Species Classification

Mohamed Ahmed

11/25/2020

Introduction:

We have two distinct species Microtus Subterraneus and Microtus Multiplex. The two species have different chromosomes. The two species overlap in the alps of southern Switzerland and northern Italy. In some regions such as Belgium, they can be found in the same elevations. The two species morphologically are indistinguishable. We know that there is not a reliable criterion. to tell apart the two species, based on the cranial morphology. It would be useful to for researchers to have a reliable way to distinguish between the two species. This project's objective is to develop a model that can help research identify which species is based on their morphometric characteristics.

Data:

The data that was used for this project was collected during a study. During the study, eight morphometric variables were collected by one of the authors (Salvioni). The data is made up of 288 specimens which was collected in central Europe and in Toscana. Out of the 288 specimens' samples only 89 specimens the chromosomes were analyzed to identify the species. The remaining 199 specimens' samples were not identified and only their morphometric characteristics are given. The table below shows a brief description of the data.

Exploratory data analysis:

From out summary statistics, we can see that we have three different groups: multiplex with 43 records, 46 subterraneus records, and unknown with 199 records. The means and the medians or all variables are pretty close, and we can say that all variables seem to have a normal distribution. **Figure 1** shows Univariate Graphs of the data such as kernel density plots, box plots, histograms, scatter plots, and bar plots. Different levels have different colors. blue represents (Unknown), Green represents(subterraneus), and red represents(multiplex). The box plots show that there is not much difference in means between different group for all variables. We can see that histograms show a normal distribution of the data. The qq plot shows normal distribution as well. scatter plots indicate that there is a strong correlation between many independent variables and the relationship between most variables seems to be linear. The bar plot shows that out of the three groups the unknown group is most occurring group which will not be good for our model because the training set will be small. Next, we constructed a heat map to look at the

correlation between independent variables. **Table 2** shows the variables with the lowest correlation between them.

Model:

Assumptions:

We will assume binary logistic regression, linearity of independent variable, and that there is small to no multicollinearity
Since the problem is a classification problem and we were asked to use generalized liner models, a logistic model is suitable for this kind of classification problems. Since we were given 89 specimens samples that are known and we want to predict the 199 unknown samples, we split our data into two different data sets. one to train the model and another dataset to use for predictions. also, we have converted our two classes into 1 and o for modeling purposes.

Model Selection:

We have selected our model by trial-and-error method, and we have attempted to use stepwise selection method using both directions which gives the model with the best AIC Value. We have the selection process by constructing a logistic model using all explanatory variables. None of the explanatory variables for this model were significant at $\alpha = 0.5$. To check the model accuracy, we calculated mean squared error and we performed 10-fold cross validation. MSE for this model was low *2.6%* and cross validation accuracy for this model was high **8.1%**. The second model that was constructed with variables that we considered to have low correlation from **Table 2**. This model had two variables that were significant at $\alpha = 0.5$ which were **Intercept & M1Left**. There rest of the explanatory variables were insignificant. The MSE for this model was **3.5%** and the cross-validation error was **5.1%**. The next step was to use the stepwise selection method. The MSE for selected model by the procedure was **2.8%** and the cross-validation error was **5.8**. After a trial and error process, we selected a simple model with two explanatory variables. This model has the lowest cross validation error **4.6%**. we used cross validation error as a criterion to select the best model. **Table 3** summarizes different error measures for different models.

Selected Model Interpretation:

This model was selected because it had the lowest cross validation error among all the models that were constructed. Also, the model has an AIC that is close to the AIC of the model that was given by the stepwise method. **Table 3** shows the **AIC** values for all models. The model has the intercept and two explanatory variables **Intercept & M1Left**. all of them were significant at an $\alpha = 0.5$ level. The AIC for the model was low but not the lowest compared to all the other models. The difference between null deviance and residual deviance gives us an idea whether we have a good fit or not. The difference between the

Null device and the residual deviance is big enough to consider our model good. from the cross validation, we conclude that this model is about 95.4% accurate.

Conclusion and Recommendations:

We have explored the Microtus data and did explanatory data analysis. From our analysis, we determined that a logistic model is suitable for this problem. We have selected the most optimal model using 10 K fold cross validation to measure the accuracy of our model and we determined that our best model had approximately **95%** accuracy. We trained our model using the 89 known specimens and we used the unknown 199 samples to make prediction. We were able to classify the 199-unknown data into the two known classes. Further details about the classified specimens can be found in **Table 5** This model can help researchers and scientist save a lot of time by using the model to classify specimens. Also, the researchers will be able to focus more on collecting unique characteristics data that can help data scientist build a model that can identify the specimens faster.

Figures And Tables:

##Data

Symbol	Description
$Group$	a factor with levels multiplex subterraneus unknown code
$M1Left$	Width of upper left molar 1 (0.001mm)
$M2Left$	Width of upper left molar 2 (0.001mm) 1=censored, 2=dead
$M3Left$	Width of upper left molar 3 (0.001mm)
$Foramen$	Length of incisive foramen (0.001mm)
$Pbone$	Length of palatal bone (0.001mm) symptomatic but completely ambulatory, 2=in bed<50% ofthe day,3=in bed>50% of the day but not bedbound, 4 = bedbound
$Length$	Condylo incisive length or skull length (0.01mm)
$Height$	Skull height above bullae (0.01mm)
$Rostrum$	Skull width across rostrum (0.01mm)

Table 1: Statistical Summary

Group	M1Left	M2Left	M3Left	Foramen	Pbone	Length	Height	Rostrum
multiplex : 43	Min. :1534	Min. :1355	Min. :1361	Min. :3155	Min. :3928	Min. :1908	Min. :700.0	Min. :375.0
subterraneus: 46	1st Qu.:17	1st Qu.:15	1st Qu.:15	1st Qu.:37	1st Qu.:48	1st Qu.:22	1st Qu.:759	1st Qu.:425

	83	03	95	51	15	27	.2	.0
unknown :199	Median :1923	Median :1570	Median :1724	Median :3932	Median :5079	Median :2312	Median :789.0	Median :450.0
NA	Mean :1935	Mean :1589	Mean :1727	Mean :3913	Mean :5082	Mean :2309	Mean :790.8	Mean :451.2
NA	3rd Qu.:2074	3rd Qu.:1660	3rd Qu.:1856	3rd Qu.:4080	3rd Qu.:5328	3rd Qu.:2388	3rd Qu.:817.8	3rd Qu.:475.0
NA	Max. :2479	Max. :1880	Max. :2187	Max. :4662	Max. :6104	Max. :2605	Max. :912.0	Max. :545.0

Figure 1: Matrix Of Plots

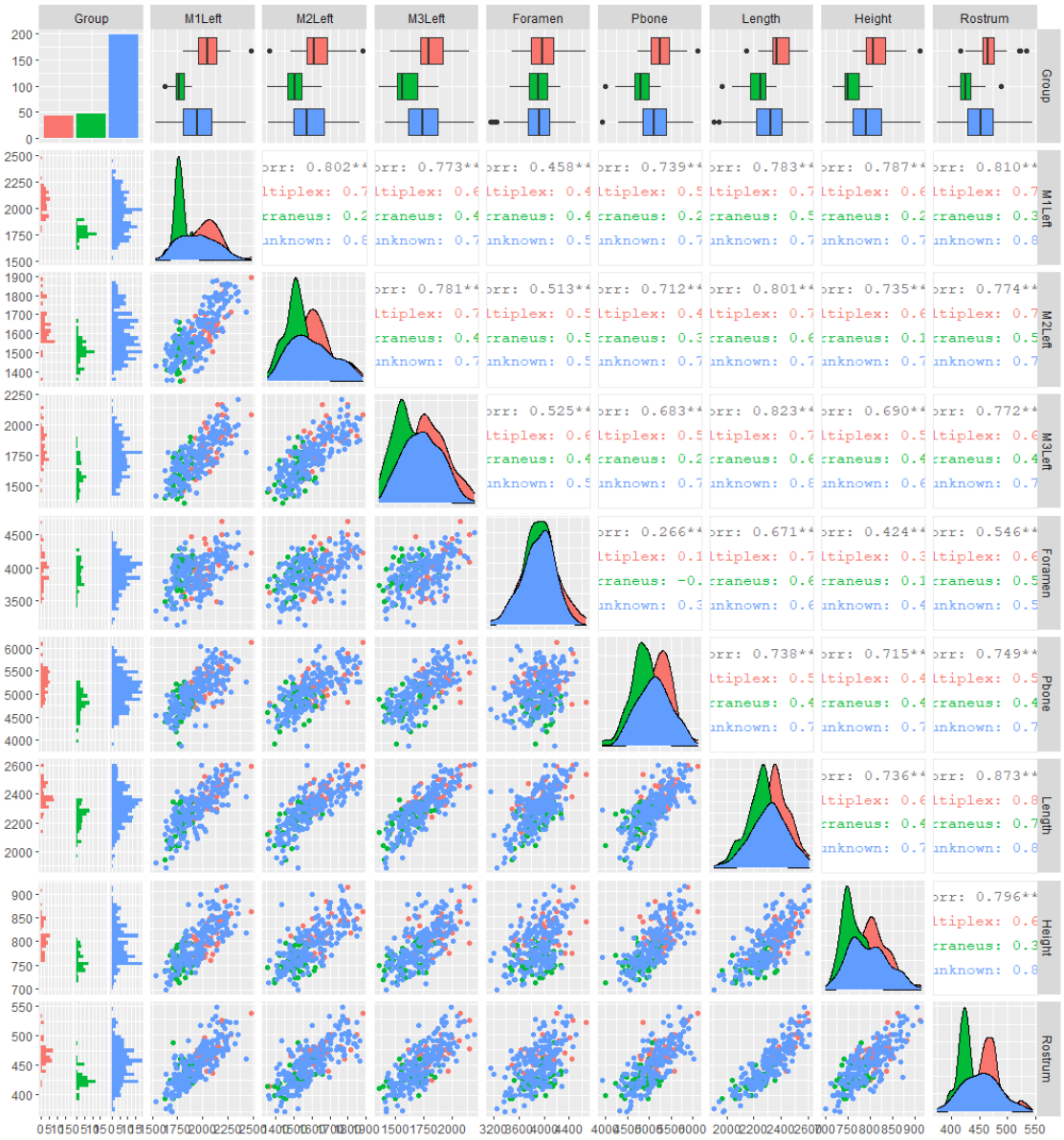

Figure 2: Correlation Matrix Graph

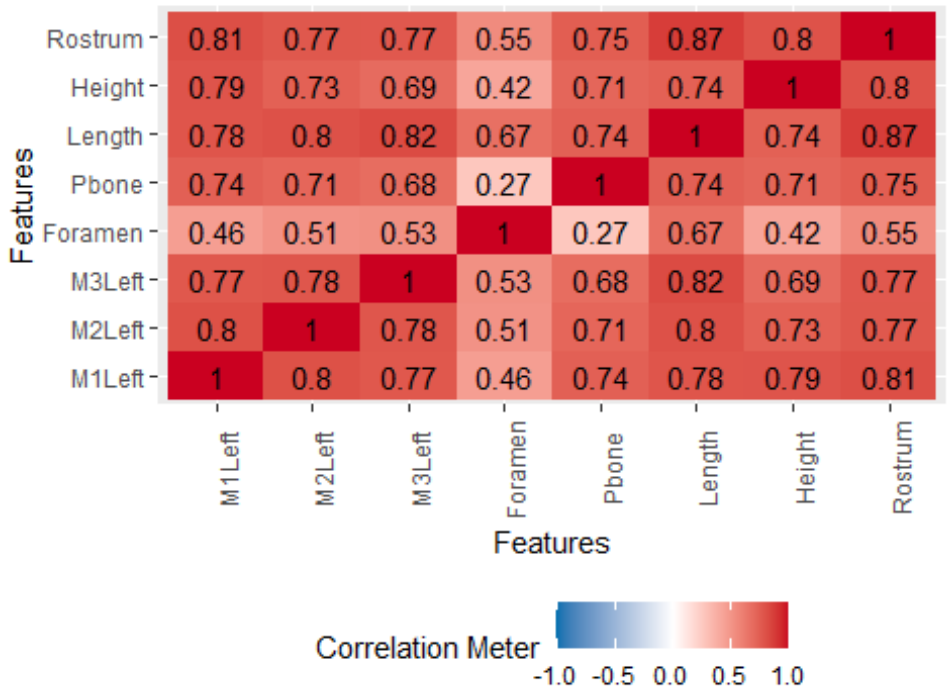

Table 2: Varaibles with low Correlations

low_corr_Variables
Pbone
Foramen
M1left
Height

Table 3: Different Models Errors

Model	MSE	Cross_Validation_Error	AIC
Full Model	0.0260759	0.0810197	32.96195
Low.Correlation.var.Model	0.0349956	0.0506961	30.93902
Stepwise Model	0.0274013	0.0582508	27.70264
Selected Model	0.0367974	0.0463364	28.04904

```
##
## Call:
## glm(formula = Group ~ M1Left + Foramen, family = binomial(),
##     data = model.data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.28036  -0.09923  -0.01058   0.01788   2.49687
##
## Coefficients:
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -62.804452  20.661080  -3.040  0.00237 **
## M1Left        0.047246   0.014091   3.353  0.00080 ***
## Foramen      -0.006637   0.003192  -2.079  0.03758 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 123.279  on 88  degrees of freedom
## Residual deviance:  22.049  on 86  degrees of freedom
## AIC: 28.049
##
## Number of Fisher Scoring iterations: 8
```

Table 4: Information on the Selected Model

Cross.Validation	AIC
0.0463364	28.04904

Table 5: Predictions Count

Var1	Freq
multiplex	121
subterraneus	78

Citations

Kassambara, Thanos, Kassambara, & Sfd. (2018, March 11). Logistic Regression Essentials in R. Retrieved December 07, 2020, from http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/

Everitt, Brian, and Torsten Hothorn. A Handbook of Statistical Analyses Using n SECOND EDITION. Taylor and Francis Group LLC, 2010.