kaggle days

# Beyond Feature Engineering and HPO

*Jean-François Puget, IBM*

LVMH Christian Dior COUTURE LOUIS VUITTON SEPHORA LOGICIA kaggle

# Jean-François Puget is CPMP

kaggle days

**CPMP**

Machine Learning

France

Joined 6 years ago · last seen in the past day

Followers 1429

Competitions Grandmaster

Home | Competitions (24) | Kernels (21) | Discussion (3,932) | Datasets | ••• | Edit Profile

| Competitions Grandmaster | | | Kernels Master | | | Discussion Grandmaster | | |
|---|---|---|---|---|---|---|---|---|
| Current Rank | Highest Rank | | Current Rank | Highest Rank | | Rank | | |
| 33 of 96,039 | 23 | | 63 of 84,561 | 12 | | 1 of 80,725 | | |
| 6 | 9 | 2 | 4 | 6 | 6 | 110 | 166 | 1581 |

LVMH Christian Dior COUTURE LOUIS VUITTON SEPHORA LOGICIEL kaggle

# What not to do

- Team without checking facts
  - I finished 11th in Toxic Comment competition, which yields a gold, then got removed because a team mate used several accounts before teaming.

    (Faulty team mate was not giba ;) )
  - It cost me a gold and about 15 ranks in competition ranking.

- Break rules
  - Using more than one account
  - Communicating code/data privately outside team

# My worflow

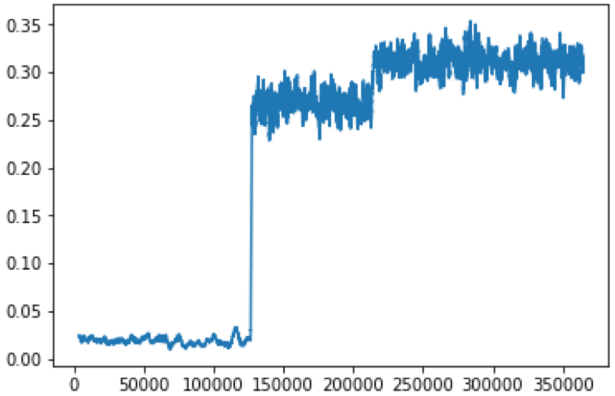
- EDA
- Baseline submission
- CV setting
- Feature engineering
- HPO
- Ensembling

LVMH Christian Dior COUTURE LOUIS VUITTON SEPHORA LOGIC

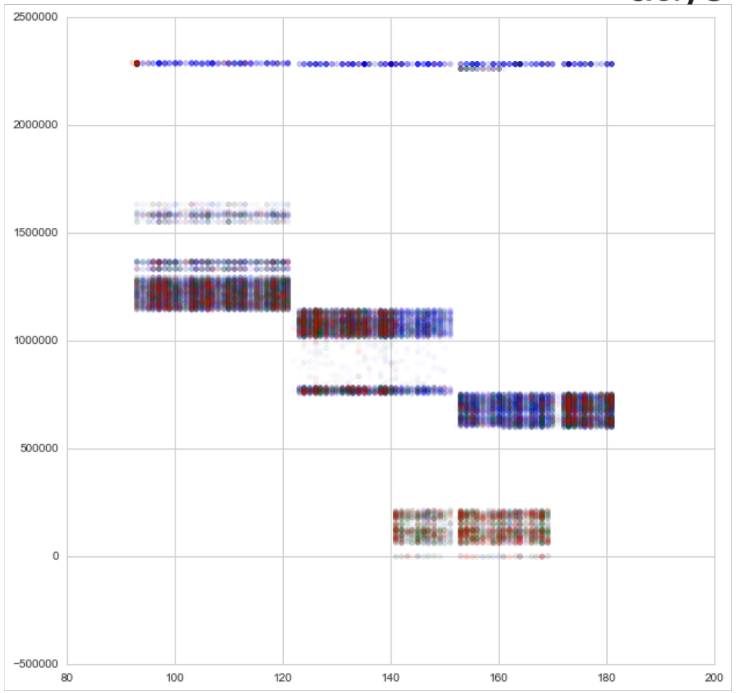
# EDA

Goal is not to produce nice graphics and basic stats to win kernel votes

Goal is to uncover patterns that can lead to new features



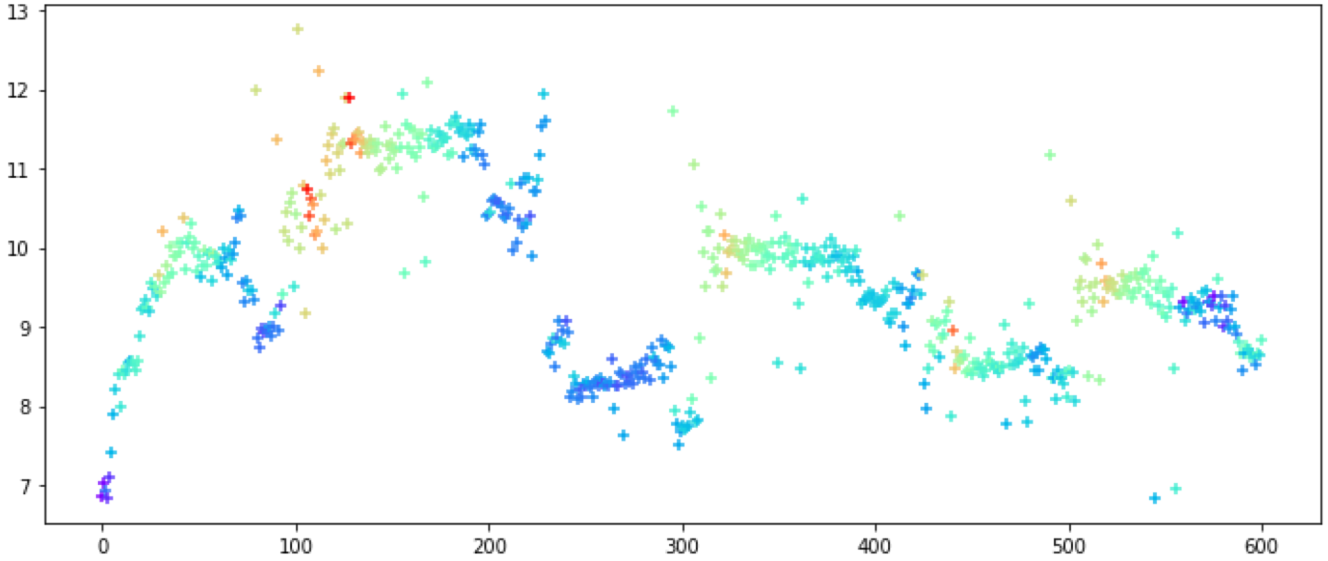

IP download rate in Talking Data




Leak in 2sigma appt rental

LVMH Christian Dior COUTURE LOUIS VUITTON SEPHORA LOGICNI kaggle

# EDA : display target, frequency




Color is target,
x is the feature,
y is frequency

kaggle days

LVMH Christian Dior COUTURE LOUIS VUITTON SEPHORA LOGICIEL kaggle

# Baseline submission

Goal is to have complete code to produce a submission quickly

Build a simple model (e.g. xgboost with raw features)

Then progressively try more complex models (NN, stacking)  and try adding engineered features

# Cross validation setting

Goal: be able to know if a new model is going to improve our private LB score

Wrong way: submit, and see it the public LB score improves

- In Porto Seguro people who  used best public kernel output moved from 100ish public to 700ish private - I moved from 1178 to 29 in that competition

Right Way:

- Use the train data to validate model or feature engineering
- K-fold cross validation (k between 3 to 10)
- Time based cross validation if time dependent data
  - Train on all train periods except last, and predict on last
  - Then retrain on all data and submit

LVMH Christian Dior COUTURE LOUIS VUITTON SEPHORA LOGICNI kaggle

# CV

Which model looks best?

| Train score: | CV score: | LB score |
| --- | --- | --- |
| | Model 1 0.98040 | |
| | Model 2 0.97937 | |

LVMH Christian Dior COUTURE LOUIS VUITTON SEPHORA LOGICNI kaggle

# CV

Which model looks best?

| Train score: | CV score: | LB score |
| --- | --- | --- |
| | Model 1 0.98040 | 0.9675 |
| | Model 2 0.97937 | 0.9694 |

LVMH Christian Dior COUTURE LOUIS VUITTON SEPHORA LOGICAI kaggle

# CV

Which model looks best?

| Train score: | CV score: | LB score |
|---|---|---|
| 0.99004 | Model 1 0.98040 | 0.9675 |
| 0.98313 | Model 2 0.97937 | 0.9694 |

Always look at the gap between train and val score. A large gap is an indication of overfiting.

When you add a feature, check that it does not widden the train/CV gap a lot.

LVMH Christian Dior COUTURE LOUIS VUITTON SEPHORA LOGICNI kaggle

# Feature engineering: try a lot and fail fast

- Try usual suspects
  - Counts
  - Target encoding
  - Clustering
  - Sin(x), cos(x) for periodic features hours, minutes, days
  - Lag variables for time series
  - No missing value imputation, no ohe

https://www.slideshare.net/HJvanVeen/feature-engineering-72376750

- *Try to understand the underlying business problem*

*More on this later*

# HPO

Tune algorithm/model parameters

Impossible without reliable validation setting

Don’t overtune your parameters: do it once, maybe twice in a competition, no more.

For XGBoost/LightGBM:

- Start with subsample=0.7, leave other values to default
- Play with min_child_weight: increase it if train/val gap is large
- Then tune max_depth or number_of_leaves
- Add regularization if LB score is way below CV

LVMH Christian Dior COUTURE LOUIS VUITTON SEPHORA LOGICNI kaggle

# Ensembling

Don’t start too early in the competition

A great model is better than an ensemble of weak models

see Talking Data
@bestfitting says it too ;)

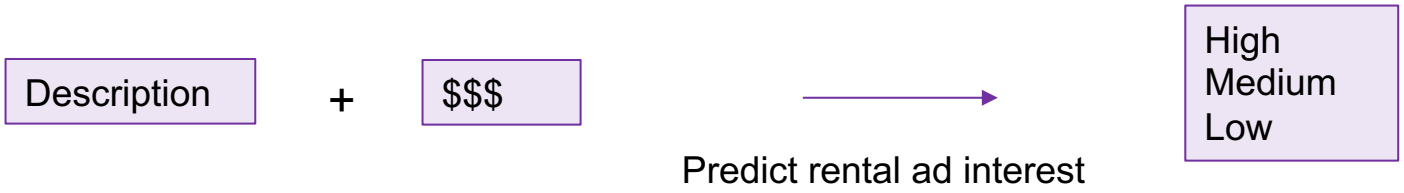
Use same folds for all models

Use out of fold predictions as feature for second level of models (stacking)

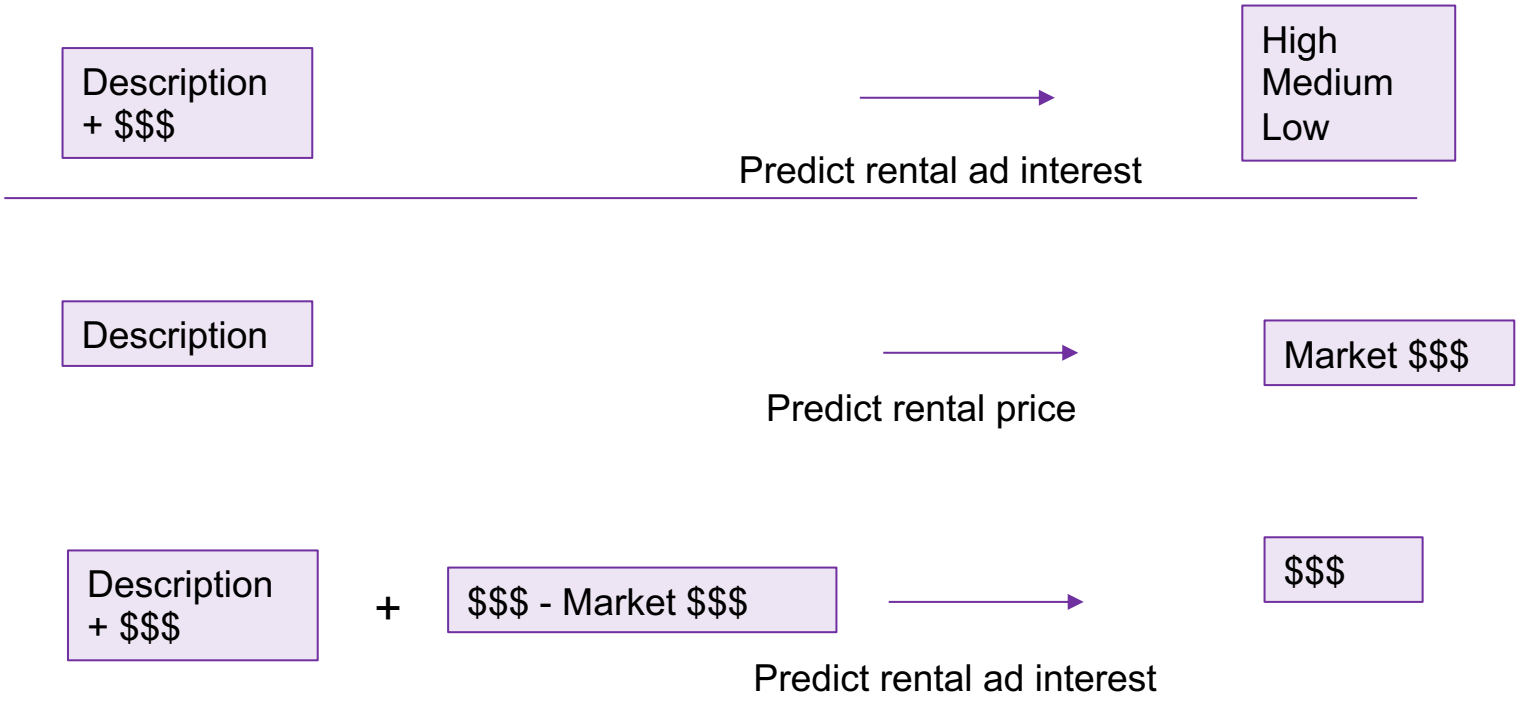
- Gap between CV and LB increases as there is some overfit

LVMH Christian Dior COUTURE LOUIS VUITTON SEPHORA LOGICIEL kaggle

# Business Problem: 2Sigma Appt Rental



# Business Problem: 2Sigma Appt Rental




LVMH Christian Dior COUTURE LOUIS VUITTON SEPHORA LOGICNI kaggle

# Business Problem: Talking Data

```
train.head()
```

| | ip | app | device | os | channel | click_time | is_attributed | click_id |
|---|---|---|---|---|---|---|---|---|
| **0** | 83230 | 3 | 1 | 13 | 379 | 2017-11-06 14:32:21 | 0 | 0 |
| **1** | 17357 | 3 | 1 | 19 | 379 | 2017-11-06 14:33:34 | 0 | 0 |
| **2** | 35810 | 3 | 1 | 13 | 379 | 2017-11-06 14:34:12 | 0 | 0 |
| **3** | 45745 | 14 | 1 | 13 | 478 | 2017-11-06 14:34:52 | 0 | 0 |
| **4** | 161007 | 3 | 1 | 13 | 379 | 2017-11-06 14:35:08 | 0 | 0 |

Target is is_attributed: the app has been downloaded following that click

Very few features

70 M clicks a day, 4 days train, 1 day test

350M rows.

# Business Problem: Talking Data

```
train.head()
```

| | ip | app | device | os | channel | click_time | is_attributed | click_id |
|---|---|---|---|---|---|---|---|---|
| **0** | 83230 | 3 | 1 | 13 | 379 | 2017-11-06 14:32:21 | 0 | 0 |
| **1** | 17357 | 3 | 1 | 19 | 379 | 2017-11-06 14:33:34 | 0 | 0 |
| **2** | 35810 | 3 | 1 | 13 | 379 | 2017-11-06 14:34:12 | 0 | 0 |
| **3** | 45745 | 14 | 1 | 13 | 478 | 2017-11-06 14:34:52 | 0 | 0 |
| **4** | 161007 | 3 | 1 | 13 | 379 | 2017-11-06 14:35:08 | 0 | 0 |

Target is is_attributed: the app has been downloaded following that click

*Once I downloaded an app then I no longer click on ads for that app…*

Killer feature: *Time to next click for same ad from same device*

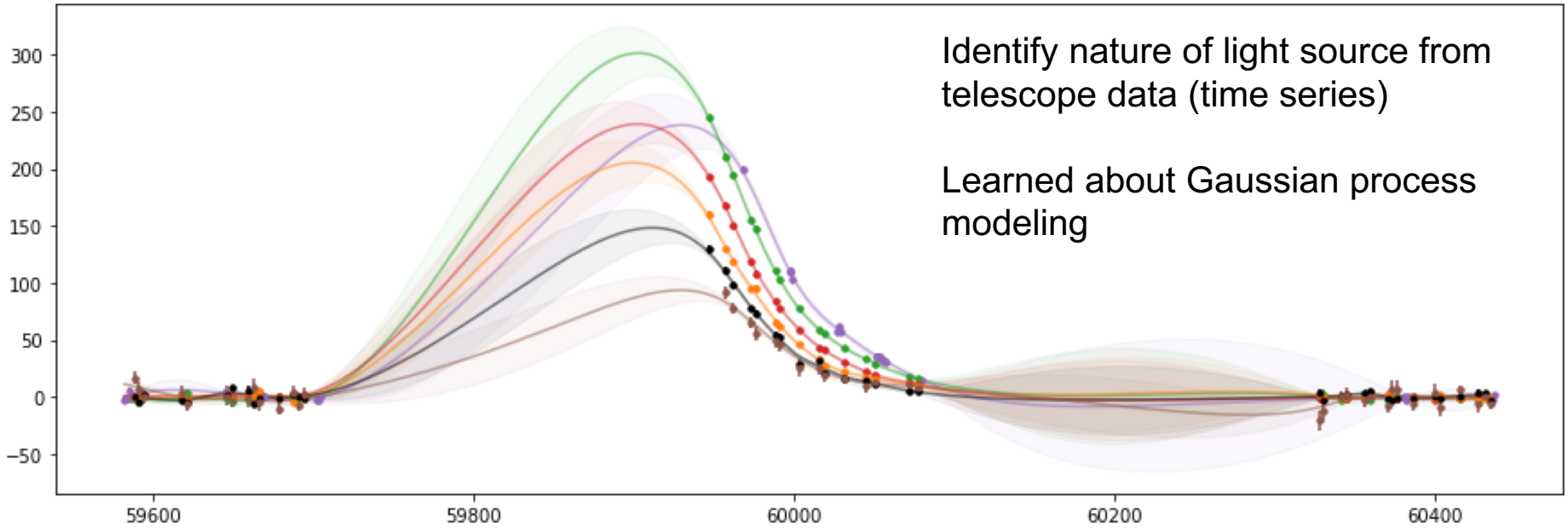
# Science Problem: PLasticc

Identify nature of light source from telescope data (time series)

Open classification:
More classes in test than in train!

LVMH Christian Dior COUTURE LOUIS VUITTON SEPHORA LOGICAL kaggle

# Science Problem: PLasticc




Identify nature of light source from telescope data (time series)

Learned about Gaussian process modeling

LVMH Christian Dior COUTURE LOUIS VUITTON SEPHORA LOGICIA kaggle

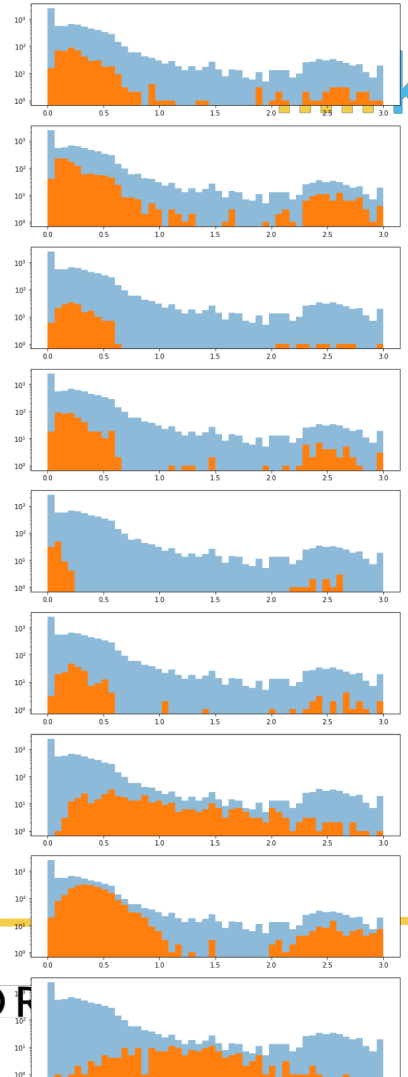
# Science Problem: PLasticc

Redshift is a fundamental property of universe

Related to distance

Universe is the same in every direction and every distance,

*Yet redshift is discriminating*



kaggle days

LVMH Christian Dior COUTURE LOUIS VUITTON SEPHORA kaggle

# Science Problem: PLasticc

Redshift is a fundamental property of universe

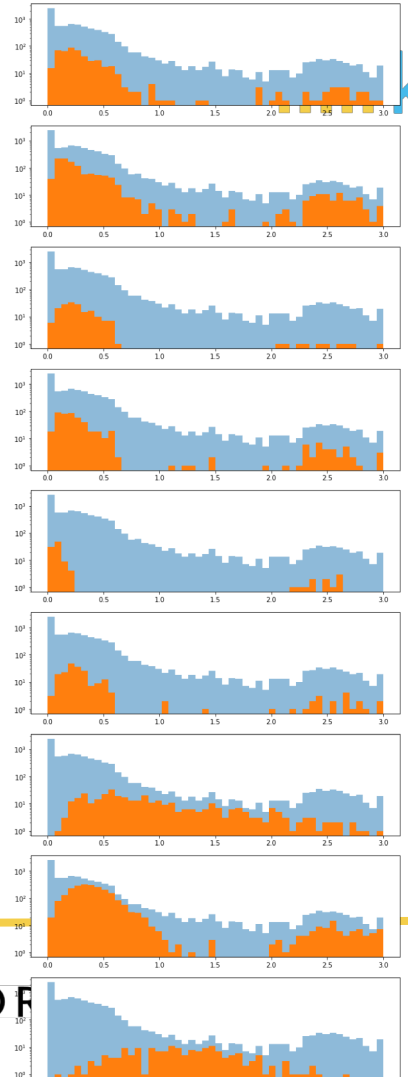
Related to distance

Universe is the same in every direction and every distance,
*Yet redshift is discriminating*

Me (6th): remove redshift

Ahmet Erdem (5th) weight train samples by redshift frequency

Mike Kim (2nd): undo redshift



kaggle days

LVMH Christian Dior COUTURE LOUIS VUITTON SEPHOR

kaggle

# Learn from deep learning

Data augmentation

Made a big difference in Plasticc for top teams

Us NN to build features

Denoising autoencodeers (Jahrer winning solution in Porto Seguro)

Factorization of frequeny matrices (My solution in Talking data)

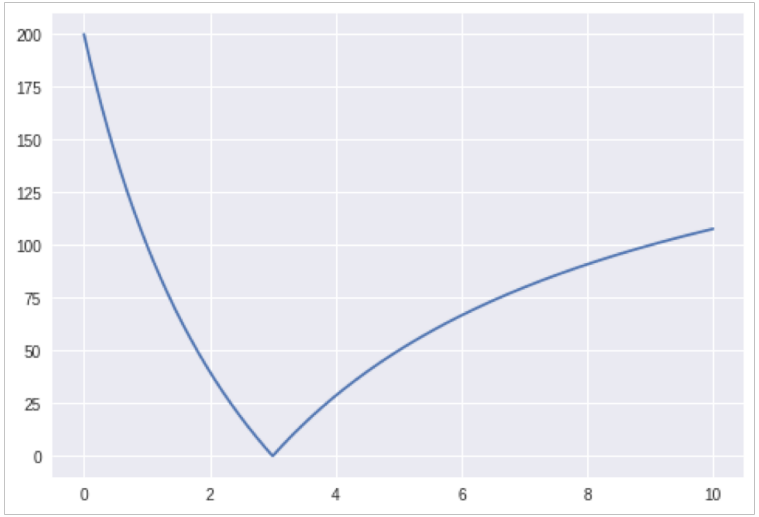
# Target Engineering: match metric

$$NWRMSLE = \sqrt{\frac{\sum_{i=1}^{n} w_i \left(\ln(\hat{y}_i + 1) - \ln(y_i + 1)\right)^2}{\sum_{i=1}^{n} w_i}}$$

Replace target by log1p(target):

yields weighted rmse, natively supported in most algorithms

# Target Engineering: match metric




SMAPE (Web Traffic Forecasting)

Becomes very close to MAE
when using log of target

LVMH
Christian Dior COUTURE
LOUIS VUITTON
SEPHORA
LOGICNI
kaggle

# Target Engineering: Can be domain specific

```
train.head()
```

| | ip | app | device | os | channel | click_time | is_attributed | click_id |
|---|---|---|---|---|---|---|---|---|
| **0** | 83230 | 3 | 1 | 13 | 379 | 2017-11-06 14:32:21 | 0 | 0 |
| **1** | 17357 | 3 | 1 | 19 | 379 | 2017-11-06 14:33:34 | 0 | 0 |
| **2** | 35810 | 3 | 1 | 13 | 379 | 2017-11-06 14:34:12 | 0 | 0 |
| **3** | 45745 | 14 | 1 | 13 | 478 | 2017-11-06 14:34:52 | 0 | 0 |
| **4** | 161007 | 3 | 1 | 13 | 379 | 2017-11-06 14:35:08 | 0 | 0 |

Many occurrences of several clicks with same features and different target !!!

*Once I downloaded an app then I no longer click on ads for that app…*

*Sort the clicks by target as well*

+ 0.004 on LB, and many ranks ☺

# Bottom line

Understand the business problem

Understand how models are evaluated

Act accordingly

Have fun!