1. **Методами машинного обучения (не статистическими тестами) показать, что разбиение на трейн и тест репрезентативно.**
   Ответ: Вопрос неоднозначный так как непонятно, что понимается под репрезентативностью разбиения. С одной стороны если в test будут те же элементы, что и в train - такое разбиение нерепрезентативно. С другой стороны если в test будут те же признаки из другого распределения, то такое разбиение тоже нерепрезентативно! Таким образом нужно чтобы разбиения train и test имели примерно одинаковое распределение при том, что отличались элементами. Репрезентативность разбиения можно оценить, например, с помощью кросс-валидации и тренеровки какой-либо модели машинного обучения на фолдах. Тут прийдется резать train на подвыборки размера test и проводить кросс-валидацию на каждом таком куске. Если выбранные метрики похожи на выбранные метрики той же модели при кросс-валидации на кусках train и test, то скорее всего разбиение репрезентативно. Не так важно, какую модель выбирать. Наверное лучше брать легкие и слабые модели типа линейной регрессии для задачи регресии и SVM с линейным ядром для классификации/кластеризации.

2. **Есть кластеризованный датасет на 4 кластера (1, 2, 3, 4). Бизнес аналитики посчитали, что самым прибыльным является кластер 2. Каждый клиент представлен в виде 10-тимерного вектора, где первые 6 значений транзакции, а оставшиеся: возраст, пол, социальный статус (женат (замужем)/неженат (не замужем)), количество детей. Нужно поставить задачу оптимизации для каждого клиента не из кластера 2 так, чтобы увидеть как должен начать вести себя клиент, чтобы перейти в кластер 2.**

Ответ: Давайте разберемся чего мы хотим, чтобы корректно сформулировать задачу оптимизации. Для начала поймем что значит "самый прибыльный". Понятно, что экономисты люди адекватные и под самым прибыльным имеется в виду "средняя прибыль кластера" или "медиана прибыли клиентов в кластере". Уточняем это, чтобы не было как в поговорке про 100 друзей, где каждый друг имеет 1 рубль, но прибыль кластера из друзей самая большая в силу их большого количества.
Итак, если мы хотим от клиентов других кластеров такую же прибль, то нам необходимо оценить какие должны быть параметры. Для вораста, пола, социального статуса и количества детей достаточно положить, что в лучшем случае они будут равны медианным значениям клиентов из класета 2.

$$x_{age}^{1,3,4} \rightarrow median(x_{age}^{2})$$

$$x_{status}^{1,3,4} \rightarrow most\_frequent(x_{status}^{2})$$

$$x_{sex}^{1,3,4} \rightarrow most\_frequent(x_{sex}^{2})$$

$$x_{children}^{1,3,4} \rightarrow median(x_{children}^{2})$$

В случае для значений транзации мы уже не можем делать такие утверждения. Однако мы можем рассмотреть сходимость матрицы с информацией о транзакциях в кластерах к такой матрице для второго кластера т.е.

$$||A^{1,3,4}||_2 \rightarrow ||A^{2}||_2$$

Ну а если делать не по эвристикам, а методами машинного обучения, то можно во-первых посмотреть на важность признаков для предсказания прибыли на 2-ом кластере (сама модель обучается на всех кластерах). В качестве такой модели хорошо пойдойдет градиентных бустинг. Далее на основе важности

признаков составить лосс-функцию, которая будет представлять из себя взвешанную сумму разности соответсвующих значений признака на кластере 2 и на другом кластере и её нужно минимизировать. В формулах:

$$Loss(x_i) = \sum_{j \in [2]} ||w \cdot (x_i - x_j)||^2 \to \min_{x_i}$$

Тут x_i - это соответствующий вектор фичей из кластеров 1, 3, 4.

Также w - вектор важности соответствующих признаков.

Ключевое, что мы минимизируем функцию по x_i до по векторам фичей из кластеров 1, 3 или 4. Таким образом решая такую задачу оптимизации мы можем получить оптимальные значения x_i.

3. Что лучше 2 модели случайного леса по 500 деревьев или одна на 1000, при условии, что ВСЕ параметры кроме количества деревьев одинаковы?
Ответ: В машинном обучении непонятно, что такое "лучше". Мы можем говорить лишь о каких-то конкретных метриках и конкретных данных. В случае с числом деревьев ответ неоднозначен. Допустим у нас есть два леса с 500 деревьями. По сути с этими лесами мы можем сделать bagging и получить уменьшение дисперсии предсказаний, но качество предсказаний врядли станет лучше. Также если вовлечь третью модель, то можно сделать stacking.

С одним лесом на 1000 деревьев мы врядли что-то сделаем. Дисперсия у него будет примерно такая же, как у bagging модели. В случае очень сложных данных количество деревьев будет решать и ещё сильнее уменьшать дисперсию, чем в случае двух лесов и bagging. В целом модель с большим количеством деревьев будет более робастной в общем случае, поэтому на сложных данных я бы сказал, что лучше иметь один лес на 1000 деревьев.

4. В наличии датасет с данными по дефолту клиентов. Как, имея в инструментарии только алгоритм kmeans получить вероятность дефолта нового клиента.
Ответ: Имея в инструментарии kmeans мы имеем возможность считать расстояния между точками. При появлении новой точки в пространстве признаков мы можем найти расстояния от нее до других точек и выбрать k ближайших расстояний. Так как наши данные разделены на два класса алгоритмом kmeans (был дефолт/не было дефолта), то мы можем посчитать:

$$prob = \frac{with\_default}{k}$$

Где with_default - количество точек из k ближайших, у которых класс "был дефолт".

Для большей справедливости можно придумать какую-то взвешанную сумму, выбирая веса обратно пропорциональные расстоянию. Но тут придеться каждый раз пересчитывать коэффициент перед суммой, чтобы при условии, что with_default вероятность была равна 1.

5. Есть выборка клиентов с заявкой на кредитный продукт. Датасет состоит из персональных данных: возраст, пол и т.д. Необходимо предсказывать доход клиента, который представляет собой непрерывные данные, но сделать это нужно используя только модель классификации.
Ответ: На самом деле часто возникает потребность оценивать непрерывные величины задачами классификации. Если визуализировать таргет мы увидим, что у нас образуются большие кластеры на определенных уровнях зарплаты, например, скорее всего довольно большой кластер будут образовывать люди с з.п. 10к, 20к, 30к, 40к, ... Т.е каждые 10 тысяч

рублей. Таким образом вместо задачи регрессии мы можем поставить задачу многоклассовой классификации, рассматривая 50 классов с зарплатами от 10к до 500к. Более высокие зарплаты являются редкостью и скорее всего будут выбросами в нашей задаче. Далее обучаем модель классификации, например, градиентный бустинг на деревьях решений.

Данный подход можно считать некоторой аппроксимацией непрерывному подходу регрессии. Скорее всего точность у такого подхода будет хуже, но в каком-то смысле так мы сможем решить задачу с приемлемой точностью.