Райффайзен БАНК

Анализируй и предсказывай:

спрогнозируй размер будущих доходов Райффайзенбанка от сотрудничества с клиентами

CRM-системы банков собирают большой объем информации, которая используется для анализа бизнес-процессов. В рамках работы над кейсом вы получите часть этих данных и решите одну из самых актуальных для Райффайзенбанка задач — предсказание размера будущих доходов от сотрудничества с клиентами. Точный прогноз не только поможет сформировать лучшие персонализированные клиентские предложения на рынке, но и позволит повысить качество среднесрочного планирования, открывая новые горизонты для развития банка.

dgtl.raiffeisen.ru

Оглавление

Введение

Артем посмотрел на офисные часы. В Москве без двадцати три, а значит, скоро начнется встреча, которую назначил Дмитрий, руководитель CRM-отдела. Обсуждать собирались новую задачу. Подробностей Артем еще не знал, но уже догадался, что проект будет амбициозным, ведь на встречу пригласили несколько старших коллег, которые занимаются анализом данных.

В Райффайзенбанк Артем устроился после стажировки Raiffeisen Evolve чуть больше года назад. Он довольно быстро адаптировался к жизни в компании, хотя поначалу многое казалось непонятным. Взять хотя бы фреймворк Scrum, который активно используют его коллеги. Сотрудники разбивались на небольшие команды и занимались одним-двумя масштабными проектами, строго планируя свою загрузку. И хотя Артем работал как Data Scientist в технологическом стартапе, но там такой системы распределения задач не встречал. Тем не менее он довольно быстро ощутил преимущества Scrum-фреймворка. Четкое разделение обязанностей, постоянный контроль качества, продуманные системы тестирования — все это позволяло в короткие сроки получить лучший результат.

В рамках Scrum-фреймворка Артем работал над созданием модели для автоматического формирования персонализированных клиентских предложений — Next Best Offer. И несмотря на то, что он присоединился к команде только на этапе тестирования, ему удалось получить опыт настройки параметров модели. На этот проект Артема позвала Оля — его ментор. Она работала над Next Best Offer в качестве ведущего специалиста и предложила привлечь молодого сотрудника. Оля, если верить спискам, тоже должна быть на этой встрече.

Артем никак не мог понять, что это за новая задача, которую они все вместе будут обсуждать. От построения гипотез его отвлекло уведомление календаря: 10 минут до встречи, пора идти. В переговорной он увидел Аню — коллегу из отдела финансового контроля, с которой они познакомились еще во время отбора на стажировку.

— Привет, давно не виделись! Я даже не знал, что ты тоже будешь на этой встрече.

— Привет! А как же без меня? Этот проект очень важен для нашего отдела. Мы хотим повысить качество финансового планирования, и для этого нам нужны специалисты по Data Science. Надо научиться предсказывать, сколько денег каждый клиент принесет компании за определенный период — год или два, а лучше даже больше.

— Это сколько же данных понадобится для такого прогноза? Точно нужна будет информация по демографическим характеристикам, возможно, кредитная история...

Артем уже собирался продолжить длинный список данных, но в комнату вошли несколько человек. Среди них была и Оля, которая махнула ему рукой в знак приветствия. Все сотрудники сели за стол, и руководитель начал обсуждение.

— Коллеги, всем привет! Начну сразу с вводных по проекту. В ближайшие 1–2 месяца мы планируем разработать и внедрить в бизнес-процессы алгоритм, который будет предсказывать будущие доходы от работы с клиентами. Пока горизонт планирования — 6 месяцев, но потом мы хотим адаптировать модель для долгосрочных прогнозов. Оля, пожалуйста, расскажи коллегам о данных и инструментарии.

— Да, конечно. Данные для построения пробной версии модели мы должны получить сегодня. Это будет информация о клиентах за 2018 год, которая находится в CRM-системе. В качестве функционала ошибки и критерия сравнения алгоритмов мы будем использовать среднюю абсолютную процентную ошибку (Mean Absolute Percentage Error, MAPE). Она дает простую интерпретацию качества полученных решений.

Оля закончила, и руководитель снова взял слово.

Многие проекты Райффайзенбанка работают по японскому принципу «Сю Ха Ри», который заключается в переходе с точного следования правилам на полную импровизацию.

— Спасибо. Добавлю еще немного вводных. Ограничений по алгоритмам у нас нет, но для нас важно не переусложнить решение. За моделью должна прослеживаться финансовая логика, иначе мы не сможем защитить решение на собрании руководителей направлений банка. Теперь предлагаю обсудить проблему более детально. У кого-нибудь есть вопросы?

Сотрудники начали активно обсуждать рабочие гипотезы в основе модели, и встреча переросла в дискуссию. Артем даже не заметил, как время вышло и встреча закончилась. Он вернулся на свое рабочее место, открыл ноутбук и сразу увидел оповещение о новом письме. Его прислал Дмитрий, руководитель отдела.

Тема: Прогнозирование будущих доходов Райффейзенбанка

От кого: Дмитрий Федоров

Кому: team@raiffeisen.ru

Коллеги, добрый день!

В 2019 году мы успешно реализовали проект Next Best Offer. Теперь мы хотим пойти дальше и начать предсказывать значения CLTV[1] для клиентов — физических лиц.

Вы получите набор данных о клиентах банка за 2018 год, который разделен на тренировочный и тестовый сеты. **В течение недели вам необходимо построить на полученных данных модель, прогнозирующую значения CLTV по клиентам на второе полугодие 2018 года, и оформить основные выводы в формате презентации.**

Чтобы решить эту задачу, вам нужно:
1. Выделить наиболее подходящие для анализа признаки.
2. Проработать гипотезы о взаимосвязи различных признаков с целевой переменной с помощью методов статистического и графического анализа.
3. Построить модели, проверить их качество, используя среднюю абсолютную процентную ошибку как метрику точности, и выбрать наилучший вариант.

Итоги работы нужно оформить в формате презентации. В ней должно быть описание пространства признаков и найденные взаимосвязи, сравнение различных алгоритмов с указанием финальной спецификации модели и выводы из полученных результатов.

Учитывайте, что ваша модель должна быть сбалансирована по трем критериям:
- **Точность.** Значение ошибки должно быть допустимым. Задача минимум — превзойти уровень точности, достигаемый наивным алгоритмом[2]. Этот компонент наиболее важен, так как позволяет повысить качество среднесрочного планирования и упростить работу с клиентами.
- **Сложность.** Логика модели должна быть настолько простой, насколько это возможно. При прочих равных одиночные модели, построенные только на внутренних данных, будут предпочтительнее ансамблевых методов.
- **Масштабируемость.** Модель можно будет применять для прогнозирования CLTV на различные периоды с сохранением достаточной точности.

Вы не ограничены в выборе алгоритма для обучения модели и можете использовать любые открытые данные (данные Росстата, колебания курсов валют и т. д.).
Важно: использование данных, появившихся после 2018 года, запрещено.
Успехов!

С уважением,
Дмитрий Федоров, руководитель отдела Customer Relationship Management

Ответить Переслать

[1] CLTV (Customer Lifetime Value, пожизненная ценность клиента) — доход, который принесет банку конкретный клиент за определенный период.
[2] Наивный алгоритм — алгоритм, прогнозирующий значение показателя в следующем периоде равным текущему.

О компании

О компании

История банка

Помогайте другим, чтобы помочь себе.

Ф. В. Райффайзен

Фридрих Вильгельм Райффайзен (1818–1888) был мэром нескольких деревень Вестервальдского района Германии в середине XIX века и делал все возможное, чтобы облегчить жизнь крестьян. Он начал с создания благотворительных кооперативов, но вскоре осознал, что христианские принципы благотворительности недостаточно эффективны, и переключился на организованную взаимопомощь, чтобы достичь поставленной цели. В 1862 году Райффайзен создал первый банковский кооператив в Анхаузене (Германия), который и стал прообразом банков Райффайзен.

В 1872 году для уменьшения финансовых рисков и обмена информацией Райффайзен объединил кредитные союзы в региональный кооперативный кредитный союз, а в 1877 году открыл центральный офис. Символом организации стали две скрещенные лошадиные головы. Этот знак по старинному обычаю прикреплялся к фронтонам крыш и защищал обитателей от бед. К моменту смерти Райффайзена в 1888 году в Германии существовало 425 созданных им обществ, в том числе около 120 из них — в Австрии.

Фридрих Вильгельм Райффайзен не был героем или революционером, однако в Австрии практически в каждом городе и деревне есть площадь Райффайзена или улица его имени. В честь него назван мост через Рейн, в городе Вайербуше находится музей Райффайзена. И конечно, его имя всегда будет ассоциироваться с организацией, которую он создал для того, чтобы помогать людям.

Структура банка

Райффайзенбанк впервые открылся в Австрии, потом его филиалы появились в Восточной Европе, а сейчас подразделения банка существуют в 28 странах. В Райффайзенбанке в России работает более 9 тыс. сотрудников, от Калининграда до Владивостока банк насчитывает 180 отделений.

В РОССИИ У БАНКА ШЕСТЬ ОТДЕЛЬНЫХ НАПРАВЛЕНИЙ:

1. Финансовая дирекция.
2. Дирекция по управлению рисками.
3. Дирекция обслуживания физических лиц и малого бизнеса.
4. Дирекция обслуживания корпоративных клиентов и инвестиционно-банковских операций.
5. Дирекция по оформлению и учету банковских операций и сопровождению бизнеса.
6. Дирекция информационных технологий.

Customer Lifetime Value

Артем решил не откладывать новую задачу и первым делом позвонил Ане, чтобы окончательно разобраться в деталях.

— Аня, привет еще раз. Есть пара минут? У меня к тебе несколько вопросов о прогнозируемом параметре.

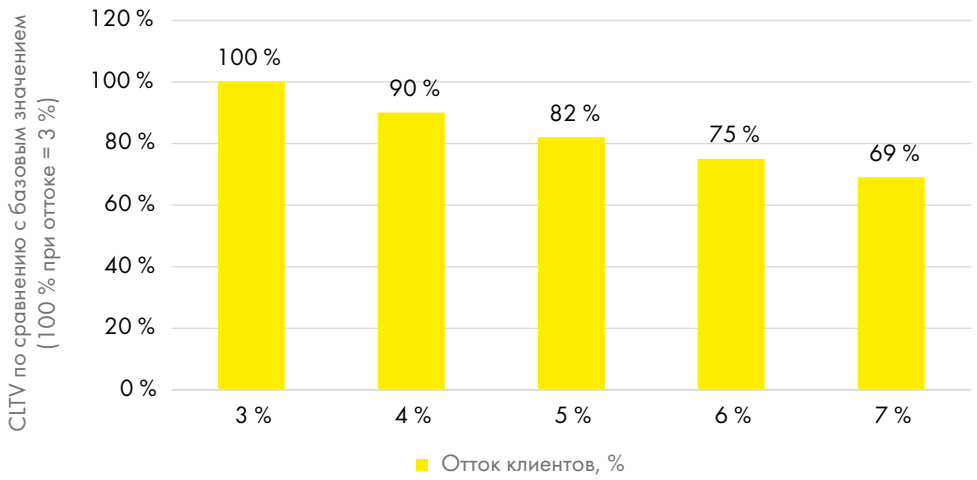
— И тебе привет! Да, время есть.

— Можешь поподробнее рассказать о CLTV?

— Да, конечно. Пожизненная ценность клиента (Customer Lifetime Value, CLTV) — ключевая бизнес-метрика при составлении стратегии по привлечению и удержанию клиентов. Основной смысл — в ответе на вопрос: «Какой ожидаемый доход принесет клиент за время работы с банком?» Если нужно оценить CLTV в целом, не за определенный период, то надо понять две вещи: как может меняться средний доход от сделок с клиентом и сколько продлится сотрудничество. Расскажу тебе о каждой.

Начнем с первого фактора. Допустим, есть компания с фиксированной стоимостью услуг — журнал или фитнес-клуб. Им довольно просто оценить доход от клиента, поскольку стоимость не меняется. У нас сложнее: приходится учитывать дополнительные условия. Клиент в любой момент может закрывать счета и открывать новые или менять сумму на каждом из счетов. Все это происходит с некими вероятностями, причем отличающимися для разных людей.

График 1. Пример зависимости CLTV от оттока клиентов



Райффайзенбанк изучает и применяет блокчейн. Заключены уже три крупные сделки с использованием технологии распределенных реестров, последняя — на проприетарной платформе R-chain.

Со вторым фактором тоже все непросто. Я сейчас скину тебе в чат пример графика, секунду... Смотри, на нем видно, насколько сильно увеличение ежегодного оттока аудитории снижает прогнозируемый CLTV. То есть учитывать это надо. Но как понять, уйдет ли конкретный клиент, если у нас есть только средние значения показателя оттока?

— Для этого вы нас и позвали, да?

— Да, вам нужно построить модель, которая предскажет CLTV с учетом всех упомянутых факторов. Причем не только в целом, но и по каждому клиенту. Для нас в финансовом контроле больше важен общий прогноз, но ребятам из маркетинга ваша модель очень бы пригодилась.

— Ясно. А что с периодом прогнозирования?

— В качестве горизонта планирования мы установили 6 месяцев, поскольку длительные прогнозы не всегда точны. Но будет здорово, если вы сможете предсказать на больший срок достаточно достоверно для нашей области применения.

— Так, кажется, я все понял. Огромное спасибо!

— Тебе спасибо, будем ждать новостей!

Райффайзенбанк проводит внутренние хакатоны (Cozy Hack), где команды банка в течение суток придумывают новые сервисы для внутренних сотрудников. Самые полезные проекты реализуются и внедряются в жизнь банка.

Применение Machine Learning в банкинге

Артем так глубоко погрузился в новый проект, что не заметил, как наступил обеденный перерыв. Если бы не Оля, он так бы и остался сидеть за ноутбуком.

— Время для небольшой паузы, — Оля была в отличном расположении духа. — Давай в кафе, а по пути расскажу тебе о небольшой задачке, которую я сейчас делаю.

Выяснилось, что Оля занимается подготовкой тренинга для стажеров. Она попросила Артема помочь ей написать текст о работе специалистов по Data Science в банке, и вместе они быстро сделали вводную часть.

В офисах Райффайзенбанка есть блиндозеры — автоматы, которые пекут блинчики.

«Банки, хранящие огромное количество информации о клиентах, быстро осознали преимущества анализа данных. Одной из первых задач в области статистического анализа стал кредитный скоринг — определение вероятности невозврата клиентом выданного кредита. В конце XX века решение о выдаче кредита практически полностью принимал искусственный интеллект, что повысило точность и скорость прогноза. Это позволило снизить сумму покрытия рисковых операций, благодаря чему освободилось больше средств для развития бизнеса. Сейчас скоринг — по-прежнему одна из главных задач машинного обучения в банках, причем его применение стало шире и теперь включает не только классификацию физических лиц, но и предсказание вероятности дефолта организаций, в частности банков партнеров.

Банки применяют искусственный интеллект и для менее специфичных для отрасли задач. Например, маркетинговые компании планируют заранее обученные модели, которые могут предсказать практически все — от наиболее релевантных для банка каналов продвижения до наилучшего времени проведения рекламных мероприятий. Одна из подзадач в этой области — оптимизация воронки продаж, которая включает формирование персонализированных предложений для клиентов. Например, для развития инвестиционной деятельности банков в клиентские приложения интегрируют робо-эдвайзеры — алгоритмы, которые генерируют инвестиционный портфель на основе данных от клиента. Подобные решения снимают часть нагрузки с персонала и оптимизируют структуру расходов банка.

Кроме того, сейчас активно развивается NLP (Natural Language Processing, анализ естественного языка), который все чаще применяется в банках. Он позволяет создавать умных помощников — приложения, которые отвечают на несложные вопросы клиента без привлечения сотрудников банка. NLP — ключевая часть алгоритмов распознавания речи, которые в перспективе можно использовать для замены колл-центров, освободив значительную часть средств, выделяемых на телемаркетинг».

В офисах Райффайзенбанка в любое время можно поиграть в кикер, PlayStation, настольный теннис и хоккей.

Обзор основных алгоритмов машинного обучения

После разговора с Олей Артем снова погрузился в проект.

«Итак, что мы имеем? У нас есть набор данных, и на его основе нужно спрогнозировать значение определенной переменной — дохода, который принесет клиент в следующем периоде. Исторические значения CLTV нам тоже доступны. То есть мы имеем дело с обучением с учителем, а конкретно с задачей регрессии — прогнозом числа. Теперь надо понять, какие методы мы можем использовать».

Артем начал вспоминать основные семейства алгоритмов машинного обучения.

Каждый из объектов выборки обладает набором признаков, иначе говоря, располагается в признаковом пространстве.

ЛИНЕЙНЫЕ МОДЕЛИ

Линейные модели представляют значение зависимой переменной в форме линейной комбинации признаков. Самый распространенный пример — метод линейной регрессии.

$$Y=X\times\beta+\varepsilon$$

где X-матрица признаков, β-коэффициенты модели, ε-случайная ошибка

Линейная регрессия не требует больших вычислительных ресурсов и легко интерпретируется, поскольку признаки получают коэффициенты. Но она проигрывает в точности моделям, учитывающим нелинейные связи, и ее применение требует соответствия данных предпосылкам, изложенным в теореме Гаусса — Маркова. В частности, признаки (данные по клиентам) должны быть линейно независимыми, а ошибка ε должна носить случайный характер, то есть описываться нормальным распределением со средним 0.

К линейным моделям также относится метод опорных векторов (Support Vector Machine, SVM). Он основан на построении в признаковом пространстве размерности N гиперплоскости размерности N-1, максимально удаленной от объектов каждого из классов. Его преимущество перед другими алгоритмами семейства заключается в умении учитывать нелинейные зависимости при описании разделяющей гиперплоскости. Однако результаты работы алгоритма сложно интерпретировать и он вычислительно затратен.

НЕЙРОННЫЕ СЕТИ

Нейронные сети появились как попытка воспроизвести устройство человеческого мозга. Так, самая первая нейронная сеть — перцептрон — имитировала работу нейрона, который, получая сигналы (электрические импульсы) от других нейронов, мог послать сигнал дальше или остановиться. В случае перцептрона сигналами от нейронов выступают признаки, которые суммируются с определенными алгоритмом весами, после чего результат передается активационной функции, в качестве которой выступает пороговая функция, преобразующая число в 0 или 1 в зависимости от того, отрицательное оно или положительное.

Райффайзенбанк активно сотрудничает с финтех-стартапами и в деле проверяет новые технологии. За 2017–2019 годы банк запустил 50 пилотов с российскими и зарубежными стартап-проектами.

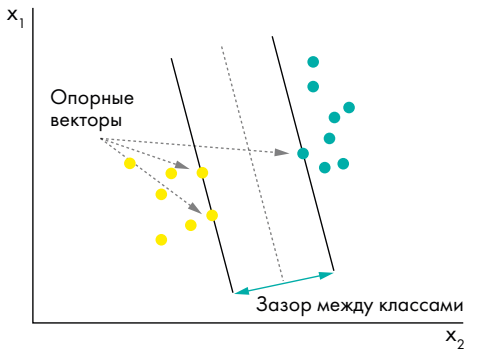


Рисунок 1. Разделяющая плоскость SVM в случае двухмерного пространства

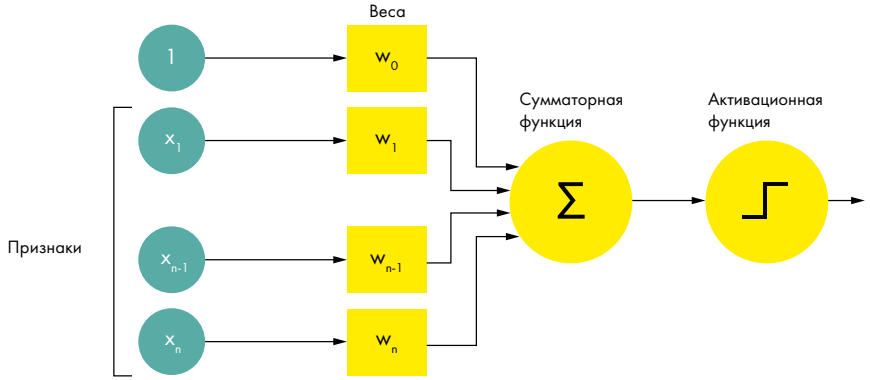


Рисунок 2. Принцип работы перцептрона

Современные нейронные сети содержат скрытые слои. Их легко представить как связки из перцептронов, которые в качестве признаков принимают сигналы от других перцептронов и так далее, в зависимости от числа скрытых слоев. Это позволяет лучше учитывать нелинейные связи и эффекты взаимодействия. Практика показывает, что нейронные сети лучше всего работают на больших наборах данных (> 105 объектов), а со значительным увеличением выборки будут показывать наилучшую точность. Но из-за сложности структуры модели будет невозможно интерпретировать влияние признаков на нейронную сеть.

МЕТРИЧЕСКИЕ МОДЕЛИ

Метрические модели основаны на вычислении расстояния между объектами выборки в признаковом пространстве. Под расстоянием предполагается любая метрика близости двух векторов признаков, например Евклидова метрика.

$$\text{Euclidean}(a,b)=\sqrt{\sum(a_i-b_i)^2}$$

Считается, что чем ближе друг к другу располагаются объекты, тем выше вероятность, что они принадлежат к одному классу. Тогда новому объекту мы присваиваем класс, который есть у большинства его ближайших соседей. Из различных спецификаций алгоритма (для числа соседей от 1 до n) мы выбираем ту, которая дает наибольшую точность. Так мы получаем простейшую метрическую модель — метод k ближайших соседей. Для задачи регрессии — прогнозирования непрерывной переменной — все несколько сложнее: в основу берется параметрическая модель (линейная регрессия), а расстояние между объектами определяет вес, с которым учитывается ошибка прогноза по каждому из объектов.

Данный тип моделей подвержен «проклятию размерности»: с увеличением числа признаков требуемый размер выборки увеличивается экспоненциально, и это сказывается на вычислительных затратах при их использовании. Метрические модели преимущественно применяются для задач с небольшим количеством признаков.

РЕШАЮЩИЕ ДЕРЕВЬЯ

Решающие деревья разделяют выборку на подмножества, которые называют листьями, таким образом, чтобы после разбиения уменьшилась энтропия информации — мера неопределенности при классификации объектов в каждом из них. У деревьев есть целый ряд преимуществ перед другими моделями: данный метод не выдвигает требований к переменным, одинаково хорошо работает с задачами бинарной и многоклассовой классификации, вычислительно прост и может распараллеливаться, что еще сильнее повышает скорость обучения. Однако деревья часто переобучаются. На рисунке 4 изображена синусоида с небольшим шумом. Дерево глубины 5 слишком подстраивается под шум, из-за чего при добавлении новых объектов, соответствующих функции синуса, будут возникать ошибки. Это называется переобученностью на шум, и по данной причине в современных алгоритмах деревья крайне редко используются поодиночке.

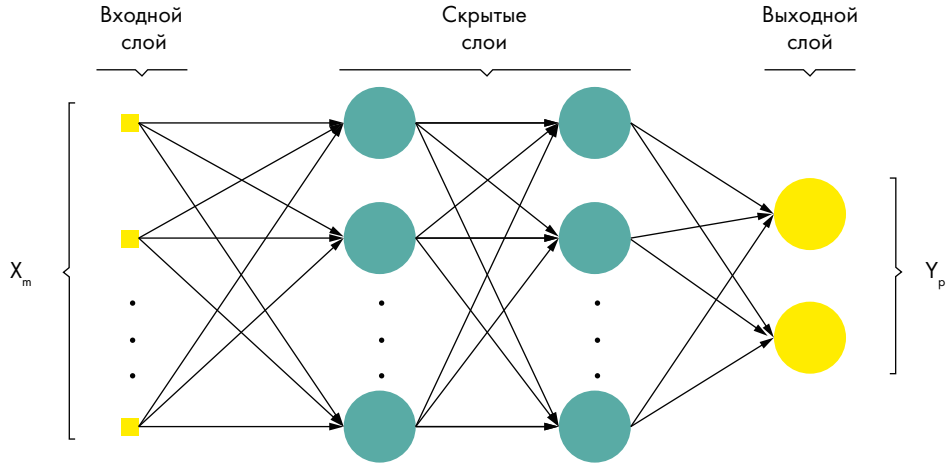


Рисунок 3. Искусственная нейронная сеть

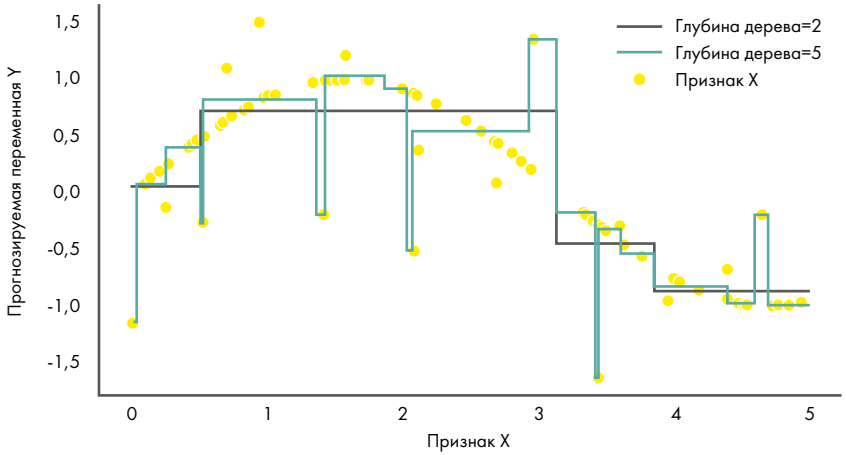


Рисунок 4. Использование решающего дерева в задаче регрессии

Компромисс между сложностью и точностью моделей

Артем пошел на кухню налить себе кофе, но мысли о задаче не отпускали. Теперь он обдумывал критерии отбора алгоритмов для прогноза CLTV. На кухне сидела Оля в компании нескольких стажеров и рассказывала о тонкостях реализации алгоритма Next Best Offer. Она заметила Артема и спросила:

— Ну как дела с работой над моделью?

— Пытаюсь понять, дадут ли классические алгоритмы достаточную точность предсказания, — ответил Артем. — Хотя пока не проверял, данные еще не собраны полностью.

— Может, одиночные алгоритмы и не дадут, — серьезно заметила Оля. — Думаешь использовать ансамбли?

— А что, ансамбли дадут лучшую точность? — задал вопрос один из стажеров.

В Raiffeisen DGTL работают около 1200 сотрудников, среди них 120 Agile-команд.

— Вполне. При построении моделей всегда есть компромисс между смещением и ошибкой, или bias-variance tradeoff. Ошибка прогноза модели состоит из трех компонентов: шума, смещения и дисперсии. Шум мы не можем устранить — это просто отклонения от реального процесса генерации данных. Смысл машинного обучения — оптимизировать смещение и дисперсию при прогнозе. Это непросто, потому что смещение снижается с увеличением сложности модели, а дисперсия падает. Отсюда и проблема недообученности и переобученности моделей.

У **недообученных алгоритмов** высокое смещение и низкая дисперсия. Они плохо улавливают зависимости в данных, из-за чего прогнозы получаются недостаточно точными. Пример — решающее дерево глубины 1, то есть разбивающее выборку только на два подмножества. Даже если выбрать оптимальное разделение, алгоритм проводит только одну разделяющую гиперплоскость, чего недостаточно для большинства задач.

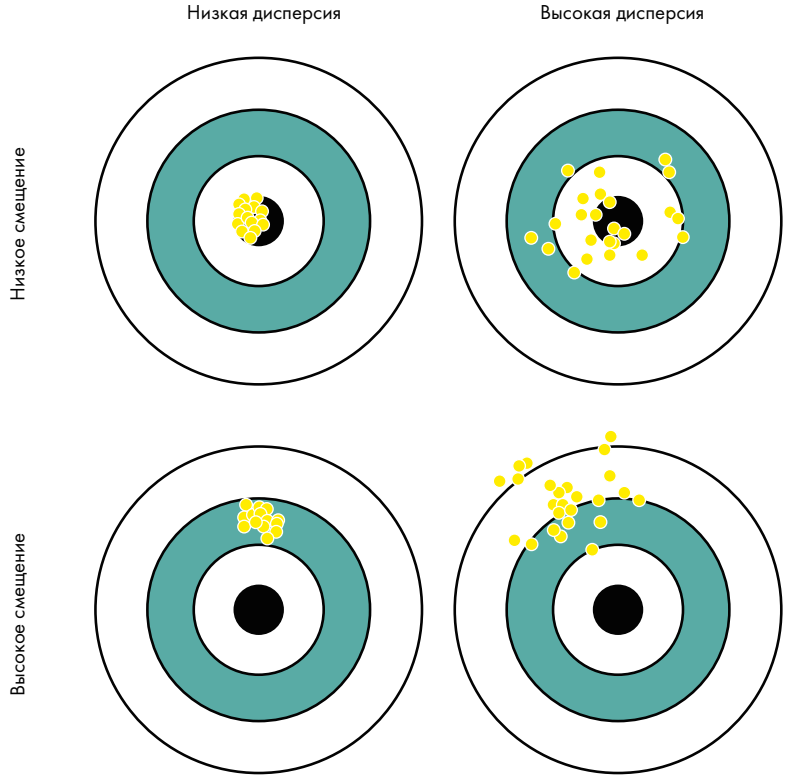
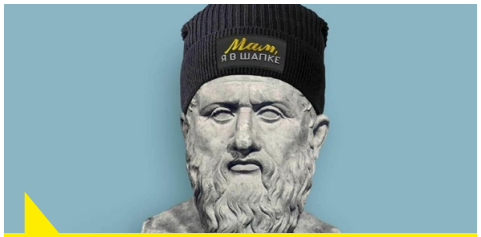


Рисунок 5. Различные модели в зависимости от смещения и дисперсии



В ноябре в некоторых офисах банка поставили банкомамы — банкоматы, которые выдавали теплые шапки в холодную погоду.

Переобученные алгоритмы характеризуются низким смещением и высокой дисперсией. Они настолько хорошо выучивают зависимости в данных, что начинают обучаться на случайный шум. С ростом сложности модели, например с увеличением глубины дерева, смещение снижается, поскольку прогнозы становятся все ближе к реальным значениям искомой переменной. Однако подобная модель все хуже воспроизводит реальный процесс генерации данных, из-за чего точность на тестовой выборке становится низкой. Именно по этой причине алгоритмы **регуляризируют**: вводят ограничения на глубину деревьев, включают коэффициенты линейных регрессий в функции потерь и так далее.

— И ансамбли решают эту проблему?

— Да, именно попытки найти компромисс между смещением и дисперсией привели к созданию **ансамблевых методов**. Суть данного подхода — в совмещении множества одиночных алгоритмов, деревьев или линейных регрессий, с целью получения сильной модели, дающей лучшие прогнозы, чем любой из алгоритмов, на которых она основана. Самые распространенные ансамблевые методы основаны на одном из двух подходов: бустинге или бэггинге.

- Бэггинг — параллельное обучение слабых моделей на подмножествах объектов и признаков, создаваемых методом бутстрэпа . Результат выдается путем голосования, то есть выбора класса, предсказанного большинством моделей. Бэггинг позволяет сильно уменьшить дисперсию ошибки: в лучшем случае в n раз, где n — число слабых алгоритмов в основе модели. Наиболее известным алгоритмом, основанным на бэггинге, является Random Forest, в котором одновременно обучаются десятки и даже сотни решающих деревьев.

- Бустинг — последовательное обучение моделей, каждая из которых обучается с учетом ошибок, совершенных на предшествующей итерации. Результаты улучшаются с увеличением числа моделей, однако бустинг в отличие от бэггинга имеет склонность к переобучению, поэтому данный алгоритм нужно постоянно контролировать на тестовых данных.

В meeting-room'ах Райффайзенбанка можно рисовать на стенах.

Один из стажеров вопросительно посмотрел на Олю и уточнил:

— Зачем вообще использовать одиночные модели, если ансамблевые методы всегда дают меньшую или такую же ошибку?

Артем улыбнулся.

— На самом деле у этого подхода тоже есть пара недостатков. Ансамблевые модели трудно интерпретировать. Проще всего показать это на сравнении решающего дерева и Random Forest. Обычное дерево дает графическое представление результатов, из которого понятно разбиение выборки в каждом листе. Визуализация случайного леса ничего не даст: слабые модели выделяют различные признаки для разбиения выборки, потом ошибка усредняется и непонятно, где конкретно в деревьях использовался признак.

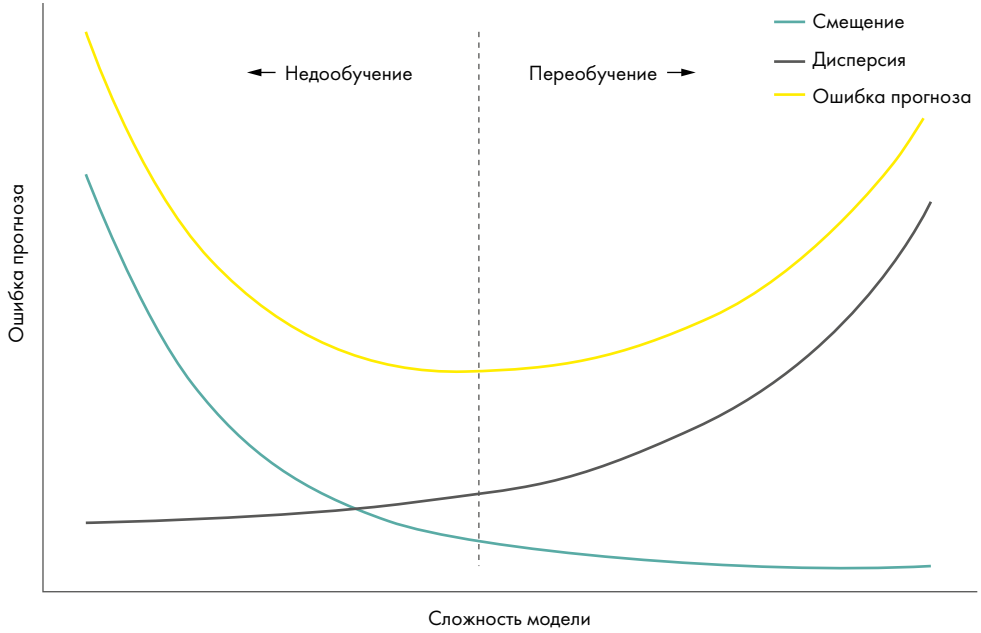


График 2. Компромисс между смещением и ошибкой

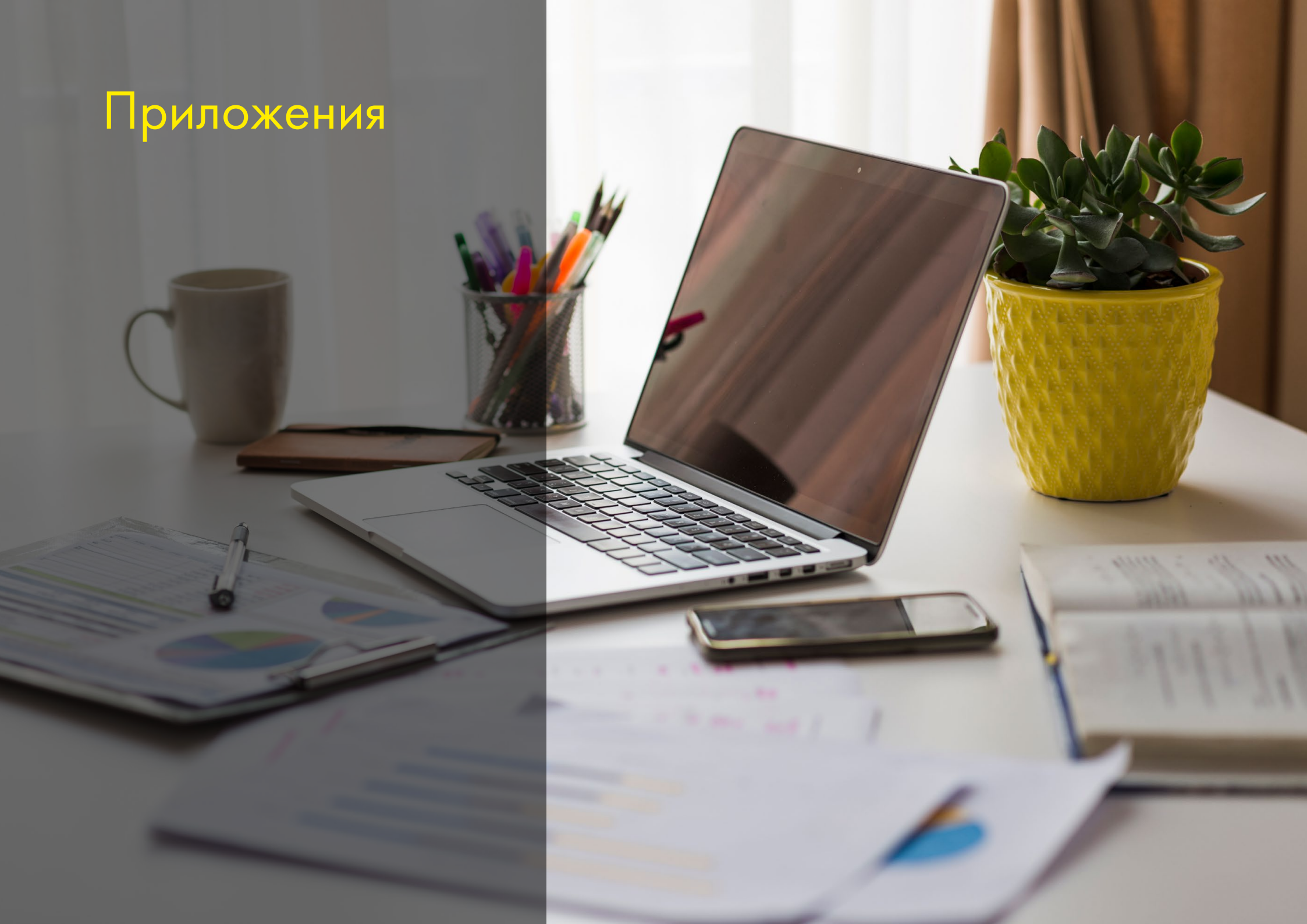
Райффайзенбанк — единственный банк, который, несмотря на экстремальные условия, семь лет эффективно работал на острове Беринга с населением около 700 человек.

— А еще ансамблевые алгоритмы являются вычислительно сложными, — добавила Оля.

— Например, бустинг требует **последовательного** построения **большого** числа моделей (> 100), вследствие чего даже простота выбранных алгоритмов не гарантирует высокой скорости работы. У нас большинство данных потоковые, каждый раз пересчитывать модель на них долго. Хотя в случае с CLTV может сработать, как считаешь?

— Понятия не имею, если честно, — сказал Артем. — Тут вопрос практики. Попробуем и поймем, что будет лучше.

Приложения



Приложение

Форма представления результата

Ответ на задачу кейса должен включать три документа:

- Работающая модель в формате .py или .ipynb, которая принимает файл CUP_IT_test_data.csv из корневой папки и записывает в эту же папку файл CUP_IT_predictions.csv, состоящий из двух столбцов: cif_id (id клиента) и cltv (прогноз Customer Lifetime Value на второе полугодие 2018 года).
- Файл CUP_IT_predictions.csv
- Презентация в формате .pdf, в которой необходимо ответить на вопросы о возможном использовании данных и построенной модели.

Файл с лучшей моделью, презентацию и файл CUP_IT_predictions.csv необходимо до 6:00 17 марта отправить на адрес first.round@changellenge.com.

Критерии оценки решений

Итоговый балл выставляется по следующей схеме:

- 75 % — оценка модели. Прогнозы модели оцениваются по критерию «Средняя абсолютная процентная ошибка» (Mean Absolute Percentage Error, MAPE). Все решения ранжируются по шкале от 0 до 5, где 0 — отсутствие решения, 5 — качество модели выше 80-го квантиля среди присланных решений (модель входит в 20 % лучших по качеству).
- 25 % — оценка презентации. Презентация оценивается по пяти критериям: широта анализа, глубина анализа, структура решения, соответствие поставленной задаче и качество презентации. Решения получают оценку от 0 до 5 по каждому из критериев, после чего балл усредняется.

Описание данных

Название признака	Описание признака
cif_id	Уникальный номер для клиента
dlk_cob_date	Временная метка данных в формате YYYY-MM-DD
gi_smooth_3m	Доход, который банк получил от данного клиента за данный месяц
big_city	Город клиента (укрупненный до вариантов Мск/Спб., миллионник, другой город)
cu_gender	Пол клиента
cu_education_level	Уровень образования клиента (может быть неактуально)
cu_empl_area	Область работы клиента (может быть неактуально)
cu_empl_level	Уровень должности клиента (может быть неактуально)
payroll_f	Флаг того, что клиент получает з/п на карту банка
cur_quantity_pl	Количество персональных кредитов
cur_quantity_mort	Количество ипотечных кредитов
cur_quantity_cc	Количество кредитных карт
cur_quantity_deposits	Количество депозитов
cur_quantity_dc	Количество дебетовых карт
cur_quantity_accounts	Количество счетов
cur_quantity_saccounts	Количество накопительных счетов
cur_quantity_mf	Количество инвестиционных продуктов
cc_balance	Баланс кредитных карт
cl_balance	Баланс автокредитов
ml_balance	Баланс ипотеки
pl_balance	Баланс кредитов
td_volume	Баланс депозитов
ca_volume	Баланс счетов

Название признака	Описание признака
sa_volume	Баланс накопительных счетов
mf_volume	Баланс инвестиций
dc_cash_spend_v	Объемы трат с дебетовых карт наличными
dc_cash_spend_c	Количество трат с дебетовых карт наличными
cc_cash_spend_v	Объемы трат с кредитных карт наличными
cc_cash_spend_c	Количество трат с кредитных карт наличными
dc_pos_spend_v	Объем POS-транзакций с дебетовых карт
dc_pos_spend_c	Количество POS-транзакций с дебетовых карт
cc_pos_spend_v	Объем POS-транзакций с кредитных карт
cc_pos_spend_c	Количество POS-транзакций с кредитных карт
ca_f	Флаг наличия текущего счета
rc_session_qnt_cur_mon	Количество сессий в онлайн-банке
cur_qnt_sms	Количество полученных СМС
active	Является ли клиент активным по определению банка
standalone_dc_f	Флаг отдельной дебетовой карты
standalone_payroll_dc_f	Флаг отдельной зарплатной карты
standalone_nonpayroll_dc_f	Флаг отдельной НЕ зарплатной карты
salary	Зарплата клиента
cu_age	Возраст клиента
cu_mob	Сколько клиент уже в банке
cu_empl_cur_dur_m	Сколько клиент работает на данной работе (может быть неактуально)
is_married	Состоит ли клиент в браке

Как стать частью команды Raiffeisen DGTL?

РАЗВИВАЕМ ТАЛАНТЫ

Вкладываемся в молодых специалистов и помогаем получить лучший профессиональный опыт, внедряем лучшие практики мира ИТ, уделяем особый фокус на T-Shape. Регулярно запускаем bootcamp'ы для студентов (Java, SAS & DWH, QA, Mobile Dev и др), а также проводим годовую программу развития.

ИННОВАЦИИ =

Работаем с финтех-стартапами и проверяем в деле новые технологии. За 2017-2019 годы запустили 50 пилотов с российскими и зарубежными стартап-проектами. Изучаем и применяем блокчейн: провели три крупные сделки, последнюю — на проприетарной платформе R-chain.

С IT НА «ТЫ»

Используем OpenSource-решения и практики DevOps, дружим с Agile и Scrum, внедряем LeSS-фреймворк. В течении года мы выступаем на конференциях, проводим хакатоны, открытые митапы и Demo-дни, где IT-команды показывают то, что только вышло в прод или даже еще не вышло, но скоро выкатится.

УСЛОВИЯ –

Мы хотим быть компанией, где каждый может найти интересные задачи и получить комфортную среду для своего развития. У нас есть программы поддержки внутреннего предпринимательства, а также инфраструктура, в которой любой сотрудник может проводить эксперименты в виртуальной среде.

Что такое Raiffeisen DGTL? Это сообщество разработчиков, тестировщиков, аналитиков, дизайнеров, инженеров и всех-всех-всех, кто стоит за цифровыми решениями банка. В команде Raiffeisen DGTL работает почти 1200 сотрудников: 900 — в Москве и 300 — в омском IT-хабе Raiffeisen TechCenter.

Райффайзенбанк меняется — мы уходим от образа классического банка, применяя digital-решения и ориентируясь на потребности клиентов. Для этого нам нужны таланты, которые мыслят технологиями, быстро реагируют на изменения, готовы двигаться вперед, искать новое и развиваться вместе с нами.

В течение года мы проводим различные программы развития, bootcamp'ы, хакатоны, конференции, открытые митапы и Demo-дни — через них ребята попадают на стажировки и на постоянные позиции. 3 из 4 стажеров остаются в штате по окончании программы.

Следи за открытыми вакансиями, анонсами bootcamp'ов и программами на наших каналах:

ВКонтакте - vk.com/raiffeisencareer

Digital-портал - dgtl.raiffeisen.ru

Страница на hh.ru - hh.ru/employer/4023

Changellenge

Кейс написан и опубликован
Changellenge >> —
ведущей организацией
по кейсам в России.

www.changellenge.com
info@changellenge.com
vk.com/changellengeglobal
facebook.com/changellenge

Raiffeisen BANK

Кейс создан по заказу
АО «Райффайзенбанк»

www.raiffeisen.ru
info@raiffeisen.ru
facebook.com/raiffeisenbankrus
vk.com/raiffeisenbankrus
twitter.com/Raiffeisen_Ru
instagram.com/raiffeisenbank_rus