Wine Quality

Andrey Bibikov | DS-27 | 2021

1. Introduction / Введение

В отчете описано создание модели для прогнозирования качества вин. Основой для создания модели прогнозирования служат измеренные физико-химические данные вина. Целью проекта является создание модели прогнозирования качества вина с высокой точностью при использовании минимума измеряемых показателей.

Для создания модели используются реальные данные вин. (Источник получения данных https://www.kaggle.com/rajyellow46/wine-quality)

СОСТАВ ОТЧЕТА

В разделе 2 представлено описание набора данных. Раздел 3 описывает исследование данных и их предобработку. В 4 разделе представлены выбранные модели и пайплайн для их обучения. В 5 разделе приведены оценки точности прогнозирования моделей и представлены дальнейшие пути повышения точности прогнозирования.

2. Input Data / Информация о наборе данных

Набор данных был загружен из репозитория машинного обучения UCI. Набор данных относится к красному и белому вариантам португальского вина "Винью Верде". Ссылка [Cortez et al., 2009]. Доступны только физико-химические (исходные данные) и сенсорные (выходные данные) переменные (например, нет данных о сортах винограда, марке вина, продажной цене вина и т. д.). Незначительное количество данных отсутствует.

Attribute Information:

For more information, read [Cortez et al., 2009].

Input variables (based on physicochemical tests):

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

Dataset состоит из 6497 строк и 13 столбцов (**табл. 1**).

табл. 1 Dataset 'winequalityN'

	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	white	7.0	0.270	0.36	20.7	0.045	45.0	170.0	1.00100	3.00	0.45	8.8	6
1	white	6.3	0.300	0.34	1.6	0.049	14.0	132.0	0.99400	3.30	0.49	9.5	6
2	white	8.1	0.280	0.40	6.9	0.050	30.0	97.0	0.99510	3.26	0.44	10.1	6
3	white	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.40	9.9	6
4	white	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.40	9.9	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6492	red	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5
6493	red	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	NaN	11.2	6
6494	red	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6
6495	red	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5
6496	red	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6

6497 rows × 13 columns

3. Features and preprocessing / Признаки и предобработка

Перед моделированием был проведен анализ данных. В **табл. 2** представлены основные статистические данные признаков.

табл. 2 Основные статистические данные

	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	6497.0000	6497.0000	6497.0000	6497.0000	6497.0000	6497.000	6497.0000	6497.0000	6497.0000	6497.0000	6497.0000	6497.0000	6497.0000
mean	0.7539	7.2166	0.3397	0.3187	5.4443	0.056	30.5253	115.7446	0.9947	3.2184	0.5312	10.4918	5.8184
std	0.4308	1.2958	0.1645	0.1452	4.7574	0.035	17.7494	56.5219	0.0030	0.1606	0.1488	1.1927	0.8733
min	0.0000	3.8000	0.0800	0.0000	0.6000	0.009	1.0000	6.0000	0.9871	2.7200	0.2200	8.0000	3.0000
25%	1.0000	6.4000	0.2300	0.2500	1.8000	0.038	17.0000	77.0000	0.9923	3.1100	0.4300	9.5000	5.0000
50%	1.0000	7.0000	0.2900	0.3100	3.0000	0.047	29.0000	118.0000	0.9949	3.2100	0.5100	10.3000	6.0000
75%	1.0000	7.7000	0.4000	0.3900	8.1000	0.065	41.0000	156.0000	0.9970	3.3200	0.6000	11.3000	6.0000
max	1.0000	15.9000	1.5800	1.6600	65.8000	0.611	289.0000	440.0000	1.0390	4.0100	2.0000	14.9000	9.0000

Столбец 'type' имеет тип данных - object, столбец 'quality' (целевая переменная) - int64, остальные столбцы - float64. В данных присутствуют пропуски (**табл. 3**)

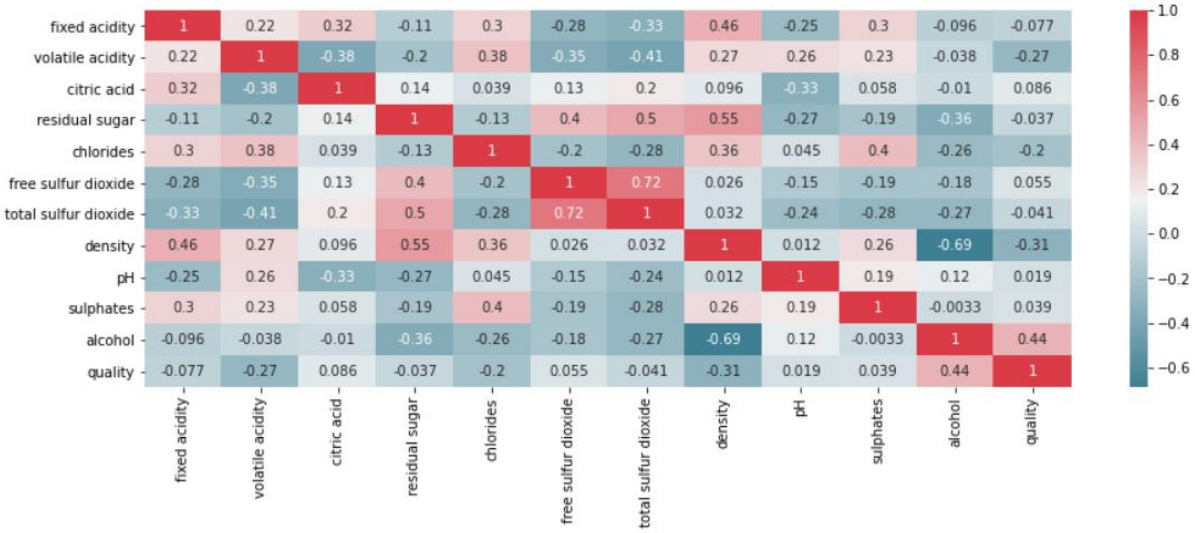
табл. 3 Оценка пропусков в данных

	Missing Values	% of Total Values	Data Types
fixed acidity	10	0.2	float64
pH	9	0.1	float64
volatile acidity	8	0.1	float64
sulphates	4	0.1	float64
citric acid	3	0.0	float64
residual sugar	2	0.0	float64
chlorides	2	0.0	float64

Пропущенные значения будут заполнены средним по переменной.

Для оценки взаимосвязи между признаками построена матрица корреляции, визуализация представлена тепловой картой (**рис. 1**).

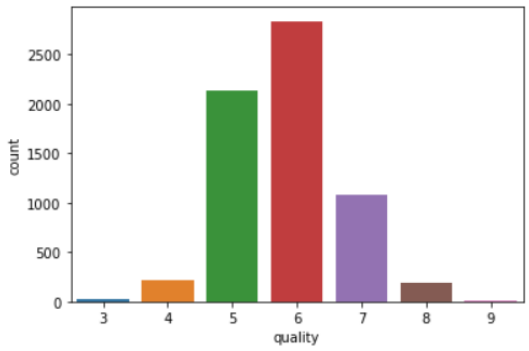
рис. 1 Матрица корреляции



Корреляция некоторых признаков превышает интервал (-0.5, 0.5). Это может усложнить и замедлить поиск зависимости моделью.

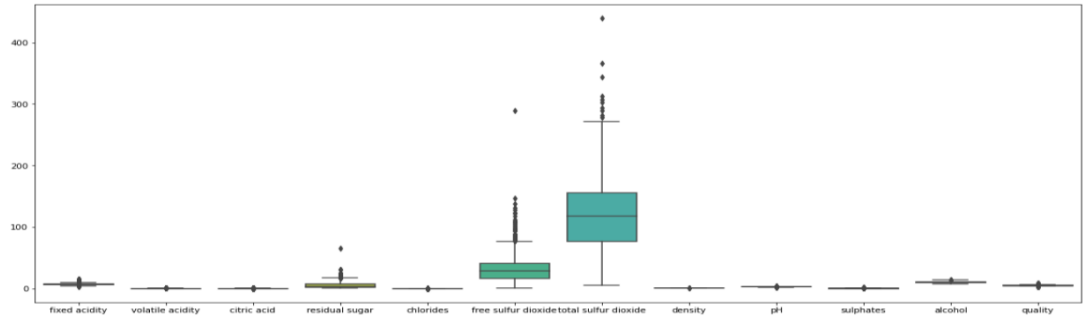
В проекте проведена оценка сбалансированности классов целевой переменной ‘quality’. Из графика (**рис. 2**) видно, что классы не сбалансированы.

рис. 2 Оценка сбалансированности классов



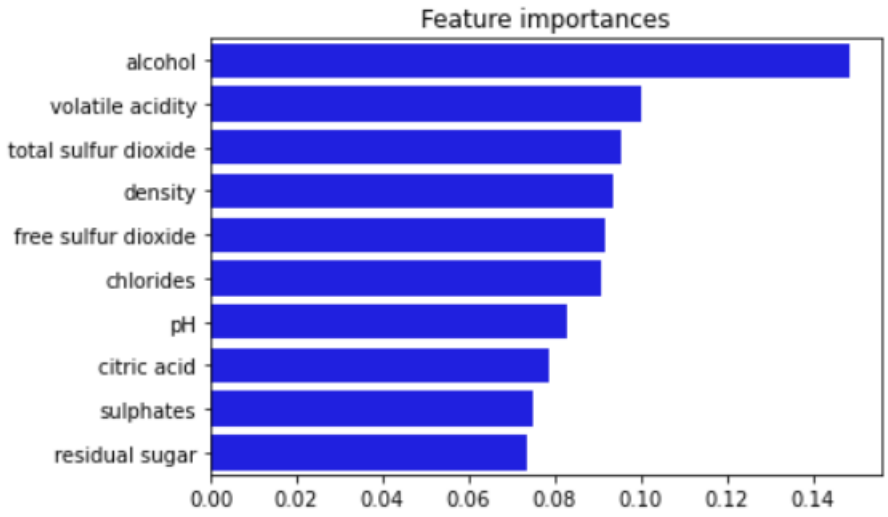
На **рис. 3** представлен график оценки выбросов.

рис. 3 Оценка выбросов



С помощью модели DecisionTreeClassifier проведена оценка важности признаков (**рис. 4**).

рис. 4 Оценка важности признаков



В данном проекте выбираем все доступные признаки для формирования итогового датасета.

Для проведения обучения, валидации и тестирования модели исходный набор данных был разделен на выборки train, val, test с использованием stratification, т.к. классы несбалансированны.

Размер обучающей выборки – (3182, 12).

Размер валидационной выборки – (1365, 12).

Размер тестовой выборки - (1950, 12).

4. Learning Models / Обучение моделей

В проекте были использованы модели из библиотеки sklearn:

- LogisticRegression
- SVC
- RandomForestClassifier
- DecisionTreeClassifier
- KNN

Во всех моделях использовались гиперпараметры по умолчанию.

Pipeline для обучения моделей состоит из StandardScaler() и модели машинного обучения.

5. Results / Результаты

Для каждой модели рассчитана оценка Score (**табл. 4**).

табл. 4 Оценки моделей

	Model	Score
2	RandomForestClassifier	0.647692
1	SVC	0.563077
4	KNN	0.555385
3	DecisionTreeClassifier	0.547179
0	LogisticRegression	0.535385

Из **табл. 4** видно, что модель RandomForestClassifier показала оценку прогнозирования лучше, чем другие модели.

Из classification_report (**табл. 5**) для RandomForestClassifier видно, что 3-й класс не определяется моделью, 4-й плохо определяется. 8-й класс тоже сложно обнаруживается моделью. 9-й класс не определился. Лучше всех определяются классы 5, 6 и 7.

табл. 5 classification_report

	precision	recall	f1-score	support
3	0.00	0.00	0.00	9
4	0.71	0.08	0.14	65
5	0.68	0.71	0.69	642
6	0.63	0.74	0.68	851
7	0.64	0.50	0.56	324
8	0.79	0.19	0.31	58
9	0.00	0.00	0.00	1
accuracy			0.65	1950
macro avg	0.49	0.32	0.34	1950
weighted avg	0.65	0.65	0.63	1950

Precision - Способность алгоритма отличать данный класс от других классов.

Recall - Способность алгоритма обнаруживать данный класс вообще.

Вывод: Оценка модели RandomForestClassifier невысока. Для повышения точности прогнозирования необходимо провести дополнительные исследования в области:

- feature engineering (заполнение пропусков, отбор признаков, генерация новых признаков, работа с выбросами);
- подобрать гиперпараметры модели;
- использовать более сложные модели (напр. ансамбли, нейронные сети);