Предсказание оценки качества вина на основе данных из физико-химических свойств

1. Введение

Данный отчет не является серьезным анализом данных и подготовлен в рамках упражнения по развития навыков правильного структурирования информации и оформления отчета по проведенному анализу данных.

Отчет представлен по результатам анализа влияния физико-химических свойств вина на экспертную оценку качества.
В анализе рассматривается набор данных с красными и белыми португальскими винами "Vinho Verde". Доступны только физико-химические и сенсорные переменные (например, нет данных о типах винограда, марке вина, цене продажи вина и т.д.).

2. Цель проводимого анализа

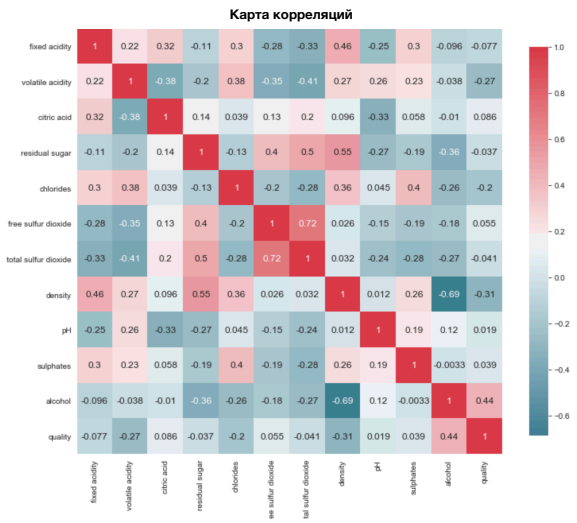
Основная цель - изучить свойства вина, которые наибольшим образом влияют на экспертную оценку качества и определить наиболее подходящий алгоритм для обучения модели, обучить модель предсказывать результаты.

3. Рассматриваемые в анализе физико-химические свойств вина

Всего в датасете 12 физико-химических свойств вина, которые мы будем использовать для предсказания целевого показателя - оценки качества вина.

- type - тип вина (красное/белое)
- fixed acidity - фиксированная кислотность
- volatile acidity - летучая кислотность
- citric acid - лимонная кислота
- residual sugar - остаточный сахар
- chlorides - хлориды
- free sulfur dioxide - свободный диоксид серы
- total sulfur dioxide - общий диоксид серы
- density - плотность
- pH
- sulphates - сульфаты
- alcohol - алкогольность

4. Изучение взаимосвязей между рассматриваемыми переменными и целевым показателем.

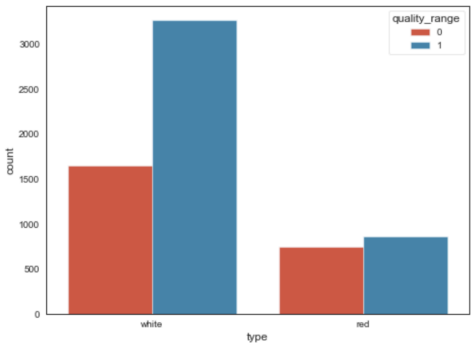


В датасете 12 числовых и одна категориальная переменные.
Наиболее высокую корреляцию с ключевым показателем имеют показатели: alcohol и density (отрицательную).

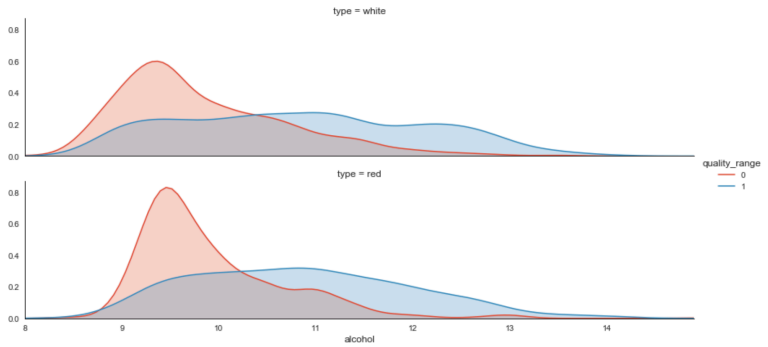
В первую очередь я перевел целевой показатель quality к бинарному виду, чтоб он принимал только два значения (0 и 1).
Далее посмотрел на взаимосвязь показателей

alcohol, type и quality. Как видно из графиков, представленных ниже, в целом в датасете представлено больше белых вин, однако хорошо заметно, что из всех белых вин около 70% получают оценку 5 и более, в то время как у красных вин только чуть более 50.

Также можно сделать вывод, что как и у красного вина, так и у белого, его алкогольность заметно влияет на оценку качества. Менее алкогольные вина получают более низкие оценки.



Распределение вин в зависимости от типа и оценки качества



Зависимость оценки качества вина от алкогольности в разбивке по типам вина

---

5. Подготовка данных

Первоначальный датасет включал в себя 6497 строк и 13 переменных, в том числе целевой показатель. В процессе подготовки данных для обучения модели мной были проделаны следующие шаги:

- категориальная переменная переведена в числовую
- так как большинство алгоритмов машинного обучения требуют, чтобы все переменные имели значения, чтобы использовать их для обучения модели, и так как пустых значений всего около 30, что составляет менее 1% от всего датасета, я их удалил
- целевой показатель был переведен к бинарному виду

В итоге финальный датасет, который я использовал для обучения модели, включает в себя 6463 строки и 14 переменных, в том числе целевой показатель.

---

6. Обучение модели и результат

После разделения датасета на 80% тренировочных данных и 20% тестовых я провел обучение модели несколькими алгоритмами и сравнил полученные результаты:
LinearRegression - 0.2
RandomForestClassifier - 0.83
KNeighborsClassifier - 0.69
XGBClassifier - 0.82
GradientBoostingClassifier - 0.76
DecisionTreeClassifier - 0.8
GaussianNB - 0.65
SVC - 0.64

Вывод: Как видно, наилучший результат показал RandomForestClassifier.
Так как основная цель проводимого анализа данных была тренировка навыков правильного структурирования самого анализа и отчета считаю, что она в той или иной степени достигнута. В дальнейшем хотелось бы проводить более подробный анализ каждой переменной, гораздо больше времени уделить обработке и подготовке данных, а также проводить более осмысленный выбор алгоритмов обучения модели.