№	Содержание	Вопросы по кейсу
Кейс № 1	Решается задача multilabel-классификации отзывов клиентов, оставленных в чате приложения.Multi-label - это как задача многоклассовой классификации, только объект может обладать сразу несколькими различными классами одновременно. Разработчиком в выборках были оставлены только те наблюдения, которые принадлежат наиболее важным в бизнес-задаче классам (одно наблюдение - один класс). Multilabel таргет был сформирован путем проставления единичного флага для класса к которому принадлежит наблюдение и нулевого для остальных (для одного наблюдения таргет представляет собой вектор вида [0, 1, 0, 0], в общем случае для multilabel задач такой вектор может содержать отличное от одного число единичных флагов: [0, 0, 0, 0], [1, 1, 0, 0], и т.д.). Применение модели выглядит следующим образом: • Тексты отзывов разделяются на слова (с исключением стоп-слов) и проходят через tf-idf (метрика, которая показывает относительную частоту встречи слова в данном тексте с учетом того, как часто оно встречается в других документах, т.е. характеризует важность, релевантность слова для данного текста) • Полученные после применения модели логистической регрессии (отдельной для каждого лейбла) скоры проходят через отсечение по порогу (бинаризация прогноза), где каждый лейбл имеет отдельный порог (подобраны на этапе обучения модели).	• Как проверить качество работы такой модели? • Какими недочетами обладает такое решение? • Каким вы видите решение подобной задачи?
Кейс № 2	Разметить магазины на форматы: супермаркеты, гипермаркеты, магазины у дома. Сегменты различаются по занимаемой торговой площади. Формальных границ каждого формата неизвестны. Могут быть использованы данные эквайрингов в магазине.	• Какие методы будете применять? • Как оценить качество? • Если планируется использовать частичную разметку, с чего начать разметку? • Как будете передавать задачу на разметку? (по какой логике им нужно разметить или как им определить класс) • По каким параметрам можно грубо разбить хоть какой-то сегмент?

Кейс № 3	Арендодателю сдающему помещение в аренду необходимо знать заранее, когда его арендатор может съехать с занимаемой площади, чтобы начать поиск нового арендатора или изменить условия выгодные обоим. Для решения могут использоваться данные транзакционной активности торговых точек в Москве за 2 последних года.	• Как определить возможный съезд арендатора по транзакционной активности • Какие данные будете собирать? • Будет ли выборка сбалансированной, какие у этого последствия? Как вы будете балансировать класс? • Предложите методы решения • Как проверить результаты модель? Какие метрики качества / ошибки следует использовать для оценки результатов предсказания? • Как бы вы будете учитывать изменения ситуации на рынке, например, кризис или ковид? • Как определить факторы, ведущие к закрытию торговой точки? Какое решение можно дать арендодателю по результатам модели?
Кейс № 4	Есть необработанные данные чеков по ресторанам города Москвы за 2 года. В чеках содержится название товарных позиций, сумма, дата и идентификатор торговой точки. Стоит задача определить нишу торговой точки. Примеры ниш: грузинский, итальянский и японский рестораны.	• Какие методы будете применять? • Как оценить качество? • Если планируется использовать частичную разметку, с чего начать разметку? • Как будете передавать задачу на разметку? (по какой логике им нужно разметить или как им определить класс) • По каким параметрам можно грубо разбить хотя бы одну нишу

Кейс № 5	Определение класса мяса и брак по фотографии.	• Будет ли проблема, связанная с несбалансированной выборкой? Если да, то как решать? • С чего начать решение? Какую архитектуру вы выберете для решения этих задач? • Как можно улучшить качество модели?
Кейс № 6	В Банке есть сервис, который парсит большинство русскоязычных новостных источников в интернете (около 30 тысяч сайтов) и сохраняет тексты новостей. Предполагается, что эту информацию можно использовать для предсказания ухудшения состояния заемщиков Банка. Таким ухудшением может быть, например, банкротство или просто просрочка по кредиту. Предложите возможные идеи по построению модели (моделей), предсказывающей такое ухудшение с учетом того, что исторические данные по состоянию заемщика (финансовая отчетность, кредитная история etc) у Банка есть.	• Как бы примерно выглядел пайплайн обработки новостного потока после парсинга (сырой новостной поток из интернета -> ??? -> предикты модели). На какие события/темы в новостном потоке стоило бы ориентироваться? • Какую бы выбрали целевую переменную? • Какие бы использовали метрики для оценки качества модели?
Кейс № 7	Содержание: в Москве и в Санкт-Петербурге проходит возле метро соц-опрос любите ли вы кофе. Необходимо проверить гипотезу правда ли Москвичи любят кофе больше жителей Санкт-Петербурга?	• Необходимо формализовать задачу. • Как бы вы проводили исследование для того чтобы получить достоверные результаты? • Может ли получиться так, что провести достоверный эксперимент невозможно?
Кейс № 8	Содержание: В здании есть несколько лифтов, они работают по следующему принципу: находясь на этаже вы нажимаете номер этажа который Вам нужен и некоторый алгоритм показывает номер лифта который Вас довезет. Необходимо разработать алгоритм который оптимально бы выбирал лифт.	• Формализуйте задачу и определите что значит оптимально. • Какие данные Вам понадобятся и как бы вы могли их собирать? • С какими проблемами вы можете столкнуться при имплементации алгоритма в жизнь? • Какими бы конкретными методами вы бы решали ее?

Кейс № 9	Содержание: представим что перед банком встала задача выявлять определенный тип транзакций (для примера возьмем большое снятие наличных с карты). Перед Вами только текстовое описание транзакции. Необходимо как можно лучше решить данную задачу.	• Формализуйте задачу • Постройте дизайн эксперимента? С чего бы Вы начали? • Какие методы машинного обучения можно использовать а какие нет? • Какие метрики стоит использовать для оценки качества? • Если предположить, что за каждый выявленный случай таких транзакций присутствует некоторый штраф/поощрение. Изменится ли Ваш подход?
Кейс № 10	Есть датасет по кредитным заявкам физлиц, в котором содержатся следующие данные: • анкетные данные (персональные данные, владение автомобилями, владение недвижимым имуществом, стаж работы, занимаемая должность, вид деятельности работодателя клиента, семейное положение); • параметры кредитной заявки (тип продукта, сумма кредита, срок кредита, первоначальный взнос); • данные о доходах и расходах клиента. Требуется разработать модель, позволяющую автоматически определять, является ли кредитная заявка аномальной. Данных о том, какие заявки аномальные, нет.	• Какова бизнес цель модели? • Как можно подойти к решению задачи в случае отсутствия разметки? • Как можно использовать информацию о наличии разметки для части датасета и как можно решить задачу в случае наличия разметки для всего датасета? • Какие метрики качества можно использовать в случае, когда есть таргет
Кейс № 11	Сейчас во всех приложениях для заказа такси есть функционал предложения оптимального места для посадки, который показывает, в какую точку лучше всего вызвать такси. Представьте, что вам необходимо этот функционал реализовать.	• Как решать такую задачу? • Зачем могла потребоваться реализация такой функции? • Какие данные необходимо собрать? • Как измерять качество реализованной модели? • Как можно улучшить качество ее предсказания?

Кейс № 12	Представьте, что для магазина приложений на смартфоны вам необходимо обучить модель, которая бы предсказывала, какое приложение хочет установить пользователь, когда он открывает ваш магазин.	• Как решать такую задачу? • Какие данные необходимо собрать? • Что делать, если о пользователе нет информации (к примеру, он только что купил смартфон)? • Как измерять качество такой модели? • Как можно улучшить качество ее предсказания?
Кейс № 13	Дана задача предсказания цен жилой недвижимости в новостройках. Целевая переменная – относительная разница цены аналога с фактической ценой объекта. Данные – из открытых источников	• Дать подробное описание этапов подготовки данных – источники, признаки, таргет. • Какую метрику выбрать? • Как отбирать признаки? • Работа с пропусками и аномальными значениями в данных. • Какие алгоритмы подойдут? • Проверка качества модели.
Кейс № 14	У банка есть данные о текущем расположении банкоматов в городе. Есть задача расширения сети банкоматов, для этого необходимо построить модель поиска наиболее эффективного расположения	• Как вы думаете, что является оптимальным расположением. • Опишите алгоритм построения модели с применением графовых методов и без. • При построении модели какие признаки нам необходимо собрать (например, характеристики места, проходимость…)