

Проект по ML

Атабекян Эдгар
Сенаторов Пётр

Цель

Предсказать качество вина (целевые классы от 3 до 8) на основе химических характеристик, используя тренировочные данные.

Целевая метрика:

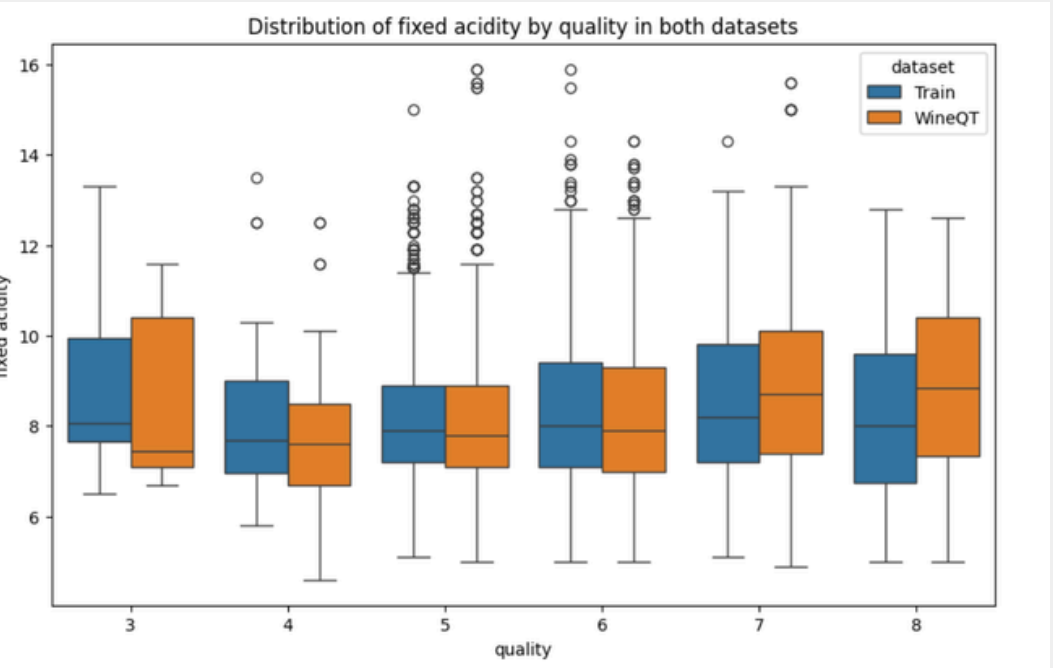
Quadratic Weighted Kappa (QWK).

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}.$$

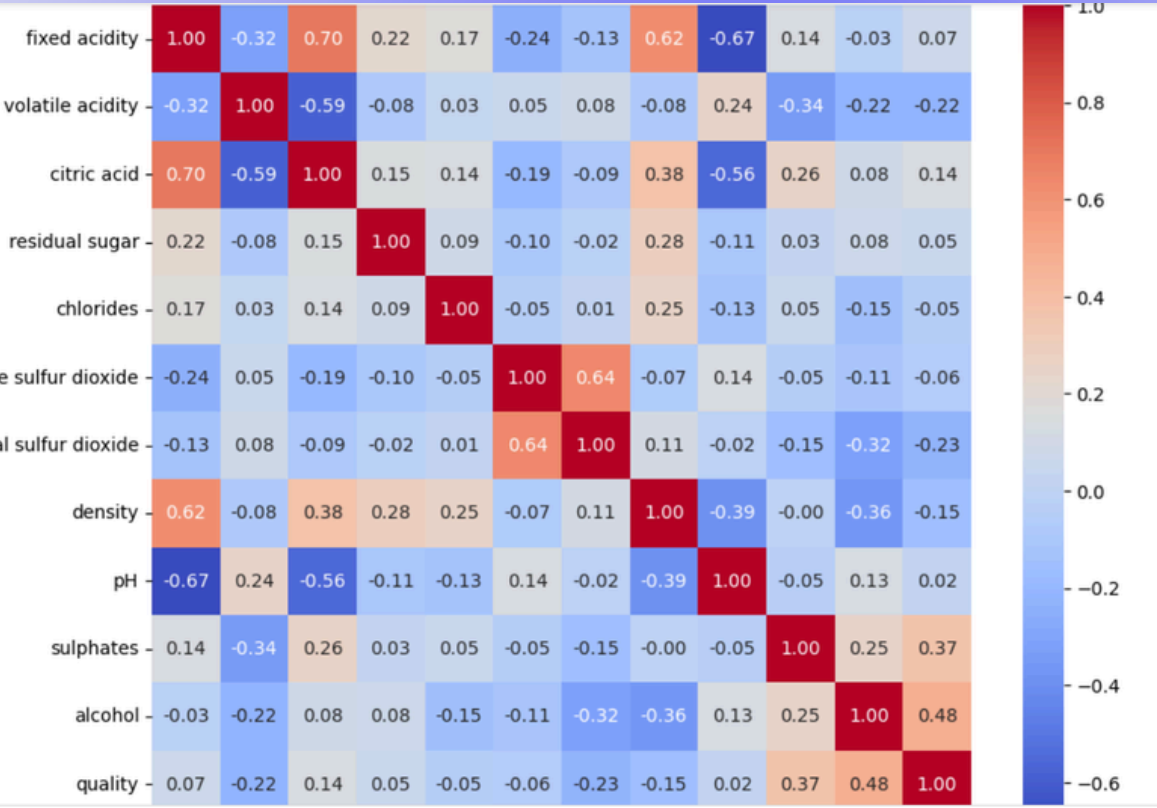
Данные:

- Тренировочный набор: химические параметры (11 признаков) и целевой признак quality.
- Тестовый набор: те же признаки, без значений quality.
- Дисбаланс классов: большинство данных приходится на классы 5 и 6.
- Пропуски отсутствуют.
- Выбросы есть, в большинстве своем у классов 5 и 6.
- Наибольшая корреляция с полем alcohol.

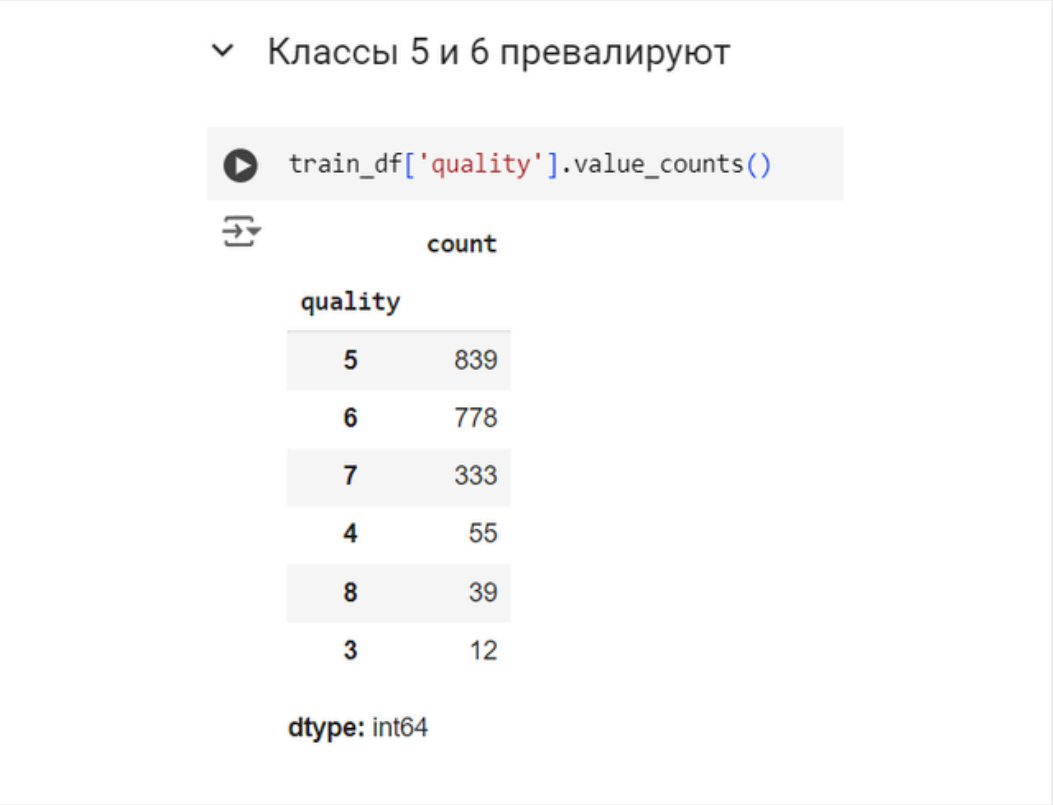
EDA



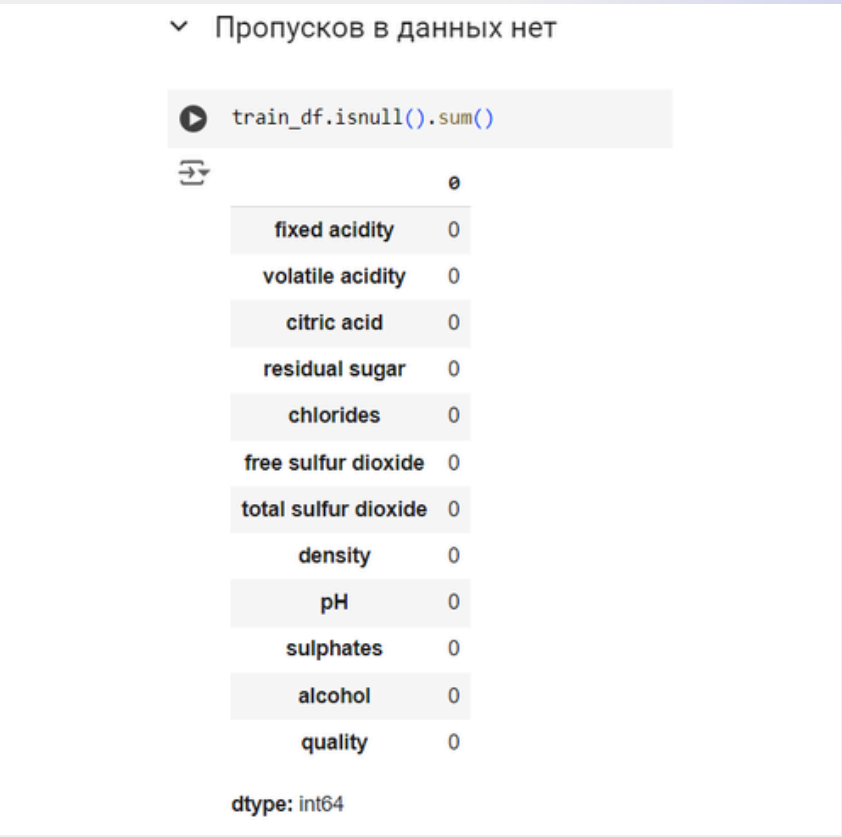
Выбросы



Корреляция



Количество элементов в
классе



Пропуски

Бейслайн:

- Модель: LightGBM Classifier с базовыми параметрами.
- Использованы все признаки без изменений.
- Добавлены новые признаки.
- Также использовался catboost для бейслайна, но цифры хуже.



LGBM_BaseLine.csv

Complete (after deadline) · 14s ago

0.52963



predictions.csv

Complete (after deadline) · 15d ago

0.52116

0.49306




Catboost baseline

0.52


Усиление бейслайна

LGBMClassifier

- Добавление новых данных
- Подбор гиперпараметров LightGBM с помощью Optuna.





	LGBM.csv Complete (after deadline) · 19m ago	0.49187	0.45166
---	--	----------------	----------------

- Кросс валидация с использованием K-Fold и выбор самого частого предсказания

	LGBM_KFold.csv Complete (after deadline) · 18m ago	0.51054	0.49657
---	--	----------------	----------------

LGBMRegressor с округлением параметров


- Добавление новых данных
- Подбор гиперпараметров LightGBM с помощью Optuna.
- Кросс валидация с использованием K-Fold и выбор самого частого предсказания
- Кросс-валидация+ Optuna

	LGBM_Reg_KFold_3.csv Complete (after deadline) · 39m ago	0.59283	0.53906
	LGBM_Reg_KFold_2.csv Complete (after deadline) · 39m ago	0.58163	0.54288
	LGBM_Reg_KFold_1.csv Complete (after deadline) · 40m ago	0.58070	0.51470
	LGBM_Reg.csv Complete (after deadline) · 40m ago	0.57761	0.54885

Усиление бейслайна

NN + Optimized Rounder

- Добавление новых данных
- Небольшая модель с одной сверткой, Dropout и Batch Norm
- После получения предсказаний прогон через Optimized rounder для получения оптимальных границ для разбиения на классы

 predictions_opt_rounder7.csv Complete (after deadline) · 19h ago	0.57900	0.56990	<input type="checkbox"/>
--	---------	---------	--------------------------

- Также была попытка усложнить модель, но результаты лучше не стали
- Помимо ранее названных методов сначала были испробованы стекинг, max voting, удаление выбросов (стало сильно хуже) и скейлинг некоторых признаков (также привело только к ухудшению)

 predictions_stack.csv Complete (after deadline) · 13d ago	0.52364	0.52169	<input type="checkbox"/>
---	---------	---------	--------------------------