Отчет по итоговому заданию из курса «Введение в машинное обучение»

Подход 1: градиентный бустинг "в лоб»

>> Какие признаки имеют пропуски среди своих значений? Что могут означать пропуски в этих признаках (ответьте на этот вопрос для двух любых признаков)?

Ниже приведен список признаков, имеющих пропуски, а также процент пропусков

Признак	Процент пропусков
first_blood_time	20.1
first_blood_team	20.1
first_blood_player1	20.1
first_blood_player2	45.2
radiant_bottle_time	16.1
radiant_courier_time	0.7
radiant_flying_courier_time	28.3
radiant_first_ward_time	1.9
dire_bottle_time	16.6
dire_courier_time	0.7
dire_flying_courier_time	26.8
dire_first_ward_time	1.9

Данные признаки описывают события те или иные события, которые наступили в течение первых 5 минут матча. В случае, если событие не произошло в течение данного периода времени, значения признаков отсутствуют. Этим объясняются пропуски.
Например, пропуск для признака first_blood_time означает, что за первые 5 минут матча не произошло события «Первая кровь». А пропуск для признака radiant_bottle_time означает, что за первые 5 минут матча команда radiant не приобрела объект bottle.

>> Как называется столбец, содержащий целевую переменную?

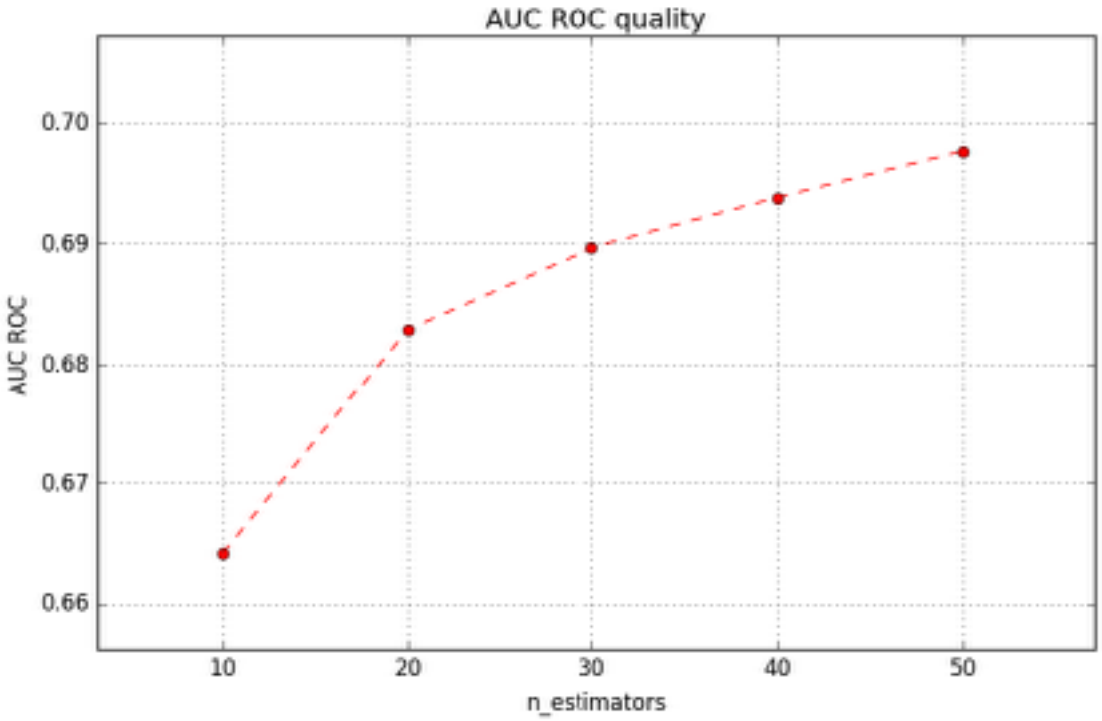
Целевую переменную содержит столбец с названием «radiant_win».
Если значение 0 - выиграла команда dire, если 1 - выиграла команда radiant.

>> Как долго проводилась кросс-валидация для градиентного бустинга с 30 деревьями?

Кросс-валидация по 5 блокам для градиентного бустинга с 30 деревьями потребовала 165.7 секунд.

>> Имеет ли смысл использовать больше 30 деревьев в градиентном бустинге?

Использование количества деревьев больше 30 имеет смысл. Ниже приведен график метрики качества (AUC ROC) для значения параметра деревьев из списка [10, 20, 30, 40, 50]:



Как видно из графика, качество классификатора продолжает расти с ростом числа деревьев.

>> Что бы вы предложили делать, чтобы ускорить его обучение при увеличении количества деревьев?

В подобной ситуации можно пробовать несколько способов ускорения обучения:

- ограничивать глубину обучаемых деревьев
- обучать деревья на случайно-выбранном подмножестве объектов
- обучать деревья на случайно-выбранном подмножестве признаков

Подход 2: логистическая регрессия

>> Какое качество получилось у логистической регрессии над всеми исходными признаками? Как оно соотносится с качеством градиентного бустинга? Чем вы можете объяснить эту разницу? Быстрее ли работает логистическая регрессия по сравнению с градиентным бустингом?

Максимальное качество, полученное методом градиентного бустинга для метрики площади под ROC кривой равно 0.697641004776. Классификатор был обучен с 50 деревьями. Время обучения составило 301 секунду.

Для логистической регрессии над всеми исходными признаками было получено наилучшее значение метрики 0.716606261879. Классификатор был обучен со значением параметра регуляризации C = 51. Время обучения составило 21 секунду.

Исходя из результатов можно заметить, что логистическая регрессия дала лучшее качество в сравнении градиентным бустингом. Это вызвано недостаточным количеством деревьев в градиентном бустинге. Однако, уже при 50 деревьях время на обучение градиентного бустинга требуется в 15 раз больше чем на обучение логистической регрессии.

>> Как влияет на качество логистической регрессии удаление категориальных признаков (укажите новое значение метрики качества)? Чем вы можете объяснить это изменение?

Логистическая регрессия по всем признакам дала наилучший результат метрики качества 0.716606261879. После удаление категориальных признаков наилучшая метрика качества составила 0.716632614755. Как видно, разница практически отсутствует. Данный факт вызван тем, что логистическая регрессия предполагает линейную зависимость результата от признаков. Категориальные признаки, выраженные в числах, не дают линейной зависимости в выходной переменной. Таким образом, при обучении лог. регрессии по всем признакам, данным признакам будет соответствовать близкое к нулю значение. Что по сути и значит - удаление этих признаков, то есть они не влияли и в первом случае на результат.

>> Сколько различных идентификаторов героев существует в данной игре?

В игре было зафиксировано 108 различных героев.

>> Какое получилось качество при добавлении "мешка слов" по героям? Улучшилось ли оно по сравнению с предыдущим вариантом? Чем вы можете это объяснить?

После добавления «мешка слов» наилучшее значение метрики качества составило 0.751873065537, что добавило примерно 4% к качеству. Данный факт объясняется тем, что герои в игре имеют различные особенности. И учет того, какой герой был использован в матче обеспечил дополнительную полезную информацию для построения модели.

>> Какое минимальное и максимальное значение прогноза на тестовой выборке получилось у лучшего из алгоритмов?

Минимальное значение прогноза равно 0.008429, максимальное 0.996371.

Проверка финальное модели

В качестве результата прогнозирования выигрыша команды был сформирован файл «result.csv» согласно требованиям. После отправки файла на kaggle.com была получена оценка качества 0.75529 (350 место на момент отправки).