МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное
Учреждение высшего образования
«ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт высоких технологий и пьезотехники

Кафедра прикладной информатики и инноватики

Направление подготовки:
09.03.03 "Прикладная информатика"

Отчёт

По дисциплине: «Большие данные»

Анализ данных по популяциям крабов

Выполнили студенты 3 курса 21ВТ-09.03.03.01-о3 группы:

_____ Рубашевская А. А
подпись

_____ Бугай Д. А
подпись

Проверил старший преподаватель _____ Яценко Д. В
подпись

Ростов-на-Дону – 2024

СОДЕРЖАНИЕ

1. Постановка задачи

Данный кейс предлагает задачу регрессионного анализа возраста крабов, что предполагает оценку различных физических характеристик крабов и их взаимосвязи с возрастом.

Задача регрессии заключается в предсказании непрерывной величины на основе входных данных. Она является одной из самых распространенных задач в машинном обучении и имеет широкое применение в различных областях.

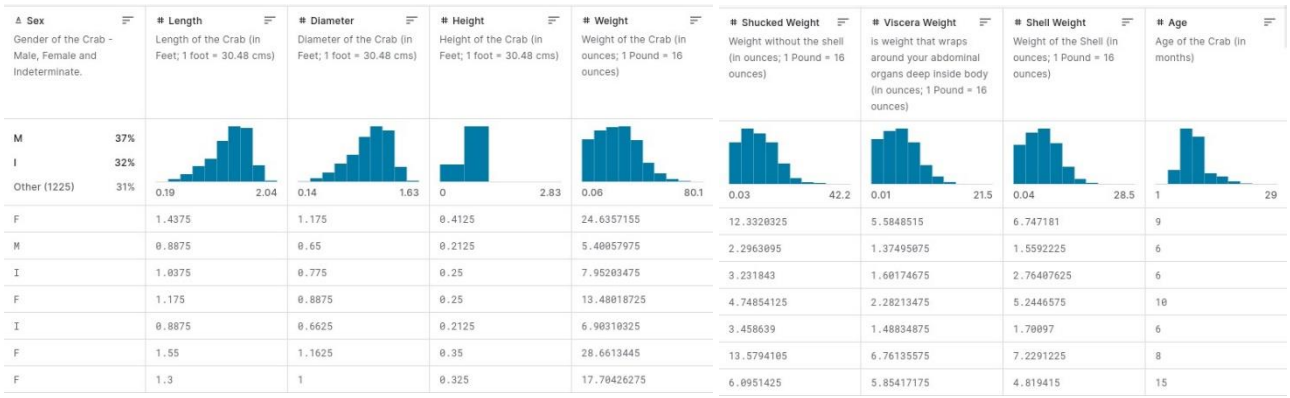
Исходя из этого, можно провести следующий анализ:

- Ознакомиться с датасетом, изучить все его поля и визуализировать;
- Предсказать возраст крабов с помощью метрических признаков различными методами регрессии и проанализировать их эффективность на датасете CrabAgePrediction;
- Проанализировать признаки и определить наиболее влиятельный из них;
- Визуализировать упомянутый датасет и результаты работы.

2. Описание датасета

Датасет CrabAgePrediction содержит 3894 записи и представляет собой набор метрических данных о крабах, а также информацию о поле и возрасте. В нем содержатся следующие поля:

1. Sex - Пол краба (мужской, женский или неопределенный)
2. Length - Длина краба (в лапах; 1 фут = 30,48 см)
3. Diameter - Диаметр краба (в лапах; 1 фут = 30,48 см)
4. Height - Высота краба (в лапах; 1 фут = 30,48 см)
5. Weight - Вес краба (в унциях; 1 фунт = 16 унций)
6. Shucked Weight - Вес краба без оболочки (в унциях; 1 фунт = 16 унций)
7. Viscera Weight - Вес брюшной полости (в унциях; 1 фунт = 16 унций)
8. Shell Weight - Вес оболочки (в унциях; 1 фунт = 16 унций)
9. Age - Возраст краба



Sex Gender of the Crab - Male, Female and Indeterminate.	Length Length of the Crab (in Feet; 1 foot = 30.48 cms)	Diameter Diameter of the Crab (in Feet; 1 foot = 30.48 cms)	Height Height of the Crab (in Feet; 1 foot = 30.48 cms)	Weight Weight of the Crab (in ounces; 1 Pound = 16 ounces)	Shucked Weight Weight without the shell (in ounces; 1 Pound = 16 ounces)	Viscera Weight is weight that wraps around your abdominal organs deep inside body (in ounces; 1 Pound = 16 ounces)	Shell Weight Weight of the Shell (in ounces; 1 Pound = 16 ounces)	Age Age of the Crab (in months)
M 37% I 32% Other (1225) 31%	0.19 2.04	0.14 1.63	0 2.83	0.06 80.1	0.03 42.2	0.01 21.5	0.04 28.5	1 29
F	1.4375	1.175	0.4125	24.6357155	12.3320325	5.5848515	6.747181	9
M	0.8875	0.65	0.2125	5.40057975	2.2963095	1.37495075	1.5592225	6
I	1.0375	0.775	0.25	7.95203475	3.231843	1.60174675	2.76407625	6
F	1.175	0.8875	0.25	13.48018725	4.74854125	2.28213475	5.2446575	10
I	0.8875	0.6625	0.2125	6.90310325	3.458639	1.48834875	1.70097	6
F	1.55	1.1625	0.35	28.6613445	13.5794105	6.76135575	7.2291225	8
F	1.3	1	0.325	17.70426275	6.0951425	5.85417175	4.819415	15

Рисунок 1. Содержание датасета CrabAgePrediction

3. Ход работы

3.1 Гипотеза

Исходя из описания датасета, была определена целевая переменная – возраст, а также признаки, на основе которых и будет строиться предсказание.

В ходе анализа поставленной задачи были сформулированы две основные гипотезы:

1. Существует статистически значимая взаимосвязь между определенными физическими характеристиками крабов и их возрастом;
2. Разработанные модели машинного обучения способны с определенной точностью прогнозировать возраст крабов на основе их физических характеристик.

Первым этапом работы являлось визуализация данных датасета для определения наличия аномалий в виде диаграммы рассеяния.

После первоначального анализа и нахождения примеров случаев аномального возраста, следующим этапом работы являлась предобработка исходных данных путем индексации. Она необходима из-за наличия в датасете категориальных данных (Sex - Пол краба), которые впоследствии были преобразованы так, что новый набор данных состоял из непрерывных величин.

Затем на основе обновленных данных были построены различные модели машинного обучения (RandomFrorestRegressor, GBTRegressor, LinearRegression, DecisionTreeRegressor, IsotonicRegression, FMRegressor, GeneralizedLinearRegression), которые впоследствии предсказывали возраст крабов с определенной точностью. Она оценивалась с помощью метрики MSE (среднеквадратичная ошибка), которая измеряет среднюю разницу между фактическим значением целевой переменной и ее прогнозом.

Финальным этапом является сбор полученных данных, которые были впоследствии структурированы и сохранены для дальнейшего построения отчета, включающего в себя различные диаграммы.

3.2 Визуализация данных по популяциям крабов

Построив диаграмму рассеяния по датасету CrabAgePrediction мы можем наблюдать следующие аномалии:

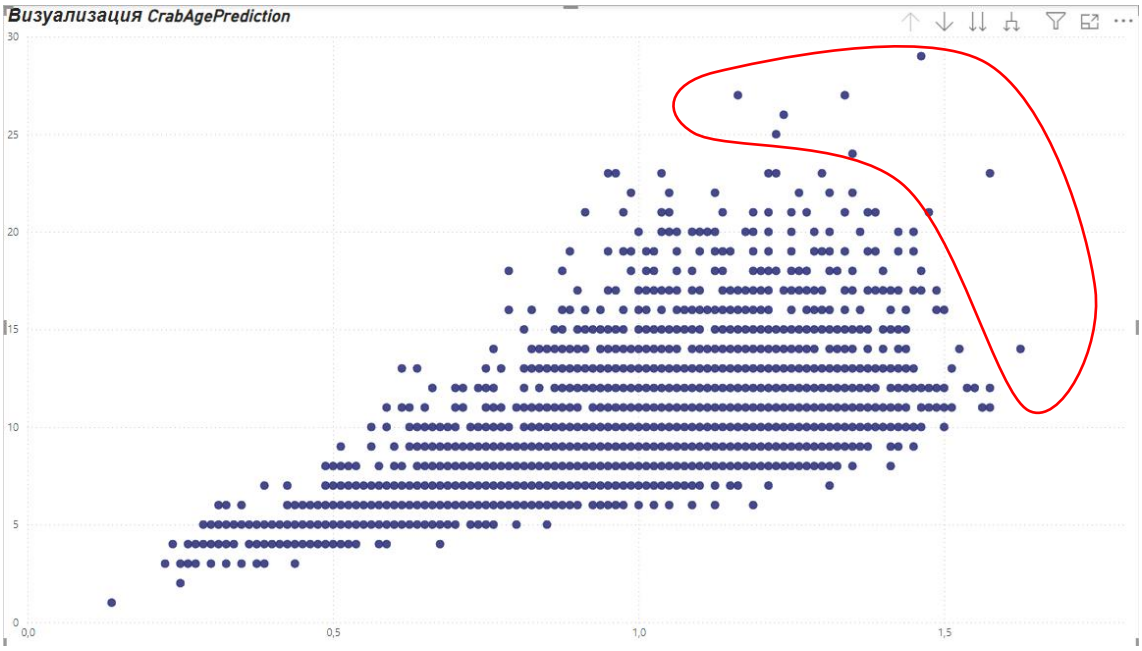


Рисунок 2. Визуализация датасета для дальнейшего анализа

На данном изображении по оси Y – отложен возраст крабов в месяцах, на оси X – все остальные параметры исходного датасета.

На данном изображении мы можем наблюдать то, что аномалий не настолько много, чтобы они могли оказать значительное влияние на результаты предсказания различными моделями машинного обучения.

3.3 Статистические данные

Анализируя представленные ниже круговые диаграммы, мы выявили наиболее значимые физические характеристики крабов для некоторых моделей машинного обучения:

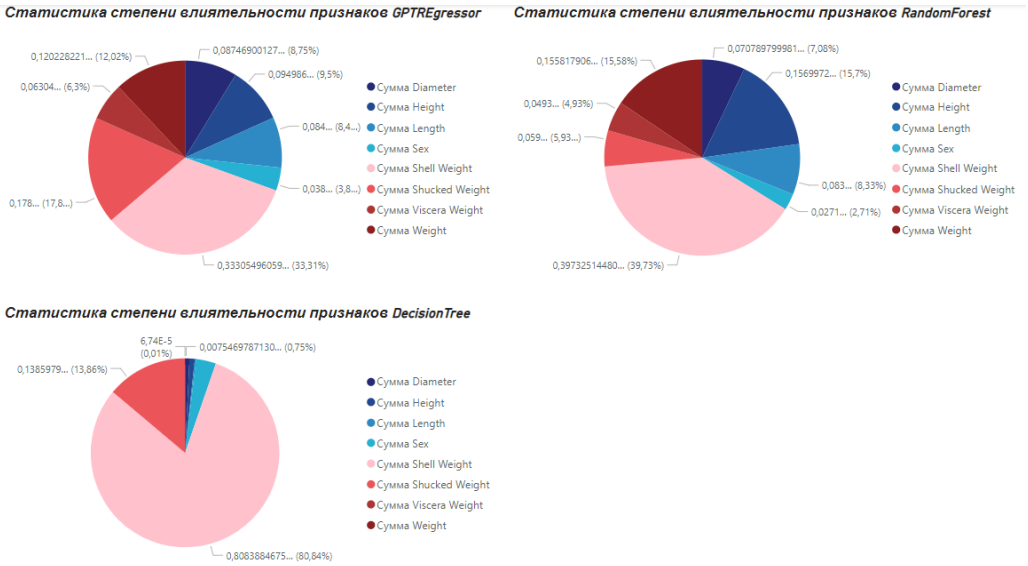


Рисунок 3. Оценка степени влиятельности признаков

Наиболее значимым признаком для всех представленных моделей является вес панциря (Shell Weight).

3.4 Результаты проделанной работы

В результате получаем следующие таблицы:

Таблица 1. Результаты предсказаний возраста крабов каждой модели

Age	predictionsRF	predictionsGBT	predictionsLR	predictionsDT	predictionsIR	predictionsFM	predictionsGLR
7	6.332898894704909	6.3856592052447505	7.160321089352263	6.363636363636363	12.0	1.1590307649514169	6.737955190427437
7	7.2443224174439775	8.328145290587976	7.284946146650438	8.306122448979592	12.0	1.2226059413432528	6.862932318748824
6	7.127882315489858	8.328145290587976	7.206448484146021	8.306122448979592	12.0	1.2993159746448828	6.740405495766632
6	7.8158080713338505	8.328145290587976	7.501189937532092	8.306122448979592	12.0	1.524598838368934	7.236090133153725
9	8.581493957648528	8.328145290587976	7.642435500897497	8.306122448979592	12.0	1.4928202500865342	7.490798254321438

only showing top 5 rows

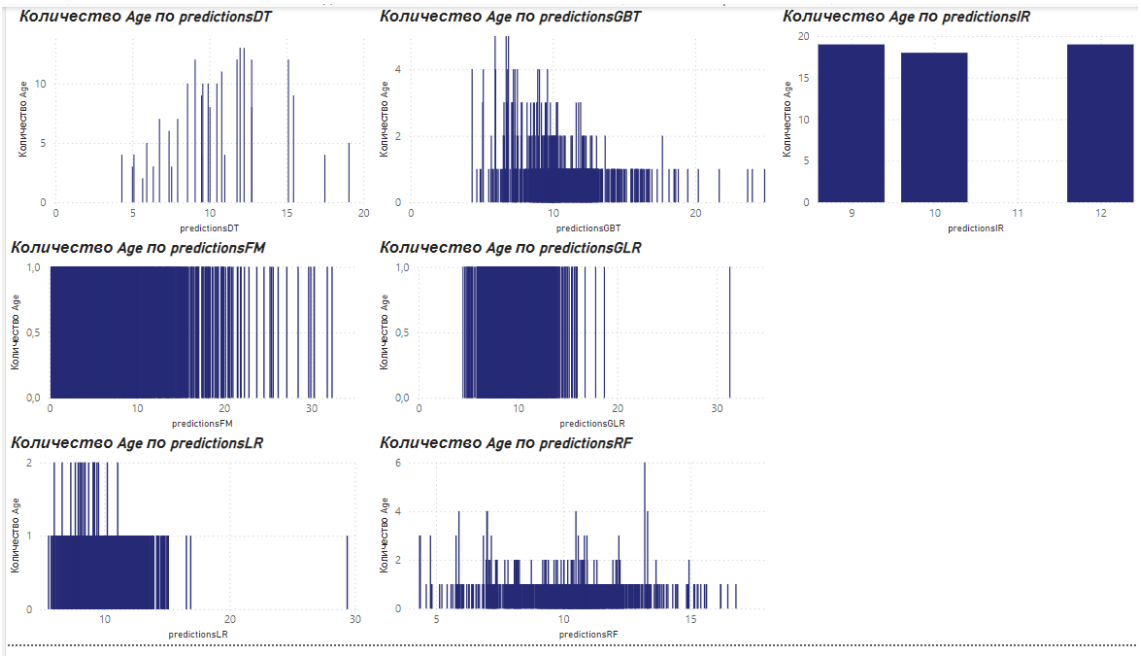


Рисунок 4. Результаты предсказаний возраста каждой модели машинного обучения

Таблица 2. Значения среднеквадратичной ошибки каждого метода

RandomForestRegressor	GBTRegressor	LinearRegression	DecisionTreeRegressor	IsotonicRegression	FMRegressor	GeneralizedLinearRegression
2.260287373287712	2.2803200991143204	2.47864214309764	2.3097527434255354	3.3033172371601327	5.368003698897728	2.283799410499867

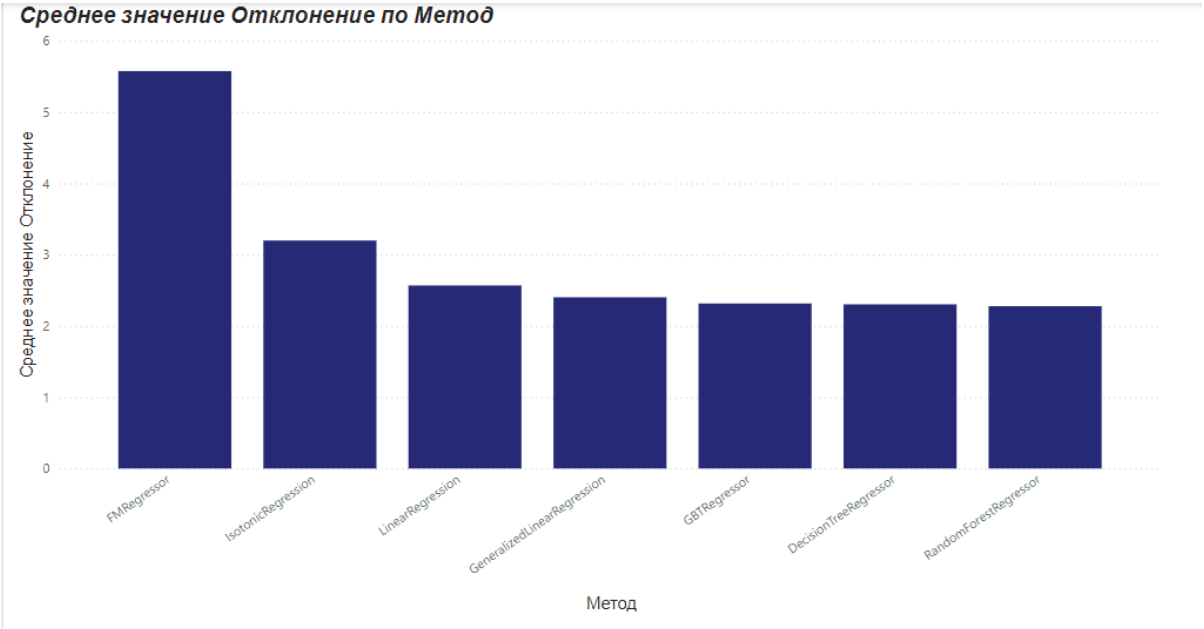


Рисунок 5. Значения среднеквадратичной ошибки каждого метода

4. Выводы

В ходе решения задачи, было сделано:

- Проанализирован и визуализирован датасет CrabAgePrediction;
- Предсказан возраст крабов с помощью метрических признаков различными методами регрессии и проанализирована их эффективность;
- Проанализированы признаки и определен наиболее влиятельный из них;
- Визуализированы результаты работы.

Исходя из этого можно подчеркнуть, что гипотезы, выдвинутые ранее, подтвердились. Цели, которые ставились в ходе проекта, выполнены.

5. Список литературы

1. Kaggle CrabAgePrediction [Электронный ресурс] – URL: https://www.kaggle.com/datasets/sidhus/crab-age-prediction/
2. Apache Spark Classification and regression [Электронный ресурс] – URL: https://spark.apache.org/docs/latest/ml-classification-regression.html/