Аналитический отчёт по исследованию

«Прогнозирование качества вина по физико-химическим данным»

Павлов Николай Геннадьевич (nikolay-pavlov-96@mail.ru)

Студент группы DS университета Нетология

1. Введение

В настоящей статье представлены основные этапы решения задачи предсказания качества вина по физико-химическим данным на основе данных красного и белого португальского вина "Vinho Verde". Решение рассматривается как задача классификации.

Структура настоящего отчёта следующая: во втором разделе разобраны представленные данные, в третьем – показаны выявленные зависимости, в четвёртом – собран датасет для моделирования, в пятом – протестированы и выбраны лучшие модели на основе метрики качества, в шестом – итоги решения.

Задача решалась по данным из ресурса Kaggle. Посмотреть более полную информацию по данным, а также сами данные можно по представленной ссылке: Ссылка на Kaggle

Решение автора отчёта представлено по ссылке: Ссылка на решение

В приложении небольшой интерактивный отчёт по исследуемым данным для разделов 2 и 3, разработанный в среде Power BI.

2. Описание данных

В таблице 1 представлено признаковое пространство датасета, с которым была проделана дальнейшая работа.

Таблица 1. Датасет

№	Признак	Перевод	Количество ненулевых значений	Тип данных	Количество уникальных значений
1	type	тип вина	6497	object	2
2	fixed acidity	фиксированная кислотность	6487	float64	106
3	volatile acidity	летучая кислотность	6489	float64	187
4	citric acid	лимонная кислота	6494	float64	89
5	residual sugar	остаточный сахар	6495	float64	316
6	chlorides	хлориды	6495	float64	214
7	free sulfur dioxide	свободный диоксид серы	6497	float64	135
8	total sulfur dioxide	диоксид серы общий	6497	float64	276
9	density	плотность	6497	float64	998
10	pH	Мера кислотности	6488	float64	108
11	sulphates	сульфаты	6493	float64	111
12	alcohol	алкоголь	6497	float64	111
13	quality (target)	качество (целевая переменная)	6497	int64	7

На рисунке 1 продемонстрированы первые 5 строк из описанного датасета.

	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	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	white	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	white	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

Рис.1. Пример данных из датасета

Данные состоят 6497 строк и 13 столбцов, один из которых представляет целевую переменную – quality. Одной из особенностей данных является то, что они нормально распределены по качеству вина и неравное распределение по типам вина данных на рисунке 2 и 3.

Рис. 2. Нормальное распределение данных

Рис. 3. Данные по типам вина

3. Зависимости в данных

Были составлена тепловая карта корреляции признакового пространства датасета для определения наиболее скореллированных признаков для предотвращения возможности переобучения (см. Рис.4). Также была обнаружена интересная зависимость, что чем более крепкое вино, тем больше у него оценки качества.

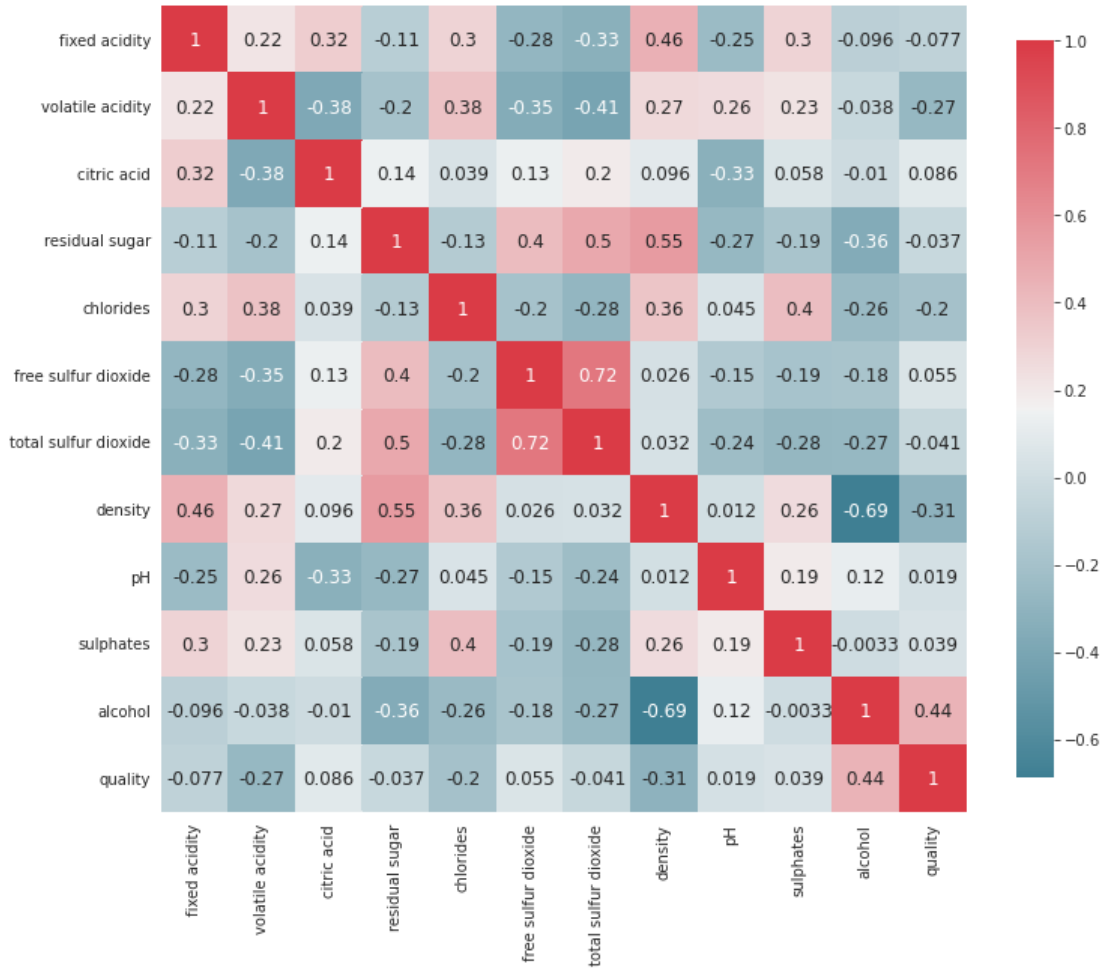

Рис. 4. Корреляция данных

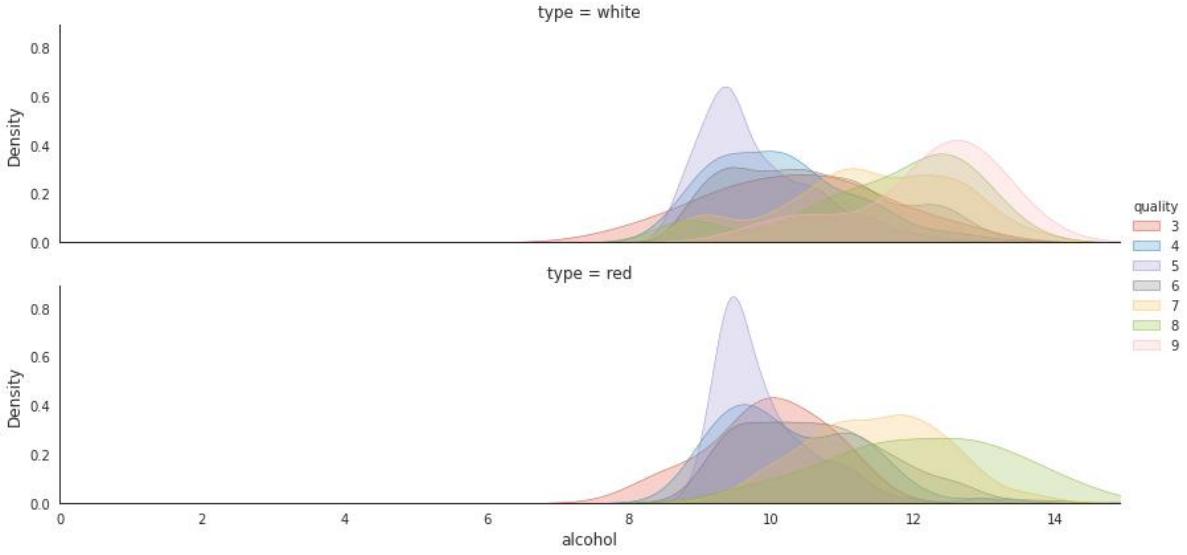

Рис. 5. Распределение density по alcohol по типам вина

4. Датасеты для моделирования

При подготовки исследуемого датасета для модели были проделаны следующие шаги:

1) преобразование категориальных переменных в числовые
2) заполнение пропущенных значений в данных средними
3) конкатенация с учётом предыдущих шагов и первоначального датасета
4) разделение конечного датасета на X (признаковое пространство) и на Y (целевая переменная), в этом же шаге была определена важность признаков (см. Рис.6)
5) разделение данных на train(2448), valid(1049), test(3000)

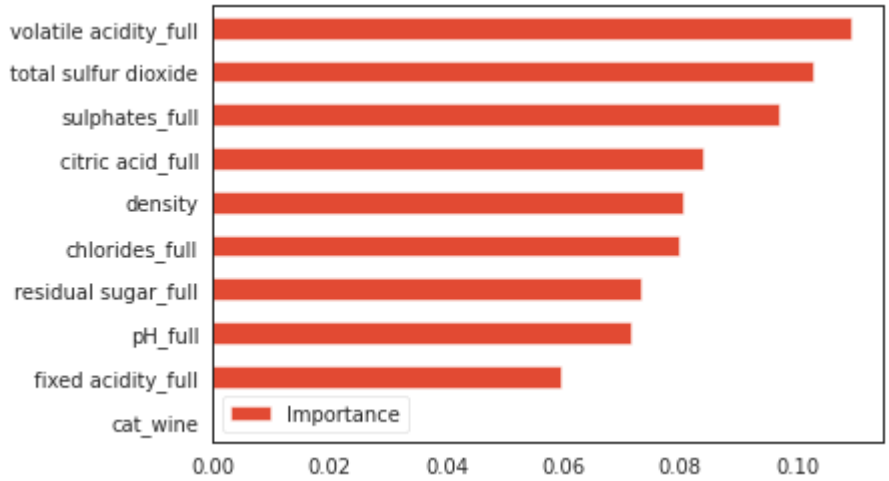

Рис.6. Определение важности признаков

5. Выбор и оценка моделей

При обучении сравнивались 8 различны моделей, все модели и их сокращенные наименования даны в таблице 2, а диаграмма сравнения их качества по различным выборкам представлена на рис.7.

Таблица 2. Исследуемые модели обучения

№	Модель	Сокращенное наименование
1	DecisionTreeClassifier(max_depth=10)	'Tree'
2	LogisticRegression()	'LogReg'
3	KNeighborsClassifier()	'Kneigh'
4	GaussianNB()	'GaussNB'
5	SVC()	'SVC'
6	LinearSVC()	'linSVC'
7	RandomForestClassifier(max_depth=10)	'Forest'
8	GradientBoostingClassifier()	'Booster'

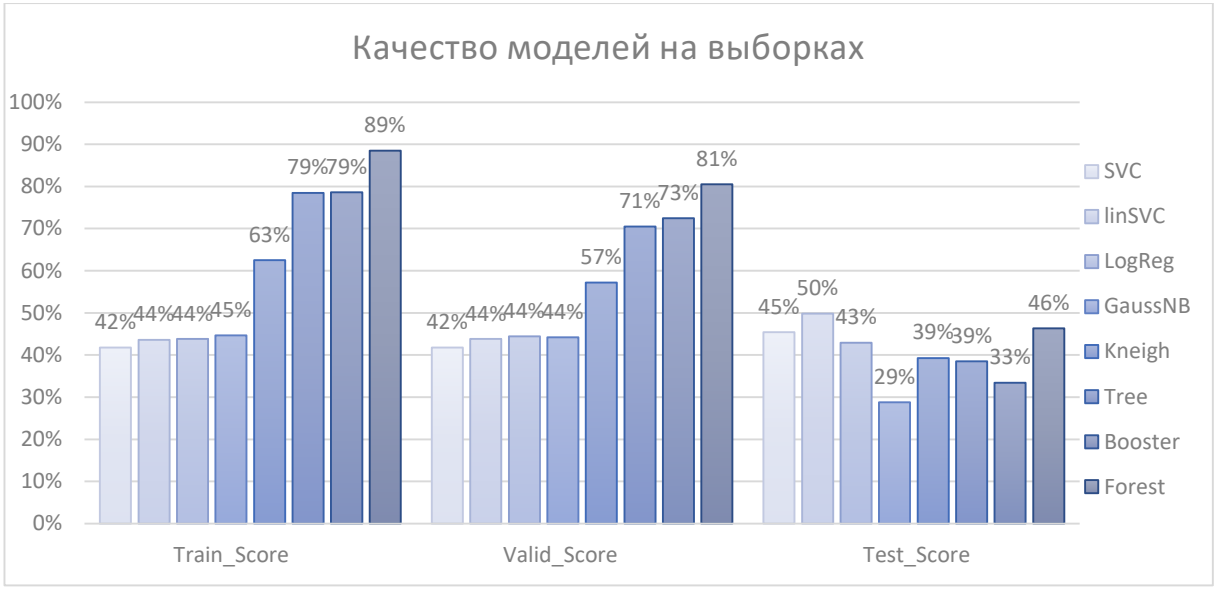

Рис.7. Диаграмма сравнения качества моделей на различных выборках

6. Заключение

Подведём итог, участвовало 8 моделей, из них лучше всего на valid выборке справились топ-5: Forest(0,805), Booster, Tree, Kneigh, LogReg (0,444). Однако на тестовой выборке на финале лучше всего справились топ-3: linSVC(0,498), Forest(0,463), SVC(0,454). Из всех них была выбрана модель Forest (RandomForestClassifier(max_depth=10)) как лучше всего справившуюся на всех выборках.

Надо отметить и то, что модели были применены в «коробочном» варианте, это значит, что можно поэкспериментировать с гиперпараметрами моделей, попробовав добиться лучшего качества. Также увеличив количество данных, можно также повысить качество обучения моделей и, соответственно, их более точного предсказания.