Assignment 3

Alexander Dunbar
Dept of Mathematics
University of New South Wales
Sydney, Australia
alex.b.dunbar@gmail.com

***Abstract*—Neural Network implementation was investigated to determine the age of abalone from length, weight and sex features. Age was calculated directly (linearly) from the count of rings in the cross-section of the abalone shell. The ages were re-distributed into four class labels (0-7, 7-10, 10-15, 15+). A sequential model neural network was used with optimizer (SGD, Adam), learning rate, number of hidden layers and number of neurons per hidden layer investigated for best result. Best result was determined through analysis of the confusion matrix. The best confusion matrix parameters produced an accuracy of 72%.**

***Index Terms*—neural networks, machine learning, classifier**

I. Introduction

a) Motivation: The data-set is from a study undertaken in 1994 [1]. The original intent of the study is to investigate whether Blacktip abalone off the north coast of Tasmania (Bass Strait) were being over fished. Abalone from five separate sites were collected, measured and weighed (Table I). A crucial aspect of determining the health of abalone stock is calculating age which is not a straight forward process. One method is to count the number of rings existing in the cross-section of the shell. Adding to the complexity of age determination is "weather patterns and location (hence food availability)" [2]. This report will study the feasibility and accuracy of predicting the age of Backflip abalone using Neural Network machine learning using the features measured in the original report.

II. Methodology

Investigate the implementation of neural networks on the classification of abalone age. The neural network is a sequential model with investigation of stochastic gradient descent and Adam optimizers with hyperparameter tuning of learning rate, number of hidden layers and number of neurons per hidden layer. Evaluation of the best performing optimizer and hyperparameter is through the results of the confusion matrix and by extension precision and recall percentages.

a) Model Build: Initialised with random weights

A. Analyse and visualise

a) Train Test Validation Split: The data set was split into a training set, a validation set and a test set (Table II). Of the 4177 instances, 40% were set aside for final testing. The test set was not seen in the training or validation stage.

Name	Data Type	Meas.	Description
Sex	nominal		M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer		+1.5 gives the age in years

TABLE I
Features

	Instances	Features	Labels
Training	2004	7	4
Validation	501	7	4
Testing	1670	7	4

TABLE II
Train, validation, test split of the data set

b) Exploratory Data Analysis (EDA): There are nine features (attributes) in the raw data file including the labels which are integer counts of ring growths with an unequal distribution (see Age distribution in figure 14). EDA included histogram plots to get a feel for distributions of numerical features, scatter plots to get a feel for relationships between features and box plots to understand extent of any outliers. All continuous features had normal distributions with some element of skew. Interestingly, length measurements ('Height', 'Length' and 'Diameter') have right skew distributions (Figure 1), while weight measurements ('Whole weight', 'Shucked weight', 'Visceral weight' and 'Shell weight') have left skew distributions (Figure 2). 'Height' was the only feature that had significant outliers which were subsequently investigated (see Outliers).

c) Scaling: It became obvious that the weights and lengths measurements of the abalone in the raw data file didn't match the weights and measurements cited in the original report [1]. For example, shell diameter measured in millimetres is approximately between 40-150mm. The range in the dataset is 0.06-0.65mm. This re-scaling is found on all continuous measurements. The UCI data set website states that, "the ranges of the continuous values have been scaled for use with an ANN (by dividing by 200)" [2]. This factor of 200 was investigated and found that it is not consistent with the original data ranges and the ranges found on the data set.

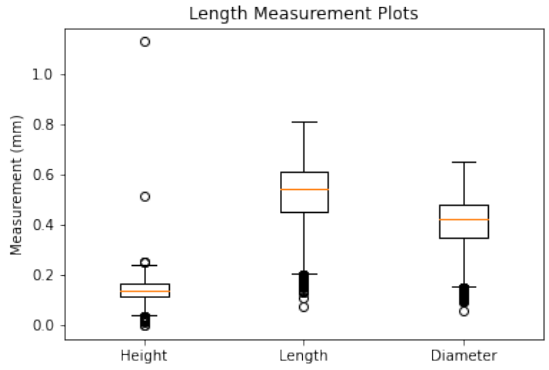

Fig. 1. Features having a length measurement. Apart from the large outliers in 'Height' (which were ultimately removed), the skew on length based features is to the left.

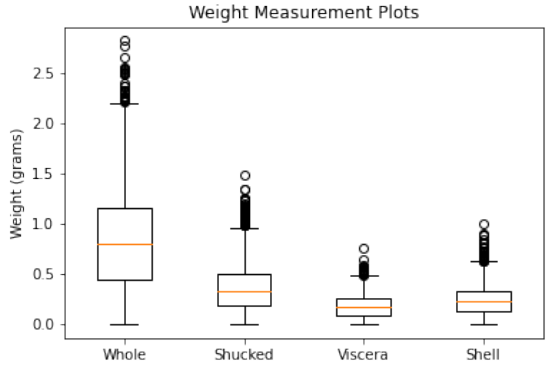

Fig. 2. Features having a weight measurement. The weight based features display a right skew.

After comparison of histograms from the original paper and histograms produced by this dataset, it was found that a factor of 200 isn't consistent across the features that have been scaled. Table III lists the approximate scale factors where it is possible to compare features directly. There is no record of the type of scaling that was done (normalistation, 0-1). The minimums are all close to 0 but the various maximums range between 0.65-2.83. Accuracy was compared with data scaled (normalisation) and as it is in its current scaling with no discernible improvement in either. Metrics presented here are in the data sets condition at download.

Measurement	Approx. Scale Factor
Length	200 (?)
Diameter	235
Height	200
Whole weight	210
Shucked weight	160
Viscera weight	200 (?)
Shell weight	200 (?)

TABLE III
APPROXIMATE SCALING FACTORS TO RETURN DATA SET VALUES TO ORIGINAL MEASUREMENTS AS REPORTED IN [1]

d) Correlation: There was an overall strong linearity among all the numerical features. A principle component analysis (PCA) of the data revealed that greater than 97% of all variation is encompassed in the first principle component. Very strong collinearity was seen between 'Length' and 'Diameter' (0.99), with diameter being measured perpendicular to length (longest shell measurement). Relationships between the different weights were very strongly correlated (0.96-0.97) and between the weights and length and diameter were 0.93. The poorest correlation was between the number of rings (age) and all other features where there was an average correlation of 0.547 (0.42-0.63) (Figure 3)

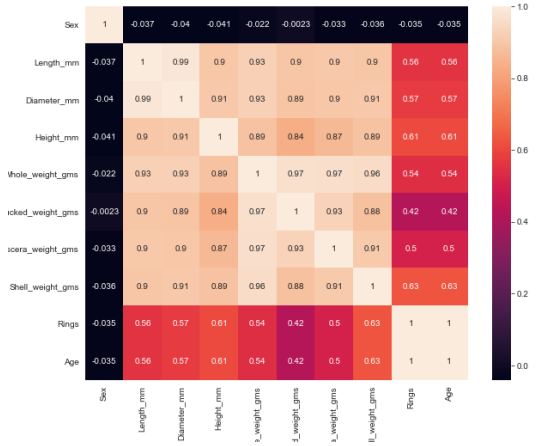

Fig. 3. Correlation coefficients. There is a strong positive correlation in all features apart from 'Rings' (i.e. Age).

e) Outliers: Box plots of all continuous numerical features displayed outliers. 'Height' had two significant outliers in the high end and the effect and processing of these was investigated (Figure 4). The two high value outliers in 'Height' would not appear to be outliers in other distributions (Figure 5, Figure 6). Three courses of action would be to 1. remove the rows completely or 2. Edit the height outliers to an average value or 3. Leave them in. For two instances in 4177 (0.05%), it is unlikely that it would make much difference. The effect of the outliers can be seen in the histogram and in scatter plots involving the other features. Potentially the largest effect would be on scaling the data. The two largest outliers on the high end were removed and the three on the low end were left in.

f) Reasons for poor age correlation: One of the reasons for having poor age-feature correlation is that the original study sampled abalone from five Bass Strait locations [1]. It can be seen in the scatterplot (Figure 9) from the original study that age vs length of one location seems to have a stronger correlation than age vs length from all locations (Figure 8). Figure 7 from the original paper clearly shows significant differences in 'Length' distributions across the five sites. Mean, median and standard deviation varying significantly across the sites. Table 4 from the original paper reports r-

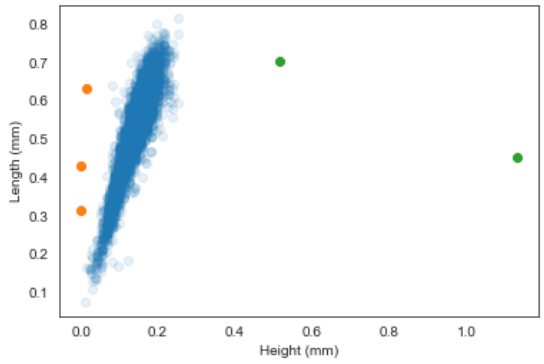

Fig. 4. Outliers in Height. Upper outliers coloured green, lower outliers coloured orange

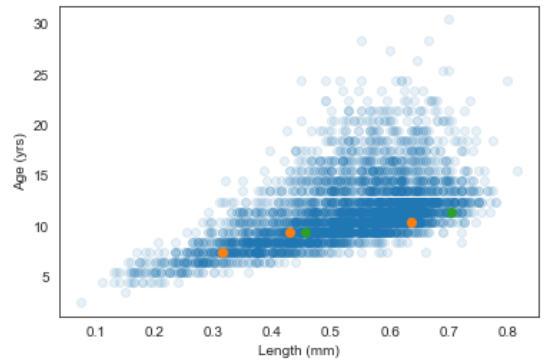

Fig. 5. Length vs Age. The rows where 'Height' contains outliers coloured orange and green

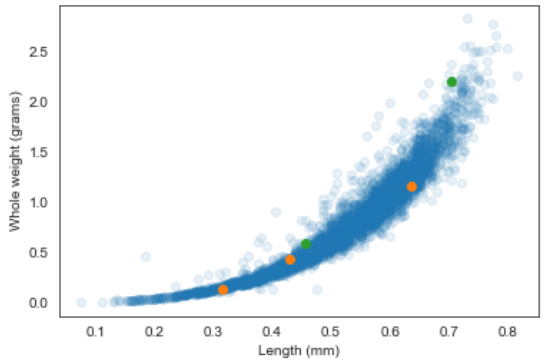

Fig. 6. Length vs Weight. The rows where 'Height' contains outliers coloured orange and green

squared values for Age-Length to be 0.598-0.770 (location) and even 0.638-0.811 (male+imm+location) and 0.641-0.824 (female+imm+location). See appendix 17 for table of original correlations. If the location data was available then I believe the age correlation to other features would reflect the original study.

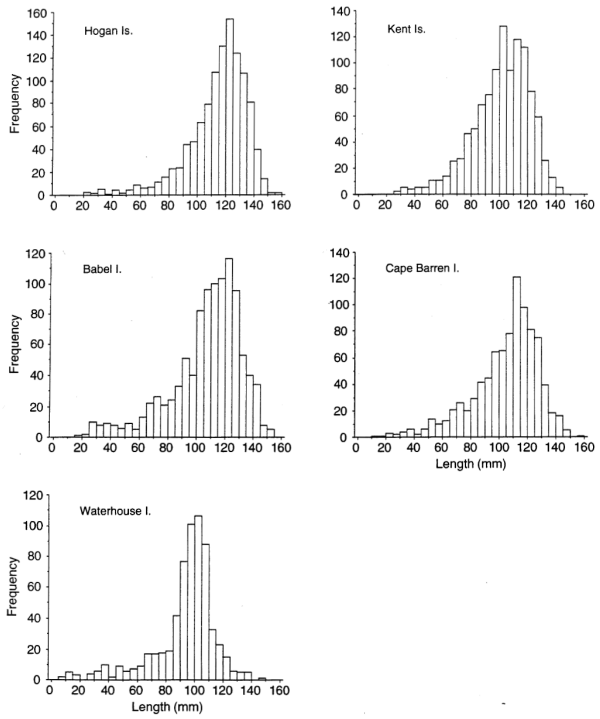

Fig. 7. Figure 18 from original paper [1]. Length histograms from the five location sites of the original paper.

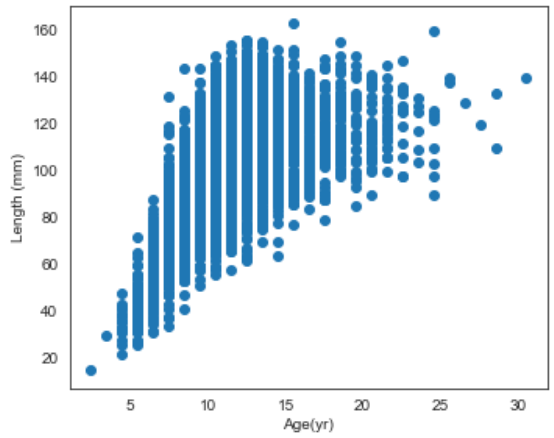

Fig. 8. Age vs Length scatterplot for all instances. Compare with 8 where Age vs Length is location specific. There is a clear growth (length) variability from site to site.

III. Investigations

Impact of optimizer, learning rate, hidden layers and number of neurons in hidden layers were investigated. Table IV lists the parameter of investigation and its corresponding values. Figure 10, 11, 12 shows the calculated values of true positive (tp), true negative (tn), false positive (fp) and false negative

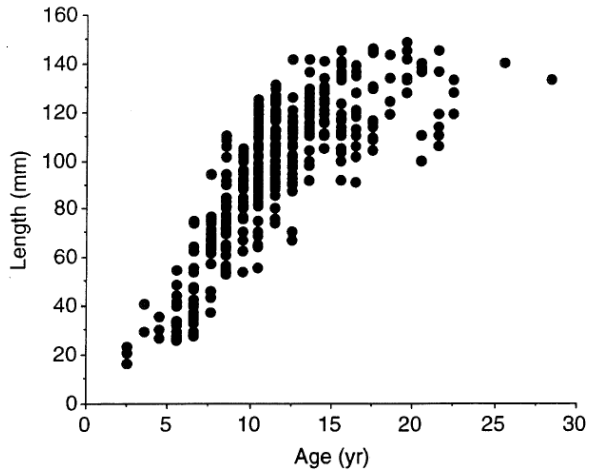

Fig. 9. Figure 16 from original paper [1]. "Relationship between age and shell length of Haliotis rubra from Babel Island, illustrating the variation in growth rate that exists."

(fn) after each run through of the model. In order of importance, optimizer, learning rate, hidden layers, number of neurons per hidden layer had the greatest impact on accuracy.

optimizer	Stochastic gradient descent (SGD), Adam	Adam
learning rate	8e-4, 1e-3, 2e-3–8e-3, 1e-2, 2e-2–6e-2	2e-2
hidden layers	1, 2, 3, 4, 5, 6, 7, 8	2
neurons	8, 16, 24, 32, 40, 48, 54, 60	40

TABLE IV
PARAMETERS INVESTIGATED FOR TUNING

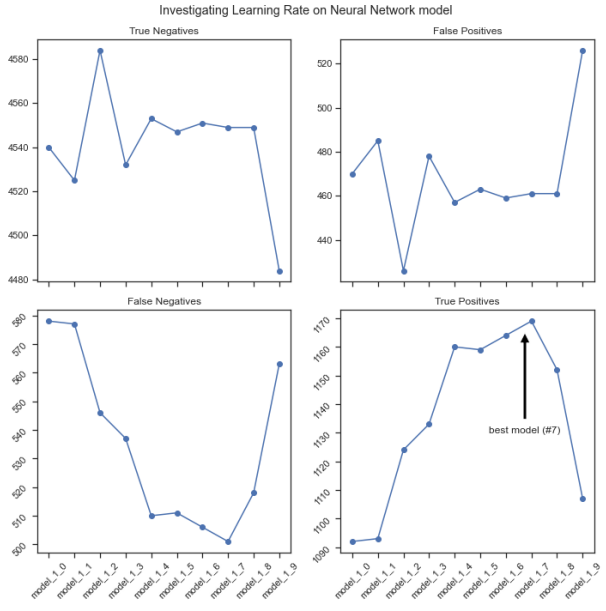

Fig. 10. Learning Rate parameter tuning. The best model learning rate is 0.02, Model 7 (counting from 0)

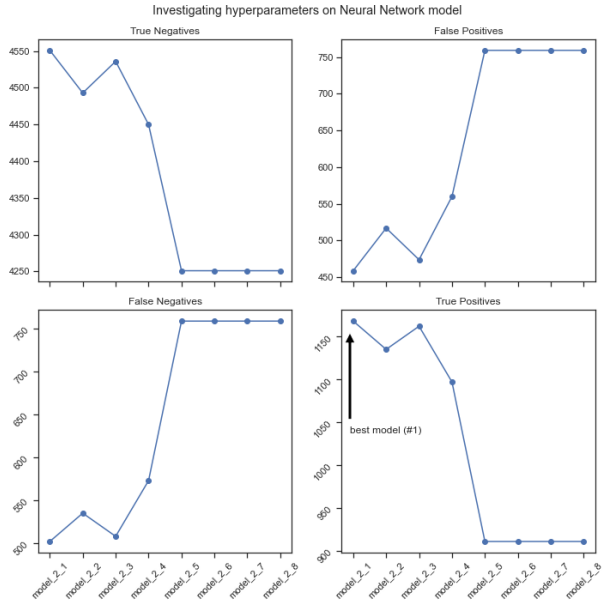

Fig. 11. Hidden Layers parameter tuning. The number of hidden layers leading to the best model is 1. It was found that tuning the number of neurons produced a slightly better model with 2 hidden layers.

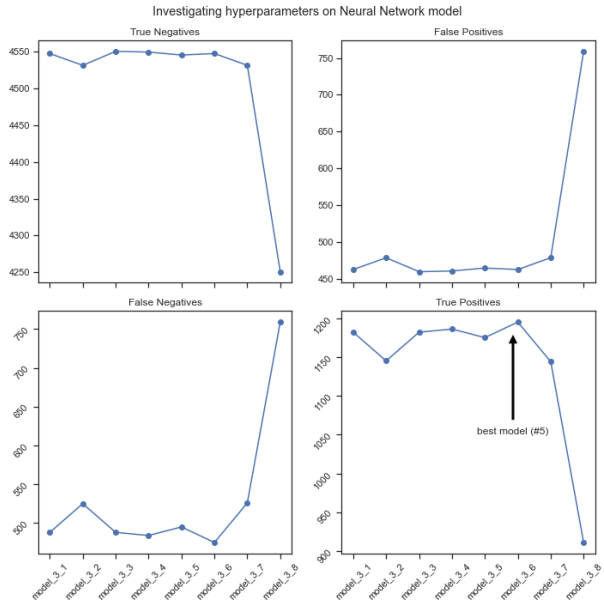

Fig. 12. Number of neurons. The best model was with 40 neurons per hidden layer (2).

a) Evaluation of the best model: Determination of the best model was evaluated using the confusion matrix. The specific model that became the best model is shown in the evolution of parameter tuning plots in figures 10, 11, 12. In this particular run of models, the best model became Model-3-5 (Learning rate=0.02, hidden layers=2, number of neurons=40). In this plot the best model displayed the highest percentage of true positives and true negatives and the lowest percentage

of false positives and false negatives. The confusion matrix for the best model is presented in figure 13. The final values of precision and recall can be seen in the classification table (Table V). There is significant bleed of predicted results into the neighbouring classification classes. This is due to the overlapping age distribution (Figure 14) and reflects the effects of missing location data which would improve the prediction accuracy.

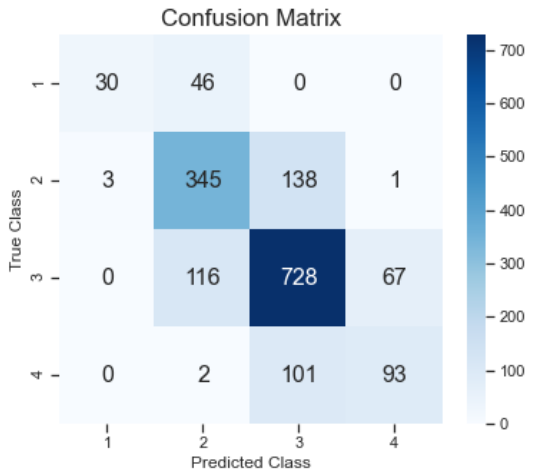

Fig. 13. Confusion Matrix

	precision	recall	f1-score	support
Class 1	0.91	0.39	0.55	76
Class 2	0.68	0.71	0.69	487
Class 3	0.75	0.80	0.78	911
Class 4	0.58	0.47	0.52	196
accuracy			0.72	1670
macro avg	0.73	0.59	0.63	1670
weighted avg	0.72	0.72	0.71	1670

TABLE V
CLASSIFICATION REPORT

b) Ten Model Average: The final best model was run tens times with randomised weights initialised each time. The average, standard deviation and 95% confidence interval for tp, fp, tn, fn, accuracy and auc are presented in Table VI.

	tp	fp	tn	fn	accuracy	auc
mean	1450	530	5482	554	0.729	0.929
std	11.65	10.01	10.01	11.65	0.005	0.002
95% ci (low)	1442	523	5475	546	0.726	0.928
95% ci (upper)	1458	537	5489	562	0.733	0.930

TABLE VI
TEN MODEL STATISTICS

IV. CONCLUSION

A number of parameters were investigated in order to find the best model based on confusion matrix results. Precision and recall rates varied across the classifications with class 2

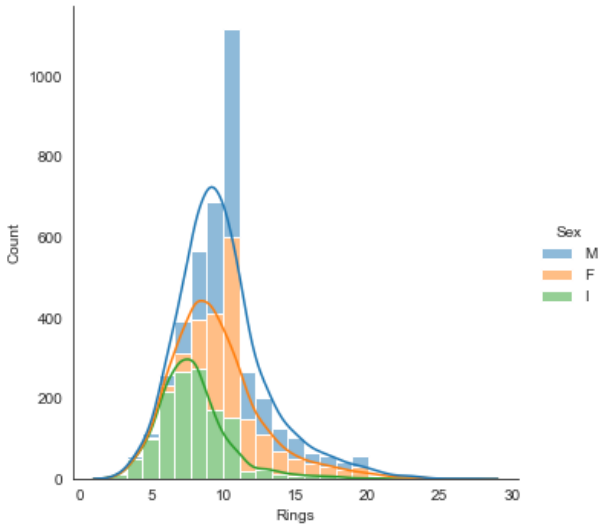

Fig. 14. Distribution of 'Rings' (i.e. age) categorised by 'Sex'.

and class 3 achieving good results (0.68, 0.71 & 0.75, 0.80 respectively). Class 1 achieved very good precision (0.91) but poor recall (0.39). Class 4 achieved the worst precision and recall (0.58, 0.47). Overall accuracy was 0.72 and an average area under curve (auc) value of 0.929 was achieved. Accuracy and therefore confusion matrix results would have improved if the location data for each instance was available.

REFERENCES

[1] Warwick J. Nash, Tracy L. Sellers, Simon R. Talbot, Andrew J. Cawthorn and Wes B. Ford, "The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and the Islands of Bass Strait.", 1994

[2] https://archive.ics.uci.edu/ml/datasets/abalone

APPENDIX

	loss	tp	fp	tn	fn	accuracy	precision	recall	auc	prcurve
0	0.626472	1424.0	543.0	5469.0	580.0	0.721058	0.723945	0.710579	0.925533	0.789643
1	0.615300	1430.0	538.0	5474.0	574.0	0.722555	0.726626	0.713573	0.927944	0.797323
2	0.611691	1442.0	535.0	5477.0	562.0	0.726547	0.729388	0.719561	0.928550	0.798599
3	0.609457	1443.0	540.0	5472.0	561.0	0.728044	0.727685	0.720060	0.928944	0.799698
4	0.607951	1447.0	534.0	5478.0	557.0	0.727046	0.730439	0.722056	0.929222	0.800713
5	0.608463	1445.0	530.0	5482.0	559.0	0.729042	0.731646	0.721058	0.929265	0.801019
6	0.605598	1438.0	540.0	5472.0	566.0	0.725050	0.726997	0.717565	0.929764	0.802168
7	0.604424	1439.0	532.0	5480.0	565.0	0.727046	0.730086	0.718064	0.930009	0.802999
8	0.606376	1453.0	528.0	5484.0	551.0	0.729541	0.733468	0.725050	0.929882	0.803265
9	0.605190	1446.0	529.0	5483.0	558.0	0.729541	0.732152	0.721557	0.929942	0.803070

Fig. 15. Ten models

	loss	tp	fp	tn	fn	accuracy	precision	recall	auc	prcurve
count	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000
mean	0.610092	1440.700000	534.900000	5477.100000	563.300000	0.726547	0.729243	0.718912	0.928906	0.799850
std	0.006631	8.485936	5.195083	5.195083	8.485936	0.002891	0.002918	0.004234	0.001360	0.004104
min	0.604424	1424.000000	528.000000	5469.000000	551.000000	0.721058	0.723945	0.710579	0.925533	0.789643
25%	0.605793	1438.250000	530.500000	5472.500000	558.250000	0.725424	0.727169	0.717690	0.928649	0.798874
50%	0.608207	1442.500000	534.500000	5477.500000	561.500000	0.727046	0.729737	0.719810	0.929244	0.800866
75%	0.611133	1445.750000	539.500000	5481.500000	565.750000	0.728792	0.731344	0.721432	0.929852	0.802791
max	0.626472	1453.000000	543.000000	5484.000000	580.000000	0.729541	0.733468	0.725050	0.930009	0.803265

Fig. 16. Ten models statistics

	Gompertz					von Bertalanffy				
Area	l_0	G	g	L_∞	r^2	L_∞	K	t_0	r^2	n
Hogan Is.										
all	0.459	5.637	0.273	128.8	0.695	135.7	0.168	3.508	0.686	413
fem+imm	0.487	5.589	0.266	130.3	0.750	138.9	0.156	3.446	0.737	244
male+imm	1.159	4.711	0.246	128.8	0.740	138.6	0.144	3.179	0.735	234
Kent Is.										
all	5.356	3.111	0.207	120.2	0.598	127.0	0.135	2.388	0.596	356
fem+imm	4.461	3.278	0.211	118.3	0.641	126.4	0.132	2.423	0.636	234
male+imm	11.500	2.421	0.154	129.5	0.638	143.4	0.089	1.276	0.639	217
Babel I.										
all	0.542	5.483	0.275	130.4	0.770	146.1	0.136	2.863	0.751	414
fem+imm	0.333	5.982	0.280	131.9	0.803	153.0	0.124	2.832	0.776	245
male+imm	1.345	4.594	0.241	133.0	0.811	155.4	0.110	2.572	0.795	256
Cape Barren I.										
all	2.743	3.852	0.222	129.2	0.767	139.4	0.129	2.634	0.761	344
fem+imm	3.048	3.748	0.216	129.3	0.787	142.4	0.119	2.480	0.781	210
male+imm	4.548	3.374	0.193	132.8	0.797	148.1	0.104	2.258	0.792	218
Waterhouse I.										
all	0.582	5.265	0.281	112.6	0.736	121.6	0.151	2.545	0.714	303
fem+imm	0.623	5.179	0.279	110.6	0.824	121.6	0.139	2.392	0.798	143
male+imm	1.181	4.586	0.254	115.9	0.783	128.5	0.129	2.369	0.766	207

Fig. 17. From original paper: "Gompertz and von Bertalanffy growth parameters of Haliotis rubra at the five Bass Strait survey sites, assuming one major shell growth ring per year. Samples from sites within each area were combined for these analyses. Analyses were conducted on the separate sexes (fem+imm, male+imm) and the sexes combined.