**Цель:**

• Построить модель PD (Probability of default) в Python на основе количественных факторов по алгоритму, описанному ниже;

• Заполнить форму отчета, представленную ниже. Отчет должен содержать выводы по каждому разделу. Длина разделов и подробность изложения на усмотрение кандидата;

• Итоговый пакет должен содержать заполненный отчет и код на языке Python.

• Допускается отходить от пайплайна и представить решение любым алгоритмом.

**Данные:** файл “Data2.csv” содержит перечень количественных факторов (qn01,…qn11) и флаг дефолта (target) для выборки из 1000 клиентов. Факторы имеют разную размерность.

**Алгоритм построения модели**

**• Однофакторный анализ**

*Доступность*

• Доступность данных (Availability) рассчитывается как количество доступных данных (не NA) к общему количеству наблюдений по фактору. Доступность фактора должна быть не менее 80%. Факторы, не удовлетворяющие данному критерию, исключаются из дальнейшего анализа.

*WoE – преобразование*

**Если понятие WOE не знакомо или пункты 1.2 – 1.4 вызывают затруднение при выполнении, можно подготовить факторы для многофакторного анализа любым подходящим способом, описав его в отчете.**

• Проводится WOE-преобразование факторов. Пропуски необходимо определить в отдельный бакет, т.е. посчитать отдельное WOE-значение для пропущенных значений. Перед проведением WOE-преобразования факторы разбиваются на интервалы с помощью деревьев решений. Параметры дерева подбираются на кросс-валидации на 5 фолдах. Необходимо рассмотреть деревья с максимальной глубиной (max\_depth) не более 2 и следующими значениями минимального числа наблюдений в листе (min\_samples\_leaf): 50, 100, 150, 200, 250, 300. Выбирается параметр, максимизирующий коэффициент Джини (Gini) преобразованного фактора на кросс-валидации.

• Преобразованные факторы со значением Джини менее 10% на кросс-валидации на 5 фолдах исключаются из дальнейшего анализа.

• Оцените полученное WOE-преобразование с экспертной точки зрения. Хорошее ли разбиение факторов получилось?

**• Многофакторный анализ**

• На первом этапе проводится отбор переменных при помощи L1-регуляризации. Подбор оптимального обратного параметра регуляризации и оценка качества осуществляются на кросс-валидации с 5 фолдами. Метрика качества – коэффициент Джини.

• На оставшихся переменных проводится дополнительная корректировка коэффициентов при помощи L2-регуляризации.

• Качество полученной модели проверяется на основе следующих критериев:

- Проводится тест на значимость коэффициентов

- Оценивается ранжирующая способность модели (коэффициент Джини скора)

**Отчет о построении PD - модели**

**Раздел I. Однофакторный анализ**

**Раздел II. Многофакторный анализ**