レガレイラ (2.0.1.1) 8277.9万 (有)サンデーレーシング 木村 哲也（美浦） 父：スワーヴリチャード 母：ロカ (母の父：ハービンジャー)	牝3/鹿 55.0kg C.ルメール 113	2024年4月14日 中山 皐月賞 GI 6着 17頭 10番 1番人気 北村 宏司 55.0kg 2000芝 1:57.6 良 109 456kg 14 15 14 13 3F 33.9 ジャスティンミラノ (0.5)	2023年12月28日 中山 ホープフルS GI 1着 16頭 13番 1番人気 C.ルメール 55.0kg 2000芝 2:00.2 良 113 454kg 14 14 11 10 3F 35.0 シンエンペラー (0.1)	2023年10月21日 東京 アイビーS L 3着 6頭 3番 1番人気 C.ルメール 55.0kg 1800芝 1:48.4 良 104 456kg 3 3 3 3F 32.7 ダノンエアズロック (0.2)	2023年7月9日 函館 新馬 1着 9頭 6番 1番人気 C.ルメール 55.0kg 1800芝 1:49.8 良 452kg 5 6 6 5 3F 34.3 セットアップ (0.2)
2	3	ジューンテイク (3.1.0.5) 9820.5万 吉川 潤 武 英智（栗東） 父：キズナ 母：アドマイヤサプリナ (母の父：シンボリクリスエス)	牡3/黒鹿 57.0kg 岩田 望来 111	2024年5月4日 京都 京都新聞杯 GII 1着 15頭 1番 8番人気 藤岡 佑介 57.0kg 2200芝 2:11.2 良 110 486kg 3 4 4 5 3F 33.6 ウエストナウ (0.2)	2024年3月16日 阪神 若葉S L 5着 9頭 2番 4番人気 和田 竜二 57.0kg 2000芝 2:00.3 良 101 486kg 5 5 4 6 3F 34.8 ミスターシービー (0.6)	2024年2月24日 阪神 すみれS L 2着 10頭 2番 3番人気 岩田 望来 57.0kg 2200芝 2:12.3 良 105 488kg 7 7 8 7 3F 34.4 サンライズアース (0.3)	2023年12月17日 阪神 朝日杯FS GI 4着 17頭 14番 11番人気 M.デムーロ 56.0kg 1600芝 1:34.0 良 111 482kg 15 8 3F 34.9 シャンタルマンタル (0.2)

取得したデータから取捨選択をして、
モデルに食べさせる自分だけの競馬新聞
（情報）を作成する。

人が予想する際に使っている情報や
苦労して分析したデータを用意しても
案外モデルの精度が上がらないのが
とってもツライ

作成したモデルについて

手法：LightGBM（lambdarank）

目的変数：3着以内に入る確率が高い馬（順位）

特徴量：129項目

→作成したデータファイルの項目自体は606項目

それから、Null importancesやboruta等を使って特徴量選択

データ量：42,491レース(596,868行)

評価指標：NDCG

実際に運用してみる

実際に5/26(日)に東京競馬場へ行ってモデルの予想で馬券を買ってみた

朝やること1

当日の朝にパソコンで起動するとSlackに当日対象のレース情報が通知される

Python_bot アプリ 09:44

予想対象レース情報

東京08R, レース名:青嵐賞, 出走時間:1345, 距離:2400

東京09R, レース名:むらさき賞, 出走時間:1420, 距離:1800

東京11R, レース名:東京優駿・Ｇ１, 出走時間:1540, 距離:2400

京都11R, レース名:白百合ステークス, 出走時間:1605, 距離:1800

東京12R, レース名:目黒記念・Ｇ２, 出走時間:1700, 距離:2500

朝やること2

モデルの過去成績から勝てそうな券種と頭数を確認する

東京競馬場 クラス:OP(5) 芝:2400m

モデルの予想の上位6頭で買う場合

的中率(hit):73.9%

回収率(ret):133.9%で

馬単(BOX)買った場合に勝てそう！！

belongs_class	distance		type	hit	ret	win
5	2400.0	0	単勝	0.869565	1.342029	18520.0
		1	複勝	1.000000	0.813043	11220.0
		2	ワイド	0.869565	0.875072	30190.0
		3	馬連	0.739130	0.999420	34480.0
		4	馬単	0.478261	0.868696	29970.0
		5	馬単_BOX	0.739130	1.339420	92420.0
		6	3連複	0.478261	1.080000	49680.0
		7	3連単	0.130435	0.845000	38870.0
		8	3連単_BOX	0.478261	1.539275	424840.0

当日の朝に見た過去成績

馬券購入

レース開始の15分ぐらい前に予想結果が通知されるので、馬券を購入する。

Python_bot アプリ 15:20

東京11R, レース名:東京優駿・Ｇ１, 出走時間:1540, 距離:2400

0,馬番:2,予想値:-1.294

1,馬番:15,予想値:-1.715

2,馬番:12,予想値:-2.675

3,馬番:6,予想値:-3.385

4,馬番:9,予想値:-4.234

5,馬番:13,予想値:-4.249

6,馬番:8,予想値:-4.563

7,馬番:18,予想値:-6.525

実際の予想結果

購入した馬券

結果（負けました！！！）

レースの結果

順位	馬番	人気	馬名
1	5	9	ダノンデザイル
2	15	1	ジャスティンミラノ
3	13	7	シンエンペラー
4	1	15	サンライズアース
5	2	2	レガレイラ
6	6	6	コスモキュランダ

払い戻し

単勝	5	4,660円	9人気
複勝	5	700円	9人気
	15	120円	1人気
	13	380円	7人気
枠連	3 - 7	820円	3人気
馬連	5 - 15	6,860円	20人気
ワイド	5 - 15	1,380円	14人気
	5 - 13	8,000円	53人気
	13 - 15	660円	6人気
馬単	5 > 15	21,490円	52人気
3連複	5 - 13 - 15	21,250円	55人気
3連単	5 > 15 > 13	229,910円	504人気

その他の成績

当日対象の5レースのうち、3レースで完全的中

馬単6頭BOXで全レースを買っていた場合、

回収率は58%（配当:8,830円/購入:15,000円)だった（なお本人は全負け）

	レース名	予想	結果(人気)	配当
東京08R	青嵐賞	4,5,10,3,11,8	11(4),5(2),4(1)	1,2,3着的中！　馬単：5,260円
東京09R	むらさき賞	10,1,8,6,5,3	10(1),1(2),5(5)	1,2,3着的中！　馬単：480円
京都11R	白百合ステークス	3,2,8,1,7,9	8(3),2(2),3(1)	1,2,3着的中！　馬単：3090円
東京12R	目黒記念	4,9,2,8,5,12	9(1),6(10),4(2)	1,3着的中！　馬単：0円

作ってみて思ったこと

モデルの予想結果だけでそのまま購入しても回収率が100%を超えられない

原因1:モデルの予想結果を分析すると競馬場毎、距離毎、クラス毎に得意不得意がある

対策1:モデルを競馬場毎、距離毎、クラス毎に分けて作成する

（レースの数が減ることでデータ量も少なくなるジレンマ）

原因2:購入点数が多くなる(6頭ボックスで買うと当たるが購入金額でマイナスになる)

対策2:モデルの傾向を分析して自分で購入時に買目を絞る

（実はランクで4~5番目の馬が当たりやすい・・・ちゃんと予想に反映して？？）

→これらをすることで回収率が100%を超えた！

回収率が高くなる買目を判断する仕組みを作って自動化するとより勝ちやすくなる

困っていることや課題

現状の課題:モデルの予測精度向上

困っていること:テストデータの数が足りない

競馬場毎、距離毎、クラス毎に分けるとレース数が限られてくるので、

学習に使用するデータが少なくなってしまう。

→水増しするにもダミーデータの整合性をどう取れば良いかわからない・・・

今後の展望

- 未実装機能の実装

  リアルタイムオッズの取得や、アプリから馬券を買うための以下の機能を作成をしたい

  1. オッズ予測
  2. 配当予測（荒れるか判定）
  3. 買目判定
  4. 自動購入機能

- 新しいデータの活用

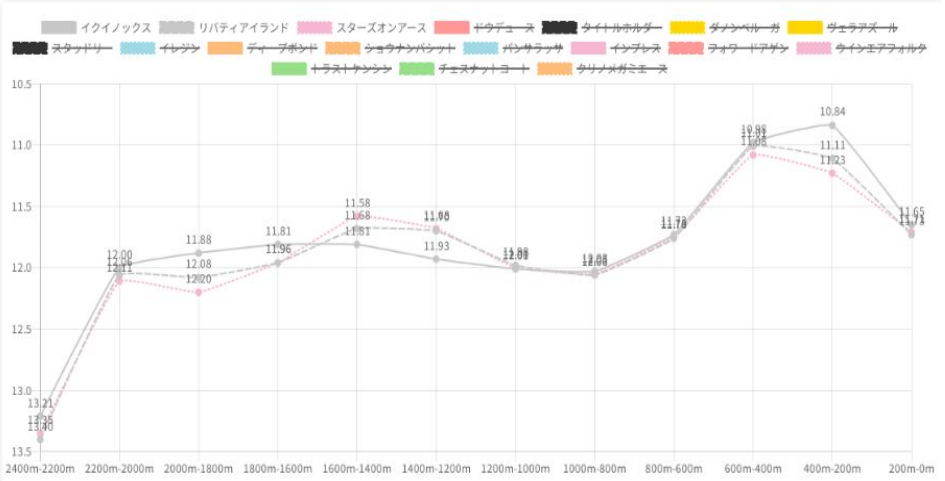
  JRDBで取得できるデータ以外を活用してみたい

  1. 各馬のラップタイム（DERBY ROOM）

     ※値段が高い・・・

  2. パドックやレース映像

     ※固定のアングルじゃない・・・



まとめ

やっぱり簡単じゃなかった！！

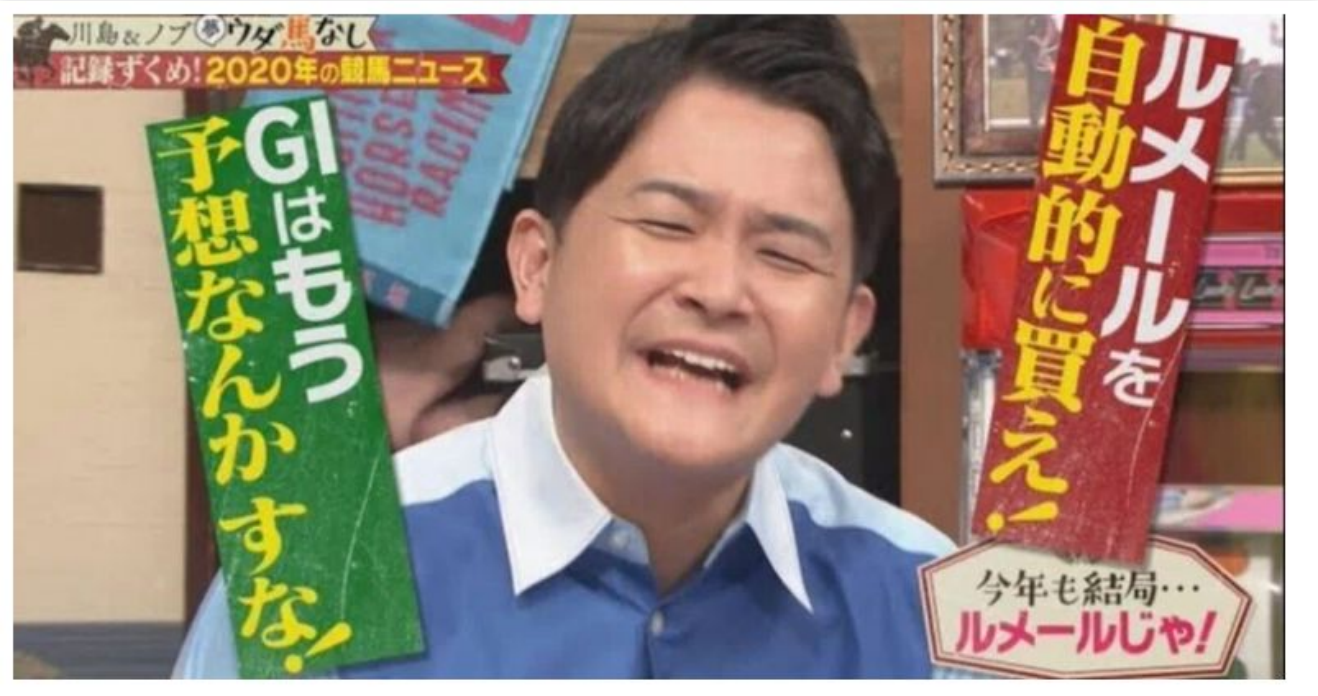
データ自体はサイトからすぐ取得できるので
機械学習の題材としては良い！

データに現れない要素も多いので、
勝つにはしっかりとした分析や仮説検証が必要！！

競馬は実際にやっても面白いので、ぜひ1度は競馬場に遊びに行ってみて！！

おまけ

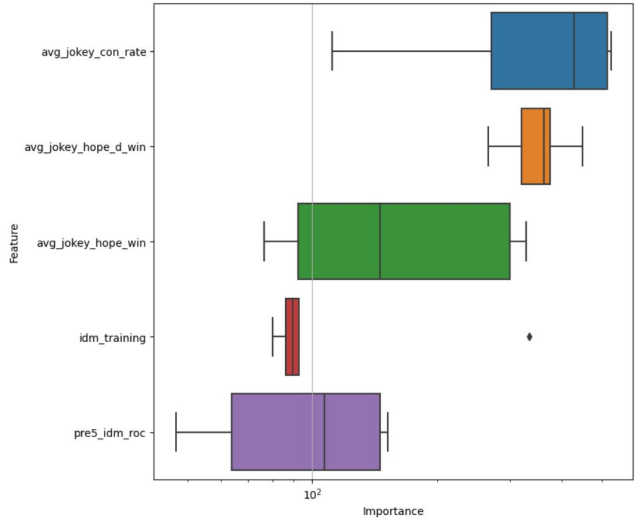
ルメールを自動的に買え！！は本当？



うちのモデルでいうと・・・

うちのモデル※で一番重要視している特徴量は騎手の連帯率（複勝率）

優先度	特徴量名(物理)	特徴量名（論理）
1	avg_jokey_con_rate	騎手期待連対率
2	avg_jokey_hope_d_win	騎手期待3着内率
3	avg_jokey_hope_win	騎手期待単勝率



※東京競馬場,芝,2勝クラス以上,1800m<=2400mの場合

去年(2023)のG1での騎手期待連対率

去年(2023)のG1で騎手期待連対率が一番高かったのは？

騎手名	騎手期待連対率平均値(%)
川田将雅	36.37
C. ルメール	36.16
R. ムーア	27.73

結論！！

川田とルメールを
自動的に買え！！

参考文献

■http://stockedge.jp/の中の人による技術メモ

https://stockedge.hatenablog.com/entry/2016/01/03/103428

■AlphaImpact 開発者ブログ
https://alphaimpact.co.jp/blog/

https://alphaimpact.co.jp/downloads/pydata20161124.pdf

■netkeiba

https://netkeiba.com/

■JRDB

http://www.jrdb.com/

■PC-KEIBA

https://pc-keiba.com/wp/