A RESEARCH ON CLASSIFICATION MODEL OF FISH BIOLOGICAL CHARACTERISTICS

An Example Based on Chinook Salmon and Native Trout in The Great Lakes

Kaiye Yang
Hzhx55@durham.ac.uk

Contents

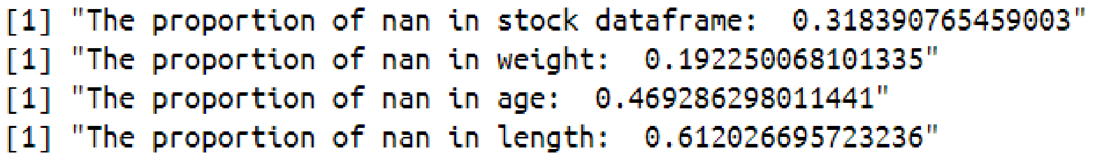
Research Background

Located in the north-eastern district of America, the Great Lakes Area is the perpetual inhabitancy of both native species and invasive creatures (Alexander & Chronicle, 2019). Despite the geographic structure of Great Lakes was formed through a 3000-year-ago crustal movement, its ecological environment is completely manipulated by human intervention, particularly in the species introduction policies throughout the time. To solve the catastrophic consequences of Atlantic salmon's extinction in the lake district (Caroline, 2021), the US government has introduced Chinook salmon as substitute, which is a typical strong and hard-running fish that satisfies both sport and entertainment demands of tourists, contributing economic growth in lake regions. Due to lacking predators, Chinook salmon once became as the most gigantic species in the Great Lakes, with average 15 kilograms in a 25-centimetre body (FAC, 2021). However, the dominance of Chinook salmon was temporary since the new invasion occurred at the upstream of its food chain---human-imported mussels have filtered the nutrition inside the lake water, which promptly shrinks the population of Chinook salmon's diet species because the nutritious plankton are their exclusive food source (Matheny, 2012). Distinct and inadequate food sources have led to a rapid decline in Chinook salmon populations in recent years, so it is rational for stocking programmes on ecological perspective. For instance, how to teach amateur fishmen to identify Chinook salmon without a visualised reference book when comparing to the other native commercial fishes (e.g., Brook Trout, Brown Trout, and Lake Trout), especially they have similar biological characteristics (Micigan.gov, 2021).

Data Source Introduction

Therefore, a Great Lakes Fish Stocking Database (2021), recording about information of both stocked and tradable fish in the reserve from 1950 to 2017, was introduced to classify Chinook salmon out from the trout groups through the computational automation of its biometrical data. To assess the stock management in fishery development, scientists customarily collect some biometrical features from a targeted species for long-term tracking about their life history and habitats, such as weight, age, and length, which is ultimately beneficial to confirm a certain fish type (Famoofo and Abdul, 2020; Kasapoglu and Duzgunes, 2014). Thus, several data columns, 'SPECIES', 'AGEMONTH', 'WEIGHT', and 'LENGTH' were selected from the database, while 'SPECIES' included two classical parameters, truant and salmon (abbreviation as TRT and CHS respectively), while 'AGEMONTH' was measured by the number of month that stocked fish lived around the lakes, and 'WEIGHT' and 'LENGTH' were the average weight and length of the research fish captured per month, measured by kilograms and centimetres separately (ibid). Four variables build an original dataset containing about 29,368 observations.

Because the whole digital information was achieved from an exclusive database which operated in a time span of over half century, data incompleteness and information missing are inevitable due to staff's input errors or special events accumulated for years. Though missing data is customary in the exploratory sections of quantitative research than being considered as mistakes or exceptions (Dong and Joanne, 2013). According to Enders's survey on educational studies (2010), average 15-20% of data missing was acceptable to demonstrate the relationships among variables. While Peng et al. (2006) stated the psychological studies have more tolerance since the higher investigation cost margin: 48% published papers were found of missing data, and 16% of them could not be determined since unknown data sources. Although missing information would not make the dataset unfunctional, it is significant to illustrate the proportion of missing data in each targeted variable to determine whether the further optimisation is needed. The percentages of missing data in three variables vary from 19% to 61% (Fig.1), which overwhelms the 10% limit for missing consequential data in the general scientific research (Schafer, 1999) which means that it would probably affect the quality of statistical inferences. Fig. 2-5 verified that this was a vital concern towards the data structure---data in each variable was exposed by having a long right tail, which should be deemed as the 50%-of-data-away outliers mainly prompted by record failure, that gives the dataset of variable in right-skewed distributions.



```
[1] "The proportion of nan in stock dataframe:  0.318390765459003"
[1] "The proportion of nan in weight:  0.192250068101335"
[1] "The proportion of nan in age:  0.469286298011441"
[1] "The proportion of nan in length:  0.612026695723236"
```

Fig.1 The Proportion of Missing Data in The Stocking Fish Dataset

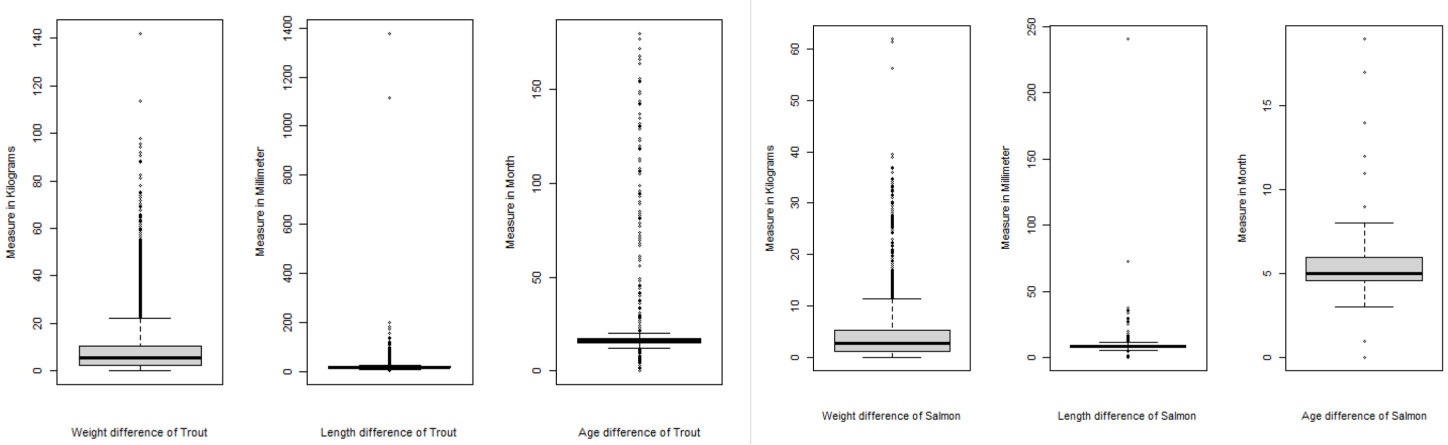


Fig.2&Fig.3 The Boxplot of Data Structure of Trout and Salmon Datasets

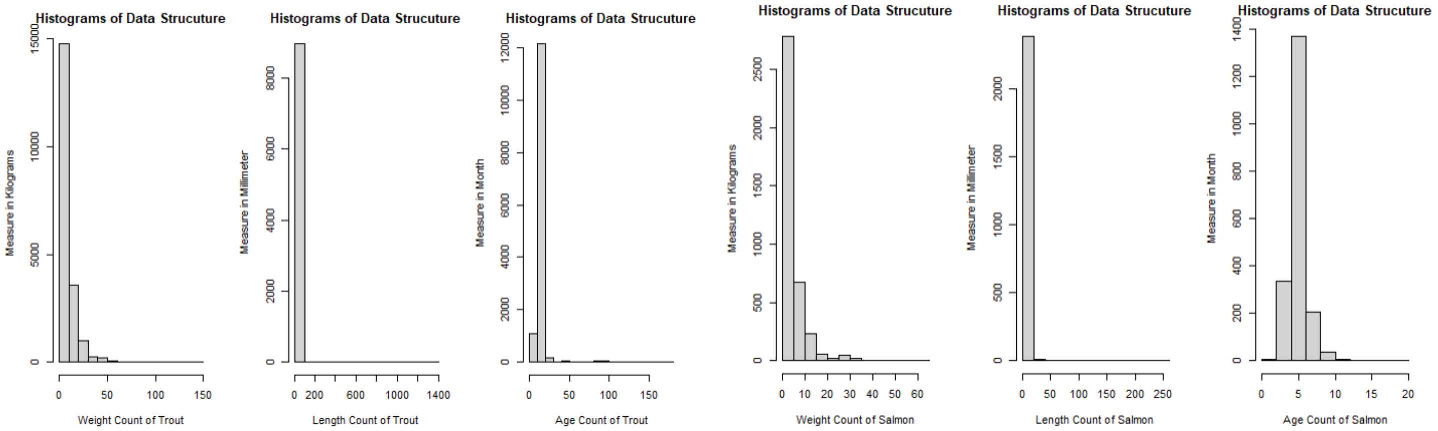
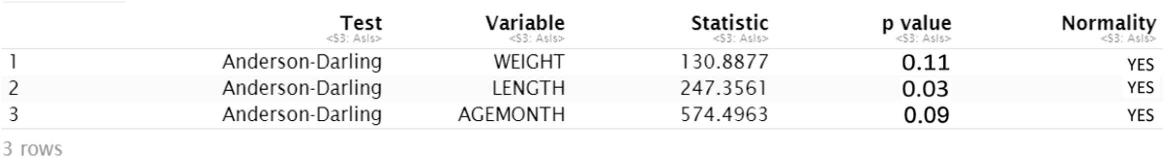


Fig.4&Fig.5 The Boxplot of Data Structure of Trout and Salmon Datasets

Scientists have developed several approaches on solving the problem of data missing (Kang, 2013): Deletion, Imputation of Mean or Median, and Linear Imputation. Deletion is the most easy-to-understand method that completely omits the Not-a-Number information by adopting the residual data instead. However, the assumptions of complete case deletion should be based on a scenario of Missing Completely at Random (MCAR), meaning that the missing data ratios are irrelevant to the values or sets of observations (Mack et al., 2018). This technique is inappropriate to apply on the fish dataset since the proportions of missing data of variables are uneven, that might be triggered by case research or biased recording habits; While the single imputation replaces the missing data through implementing certain rules, generally in median or mean (Dziura et al., 2013). In this case, the long-tail information will overestimate the total mean if it fills those invalid values, but the median will enhance the weight of meddle value, which is effective to amend the skewness of distribution. Nevertheless, average over 20% of values are problematic makes this manual method unreliable, because these created values were not representatives of a real fish, which might severely affect the performance of classifiers when their numbers are excessive; Linear imputation is assumed on the data following on time-series patterns or seasonality (Swalin, 2018). However, a fish's biological characteristics are independent of other fish in terms of time or season, compared with genetic information shared within its species. But the genetic information was not in the range of measurements, so this method was also rejected.

Although all the techniques were unsuitable to the fish dataset, the complete-case deletion method was eventually preferred in consideration of remaining the true information of observations, even it triggered a large sampling bias which would be discussed in limitation paragraphs.

After deleting the Not-a-Number information, Multivariate Normality Test (MNT) validated whether the data packages of multiple variables follow Central limit theorem by setting a null hypothesis that they are similar to the normal distribution (Zhou & Shao, 2014). Assumed the significant level α=0.01, the p-values of three variables were accepted by the null hypothesis, while the length variable seemed to be skewed than the other two dimensions (Fig.6).



	Test <S3: AsIs>	Variable <S3: AsIs>	Statistic <S3: AsIs>	p value <S3: AsIs>	Normality <S3: AsIs>
1	Anderson-Darling	WEIGHT	130.8877	0.11	YES
2	Anderson-Darling	LENGTH	247.3561	0.03	YES
3	Anderson-Darling	AGEMONTH	574.4963	0.09	YES

3 rows

Fig.6 The Multivariate Normality Test of Measured Variables

Modelling Processes

Before selecting a suitable model, it is crucial to divide the processed dataset into training set and test set. A standard modelling mechanism is to train the expected algorithms only through a large-scale training set, while then measuring the model's performance after putting the test data into the designed algorithm through comparing the calculated results and the actual values (Gutierrez, 2013). Test sets should be used independently, otherwise the information they carry would contaminate the entire model, leading to poor generalisation outside the experimental environment. Thus, sample.split function was adapted to split the fish dataset in a predefined ratio (0.75) while preserving the relative ratios of four labels in two new subsets.

Logistic Regression

Because its prediction is either salmon or trout, the classifier should distinguish the results binarily after the internal calculation of variables' information. Several statistical means were recommended for their specific-designed classifications on binary problems: Logistic Regression (LR), Linear discriminant analysis (LDA), Quadratic Discriminant Analysis (QDA). LR is a special type of linear regression, that it has some commons on the linear relationships between exploratory variables and response variable, only differing in the expectation construction. Possibility dominates the formula of LR $p(X) = \frac{exp(\beta 0+\beta 1X)}{1+exp(\beta 0+\beta 1X)}$, rather than linear expectation used in linear regression $E(X) = exp(\beta 0+\beta 1X)$. In order to reflect the concept of linearity in probability formulas, LR applies a complex logit transformation that solids the predicted values between 0 and 1 within a Sigmoid curve, which could not be managed by linear regression as it will transcend the boundaries (Sperandei, 2014).

Since it is abstract about the formula construction on mentioning the relationships about species of fish and its biological characteristics, thus both addition and multiplication would be employed to connect 'SPECIES' and the other three numeric vectors, which is because that both functions remain the value of each independent variable and the relationship between variables. Two logistic regression model were built through the correlation of addition and multiplication (Appendix.1 & Appendix.2):

ADDITION:

glm(as.numeric(New_stock$SPECIES=="CHS")~New_stock$WEIGHT+New_stock$LENGTH+New_stock$AGEMONTH, data = train, family = "binomial")

MULTIPLICATION:

*glm(as.numeric(New_stock$SPECIES=="CHS")~New_stock$WEIGHT*New_stock$LENGTH*New_stock$AGEMONTH, data = train, family = "binomial",maxit = 100)*

The value of Akaike Information Criterion (AIC) (how well the model reproduces the data) of the Multiplication Method exceeds 6000, which was almost eleven times as the AIC value generated by the Addition Method, which implies that Multiplication Method has created over-complicated relationships among variables, the purpose of Multiplication Method was rejected.

Although a model was successfully developed, all participating data needs to be validated on whether they have met the minimum standard rules of applying a certain type of classifier. LR operates in the following five hypotheses:

1. Binary Outcome Type and Numeric Variables
2. No Multicollinearity among explanatory variables
3. Independence of observations from a large-scale data pool
4. Linearity of independent variables and log-odds
5. Absence of strongly influential outliers

Because the 'salmon or trout' question was typically binary and each exploratory data were numeric (Appendix.3), the hypothesis 1 was met; Variance inflation factor (VIF) could measure the intensity of multicollinearity through a set of multiple regression variables, which depicts the internal correlation among three variables are similar at around 1.5, also being visualised into the horizontal or vertical scatters (Fig.7), so hypothesis 2 was passed; Because the bioinformation of each fish was recorded independently, the observations could not be duplicates. Through numerous missing data was deleted and training sets was smaller than the original dataset, it still contained 3,397 observations divided equally into four categories, which is distant away from the red line suggested by Leung (2021)---at least 10 observations from each independent variable and 500 of observations in total; However, hypothesis 4 was refused when Box-Tidwell test measured the linearity between the predictors and the logit: 'It is done by adding log-transformed interaction terms between the continuous independent variables and their corresponding natural log into the model'(ibid). The probability expectation of the model made by Addition Method was negative after the logit transformation, which could not pass box-Tidwell test. Therefore, the hypothesis failed, and the constructed model made by LR was invalid.

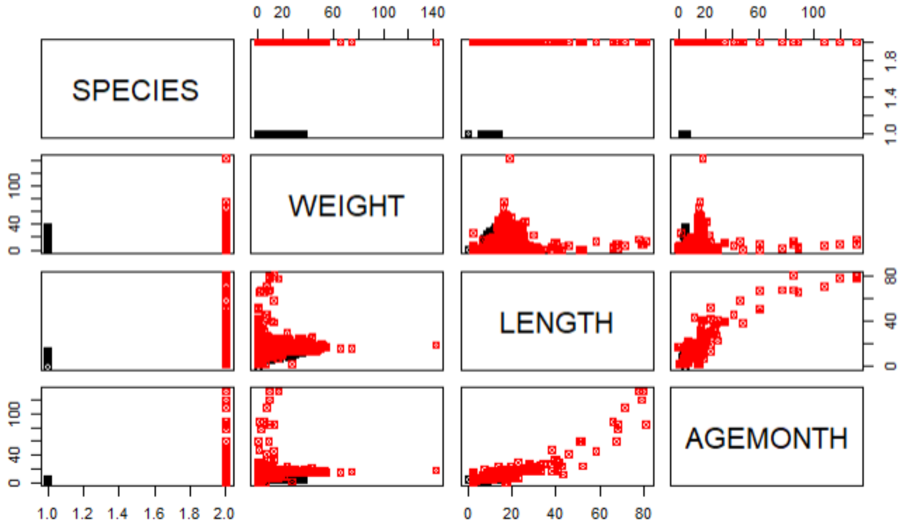


Fig.7 The Relationship Among the Measured Variables

LDA and QDA

Contrasted with the linearity insisted by LR, both LDA and QDA algorithms are established on the native Bayes theorem which believes the possibility of A is influenced by the event B when B occurs, formula as $\Pr(Y=k \mid X=x) = \frac{\Pr(X=x|Y=k) * \Pr(Y=k)}{\sum_{p=1}^{p=k} \Pr(X=x|Y=p) * \Pr(Y=p)}$ (Shekhar, 2018). In normal, discriminant analysis method is more robust to the dataset whose classes of variables are well-divided than LR, especially for small sample size, just like the salmon samples, which were a few black points illustrated in Fig.7.

It is noticeable that both LDA&QDA have more restrictive assumptions on data structure than logistic method:

1. Both LDA and QDA assume the exploratory variables are drawn from a multivariate normal distribution.
2. Both LDA and QDA need the number of exploratory variables should be less than the sample size.
3. LDA requires equality of covariances among the explanatory variables towards all response classes, while QDA has no such restriction.

Assumption 1 was qualified due to an early validation about the normal distribution of variables; The sample sizes of train and test datasets were 3,397 and 1,132, both were far more than three; Nevertheless, LDA was rejected since the variances of the three variables are significantly different, as the variances of the Weight variable were close to the sum of the variances of the other two dimensions (Fig.8). In conclude, only QDA could be applied on modelling construction.

Addition and multiplication methods brought two different models with corresponding predictions when comparing the computed expectation and the actual values in terms of confusion matrixes (Fig.9 & Fig.10):

Addition: qda (train$SPECIES ~ train$WEIGHT+train$LENGTH+train$AGEMONTH, data=train)

*Multiplication: qda(train$SPECIES ~ train$WEIGHT*train$LENGTH*train$AGEMONTH, data=train)*

Detailed information about these two models is in the Appendix.4 & Appendix.5.

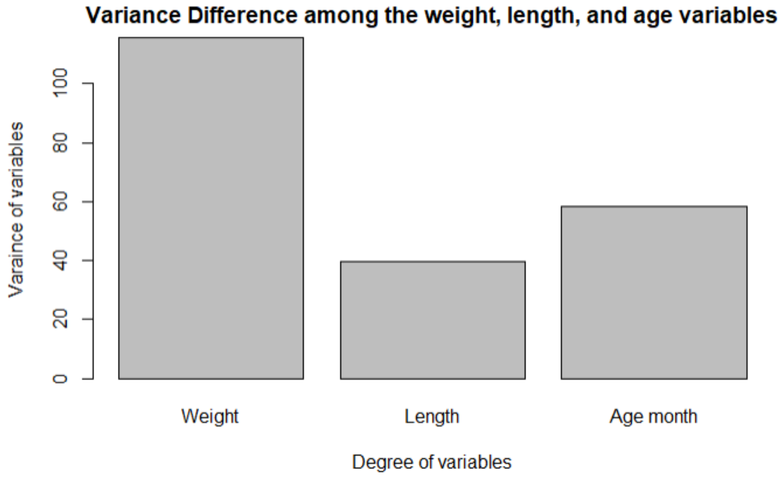
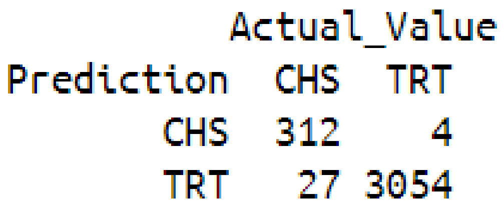
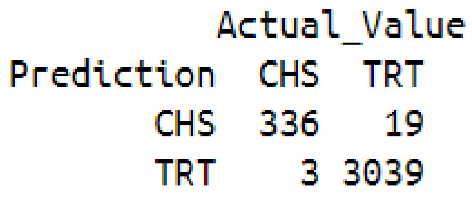


Fig.8 Variance Difference Among the Three Measured Variables



```
           Actual_Value
Prediction  CHS  TRT
       CHS  312    4
       TRT   27 3054
```



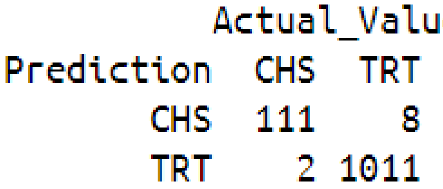
```
           Actual_Value
Prediction  CHS  TRT
       CHS  336   19
       TRT    3 3039
```

Fig.9 & Fig.10 The Confusion Matrixes of Addition Method and Multiplication Method in The Training Dataset

It is necessary that the appropriate selection of classifier's performance metrics determines the validity of studied classification model. Confusion matrix in the classifier could generate four sections on describing the similarity and difference between the real-world data and predicted data into four sections: True Positive (TP means it is true in reality and being validated as positive in the prediction), True Negative (TN means it is true in reality but being validated as negative in the prediction), False Positive (FP), and False Negative (FN) could be explained vice versa. Then, an evaluative score on balancing the model's performance would be created through comparing and calculating these four dimensions, ranging from 0 (definitely negative) to 1(definitely positive), with a valid threshold at 0.5 (Brownlee, 2014). Accuracy is a conventional statistical measure of how well the true predictions have dominated the total population, whose formula is $Accuracy = \frac{TP + FN}{TP + TN + FP + FN}$, which is reflected as 0.9909 and 0.9920 in Addition Method and Multiplication Method respectively. However, this single indicator could not determine which model should be chosen because the large amount of FN has polluted the weight of the numerator, but it is unimportant to consider how many trout are being properly classified. Indicators about salmon matter, particularly in those salmon's samples that were misclassified. Precision and recall thus are preferred since their appropriateness in this case---one measures the proportion of correct prediction from all positive class, $Precision = \frac{TP}{TP + FN}$, and another calculates the percentage of how many observations are truly positive from the observations being marked as positive, $Recall = \frac{TP}{TP + NP}$.

To prevent these two tests from being too extreme, the F-Score neutralises them by imposing more punishment, as $F - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$, showing at 0.1679 and 0.1806 for two classification models. Furthermore, 28 salmons were misclassified as trout in Addition Method, which was 24 more in absolute terms than Multiplication Method, which might result in them being caught by fishmen since lacking efficient warnings, thus the performance of the Addition Method is inferior to that of the Multiplication Method. To conclude, the QDA model of Multiplication Method was designated as the fitted model eventually.

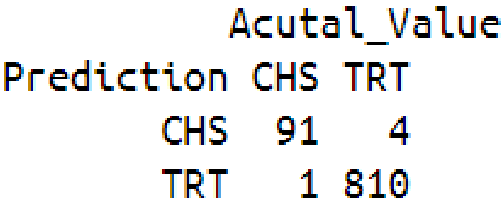
After the algorithm construction, the unused test data, which was also the rest 25% information from the original dataset (Appendix.6), was fitted into the designed model (Fig.11), resulting in F-Score at 0.1823, which was better than the prediction than the training dataset though it is imperfect that two salmon were misrecognised as Trout.



```
           Actual_Value
Prediction  CHS  TRT
       CHS  111    8
       TRT    2 1011
```

Fig.11 The Confusion Matrixes of Multiplication Method in The Test Dataset

However, it is problematic that artificial manipulation of the sample split threshold since the selecting randomisation of training or test datasets could not be eliminated, which provides less information of conclusion's confidence and dubious prediction. Therefore, K-fold resampling method should be advised as it could qualify both the population parameters and the uncertainty of the estimate through a small-sized collected datasets (Montero, 2021). The original dataset is randomly separated into K same-sized folds. Each round, K-1 folds are stochastically selected as the training set, and the only remaining fold is used as the test set. When the round was completed, K-folds repeats until the end of rounds. Although there isn't any available model waiting to being validated on test error, the K-fold method was considered a sampling alternative, which consequences in better performance that only one of salmon was mismatched this time (Fig.12)



```
              Acutal_Value
Prediction CHS TRT
       CHS  91   4
       TRT   1 810
```

Fig.11 The Confusion Matrixes of Multiplication Method in The Test Dataset by Using The K-fold Method

Discussion & Limitation

From a statistical point of view, this model designed for the fish dataset, was very efficient at properly identifying Chinook salmon and trout species, as the results for each of the above sampling methods are more than 99% accurate. Take the results of K-fold's resampling as an example, if a fisherman uploads the biometric information of 1,000 fish on a given day, 994 will be accurately classified; 99 Chinook salmon will be released back into the water; Only one Chinook salmon would be mistaken as a tradable trout, with a probability of 1.1%% percent, which is much less than the 5 percent threshold for an event with a low probability that salmon's ethnical system would not be disturbed by human factors.

Nevertheless, such a model might not be meaningful on problem-solving in the real life, because the operation of complete-event deletion has ruined the reliability of source dataset. Its data structure was unqualified in the MCAR's hypothesis, which implies that the case-wise deletion method would increase the standard errors of parameter estimates when nearly half the sample were discarded (Afghari et al., 2019), especially for a set of observational data that only failed one variable but has to be abandoned entirely (Lord, 2006; Lord and Miranda-Moreno, 2008). Hence it is unrealistic that the average 90% of accuracy rate prediction could be generated by the classification model since it was based on a biased assumption. For instance, due to a slight right-skewness of the Length variable, the author had to set the significant level of as 0.01 during its normal distribution test, rather than 0.05, a common threshold of validation, was because that its p-value is insignificant in statistical meaning that would break the null hypothesis.

Conclusion

This paper established a classification model for Chinook salmon and other species s living in the Great Lakes by classifying their biometric information through data filtering, model selection & validation, and resampling procedures. Three typical classification tools on binary problems have been fully discussed on their adaptive assumptions and appropriateness towards the fitting data. Although the processed dataset was slightly biased, it is surprising that the accuracy of the model is approximate at 99.021% and F1-score is around 0.1832, that only 1%% of salmon would be misclassified as the other species.

References

Afghari, A.P., Washington, S., Prato, C. and Haquea, Md.M. (2019 'Contrasting case-wise deletion with multiple imputation and latent variable approaches to dealing with missing observations in count regression models', *Analytic Methods in Accident Research,* Volume 24.

Alexander, J. & Chronicle, M. (2019) *Collapse of Lake Huron salmon fishery offers lessons.* Available at: https://www.mlive.com/outdoors/2011/04/collapse_of_lake_huron_salmon.html (Accessed: 12 December 2021)

Brownlee, J. (2014) *Assessing and Comparing Classifier Performance with ROC Curves.* Available at: https://machinelearningmastery.com/assessing-comparing-classifier-performance-roc-curves-2/ (Accessed: 15 December 2021)

Caroline, F. (2021) *Pacific Salmon: The "King" Species of the Great Lakes.* Available at: https://storymaps.arcgis.com/stories/381bf0625c2f4654a680bcd513962edb (Accessed: 12 December 2021)

Dong, Y. and Joanne, C. (2013) 'Principled missing data methods for researchers', *Springerplus,* 2(222).

Dziura, J.D., Post, L.A., Zhao, Q., Fu, Z. and Peduzzi, P. (2013) 'Strategies for dealing with missing data in clinical trials: from design to analysis', *Yale Journal of Biology and Medicine,* 86(3), p. 343–358.

Enders, C. (2010) 'Using the Expectation Maximization Algorithm to Estimate Coefficient Alpha for Scales with Item-Level Missing Data', *Psychological Methedology,* 8(3), p. 322–337.

Famoofo, O.O. and Abdul, W.O. (2020) 'Biometry, condition factors and length-weight relationships of sixteen fish species in Iwopin fresh-water ecotype of Lekki Lagoon, Ogun State, Southwest Nigeria', *Heliyon,* 6(1).

Fish and Aquatic Conservation (2021) *Chinook salmon.* Available at: https://www.fws.gov/fisheries/freshwater-fish-of-america/chinook_salmon.html (Accessed: 12 December 2021)

Gutierrez, D. (2013) *TECH TIP: The Importance of Training and Test Set Separation.* Available at: https://insidebigdata.com/2013/11/08/tech-tip-importance-training-test-set-separation/ (Accessed: 14 December 2021)

Kang, H. (2013) 'The prevention and handling of the missing data', *Korean Journal of Anesthesiology,* 64(5), p. 402–406.

Kasapoglu, N. and Duzgunes, E. (2014) 'Length-weight relationships of marine species caught by five gears from the Black Sea', *Mediterranean Marine Science,* 15(1), pp. 95-100.

Leung, K. (2021) *Assumptions of Logistic Regression, Clearly Explained.* Available at: https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290 (Accessed: 14 December 2021)

Mack, C., Su, Z. and Westreich, D. (2018) *Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide.* 3rd Edition ed. s.l.:Rockville (MD): Agency for Healthcare Research and Quality (US).

Matheny, K. (2012) *King salmon reign becomes more precarious on changing Great Lakes.* Available at: https://eu.freep.com/story/news/local/michigan/2017/10/23/king-chinook-salmon-great-lakes-fish/780231001/ (Accessed: 12 December 2021)

Micigan.gov (2021) *Brook Trout.* Available at: https://www.mastersindatascience.org/learning/how-to-deal-with-missing-data/ (Accessed: 12 December 2021)

Montero, M. (2021) '*Resampling Methods for Machine Learning modeling.* Available at: https://medium.com/geekculture/resampling-methods-for-machine-learning-modeling-d2cdc1d3640f (Accessed: 15 December 2021)

Peng, C., Harwell, M., Liou, S. and Ehman, L. (2006) 'Real data analysis', *Advances in missing data methods and implications for educational research,* pp. 31-78.

Schafer, J. (1999) 'Multiple imputation: a primer', *Statistical Methods in Medical Research,* 8(1), pp. 3-15.

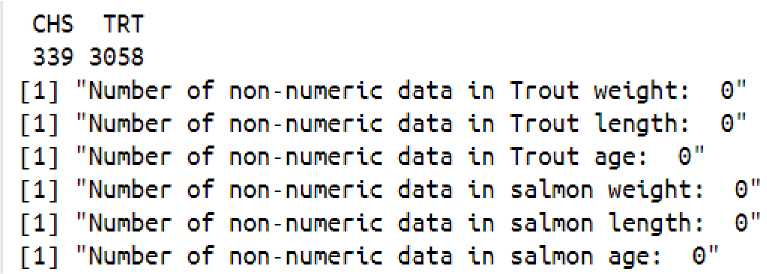
Shekhar, P. (2018) *How to perform Logistic Regression, LDA, & QDA in R.* Available at: https://datascienceplus.com/how-to-perform-logistic-regression-lda-qda-in-r/ (Accessed: 15 December 2021)

Sperandei, S. (2014) 'Understanding logistic regression analysis', *Biochemia Medica,* 24(1), p. 12–18.

Swalin, A. (2018) *How to Handle Missing Data.* Available at: https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4 (Accessed: 14 December 2021)

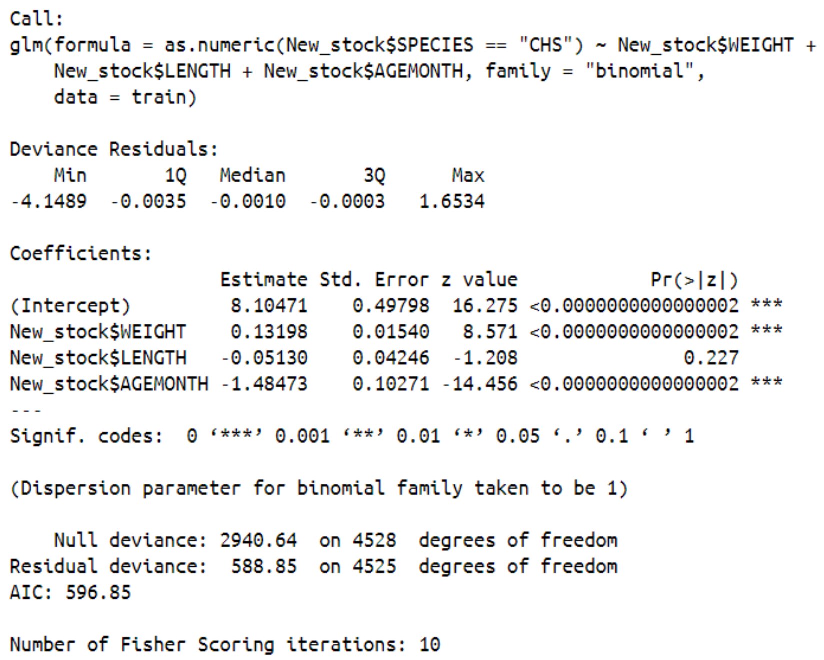
Zhou, M. and Shao, Y. (2014) 'A Powerful Test for Multivariate Normality', *Journal of Applied Statistics,* 41(2), p. 351–363.

Appendices



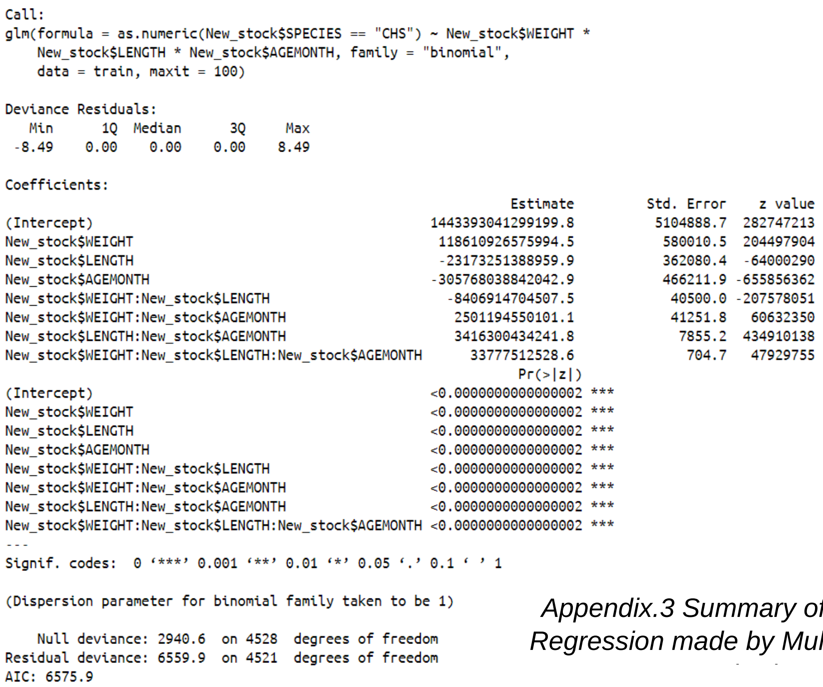
```
 CHS  TRT 
 339 3058 
[1] "Number of non-numeric data in Trout weight:  0"
[1] "Number of non-numeric data in Trout length:  0"
[1] "Number of non-numeric data in Trout age:  0"
[1] "Number of non-numeric data in salmon weight:  0"
[1] "Number of non-numeric data in salmon length:  0"
[1] "Number of non-numeric data in salmon age:  0"
```

Appendix.1 The Outcome Type and Number of Non-Numeric Data in Variables



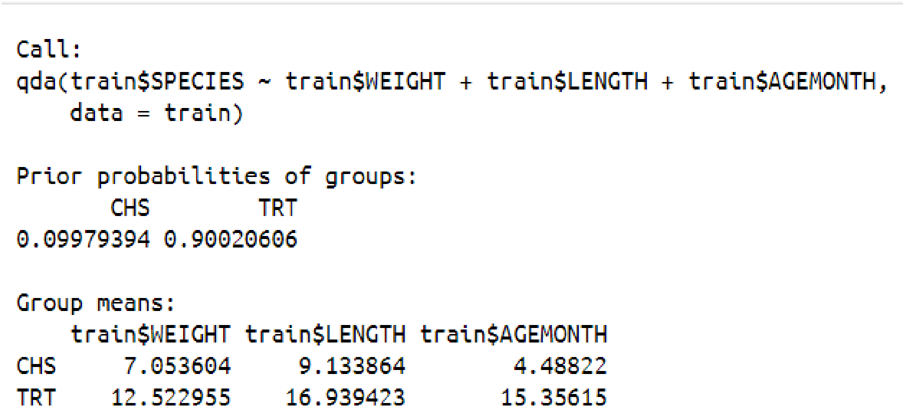
```
Call:
glm(formula = as.numeric(New_stock$SPECIES == "CHS") ~ New_stock$WEIGHT + 
    New_stock$LENGTH + New_stock$AGEMONTH, family = "binomial", 
    data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-4.1489  -0.0035  -0.0010  -0.0003   1.6534  

Coefficients:
                   Estimate Std. Error z value             Pr(>|z|)    
(Intercept)         8.10471    0.49798  16.275 <0.0000000000000002 ***
New_stock$WEIGHT    0.13198    0.01540   8.571 <0.0000000000000002 ***
New_stock$LENGTH   -0.05130    0.04246  -1.208                0.227    
New_stock$AGEMONTH -1.48473    0.10271 -14.456 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2940.64  on 4528  degrees of freedom
Residual deviance:  588.85  on 4525  degrees of freedom
AIC: 596.85

Number of Fisher Scoring iterations: 10
```

Appendix.2 Summary of Logistic Regression made by Addition Method



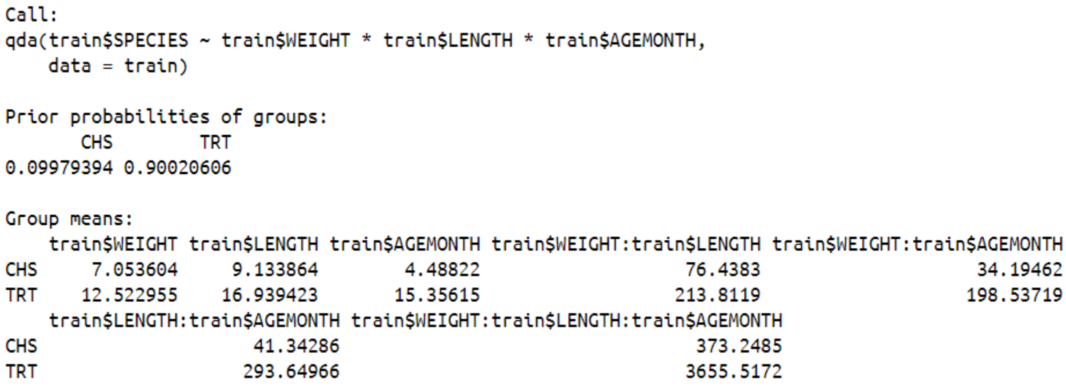
```
Call:
glm(formula = as.numeric(New_stock$SPECIES == "CHS") ~ New_stock$WEIGHT * 
    New_stock$LENGTH * New_stock$AGEMONTH, family = "binomial", 
    data = train, maxit = 100)

Deviance Residuals: 
   Min      1Q  Median      3Q     Max  
 -8.49    0.00    0.00    0.00    8.49  

Coefficients:
                                                               Estimate         Std. Error    z value
(Intercept)                                          1443393041299199.8          5104888.7  282747213
New_stock$WEIGHT                                      118610926575994.5           580010.5  204497904
New_stock$LENGTH                                      -23173251388959.9           362080.4  -64000290
New_stock$AGEMONTH                                   -305768038842042.9           466211.9 -655856362
New_stock$WEIGHT:New_stock$LENGTH                      -8406914704507.5            40500.0 -207578051
New_stock$WEIGHT:New_stock$AGEMONTH                     2501194550101.1            41251.8   60632350
New_stock$LENGTH:New_stock$AGEMONTH                     3416300434241.8             7855.2  434910138
New_stock$WEIGHT:New_stock$LENGTH:New_stock$AGEMONTH      33777512528.6              704.7   47929755
                                                                Pr(>|z|)    
(Intercept)                                          <0.0000000000000002 ***
New_stock$WEIGHT                                     <0.0000000000000002 ***
New_stock$LENGTH                                     <0.0000000000000002 ***
New_stock$AGEMONTH                                   <0.0000000000000002 ***
New_stock$WEIGHT:New_stock$LENGTH                    <0.0000000000000002 ***
New_stock$WEIGHT:New_stock$AGEMONTH                  <0.0000000000000002 ***
New_stock$LENGTH:New_stock$AGEMONTH                  <0.0000000000000002 ***
New_stock$WEIGHT:New_stock$LENGTH:New_stock$AGEMONTH <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2940.6  on 4528  degrees of freedom
Residual deviance: 6559.9  on 4521  degrees of freedom
AIC: 6575.9
```

Appendix.3 Summary of Logistic Regression made by Multiplication



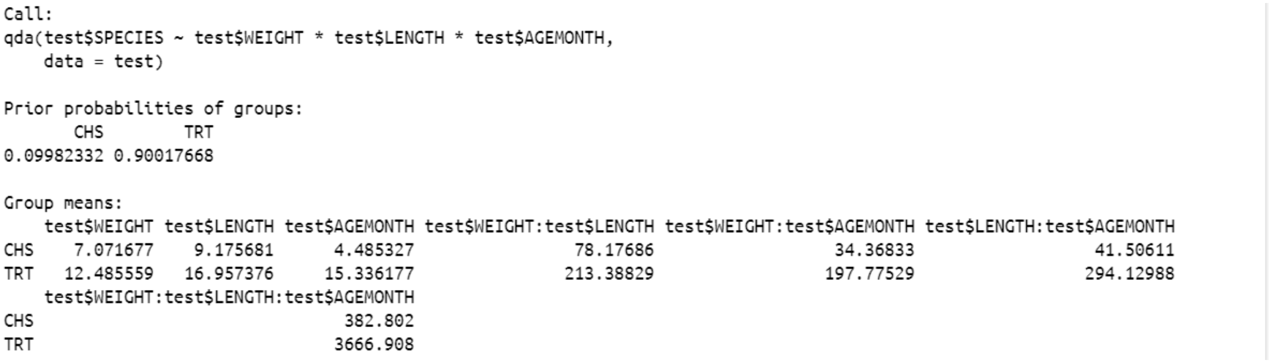
```
Call:
qda(train$SPECIES ~ train$WEIGHT + train$LENGTH + train$AGEMONTH,
    data = train)

Prior probabilities of groups:
       CHS        TRT
0.09979394 0.90020606

Group means:
    train$WEIGHT train$LENGTH train$AGEMONTH
CHS     7.053604     9.133864        4.48822
TRT    12.522955    16.939423       15.35615
```

Appendix.4 Summary of QDA made by Addition Method



```
Call:
qda(train$SPECIES ~ train$WEIGHT * train$LENGTH * train$AGEMONTH,
    data = train)

Prior probabilities of groups:
       CHS        TRT
0.09979394 0.90020606

Group means:
    train$WEIGHT train$LENGTH train$AGEMONTH train$WEIGHT:train$LENGTH train$WEIGHT:train$AGEMONTH
CHS     7.053604     9.133864        4.48822                   76.4383                    34.19462
TRT    12.522955    16.939423       15.35615                  213.8119                   198.53719
    train$LENGTH:train$AGEMONTH train$WEIGHT:train$LENGTH:train$AGEMONTH
CHS                    41.34286                                 373.2485
TRT                   293.64966                                3655.5172
```

Appendix.5 Summary of QDA made by Multiplication Method in The Training Dataset



```
Call:
qda(test$SPECIES ~ test$WEIGHT * test$LENGTH * test$AGEMONTH,
    data = test)

Prior probabilities of groups:
       CHS        TRT
0.09982332 0.90017668

Group means:
    test$WEIGHT test$LENGTH test$AGEMONTH test$WEIGHT:test$LENGTH test$WEIGHT:test$AGEMONTH test$LENGTH:test$AGEMONTH
CHS    7.071677    9.175681      4.485327                78.17686                  34.36833                  41.50611
TRT   12.485559   16.957376     15.336177               213.38829                 197.77529                 294.12988
    test$WEIGHT:test$LENGTH:test$AGEMONTH
CHS                               382.802
TRT                              3666.908
```

Appendix.6 Summary of QDA made by Multiplication Method in The Test Dataset