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

Ответ: Предлагаю обучить классификатор на 2 класса – в качестве первого класса передать данные из трейна, в качестве 2 класса передать данные из теста. Если данные имеют одинаковое распределение, то их нельзя будет различить между собой и классификатор будет отрабатывать на уровне константного (скажет, с точность 0.7 если у нас трейн составлял 0.7 от всей выборки). Если результаты классификации будут лучше константного классификатора, то значит данные в трейн и тесте отличаются по параметрам распределения фичей и разделение не было репрезентативным

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: так как повлиять на последние 4 параметра возраст, пол, социальный статус (женат (замужем)/неженат (не замужем)), количество детей клиента нельзя, то нужно по очереди фиксируя эти значения оптимизировать первые 6 значений, отвечающие за транзакции. Например, берем человека из 1/2/3 кластера в возрастной категории 30-40 лет, мужской род, женат, детей нет.

Посмотрим:
какие значения транзакций характерны для людей из второго кластера, у которых возраст от 30 до 40 лет?
какие значения транзакций характерны для людей из второго кластера мужского рода?
какие значения транзакций характерны для людей из второго кластера, у которых социальное положение - женат?
какие значения транзакций характерны для людей из второго кластера, у которых нет детей?
Ставим задачу оптимизации с учетом полученных 4 ограничений, решаем ее стандартными способами, на выходе получаем области значений для транзакций данного индивидуума, какими они должны быть, чтобы можно было отнести его ко 2 кластеру.
Геометрически это можно интерпретировать следующим образом: мы ищем область 6 мерного пространства, в котором могут находиться значения фичей для такого сочетания 4х параметров, отличных от транзакций. Текущее значение 6мерного вектора – его положение на данный момент. 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: Две эти модели будут работать одинаково (как для случая регрессии, так и для случая классификации)

Посмотрим на параметры случайного леса, которые описаны в документации sklearn для классификации:

class `sklearn.ensemble.`**`RandomForestClassifier`**(*n_estimators=100, *, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='sqrt', max_leaf_nodes=None, min_impurity_decrease=0.0, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None*)

В условии дано, что все параметры кроме *n_estimators* одинаковы, а значит нет разницы, как мы будем осуществлять усреднение ответов – по 1000 деревьям сразу или по 500 два раза а потом два полученных ответа снова усреднять.

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

Ответ:
Можно так сделать в общем случае: у нас есть разбиение на n кластеров, по каждой точке есть информация - дефолт или нет. По каждому кластеру можно посчитать вероятность дефолта для участников данного кластера - доля дефолтных точек к общему числу точек в кластере. Потом пытаемся предсказать вероятность дефолта для новой точки: смотрим, к какому из центров ближе наша новая точка и берем вероятность по этому кластеру. Вариант определения принадлежности точки к одному из кластеров можно усложнять, например, смотреть расстояния до центров 2 ближайших кластеров и определять вероятность дефолта, обратным образом взвесив эти вероятности на расстояния.

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: у меня была похожая задача на работе – нужно было определить возраст человека по аудио с его голосом, расскажу, как решала ее.
Мы определили, сколько классов со стороны бизнес логики необходимо различать, в нашем случае – это было 3 возрастных категории: молодежь, люди среднего возраста и пенсионеры. Обслуживание этих трех возрастных категорий имеет различия, дальнейшее уточнение возраста не имеет практического смысла, а только увеличивает вероятность ошибки классификации.
Далее я объединила данные в 3 класса по метке возраст и обучила классификационную модель. В зависимости от характера данных можно выбирать вид модели, для персональных данных (возраст, пол и тд) я бы сначала попробовала catboost, так как это модель хорошо обрабатывает категориальные признаки (среди персональных данных есть признаки, для которых one-hot кодирование добавит категорию отношения для фичей, которой изначально в данных не было, например, м -> 1, ж-> 0 или наоборот).