MATH333 Group Project

Project 1: Dog class

Adam Liversidge, Diya Guo, James Robinson

0.1 Introduction

A researcher in animal psychology is investigating the effectiveness of group obedience classes on the behaviour of pet dogs that have been showing signs of aggression and general bad behaviour. These dogs have attended a 3 month course of classes, and data about their characteristics and background has been collected, as well as an indication of whether the dog was thought to have improved by the end of the course.
We are interested in predicting the probability of whether a dog with a certain combination of explanatory variables will improve or not by the end of the course. For example, we may want to find out whether a dog who goes through the program as a puppy is more or less likely to improve than an adult or senior dog. Furthermore, we'd like to analyse if our methods of predicting these probabilities is accurate, such that we could continue to use our methods to identify things such as high risks groups, or whether the classes are effective for certain types of dogs.

0.2 Methods

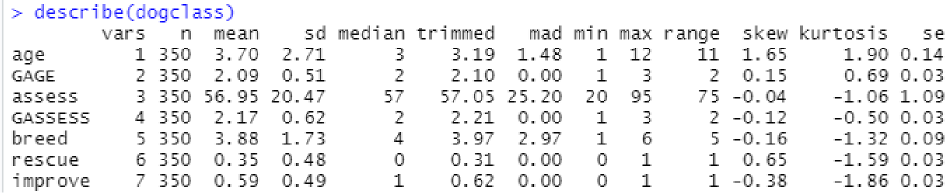
Our original dataset shall be analysed in many ways throughout this report. Firstly, we shall see what conclusions we can draw from our initial data; this shall include calculating the mean values of each variable, as well as standard deviation/error, among others. Furthermore, we shall determine what relationships exist between each variable in our dataset by creating graphs (our first 6 Figures) in the hope of spotting initial correlations. Secondly, we shall use R to determine whether our initial dataset is missing any values, as this could ruin our model and cause bugs in our code. This shall be done through the use of logistic regression.

Once we have determined if we are missing any values, we shall use correlation analysis to prove whether any relationships do in fact exist between our variables, whether they are weak correlations or strong ones. Furthermore, we shall use Logistic Regression Analysis to train our model in R, this will allow us to see whether any variables have a statistical effect on each other by monitoring the p-values of our model, this could indicate which variables effect a dog's improvement. Finally, we shall run ANOVA model in order to analyse the deviance of each variable before calculating the predicting capability of our fitted model to determine it's reliability; then conclude our results.

0.3 Analysis

0.3.1 Descriptive Statistics and Relationships Between Each Variable

Firstly, we have taken our original dataset, which comprises of data from 350 dogs, and derived some descriptive statistics. These results are as shown below:



```
> describe(dogclass)
        vars   n  mean    sd median trimmed   mad min max range  skew kurtosis   se
age        1 350  3.70  2.71      3    3.19  1.48   1  12    11  1.65     1.90 0.14
GAGE       2 350  2.09  0.51      2    2.10  0.00   1   3     2  0.15     0.69 0.03
assess     3 350 56.95 20.47     57   57.05 25.20  20  95    75 -0.04    -1.06 1.09
GASSESS    4 350  2.17  0.62      2    2.21  0.00   1   3     2 -0.12    -0.50 0.03
breed      5 350  3.88  1.73      4    3.97  2.97   1   6     5 -0.16    -1.32 0.09
rescue     6 350  0.35  0.48      0    0.31  0.00   0   1     1  0.65    -1.59 0.03
improve    7 350  0.59  0.49      1    0.62  0.00   0   1     1 -0.38    -1.86 0.03
```

In order to determine the relationships between several variables and our main focus - improvement, we decided to draw out some graphs to show their relationships, as seen below:

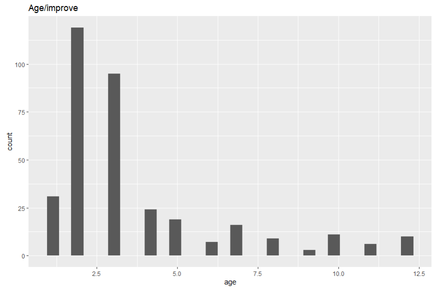


Figure 1: Relationship Between Age and Improvement

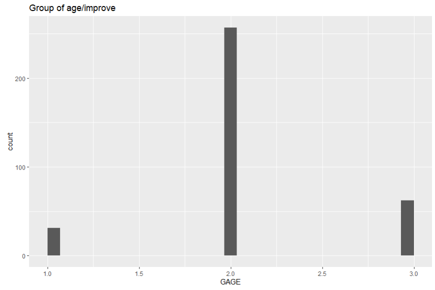


Figure 2: Relationship Between GAGE and Improvement

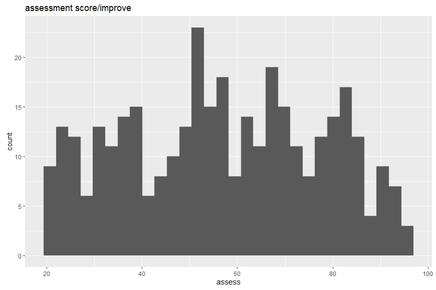


Figure 3: Relationship Between Assess and Improvement

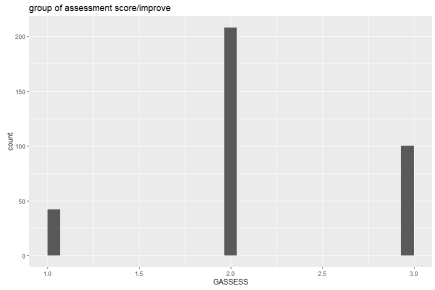


Figure 4: Relationship Between GASSESS and Improvement

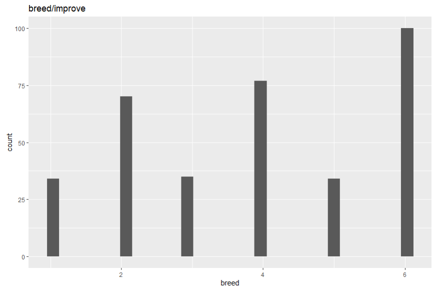


Figure 5: Relationship Between Breed and Improvement

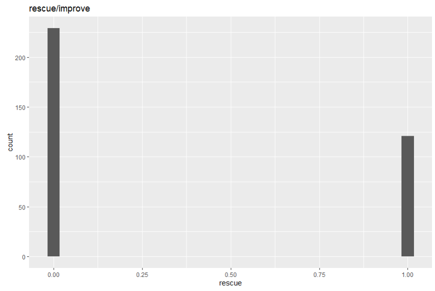


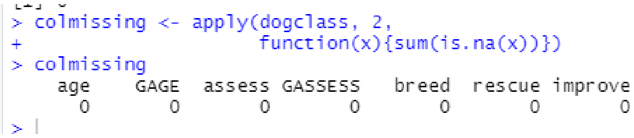
Figure 6: Relationship Between Rescue Status and Improvement

By investigating the frequencies of our 6 variables over improvement, we are able to determine whether we find similar results to those we have gained through the analysis of our descriptive statics. Furthermore, we hope to be able to spot which variables appear to have the largest influence on a dog's improvement over the 3 month course. We don't yet know whether any of these conclusions are a result of the distribution of each variable; therefore, we will run logistic regression in the following steps to deeper dig in the relationships among them.

0.3.2 Missing Values and Correlation Analysis

Missing values have the potential to ruin our model, by causing bugs in code, or by negatively impacting the accuracy of our model. Correlation analysis can be performed to identify highly related variables that might autocorrelated.

Before running logistic regression, we must determine if our dataset has any missing values. The following R-code shows the expected missing values from each category:



```
> colmissing <- apply(dogclass, 2,
+                     function(x){sum(is.na(x))})
> colmissing
    age    GAGE  assess GASSESS   breed  rescue improve
      0       0       0       0       0       0       0
> 
```

Now we know that our data is unlikely to be missing values, we can plot the correlation matrix:

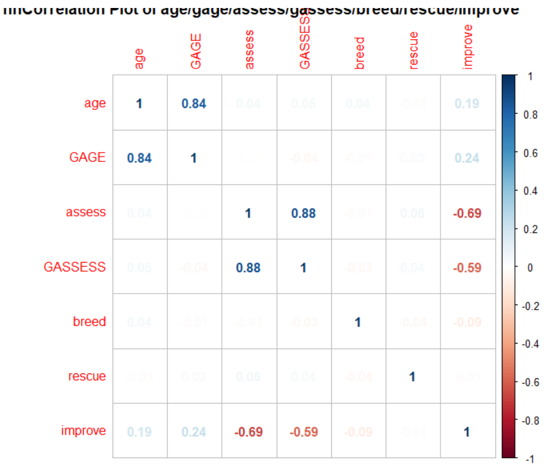


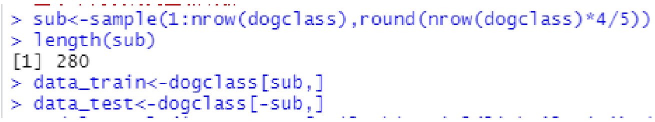
Figure 7: Correlation Plot of Age/GAGE/Assess/GASSESS/Breed/Rescue/Improve

From this matrix we are able to determine whether or not each variable has a correlation with another, as well as the nature of said correlation. The following table can be used to evaluate the correlations between variables:

Correlation Coefficient	Interpretation
0.90 - 1.00	Very Strong Correlation
0.70 - 0.89	Strong Correlation
0.40 - 0.69	Moderate Correlation
0.10 - 0.39	Weak Correlation
0.00 - 0.10	Negligible Correlation

Table 1: Correlation Between Variables

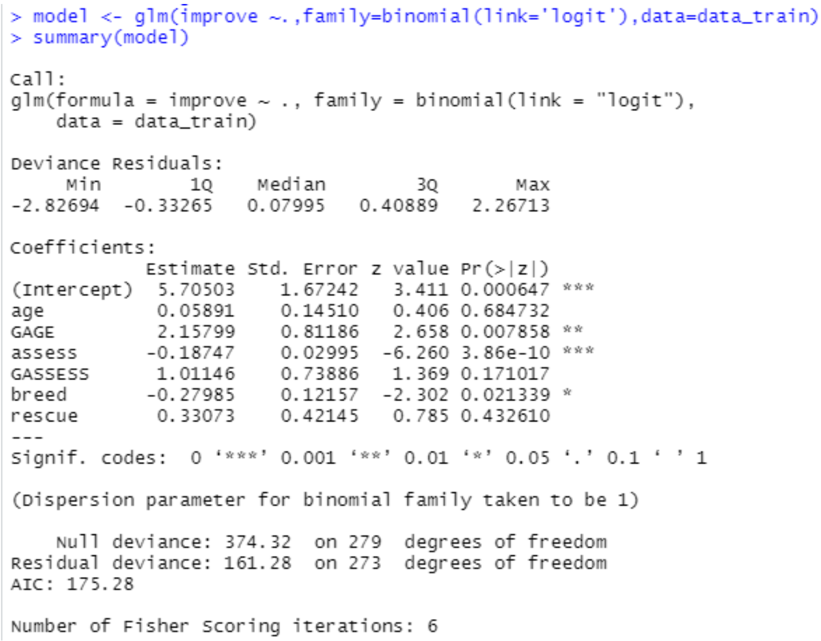
Given we have no missing values in our data, the next step will be implementing the model directly. In order to test the capability of our trained logistic regression model, we split our dataset into two parts: train dataset with 80% data (280), test dataset with 20% data (70), as seen below:



```
> sub<-sample(1:nrow(dogclass),round(nrow(dogclass)*4/5))
> length(sub)
[1] 280
> data_train<-dogclass[sub,]
> data_test<-dogclass[-sub,]
```

0.3.3 Logistic Regression Analysis

Now we train our model and the result is as follows:

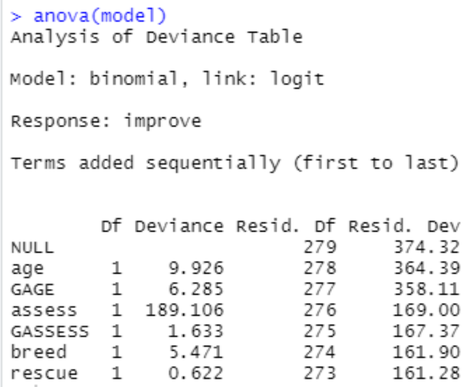


```
> model <- glm(improve ~.,family=binomial(link='logit'),data=data_train)
> summary(model)

Call:
glm(formula = improve ~ ., family = binomial(link = "logit"), 
    data = data_train)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.82694  -0.33265   0.07995   0.40889   2.26713  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  5.70503    1.67242   3.411 0.000647 ***
age          0.05891    0.14510   0.406 0.684732    
GAGE         2.15799    0.81186   2.658 0.007858 ** 
assess      -0.18747    0.02995  -6.260 3.86e-10 ***
GASSESS      1.01146    0.73886   1.369 0.171017    
breed       -0.27985    0.12157  -2.302 0.021339 *  
rescue       0.33073    0.42145   0.785 0.432610    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 374.32  on 279  degrees of freedom
Residual deviance: 161.28  on 273  degrees of freedom
AIC: 175.28

Number of Fisher Scoring iterations: 6
```

From this we will be able to conclusively show in our results which variables have the largest influence on a dog's improvement, and begin to make predictions as to which dog's would be most likely to improve in a later class.

0.3.4 ANOVA model and Predicting Capability

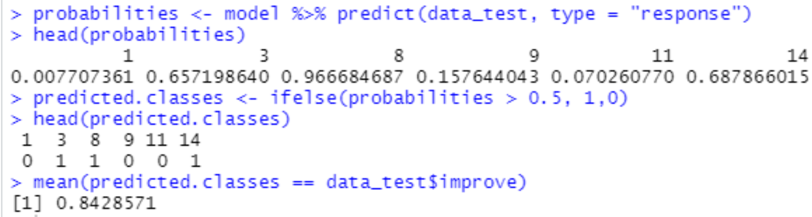
Now we run ANOVA model to analyze the deviance:



```
> anova(model)
Analysis of Deviance Table

Model: binomial, link: logit

Response: improve

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev
NULL                      279     374.32
age      1    9.926       278     364.39
GAGE     1    6.285       277     358.11
assess   1  189.106       276     169.00
GASSESS  1    1.633       275     167.37
breed    1    5.471       274     161.90
rescue   1    0.622       273     161.28
```

Using these results we shall be able to analyse the null diviance against other deviances, allowing us to further determine whether our model is in fact a good fit.

Finally, we want to know the predicting capability of our fitted model. So, we let our model output the probability of improvement given several variable inputs from our test dataset and let the probability of imrovment equal 1 if it greater than 0.5, and 0 otherwise; the results of this can be seen below:



```
> probabilities <- model %>% predict(data_test, type = "response")
> head(probabilities)
          1           3           8           9          11          14 
0.007707361 0.657198640 0.966684687 0.157644043 0.070260770 0.687866015 
> predicted.classes <- ifelse(probabilities > 0.5, 1,0)
> head(predicted.classes)
 1  3  8  9 11 14 
 0  1  1  0  0  1 
> mean(predicted.classes == data_test$improve)
[1] 0.8428571
```

Using this model we are able to determine the prediction accuracy of our fitted model, and as a result gain the wrong classification error too.

0.4 Results

0.4.1 Descriptive Statistics and Relationships Between Each Variable

Certain values within our descriptive statistics are worth noting. For instance, we can see that the mean value for age and GAGE (age groups) in our dataset are 3.70 and 2.09 respectively, which means that most dogs in the our data are adults, centralized around 3.7 years old. When looking at the standard error of for ”assess” (20.47), we can see that there exists a great difference in the individual level of our dogs.

From our first 6 Figures (labelled in the analysis section of this report) we can see that we find similar results and come to similar conclusions, to those taken from our descriptive statistics. For instance, the largest improvement in dogs are mainly within those concentrated near age 2.5 and group 2 of GAGE.

However, we also find that initial assessment score of the dog's behaviour (”assess”) are far more influential on the improvement of a dog. Moreover, dogs that have not been rescued, and are within breeds 2,4 and 6 have a larger probability of improvement than others. However, we don't yet know whether this is because of the distribution of each variable.

0.4.2 Missing Values and Correlation Analysis

By interpretation of our R code, as seen in the analysis section of this report, we have determined that our original data is not missing any values, which means there is no need to worry about a multicollinearity problem.

0.4.3 Correlation Analysis

From Figure 7 in the analysis section of this report, we can see that only ”age” and ”GAGE”, ”assess” and ”GASSESS” are correlated due to their characteristics. Other variables are all in weak or negligible correlation. This is obvious by following the rule of thumb applied to evaluate the correlations between variables, as seen in our table below Figure 7.

0.4.4 Logistic Regression Analysis

First, we can see that the variables ”age”, ”GASSESS” and ”rescue” do not have a statistical effect on improvement. As for statistically significant variables, ”assess” has the lowest p-value (as seen in our logistic regression analysis in R) which implies that a dog's initial assessment has a strong link to it's improvement. Also, the negativeness of the estimate of ”assess” shows that if all other variables were to remain the same, higher values for ”breed” and ”assess” might result in a lower probability of improvement. We can roughly guess that if a dog does not improve, it is unlikely to be in age group (GAGE”) 2; have high assessment value (”assess”); a higher number for breed; and is not rescued.

0.4.5 ANOVA model and Predicting Capability

As we know, the higher the difference between null deviance and other deviances, the better. By analyzing the chart generated in R in our analysis section, we can get the circumstance of deviance when adding one more variable. Furthermore, adding ”age”, ”GAGE” and ”assess” can significantly reduce residual deviances. Finally, by interpreting our final R code, we can see that the prediction accuracy of classification is around 84%, which means a wrong classification error of around 16%.

0.5 Conclusion

Following from our prediction accuracy of around 84%, we can argue that our model is a good predictor of whether a certain class has an effect on each dog's improvement. As well as this, we can see from our results section that some of our variables have strong correlations, as shown in the matrix within our analysis section. Furthermore, we have shown that certain variables have very little effect on the improvement of a dog, such as the rescue status of a dog, and their ”GASSESS” score. Moreover, we have also seen that multiple variables do have a significant statistical effect on the improvement of a dog. It seems that ”assess” is the variable who's value mostly effects the ”improve” variable, due to its low p-value. All in all, we have shown which variables are most likely to effect the improvement of a dog, and in what way, whether it be a positive or negative impact, as well as creating an accurate model in determining the success of individual dogs.