1. Методами машинного обучения (не статистическими тестами) показать, что разбиение на трейн и тест репрезентативно.

Ответ: Как вариант можно обучить классификатор следующим образом: 1 класс заполнить данными из трейна, в качестве 2 класса передать данные из теста. Если данные имеют одинаковое распределение, то их нельзя будет различить между собой. Если результаты классификации будут отличаться (будут лучше некоего константного классификатора), то значит данные в трейн и тесте отличаются, то есть не репрезентативны.

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-тимертного 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: По сути, это, наверное, задача выпуклой оптимизации с заданными ограничениями. В нашем случае этих ограничений 4. Мы знаем, к какому кластеру принадлежит клиент, то можем определить его положение в пространстве. Далее, нужно сделать так, чтобы эта точка переместилась внутрь найденной области допустимых значений. Такие изменения его вектора с транзакциями определяют его поведение, помогающее ему перейти в успешный кластер.

3. Что лучше 2 модели случайного леса по 500 деревьев или одна на 1000, при условии, что ВСЕ параметры кроме количества деревьев одинаковы?

Ответ: Наверное, все зависит от задачи и данных.
Например знаем, что с ростом деревьев растёт точность прогноза, то при прочих равных одна модель на 1000 деревьев будет давать более точный прогноз, чем две по 500.
Но с другой стороны, если мы обратимся к документации sklearn: все параметры кроме n_estimators одинаковы и нет разницы, как мы будем усреднять.

4. В наличии датасет с данными по дефолту клиентов. Как, имея в инструментарии только алгоритм kmeans получить вероятность дефолта нового клиента.
Ответ:
Проведем следующие операции:

1. Разбиваем данные по кластерам без метки дефолта
2. Затем частотным методом определяем вероятность дефолта в каждом кластере
3. Приписываем новому клиенту ближайший кластер

И тогда вероятность дефолта клиента будет равна вероятности дефолта в ближайшем для него кластере.

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.

???