1. Какими методами машинного обучения можно показать, что разбиение на трейн и тест репрезентативно?

Можно применить t-тест, chi-square тест для сравнения средних и распределений в выборках.
Можно визуализировать данные и сравнить распределение признаков в выборках.

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.

Добавим в датасет столбец, указывающий принадлежность каждого клиента к конкретному кластеру (1, 2, 3, 4) в качестве таргета.

Обучим модель на этом датасете предсказывать принадлежность к кластеру используя остальные признаки в качестве features.

Проанализируем какие из признаков имеют наибольшее влияние на принадлежность к кластеру 2. Изменение или учёт этих характеристик может повысить вероятность принадлежности к нужному кластеру для клиентов.

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

Обучение и предсказание на двух моделях по 500 деревьев, вероятно, будут более быстрыми, чем на одной модели на 1000 деревьев. Это может быть важным фактором в случае больших наборов данных или при ограниченных вычислительных ресурсах.

С другой стороны один больной случайный лес может быть аггрегирован в алгебраическую диаграмму принятия решений (Algebraic Decision Diagram) методами описанными здесь https://arxiv.org/pdf/1912.10934.pdf, что существенно улучшит производительность.

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

Применить алгоритм kmeans и сформировать кластеры используя наш датасет . Количество кластеров, которые нужно найти можно указать 2 (в самом простом варианте).

После того, как алгоритм обучен на наборе данных и сформировал кластеры он назначает каждому клиенту в датасете свой кластер.

Для нового клиента можно использовать обученную модель для определения, к какому кластеру этот новый клиент будет наиболее близок. Необходимо, чтобы у нового клиента были те же признаки, которые использовались при обучении.

После того, как назначим кластер клиенту надо проанализировать данные о дефолте других клиентов из этого же кластера. Если большинство клиентов из этого кластера ранее столкнулись с дефолтом, то можно предположить, что и для нового клиента вероятность дефолта будет высокая и наоборот. Самый банальный способ - посчитать отношение дефолтных клиентов в кластере ко всем клиентам в 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.

Для прогнозирования дохода, используя классификацию, можно следовать такой схеме:
Разбиение на категории - преобразовать непрерывные данные о доходе в категории (например, диапазоны доходов).
Классификация - обучение классификатора для предсказания категории дохода.
Постобработка - вычисление предсказания дохода, как среднего или медианы диапазона предсказанной категории.