Sberbank Data Science Contest

Задача B

Предсказание суммарных трат всех клиентов по категориям

Желубенков Александр
19 ноября, 2016

Данные

- Таблица с транзакциями
- Таблица с описанием mcc-кодов транзакций

	customer_id	tr_datetime	mcc_code	tr_type	amount	term_id
0	39026145	0 10:23:26	4814	1030	-2245.92	NaN
1	39026145	1 10:19:29	6011	7010	56147.89	NaN
2	39026145	1 10:20:56	4829	2330	-56147.89	NaN
3	39026145	1 10:39:54	5499	1010	-1392.47	NaN
4	39026145	2 15:33:42	5499	1010	-920.83	NaN

	mcc_code	mcc_description
0	742	Ветеринарные услуги
1	1711	Генеральные подрядчики по вентиляции, теплосна...
2	1731	Подрядчики по электричеству
3	1799	Подрядчики, специализированная торговля — нигд...
4	2741	Разнообразные издательства/печатное дело

- Таблица с информацией о поле клиентов
- Таблица с описанием типов транзакций

Количество транзакций: ~6.8 млн.
Количество различных категорий(mcc-кодов): 184.
Количество клиентов: 15.000.
Период наблюдения: 15 месяцев (457 дней).

Задача

Требуется:

Предсказать объем трат по каждой из 184 категорий
на каждый день следующего месяца.

Объем трат в конкретной категории считается как сумма всех расходных транзакций в текущей категории по всем пользователям.

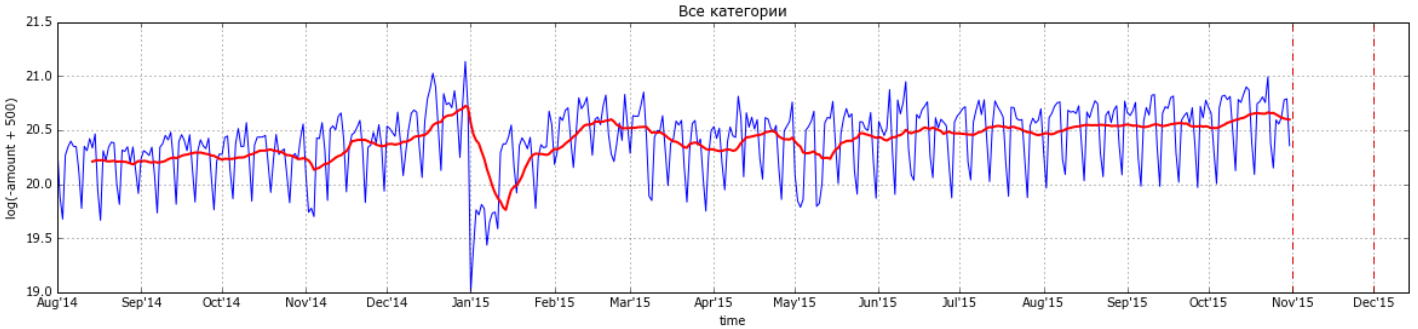
Количество предсказаний: 184 * 30 = 5520.

Метрика: RMSLE (со смещением 500)

$$\mathrm{RMSLE}_{500} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log(y_i + 500) - \log(|\tilde{y}_i| + 500)\right)^2}.$$

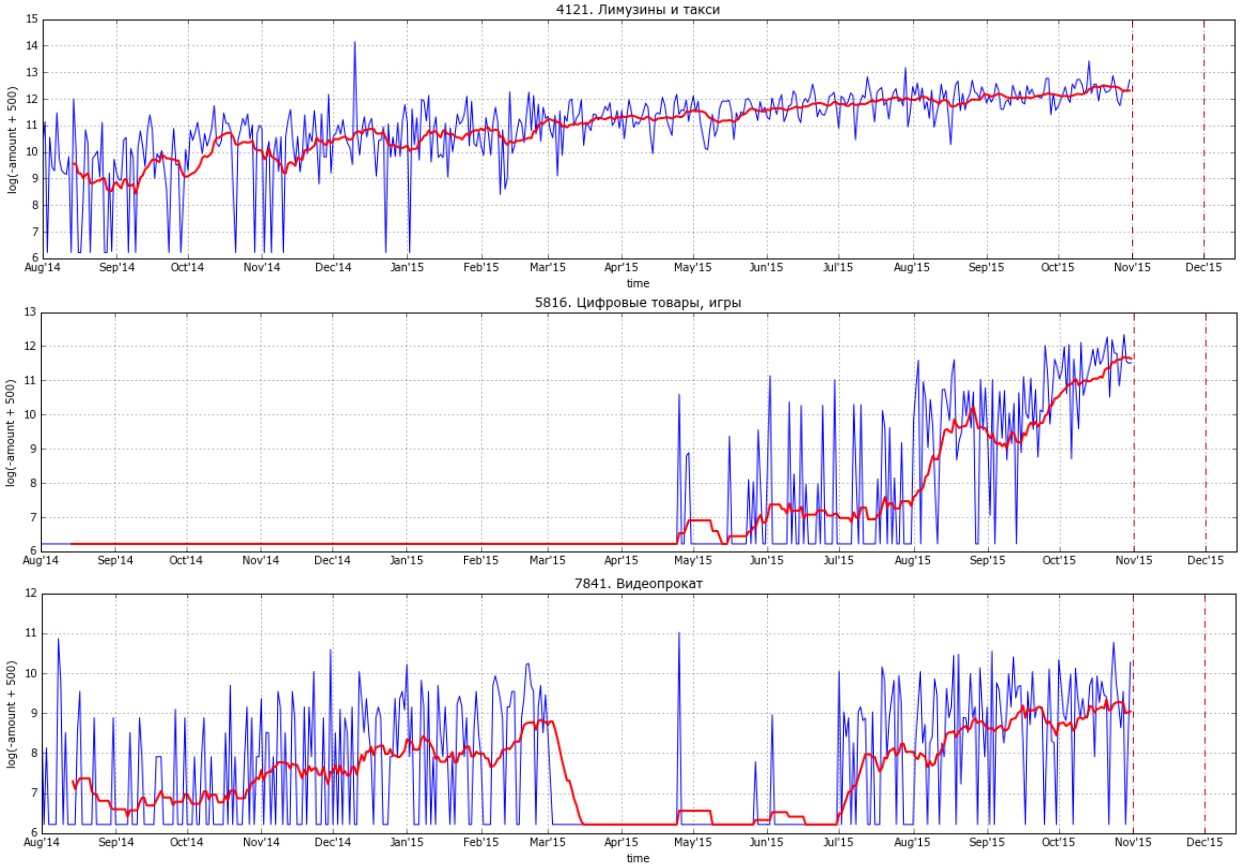
Предсказание суммарных трат

Все категории:



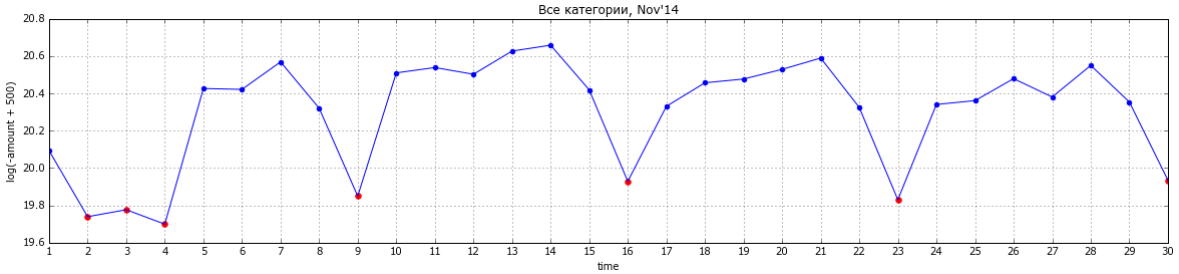
Отдельные категории:

Предсказание суммарных трат

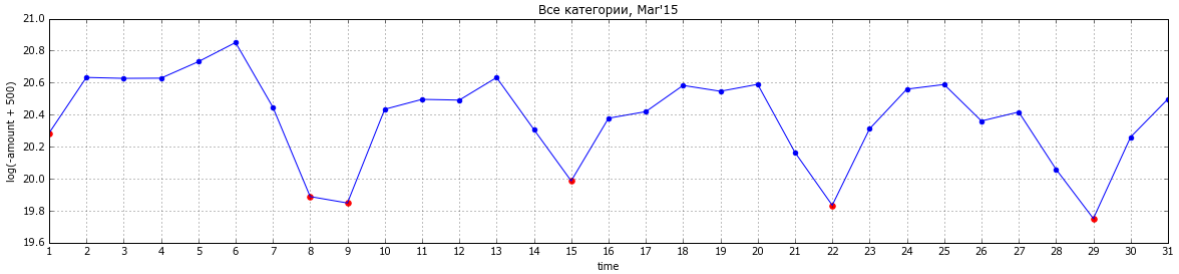


Идентификация периода времени

Период наблюдения: 15 месяцев: август 2014 – октябрь 2015.



Ноябрь						
пн	вт	ср	чт	пт	сб	вс
27	28	29	30	31	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
1	2	3	4	5	6	7



Март						
пн	вт	ср	чт	пт	сб	вс
23	24	25	26	27	28	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

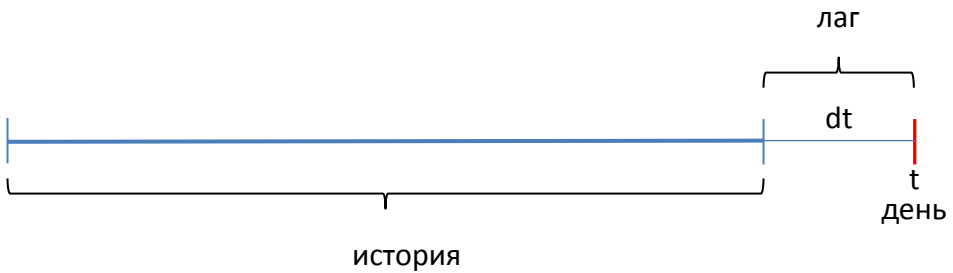
Период для предсказания: ноябрь 2015.

Ноябрь						
пн	вт	ср	чт	пт	сб	вс
26	27	28	29	30	31	1
2	3*	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	1	2	3	4	5	6

Прогнозирование временного ряда

Цель: прогнозировать временные ряды на 1 месяц вперед (30 дней вперед)

Объект: тройка (категория, день, временной лаг)



Временной лаг: номер дня в месяце

Факторы(1)

(1) Ключ – (день):

- Номер дня от начала периода наблюдения (0..486).
- Номер месяца (1..12).
- День недели (0..6).
- 3 бинарных категории праздничных дней (ручная разметка):
  a. (15 дней): '2014-11-04', ['2015-01-01'-'2015-01-08'], '2015-02-23', '2015-03-08', '2015-05-01', '2015-05-09', '2015-06-12', '2015-11-04'
  b. (20 дней): a. + 5 дней (нерабочие будние дни)
  c. (33 дня): b. + 13 дней (выходные дни, соседние с праздничными)
- 2 бинарные категории пред/после праздничный день.
- Составные бинарные категории: «сб + вс», «сб + вс + праздничный день», «вс + праздничный день», «пред-праздничный день + праздничный день»,..
- Количество дней до/после праздника.
- + временной лаг

Итого: 15 факторов.

Факторы(2)

(2) Ключ – (категория + день):

Средние значения для следующих разрезов (7 разрезов):

- (категория, *)
- (категория, день недели)
- (категория, «сб + вс»)
- (категория,«сб + вс + праздничный день»)
- (категория, «вс + праздничный день»)
- (категория, «праздничный день»)
- (категория, «пред-праздничный день + праздничный день»)

Размеры окон (12 окон): (1, 2, 3, 4, 8, 10, 13, 16, 20, 26, 52 + все) предшествующие недели.

Временной лаг: номер дня в месяце

Итого: 7 * 12 = 84 фактора.

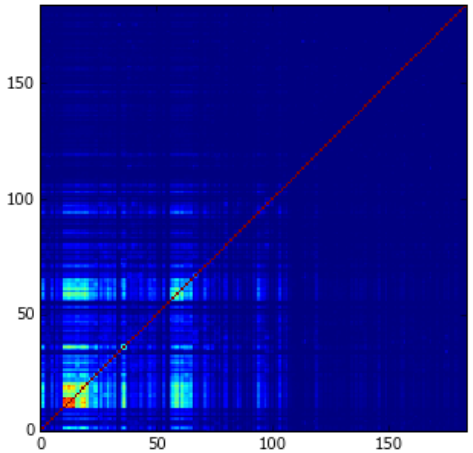
Кластеризация категорий

Матрица связей категорий:

$$sim_{i,j} = \frac{{c_{i,j}}^2}{c_i c_j}$$, где

c_i - количество клиентов, у которых встретилась категория i,

$c_{i,j}$ - количество клиентов, у которых встретились обе категории i, j.



Спектральная кластеризация (sklearn.cluster.SpectralClustering)

Количество кластеров: 46

- 10 единичных кластеров;
- Один общий: 75 категорий;
- 35 адекватных кластера:
  - 'Авиалинии, авиакомпании', 'Авиалинии, авиакомпании, нигде более не классифицированные', 'Туристические агентства и организаторы экскурсий';
  - 'Ветеринарные услуги', 'Зоомагазины';
  - 'Готовая женская одежда', 'Одежда для всей семьи', 'Обувные магазины', 'Магазины мужской и женской одежды', 'Различные магазины одежды и аксессуаров', 'Магазины косметики'.
  - 'Финансовые институты — снятие наличности автоматически', 'Финансовые институты — снятие наличности вручную', 'Бакалейные магазины, супермаркеты', 'Денежные переводы', 'Звонки с использованием телефонов, считывающих магнитную ленту'.

Факторы(3-4)

(3) Ключи – (кластер категорий + день), (все категории + день):

Усредненные факторы(2) для ключей (категория + день)

Итого: 2 * 84 = 168 факторов.

(4) Бинаризация категориальных факторов:

- категории (184 фактора)
- группы категорий (46 факторов)
- «особенные дни» (выходные, пред/праздничные дни, 14 февраля, 1 сентября) (128 факторов)

Общее число факторов: 631

Обучение и тестирование

Кросс-валидация: 9 месяцев – (февраль 2015 – октябрь 2015)

Тренировочная выборка: 14 месяцев (сентябрь 2014 – октябрь 2015), размер ~78.000.

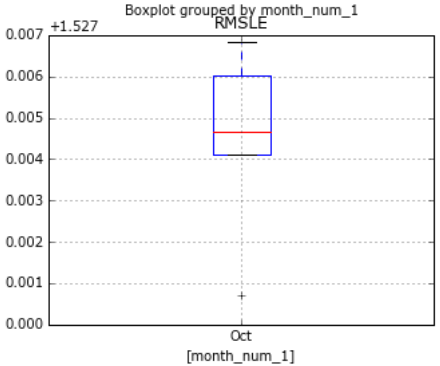
Обучение и тестирование

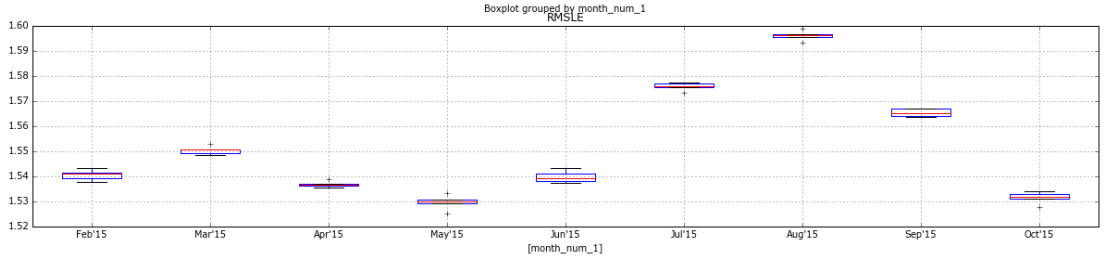
Модель: xgboost

Параметры:

- max_depth: [3, **4**, 5, 6];
- min_child_weight: [10, **20**, 40];
- n_estimators: [**80**, 100, 150];
- learning_rate: **0.1**;
- colsample_bylevel: [0.05, **0.1**, 0.2].

5 запусков с разным значением параметра 'seed'.





Финальная модель

Параметры одной модели xgboost:

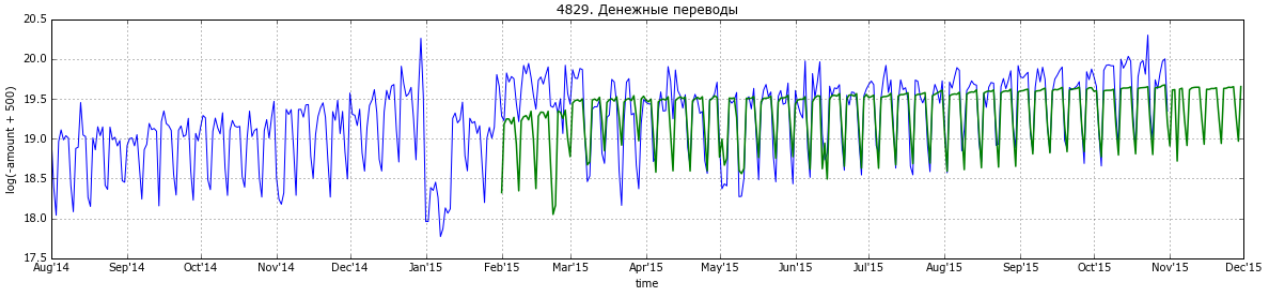
- max_depth: 4;
- min_child_weight: 20;
- subsample: **0.8**;
- n_estimators: 80 -> **350**;
- learning_rate: 0.1 -> **0.03**;
- colsample_bylevel: 0.1;
- seed: ?;

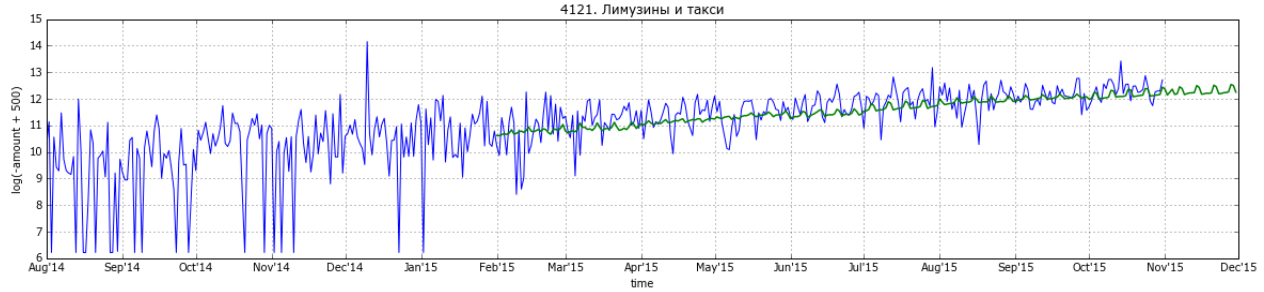
Факторы	#моделей	CV-score	Окт' 15	Public	Private
(1),(2): (day + mcc_code)	1	1.553	1.536	1.621	?
+ (4): (categorical)	12	1.549	1.529	1.618	?
+ (3): (all, cluster)	12	1.548	1.527	1.616	1.540

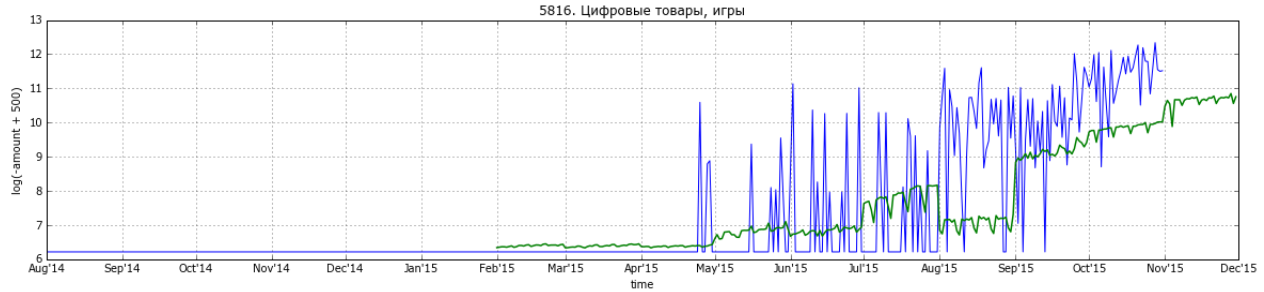
Генерация разных выборок: Временной лаг для каждого дня - randint(1, 31)

Усреднение 12 моделей: 4 выборки с разными временными лагами * 3 разных seed-а.

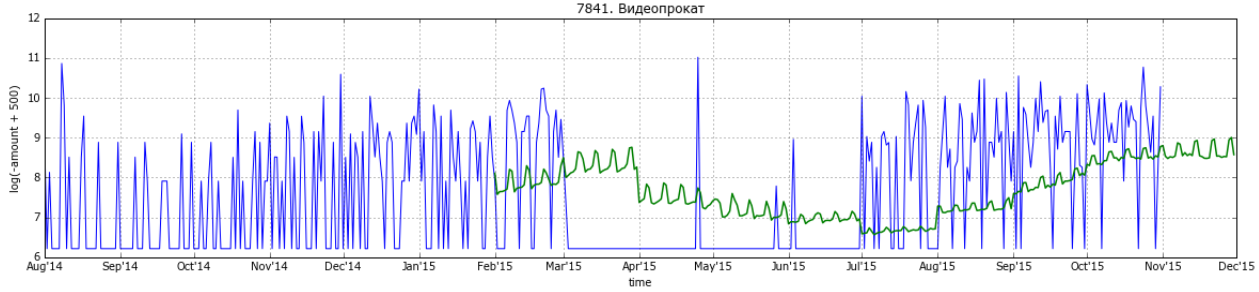
Прогнозы модели

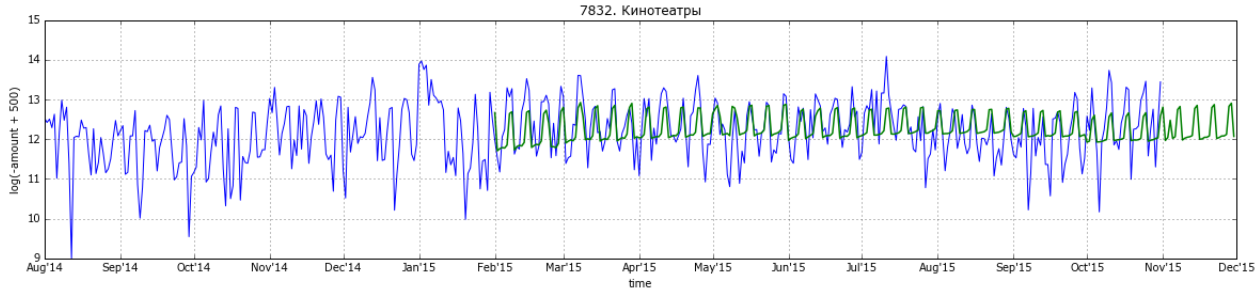






Прогнозы модели





Технические детали

Инструменты:

- IPython notebook
- pandas, numpy, scipy, sklearn, xgboost
- Asus, Core-i7, 8GB RAM

Время:

Генерация факторов: ~40 мин

Обучение финальной модели: ~20 мин

Ссылка на код:

https://github.com/Topspin26/SberbankDataScienceContest_2016

Другие решения и идеи

- Для ключа (категория, день) использование:
  - среднего взвешенного, медианы, ст. отклонения;
  - линейных регрессий;
- Линейные комбинации базовых моделей отдельно для каждой категории.
- Label encoding вместо one-hot encoding
- Модели авторегрессии-скользящего среднего
- Связи между разными категориями.
- Положительные(доходные) транзакции?

1 место по задаче B:
https://github.com/VasiliyRubtsov/SberbankSol

#	Участник	Общий балл	Задача A	Задача B	Задача C
3	rubcovvasilii	850.9795	0.896056 (168.0613)	1.535778 (300.0000)	1.276845 (382.9181)
28	mpyat	723.71	0.891389 (142.0432)	1.537844 (294.2893)	1.339443 (287.3775)
37	sorokinv	670.6694	0.885995 (111.9721)	1.538344 (292.9072)	1.353587 (265.7901)
10	Rety	794.0239	0.889795 (133.1568)	1.539828 (288.8052)	1.283958 (372.0619)
2	mango	852.3675	0.895825 (166.7735)	1.539992 (288.3519)	1.26746 (397.2421)
4	Topspin26	838.3445	0.895124 (162.8655)	1.540366 (287.3181)	1.27341 (388.1608)
16	atikhonov	763.7522	0.888832 (127.7882)	1.540953 (285.6956)	1.298237 (350.2684)
8	alexey	806.0317	0.889764 (132.9840)	1.541589 (283.9576)	1.272788 (389.1102)
14	akulov	768.9838	0.887683 (121.3826)	1.541597 (283.9155)	1.289446 (363.6858)
6	maxim.voevodsky	818.1524	0.892151 (146.2913)	1.541888 (283.1111)	1.273024 (388.7500)

Спасибо за внимание!

Желубенков Александр
zhelubenkovalexandr@gmail.com