Предсказание оценки качества для вина по физико-химическим показателям

1. Введение

Качество вина определяется множеством факторов, от физико-химического состава до субъективных восприятий. Располагая данными только о физико-химических показателях, провели исследование об их влиянии на выставляемую продукту оценку качества. Построили несколько моделей, способную предсказывать оценку, которую получит вино, по его составу.

2. Анализ данных

Выполнили необходимую предобработку данных, посмотрели на статистические показатели, определили наличие выбросов в данных, оценили линейные взаимосвязи показателей со значением оценки качества вина – сильных линейных взаимосвязей не выявлено. Посмотрели на распределение оценок по красному и белому вину (рис.1). наиболее часто выставляемые оценки – пять и шесть, это средние оценки.

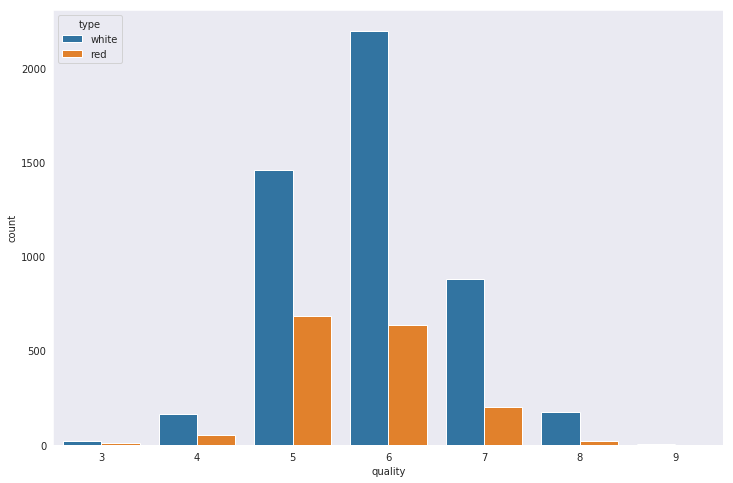


Рисунок 1. Распределение оценок по видам вина

3. Моделирование и оценка результатов

Построили две модели – логистической регрессии и случайного леса. Обе модели показали точность примерно 55%, что является средним результатом. При этом из-за несбалансированности данных модели хорошо работают с классами наиболее часто встречающихся оценок и совсем не работают с редкими оценками. После уравновешивания данных модели стали работать со всеми классами оценок, но хуже, и точность моделей снизилась на 20%.

4. Выводы

Исходя из того, что потребители в большинстве случаев выставляют средние оценки, важно, чтобы модели хорошо работали именно с этими классами оценок. В перспективе, накопив больше данных, можно повысить качество моделей.