Отчёт по выполнению задания

Процесс выполнения задания состоял из нескольких этапов:

* препроцессинг данных
* выбор моделей и их обучение
* анализ получаемых метрик и результатов
* тестирование статистической значимости того, какая из построенных моделей оказалась лучше

На первом этапе было необходимо перекодировать бинарные признаки в 0 и 1, а к категориальным применить one hot encoding. Остальные признаки были отскалированы, дабы уменьшить в них шум и привести к одинаковому масштабу.

Пропорции классов в наборе данных 1:4, то есть датасет несбалансирован. В качестве out-of-sample данных было оставлено 15% данных (для валидации моделей), а из оставшихся 85% был сделан несбалансированный датасет, но с весами классов, а также сбалансированный датасет с помощью метода k-соседей (функция SMOTENC из библиотеки imbalanced learning). Таким образом, имеется 2 обучающих набора, чтобы в дальнейшем оценить, какой из них позволит получить более качественную модель.

В качестве моделей были выбраны экстремальный градиентный бустинг и случайный лес. Для каждой из моделей был произведен подбор гиперпараметров на каждом из двух наборов данных. Параметры, оставленные в GridSerach, являются оптимальными. Для оценки качества моделей использована метрика ROC AUC. Качество моделей оценивалось как и на кросс-валидации, так и на валидационном датасете (те 15% исходного набора данных), чтобы контролировать переобучение.

В качестве дополнительных признаков были добавлены попарные произведения переменных HasCrCard, Gender и IsActiveMember. Были изучены feature importance и PCA, чтобы исключить часть несущественных признаков.

В качестве прогнозных значений были использованы вероятности принадлежности объектов к классу 1, то есть predict\_proba. Помимо этого были отрисованы гистограммы получаемых прогнозных значений, а также confusion matrix.

Если говорить о гистограммах, то распределение прогнозов, получаемых с помощью бустинга, напоминает U-shape distribution (стоит учитывать пропорции классов, поэтому правый хвост ниже левого), в то время как прогнозы случайного леса больше распределены нормально или равномерно, то есть для значительной части объектов вероятность принадлежности к классу 1 расположена в пределах [0.35; 0.65], что означает, что модель «в замешательстве» и не смогла так хорошо промоделировать объекты, в отличие от бустинга (несмотря на близкие метрики). Объяснением этому может быть bias-variance decomposition: случаный лес склонен уменьшать дисперсию, жертвуя при этом смещением. Причина — каждое дерево из ансамбля работает с малым набором данных, что ведет к увеличению смещения, но при этом в ансамбле много деревьев, что позволяет уменьшать дисперсию

Помимо этого были изучены confusion matrices, и бустинг показывал похожие значения FN и FP, в то время как соотнешие FN/FP у случаного леса было далеко от единицы. Также бустинг показывал более хорошие результаты TN и немного худшие результаты TP.

Таким образом, предпочтение было отдано бустингу. Для оценки статистической значимости, валидационный датасет был разделен на 10 частей, и каждая из 4-х моделей была протестирована на каждой из этих частей. Для полученных векторов метрик были рассчитаны t-статистики Стьюдента для зависимых выборок и оценены p-values. После повторения указанной процедуры было выяснено, что бустинг, обученный на oversampled данных показал наилучшие результаты.