Суть алгоритма:

Для нахождения клиентов, которые могут иметь несколько SIM-карт, было решено использовать векторное расстояние между каждым из объектов (абоненты). Те абоненты, векторное расстояние между которыми наименьшее среди всех абонентов (или векторное расстояние меньше какого-то значения) могут являться одним и тем же физ.лицом и иметь несколько SIM-карт.

Векторное расстояние рассчитывалось с помощью алгоритма KNN.

Также, предварительно были обработаны имеющиеся данные:

- Группировка вида 1 строка - 1 абонент;
- Вычислены и добавлены в модель различные аггрегаты по каждому абоненту (см. скрипт);

Анализ результатов:

Среднее значение векторного расстояния между абонентами с 2 SIM получилось несколько меньше, чем среднее значение векторного расстояния между случайными абонентами (см. скрипт -график).

Среднее значение (2 SIM): 5.82
Среднее значение(Rand) *: 7.86

Однако в довольно большом количестве случаев (25-30%) векторные расстояния между случайными абонентами оказывались меньше, чем расстояния между абонентами с 2 SIM.

(*) Среднее значение было получено из 10 случайных выборок.
Для оценки статистической значимости полученных различий в ср.значениях разных групп, необходимо было проводить t-test, предварительно приведя выборки к нормальному распределению, а также убрав выбросы.

Возможное улучшение:

- В качестве выходных данных можно выдавать не абонента с минимальным векторным расстоянием, а например топ-10 абонентов, у которых оно минимально для данного абонента.
  После того, как эти абоненты отобраны, из этих абонентов отбирается максимально похожий на нашего абонента - например у него может быть минимальный временной лаг по регистрации на BS, максимально похожий набор BS, минимально различающиеся углы и т д;
- Обогащение данными - например добавить геолокацию в одно и то же время для разных абонентов;