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

Можно сделать трейн и тест равного размера (или примерно равного, 60/40), обучить модель сначала на трейне и проверить на тесте, затем вторую модель (с теми же параметрами) обучить уже на исходном тесте и проверить на трейне. Если качество работы примерно одинаковое, то и представленность, например, разных классов в выборке одинаковая.

Если же сохранять пропорции 80/20, то можно в цикле обучить, например, 50 моделей на каждом новом разбиении сета на трейн и сплит и сравнить их результаты.

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.

Поведение клиента здесь можно ограничить набором транзакций, так что задача оптимизации ставится на первые 6 значений вектора. Может быть, релевантным является меньшее количество признаков, можно это предварительно выяснить и уменьшать расстояние до них. Если n – количество релевантных признаков, то клиенты кластера 2 находятся в некотором n-мерном пространстве, и надо минимизировать расстояние от точки каждого клиента класса 1,3,4 до этого пространства.

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

В случайном лесе предсказания усредняются; если возьмем два леса по 500 деревьев, то мы возьмем их предсказания и усредним их: (D1 + D2)/2. По сути, это тоже самое, что получить предсказание одного ансамбля из 1000 деревьев, поэтому разницы нет

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

kmeans значит, что мы кластеризуем множество клиентов, то есть разбиваем их на группы по общему признаку. пусть есть два кластера – дефолт и недефолт. Для нового клиента посчитать расстояние до центра кластера 1 и кластера 2, вероятность дефолта = расстояние до дефолтного кластера / сумму расстояний

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

разбить клиентов на группы по доходу (min-1/10 max, 1/10 max-2/10max, …, 9/10 max-max), обучить модель на классификацию клиентов по группам дохода, классифицировать нового клиента. Допустим, он попал в группу 1: min-1/10 max. Теперь можно разбить уже эту группу на 10 подгрупп – (min-1/100 max, 1/100 max – 2/100 max, …, 9/100 max- 1/10 max). Далее можно опять

разбить на подгруппу и повторить. Зачем делать множество моделей вместо 1 большой – на мой взгляд, так повысится точность.

Можно, как в вопросе 2, попробовать поиграть с расстояниями от одного класса до другого. Сделать одну модель на классификацию тех же 10 классов, найти расстояние от человека с минимальным окладом (А) до человека с максимальным (Б). Для предсказания дохода клиента найдем его позицию в n-мерном пространстве, найдем его проекцию на прямую (?), соединяющую А и Б (можно ее отшкалировать по доходу, как линейку), точка, на которую он попадет, будет показывать его доход.