Ответьте на вопросы по первому этапу задания (градиентный бустинг):

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

Время firstblood – первого убийства за 5 минут могло и не быть

И время покупеи летающего курьера – могли не покупать.

1. Как называется столбец, содержащий целевую переменную?

radiant\_win в файле features.csv

1. Как долго проводилась кросс-валидация для градиентного бустинга с 30 деревьями? Инструкцию по измерению времени можно найти выше по тексту. Какое качество при этом получилось?

Кросс-валидация для градиентного бустинга с 30 деревьями заняла 0:00:54.

Показатель метрики качества AUC-ROC в таком случае равен 0.69.

1. Имеет ли смысл использовать больше 30 деревьев в градиентном бустинге? Что можно сделать, чтобы ускорить его обучение при увеличении количества деревьев?

Для достижения большего качества имеет смысл использовать более 30 деревьев.

Для ускорения обучения можно использовать не всю выборку, а некоторое ее подмножество — например, половину объектов.

Также можно уменьшить глубину деревьев (max\_depth).

Ответьте на вопросы по второму этапу задания:

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

Наилучшее значение показателя AUC-ROC достигается при C = 0.01 и равно0.72

Это сравнимо с градиентным бустингом по 250 деревьям, при этом логистическая регрессия работает заметно быстрее 14 секунд вместо 5 минут.

1. Как влияет на качество логистической регрессии удаление категориальных признаков (укажите новое значение метрики качества)? Чем можно объяснить это изменение?

Удаление категориальных признаков практически не повлияло на качество предсказания (изменения только в 5 знаке после запятой).

Наилучшее значение показателя AUC-ROC так же достигается при C = 0.01 и равно 0.71.

Получается, что в предыдущей модели эти признаки никак не влияли на результат предсказания.

1. Сколько различных идентификаторов героев существует в данной игре?

число героев в игре 112

1. Какое получилось качество при добавлении "мешка слов" по героям? Улучшилось ли оно по сравнению с предыдущим вариантом? Чем можно это объяснить?

После добавления "мешка слов" по героям качество заметно улучшилось.

Наилучшее значение показателя AUC-ROC достигается при C = 0.1 и равно 0.75

Это объясняется тем, что вместо отсутствия данных о героях или случайного шума из id мы имеем осмысленную разреженную матрицу для построения предсказания.

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

При логистической регресси ( лучший алгоритм) мин знач 0.6931532

Максимально при использовании мешка слов = 0.75607342