Отчет по контесту «Предсказание затрат»

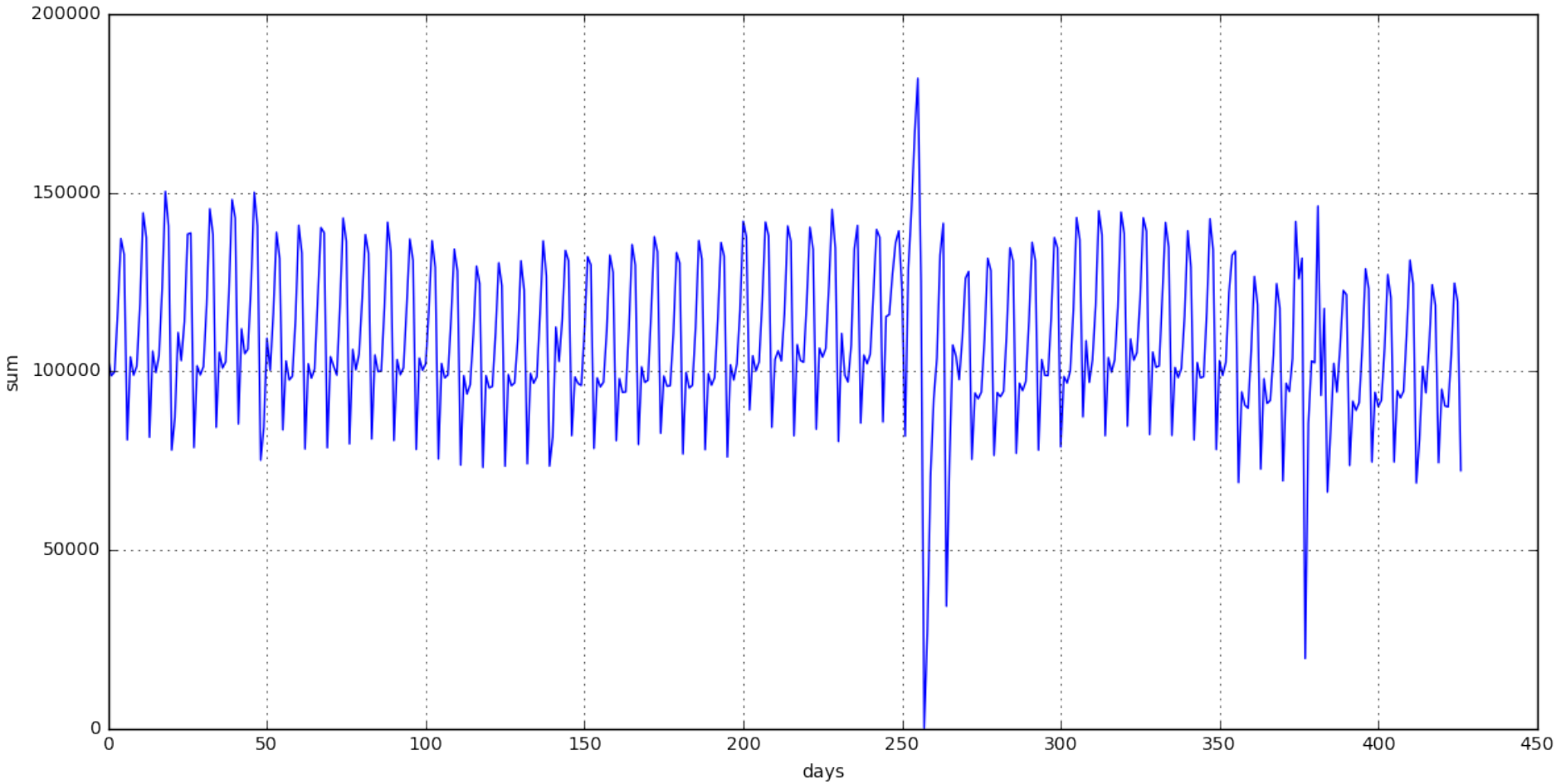
Таскынов Ануар, 517 группа

Курс «Прикладные задачи анализа данных»

5 октября 2017 г.

Постановка задачи: дана статистика посещения клиентов одного магазина "X"и необходимо было предсказать сумму следующей покупки, которая будет совершена в течение семи дней. Если клиент не придет в магазин, то ответ 0.
Данные: клиентов было 110000, история покупок содержала 438 дней, а суммы покупок были категоризированы и если клиент не совершал покупку, то сумма покупок равна 0.
Метрика качества: Accuracy.
Проверка качества: проверка качества производилась локально удалением последней недели и по Public.

Проверка на адекватность данных.



Видна периодичность, однако каждый клиент ведет себя "по-своему"и не поддается данному закону. И лучше всего рассматривать их отдельно друг от друга. Для того, чтобы лучше представлять недели, решено удалить самые первые дни.

Идея 1: пытаться спрогнозировать день, когда придет клиент, дальше выдать в качестве ответа суммы покупок моду. Рассматривается следующая вероятностная модель поведения:

$$\hat{p}_i = p_i \prod_{k=1}^{i-1} \big(1 - p_k\big),$$

где p_i – вероятность посещения в i-й день, оцененная по последним неделям.
Результат: **0.25557** на Public.

Идея 2: забить на предсказание дня и спрогнозировать модой. Заметно, что некоторые клиенты перестали ходить в магазин. Для них выдать 0, а для остальных спрогнозировать модой.
Результат: **0.35018** на Public.

Идея: рассмотреть по всем неделям только первые покупки и сложить их с затухающими весами. Далее перенормировать, чтобы интерпретировать вероятность суммы покупки.
Были выбраны следующие весовые схемы:

$$w_i = \frac{1}{i^\gamma}$$

$$\hat{w}_i = \beta^{i^\delta},$$

где i - это номер недели, отсчитанный с конца.
Результаты: первая весовая схема с $\gamma = 0.6$ дала на Public **0.39060**. Вторая весовая схема с $\beta = 0.99, \delta = 1.5$ дала на Public **0.39303**.
Можно было смешать вероятности данных двух схем с коэффициентом α:
$\alpha * w + (1 - \alpha) * \hat{w}$.
Результат: **0.39436** с коэффициентом $\alpha = 0.4$.

Идея: забить на схемы и запустить XGBoost на всех данных, пусть сам разбирается. Валидироваться по клиентам.
Результат: **0.35284** на Public при стандартных настройках XGBoost.
Добавление/удаление фичей:

- оставить только первые покупки по неделям.
- средняя покупка за последний месяц в дни, когда были сделаны покупки.
- средние покупки по всем неделям в дни, когда были сделаны покупки.

Результат: **0.38842** на Public.
Идея: изначально задача была регрессионной, тогда можно было бы попытаться запустить линейную регрессию и подать их ответы в качестве признаков XGBoost. Однако из-за "неустойчивых"клиентов возможно не было возрастания качества.
Итоговое решение: смешать смесь двух весовых схем и XGBoost с весами 0.6 и 0.4 соответственно – **0.39503** на Public.

Выводы.

- клиенты очень "нестабильны"и валидироваться по ним жизнеопасно.
- весовые схемы работают лучше, чем XGBoost. Однако, возможно, это из-за маленького числа сгенерированных признаков и неправильной валидации по клиентам.
- подставлять значения регрессора в качестве признаков - плохая идея.

Что за данные?