Recruit Restaurant Visitor Forecasting

Churkin Nikita
Gleb Filatov

From the description

We are provided a time-series forecasting problem centered around restaurant visitors

From the description

We are provided a time-series forecasting problem centered around restaurant visitors

Data comes from two separate sites:

Hot Pepper Gourmet: similar to Yelp, here users can search restaurants and also make a reservation online

AirREGI/Restaurant Board: similar to Square, a reservation control and cash register system

From the description

We are provided a time-series forecasting problem centered around restaurant visitors

Data comes from two separate sites:

Hot Pepper Gourmet: similar to Yelp, here users can search restaurants and also make a reservation online

AirREGI/Restaurant Board: similar to Square, a reservation control and cash register system

We must use the reservations, visits, and other information from these sites to forecast future restaurant visitor totals on a given date

Competition results

1. Massive shake-up
2. Top ODS teams:

Andrey Filimonov (24th),

Sergey Novik (31th),

N & G (35th)

22	▲ 18	买两张彩票		0.511	37	2mo
23	▲ 37	To Train Them Is My Cause		0.511	38	2mo
24	▲ 25	slonoslon		0.512	69	2mo
25	▲ 671	ZABURO		0.512	29	2mo
26	—	Li-Der		0.512	27	2mo
27	▼ 26	Oncu Kayalar		0.512	144	2mo
28	▲ 55	Don Lima		0.513	18	2mo
29	▲ 33	ShengZhao		0.513	87	2mo
30	▲ 20	ShinnaMashiro		0.513	16	2mo
31	▲ 41	snovik		0.513	167	2mo
32	▼ 9	winner winner chicken dinner		0.513	98	2mo
33	▲ 12	Ben Ogorek		0.513	50	2mo
34	▼ 14	YECK		0.513	246	2mo
35	▲ 36	N & G		0.513	89	2mo
36	▼ 11	YiTang Overfitting		0.513	42	2mo
37	▲ 28	Forecaster		0.513	20	2mo
38	▲ 923	Gen X		0.513	79	2mo

Competition details

- 8 data files: AIR/HPG restaurants info,

reservations, visits, AIR-HPG mapping, holidays info

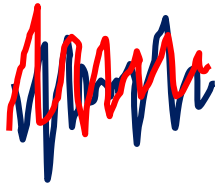
- RMSLE metric
- ~2K participants

Competition details

- AIR/HPG restaurants info files contain: area name, genre name, latitude and longitude of restaurants
- AIR/HPG reservation files contain: restaurant id, visit datetime, reserve datetime and reserve visitors.

Competition details

Multiple time series



Multiple time series

- 829 restaurants in AIR system – train (train is AIR exclusive)
- 821 restaurants in AIR system – test (test is AIR exclusive)
- 100% intersection

- 4.7K restaurants in HPG system
- But no target for HPG restaurants!

Train data – 2016-01-01 -> 2017-04-22

Test data – 2017-04-23 -> 2017-05-31

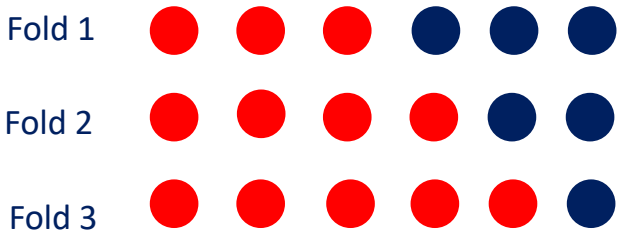
Competition details

Multiple time series

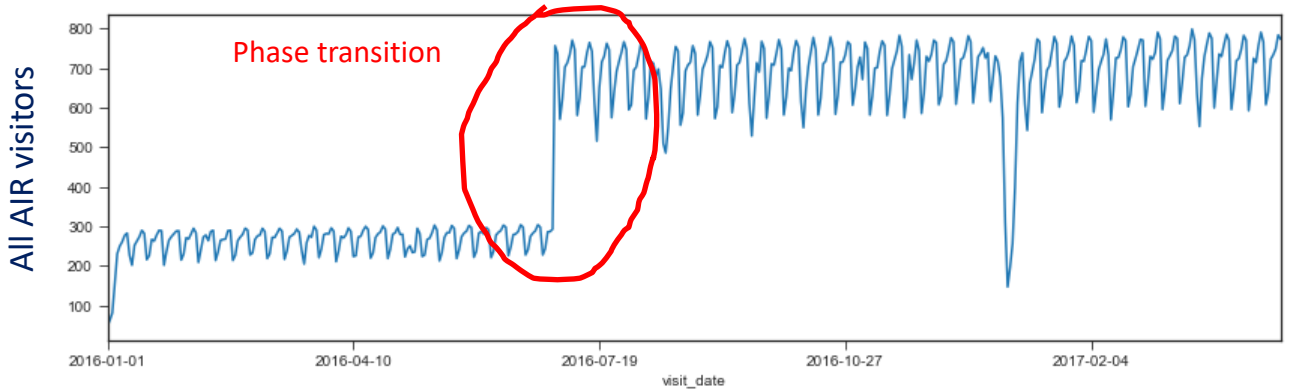
Different validation approaches

Different validation approaches

- Vanilla KFold (1st place of public LB)
- Holdout set (April 2017 for example)
- Rolling validation:

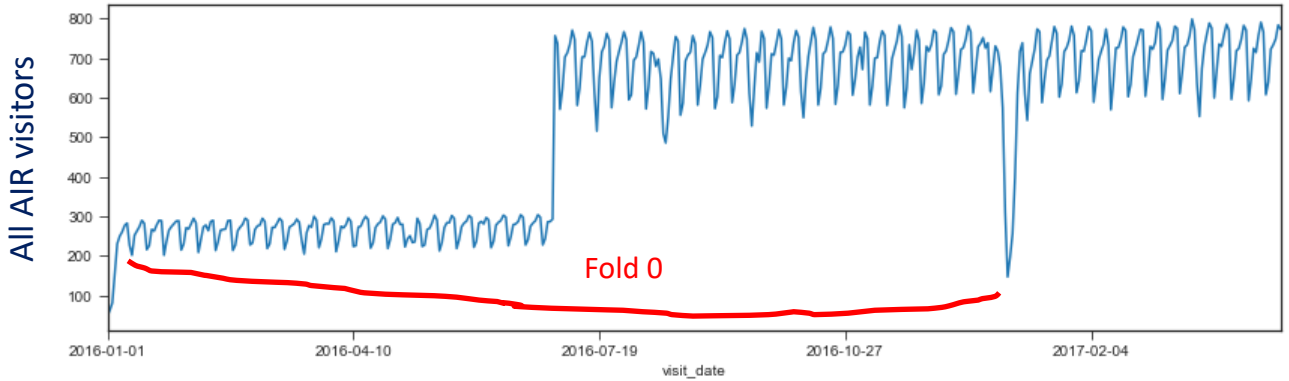


Phase transition: new restaurants in the AIR system



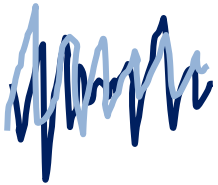
Our validation scheme

Rolling validation (month) with initial data as Fold 0



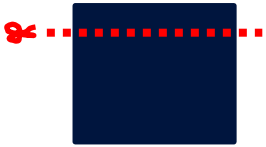
Competition details

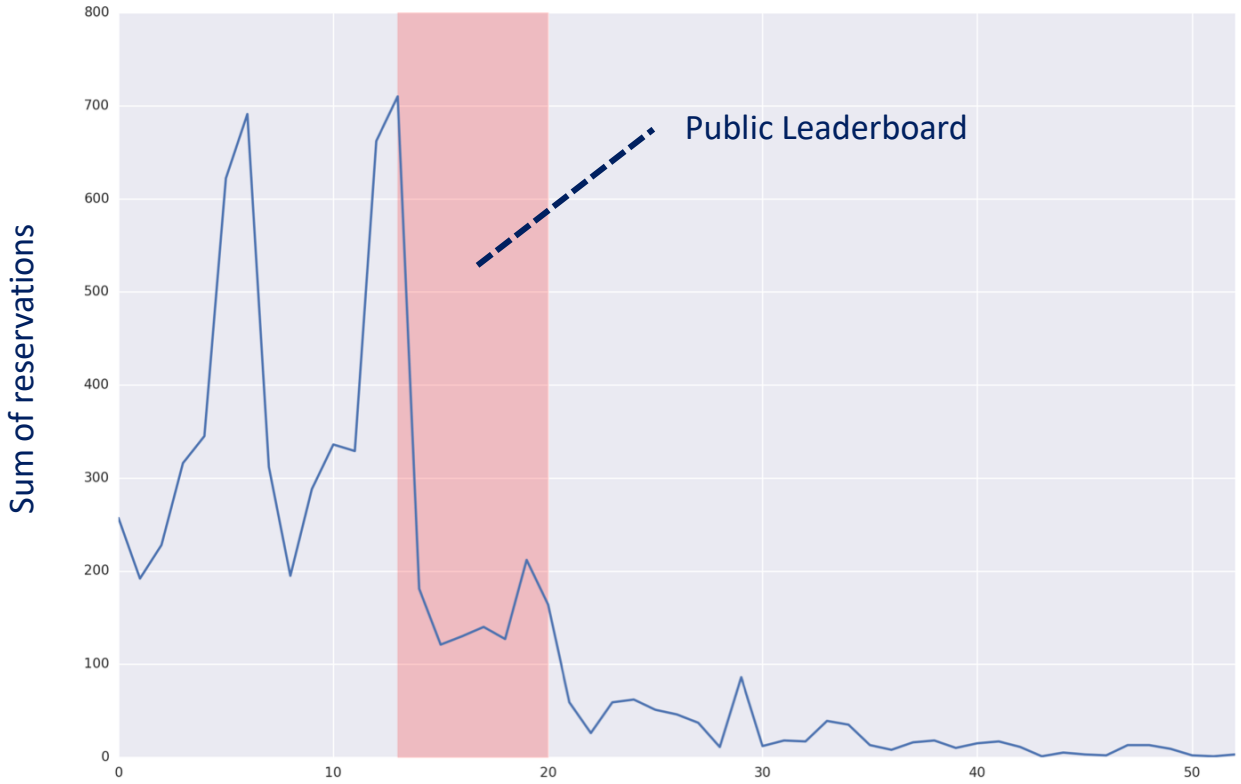
Multiple time series



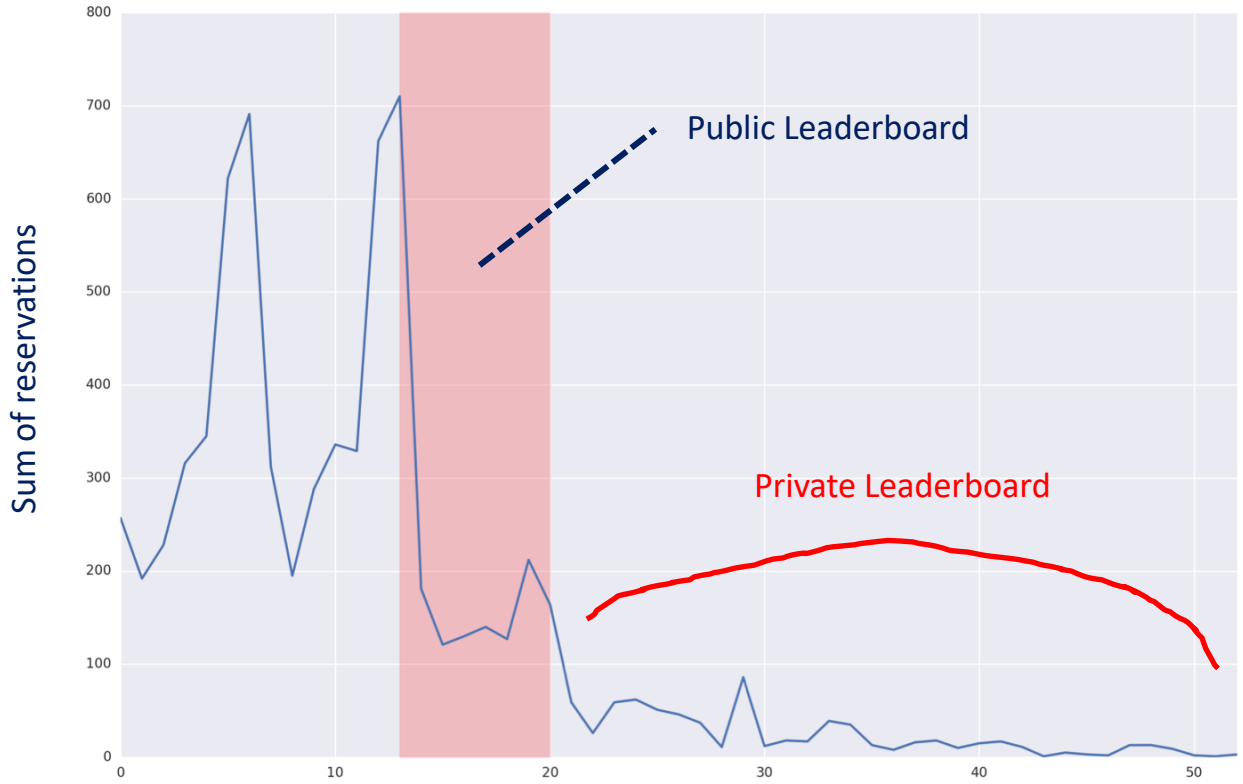
Different validation approaches

Whimsical data





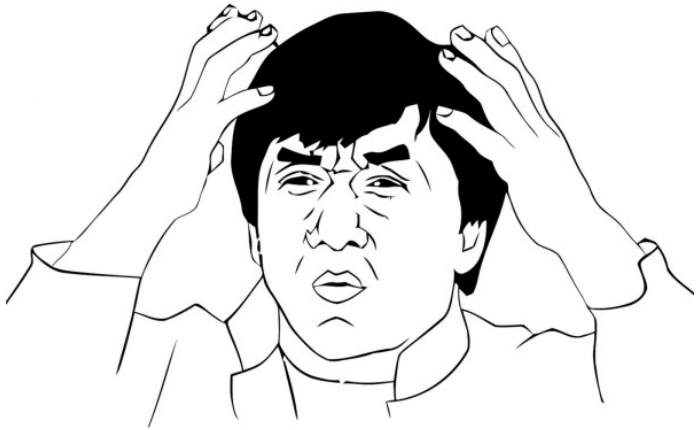
Reservations – present on train and public leaderboard, but fall off quickly further in time



Reservations – present on train and public leaderboard, but fall off quickly further in time

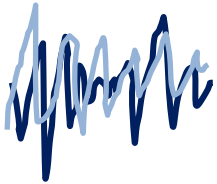
VERY good on local validation/public leaderboard, but will be much worse on private leaderboard

We have the same drop in HPG reservations!



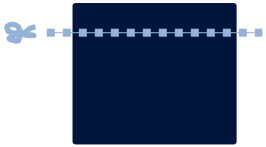
Competition details

Multiple time series



Different validation approaches

Whimsical data



“External” whether datasets

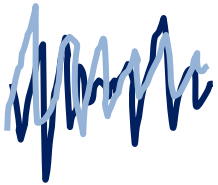
“External” whether datasets

Hunter McGushion has parsed the site and has created a lot of datasets based on the weather data, so thanks!

https://www.kaggle.com/huntermcgushion
https://www.kaggle.com/huntermcgushion/exhaustive-weather-eda-file-overview

Competition details

Multiple time series

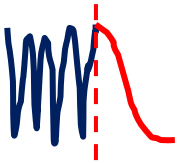


Different validation approaches

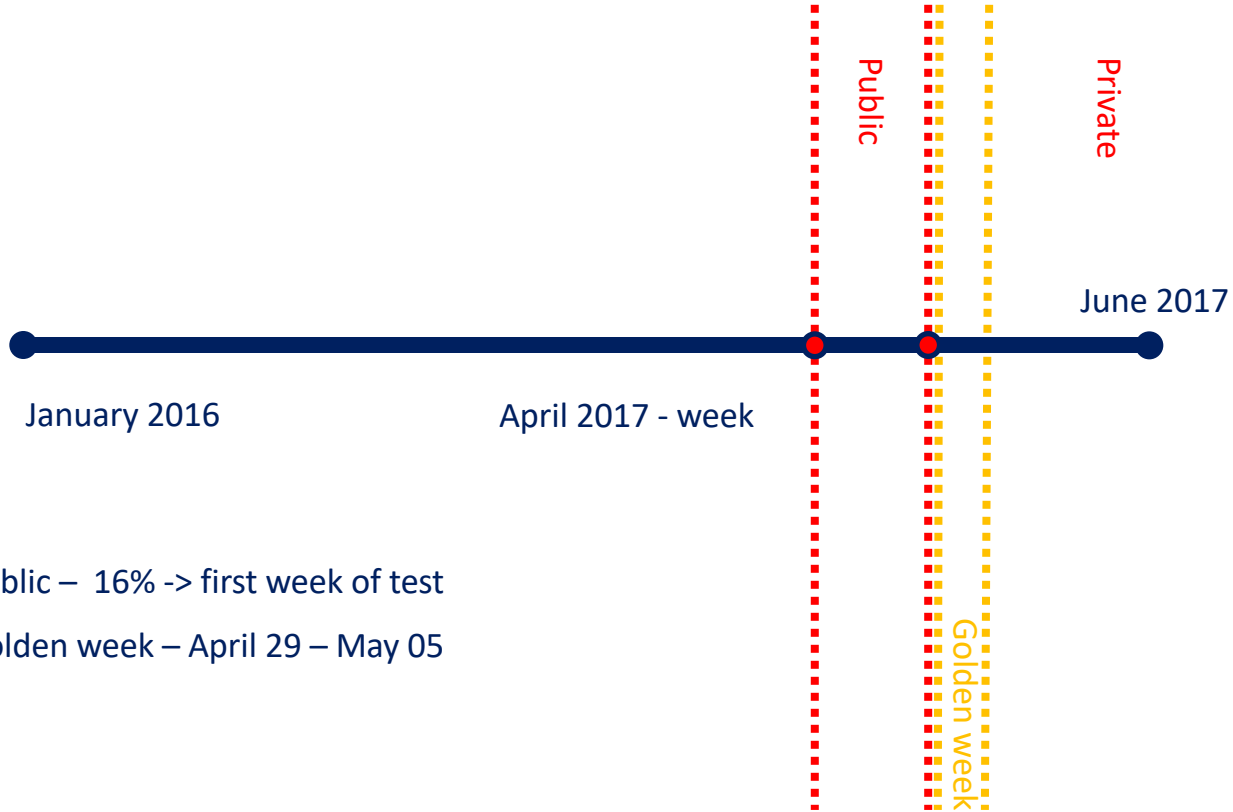
Whimsical data

“External” whether datasets

Golden week issue

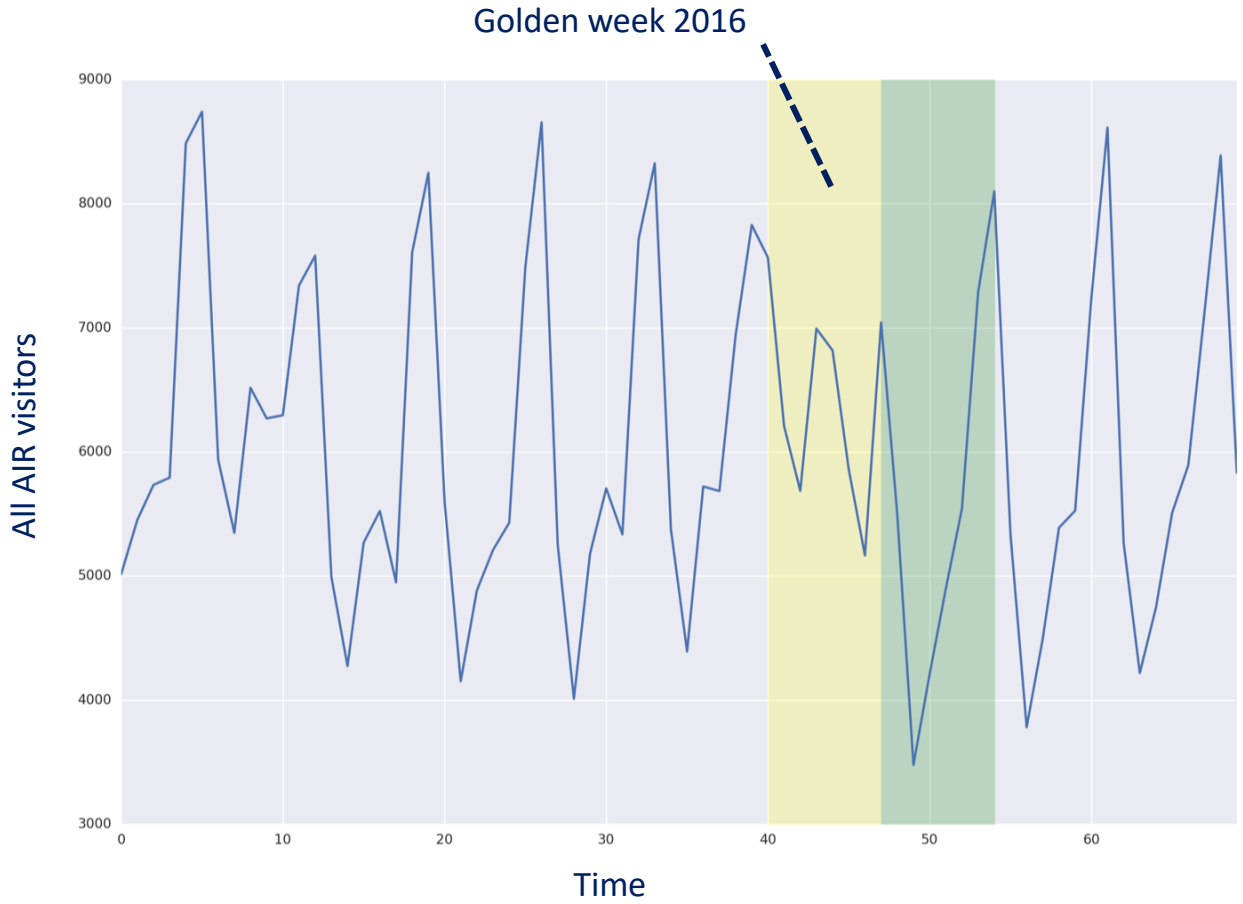


Golden week issue



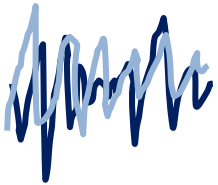
Public – 16% -> first week of test

Golden week – April 29 – May 05


Golden week 2016
All AIR visitors
Time
9000
8000
7000
6000
5000
4000
3000
0
10
20
30
40
50
60

Competition details

Multiple time series

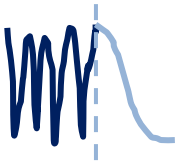


Different validation approaches

Whimsical data

“External” whether datasets

Golden week issue



Tiny public test set

84% 16%

Tiny public test set

Only one pair (restaurant_id, weekday) in public LB

Flavours of Physics	0.038	0.013	wAUC	0.30
Prudential Life	0.043	0.066	QWK	0.30
Rossman	0.044	0.041	RMSPE	0.39
Russian Housing	0.045	0.064	RMSLE	0.35
Grupo Bimbo	0.061	0.078	RMSLE	0.51
Restaurant Visitor Forecasting	0.062	0.105	RMSLE	0.16
Loan Default (Imperial)	0.065	0.012	MAE	0.20
Amazon from Space	0.067	0.140	F_2	0.66
Liberty Mutual 1 (Fire Peril)	0.073	0.077	NWGC	0.50
Allstate (Purchase Pred)	0.076	0.023	Accuracy	0.30
DonorsChoose.org	0.078	0.066	AUC	0.45
Event Recommendation Engine	0.079	0.038	MAP@200	1.00
Decoding Human Brain	0.092	0.101	Accuracy	0.43
Stumbleupon	0.095	0.184	AUC	0.20
Two Sigma Financial	0.100	0.085	R value	0.37
Winton Stock Market	0.109	0.286	wMAE	0.25
Santander (Cust Satisfaction)	0.109	0.314	AUC	0.50
Mercedes-Benz	0.146	0.240	R^2	0.19
Restaurant Revenue	0.158	0.256	RMSE	0.30
MLSP2014 Schizophrenia	0.240	0.385	AUC	0.52
Big Data Combine	0.300	0.592	MAE	0.30

Table by BreakfastPirate

Our features

0. air_store_id (label encoded)
1. Day of month, day of week, holiday_flag, holiday_tomorrow_flag
2. Combinations: id|weekday, id|year, id|year, id|holiday, id|holiday_tomorrow, weekday|area, weekday|holiday
3. Target encoding: Emean of visitors (target) by id|weekday, id|year, id|holiday_tomorrow, id, holiday_tomorrow, 5, etc.
4. Count encoding: count (expanding way) of some categorical feats
5. Sum of reserved seats in particular restaurant for particular day* (tricky feature!)

Our features

- Manual selection of features
- Building features for date X using only date Y: Y <= X (avoiding looking into the feature)
- No weather data or data from HPG-related files
- No latitude or longitude from original AIR datasets
- No genre or location data from original AIR datasets

Public models

“Surprise Me”-series kernel:

All available data + “leaky approach” + Xgboost + Sklearn GradientBoosting + KNN on unnormalized data + MLP + … + averaging with baseline script with weights based on the public LB performance.

Not so bad mixed with other kernels

? CV / 0.482 Public / 0.525 Public

https://www.kaggle.com/tunguz/surprise-me-2

Our models

It is better start with something simple:

id_weekday expanding mean baseline

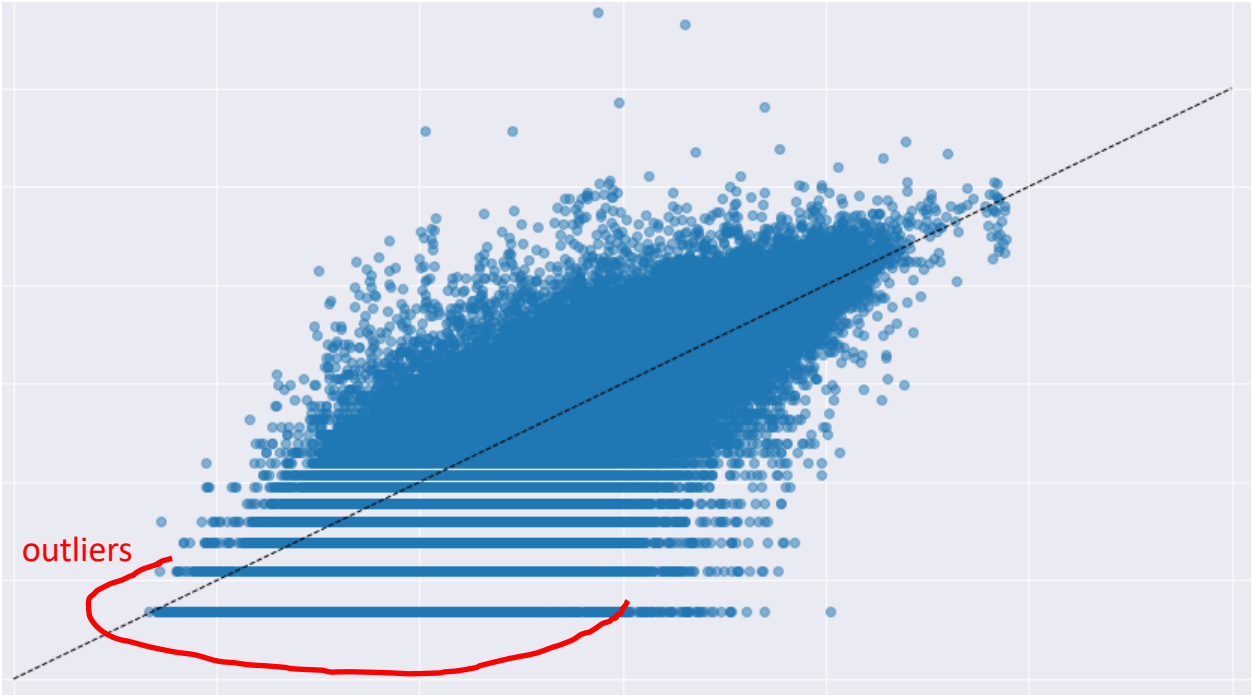
0.543 CV / 0.524 Public / 0.559 Private

Better than some boosting models on the start of the competition

Our models

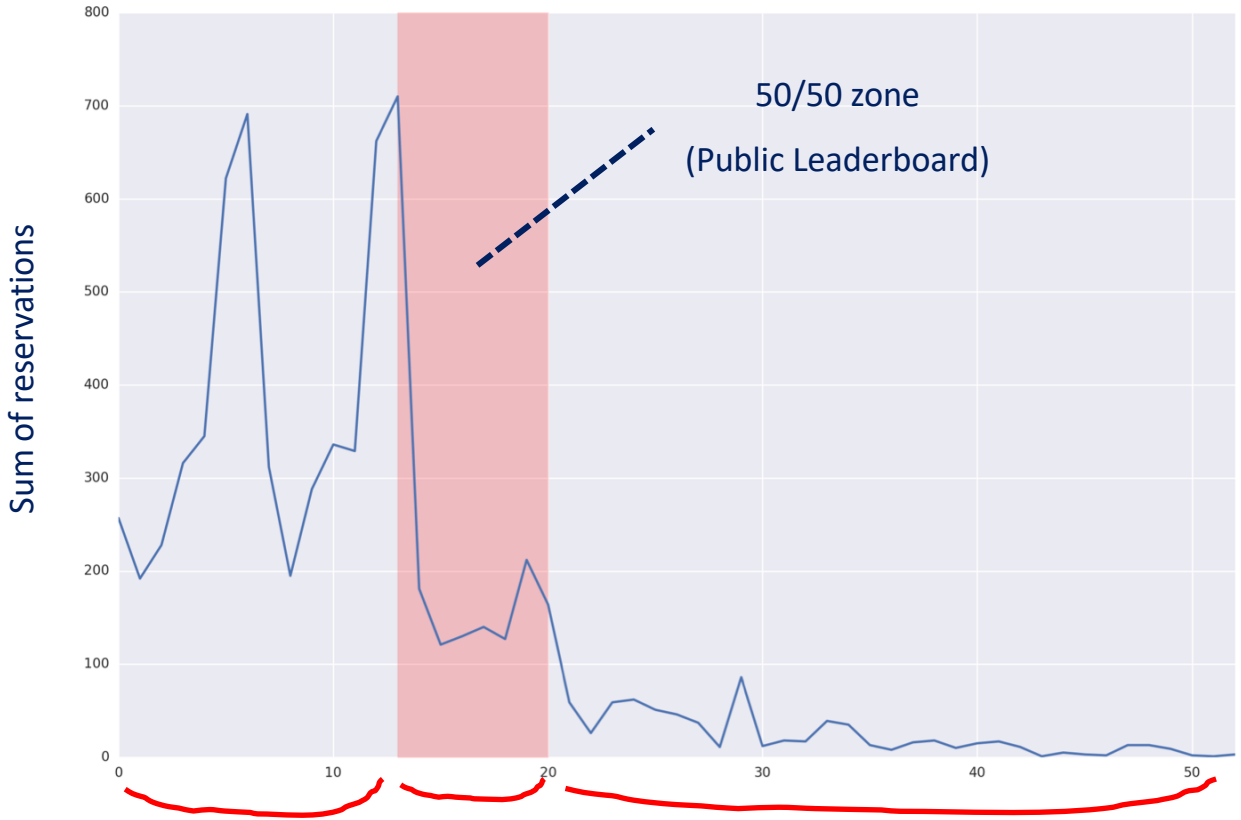
LGB - XGB convex combination

0.498 CV / 0.481 Public / 0.516 Private



And what is the deal with the tricky feature: sum of reserved seats in particular restaurant for particular day?

… recall



Nice and sound zone

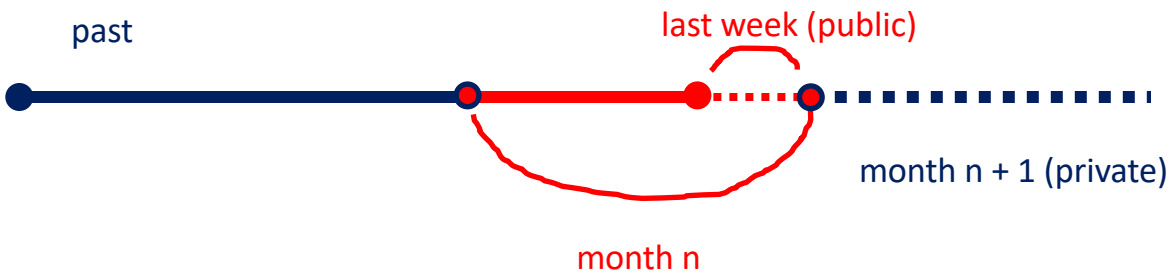
(Validation)

Dangerous zone

(Private Leaderboard)

Pseudo-public and pseudo-private scheme
for checking reservation quality

Feature construction and model fitting only on the “past”



Initial past := January 2016 – November 2016

Month 0 := December 2016

Reservations

Fold 201701 public: 0.6335

Fold 201701 private: 0.5328

Fold 201702 public: 0.4826

Fold 201702 private: 0.5217

Fold 201703 public: 0.4798

Fold 201703 private: 0.5471

Fold 201704 public: 0.5144

Fold 201704 private: 0.5208

Public CV: 0.528 Private CV: 0.531

Weak reservations*

*use lag ~ 30 days to calculate reservations sums

Fold 201701 public: 0.6628

Fold 201701 private: 0.5318

Fold 201702 public: 0.4897

Fold 201702 private: 0.515

Fold 201703 public: 0.4853

Fold 201703 private: 0.5334

Fold 201704 public: 0.5239

Fold 201704 private: 0.5186

Public CV: 0.54 Private CV: 0.524

Our models

LGB - XGB convex combination

0.498 CV / 0.481 Public / 0.516 Private

LGB - XGB convex combination with “weak” reservations

0.516 CV / 0.491 Public / 0.513 Private

LGB/XGB Tuning

- Manual hyperparameter tuning
- Train RMSLE close to validation RMSLE but no too close
- No early stopping (introduces a leakage, leads to wide range of optimal number of iterations for different validation folds)
- Using “categorical_feature” in LGB model
- Relatively high max_depth and relatively heavy regularization for XGB

Summary

- Solid validation approach (rolling CV + pseudo-sets)
- Minimalistic feature space: small fraction of generated features (basically, some categorical data + target encoding + weak reservations)
- Careful examinations of reservations
- Ignoring Golden week issue (both benefit and drawback)
- Manual parameter tuning
- Simple blending of 2 models

Some better solutions (7,8 and 12 place)

7 place (up ~550 places!):

- ”Expected visitors” using reservations –

average everything, add std

- NN as stacker model

12 place:

- Creative day of week treatment: treat every

day preceding holiday as Friday

Golden week Hack:

- Treat days in golden week not as holidays,

but as Sundays/Saturdays (based on EDA)

- https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting/discussion/49259 *
- https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting/discussion/49251 **
- https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting/discussion/49100 ***

Thank you for attention

nikita_churkin @ opendatascience
gleb_filatov @ opendatascience