# 実践・競馬データサイエンス

## Practical Data Science for Horse Racing

AlphaImpact
NUKUI Shun
@PyCon JP 2018

# Profile

- 貫井 駿 (NUKUI Shun, @heartz2001)
- Speciality : Machine Learning
- Jobs:
  - Fringe81 Co.,Ltd.
    - Ad Tech
    - HR Tech
  - プロ競馬予想家 (Professional tipster)
- Experience of horse racing: 12 years (馬券は二十歳から)
- Favorite horse: ハーツクライ (Heart's Cry)


第１回電脳賞（春）出場当時

# AlphaImpact

- Developing horse racing AI (2016/06~)
- 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/
  - Sell predictions on netkeiba.com ウマい馬券(2017/03~)
    - http://yoso.netkeiba.com/

# Agenda

- 競馬データについて (Data of horse racing)
- 目的変数の設計 (Design of objectives)
- 特徴量作成 (Feature Engineering)
- 予測モデルの学習 (Training of prediction models)
- 予測モデルの評価 (Evaluation of prediction models)

# Agenda

- 競馬データについて (Data of horse racing)
- 目的変数の設計 (Design of objectives)
- 特徴量作成 (Feature engineering)
- 予測モデルの学習 (Training prediction models)
- 予測モデルの評価 (Evaluation of prediction models)

# What is Horse Racing

- 騎手の乗った馬が着順を競い合う熱いスポーツ

  An exciting sport that horses with jockeys compete

- その着順を予想するギャンブル

  A gambling to predict its results

# Why Horse Racing x Data Science?

- 毎週（毎日）解くべき問題となるデータが新しく追加される

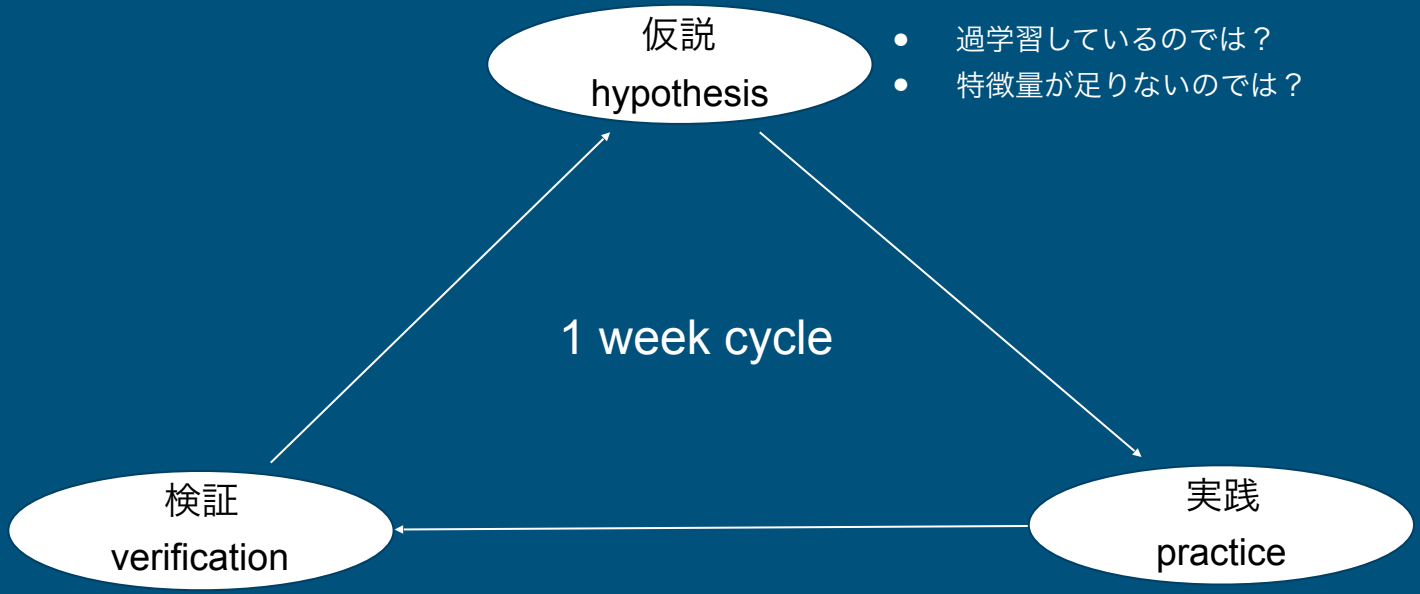
  New data is added weekly or everyday
- 結果が出る過程をリアルタイムに映像で観ることができる

  We can watch the process of output from live streaming
- エンジニアリングのしがいがある豊富なデータ

  Variety of data
- 楽しい！（重要）

  Exciting (important!!)

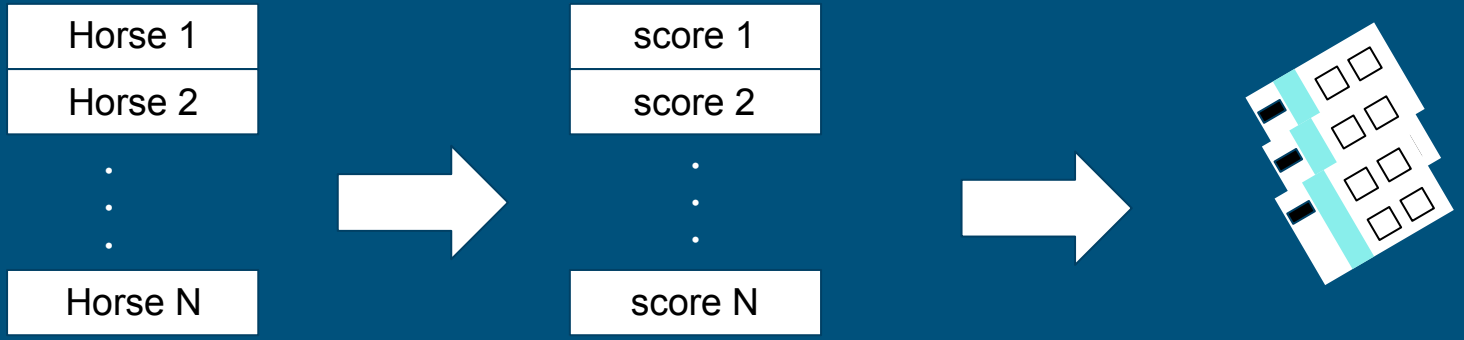
# Hypothesis・Practice・Verification



# Agenda

- 競馬データについて (Data of horse racing)
- 目的変数の設計 (Design of objectives)
- 特徴量作成 (Feature engineering)
- 予測モデルの学習 (Training prediction models)
- 予測モデルの評価 (Evaluation of prediction models)

# Problem Setting




data of N horses in race

score of performance

馬券の買い目

bets

Focus on

# Selection of Objectives

- 「何を解くか」は特徴量作成やモデル選択よりも重要

  What to solve is more critical than creating features or models

- 目的変数に使える基本的な変数 (basic objectives)
  - 着順 (order)
  - 走破タイム (finishing time)
  - 1着からの秒差 (delta time with winner)
  - 賞金 (prize)

# Engineering of Objectives

- レース内標準化 (normalization in race)
  - 走破タイムは距離や馬場の状態に依存する部分が大きい
  - 標準化して環境バイアスを消すことで解きやすくする
- 馬券外の馬のスコアは同じにする (identify unplaced horses)
  - 馬券に絡まない部分の着順を高精度で当てても価値が無い

**良い目的変数を作るためには、ドメイン知識を頭に入れ、問題を正しく理解する**
**Good objectives require the domain knowledge and understanding of the problem**

# Agenda

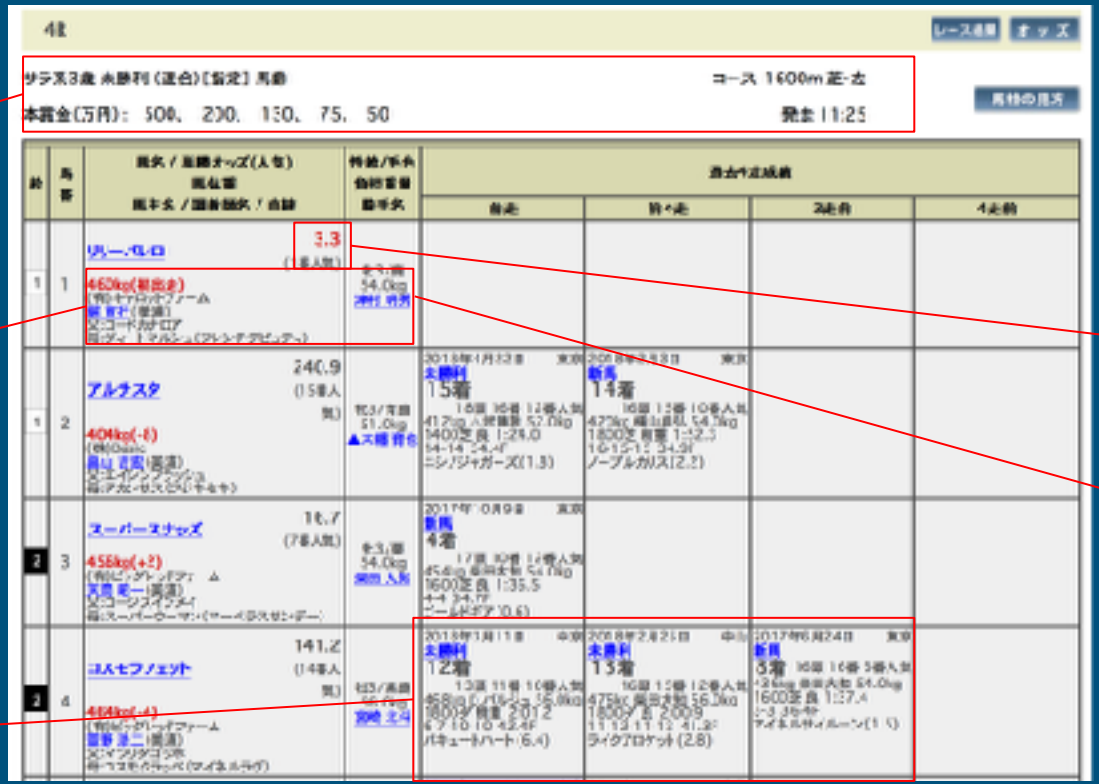
- 競馬データについて (Data of horse racing)
- 目的変数の設計 (Design of objectives)
- 特徴量作成 (Feature engineering)
- 予測モデルの学習 (Training prediction models)
- 予測モデルの評価 (Evaluation of prediction models)

# 馬柱 (Horse Table)

レース情報
race info

馬の属性
attributes of horse

出走履歴
race history



オッズ
Odds

騎手
Jockey

# Flow of Data Processing

- 多数あるテーブルから直接特徴ベクトルを作ろうとすると保守性悪化
  Processing features from collected data directly lowers the maintainability


collected data

feature vectors

# Flow of Data Processing

- 「馬柱」としてまとめることでインターフェースを簡潔に記述

  Horse table data simplify the interfaces of data processing


collected data

horse table data

feature vectors

# 出走履歴データ (Past Race History)

- 過去X走までの情報をそのまま特徴量に加える (Use past X histories as feature)
  - Xが増えると欠損データが増えてスパースになる
  - 例) 過去2走前馬体重、過去3走前走破タイム
- 過去X走の成績を集計する (Summarize past X histories)
  - 次元数はXに依存しない
- 過去Xヶ月の成績を集計する (Summarize histories in the past X monthes)

# Categorical Data

- Categorical data in horse racing
  - 馬名、騎手名、調教師、父馬、母父馬、競馬場、トラック種別
- Categorical data able to be used as numeric
  - 馬番、レース番号、距離
  - 迷ったら数値、カテゴリを別々の特徴量として入れても良

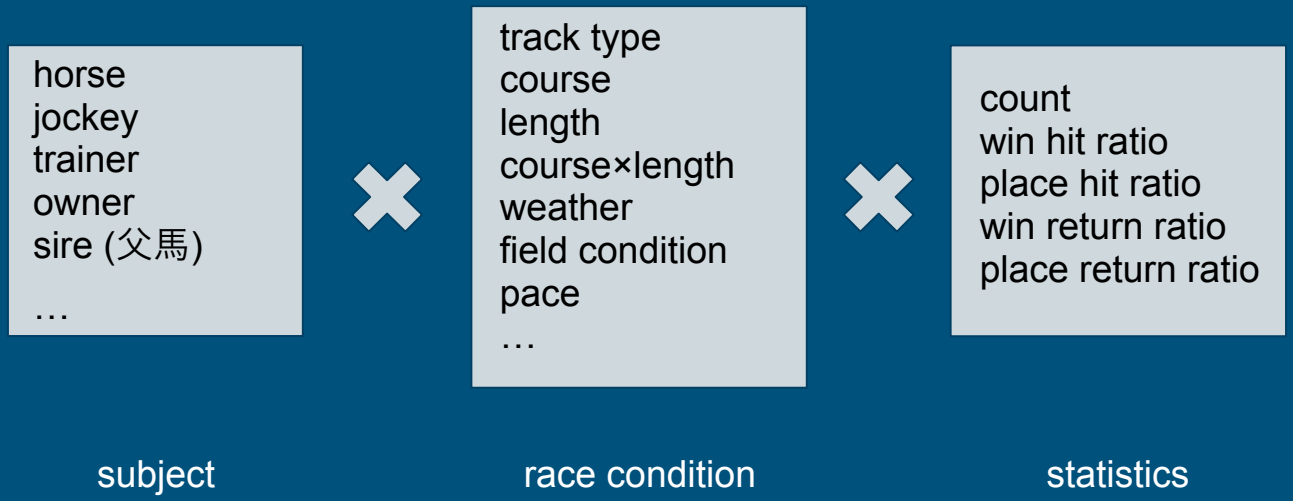
    Can use both of categorical and numeric as a feature

# Encoding of Categorical Data

- One-Hot-Encoding
  - one hot vector
  - 次元数が増えるので、出現回数で足切るなどの工夫
- Target Encoding
  - 過去データにおける該当カテゴリの目的変数の集計値（mean, count, sum, …）
    - 例) 同父馬の平均着順
  - 目的変数に限らず、的中率 (hit ratio)、回収率 (return ratio)などで集計することも可能

# Automated Achievement Features

- We made more than 1500 achievement features



# Smoothing Ratio Features

- 該当カテゴリが少ない場合は、集計値を全体平均に近づける

  When the categories is small, set the aggregate value closer to the overall average
- αはカテゴリごとに最適値を選択

  Select optimal α for each category

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

count=2
hit ratio=0.5

count=100
hit ratio=0.4

hit ratio=0.2

low confidence

A

B

average

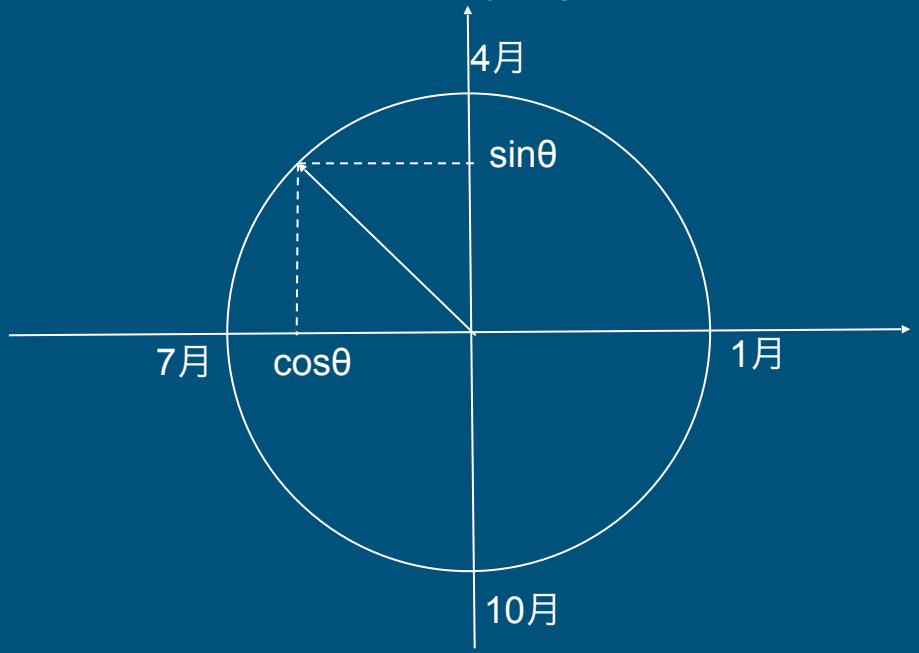
where α=0.1

**hit ratio=0.254**

**hit ratio=0.399**

# 季節特徴量 (Seasonal Features)

- 周期性のあるものは三角関数で表現できる
  Cyclic features can be represented by trigonometric functions



# Agenda

- 競馬データについて (Data of horse racing)
- 目的変数の設計 (Design of objectives)
- 特徴量作成 (Feature engineering)
- 予測モデルの学習 (Training prediction models)
- 予測モデルの評価 (Evaluation of prediction models)

# LightGBM

- 勾配ブースティングの高速・軽量・高精度な実装

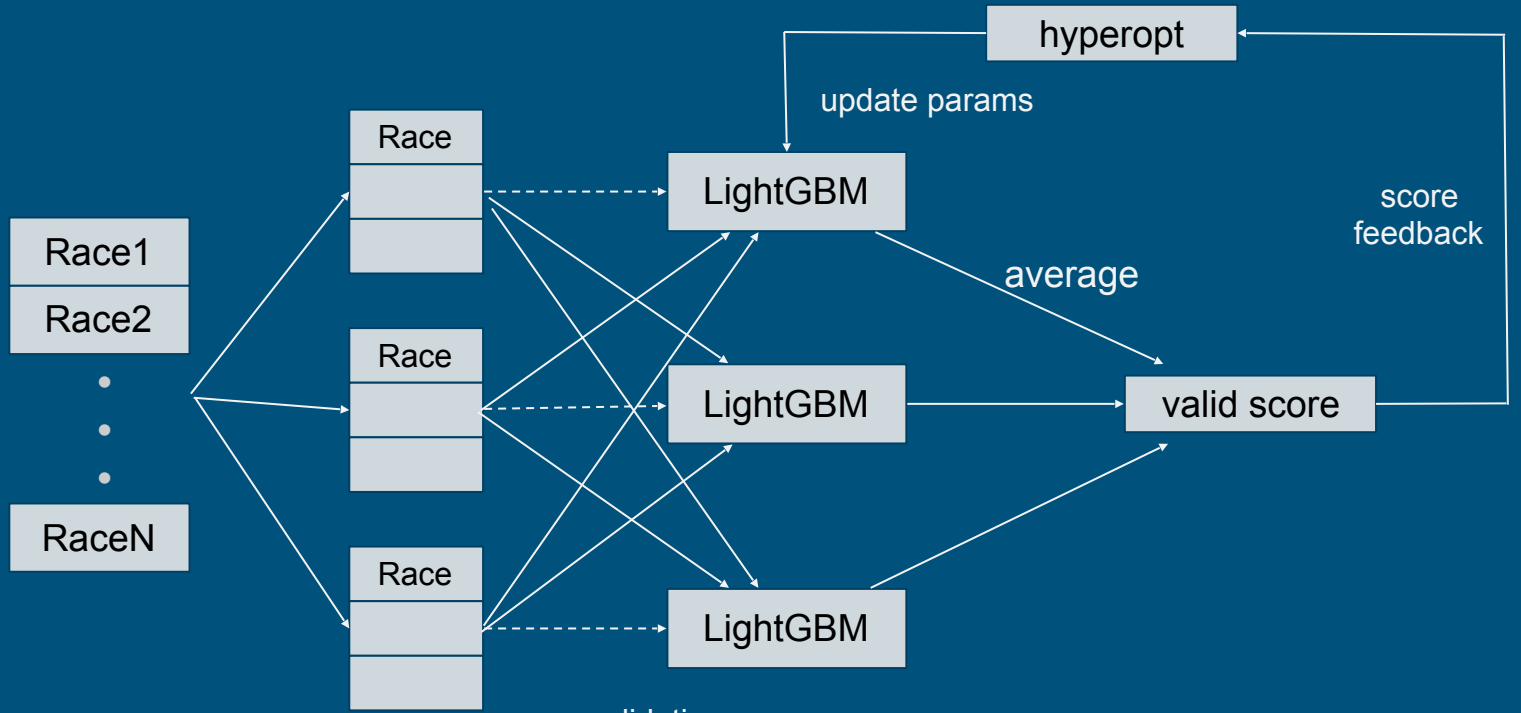
  The fast, light, accurate implementation of gradient boosting
- カテゴリ変数をカテゴリ変数として扱うことができる

  Category variables can be treated as categorical variables
- 欠損値を欠損値として扱うことができる

  Missing values can be treated as missing values
- AlphaImpactのOMOTOはLightGBMのcommitter

  We have a committer of LightGBM

# モデル学習方法 (Model Training)



# hyperopt

- https://github.com/hyperopt/hyperopt
- Tree-structured Parzen Estimator（TPE）
- Grid SearchやRandom Searchに比べてパラメータの探索と活用を効率良く実行できる

  More efficient than grid search or random search

# Efficient Search with GCE × hyperopt

- LightGBMはハイパーパラメータ数が多いのでhyperoptでも探索に時間がかかる

  Since LightGBM has a lot of hyper parameters, it takes much time to search even with hyperopt
- Google Cloud Engine (GCE) のプリエンプティブインスタンスを活用

  The price of the GCE preemptible instance is reasonable, but the instance might be shut down at any time
  - GCEの計算余剰資源を利用しているためお値段約70%オフ
  - ただし、いつインスタンスが落ちるかわからない（最大24時間）
- hyperoptの探索の中間状態をTrialsオブジェクトに保持してepochごとにpkl保存しておけば、途中から探索再開可能

  If you save the intermediate state as Trials object for each epoch, you can resume searching on the way

# Tips of LightGBM Tuning

- カテゴリ変数はダミー化したほうが精度が出ることが多い

  Dummying categorical variables is often more accurate
- early stoppingしないと簡単に過学習してしまうので必ず使う

  Early stopping is necessary to avoid overfitting
- random_stateによって精度が結構変わる

  Accuracy varies considerably with different random_state

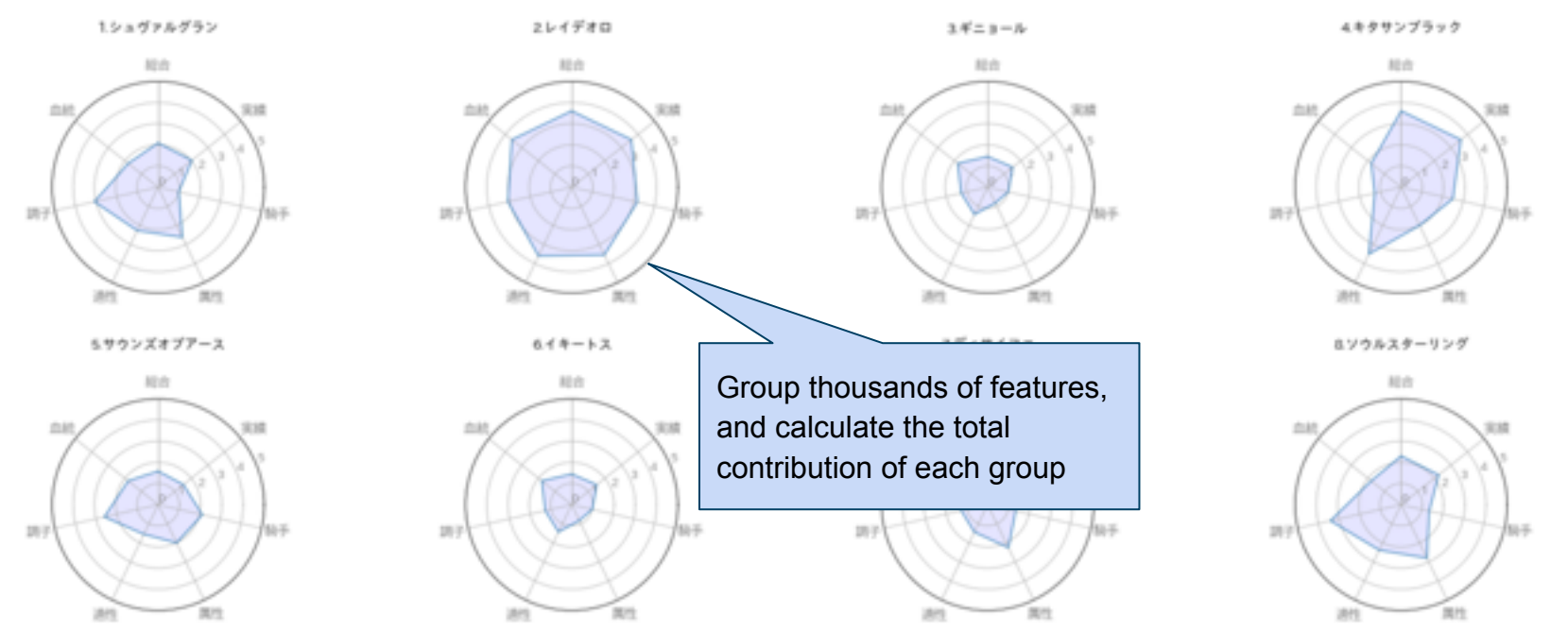
# Feature Analysis with LightGBM

- 特徴量の重要度を見る

  Check feature importance
  - マクロな特徴量分析
- 入力データにおける予測の特徴量の寄与を見る

  Check the contribution of features for predictions
  - ミクロな特徴量分析
  - あるレースにおける予測の根拠が出せる
  - cf. SHAP (SHapley Additive exPlanations)
    - https://github.com/slundberg/shap

# The Contribution of Features for Predictions



# Agenda

- 競馬データについて (Data of horse racing)
- 目的変数の設計 (Design of objectives)
- 特徴量作成 (Feature engineering)
- 予測モデルの学習 (Training prediction models)
- 予測モデルの評価 (Evaluation of prediction models)

# Evaluation Metrics

- ランキング問題でよく使われるnDCGを利用

  Apply nDCG, a well-used metric for ranking problem
- 高い関連度がより上位に予測できていれば大きな値になる（最大値1）

  High relevant score should be positioned at high rank
- 関連度はいろいろな観点をつくるべき

  Should define the variety of relevant scores

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

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

# The Relevant Score of nDCG

- 着順の逆数 (inverse of order)
  - 1/1, 1/2, …, 1/N
  - consider the whole ranking
- 賞金 (prize)
  - 15000,  6000,  3800, 2300, 1500, 0, 0, ..., 0
  - consider only top 5
- 賞金@3 (prize@3)
  - 15000,  6000,  3800, 0, 0, 0, 0, ..., 0
- 複勝払戻し (place payoff)
  - emphasize the dark horses
- 単勝支持率 (win betting share)
  - how close to popularity

# Comparison of Models with nDCG

all turf races in 2017

着順逆数
本賞金
本賞金@3
本賞金@1
複勝払戻し
単勝支持率

(A)

---- 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 |

(B)

---- 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 |

- (A)のほうが(B)よりも的中精度が高い

  (A) is more accurate than (B)
- 単勝支持率(=オッズ)との一致性は(B)のほうが高い

  (B) is closer to win betting share
- (A)は(B)に比べて収益性も高い予測になっている

  (A) is more profitable than (B)

# Evaluation of Top-N Box Betting

all turf races in 2017

Box馬券:

総当たりの組み合わせの馬券

---- 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 | 0.388 | 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 | 0.704 | 4.165 |

的中率、回収率、回収率の偏差

hit ratio, return ratio, std of return ratio

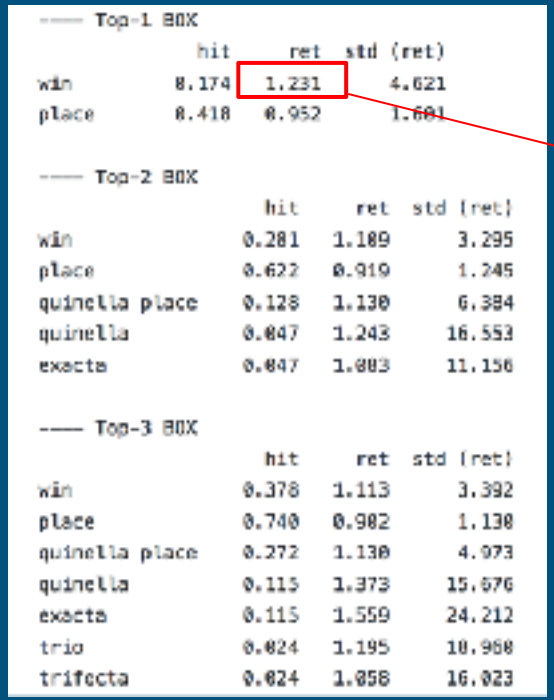
馬券種:

- win: 単勝
- place: 複勝
- quinella place: ワイド
- quinella: 馬連
- exacta: 馬単
- trio: 3連複
- trifecta: 3連単

# Evaluation of Top-N Box Betting

all turf races in 2017

As a result of the effort of objective and feature engineering...



---- Top-1 BOX

| | hit | ret | std (ret) |
|---|---|---|---|
| win | 0.174 | 1.231 | 4.621 |
| place | 0.418 | 0.952 | 1.601 |

---- Top-2 BOX

| | hit | ret | std (ret) |
|---|---|---|---|
| win | 0.281 | 1.109 | 3.295 |
| place | 0.622 | 0.919 | 1.245 |
| quinella place | 0.128 | 1.130 | 6.384 |
| quinella | 0.047 | 1.243 | 16.553 |
| exacta | 0.047 | 1.003 | 11.156 |

---- Top-3 BOX

| | hit | ret | std (ret) |
|---|---|---|---|
| win | 0.378 | 1.113 | 3.392 |
| place | 0.740 | 0.902 | 1.130 |
| quinella place | 0.272 | 1.130 | 4.073 |
| quinella | 0.115 | 1.373 | 15.676 |
| exacta | 0.115 | 1.559 | 24.212 |
| trio | 0.024 | 1.195 | 10.988 |
| trifecta | 0.024 | 1.058 | 16.023 |

Return ratio of win betting is 123%!!

# 定性評価 (Qualitative Evaluation)

- 評価データのうち代表的な数レースをピックアップして予測を目で見る

  See the predictions for a few of representative races
- 概ねまともな予測になっているかどうか

  Check if the predictions make sense
- 過信は過適合につながるので参考程度に

  Overconfidence will lead over fitting
- **実際の予測を見るとテンションが上がりモチベーションUP**

  **Seeing the actual predictions lifts motivation**

# Summary

- 競馬は機械学習のテーマして最高

  Horse racing is supreme as a theme of machine learning
- 目的変数は問題設定にあわせてエンジニアリングする

  Objective variables should be engineered to fit the problem setting
- 特徴量エンジニアリングは精度向上するために必須

  Feature engineering is required for improving the accuracy
- LightGBMは競馬予測に非常に有効

  LightGBM is very effective for horse racing prediction
- モデル性能は定量と定性の両方で評価する

  Model performance is evaluated by both quantitative and qualitative

# Thank you