HOW OLD IS MR KRABS?

Optimize your farming with Crab Age Prediction

ROZZ CHARLES BANQUERIGO LT3

Why is this important?

FOR A COMMERCIAL MUD CRAB **"ALIMANGO"** (SCYLLA SERRATA) FARMER KNOWING THE RIGHT AGE OF THE CRAB HELPS THEM DECIDE IF AND WHEN TO HARVEST THE CRABS. BEYOND A CERTAIN AGE, THERE IS NEGLIGIBLE GROWTH IN CRAB'S PHYSICAL CHARACTERISTICS AND HENCE, IT IS IMPORTANT TO TIME THE HARVESTING TO REDUCE COST AND INCREASE PROFIT.

IDEAL WEIGHT
700G - 1KG MALE
400-500G FEMALE
HARVEST TIME 4-6 MONTHS

Mud Crab Farming in The Philippines

1

MOST IMPORTANT CRAB IN THE COUNTRY DUE TO ITS HUGE DEMAND LOCALLY AND ABROAD

2

CRAB SEEDS ARE ABUNDANT IN: CATARMAN, SAMAR; APARRI, CAGAYAN; BULAN, SORSOGON; MASBATE AND CAMARINES SUR IN PANAY ISLAND, BORATOC NUEVO AND DUMANGAS, ILOILO; PONTAVEDRA, CAPIZ; AND NEW WASHINGTON, AKLAN

3

POLYCULTURE GROWN WITH MILKFISH (BANGUS) OR TIGER PRAWNS

DATASET

7 Features

Sex - Gender of crab (M, F, I)

Length - Lenght of crab in Feet

Diameter - Diameter of crab in Feet

Height - Height of crab in Feet

Weight - Weight of crab in Ounces

Shucked Weight - Weight of crab without shell in Ounces

Viscera Weight - is the weight that wraps around your abdominal organs deep inside the body in Ounces

Shell Weight -Weight of shell in Ounces

Target

Age - Age of crab in months

	Sex	Length	Diameter	Height	Weight	Shucked Weight	Viscera Weight	Shell Weight	Age
0	F	1.4375	1.1750	0.4125	24.635715	12.332033	5.584852	6.747181	9
1	M	0.8875	0.6500	0.2125	5.400580	2.296310	1.374951	1.559222	6
2	I	1.0375	0.7750	0.2500	7.952035	3.231843	1.601747	2.764076	6
3	F	1.1750	0.8875	0.2500	13.480187	4.748541	2.282135	5.244657	10
4	I	0.8875	0.6625	0.2125	6.903103	3.458639	1.488349	1.700970	6
5	F	1.5500	1.1625	0.3500	28.661344	13.579410	6.761356	7.229122	8

	Length	Diameter	Height	Weight	Shucked Weight	Viscera Weight	Shell Weight	Age
count	3893.000000	3893.000000	3893.000000	3893.000000	3893.000000	3893.000000	3893.000000	3893.000000
mean	1.311306	1.020893	0.349374	23.567275	10.207342	5.136546	6.795844	9.954791
std	0.300431	0.248233	0.104976	13.891201	6.275275	3.104133	3.943392	3.220967
min	0.187500	0.137500	0.000000	0.056699	0.028349	0.014175	0.042524	1.000000
25%	1.125000	0.875000	0.287500	12.672227	5.343881	2.664853	3.713785	8.000000
50%	1.362500	1.062500	0.362500	22.792998	9.539607	4.861939	6.662133	10.000000
75%	1.537500	1.200000	0.412500	32.786197	14.273973	7.200773	9.355335	11.000000
max	2.037500	1.625000	2.825000	80.101512	42.184056	21.545620	28.491248	29.000000

ML PIPELINE

EDA → Run Models → Analysis of Output → Feature Engineering / Parameter Tuning → Output Good?

- NULL CHECKS
- DATA DISTRIBUTION
- CORRELATIONS
- BOX PLOTS

- GridSearchCV Pipeline
- OneHotEncoding
- 3 Different Runs
  - Normalizer
  - StandardScaler
  - MinMaxScalar
- KNN
- LinearRegression
- Ridge
- Lasso
- DecistionTreesRegressor
- RandomForestRegressor
- GradientBoostingRegressor
- AdaBoostRegressor
- XGBRegressor
- LGBMRegressor
- CatBoostRegressor

- Summary Table
  - Best Scores
  - Best Parameters
  - Run Time

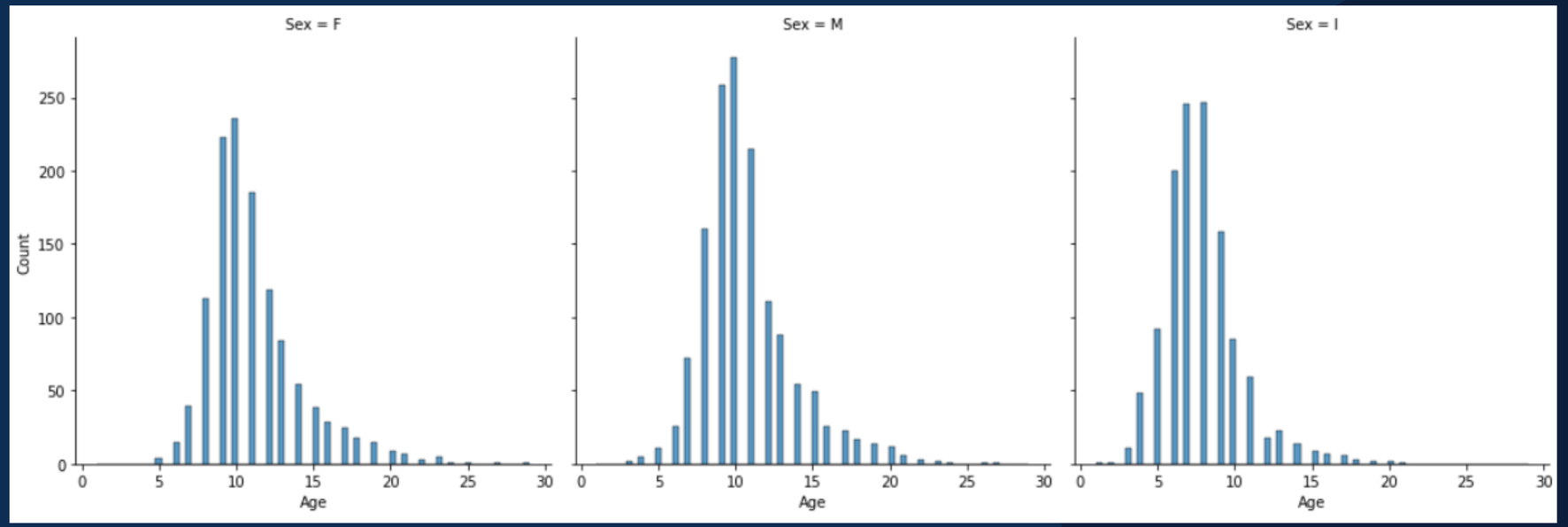
- Added Derived Features
- Added Derived Features

Output Good? → Model Selection and Deployment

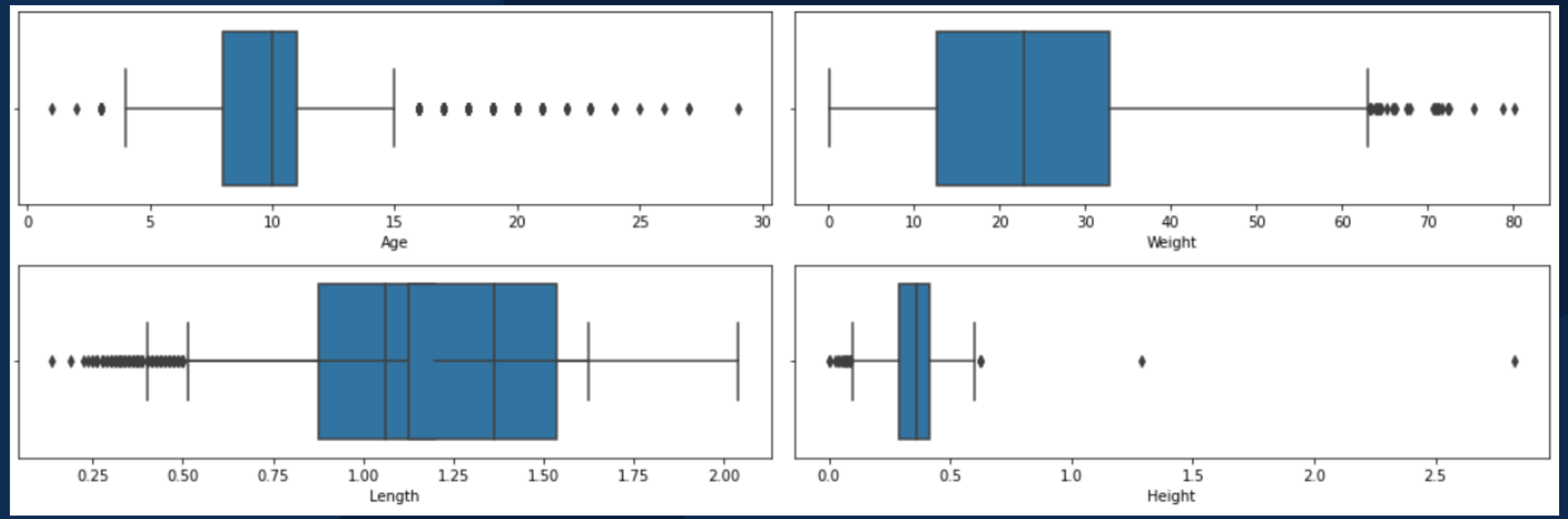
TRAIN 70% VAL 20%
TEST 10%

TRAIN 75% TEST 25%

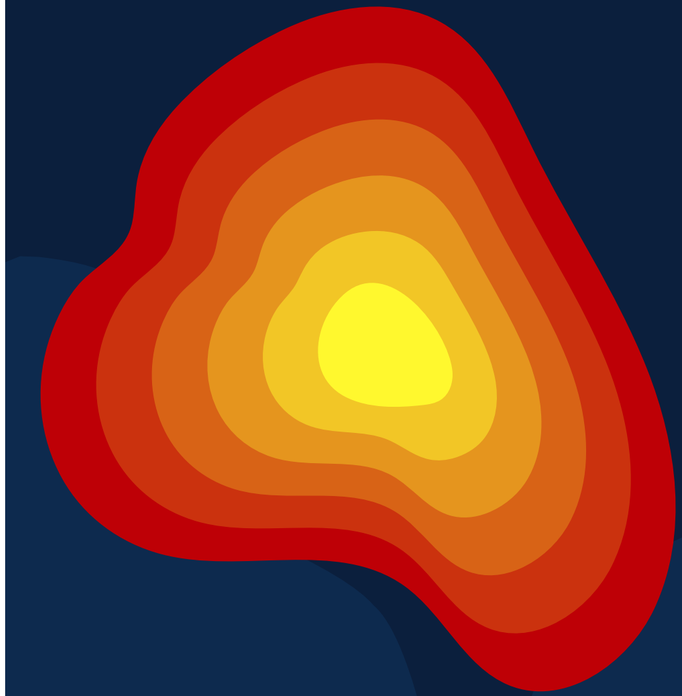
Age Distribution

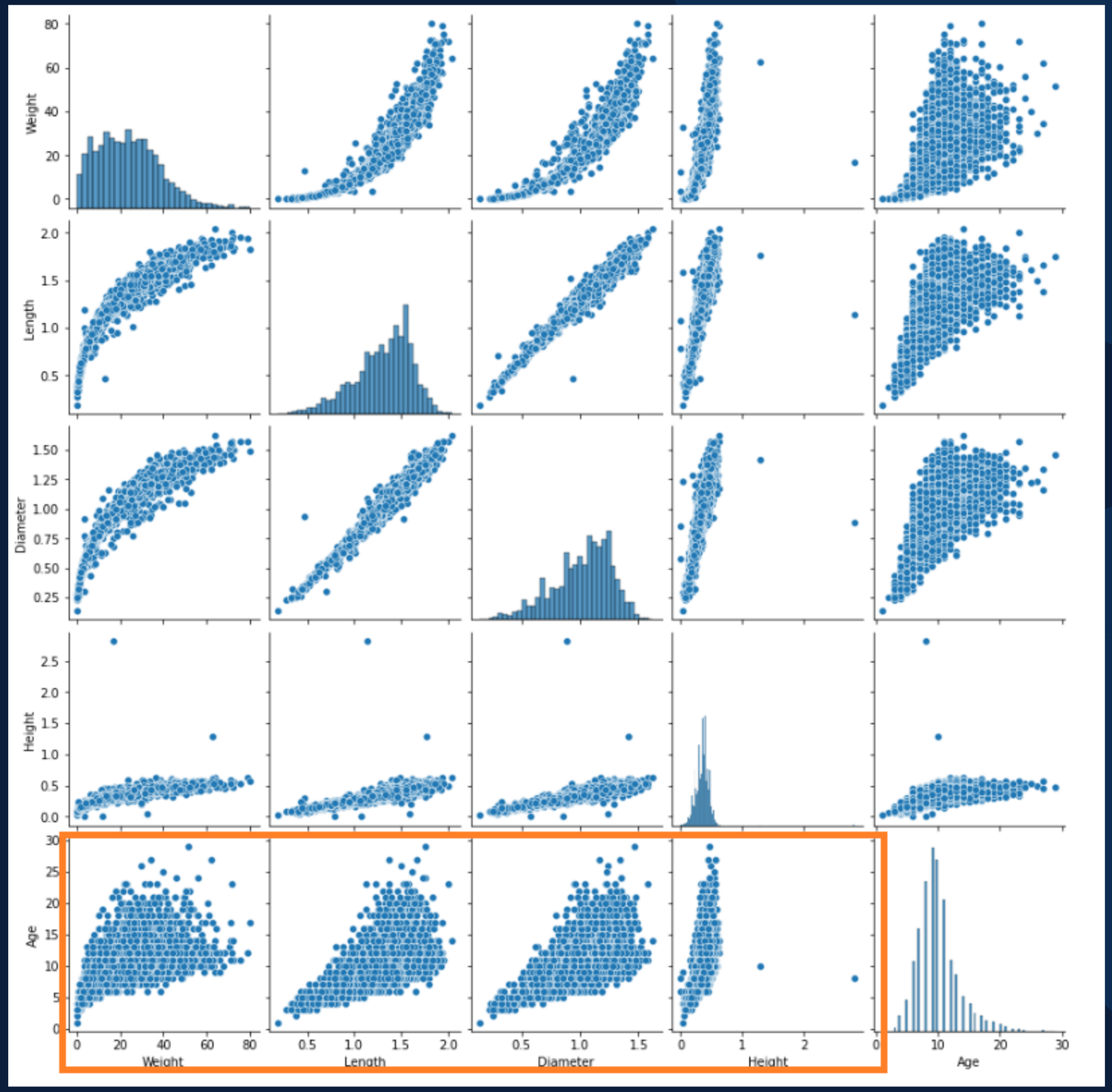


Box Plot



Correlation Heatmap





Pairplot

WE CAN SEE MULTICOLINEARITY BETWEEN THE FEATURES

HEIGHT FLATTENS OUT

AGE DOES NOT CLEAR RELATIONSHIP AS THE CRAB BCEOMES OLDER

	Model	Scaler	Train Score	Validation Score	Test Score	Top Predictor	Fit Time	Score Time
8	kNN	Normalizer()	1.000000	0.561860	0.593538	N/A	0.003328	0.019425
13	RF	None	0.735134	0.574390	0.591767	Weight	0.796414	0.008801
18	Catboost	None	0.636311	0.548644	0.589259	Shell Weight	1.617711	0.003759
16	XGB	None	0.653462	0.548565	0.586273	Shell Weight	0.175234	0.002505
14	GBM	None	0.675365	0.541064	0.586055	Shell Weight	2.197905	0.007675
17	LGB	None	0.700352	0.555301	0.581842	Shucked Weight	0.237092	0.024559
7	Ridge	MinMaxScaler()	0.532228	0.520139	0.580479	Weight	0.001748	0.000654
3	Ridge	StandardScaler()	0.532686	0.520057	0.580282	Weight	0.001770	0.000640
2	Lasso	StandardScaler()	0.532708	0.520123	0.580098	Weight	0.015851	0.001029
6	Lasso	MinMaxScaler()	0.531885	0.520191	0.580092	Weight	0.013725	0.001149
1	Linear Regression	StandardScaler()	0.532729	0.520110	0.580006	Weight	0.010554	0.001374
5	Linear Regression	MinMaxScaler()	0.532622	0.520193	0.579994	sex_M	0.004400	0.000681
4	kNN	MinMaxScaler()	1.000000	0.524147	0.564601	N/A	0.004981	0.010257
0	kNN	StandardScaler()	1.000000	0.516262	0.555271	N/A	0.004781	0.017710
12	DT	None	0.548012	0.469928	0.525224	Shell Weight	0.006809	0.002086
10	Lasso	Normalizer()	0.503200	0.493495	0.519184	Shell Weight	0.009674	0.001822
11	Ridge	Normalizer()	0.503167	0.493804	0.518667	Shell Weight	0.001702	0.000647
9	Linear Regression	Normalizer()	0.507745	0.472291	0.518372	Height	0.004518	0.000720
15	AdaBoost	None	0.479809	0.449434	0.512227	Shell Weight	1.296877	0.030407

AUTO ML

BASELINE RUN

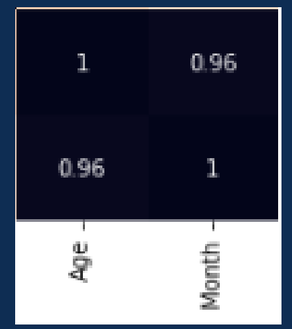
WE ARE ALSO STORING THE BEST HYPERPARAMETERS!

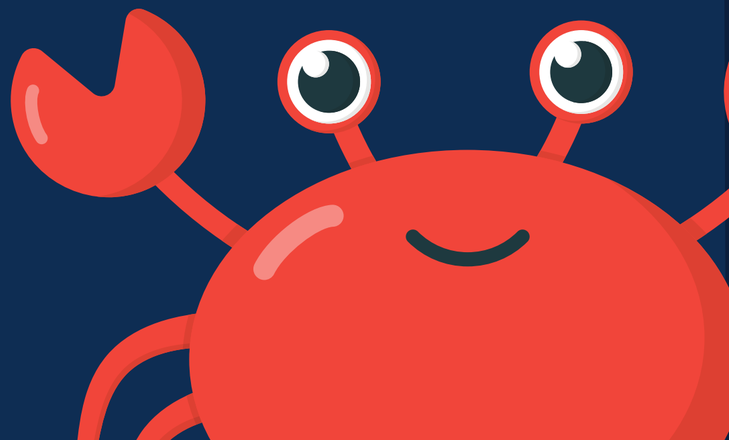
	Model	Scaler	Train Score	Validation Score	Test Score	Top Predictor	Fit Time	Score Time
17	LGB	None	0.894390	0.860462	0.855851	Shucked Weight	0.033300	0.002192
18	Catboost	None	0.881268	0.862864	0.855314	Month	2.051218	0.002027
16	XGB	None	0.902840	0.864460	0.855280	Month	0.249648	0.002535
14	GBM	None	0.888898	0.864358	0.854021	Month	0.339306	0.002184
0	kNN	StandardScaler()	1.000000	0.842419	0.850622	N/A	0.004322	0.013831
4	kNN	MinMaxScaler()	1.000000	0.841731	0.849932	N/A	0.006721	0.034945
12	DT	None	0.868149	0.854812	0.842661	Month	0.006951	0.001360
15	AdaBoost	None	0.844827	0.841067	0.833402	Month	0.402323	0.009096
5	Linear Regression	MinMaxScaler()	0.836982	0.834403	0.832155	Month	0.006242	0.000800
1	Linear Regression	StandardScaler()	0.837053	0.834151	0.831896	Month	0.004994	0.000739
6	Lasso	MinMaxScaler()	0.837049	0.834361	0.831843	Month	0.016040	0.001675
2	Lasso	StandardScaler()	0.837029	0.834348	0.831817	Month	0.017720	0.000912
3	Ridge	StandardScaler()	0.837030	0.834344	0.831746	Month	0.005398	0.001179
7	Ridge	MinMaxScaler()	0.837020	0.834383	0.831711	Month	0.001798	0.000661
13	RF	None	0.958073	0.814309	0.811611	Length	3.905877	0.038979
8	kNN	Normalizer()	1.000000	0.668959	0.692137	N/A	0.004130	0.014809
9	Linear Regression	Normalizer()	0.583518	0.515912	0.570186	sex_F	0.005591	0.000785
10	Lasso	Normalizer()	0.572135	0.559158	0.562251	Month	0.013111	0.000777
11	Ridge	Normalizer()	0.571816	0.557594	0.561887	Month	0.002043	0.001126

AUTO ML

AFTER FEATURE DERIVATION

BIN THE AGE BY 6 MONTHS
1-6 = 1
7-12 = 2
13-18 = 3



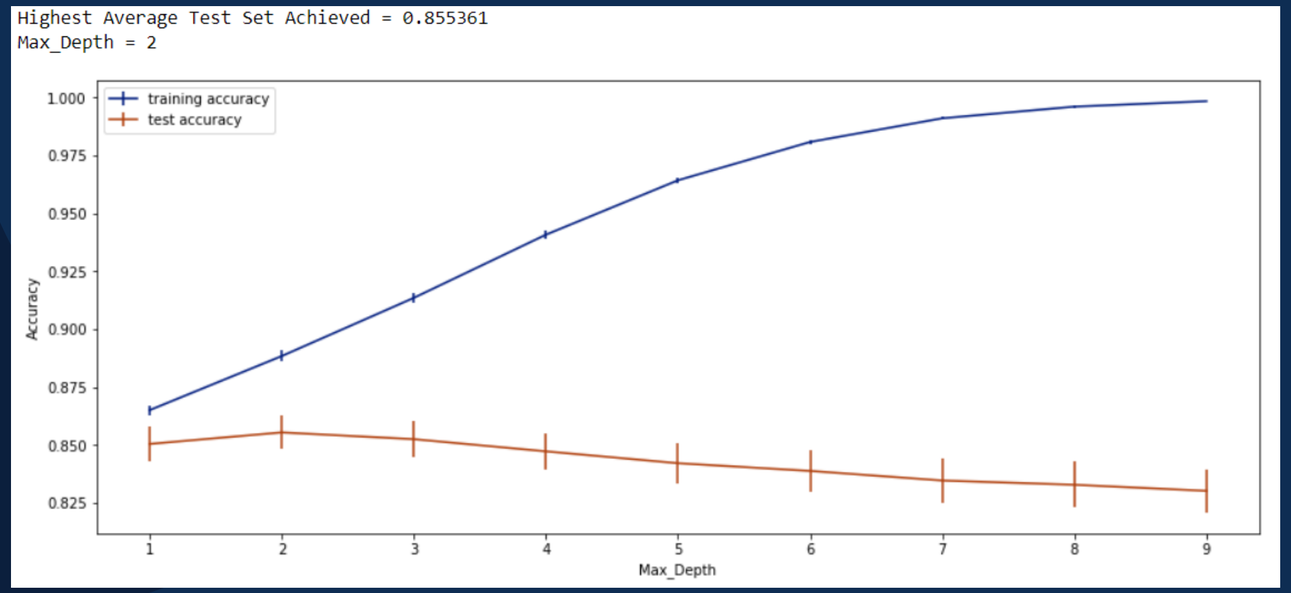


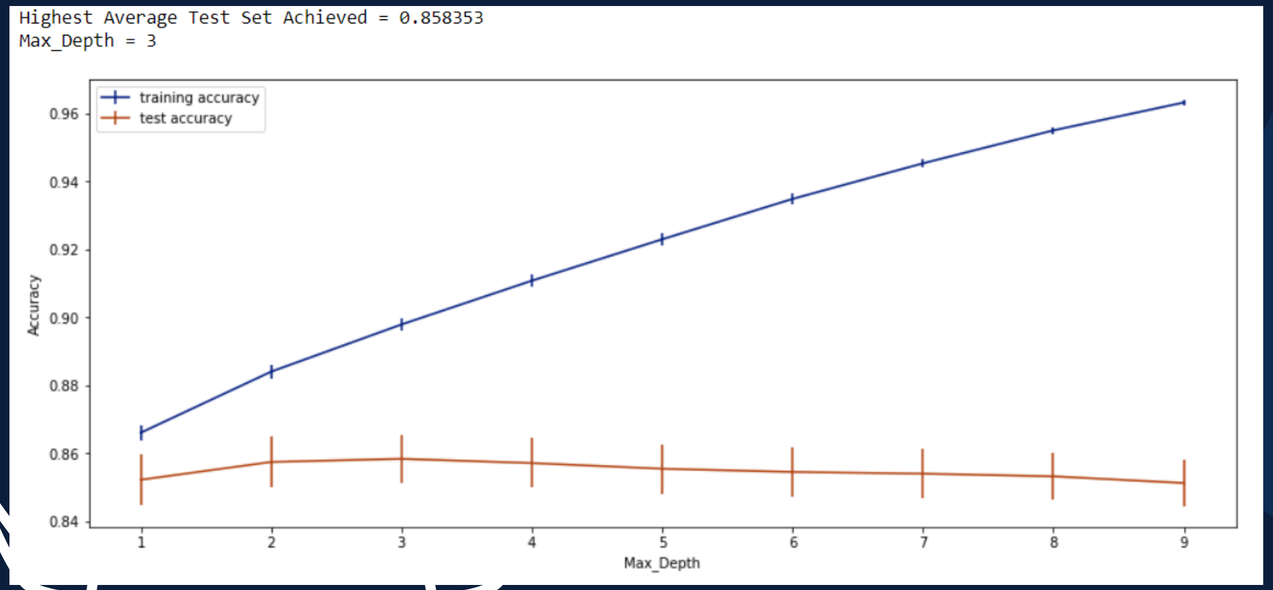
RESULT

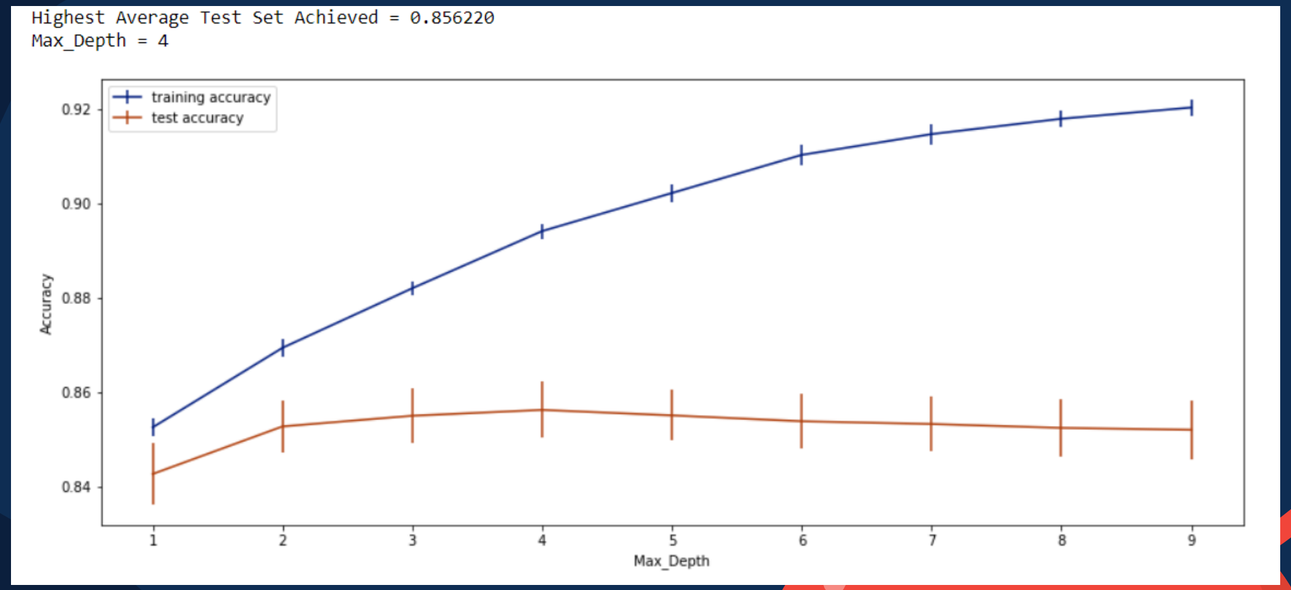
CATBOOSTREGRESSOR =0.858353
MAX DEPTH = 3

LGBMREGRESSOR =0.856220
MAX DEPTH = 4

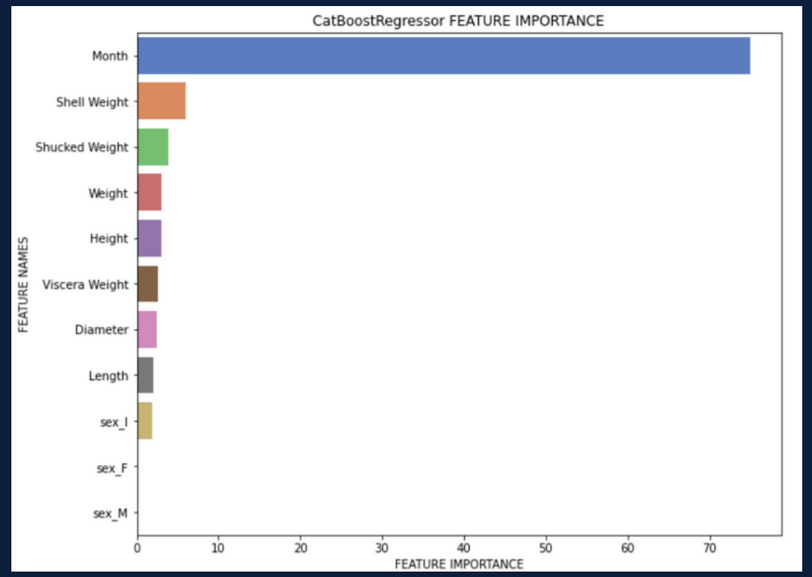
XGBREGRESSOR = 0.855361
MAX DEPTH = 2

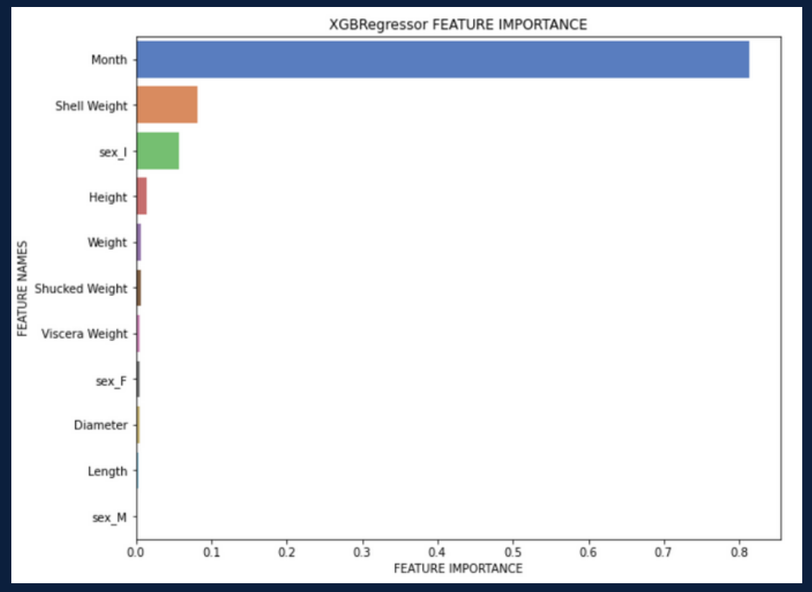


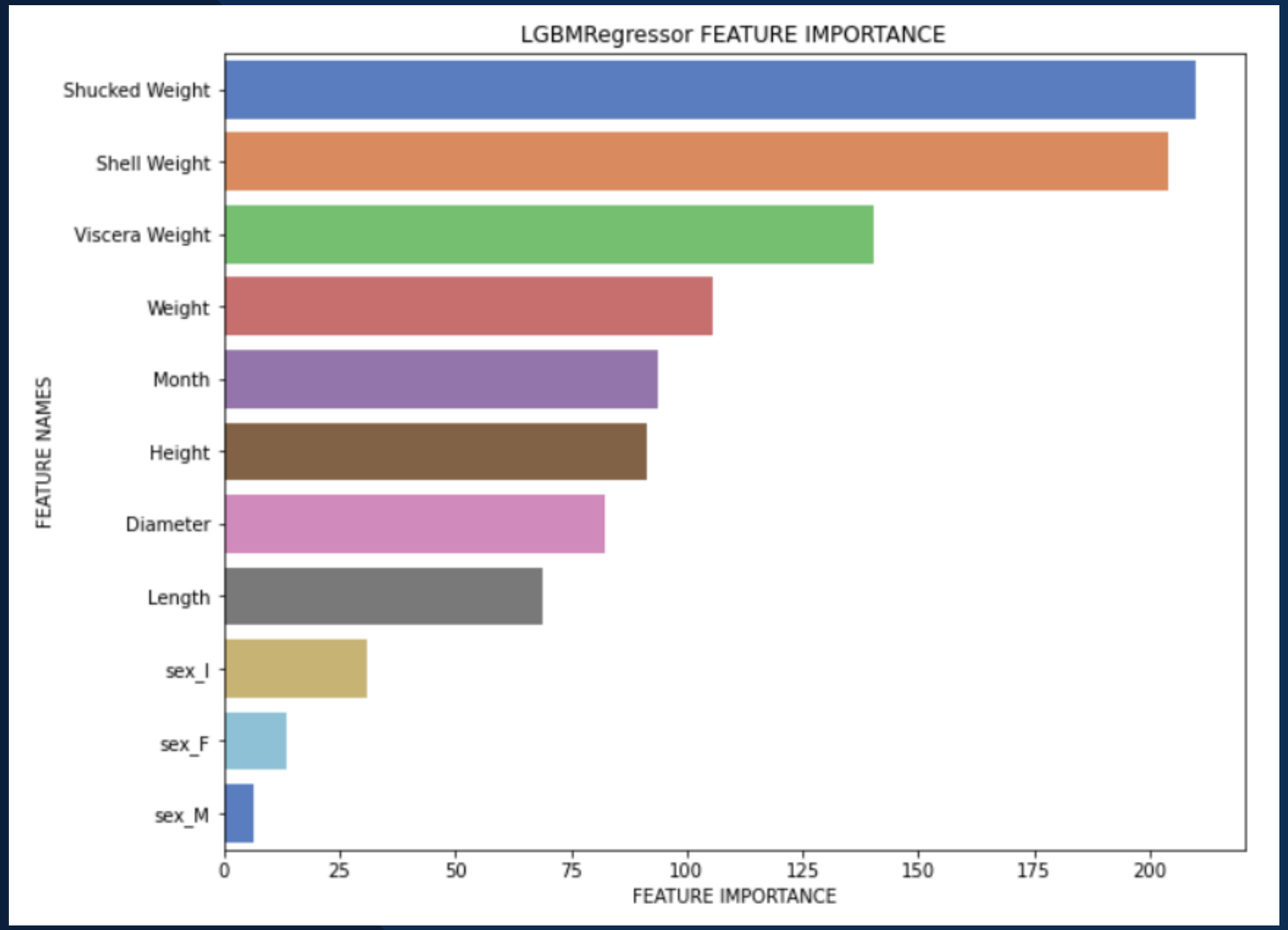




FEATURE IMPORTANCE







Mean Absolute Error

BEFORE TUNING

1.5811040944232 69 (MONTHS +-)

AFTER TUNING

0.9720435246678673 (MONTHS +-)

Learnings and Highlights

1. GRIDSEARCHCV + PIPELINE SAVES A LOT OF TIME AND NEEDS LESS SCRIPT.
2. DERIVING FEATURES HELPS IMPROVE ACCURACY AS LONG AS ITS PROPERLY DEFINED.
3. NOTE WEIGHT IS A DERIVED FEATURE WHERE ITS THE SUM OF THE 3 OTHER WEIGHTS HOWEVER SUMMING DOES NOT GIVE ANY EXTRA SIGNIFICANCE AND DOES NOT IMPROVE ACCURACY.
4. THERE ARE ALSO SOME GUIDELINES TO DROP FEATURES WHERE ITS DERIVED FROM. HOWEVER, DROPPING THE OTHER WEIGHTS, IN THIS CASE, IS NOT A GOOD IDEA SINCE THOSE WEIGHTS PROVIDE MORE DISTINCTION TO THE CRABS AGE.
5. FEATURES THAT HAVE COLINEARITY (MULTI IN THIS DATASET) CAN BE DROPPED. LENGTH, HEIGHT, AND DIAMETER ARE COLINEAR, AND DROPPING 1-2 OF THEM DOES NOT AFFECT THE ACCURACY.
6. FEATURES THAT DEFINES A PHYSICAL ATTRIBUTE TENDS TO PERFORM POORLY AT LATER AGES OF MOST LIVING THING DUE TO IT FLATTENS OUT AT A CERTAIN POINT. WE NEED ADDITIONAL FEATURES TO HELP PREDICT IT BETTER.
7. USE CASE IF A FARMER WANTS TO USE THE SAME POND AND PUT CRABS IN DIFFERENT PERIODS THIS CAN HELP PREDICT THEIR AGE.


Thank you!