Отчет по разработке модели предсказания класса вина

1. Вступление

Учебная задача по определению класса вина: балл от 0 до 10.
Отчет состоит из следующих разделов:

- Описание датасета. Описание из каких параметров состоит датасет.
- Предварительная обработка. Описание какая предобработка данных была необходима и как она выполнялась
- Обучающая модель. Какие модели использовались при обучении и почему
- Результаты. Какие результаты были получены в результате обучения модели
- Обсуждение и доработки модели. Насколько хорошо сработала модель, как ее можно доработать
- Планы на будущее. Планы по улучшению модели и ее использованию

2. Описание датасета

Два набора данных относятся к красному и белому вариантам португальского вина «Vinho Verde». Ссылка [Cortez et al., 2009]. Из-за проблем с конфиденциальностью и логистикой доступны только физико-химические (входные) и органолептические (выходные) переменные (например, нет данных о сортах винограда, марке вина, продажной цене вина и т. д.).

Эти наборы данных можно рассматривать как задачи классификации или регрессии. Классы упорядочены и не сбалансированы.

Два набора данных были объединены, и несколько значений были удалены случайным образом.

Входные переменные (на основе физико-химических тестов):
1 - fixed acidity фиксированная кислотность
2 - volatile acidity летучая кислотность
3 - citric acid лимонная кислота
4 - residual sugar остаточный сахар
5 - chlorides хлориды
6 - free sulfur dioxide свободный диоксид серы
7 - total sulfur dioxide диоксид серы общий
8 - density плотность
9 - pH
10 - sulphates сульфаты
11 - alcohol алкоголь
Выходная переменная (на основе сенсорных данных):
12 - quality качество (оценка от 0 до 10)

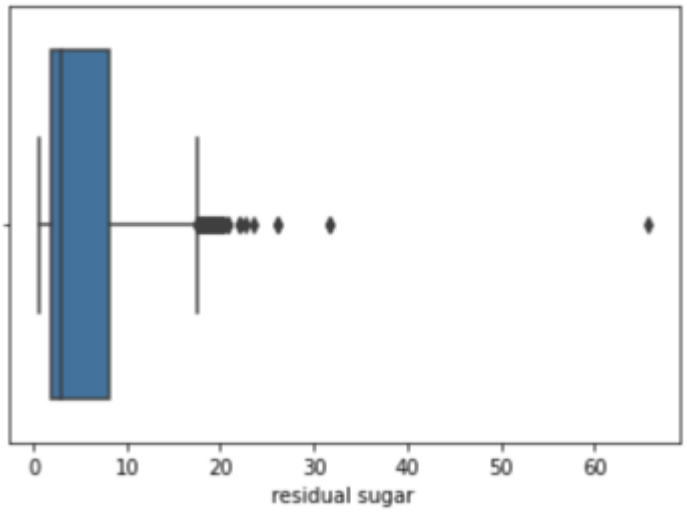
3. Предварительная обработка данных

В данных имеются пропуски:

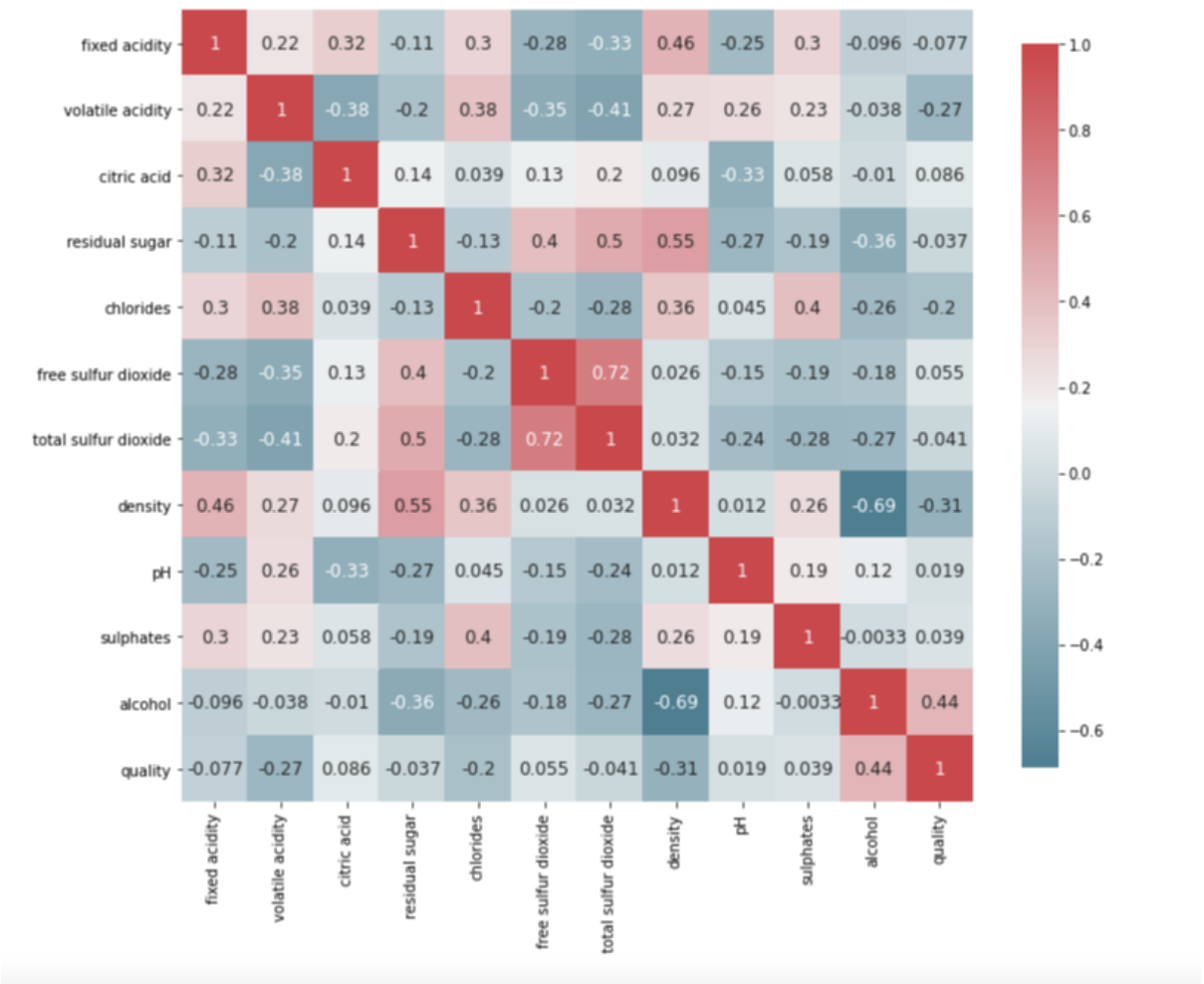
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6497 entries, 0 to 6496
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   type                  6497 non-null   object 
 1   fixed acidity         6487 non-null   float64
 2   volatile acidity      6489 non-null   float64
 3   citric acid           6494 non-null   float64
 4   residual sugar        6495 non-null   float64
 5   chlorides             6495 non-null   float64
 6   free sulfur dioxide   6497 non-null   float64
 7   total sulfur dioxide  6497 non-null   float64
 8   density               6497 non-null   float64
 9   pH                    6488 non-null   float64
 10  sulphates             6493 non-null   float64
 11  alcohol               6497 non-null   float64
 12  quality               6497 non-null   int64  
dtypes: float64(11), int64(1), object(1)
memory usage: 660.0+ KB
```

Пропусков немного - менее 1%, поэтому мы их просто удалили

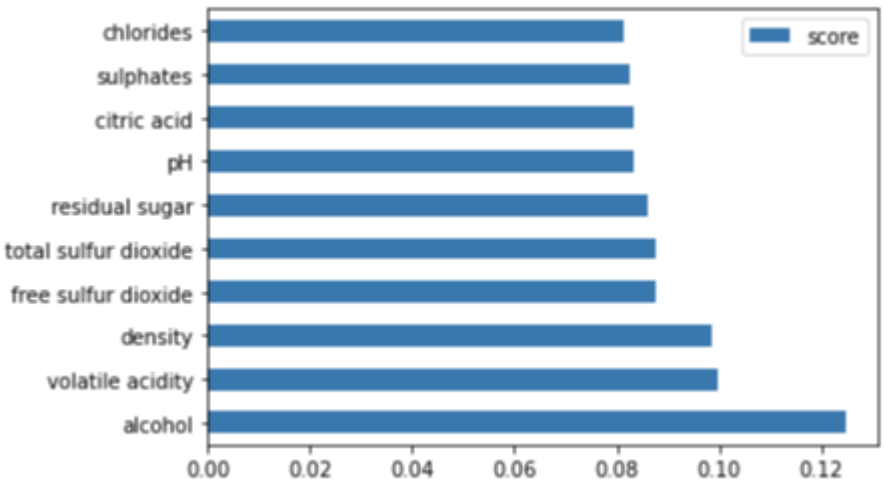
В данных residual sugar есть выбросы:



Экземпляры со значением более 20 можно считать выбросами - удалили их.
На основе матрицы корреляции посмотрим нет ли сильных зависимостей между целевым параметром и остальными



Сильных корреляций нет. Проверили важность признаков через feature_importances



Самая большая ценность у параметра alcohol, остальные параметры имеют практически одинаковое значение.
Исходя из вышеуказанных исследований мы сделали вывод, что будем использовать все параметры для обучения.

Параметр type - категориальный. Мы его преобразовали через OneHotEncoding в числовой.

4. Обучающая модель

В качестве модели обучения выбрали логистическую регрессию.
Модель логистической регрессии - это линейная модель, которая считает вероятность принадлежности объекта тому или иному классу. Модель максимально простая и должны быстро обучиться. Модель показала низкую точность в районе 0.445.
Вероятной причиной может недостаточное количество данных - маленький датасет.

Мы попробовали использовать другу модель: случайный лес - это ансамбль из деревьев решений. Данная модель будет считать дольше чем логистическая регрессия, более склонна к переобучению, но может сработать более точно, чем логистическая регрессия.
Результат обучения этой модели существенно выше 0.67 балла точность на проверочной выборке.

5. Результаты

Итоговым результатом является обученная модель случайного леса, которая предсказывает качества вина с точностью 67%.
Качество невысокое - модель нужно улучшать.

6. Обсуждение и доработки модели

Модель показывает невысокое качество - 67% точность. Чтобы ее применять, ее обязательно нужно доработать.
Одним из вариантов ее улучшения - это генерация новых объектов для обучения. Классы не сбалансированы, имеет смысл искусственно добавить объекты их дублированием в классы, которые имеют крайне низкое количество примеров.
Провести исследование и проработать разные параметры обучающей модели: количество уровней деревьев, максимальное количество объектов в одном листе, и найти оптимальные значения.
Провести исследование и перебрать параметры для обучения. Возможно использование некоторых параметров ухудшает качество модели.

7. Планы на будущее

Необходимо повысить качество модели, применив указанные выше приемы. Качество модели нужно довести до точности 80%. После этого ее можно использовать на производство для автоматической оценки качество вина. Для этого можно разработать инструмент оценки качества вина, который будет использовать обученную модель - в него нужно будет только ввести входные параметры вина.