Анализ качества вина по его химическим свойствам

Презентация для аналитиков торговой сети алкомаркетов «Норман»

Ситуация и цели проекта

- Ситуация

1. Через магазины торговой сети «Норман» проходит огромное количество разнообразного вина.
2. Торговая сеть желает более точно оценивать качество вина, которым торгует, чтобы лучше позиционировать его на рынке.
3. Торговая сеть желает иметь возможность оценивать вино, предлагаемое им различными алкогольными компаниями без привлечения каждый раз дорогостоящих экспертов.

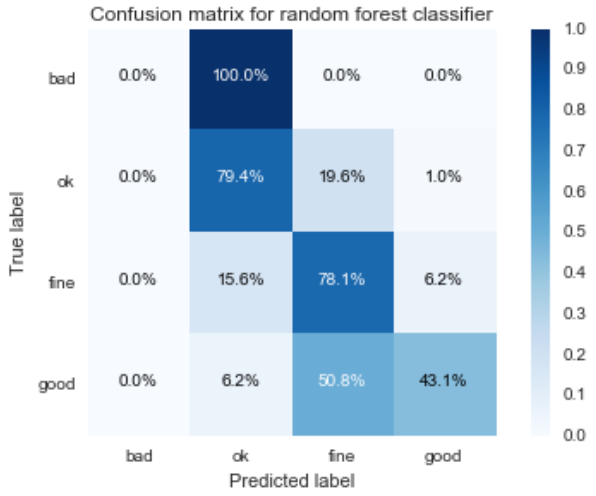
- Цели проекта «Анализ качества вина»

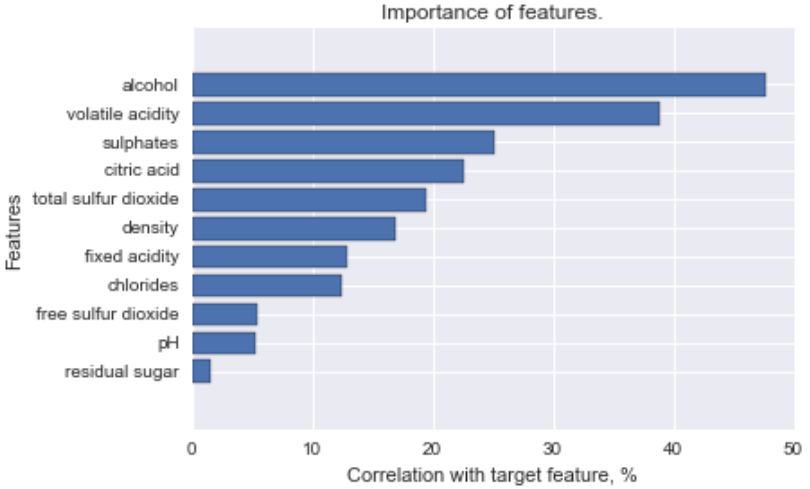
1. Разработать модель, предсказывающую качество вина по его химическим свойствам
2. Качество оценки модели должно быть не ниже, чем у эксперта «средней руки»
3. Модель должна быть достаточно производительной, чтобы иметь возможность достаточно быстро обрабатывать массивы информации о вине, приходящие в торговую сеть

Заключения исполнителя

Создана модель, которая может в течение секунд определить качество вина, что позволяет оптимизировать работу менеджеров по закупкам

- **Модель сравнительно хорошо справляется с определением качества вина**
  1. Хорошо получается определять среднее вино(класс «ok») - точность 75% .
  2. Модель с невысокой точностью определяет хорошее вино(класс «fine») - точность 67% и отличное вино(класс «wine») - точность 67%
  3. Модель не способна определять плохое вино(класс «bad»)
- **Выявлены закономерности, позволяющие на высоком уровне оценивать качество вина :**
  1. Повышенное содержания алкоголя в вине положительно влияет на качество вина (корреляция 47%)
  2. Повышенное содержание летучих кислот отрицательно влияет на качество вина (корреляция 37%)
- **При наличии данных о химических свойствах вина, модель способна определять качество вина в течении нескольких секунд и способна запускаться на любом компьютере.**





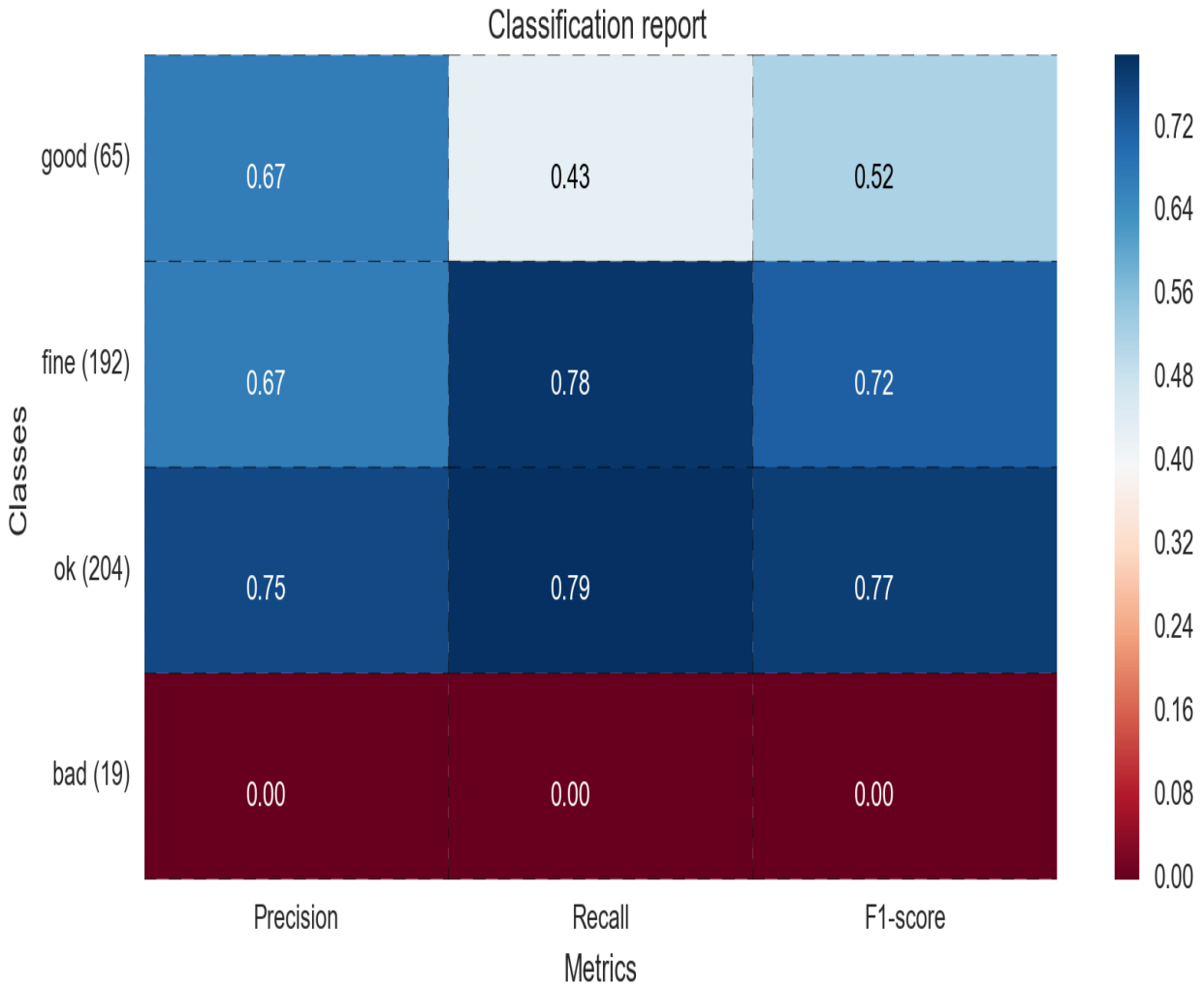
Подход

- **Использованы данные, полученные при исследовании 1599 видов и сортов вин, производимых на севере Португалии**
- **Опробованы различные модели анализа данных для получения наиболее точной оценки**
  - Регрессия зависимой переменной по признакам
  - Классификация вина по качеству с помощью логистической регрессии, дерева решений, случайного леса дерева решений
- **Разработана модель анализа качества вина**
  - Идентифицированы самые значимые при оценке вина параметры
  - Получена удовлетворительная предсказательная сила на основе легко получаемых химимических свойств вина
  - Модель может быть легко адаптирована под большее количество признаков и данных
- **Проведены тестовые запуски модели для определения её производительности**

Описание модели

- **Обзор методологии:**
  - предсказать качество вина(плохое, среднее, хорошее или отличное) лучше, чем эксперт «средней руки»
- **Модель:**
  - случайный лес решающих деревьев
- **Зависимая переменная:**
  - качество вина
- **Выборка:**
  - Тренировочная: 1119 образцов
  - Тестовая: 480 образцов
- **Разработанная модель имеет предсказательную способность не хуже, чем эксперт «средней руки»**
  - В 2001 году Гил Мэротт провел исследования, в котором 54 эксперта по вину определили белое вино, покрашенное красителем в красный цвет, как красное[1]

Ключевые показатели



Детали модели

- **В качестве классификатора выбран случайный лес деревьев решений**
  - Позволяет бороться с недообучением, которое показывает на этих данных логистическая регрессия
  - Позволяет частично бороться с переобучением, которое показывают глубокие дерьевья решений
- **Данные очень сильно несбалансированны**
  - Плохого и отличного вина в сэмпле на порядок меньше чем среднего и хорошего
- **Работа с данными которая проводилась и улучшила модель:**
  - Масштабирование данных
  - Сокращение числа классов
- **Работа с данными, которая проводилась, и не улучшила модель:**
  - Добавление новых полиномиальных признаков из старых
  - Балансировка классов
- **Подбор гиперпараметров проводился по сетке**
  - Поэтому при добавлении новых тестовых данных, нужно будет заново подбирать гиперпараметры

Рекомендации

- **Проведите пилотный запуск модели**
  - Позволит оценить её качество в реальной эксплуатации и обнаружить проблемы на ранней стадии проекта
- **Используйте свои возможности и предоставьте больше образцов вина для модели**
  - Сейчас модель обучена только на данных вина из одного региона, это может снизить её предсказательную способность для вина из других регионов
  - Добавьте признаки, относящиеся к винограду, из которого делали вино
- **Для фильтрации откровенно плохого вина возможно придется проводить дегустации вина, которую модель оценила как среднее**