Отчет по анализу дата-сета вин

1. Введение

Винная продукция пользуется большим спросом во многих странах, а потребление данного напитка растет несколько лет подряд, а вкусы и запросы потребителей становятся все более сложными и необычными. Рынок становится очень привлекательным и для новых игроков, что повышает уровень конкуренции. Чтобы сохранить и даже увеличить долю рынка требуется провести анализ параметров вин для выявления лучшего состава, при котором оценка качества будет на максимально высоком уровне.

2. Описание дата-сета

В анализе использовался дата-сет объединяющий набор данных белых и красных вин различного качества и химического состава. В исходных данных было представлено 6497 различных сортов вина. Общее количество входных переменных 11 штук, прогнозируемая переменная – качество – оценивалось по десятибалльной шкале.

Пример входных данных представлен в таблице 1.

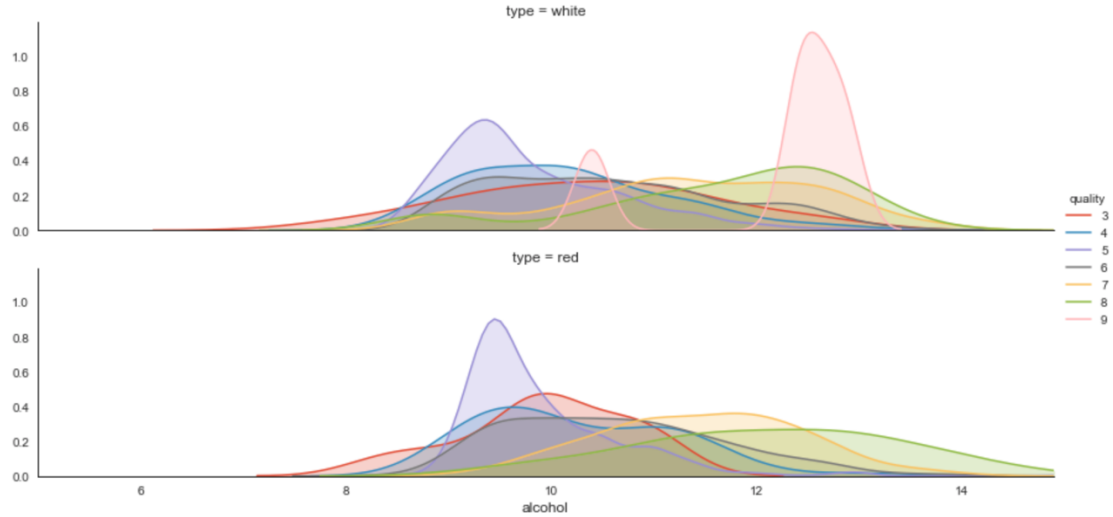
Таблица 1

type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
white	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8
white	[illegible]	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5
white	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1
white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9
white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9

utput; double click to hide output

Целевым признаком является качество вин. Зависимость распределения алкоголя и качество вина в разрезе типа вина представлено на рисунке 1.

Рисунок 1. Зависимость распределения алкоголя на качество вин



3. Подготовка данных

Перед созданием модели дата-сет был предварительно обработан. Для первичного анализа и построения модели были удалены выбросы и пропущенные значения, которых было незначительное количество, что не должно повлиять на результат оценки модели.

Также категориальные признаки были переведены в числовые значения для повышения обучаемости модели.

4. Обучение модели

Целью моделирования является нахождение оптимального набора показателей, при котором модель показывает лучшие результаты.

В качестве модели была выбрана модель логистической регрессии, данные были разделена на тренировочные и тестовые, 70% и 30% соответственно.

В результате обучения был достигнут показатель в 48%.

5. Анализ результатов моделирования

Показатель модели получились невысокие, требуется дальнейшая работа с данными для повышения качества модели.

6. Планы

В дальнейших планах провести повторный анализ данных, обогатить данные из внешних источников. Также следует провести эксперименты с другими моделями, возможно модель логистической регрессии является не самой подходящей для этой задачи.