**Отчёт по выполнению задания**

На первом этапе был произведен первичный анализ данных с помощью profile\_report (результат находится в DataReport.html). Исходя из полученных данных, было принято решение убрать некоторые столбцы. После этого из некоторых категориальных переменных были сделаны их бинарные аналоги (из-за пропорций принимаемых значений), а к другим категориальныим переменным было применено one-hot кодирование. Также были созданы новые признаки. После этого данные были отмасштабированы (к тестовой выборке принялись коэффициенты масштабирования, рассчитанные по обучающей выборке).

Для проведения сегментации сначало было проведено изучение оптимального количетсва класеров на основе дендограммы и правила локтя. На немасштабированных данных график в праввиле локтя получился более информативным, и в итоге данные были разделены на 13 кластеров (см. файл data\_train\_clustered.xlsx) с помощью метода К-соседей.

После этого был произведен анализ имеющихся признаков (парные красные и синие графики) для обучения, PCA и feature importance. Совсем не существенные признаки были отброшены.

В качестве моделей для тестирования были выбраны бустинг, случайный лес и лог. регрессия. Стоит отметить, что классы в обучающей выборке несбалансированы, поэтому каждая из моделей была протестирована как на исходных данных (с весами классов), так и на oversampled (с помощью SMOTENC) данных. По итогу, oversampled данные не дали улучшения в качетве прогнозов, поэтому выбор был сделан в пользую исходных данных. От обучающей выборки было оставлено 15% в качестве валидационных данных (чтобы контролировать переобучение после подбора параметров на кросс-валидации). Для каждой модели был произведен поиск оптимальных параметров с помощью GridSearch, финальные их значения оставлены в соотвествующих ячейках с моделями. Качество моделей оценивалось не только с по метрике roc auc, но также и по confusion matrix и распределению прогнозных вероятностей. Последнее является достаточно полезным — чем больше прогнозов сосредоточено вокруг 0.5, тем в большей неопределённости находится модель. В итоге, предпочтение было отдано бустингу, так как он давал самые сбалансированные результаты с точки зрения всех критериев качества.

Для получения флагов дефолта бустинг был обучен уже на всей обучающей выборке. В качестве скорингового балла было решено использовать величину (вероятность дефолта)\*(сумма выданного кредита)/(доход семьи). Бустинг хорошо классифицировал клиентов, которые не выходили в просрочку, но противополжных им классифировал не совсем точно, поэтому клиентов, по которым флаг дефолта стоит 0, надо ранжировать по скоринговому баллу и и присваивать флаг дефолта 1 тем из них, для которых скоринговый балл находится выше некоторого порогового значения.

Для повышения качества прогнозов, по моему мнению, было бы также полезно иметь размер назначенного каждому клиенту аннуитета, а также, возможно, некоторую кредитную оценку, полученную от сторонней организации. Еще, например, если возможо, то по каждому клиенту можно было бы подтянуть флаг попадания в просрочку по имеющимся или уже закрытым кредитам (чтобы было понимание предрасположенности клиента к выходу на просрочку).Помимо этого, полагаю, более точные прогнозы можно было бы получить с помощью нейронных сетей.