Вы – аналитик в компании **"MarineAge Analytics"**, специализирующейся на разведке и селекции морских улиток (абалонов) для дальнейшей переработки в украшения и деликатесы. Компания недавно инвестировала в современные технологии для оптимизации производственного цикла и контроля качества продукции. Одной из ключевых задач стало определение возраста абалона, поскольку именно от возраста зависит его ценность и пригодность для дальнейшей обработки.

Измерения абалонов (длина, диаметр, высота, вес и прочее) собраны в датасете, который вы получили от отдела контроля качества. Ваша цель – на основе этих данных разработать модель регрессии, способную предсказывать возраст абалона (определяемый количеством колец). Это поможет компании оптимизировать процессы селекции, закупок и продаж, а также повысить качество конечного продукта.

Задание

Создайте полный пайплайн анализа и моделирования для задачи регрессии на датасете и ***реализуйте 3 итерации улучшения модели***. Работа должна быть оформлена в виде Jupyter Notebook или Orange + презентация с подробными комментариями и обоснованием каждого решения и шага.

Обязательные этапы пайплайна:

1. **EDA (Exploratory Data Analysis):**
    - Проведите визуализацию распределения признаков и целевой переменной.
    - Определите наличие пропущенных значений, выбросов, нелинейных зависимостей.
    - Сформулируйте гипотезы по влиянию различных признаков на возраст абалона.
2. **Предобработка данных:**
    - Обработка пропущенных значений (если они есть) и выбросов.
    - Применение масштабирования к числовым признакам.
    - Feature Engineering: создание новых признаков, которые, по вашему мнению, могут улучшить качество модели.
3. **Разделение на обучающую и тестовую выборки:**
    - Разделите данные на train/test (например, в пропорции 80/20).

4. **Выбор метрики:**

    - Обоснуйте выбор метрики (например, MSE, RMSE или MAE) для оценки качества модели регрессии.

5. **Выбор и обучение модели:**

    - Выберите интересную вам модель (для первой и второй итерации) регрессии из рассмотренных на семинаре. На третьей итерации рассмотрите 3 модели
    - Постройте модели, обучите и оцените качество на тестовой выборке.

6. **Измерение результата:**

    - Подведите итоги по выбранной метрике, сравните с базовым значением (baseline).

Итеративный процесс (3 итерации)

Итерация 1 (Baseline):

- Запустите модель без предварительной обработки данных (без масштабирования, без feature engineering).
- Зафиксируйте базовый результат по выбранной метрике.
- Объясните, почему базовый подход может иметь ограничения.

Итерация 2:

- Проведите полноценный EDA, выявите проблемы с данными (выбросы, несбалансированность распределения и т.д.).
- Примените первичный этап предобработки: масштабирование числовых признаков, очистку выбросов, обработку пропущенных значений (если имеются).
- Выполните базовый feature engineering (например, создание одного–двух новых признаков, выявленных на основе EDA).
- Обучите модель, измерьте и сравните результаты с итерацией 1.
- Объясните, какие изменения внесены, как они повлияли на качество модели, и почему эти изменения были логичными в контексте изученных данных.

Итерация 3:

- Проведите более углублённый анализ: попробуйте более сложные методы feature engineering (например, полиномиальные признаки, взаимодействия между признаками, возможно, использование логарифмических преобразований, если распределения сильно смещены).
- Проведите подбор и оптимизацию гиперпараметров *выбранных моделей* с использованием кросс-валидации.
- Снова измерьте результаты и сравните с предыдущими итерациями.
- Объясните, какие дополнительные шаги были предприняты и как они должны были способствовать улучшению метрики.

Общие требования:

- **Документирование:** На каждом этапе в ноутбуке должны быть чёткие выводы и обоснования принятых решений.
- **Использование масштабирования и feature engineering обязательно:** Объясните, почему выбран конкретный метод масштабирования и какие новые признаки были созданы.
- **Каждая итерация должна демонстрировать улучшение:** Результаты улучшения (или анализ их отсутствия) должны быть обоснованы, с обсуждением возможных причин.
- **Обоснование выбора метрик:** Почему выбрана та или иная метрика, и как она помогает оценить качество модели с точки зрения бизнес-задачи.

Критерии:

- Реализован весь пайплайн с 3-мя итерациями (макс. *6 баллов*)
- Обоснование всех шагов и принятых решених (макс. *2 балла)*
- Использование advanced методов (подбор гиперпараметров у моделей и методы фича инжиниринга, которые не обсуждали на семинаре) с прикреплением ссылок (макс. *2 балла*)

Итоговая цель

Создать рабочий, задокументированный пайплайн, который будет включать полный цикл от EDA до финальной оптимизированной модели регрессии.

Описание данных

- **Sex**: Feature, Categorical (M, F, and I (infant))
- **Length**: Feature, Continuous (Longest shell measurement), mm
- **Diameter**: Feature, Continuous (Perpendicular to length), mm
- **Height**: Feature, Continuous (With meat in shell), mm
- **Whole_weight**: Feature, Continuous (Whole abalone), grams
- **Shucked_weight**: Feature, Continuous (Weight of meat), grams
- **Viscera_weight**: Feature, Continuous (Gut weight (after bleeding)), grams
- **Shell_weight**: Feature, Continuous (After being dried), grams
- **Rings**: Target, Integer (+1.5 gives the age in years)