Добрый день,

Для данной задачи классификации, мной была выбрана модель решающего дерева. Такое решение я принял исходя из объема данных, количества фичей, затруднительности (для меня лично) предобработки анонимизированных признаков. Также дерево решений не требует нормализации и относительно терпимо к выбросам.
Относительно других моделей (бустиногов и леса) оно показало большую скорость обучения и предсказаний.
Параметры подбирались через GridSearchCV.
Показатели метрик -
f1(macro) 0.71
roc_auc 0.87

Подбор порога решения достаточно сложно осуществить без понимания бизнес логики. У меня есть предположение, что в данной задаче стоит подбирать порог для максимизации чувствительности модели (sensivity, TPR) - если мы делаем предложение о подключении услуги и это предложение для компании не несет каких либо значительных затрат. В случае выдачи специальных предложений, акций и т.д. необходим более детальный анализ затрат/возврата на основе матрицы ошибок. Но без понимания бизнес процессов, я остановился на максимизации f1 метрики, как было указано в задаче.

Пару слов о структуре проекта.

Папка Features содежит основной датафрейм с признакоми клиентов оператора МегаФон. (Не добавлен в git, из-за объёма)

Папка Modules содержит ноутбук со скриптом обучения модели, саму обученную модель, датафрейм для обучения, файл с атрибутами класса генератора признаков, а так же файл с классами обработки данных (они мне понадобились в основном скрипте).

В корне лежит тестовый датафрейм, ответы - предсказания модели и основной скрипт.