


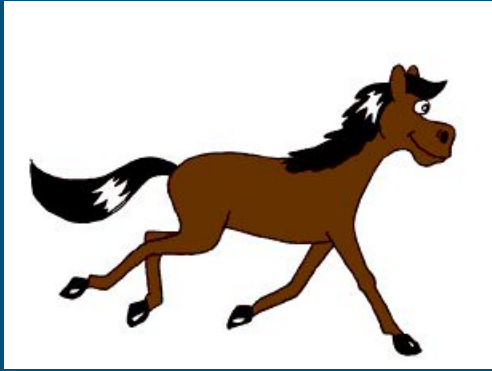


Road to Winning at Horse Racing with Data Science

2018/06/27(Wed)
AlphaImpact
NUKUI Shun



Data x



+



How to win?



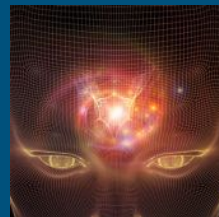
Profile

- Name: NUKUI Shun
- Speciality: Machine Learning
- Experience of horse racing: 11 years
- Bio
 - The president of horse racing club in Tokyo Tech ~2017/03
 - Participated in 電腦賞(春) 2016/03
 - **AlphaImpact (development of horse racing AI) 2016/06~**
- Favorites: Watching training progress of LightGBM, Kaggle(Home Credit Default Risk)



AlphaImpact Project

- Developing horse racing AI
- Members
 - NUKUI Shun : Machine Learning, Horse racing domain knowledge
 - OMOTO Tsukasa : Machine Learning, System Architect, One of committers of LightGBM
 - HARA Tomonori : Horse racing hacker
- Activities
 - HP: <https://alphaimpact.jp/>
 - Published AI scores for free (~2018/06/24)
 - Sell predictions on netkeiba ウマい馬券
 - <http://yoso.netkeiba.com/>



Road to winning

1. Introduction to the system of horse racing
2. Definition of objective
3. Feature engineering
4. Prediction model
5. Evaluation

Road to winning

1. Introduction to the system of horse racing
2. Definition of objective
3. Feature engineering
4. Prediction model
5. Evaluation

What is the Purpose of Horse Racing Prediction?

- Hit?

What is the Purpose of Horse Racing Prediction?

- — Hit?

- **MAKE A PROFIT!!**

Which should we bet on?

1

2

3

win proba. (勝率)

70%

20%

10%

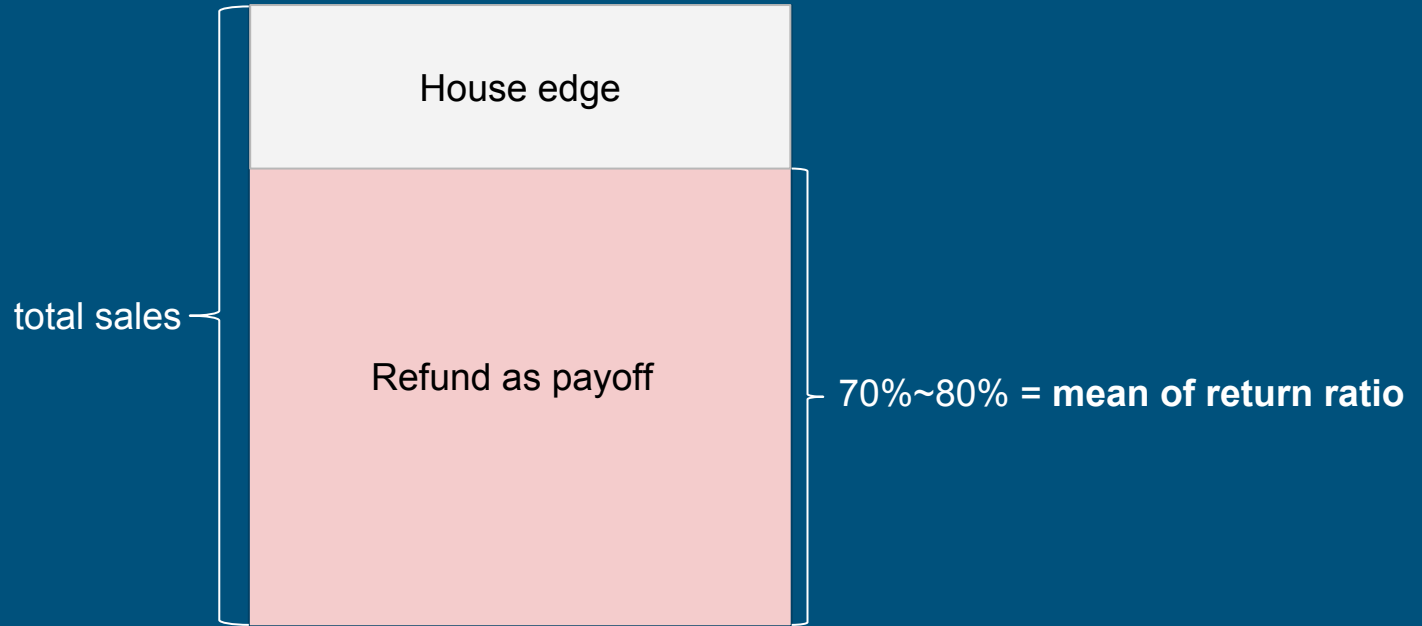
↑
best choice?

Which should we bet on?

	1	2	3
win proba. (勝率)	70%	20%	10%
odds (オッズ)	x1.1	x5.2	x8.0
exp. (期待値)	$1.1 * 0.70 = 0.77$	$5.2 * 0.20 = \mathbf{1.04}$	$8.0 * 0.10 = 0.8$

↑
best choice!!

House Edge (控除率)



Betting Type

house edge

- Win (単勝) 20.0%
- Place (複勝)

- Bracket Quinella (枠連)
- Quinella (馬連) 22.5%
- Quinella Place (ワイド)

- Exacta (馬単) 25.0%
- Trio (3連複)

- Trifecta (3連単) 27.5%

Not Impossible to Win

- Mr. 冨 (Manji) made a profit of 140 million in a 3 years
- His theory and analytical skill is great, but we believe it is not impossible to exceed him by machine learning
- Not hitting tickets are acknowledged as expenses under certain conditions



Road to winning

1. Introduction to the system of horse racing
2. Definition of objective
3. Feature engineering
4. Prediction model
5. Evaluation

What should we solve?

- classification
 - imbalance problem
- regression
 - better choice
 - important to design smoothed objective
- ranking
 - seems to be a natural choice too
 - but not better performance than regression

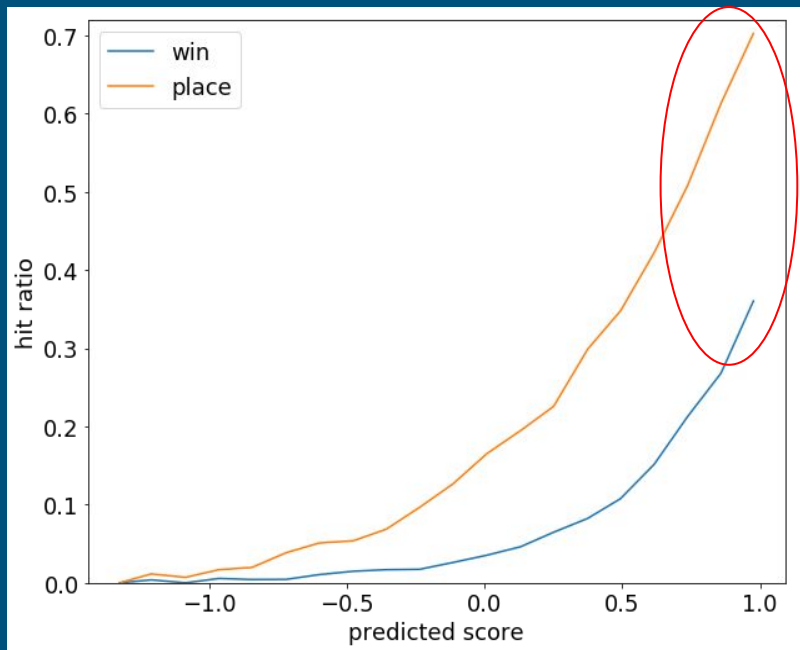


Objective of Regression

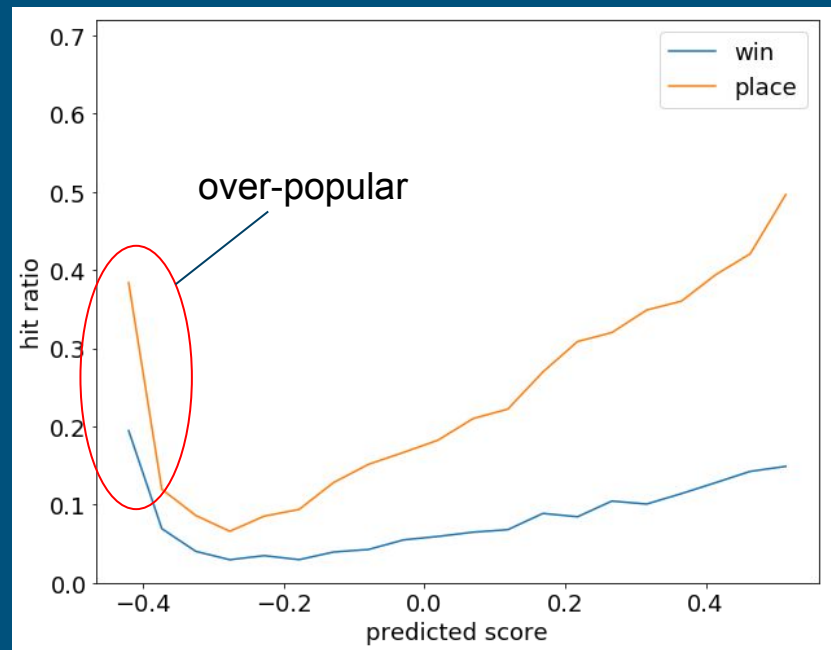
- Strength
 - place of order (着順)
 - standardized time (標準化走破タイム)
 - **standardized velocity (標準化平均速度)**
 - prize (賞金)
 - speed index (スピード指数)
- Profit
 - place (=within top3) payoff (複勝払い戻し)
 - 1st 120 yen < 3rd 540 yen
 - **dark score (business secret)**
 - **transform strength score to be high correlated with return ratio**

Two types of Objective vs Hit Ratio

all turf races in 2017



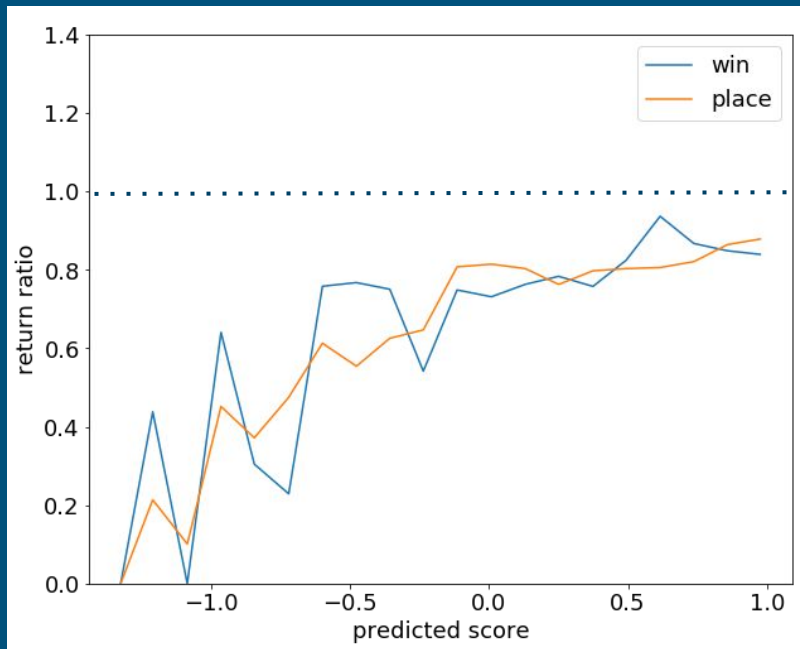
standardized velocity



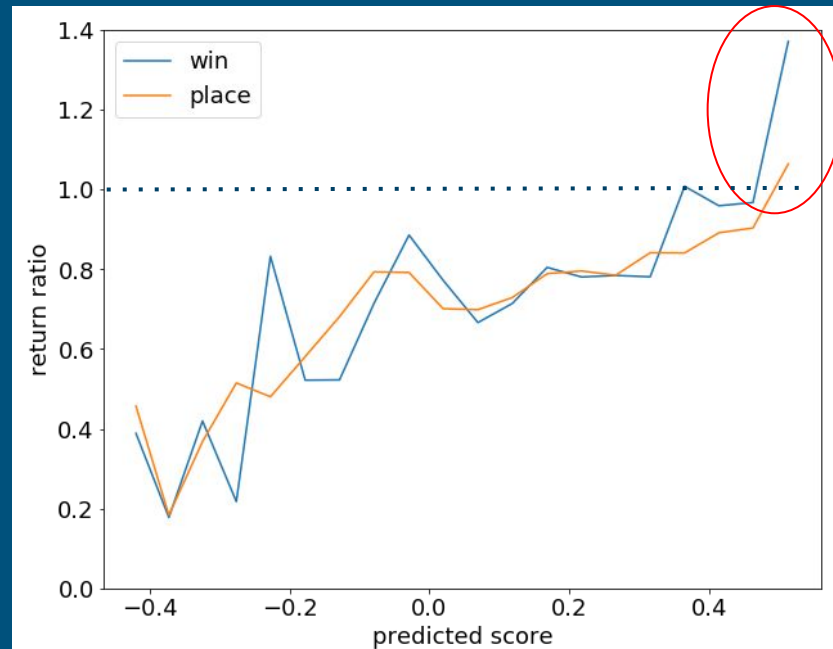
dark score

Two types of Objective vs Return Ratio

all turf races in 2017



standardized velocity



dark score

Road to winning

1. Introduction to the system of horse racing
2. Definition of objective
3. Feature engineering
4. Prediction model
5. Evaluation

Feature Engineering in Horse Racing

- Most important and most time-consuming part
- Necessary to collect data by ourselves, unlike Kaggle
- Difficult to handle complicated structured data
- Requires deep domain knowledge to horse racing

Horse Table (馬柱)

4R		レース結果 オッズ					
サラ系3歳 未勝利 (混合) [指定] 馬齢		コース 1600m 芝・左					
本賞金(万円): 500、200、130、75、50		発走 11:25					
		馬柱の見方					
枠	馬番	馬名 / 単勝オッズ(人気) 馬体重 馬主名 / 調教師名 / 血統	性別/毛色 負担重量 騎手名	過去4走成績			
				前走	前々走	3走前	4走前
1	1	リリーパレロ (有)キャロットファーム 厩: 堀宣行(美浦) 父: ロードカナロア 母: ヴィートマルシェ(フレンチデピュティ)	牝3/鹿 54.0kg 津村 明秀				
				odds			
1	2	アルチスタ (株)Basic 厩: 畠山吉宏(美浦) 父: エイシンフラッシュ 母: アカンサス(フシキセキ)	牝3/青鹿 51.0kg ▲木幡 育也	2018年4月22日 東京 未勝利 15着 18頭 16番 12番人気 412kg △武蔵雅 52.0kg 1400芝 良 1:24.0 14-14 34,4F ニシノジャガーズ(1.8)	2018年2月3日 東京 新馬 14着 16頭 15番 10番人気 420kg 横山典弘 54.0kg 1800芝 稍重 1:52.3 16-15-15 34,9F ノーブルカリス(2.2)		
				jockey			
2	3	スーパースナズ (有)ビッグレッドファーム 厩: 天間昭一(美浦) 父: ローンズインメイ 母: スーパーウーマン(マーベラスサンデー)	牝3/栗 54.0kg 柴田 大知	2017年10月9日 東京 新馬 4着 17頭 10番 12番人気 454kg 柴田大知 54.0kg 1600芝 良 1:36.5 4-4 34,7F ゴールドギア(0.6)			
2	4	コスモラフェット (有)ビッグレッドファーム 厩: 菅野浩二(美浦) 父: マツリダゴッホ 母: コスモクラッペ(マイネルラヴ)	牝3/黒鹿 56.0kg 宮嶋 北斗	2018年3月11日 中京 未勝利 12着 13頭 11番 10番人気 468kg D.バレルジュ 56.0kg 1800芝 稍重 2:01.2 6-7-10-10 42,4F パキユートハート(6.4)	2018年2月25日 中山 未勝利 13着 16頭 15番 12番人気 476kg 柴田大知 56.0kg 1800芝 良 2:00.9 11-12-11-13 41,3F ライクアロケット(2.8)	2017年6月24日 東京 新馬 8着 16頭 16番 6番人気 436kg 柴田大知 54.0kg 1600芝 良 1:37.4 3-3 36,4F マイネルサイルーン(1.5)	

race info

horse attribute

race histories

Many Types of History

horse

【出走レース】

年月日	場	レース名	距離	馬場	頭数	人気	着順	騎手	負担重量	馬体重	タイム	1着馬(2着馬)
2018.06.24	東京	3歳未勝利	芝1600	重	15	14	8	宮崎 北斗	56.0	464	1:37.2	トーセンクロノス
2018.03.11	中京	3歳未勝利	ダ1800	稍重	13	10	12	D.バルジュー	56.0	468	2:01.2	パキュートハート
2018.02.25	中山	3歳未勝利	ダ1800	良	16	12	13	柴田 大知	56.0	476	2:00.9	ライクアロケット
2017.06.24	東京	2歳新馬	芝1600	良	16	6	8	柴田 大知	54.0	436	1:37.4	マイネルサイルーン

trainer

■ 本年成績 萱野浩二 (カヤノ コウジ)

2018年6月25日現在 [▶ 戻る](#)

1 | 2 | 3 | 4 | 5 | 次の20件

年月日	場	レース名	馬名	距離	馬場	頭数	人気	着順	騎手	負担重量	馬体重	タイム
2018.06.24	函館	津軽海峡特別	ノースランドボーイ	ダ1700	良	14	10	13	勝浦 正樹	57.0	490	1:47.5
2018.06.24	東京	パラダイスS	トウショウドラフタ	芝1400	稍重	10	4	4	田中 勝春	56.0	482	1:22.6
2018.06.24	東京	3歳未勝利	コスモラフェット	芝1600	重	15	14	8	宮崎 北斗	56.0	464	1:37.2
2018.06.24	阪神	皆生特別	クラウンアゲン	芝1200	稍重	14	14	6	川又 賢治	52.0	440	1:09.7
2018.06.23	東京	3歳上500万下	ニシノジャガーズ	芝1400	重	15	4	14	戸崎 圭太	54.0	472	1:26.1
2018.06.23	東京	3歳上500万下	タイムトラベル	ダ1600	稍重	16	16	15	岩部 純二	55.0	454	1:39.7
2018.06.23	東京	3歳未勝利	ソリフロール	芝2400	良	15	6	3	北村 宏司	54.0	446	2:27.3

jockey

■ 本年成績 宮崎 北斗 (ミヤザキ ホクト)

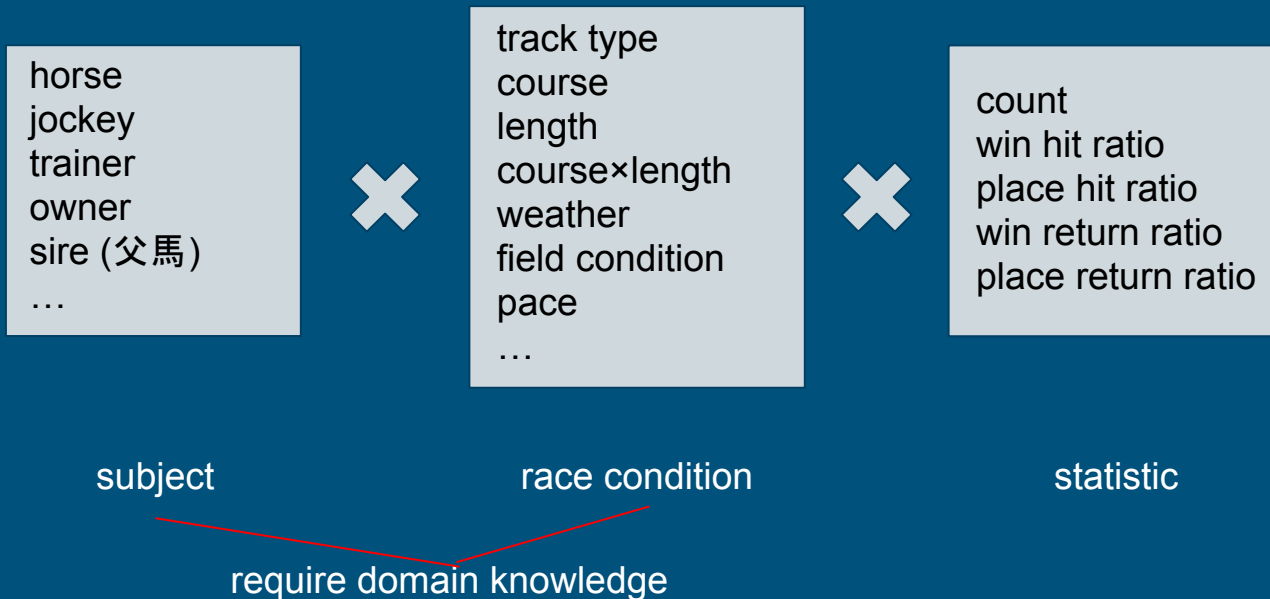
2018年6月25日現在 [▶ 戻る](#)

1 | 2 | 3 | 4 | 5 | 次の20件

年月日	場	レース名	馬名	距離	馬場	頭数	人気	着順	調教師	負担重量	馬体重	タイム
2018.06.24	東京	2歳新馬	クラウンオペラ	芝1800	重	14	14	14	伊藤 伸一	54.0	466	1:59.4
2018.06.24	東京	3歳未勝利	コスモラフェット	芝1600	重	15	14	8	萱野 浩二	56.0	464	1:37.2
2018.06.23	東京	3歳未勝利	エイワファースト	ダ1600	稍重	16	10	9	菊川 正達	54.0	452	1:40.1
2018.06.23	東京	2歳未勝利	マイネルオリエンズ	芝1600	良	14	12	9	高橋 祥泰	54.0	472	1:37.5
2018.06.17	東京	3歳未勝利	ムーンケリー	ダ1600	重	16	9	11	中川 公成	56.0	504	1:40.0
2018.06.17	東京	3歳未勝利	ブラウンファシール	ダ1400	重	16	14	16	小西 一男	54.0	450	1:27.3
2018.06.16	東京	3歳未勝利	ショウナンハイル	芝1800	重	16	11	9	田中 剛	54.0	426	1:50.6
2018.06.16	東京	2歳新馬	ウインブレイヤー	ダ1400	重	16	9	16	宗像 義忠	54.0	470	1:30.1
2018.06.16	東京	3歳未勝利	エリースコル	ダ1400	不良	16	13	12	青木 孝文	54.0	454	1:26.5
2018.06.16	東京	3歳未勝利	ザラストキャンディ	ダ1600	不良	16	13	13	天間 昭一	54.0	458	1:39.5
2018.06.12	川崎	ジュンF賞	マイネルマリボッサ	ダ1500	重	13		12	清水 英克	56.0	461	1:41.2
2018.06.10	東京	3歳未勝利	ラブチュア	ダ1600	良	16	5	8	伊藤 伸一	54.0	430	1:41.5
2018.06.03	東京	3歳未勝利	サンディアロッサ	ダ1300	良	16	15	11	伊藤 伸一	54.0	410	1:21.6
2018.06.03	東京	3歳未勝利	デザートカレン	ダ1400	良	16	11	11	和田 勇介	54.0	428	1:28.3
2018.06.02	東京	3歳上500万下	イエローブレイヴ	ダ1300	良	16	11	9	武市 藤男	54.0	522	1:19.6
2018.05.26	東京	3歳未勝利	ラブチュア	ダ1600	良	16	16	3	伊藤 伸一	54.0	432	1:40.1
2018.05.20	新潟	3歳未勝利	ラグナズルット	芝1200	重	16	6	7	松山 将樹	54.0	442	1:10.9
2018.05.19	新潟	大日岳特別	ワコマタイヨウ	芝1200	重	16	13	15	武市 藤男	55.0	434	1:12.1
2018.05.19	新潟	早苗賞	タイムムーン	芝1800	重	10	9	2	青木 孝文	54.0	458	1:50.6
2018.05.19	新潟	3歳未勝利	ベリータ	芝2000	重	15	6	12	水野 貴広	54.0	438	2:06.4

Automated Achievement Features

- We made more than 1500 achievement features



Smoothing Ratio Features

- Eliminate noise of unpopular subject×condition

$$R_{smooth} = (1 - \exp(-\alpha N)) \cdot R + \exp(-\alpha N) \cdot R_{average}$$



where $\alpha=0.1$

hit ratio=0.254

hit ratio=0.399

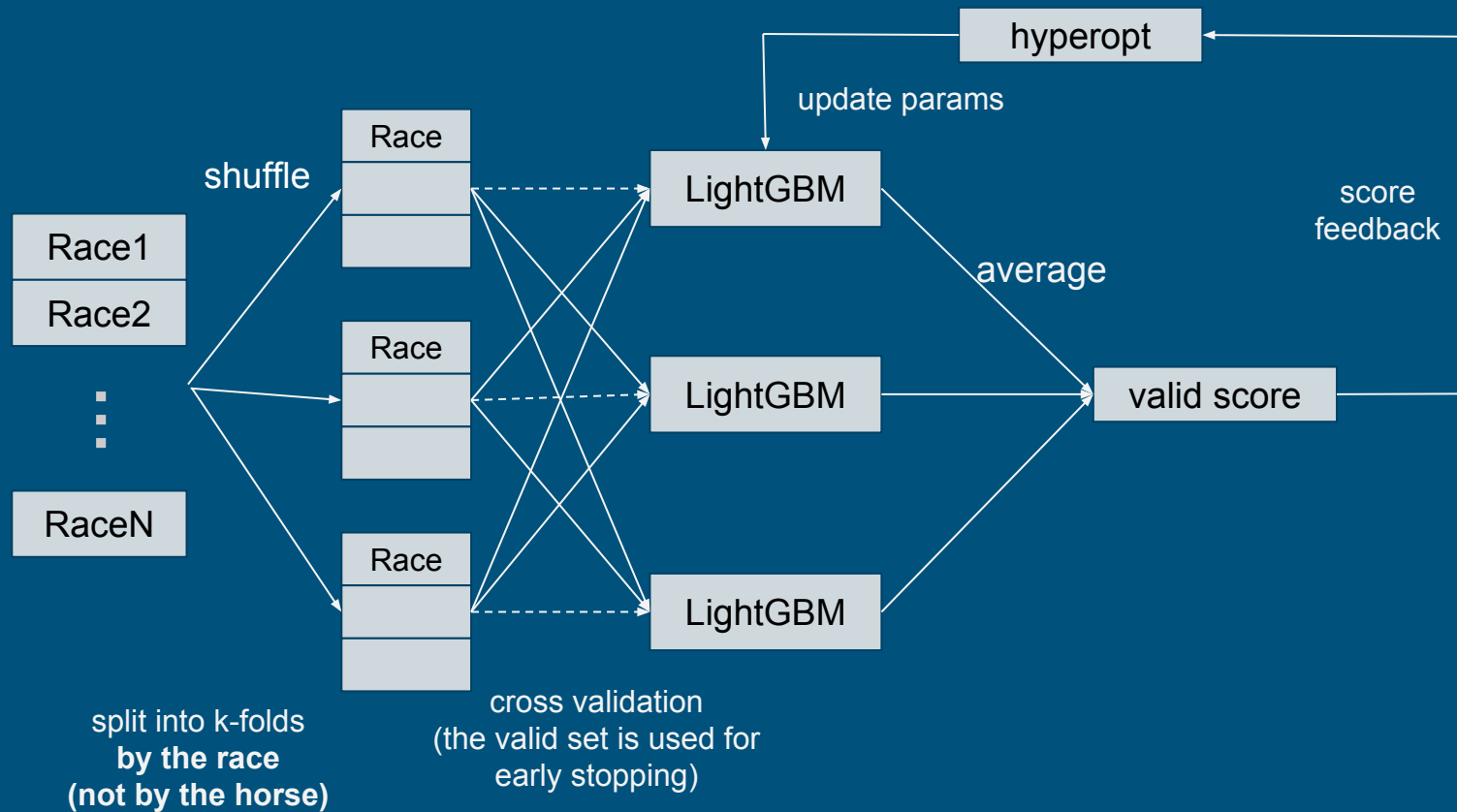
Road to winning

1. Introduction to the system of horse racing
2. Definition of objective
3. Feature engineering
4. Prediction model
5. Evaluation

LightGBM

- One of the most popular frameworks in Kaggle
- Very strong to structured data
- Not affected by scales of features
- Can handle the missing value
- Robust to meaningless features
- We have a committer of LightGBM

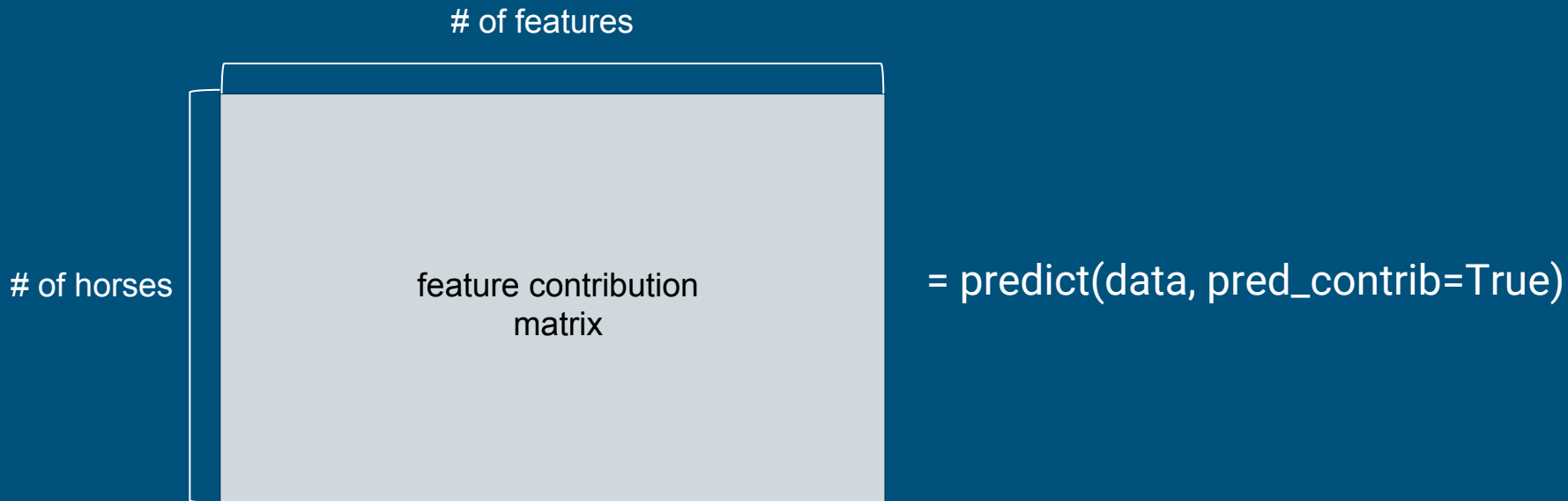
Training Architecture



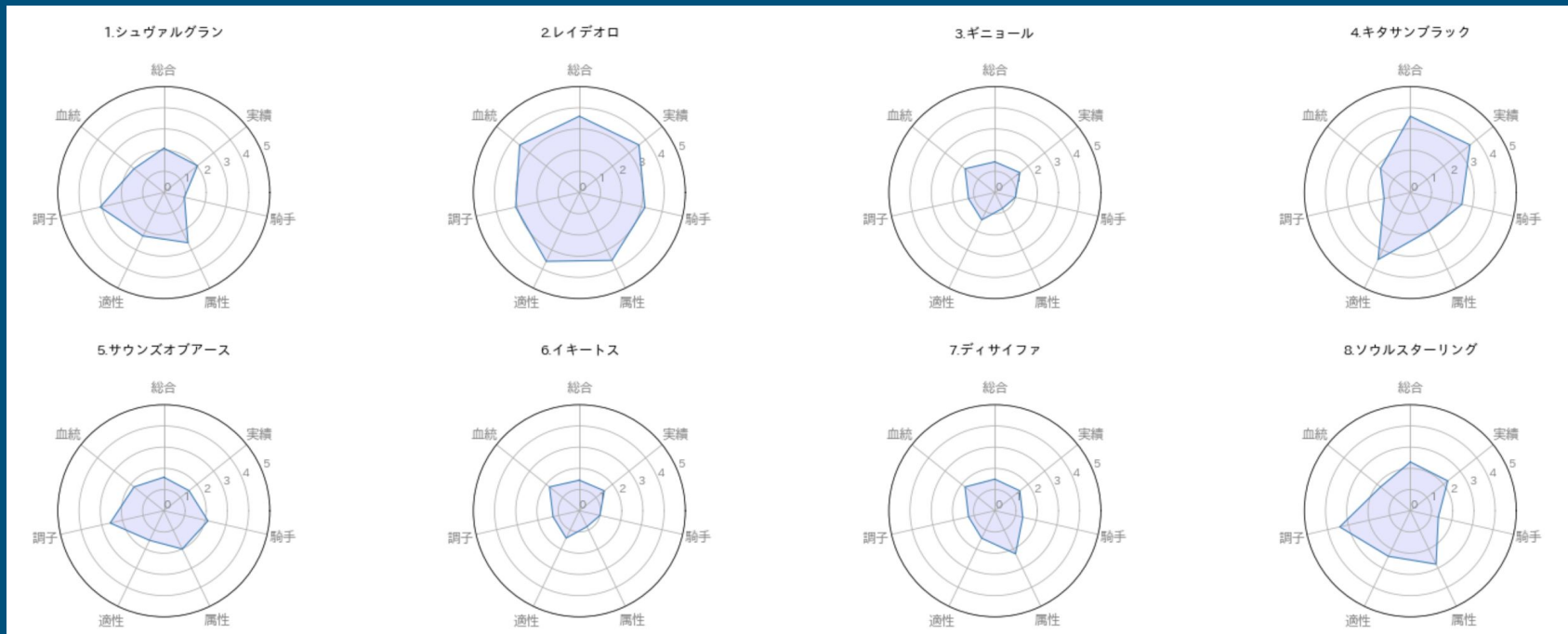
Notice of LightGBM

- Early stopping is important to avoid overfitting
- Dummied features are better than categorical ones
 - Depends on dataset
- Sensitive to `random_state`
 - May result from `subsample/colsample_bytree`?
- `pred_contrib` is useful for feature analysis

Feature Analysis with LightGBM



Feature Analysis with LightGBM



Road to winning

1. Introduction to the system of horse racing
2. Definition of objective
3. Feature engineering
4. Prediction model
5. Evaluation

nDCG

- Used in ranking evaluation
- Higher relevant score (rel_i) should be positioned at higher rank

non-negative, “the higher, the better”

$$DCG = rel_1 + \sum_{i=2}^n \frac{rel_i}{\log_2 i}$$

$$nDCG = \frac{DCG}{idealDCG}$$

The Relevant Score of nDCG

- inverse of order
 - $1/1, 1/2, \dots, 1/N$
 - consider the whole ranking
- prize
 - 15000, 6000, 3800, 2300, 1500, 0, 0, ..., 0
 - consider only top 5
- prize@3
 - 15000, 6000, 3800, 0, 0, 0, 0, ..., 0
- place payoff
 - emphasize the dark horses
- win betting share (单勝支持率)
 - how close to popularity

The Relevant Score of nDCG

all turf races in 2017

strength model1 (standardized velocity)

----- nDCG	nDCG	std
inv_order	0.919	0.052
prize	0.735	0.170
prize@3	0.712	0.189
prize@1	0.460	0.403
payback	0.616	0.178
win_share	0.975	0.035

strength model2 (standardized velocity)

----- nDCG	nDCG	std
inv_order	0.915	0.053
prize	0.732	0.170
prize@3	0.709	0.189
prize@1	0.452	0.403
payback	0.613	0.179
win_share	0.983	0.028

- model1 is more accurate than model2
- model2 is closer to win betting share
- model1 is able to hit the horses that the public cannot predict

Evaluation of Top-N Box Betting

all turf races in 2017

storength model (standardized velocity)

---- Top-1 BOX			
	hit	ret	std (ret)
win	0.328	0.848	1.368
place	0.658	0.878	0.669

---- Top-2 BOX			
	hit	ret	std (ret)
win	0.528	0.870	1.068
place	0.846	0.844	0.501
quinella place	0.308	0.824	1.451
quinella	0.154	0.894	2.625
exacta	0.154	0.883	2.725

---- Top-3 BOX			
	hit	ret	std (ret)
win	0.640	0.825	0.899
place	0.927	0.809	0.440
quinella place	0.515	0.787	1.135
quinella	0.292	0.773	1.588
exacta	0.292	0.762	1.657
trio	0.090	0.840	3.698
trifecta	0.090	0.784	4.165

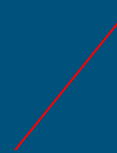
profit model (dark score)

---- Top-1 BOX			
	hit	ret	std (ret)
win	0.136	1.020	3.093
place	0.406	0.886	1.211

---- Top-2 BOX			
	hit	ret	std (ret)
win	0.259	1.015	2.140
place	0.661	0.878	0.877
quinella place	0.112	0.986	3.368
quinella	0.041	1.124	7.056
exacta	0.041	1.221	8.420

---- Top-3 BOX			
	hit	ret	std (ret)
win	0.348	0.916	1.613
place	0.789	0.850	0.764
quinella place	0.293	0.948	2.050
quinella	0.101	0.917	3.611
exacta	0.101	0.944	3.915
trio	0.016	0.906	8.868
trifecta	0.016	0.892	9.204

the lower hit,
the higher return



Summary

- The purpose of horse racing prediction is making a profit
- Design of objective is critical to performance
- Feature engineering requires domain knowledge too
- LightGBM is cool
- nDCG is useful for model evaluation

Future work

- Calculate expectation with predicted probability and real-time odds
- Auto buying system with investment strategies

Enjoy horse racing life!!