**Classifying Possum Sex with Logistic Regression**

**Possums Dataset from Kaggle**

Tess Anderson, [tanderson4@bellarmine.edu](mailto:tanderson4@bellarmine.edu)

**ABSTRACT**

In this project, I will complete Exploratory Data Analysis and Logistic Regression on a dataset about Possums containing 14 columns and 104 entries. The subjects are possums trapped in 7 different sites and the columns contain detailed information on measurements of each possum trapped. By analyzing the relationship between the different body measurements and the sex of the possum, I will create a Logistic Regression model that can predict the sex of a possum based on its measurements.

1. **INTRODUCTION**

In this project, I use a dataset from Kaggle titled “Possum Regression”. The data frame contains 9 morphometric measurements on each of the 104 mountian brushtail possums, trapped at seven sites from Southern Victoria to central Queensland. On this dataset, I will perform Exploratory Data Analysis to understand the variables contained in the set and their possible relationships to one another. Following EDA, I will conduct Classification analysis using the LogisticRegression function from the Scikit-learn library in Python. My goal in this analysis is to train a Logistic Regression model to accurately classify a possum as male or female based on measurements of its body.

1. **BACKGROUND**
   1. *Data Set Description*

I found this dataset on Kaggle, however it was originally published in the DAAG R package (datasets used in examples and exercises in the book Maindonald, J.H. and Braun, W.J. (2003, 2007, 2010) “Data Analytics and Graphics Using R”). As mentioned, there are 14 columns, 9 of which are morphometric measurements. The others are case, site, age, sex, population. Site tells at which of the seven sites the possum was trapped. Population tells to which population the possum belongs, either Victoria or Other (New South Wales or Queensland).

* 1. *Machine Learning Model*

Logistic Regression is a Machine Learning Model that provides large amounts of mathematical and statistical information about the relationships between independent variables and one dependent variable. In this case, the independent variables are all variables except sex and the dependent variable is sex. The idea behind Logistic Regression is to find a relationship between features and the probability of a particular outcome. To fit the model to the specific dataset, we must split the dataset so that we have a training set and a testing set. Using the training set, which contains most of the entries, we fit the model to the data so that it can make predictions based on the relationships in the data. Afterwards, we use the rest of the data, the testing set, to compare the model’s predictions to the outcomes shown in the data so that we can measure the accuracy of the predictions.

1. **EXPLORATORY ANALYSIS**

This dataset contains 104 entries with 14 columns with various datatypes. There were a few missing values in two columns – they were filled during data preprocessing.

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| Case | Int64 |
| Site | Int64 |
| Pop | Object |
| Sex | Object |
| Age | Float64 |
| Hdlngth | Float64 |
| Skullw | Float64 |
| Totlngth | Float64 |
| Taill | Float64 |
| Footlgth | Float64 |
| Earconch | Float64 |
| Eye | Float64 |
| Chest | Float64 |
| Belly | Float64 |

Most of the numeric variables are approximately Normally distributed, however there are a few to note that follow a bimodal or otherwise abnormal distribution which we should keep in mind throughout analysis. For example, the histograms of the Ear Conch and Tail Length variables show bimodal distributions, while Age and Total Length show significant skew in their distributions.

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, box and whisker chart

Description automatically generatedChart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Shown to the right are boxplots for each of the nine measurement variables – this allows us to easily see where measures of spread and center are for each variable, as well as the outliers present for each variable.

To explore the data more, I examined the trappings by site. More specifically, I made bar plots showing the count of possums caught at each site by sex and by population.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

1. **METHODS**
   1. *Data Preparation*

There were several steps I had to take during the Data Preprocessing stage in order to prepare the data for analysis. First, there were 2 missing entries in the age column and 1 missing entry in the footlngth column that needed to be filled. To do this, I used the median imputation method because the distributions are not Normal. Once the missing values were filled with the median value of the column, I proceeded with dropping the case column because it contains a unique value for each entry which is not useful for analysis. Finally, I prepared the categorical variables in the set by using the pd.get\_dummies() function which encodes categorical variables into binary code so that they can be easily analyzed with the scikit-learn package later.

* 1. *Experimental Design*

Table X: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | All twelve (12) raw features with 80/10/10 split for train, validate, and test |
| 2 | All four (4) normalized features with 80/10/10 split for train, validate, and test |
| 3 | All four (4) raw features with 70/15/15 split for train, validate, and test |
| 4 | All four (4) normalized features with 70/15/15 split for train, validate, and test |

* 1. *Tools Used*

Describe all of the software tools you used to perform your data preparation and model implementation. For example:

The following tools were used for this analysis: Python v3.5.2 running the Anaconda 4.3.22 environment for Apple Macintosh computer was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 0.18.1, Numpy 1.11.3, Matplotlib 1.5.3, Seaborn 0.7.1, SKLearn 0.18.1. Provide a brief explanation of why you chose these tools.

1. **RESULTS**
   1. *Classification Measures*

Provide the classification measures for each experiment using a confusion matrix and classification report.

* 1. *Discussion of Results*

Explain your model finding using a confusion matrix, the accuracy score and classification report.

* 1. *Problems Encountered*

No project goes perfectly smoothly. Discuss any problems you had with obtaining the data, preparing the data, implementing the model, or evaluating the model. **It would be highly unusual to indicate that you had not problems.**

* 1. *Limitations of Implementation*

Discuss the limitations of your model. Is there is reason it might not be the best way to model the data? What other models might work better?

* 1. *Improvements/Future Work*

What would you like to do to improve your model in future work? Do more experiments, use a different model, add/remove variables, find a different data set, etc?

1. **CONCLUSION**

Finish up with a paragraph or two of summarizing your problem, the results and your conclusions (good model, bad model, needs more work, etc.).

**REFERENCES**

List any websites, books, articles, etc. that you found useful while you worked on this project. It is not necessary to cite the references in the paper unless you specifically mention it in the text.

**Other directions:**

1. 10-pt, Times New Roman, 1” margins all around (if you use this template you are already set).
2. Ensure all tables and figures are numbered appropriately and referenced in the text. See examples above and below.

|  |  |
| --- | --- |
| **Figure 1: Comparison of X/Y from dataset (single plot) (8 pt.)** | **Figure 2: (a) Function Output (b) A against B (multiple plots) (8 pt.)** |