Santander product recommendation

Евгений Патеха

Конкурс на Kaggle
Santander Product Recommendation

- Задача – предсказание «новых» продуктов, которыми воспользуются клиенты в следующем периоде
- Периоды:
  - Train - **13 647 309** записей - январь 2015 – май 2016
  - Test – **929 515** записей - июнь 2016
- Метрика – **MAP@7**
  public lb – 30%, private lb – 70%
- Ежемесячно «новые» продукты используют **3-3.5%** клиентов

1	**idle_speculation**	**0,0314090**
2	**Tom Van de Wiele**	**0,0313171**
3	**Jack (Japan)**	**0,0313157**
4	**yoniko**	**0,0313052**
5	**Jared Turkewitz and BreakfastPirate**	**0,0312650**
6	**In Public Leaderboard We Trust**	**0,0311988**
7	**Evgeny Patekha**	**0,0311658**
8	**Alejo y Miro**	**0,0311227**
9	**raddar & Davut**	**0,0311167**
10	**colun**	**0,0310650**

Клиентские признаки

22 признака

- пол, возраст, провинция, страна резидентства, семейный доход
- дата появления нового клиента, «стаж» клиента в банке в месяцах, флаги нового клиента (первые 6 месяцев) и активности
- канал привлечения клиента, сегмент (ВИП, обычные, студенты)
- категории клиентов по отношению к банку (работники банка, акционеры)

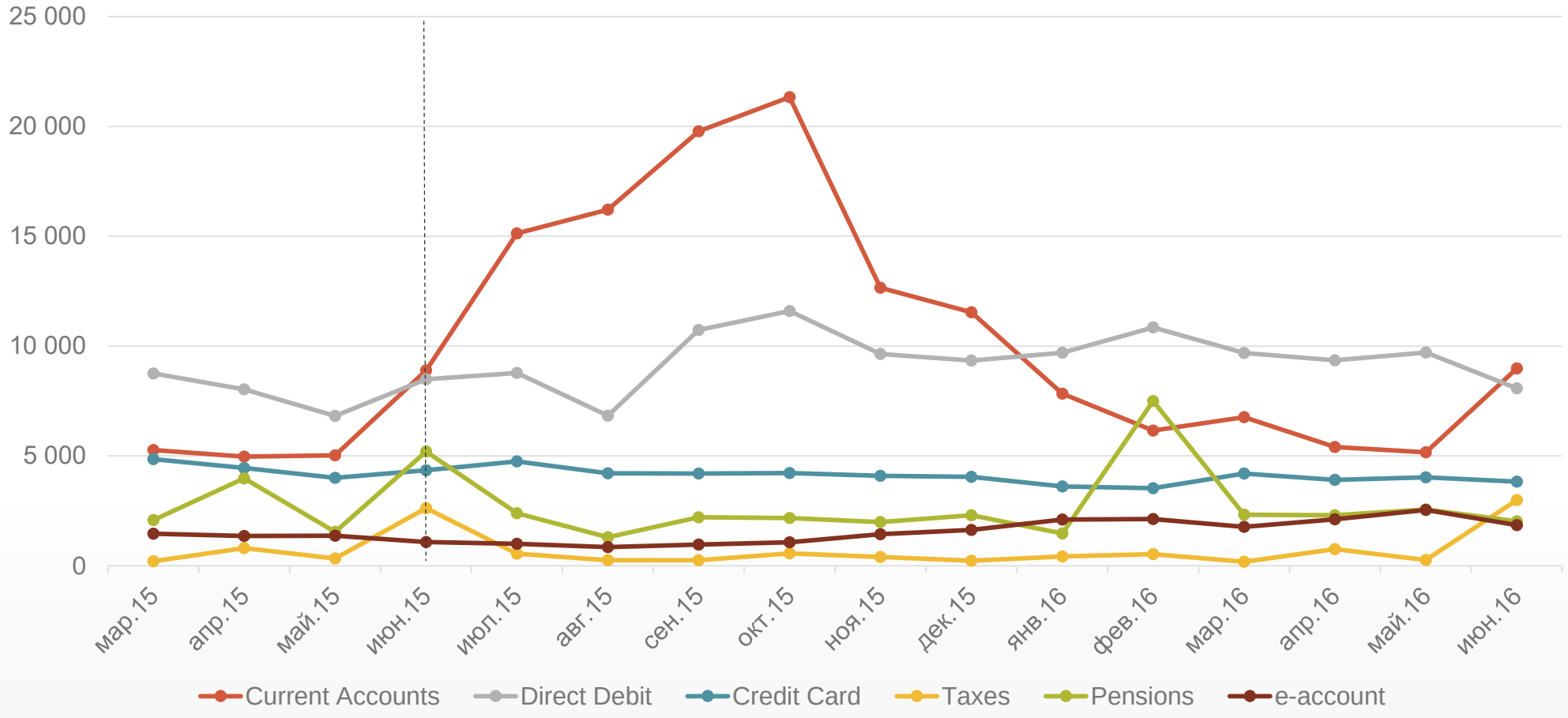
Продукты

24 продукта со значениями 1 или 0

- Различные виды счетов (текущий, зарплатный, электронный, специальные)
- Депозиты, кредиты, финансовые инструменты, налоги
- Активные операции в периоде (использование кредитных карт, банковские переводы, поступление зарплат, пенсионные взносы)

1	ind_cco_fin_ult1	Current Accounts
2	ind_recibo_ult1	Direct Debit
3	ind_tjcr_fin_ult1	Credit Card
4	ind_reca_fin_ult1	Taxes
5	ind_nom_pens_ult1	Pensions
6	ind_nomina_ult1	Payroll
7	ind_ecue_fin_ult1	e-account
8	ind_cno_fin_ult1	Payroll Account
9	ind_ctma_fin_ult1	Mas particular Account
10	ind_valo_fin_ult1	Securities
11	ind_ctop_fin_ult1	particular Account
12	ind_ctpp_fin_ult1	particular Plus Account
13	ind_fond_fin_ult1	Funds
14	ind_dela_fin_ult1	Long-term deposits
15	ind_ctju_fin_ult1	Junior Account
16	ind_hip_fin_ult1	Mortgage
17	ind_plan_fin_ult1	Pensions
18	ind_pres_fin_ult1	Loans
19	ind_cder_fin_ult1	Derivada Account
20	ind_viv_fin_ult1	Home Account
21	ind_deco_fin_ult1	Short-term deposits
22	ind_deme_fin_ult1	Medium-term deposits
23	ind_ahor_fin_ult1	Saving Account
24	ind_aval_fin_ult1	Guarantees

Анализ - востребованность новых продуктов



Анализ. Сезонность и тренды. Аномалии

- Июнь 15 – значительный всплеск по налогам
- Июнь 15 - декабрь 15 – открытие текущих счетов превышает долю банковских переводов (direct debit)
- 2016 – рост популярности e-accounts
- 2016 – резкое снижение востребованности депозитов – средне и краткосрочных – до нуля, долгосрочных – более чем в 10 раз
- Всплески «ind_nomina_ult1» (Payroll) и «ind_nom_pens_ult1» (Pensions) в апреле15, июне15, феврале16

Подготовка данных

- Перевод дат и текстовых признаков в integer
- Заполнение пустых признаков по данным из других периодов
- Удаление строк с пустыми признаками
- Ликвидация «календарных» провалов по продуктам «ind_nom_pens_ult1» (Pensions) и «ind_nomina_ult1» (Payroll) в марте и мае 15 и январе 16

ind_nom_pens_ult1	фев.15	мар.15	апр.15	май.15	июн.15	янв.16	фев.16	мар.16
не используется	26 440	26 095	25 630	25 431	24 726	20 415	20 738	20 442
продолжают использовать	39 033	39 510	40 861	39 044	39 474	38 993	37 158	45 846
перестали использовать	4 213	**5 241**	3 207	**7 523**	2 976	**13 054**	4 777	4 043
новые	5 718	4 558	**5 706**	3 406	**8 228**	2 942	**12 731**	5 073

Модели. Технические параметры

- Мультиклассовый XGBoost
  - 20 классов (19 популярных продуктов плюс хвост)
  - eval_metric = "mlogloss”, "merror“
- Только строки с «новыми» продуктами
- В случае с несколькими новыми продуктами у одного клиента, строки дублировались

Модели

- «Базовая» модель 1 - декабрь15 - январь16
- «Базовая» модель 2 - июнь15, декабрь15 - январь16
- «Налоговая» модель – июнь15
  - Прореживание в новых продуктах длинных депозитов, фондов и хвоста
- Модель для e-account - февраль16 – май16
  - Определение вероятности использования e-account и для строк выше задаваемого порога увеличены вероятности e-account в базовой модели
- Модель для новых и «пустых» клиентов - ноябрь15 – май16
  - Обучение на строках новых клиентов либо без продуктов в предыдущем периоде

Новые клиентские признаки

- Объединены в один признаки, характеризующие тип клиента по отношению к банку (работник банка, акционер, супруга работника и др)
- Добавлены признаки – флаги изменения канала и сегмента по сравнению с предыдущим периодом, 5 лаговых индексов активности
- Канал привлечения клиентов перекодирован в соответствии с частотой использования
- Удаление незначимых признаков (даты, адрес, флаг и дата праймари-клиента)

Новые продуктовые признаки

- Признак продуктов предыдущего периода, соединенных вместе в один показатель «10001000..» и закодированный в соответствии с частотой
- Признаки использования продуктов в прошлых периодах (max)
- Признаки использования продуктов с лагом 1-5 периодов назад
- Признаки изменений по продуктам 2 и 3 месяца назад (1-2, 2-3)
- Сумма месяцев использования продуктов за предыдущие 5 месяцев (sum)

Постпроцессинг

- Перенос «ind_nom_pens_ult1» (Pensions) перед «ind_nomina_ult1» (Payroll) в случае если модель предсказывала наоборот
- Удаление продуктов, встречавшихся в предыдущем периоде и редких продуктов

Валидация

- Первичный отбор признаков, параметров xgboost – обучение январь16 - апрель16, валидация - май16
- Имитация сезонной модели – апрель15 - апрель16
- Кросс-валидация

Вычислительные ресурсы, ПО и оптимизация

- Ноутбук - 2 ядра i5, 12 GB RAM
- R пакеты data.table, xgboost, Metrix

- После первичной обработки данные были разделены на клиентскую и продуктовую части
- Добавление новых показателей производилось для каждого месяца отдельно, результаты сохранялись в отдельных файлах .rds
- Перед обучением клиентская и продуктовая части объединялись по выбранным периодам

Другие решения – 1 место – Idle_speculation

- Объединение 12 моделей NN и 8 моделей GBM
- GBM – весь набор данных, 17 классов для новых продуктов (последний класс – нет нового продукта). Разные периоды для обучения, разный набор признаков.
- NN – предсказание наличия продукта (не важно новый или нет). Разные периоды для обучения, разный набор признаков, разные сиды.
- Прогноз строился 3 раза – тестовые данные прогонялись с разным признаком даты - как июнь15 (для налогов), декабрь16 (для тек счета) и июнь16 (остальное)
- Нормализация вероятностей в соответствии с probe public leaderboard
- Объединение моделей - взвешивание по результатам public leaderboard

Другие решения – 4 место - yoniko

- обучение на всех периодах, только строки с новыми продуктами
- месяц и год записи как признаки для вылавливания сезонных трендов
- **взвешенная во времени средняя использования продуктов**
- лаги продуктов, категорийных признаков, флаги изменения категорийных признаков
- custom objective function for XGBoost – “multi:map”
- сильная сторона – модель не заточена под конкретный месяц

Интересное в других решениях

- Все в Top-3 использовали вероятности для налогов из июня15 и для текущего счета из декабря15
- Использование полного набора данных (против данных только с новыми продуктами) и отдельного класса для ситуации отсутствия новых продуктов в мультиклассовых моделях
- Полезные признаки - число периодов, прошедших с момента последнего изменения по продукту
- Нормализация вероятностей перед объединением моделей
- Создание данных для обучения путем сэмплинга строк по каждому продукту до целевой пропорции

Спасибо за внимание!