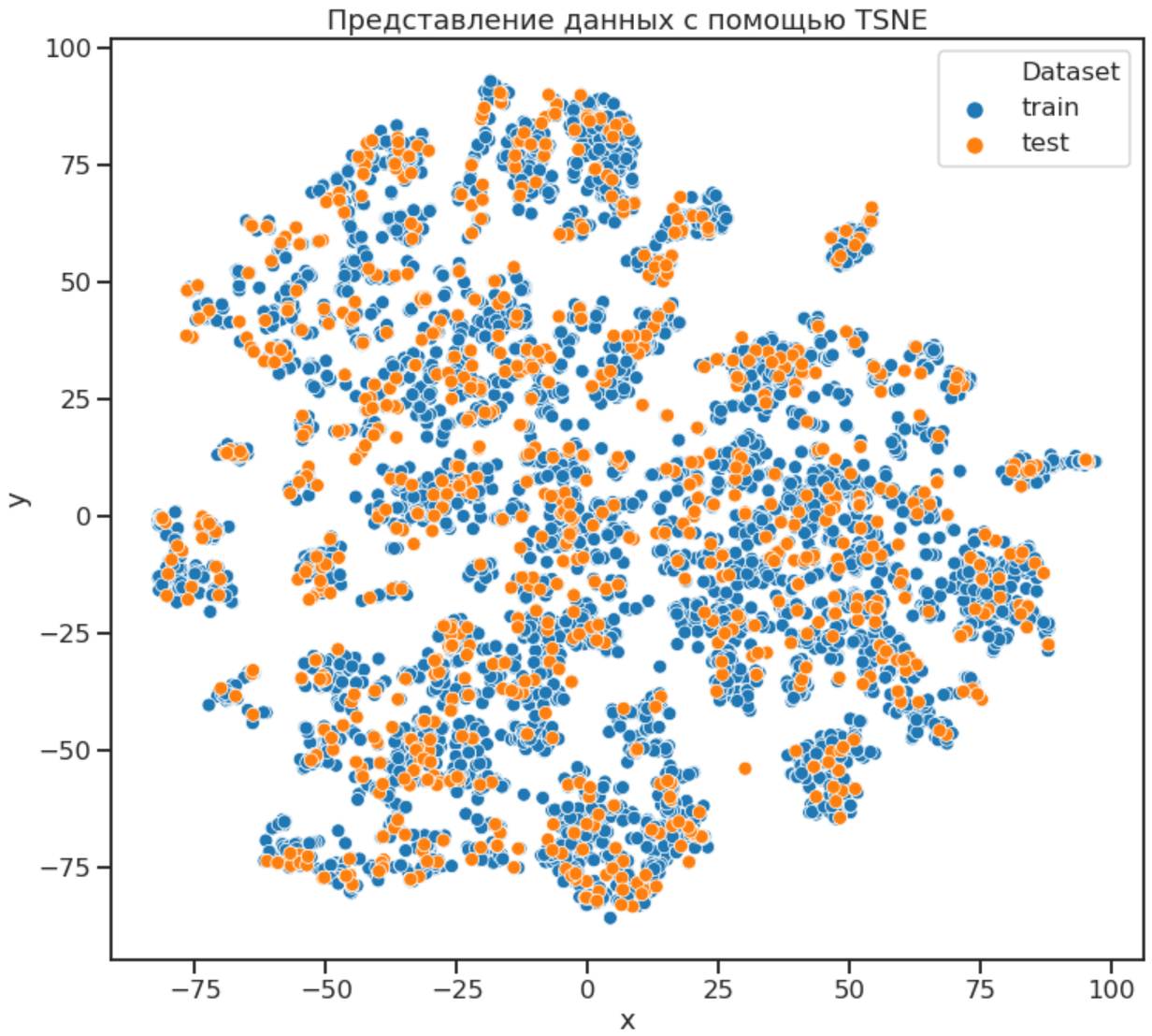
Предсказание калорийности продуктов

Решение команды "Лед под ногами майора"

Предварительный анализ данных

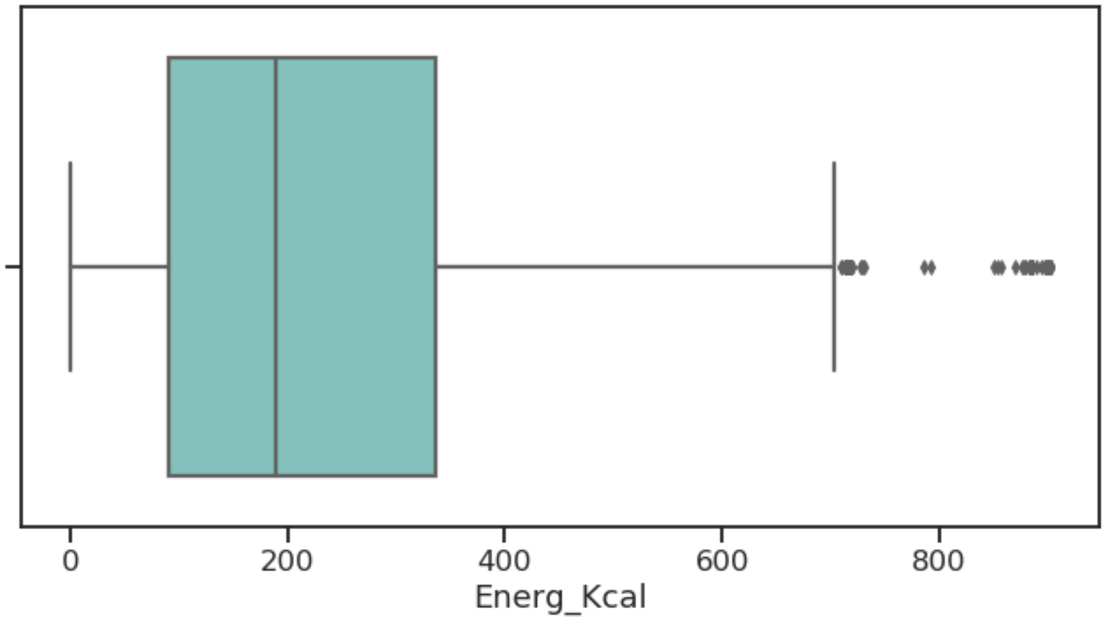


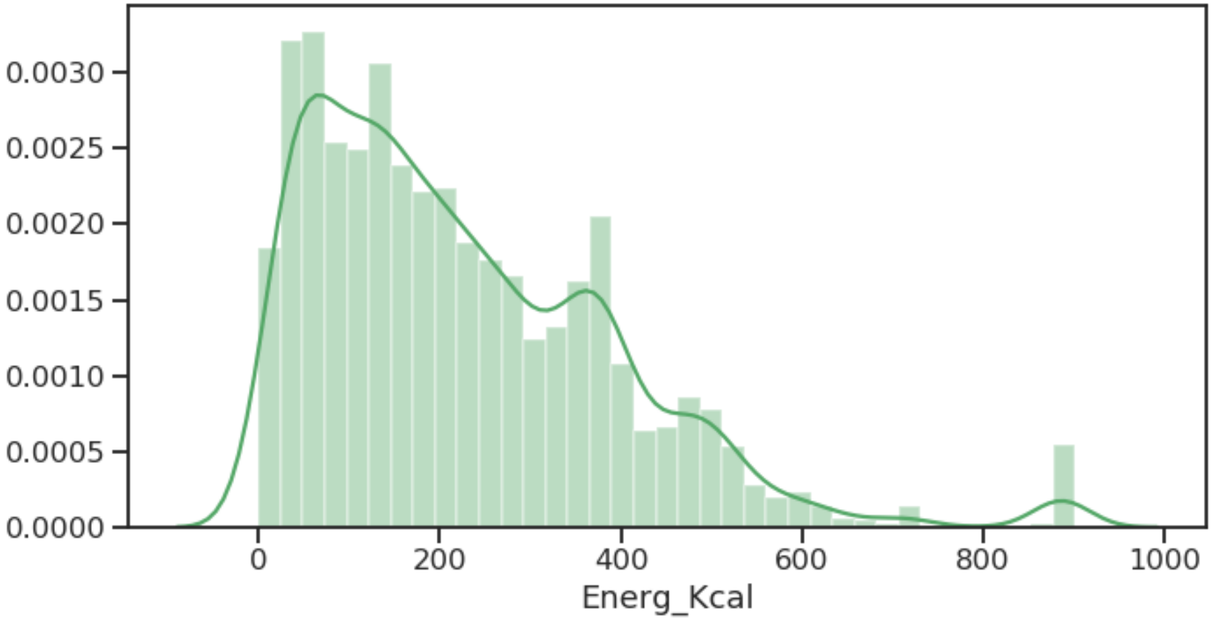
Общее представление данных

- **Похоже, что данные трейна и теста приходят из одного распределения**
- **Особенно четкой структуры кластеров не наблюдается**

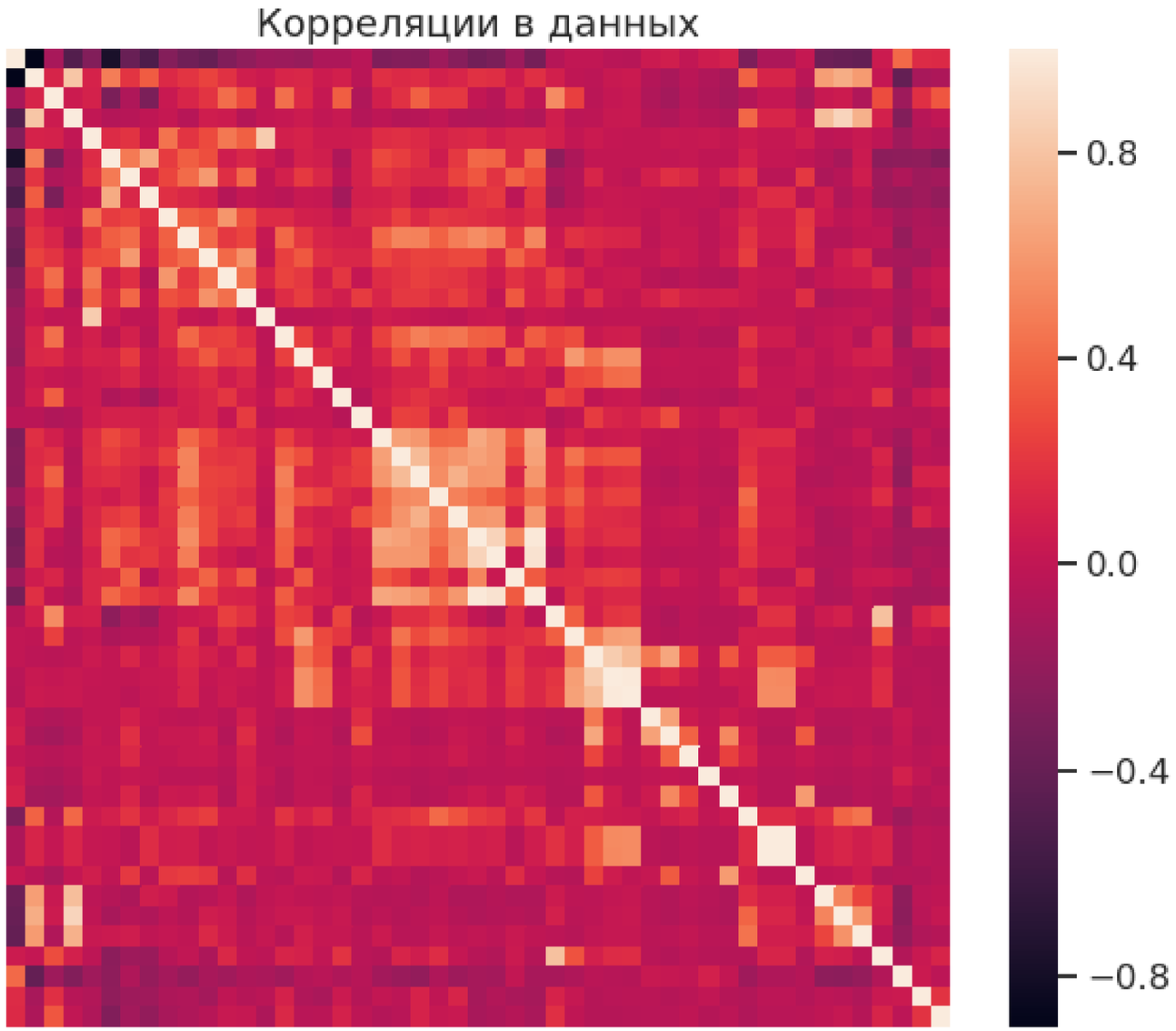
Распределение целевой переменной

- Распределение даже близко не равномерное
- На нормальное тоже не похоже
- Особенных выбросов нет





Поиск корреляций



- **Сильно коррелирующие пары есть, но их не больше десятка**
- **С целевой переменной больше всего коррелирует содержание воды (в отрицательную сторону) и жиров**

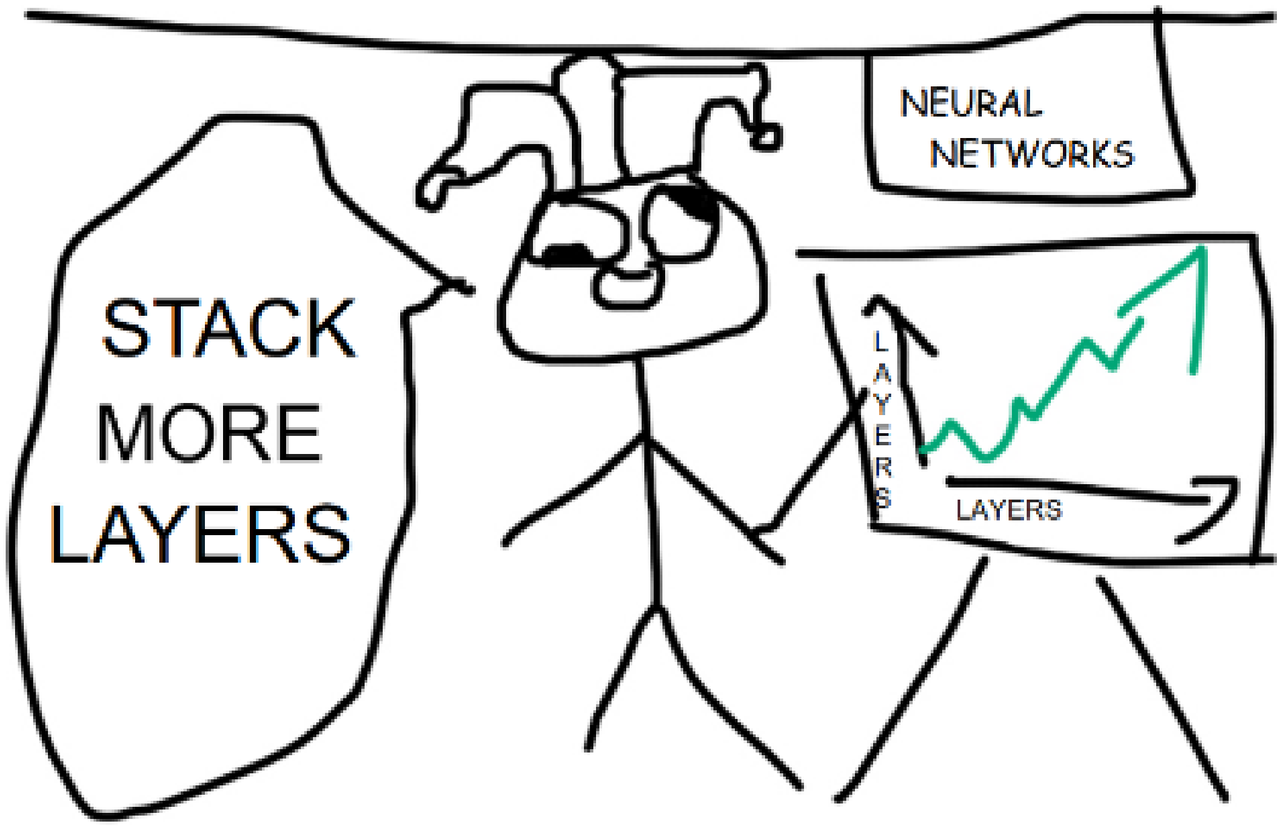
Признаки и модели

Признаки для основной и дополнительных задач

- Полиномиальные признаки
- Обработка пропусков медианами
- TF-IDF для описаний
- Mean Encodings для признаков GmWt
- Используем нормализацию признаков
- Отбрасываем сильно коррелирующие признаки

Модель для основной и дополнительной задачи №2

- Используется двухуровневый стакинг
- На первом уровне: KNN, градиентный бустинг, случайный лес, линейная регрессия
- На втором уровне: линейная регрессия



Решение дополнительной задачи №1

- Создаем корпус текстов на основе кулинарных книг
- Обучаем word2vec на этом корпусе
- Объектам из трейна присваиваем метки классов на основе косинусного расстояния между названиями меток и описания объектов (Shrt_Desc)
- Используем уже описанные признаки и новую модель (снова двухуровневый стакинг)

Результаты

- **MAE по основной задаче на кросс-валидации: 5.35**
- **MAE по дополнительной задаче №2 на кросс-валидации: 7.19**
- **Accuracy на дополнительной задаче №1 - около 0.51**

- **Для завершения необходимо выполнить более качественный подбор гиперпараметров**
- **Провести больше экспериментов с признаками и моделями**

Это все, спасибо за внимание