Быков Филипп Леонидович

«Гидрометцентр России»

О прогнозе посещаемости страниц Википедии

Web Traffic Time Series Forecasting

18 место (best public solution ~ 150 место)

Постановка задачи соревнования

Есть данные о суточных просмотрах ~145.000 страниц Википедии

1я стадия: с 1 июля 2015 по 31 декабря 2016 = 550 дней

2я стадия: они же по 10 сентября 2017 = 803 дней

Оценка по метрике $SMAPE = 2\frac{|\text{fact} - \text{forecast}|}{\text{fact} + \text{forecast}}$

Требуется дать прогноз их же на 2 месяца:

1я стадия: с 1 января по 1 марта 2017. $SMAPE$~44%

2я стадия: с 13 сентября по 13 ноября 2017. $SMAPE$~38%

Данные для 2й стадии появились 12 сентября и было всего ~ 12ч чтобы загрузить решение. ~ 3 сентября появились данные по 31 августа

Некоторые участники нашли, где данные есть в открытом доступе (не использовал)

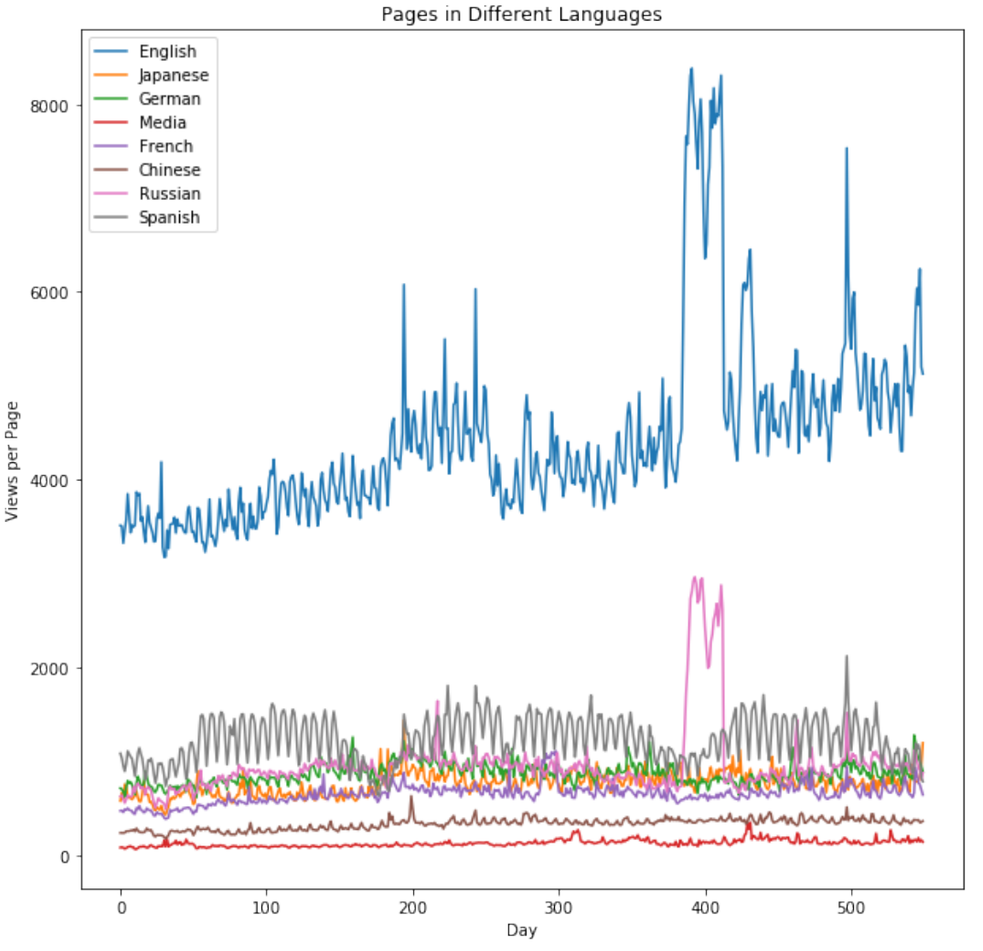
Известен адрес страницы (в т.ч. язык) и «agent» (desktop, mobile, spider)

Никто (?) информацию на странице не использовал в решении

Какие интересные наблюдения можно было сделать по данным?

1. Зависит от дня недели / рабочий ли день
2. Некоторые статьи имею выраженные ежегодные пики. Например «Нобелевская премия», статьи об известных людях
3. Много непредсказуемых пиков (смотрят информацию, связанную с новостями?)
4. Есть 0 и пропуски, при этом, в моем решении при замене пропусков на 0 оценки прогнозов хуже

5.В русской и английской вики странный выброс летом 2016



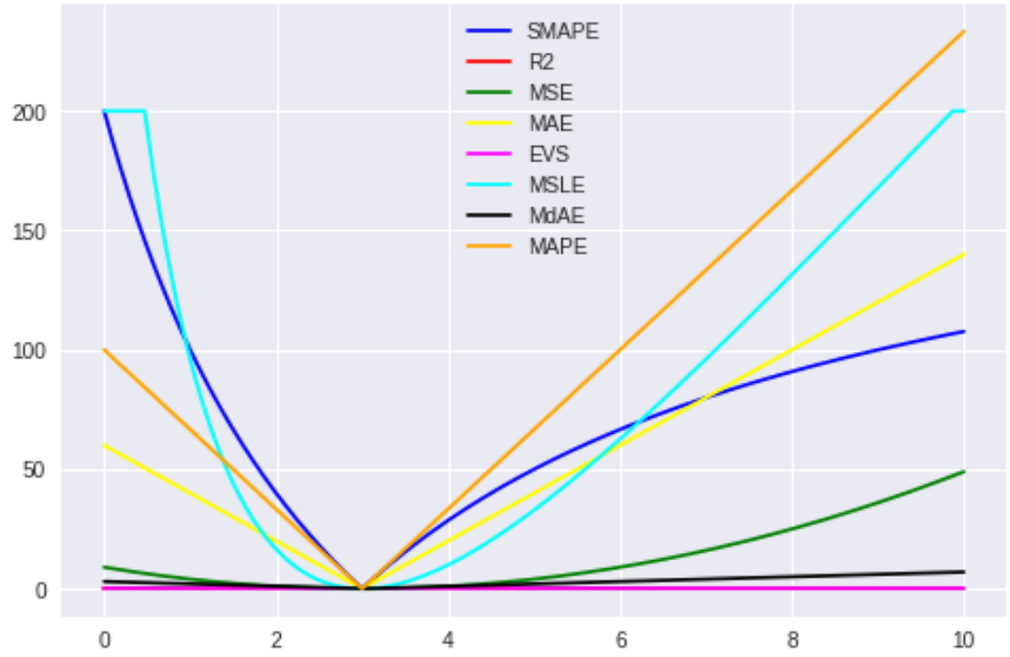
Какой способ валидации модели использовали, почему именно такой способ, насколько качество на валидации коррелировало с лидербордом?

Валидация на последнем 2 месячном периоде (такой же длины как требовалось submit)

Корреляция с лидербордом была хорошая, стабильно -1%

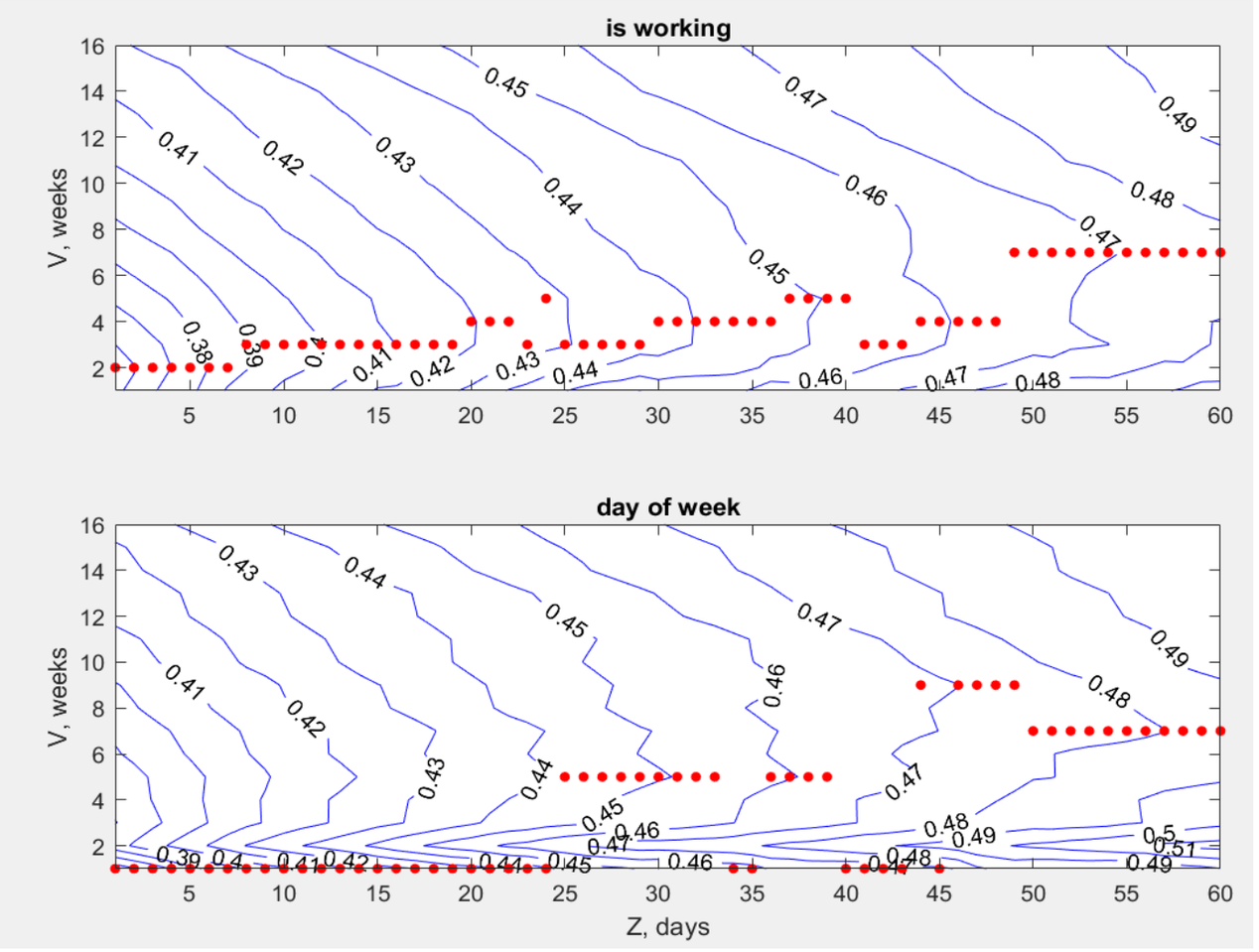
Сделал 3 подвыборки из 3000, 15000, 50000 рядов для быстрой валидации

Метрика $SMAPE = 2\frac{|\,\text{fact} - \text{forecast}\,|}{\text{fact} + \text{forecast}}$



Медиана по дням недели – хорошая стартовая модель, наилучший срок обучения = 56 дней

Зависимость отклонения от медианы от заблаговременности (Z) и длины обучения (V)



Я выбирал длины периодов обучения V = 1, 2, 4, … недель и блендил их:

Пусть T – последний день, посещаемость в который доступна. Z_{max} – максимальная необходимая заблаговременность.

1. Учим 3 группы методов (прогноз = медиана по таким же дням недели / рабочий ли день, Y = 0, 1):

   $M_1(V)$ - на отрезках $[T - V - Z_{max} - 365Y; T - Z_{max} - 365Y]$

   $M_2(V)$ - на отрезке $[T - V - 0.5Z_{max} - 365Y; T - 0.5Z_{max} - 365Y]$

   $M(V)$ - на отрезке $[T - V - 365Y; T - 365Y]$

2. Оцениваем A_1 и A_2 = SMAPE погрешность $M_1(V)$ и $M_2(V)$ соответственно. Оцениваем скорость увеличения погрешности $S = \max(0,\log(A_1/A_2))/(0.5Z_{max})$

3. Оцениваем погрешность прогноза с периодом обучения V и заблаговр. Z:

$$A(V,Z) = A_2 \exp((Z-0.25Z_{max})S)$$

В некоторый момент появилось public решение «Median of medians», где V выбирались как Фибоначчи V = 1, 2, 3, 5, 8, … недель и это было лучше, чем 2^n примерно на 1.4%

После этого я заменил в своем решении:

V = [4, 7, 11, 17, 28, 45, 72, 117, 190, 307, 496, 803] дней при Y=0

V = [21, 35, 56] дней при Y=1.

Итого 15*2 (учитывающие день недели или только рабочий или нет) = 30 прогнозов

При этом такой замене преимущество над наилучшим из public решением сохранилось на уровне ~1.7%

4.Блендим методы. Варианты:

	Crossvalidation	Leaderboard (Late Submit)
$M(56,Z)$	39.816%	
$M(V,Z)$ с наименьшим $A(V,Z)$	39.971%	
Медиана	37.410%	38.086%
Взвешенная медиана	37.195%	38.293%
Усреднение (итоговое решение) $\exp\left(\frac{\sum_V \log\left(1+M(V,Z)\right)A(V,Z)^{-2}}{\sum_V A(V,Z)^{-2}}\right)-$	37.119%	38.236%
Итоговый Leaderboard (код не сохранился)	38.315%	

MATLAB, i7-2600K = 5 мин; i5-4200U = ~15 мин (кроме предварительного преобразования данных)

На решение ушло в сумме ~7 дней. ~30 июля загрузил данные. 4-20 августа был в отпуске, ничего не делал

Решения остальных участников

1е место (SMAPE=35.480): LSTM сеть. Параметры: история просмотров, agent, язык, year-to-year autocorrelation, quarter-to-quarter autocorrelation

5е место (SMAPE=37.132): Блендинг 4 методов: 1. медиана, 2. медиана с учетом дня недели, 3. авторегрессия, 4. как год назад. Подбор коэффициентов от заблаговременности

8е место (SMAPE=37.583): Медианы + фильтр Калмана для оценки тренда