Sberbank Data Science Contest
Отчет
Василий Рубцов

1 Задача A

В данной задаче необходимо было предсказывать пол клиента по его транзакциям. Метрика оценка качества — AUC-ROC. В решение описывается набор признаков, генерируемых по данным о транзакциях для каждого клиента:

- Всевозможные статистики:
  - Количество дней между первым и последним днем транзакций
  - Разница между мединным и средним значением дней транзакций
  - Максимальная расходная транзакция
  - Максимальная приходная транзакция
  - Среднее по транзакциям
  - Среднее по расходным транзакциям
  - Медиана транзакция
  - Медиана расходных транзакций
  - Сумма всех транзакций
  - Сумма всех расходных транзакций
  - Среднее значение времени (время — количество минут от полуночи)
  - Медиана времени
  - Средне квадратичное отклонение
  - Средне квадратичное отклонение тех времен, которые не равны нулю (то есть за исключением автоматических операций, выполняемых в полночь)
  - Медиана количества транзакций в день
  - Среднее по всем среднеквадратичным отклонениям времени, которые считались только по тем дням, в которых больше одной транзакции

  - Среднее число минут между покупками

- Статистики по транзакциям по каждому отдельному коду:
  - Количество транзакций
  - Доля транзакций
  - Сумма транзакций
  - Среднее значение по выходным
  - Отношение числа покупок в буднии дни к числу покупок в выходные

- Статистики по транзакциям по каждому отдельному типу:
  - Сумма транзакций

- Траты в определенные праздники по характерным кодам:
  - Количество трат по коду 5944 (ювелирные украшения) за пару дней до нового года и 8 марта
  - Количество трат на цветы (код 5992) 8 марта.
  - Количество трат на мужскую одежду (код 5947) перед 23 февраля
  - Количество трат на подарки (код 5611) перед 23 фефраля

- Остальные признаки
  - Первая компонента по нормированным суммарным тратам в пространстве некоторых групп кодов (характерных для мужчин и женщин)
  - Наличие каждой из 180 наиболее популярных пар транзакций (если брать в порядке совершений трат и найти наиболее популярные пары по всем данным)
  - Отношение общего количества трат в выходные к количеству трат в будни

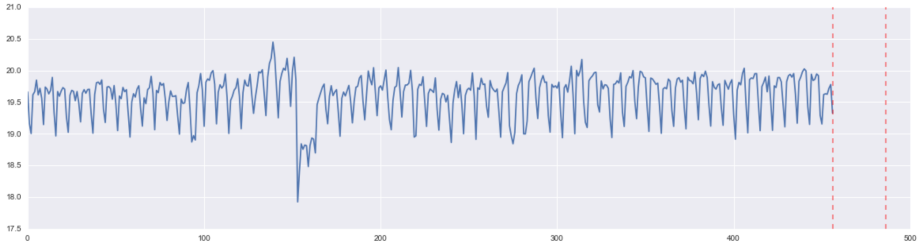
На основе этих признаков обучается xgboost. Всего — 11 моделей с разными параметрами max_depth и subsample. Затем используется блендинг.

2 Задача B

Задача состояла в прогнозировании временного ряда, где ключевой переменной было количество общих трат клиентами сбербанка за день по отдельному коду транзакций. Всего было 148 кодов. Оценкой качества была метрика RMSLE со сдвигом 500:

$$\text{RMSLE}_{500} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(y_i + 500) - \log(\hat{y}_i + 500))^2}$$

То есть, метрика RMSE по логарифмированным со сдвигом значениям. Далее в отчете подразумевается, что все значения уже прологарифмированы. Вот так выглядит временной ряд для кода 6010 (Финансовые институты — снятие наличности вручную):



Красными линиями помечен промежуток времени, на который нужно прогнозировать. На графике можно увидеть когда был новый год — 152-ой день.

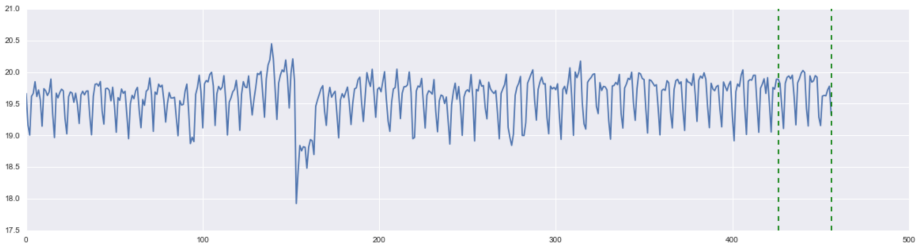
Решением задачи было использование xgboost на вход которому подавались значения простых предикторов, а также их комбинаций и нескольких дополнительных признаков.

Одним из таких признаков был код транзакций. Этот код — категориальная переменная. Обычно, label encoding таких признаков использовать лучше, чем one hot encoding при работе с моделями, основанными на деревьях принятий решений. А еще лучше использовать label encoding, при котором кодирование категорий происходит в соответствии с тем, какое среднее значение ключевой переменной в данной категории.

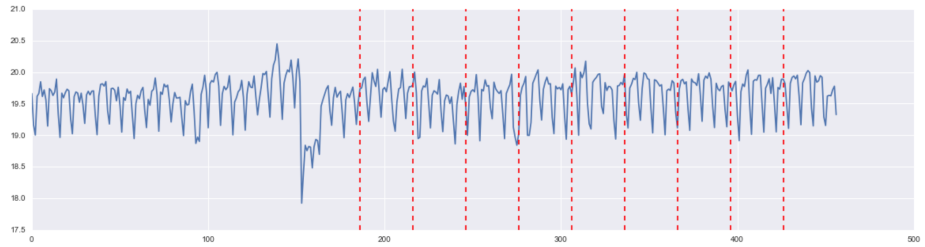
Например, возьмем первые шесть кодов транзакций: 742, 1711, 1731, 1799, 2741, 3000. Среднее значение дневных трат по ним: 7.74, 7.12, 6.69, 6.96, 6.56, 14.35. Предположим, что в данных только такие категории. Тогда была бы логично переобозначить данные коды, например, в такие числа: 4, 3, 1, 2, 0, 5. Тем самым помогаю дереву сразу делать хорошие разбиения (например сразу разбить на группы где среднее значение целевой переменной больше 7 или меньше). Однако если использовать непосредственно значения из выборки для обучения, то это привидет к

переобучению. Этого можно избежать, если считать среднее по той части истории, которая не используется в значениях ключевого признака выборки для обучения.

Для тестовой выборки было выбрано последние 30 дней:



Для обучающей выборки были выбраны промежутки в 30 дней до выборки для тестирования.



И тогда среднее значение для формирования оценки по которой можно переобозначить код можно считать на оставшемся промежутке времени (до 187-го дня).

Теперь предположим, что мы выбрали 30-ти дневный период, на который мы делаем прогноз. Составим множество простых предикторов. Например, на данном графике, показано какие значения получаются если брать в качестве предикторов линейную регрессию, среднее за все время, а также линейную регрессию, построенную на днях тех же дней недели, что и время, на которое мы прогнозируем:

Здесь мы прогнозировали на 14-ый день.

В работе использовался следующий набор предикторов:

- Среднее за все время
- Среднее за последние пол года
- Среднее за последний месяц
- Среднее за последнюю неделю
- Значение в последний известный день
- Средне взвешенное за последние 50 дней (вес тем больше, чем ближе к последней известной дате)
- Среднее по дням неделям
- Среднее по дням неделям за последние пол года
- Линейная регрессия
- Линейная регрессия по дням неделям
- Линейная регрессия на данных за последние пол года
- Линейная регрессия на данных за последний месяц

Здесь "по дням неделям" значит, что если мы прогнозируем на день, который является, например, понедельником, то мы используем только данные с понедельников.

Далее, если у нас есть набор простых предикторов — $y_1, ..., y_n$, то мы можем брать их комбинации. Например, выпуклую комбинацию:

$$y_{n+1} = w_1 y_{i_1} + w_2 y_{i_2} + ... + w_k y_{i_k},$$

где $w_1 + ... + w_k = 1$ и $w_i \geqslant 0$ для $i = 1, ..., n$.

Однако, мы понимаем что некоторые предикторы будут работать для одних кодов лучше, чем других. Точно также разные предикторы будут ошибаться по разному прогнозируя на 1 или на 30 дней вперед. Например для кода 6010, график которого представлен выше, очень хорошо работает регрессия по дням неделям, в то время как для кодов, у которых нету четко выраженных недельных циклов, она будет бесполезна. Точно также если мы рассмотрим линейную регрессию построенную на днях за прошедший месяц, то она очень не стабильна — дает хороший прогноз на первые дни, но прогнозируя на 30-ый день может сильно ошибаться и даже уйти в минус.

Поэтому, было бы хорошо брать разные веса $w_1, ..., w_k$ для каждого кода и для каждого дня. В таком случае сильно растет количество параметров, и для того, чтобы избежать переобучение, необходимо выбрать грамотную стратегию подбора весов. Например, можно брать веса в соответствии с тем, какую среднюю ошибку данная модель делала в прошлом.

Пусть $e = (e_1, ..., e_k)$ ошибки предикторов $y_{i_1}, ..., y_{i_k}$. Так как имеет смысл с большим весом брать ту модель, ошибка которой меньше, то можно взять просто $\hat{w}_i = 1/e_i$. Можно обобщить: $\hat{w}_i = 1/e_i^\alpha$, где $\alpha > 1$, подразумевая таким образом, что большая ошибка еще больше занижает вес. Далее необходимо нормировать веса: $w_i = \hat{w}_i/(\hat{w}_1 + ... + \hat{w}_k)$. Таким образом получаем следующую модель:

$$y_{n+2} = w_1(e)y_{i_1} + w_2(e)y_{i_2} + ... + w_k(e)y_{i_k},$$

где e — вектор ошибок.

Помимо того, важными признаками будет разница между значениями некоторых предикторов. Например, разница между средним за все время и средним по дням неделям будет индикатором важности предикторов, построенных по дням неделям.

Итоговый набор признаков выглядит следующим образом:

1) Код (переобозначенный).

2) День, на который прогнозируем (от 1 до 30).

3) Доля нулевых значений в истории данного кода.

4) Простые модели:

   - y_1 — Среднее за все время
   - y_2 — Среднее за последние пол года
   - y_3 — Среднее за последний месяц
   - y_4 — Среднее за последнюю неделю
   - y_5 — Значение в последний известный день
   - y_6 — Средне взвешенное за последние 50 дней
   - y_7 — Среднее по дням неделям
   - y_8 — Среднее по дням неделям за последние пол года
   - y_9 — Линейная регрессия
   - y_{10} — Линейная регрессия по дням неделям

- y_{11} — Линейная регрессия на данных за последние пол года
- y_{12} — Линейная регрессия на данных за последний месяц

5) Некоторые разницы

- $d_1 = y_1 - y_7$
- $d_2 = y_1 - y_2$
- $d_3 = y_5 - y_{10}$

6) Комбинации простых моделей:

- $y_{13} = w_1 y_8 + w_2 y_3 + w_3 y_{10}$
- $y_{14} = w_1(e) y_7 + w_2(e) y_{10} + w_3(e) y_6 + w_4(e) y_2$

Данные признаки можно подавать в xgboost. В итоге это что-то вроде стекинга простых моделей и их комбинаций, где помимо выходов моделей добавляются еще некоторые признаки.

Как можно было выяснить из данных, месяц, на который необходимо было прогнозировать — это ноябрь 2014. Тот самый месяц, в котором было резкое повышение курса доллара. В связи с этим стоит ожидать переоценку или недооценку значений моделью по некоторым кодам. В качестве таких кодов было выбрано 3501 (жилье — отели, мотели, курорты), 4722 (туристические агентства и организаторы экскурсий) и 6211 (ценные бумаги: брокеры/дилеры). Все предсказания этих кодов были изменены на константу. Для кодов 3501 и 4722 предсказания необходимо было уменьшить в связи с уменьшением туристической активности россиян в этом месяце. А для кода 6211 — увеличить, так как многие начали играть на курсе валют.

3 Задача C

Задача состояла в прогнозировании суммарных трат клиента на месяц вперед по каждому коду. Метрика оценки качества RMSLE со сдвигом 1. Основная стратегия такая же как в задаче B. Прогнозируем временной ряд на следующий месяц. В качестве признаков:

- Среднее значение
- Значение за последний месяц
- Среднее за последнии 4 месяца

- Среднее за последнии 8 месяцев
- Идентификатор того, что все предыдущие значения нулевые
- Линейная регрессия
- Среднее за те месяца, в которых была какая-либо активность клиента
- Линейная регрессия по тем месяцам, в которых была какая-либо активность
- Средне-взвешенное (вес тем больше, чем ближе к последнему известному месяцу)
- Среднее значение каждого из предыдущих оценок по всем пользователям по данной категории

Дальше, пусть у нас есть общее количество трат каждого из клиентов по каждому коду. Мы можем посчитать корреляцию между каждой парой клиентов как корреляцию между векторами их трат по каждой категории. Это будет оценкой схожести между клиентами $sim(i, j)$. Теперь для каждого клиента и для каждого кода мы можем посчитать новый временной ряд, как средневзвешенное соответствующих рядов других клиентов, наиболее похожих с данным, а в качестве весов взять их схожесть.

$$\hat{s}_{i,c} = \frac{\sum_{j \in N_i} sim(i, j) s_{j,c}}{\sum_{j \in N_i} sim(i, j)},$$

где $s_{i,c}$ — временной ряд трат клиента i по категории c, N_i — множество наиболее похожих с i клиентами, то есть те, корреляция с которыми наибольшая (в данном решение это 14 ближайших соседей), и наконец, $\hat{s}_{i,c}$ — новый ряд, который показывает средние траты клиентов, похожих на данного.

По данному ряду строятся следующие признаки:

- Линейная регрессия
- Среднее за последние 8 месяцев
- Среднее за все время

Данные признаки на ряду с другими подаются в xgboost. Строится несколько моделей с разной глубиной деревьев, а затем их блендинг.