Challenge data science ANAP-ATIH 2020:

Mieux anticiper l’augmentation des maladies chroniques!

Лучше предвидеть возникновение хронических заболеваний [Google translate]

Савватеев Сергей

10 декабря 2016 г.

О себе

- работаю аналитиком в TNS
- методы машинного обучения иногда использую в работе, но в контестах не участвовал до осени этого года
- пока не стал ходить на тренировки по машинному обучению…

once pop you can't stop

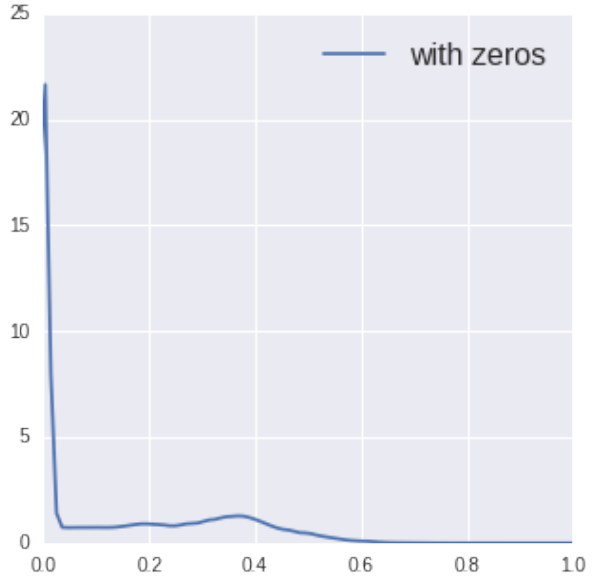
Contest

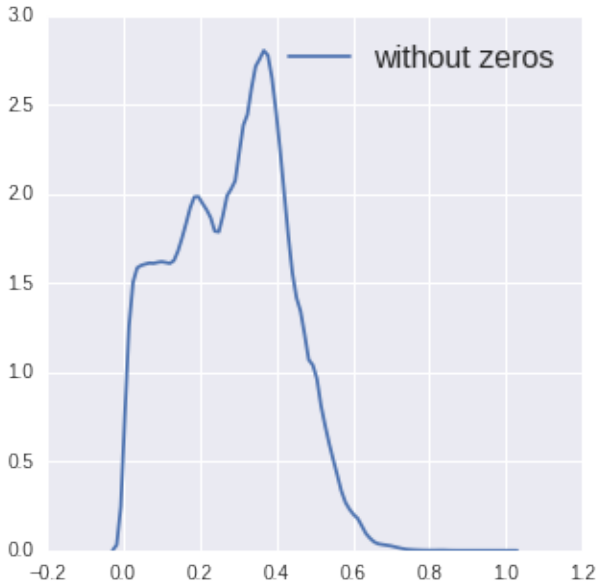

datascience.net
Accueil
Challenges
Classement
Aide
Mon compte
ANAP
appui santé & médico-social
atih
AGENCE TECHNIQUE DE L'INFORMATION SUR L'HOSPITALISATION
Challenge data science ANAP-ATIH 2020 : Mieux anticiper l'augmentation des maladies chroniques !
L'objectif du challenge proposé par l'ANAP et l'ATIH est de prévoir l'évolution à moyen terme de l'importance de la prise en charge des maladies chroniques pour les établissements de santé.
Classement
1. ➡ (1) Quentin Morel Score 1,95602%
2. ➡ (2) Romain Ayres Score 2,02876%
3. ➡ (3) Nicolas Gaude Score 2,05166%
Voir tout le classement
Résumé
Télécharger
Mes contributions
Discuter
Ce challenge est terminé.
6 000 €
3254 contributions
599 participants
terminé
Mes dernières contributions
14/11/16 19:42 Score 2,29983%
14/11/16 19:38 Score 2,40441%
14/11/16 19:28 Score 2,57185%

Особенности

- на начальном этапе был лик, его пофиксили
- целевая переменная:

среднесрочная эволюция важности лечения хронических заболеваний для медицинских учреждений





распределена от 0 до 1
около половины
значений - нули

Особенности

- были доступны “данные из будущего”
  “ex post” analysis?
- призы – только резидентам Франции
- метрика - RMSE

Данные

Базовые данные – 8 переменных:

- "ID", название клиники
- место жительства пациента (департамент), пациент младше / старше 75 лет
- область деятельности
- число долгосрочных пребываний, общее число пребываний
- год (2008 – 2013 для обучения, 2014 – 2015 для теста)

  тест делился на public и private части

1.88M записей в обучающих данных, 670K в тестовых

Данные

Opendata (xls-файлы по годам)

- HD – 178 показателей
- PDMREG – 6 показателей
- PDZMA – 5 показателей

  для пар госпиталь + район

Indicateur A1

A1	Part de marché globale en médecine sur la zone d'attractivité	1er niveau d'analyse
Cet indicateur permet d'appréhender la capacité de l'établissement à s'imposer sur sa zone d'attractivité en médecine et de mettre en évidence ses concurrents. La définition de la zone d'attractivité retenue est la suivante : la zone d'attractivité d'un établissement est le territoire défini par la liste des codes postaux dans lesquels l'établissement réalise les taux d'hospitalisation (Nb séjours médicaux/ Nb habitants) les plus élevés. Ces codes postaux sont classés de manière décroissante. Sont retenues dans la zone d'attractivité de médecine, les localités dont le cumul des séjours représente 80% de l'activité de l'établissement.		

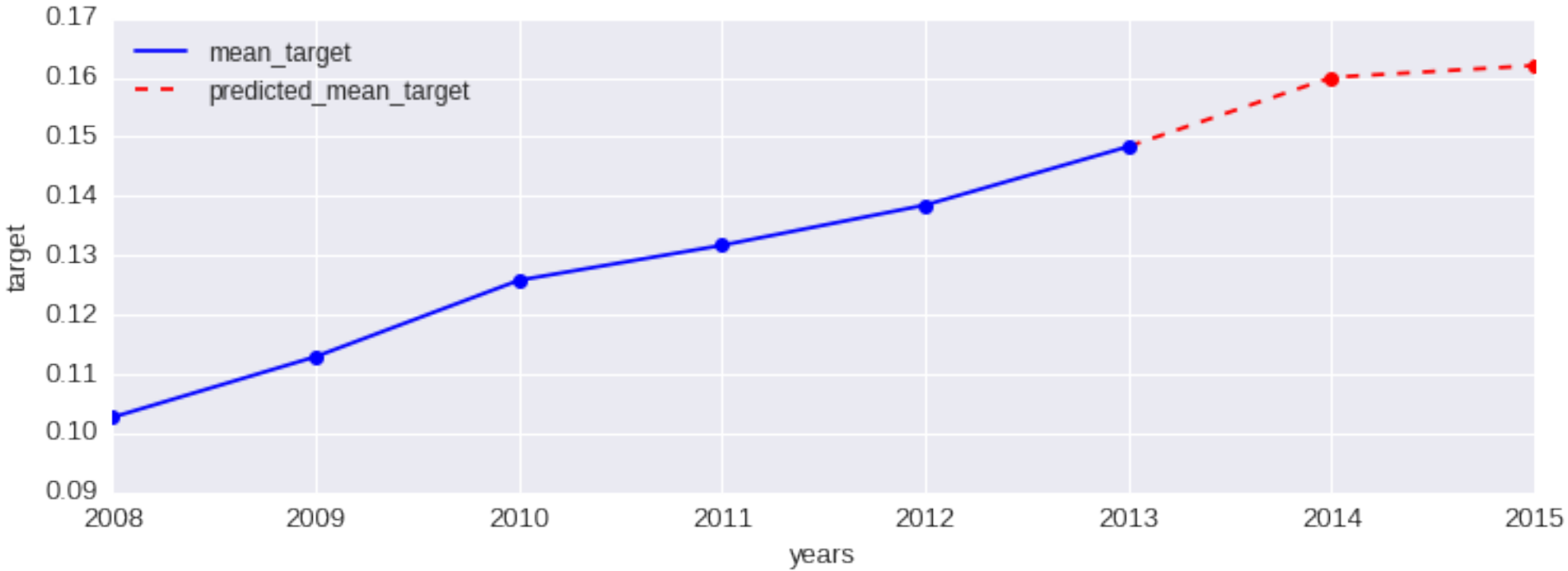
Source	PMSI / PMSI 1 (V11G, classification ASO M) ☑ Publics ☑ ESPIC ☑ Privé	Rang 1	S'exprime en %
Numérateur	Nombre de séjours de médecine réalisés dans l'établissement pour des patients résidant dans la zone d'attractivité	L'ensemble des séjours sont pris en compte, y compris les séjours ambulatoires Le nombre de séjours médicaux s'entend au sens de la classification ASO hors CMD 15 et CMD28. Les séjours classés en erreur ainsi que les prestations inter-établissement sont supprimés.	
Dénominateur	Nombre total de séjours de médecine pour des patients résidant dans la zone d'attractivité	L'ensemble des séjours sont pris en compte, y compris les séjours ambulatoires Le nombre de séjours médicaux s'entend au sens de la classification ASO hors CMD15 et CMD28. Les séjours classés en erreur ainsi que les prestations inter-établissement sont supprimés.	
Interprétation	La comparaison des parts de marché de l'établissement avec celles des établissements de sa zone permet d'identifier les établissements concurrents (la part de marché des établissements concurrents est en effet calculée sur la zone d'attractivité de l'établissement étudié). La zone d'attractivité permet d'appréhender le rayonnement l'établissement (local, régional, national) A noter : La part de marché sur la zone d'attractivité est d'autant plus faible que la zone d'attractivité est grande. Cet indicateur doit donc être mis en regard de la taille de la zone d'attractivité. La zone d'attractivité est la même pour les 4 années analysées (La zone d'attractivité retenue est celle de la dernière année disponible). L'évolution de cet indicateur indique s'il gagne ou perd des PdM sur cette zone. Il est pertinent, en second niveau d'analyse, d'étudier les parts de marché de chaque spécialité ou valorisées. L'analyse de cet indicateur doit être complétée afin de savoir si les évolutions constatées sont identiques en valorisation. De l'établissement (local, régional, national) A noter : La part de marché sur la zone d'attractivité est d'autant plus faible que la zone d'attractivité est grande. Cet indicateur doit donc être mis en regard de la taille de la zone d'attractivité. La zone d'attractivité est la même pour les 4 années analysées (La zone d'attractivité retenue est celle de la dernière année disponible). L'évolution de cet indicateur indique s'il gagne ou perd des PdM sur cette zone. Il est pertinent, en second niveau d'analyse, d'étudier les parts de marché de chaque spécialité ou valorisées. L'analyse de cet indicateur doit être complétée afin de savoir si les évolutions constatées sont identiques en valorisation.		
A croiser	A13 : Taux d'utilisation / occupation des lits en médecine P1 : IP-DMS en médecine A1bis : Part de marché en médecine sur la région		
Commentaire	Calcul de la zone d'attractivité : on classe les zones géographiques (un code postal ou un regroupement de codes postaux) par taux de pénétration décroissant (nombre d'hospitalisations pour 1000 habitants) de l'établissement concerné. On sélectionne les zones qui ont les plus forts taux de sorte que le recrutement sur ces entités géographiques représente 80% du recrutement total de l'établissement (ou 80% du recrutement sur des codes géographiques connus dans le cas où l'établissement possède beaucoup de codes erronés ou étrangers). Ces calculs se font sur toutes les hospitalisations à l'exclusion des séances. La zone ainsi déterminée représente la zone d'attractivité de l'établissement et c'est sur cette zone que sont calculées les parts de marché. Il est possible pour en avoir une représentation (même si elle n'est pas identique dans son calcul) d'aller sur http://cartographie.atih.sante.fr		

данные были ключевыми в задаче

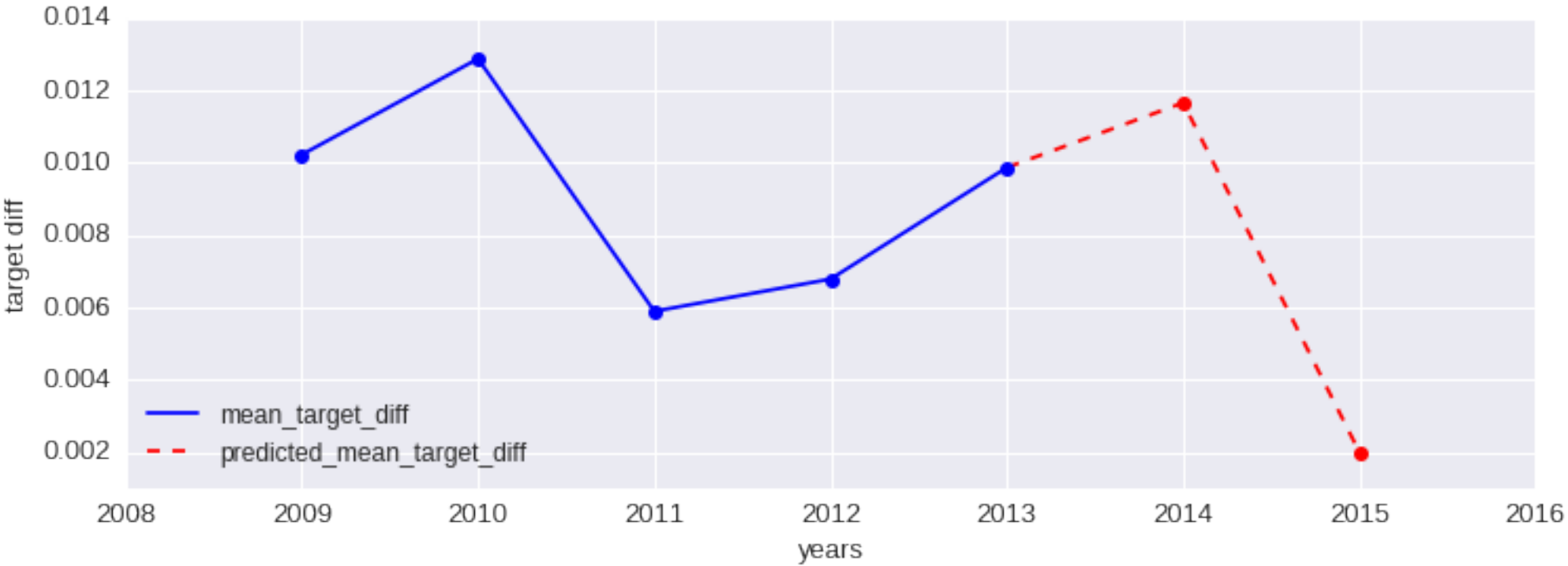
часть важных показателей для 2015 года были N/A

Целевая переменная

`число длит-х визитов == 0  <=>  таргет == 0`

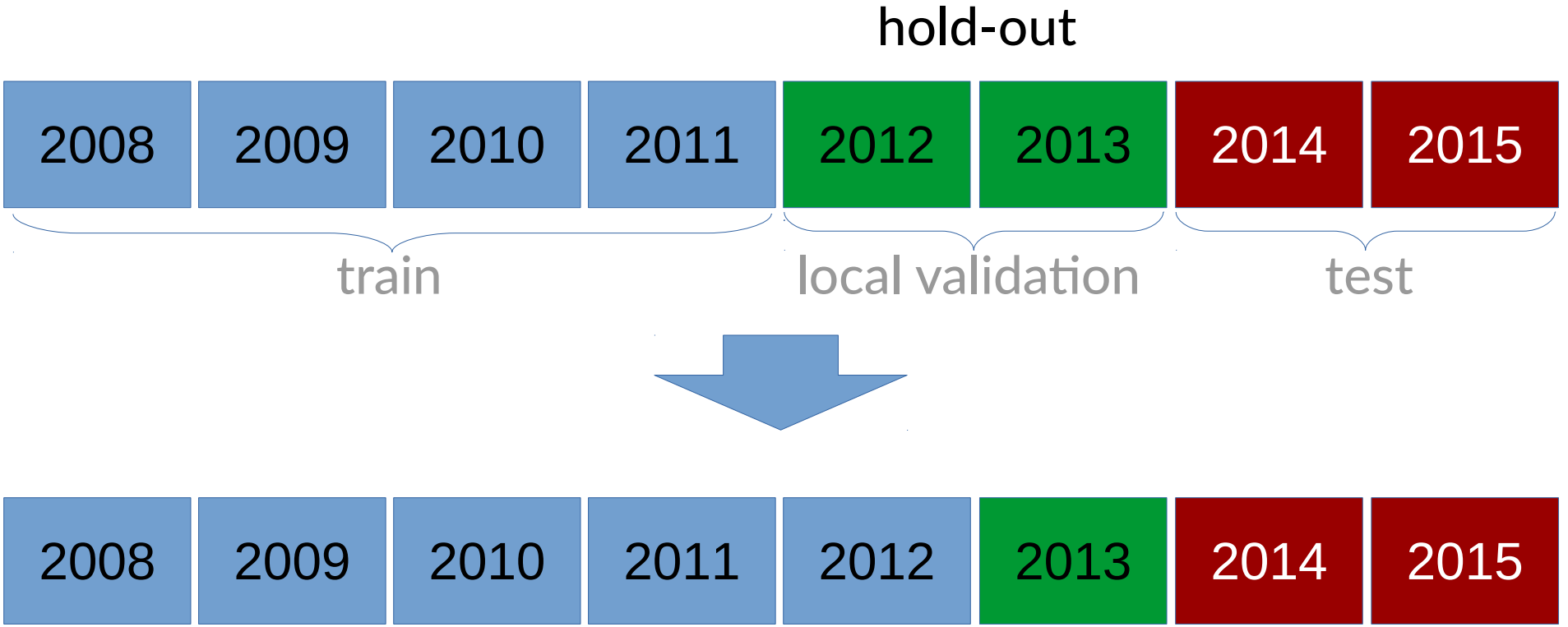


Целевая переменная (разность)



- прогноз выгодно было домножать
  коэффициент подбирался по лидерборду
- достаточно один раз было подобрать среднее за год

Валидация


hold-out
2008
2009
2010
2011
2012
2013
2014
2015
train
local validation
test
2008
2009
2010
2011
2012
2013
2014
2015

Признаки

finess

310781067

provenance

31-Haute-Garonne

- выделение департаментов из finess и provenance
- “is aborigen” - живет ли пациент там же, где госпиталь
- отношение числа длительных визитов к общему числу визитов (quot)

  линейная корреляция с таргетом - 74%
- признаки из opendata xls

Признаки

- кодирование таргетом (OOF), quot
  mean, median
  с группировкой по разным сочетаниям полей
- таргет, quot с лагом на один и два года
  с группировкой по finess+domaines+provenance+age
  2008 и 2009 год исключался из обработки
  для 2015 лаг-1 брался прогноз на 2014

В результате датасет уже содержал ~250 признаков

“Блочный” Add-Del

- жадный алгоритм поочередного добавления/удаления признаков не работал (как хотелось бы)
- разбивать признаки на блоки по смыслу и уже внутри них запускать Add-Del – получилось лучше

  252 признака → 29 признаков без потери качества

Алгоритмы

Основные модели – “сладкая парочка”

- XGBoost
- LightGBM

  после небольшой настройки давал сравнимое качество, при этом работал в 5 раз быстрее

Давали сильно худшее качество

- RandomForest и ExtraTrees

Две модели

Модели строились отдельно для 2014 и 2015

- важную фичу y_lag1 приходилось брать из прогноза на 2014
- часть фичей в xls отсутствовала для 2015

  взять их из 2014 не помогало

В результате того, что модель для 2015 года строилась без части важных фич, она была сильно хуже по качеству и преодолеть этот разрыв в ходе соревнования так и не удалось

Блендинг

Простое усреднение различных моделей давало хороший прирост

- XGBoost и LightGBM
- “широкий” и узкий датасеты и т.д.

Стекинг

- с помощью XGBoost, LightGBM и RandomForest добавил фич по годам, но не успел отправить...

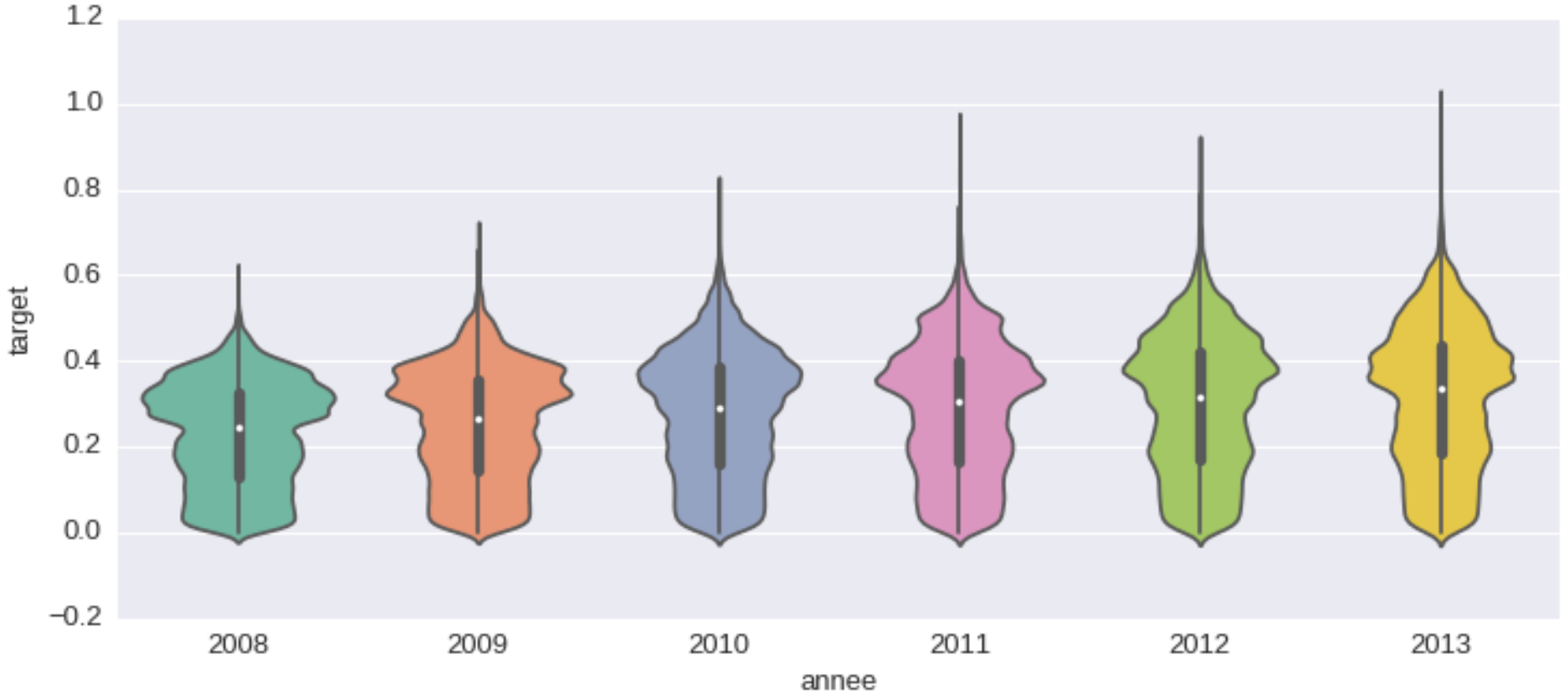
5 points about LightGBM

Released 2016-10-17

- изначально был только CLI, поэтому все пользовались pyLightGBM - сторонним враппером вокруг него
- уже появилась бета-версия родного python-враппера, без оверхеда на сохранение датасета и с полезными плюшками
- API похож на XGBoost, удобно пользоваться
- дискретизирует признаки (histogram based)
- leaf-wise tree growth

Как учесть тренд

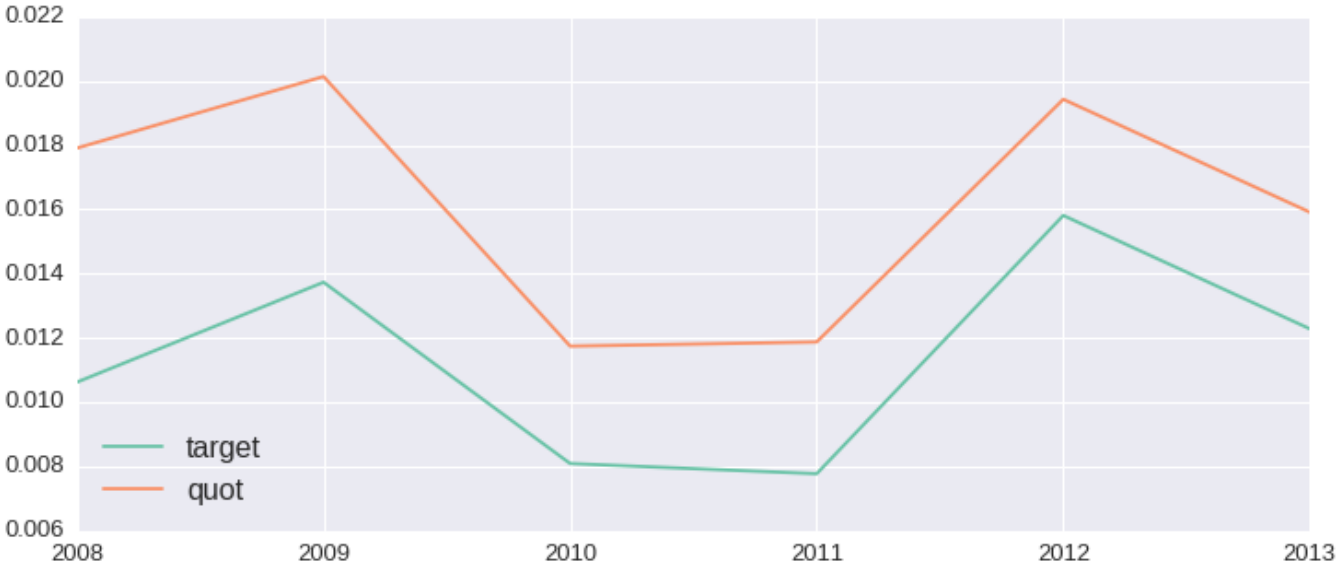
Деление на средний таргет по годам не давало прироста по качеству



Как учесть тренд

ID клиники	депар-тамент пациента	возраст	обл-ть деят-ти	год	число долговр. визитов	общее число визитов	таргет

взглянуть на данные, как на 390K временных рядов



простой признак:

таргет с лагом, но смасштабированный по quot

$$y'_t = \frac{y_{t-1}}{quot_{t-1}} * quot_t$$

Как учесть тренд

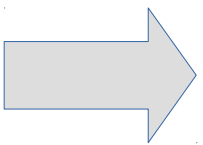
Регрессия в рамках группы

- без учета года строим модель регрессии (quot и еще какой-нибудь признак, как предикторы)
- прогнозируем следущий год - $\mathbf{pred_1}$
- строим еще одну регрессию уже на остатках (только год в качестве предиктора)
- прогнозируем следующий год - $\mathbf{pred_2}$
- сумма двух прогнозов – $\mathbf{pred_1 + pred_2}$ используем как признак, вместе с R^2 от обоих регрессий

4-е место Дмитрий Дрёмов

- join по finess без учета года (для важных признаков из xls)
  т.е. для каждого finess и признака “col”:

year	col
2008	val1
2009	val2
...	
2014	val7
2015	val8



year	col1	col2		col7	col8
2008	val1	val2	...	val7	val8
2009	val1	val2	...	val7	val8
...	...	...	...	...	...
2014	val1	val2	...	val7	val8
2015	val1	val2	...	val7	val8

- жадное удаление признаков
- модели по годам получились более ровные

3-е место Nicolas Gaude

almost 100% the same but

- huge effort on lag value of cible accross different aggregation (by finess by activity by age etc…)
- plus much different models to give my final blend more diversity

  linear regression, neural network, randomforest
- having separate model + lag value was the key

1-е место Quentin Morel

- 2012 & 2013 for training
- Same base features
- Open data from Insee was used to add sociodemographic features but it was not really useful
- Missing values for 2015 was replaced with the values of 2014 and it gave a very significant gain
- 40 xgboost on different subsets of features with different parameters and 1 random forest
- Ridge Regression to stacking

Резюме

- join-фичи
- больше фич с лагом и прогнозом внутри групп
- стекинг/блендинг снова решает

Спасибо за внимание


s.savvateev@gmail.com


@sswt opendatascience.slack.com