Отчет по исследованию Wine quality

1. Введение.

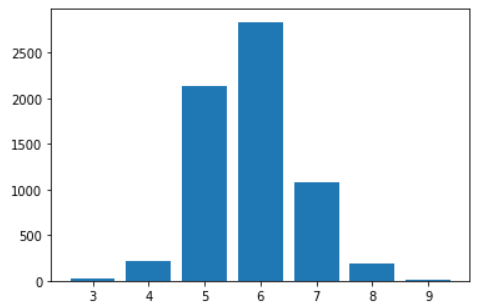
Определение степени качества – важная стадия в процессе подготовки продукта к его реализации. В случае компаний, производящих винную продукцию, наиболее подходящим является экспертный метод оценки качества, включающий органолептические, физико-химические показатели, показатели безопасности и возможную фальсификацию. При этом не всегда наличествуют абсолютно все показатели, по которым можно было бы определить качество продукта, поэтому возникают трудности оценивания. Так в данном проекте возникли проблемы с конфиденциальностью и логистикой, что повлекло за собой потерю массива данных. Доступными оказались только физико-химические (исходные) и сенсорные (выходные) переменные; остальные оказались утеряны.

Цель работы: предсказать экспертную оценку качества вина по шкале от 0 до 9.

1. Описание проблемы и набор данных.

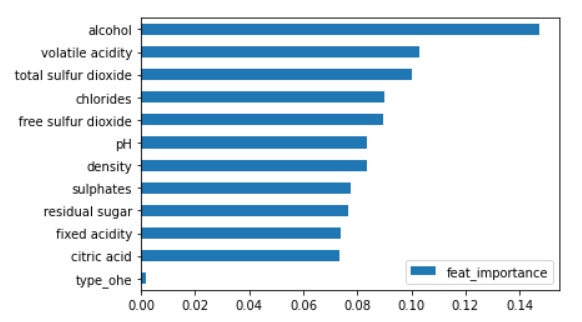
Основной проблемой стали пропуски в данных и дисбаланс классов размеченных данных. Так, например, 0-2 классов в датасете было не представлено, а объем данных класса 6 вместил бы все остальные классы вместе взятые:

Дисбаланс классов целевой переменной



1. Важность фичей и препроцессинг.

В ходе работы было выяснено, что не все признаки одинаково ~~полезны~~ важны для модели.

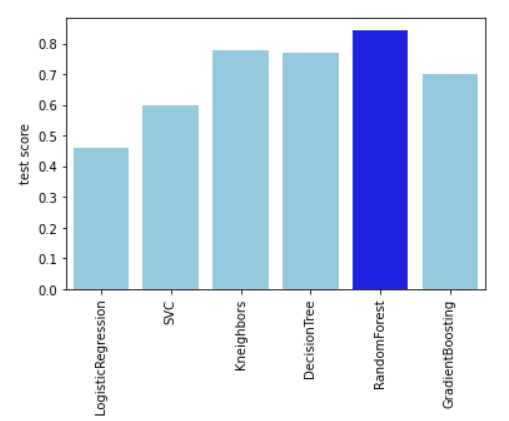
Важность признаков при обучении модели по убыванию

Наиболее важные признаки были предобработаны и выбраны для обучения модели.

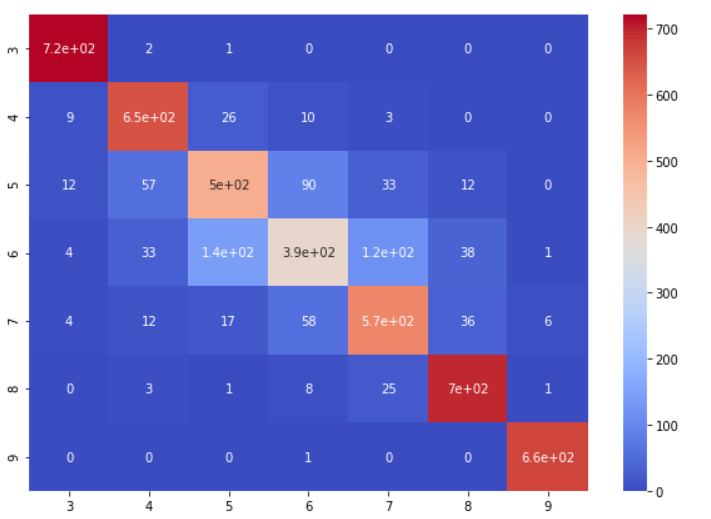
1. Обучение моделей

На данном этапе был осуществлен выбор модели с наилучшим скором.

Сравнение скора тестируемых моделей



1. Результат обучения случайного леса

В 84,6% модель успешно предсказывает верный класс.

* вина класса 5, 6 и 7 модель определяет хуже остальных
* вина классов 3, 8, 9 и 4 модель выявляет лучше всего

1. Рекомендации и планы

В перспективе планируется обучение бустинговых моделей, таких как LightGBM, CatBoost. Возможно, они покажут более высокий скор.