Предсказание потребительского поведения на основании данных о совершенных транзакциях по банковским картам

Настоящее исследование состоит из 2 частей,

- в 1 части будут описаны подходы к классификации и организации пространства признаков потребительского поведения, применимость отдельных признаков классификации под задачу на примере признаков, которые возможно извлечь из данных о платежах по банковским картам.
- во 2 части будет описано решение конкретной бизнес задачи поведенческого скоринга, по построению инструментов предсказания дефолта по кредитной карте на основании данных о транзакциях, предшествовавших выдаче.

Часть 1
Принципы классификации признаков потребительского поведения

1.1. Виды потребительского поведения

Модели потребительского поведения можно классифицировать на 2 основные группы:

- поведение по привычке
- поведение в изменениях

Поведение по привычке - это та часть наших действий и расходов, которая связана с обычными / ежедневными потребностями, о которых мы не задумываемся (покупаем продукты, хозтовары, ходим в кино, слушаем музыку). Эти действия формируются под влиянием уже сложившейся структуры нашей личности, обычаев и привычек. Изменение состава потребляемых товаров и услуг несущественно и происходит:

- в рутине в основном за счет внешних факторов (например: реклама, общественное мнение, и т п.) ;
- (реже) в рамках программ "волевого" самостоятельного изменения образа жизни - но это уже другая ветвь классификации)

Поведение в изменениях предполагает наличие (одного или нескольких) следующих отличительных признаков:

- решение о покупке отражает
  - либо личное стремление потребителя к изменению образа жизни (например: покупка недвижимости)
  - либо инструмент выхода из кризисной ситуации (например: решение о кредите)
- решение о покупке, как правило, принимается редко
- решение о покупке критически важно для потребителя,

(и главное) …

в ходе подготовки к решению о покупке человек меняется в комплексе, совершает необычные покупки (например начинает себя баловать) и изменяет структуру расходов под новый образ жизни.

Далеко не всегда различение между поведением в привычках и поведение в изменениях проходит по сумме покупки, например:

- покупка первого автомобиля в семье - 'модель изменений", а покупка автомобиля каждые 3 – 5 лет - есть обычное предсказуемое решение, подтверждающее сложившиеся привычки,

- решение о прохождении обучения и освоении новой профессии не всегда стоит дорого, но требует существенной готовности к изменениям в образе жизни.

- получение кредита на погашение другого кредита – есть обычная сделка оптимизации расходов, не повлияет на платежное поведение,

… и тому подобные по сути различия

2. Подходы к определению моделей предсказания поведения

При построения моделей машинного обучения в отношении представленных типов потребительского поведения мы увидим следующие существенные различия:

Модели привычек:

При анализе моделей привычек мы сталкиваемся регулярно повторяющимися похожими действиями в отношении одной (или нескольких смежных) группы товаров / услуг, которые рассматриваются как целое обособленная часть образа жизни, Все события окружающего мира, происходящие и фиксируемые одновременно, воспринимаются как контекст или ассоциативный ряд для исходного события. Способ осуществления покупки является неотъемлемой частью привычки, поэтому вопрос как и в каких условиях осуществлена покупка не менее важно, чем что куплено.

Дополнительно следует отметить, что: для предсказания текущих расходов достаточно видеть метрики в отношении отдельных (одной или нескольких) товарных групп, базовую модель можно построить на основании данных одной торговой точки или сайта. предметом исследования является каждая покупка / чек.

Для предсказания в рамках моделей привычек применимы простые статистические метрики для оценки вариативности обычного поведения, а также ассоциативные правила / рекомендательные системы для предсказание перетока спроса между схожими товарными группами.

Модели изменений:

Предсказание «необычных расходов» основан на анализе «созревания» клиента, как личности в комплексе, до критически важного решения. Ключевым отличием данного типа моделей (предсказания) поведения является то, что мы предсказываем событие которое (как правило) никогда раньше не совершалось на основании прошлых качественно других событий в жизни этого клиента, выступающих отдельными индикаторами изменений в поведениии - фичами модели.

Применение Моделей изменений требует информации об изменении структуры потребительского поведения в целом по индивиду. Источников такой информации два, это прежде всего:

- информация о фактически произведенных расходах (в наличии у обслуживающего банка / возможно процессингового центра);

- информация о посещенных интернет сайта (в наличии у интернет провайдера) - является менее надежным источником информации, потому что отражает только намерения, неясного уровня существенности.

Предметом исследования в данном случае будет структура расходов в целом по клиенту и динамика изменения такой структуры. Ключевыми метриками таких моделей будут структурные метрики, такие как:

- временной ряд долей расходов клиента, разложенный по категориальному признаку или

- структура расходов за период, отнесенная к средней, обычной структуре расходов клиента.

В Моделях изменений способы осуществления (разных) покупок являются гораздо менее значимыми метриками, отражающими текущий контур привычек, обычный социальный и платежный статус, часто выступает дополнительным шумом

Модели изменений уже нельзя обьяснить простыми статистическими метриками, рекомендательные системы товаров, которые ранее никогда не покупались не работают. Здесь необходимы более сложные инструменты, основанные на градиентном бустинге, многомерных временных рядах.

1.3. Нормализация данных в моделях изменений

При анализе изменений при построении моделей машинного обучения мы используем данные в целом по рынку / платежной системе. Неизбежно столкновение с большим количеством абсолютно независимых участников наблюдений, у каждого из которых набор потребительских предпочтений индивидуален. Структура потребления и размер доходов различаются по регионам и социальным группам. И вместе с тем, нам нужно сравнить десятки или сотни тысяч моделей поведения сильно различающихся между собой.

Для решения такой задачи предлагается использовать в качестве ключевой метрики в Моделях изменения поведения следующий показатель:

$$\frac{\text{Расходы за период по категориальному признаку}}{\text{Средняя сумма расходов за аналогичный период в прошлом}}$$

При использовании показателя, в отношении каждого категориального признака каждого участника наблюдений мы получаем:

- за отчетный период – коэффициент значимости каждого вида расходов, взвешенный к обычной значимости этой статьи для конкретного клиента

- за последовательность периодов – изменение значимости по виду расходов «вокруг единицы» / средней;

- в целом по клиенту за последовательность периодов - комплексную нормированную картину / матрицу изменения потребительских предпочтений.

Такие матрицы гораздо проще классифицировать и выявлять закономерности потому, что в них не содержится индивидуальных моделей поведения, а содержится только показатели отклонений в поведении индивида от его обычного и привычного образа.

1.4. Отбор категориальных признаков

Разберем актуальность признаков поведения в Моделях изменений на примере базы данных транзакций по пластиковым картам

Главный классификатор	Метрика		Сопровождающие интуиции
	Содержание	Признаки	
Цель	назначение платежа	mcc	Ключевой параметр классификации, определяющий вектор поведения / цель события
Значимость	сумма платежа	amnt	сумма определяет субъективную важность / значимость события
Актуальность	срок с даты платежа до даты заявки на кредит	days_before	Чем больший период прошел с даты транзакции, тем меньше он повлиял на выдачу и использование карты
Обстоятельства | способ	Банковский продукт	operation_ type, operation_ kind, ecommerce, payment system	предпочтения клиента по банковским продуктам
Обстоятельства | время	Время и периодичность платежей		Предпочтения клиента по дням недели и времени осуществления покупок, возможно ключевым датам и праздникам
Обстоятельства | место	Место платежа	Country, city	

Ключевым показателем в любой модели описания поведения на основании транзакций будет цель и значимость платежа. Использование mcc является весьма удобным инструментом классификации вида потребительских расходов. классификация уже есть, описана весьма подробно и непротиворечиво и позволяет в целом сложить комплексную картину о потребительских предпочтениях индивида https://mcc-codes.ru/code.

Вопрос актуальности платежа во многом зависит от цели исследования, чем значимее изменение ,тем больший период анализа необходимо включать в исследование. Так, например, предполагаю, что для анализа решения о предстоящем получении ипотечного кредита период для анализа может достигать 12-18 месяцев, для решения о получении кредитной карты - 1-3 месяца.

При определения актуальности платежа следует принимать во внимание, что в отношении значительной части клиентов / индивидов справедливы утверждения:

- структура доходов и расходов в целом стабильна на интервале 1 месяца (30-31 любых последовательных дней по календарю)

- структура (совсем обычных) расходов на быт и питание условно стабильна еженедельно / на интервале 7 последовательных дней по календарю

Значит агрегирование расходов на месяц или неделю (в зависимости от целей исследования) способно сгладить динамические различия в структуре расходов, при наименьших потерях в качестве модели.

Обстоятельства платежа:

- важны для моделей предсказания действия, аналогичного многократно повторявшимся действиям в прошлом;

- (часто) являются шумом для предсказания действий, которые ранее никогда не совершались (или аналогичные по типу действия невозможно идентифицировать в датасете).

Поэтому признаки обстоятельств платежа в моделях изменений могут быть использованы при условии, что значения таких признаков следует:

- агрегировать в логике, отличной от описанной выше;

- обособить в отдельную модель, используемую для предсказания в рамках ансамбля моделей

В итоге:

Базовой метрикой, представляющей полную картину изменений в потребительском поведении предлагаю использовать показатели:

Расходы за период по mcc кодам

Средняя сумма расходов за аналогичный период в прошлом

Преимуществами данной метрики, являются:

- метрика способна прямо или косвенно описать любое изменение в потребительских предпочтениях индивида на верхнем уровне анализа ;

- метрика исключает шумы, образующиеся за счет различий в потребительском поведении и платежном статусе многих индивидуальных потребителей;

- классификатор целей расходов уже написан, практически используется, данные размеченные по классификатору уже есть в учетных системах банковский карт за длительный период в прошлом

- в отношении данных по метрике корректно применять методы StandartScaler, UnderSampling, что было недопустимо в отношении базовых данных датасета.

Использование указанной метрики позволило повысить точность предсказания в модели поведенческого кредитного скоринга с 78,6 до 95+% ,

но об этом уже в части 2.

Часть 2

Задача поведенческого скоринга

Предсказание выхода клиента в дефолт исходя из истории транзакций по банковским картам

Настоящее исследование основано на задаче, представленной "Альфа Банк" АО (TOP10, S&P BB+) в рамках открытого чемпионата по анализу данных AlfaBattle 2.0 проводимого на сайте Boosters.pro.

Состав исходных данных датасета

В условии задачи представлены данные об истории транзакций по банковским картам клиентов, представивших заявку на получение кредитной карты(в том числе сумма и назначение платежа ,обстоятельства платежа подробно описанные в части 1 настоящей статьи), а также отметка о попадании кредитной карты в дефолт в будущем (12 мес). Поставлена задача рассчитать вероятность дефолта.

Информация представлена в отношении 270 450 тыс. транзакций, осуществленных 963.8 тыс. клиентов, В отношении состава и полноты представленных данных необходимо отметить:

-информация о транзакциях представлена за период от нескольких дней - до 1 года (по каждому клиенту период разный),

- выборка наблюдений несбалансированна в плоскости различения хороших и плохих наблюдений (доля плохих наблюдений - 2,7 %)

Предварительный анализ данных и общие логики отбора переменных в модель.

В качестве базового признака модели нами принята следующий синтетический признак:

Расходы за период по mcc кодам

Средняя сумма расходов за аналогичный период в прошлом

С учетом условий продукта "Кредитная карта", которые состоят в следующем:

Кредитная карта – это продукт, подразумевающий возможность Клиента:

- быстро получить небольшую сумму денег (3-5 среднемесячных доходов)
- без контроля потратить
- не гасить тело кредита достаточно долго (только при нарушении грейс периода)

решение получения кредитной карты (как правило) стихийно и формируется быстро.

В составе модели принято решение о следующих метриках и учете фактора времени в составе исходной формулы:

- период транзакций для расчета среднего уровня и структуры расходов - 3 месяца
- период транзакций для расчета характерного уровня и структуры расходов -1 месяц
- если клиент в течение 30 дней до заявки не осуществлял транзакции по карте – информации для анализа недостаточно.

Такой подход позволил ограничить пространство признаков до лаконичной комплексной картины изменений потребительского поведения клиента за последние месяцев, представляющий комплексный срез структуры расходов в одной плоскости.

Построение модели.

В рамках решения поставленной задачи мы применили следующие подходы:

а) SGD Classifier

пайплайн решения состоял из следующих этапов:

- нормализация данных с помощью инструмента StandartScaler библиотеки Scikit learn,

- обучение модели при помощи инструмента SGDClassifier библиотеки Scikit learn, обеспечивающий обучение линейных моделей с применением стохастического градиентного спуска. в рамках обучения применены следующие настройки: функция потерь - 'huber', настройки learning rate - optimal, с применением l1 регуляризации.

Результат обучения модели на кросс валидации (5 fold) по метрике ROC-AUC - 97.19%

б) MLP Classifier

пайплайн решения состоял из следующих этапов:

- нормализация данных с помощью инструмента StandartScaler библиотеки Scikit learn,

- обучение модели при помощи инструмента MLP Classifier библиотеки Scikit learn, в рамках обучения применены следующие настройки: solver -sgd, настройки learning rate - invscaling, init = 0.0005, rate = 0.001.

Результат обучения модели на кросс валидации (5 fold) по метрике ROC-AUC - 97.18%

Основные выводы:

Применение подходов организации метрик потребительского поведения, основанных на

структуре расходов индивида в разрезе mcc кодов на горизонте в 1 месяц в сопоставлении со средней индивидуальной структурой за более длительный период,

способно представить ясную комплексную картину состояния и динамики потребительского поведения,

которая просто детектируется и классифицируется в рамках традиционных моделей машинного обучения, основанных на логистической регрессии и стохастическом градиентном спуске.

В частности, применение подходов позволило увеличить точность моделей поведенческого кредитного скоринга

с уровня лучших практик / победителей соревнования AlfaBattle 2.0 - на уровне 78.6 %,

до уровня 97.19% , достигнутых с применением описанных выше подходов.

Пути совершенствования модели

Основными направлениями развития модели мы видим сегодня:

- повышение эффективности модели за счет имплиментации в модель традиционных фичей кредитного скоринга, в части социальных метрик;

- создание и развитие моделей основанных на косвенных признаках социального профиля, содержащихся в составе информации о транзакциях (например: количество посещенных стран и городов за период, склонность к покупкам в

определенные дни и часы, интенсивность использования отдельных банковских продуктов) в сочетании с реклассификацией категориальных признаков (в случае если таковая информация будет в наличии), ансамблирование моделей косвенных признаков поведения с базовой моделью настоящего исследования;

- применение рекуррентных нейронных сетей на основе представленной системы метрик в рамках более широких задач предсказания потребительского поведения.

Ноутбук
https://colab.research.google.com/drive/1gypAskBkEQy9mQmkM4KFLJRaTLDvUlft?usp=sharing