Case Study Report

Introduction

Purpose: The objective of this study is to predict:

- Sex of the subjects
- Dominant male subject

Dataset: The data set is a collection of observer recordings from June 13th - July 10th, 2019. The observations are from a study of a group of 20 Baboons at the Primate Center in France data set (Gelardi et al., 2020, p. 20190737), OBS_data.txt (SocioPatterns, 2020). There is an additional dataset included with the study recording the RFID proximity sensors. This data does not include the type of behavior but may be included to supplement the interactions recorded in the OBS_data.txt.

The Data set contains the following columns:

- DateTime - Time stamp of the event
- Actor - Name of the initiator of the interaction
- Recipient - Name of the target of the interaction
- Behavior - Behavior of the Actor towards the Recipient- 14 Values
    - Resting
    - Grooming
    - Presenting
    - Playing with
    - Grunting-Lip-smacking
    - Supplanting
    - Threatening
    - Submission
    - Touching
    - Avoiding
    - Attacking
    - Carrying
    - Embracing
    - Mounting
    - Copulating
    - Chasing
- Category - Classification of the Behavior

- Duration - Duration of the behavior
- Localization - Enclosure zone
- Point - Point event (Yes or No)

Phase Summary

The Case Study was conducted in three phases. The following sections describe the effort and results of each phase of the project.

Part 1. Exploratory Data Analysis

The first phase consists of conducting an exploratory analysis of the dataset. Identifying obvious data relationships and any apparent anomalies with the data itself. From the data relationship, there are defined sub-groups within the population. Also, there may be other significant relationships, such as those subjects who may appear to be the group leaders or the dominant subjects. This assumption is based upon the amount and type of behavior displayed by the subjects and the targets of the behavior. Several network graphs were created to assist in identifying sub-groups or cliques to further refine the subject and target relationships. The behavior types and the number of occurrences by each subject, otherwise known as Actor, are shown in a bar plot by Figure 1 Behavior Counts by Actor.

Another task conducted during this phase was a preliminary dataset refinement. Based upon the 2 objectives, the column identifying the Sex of each Actor was added to facilitate the first prediction test. Additionally, a Rank column was added to store the Ordinal values for each male subject for the Dominant male prediction. A refinement of the dataset included removing the Date Time, Category, Duration, Localization, Point features. These features did not contribute any significant relationships in fulfilling the prediction scenarios. The rows with the behavior types of Other and Resting were deleted. The counts for these values are far greater than the other behavior types' of counts. This caused an imbalance with the data. The resultant dataset is illustrated by Figure 2 Full Dataset. This dataset is referred to as the Full dataset during the model execution and evaluation tasks. The Rank column is not shown in this figure. The Rank column was added during the model selection task for the Dominant Male test.

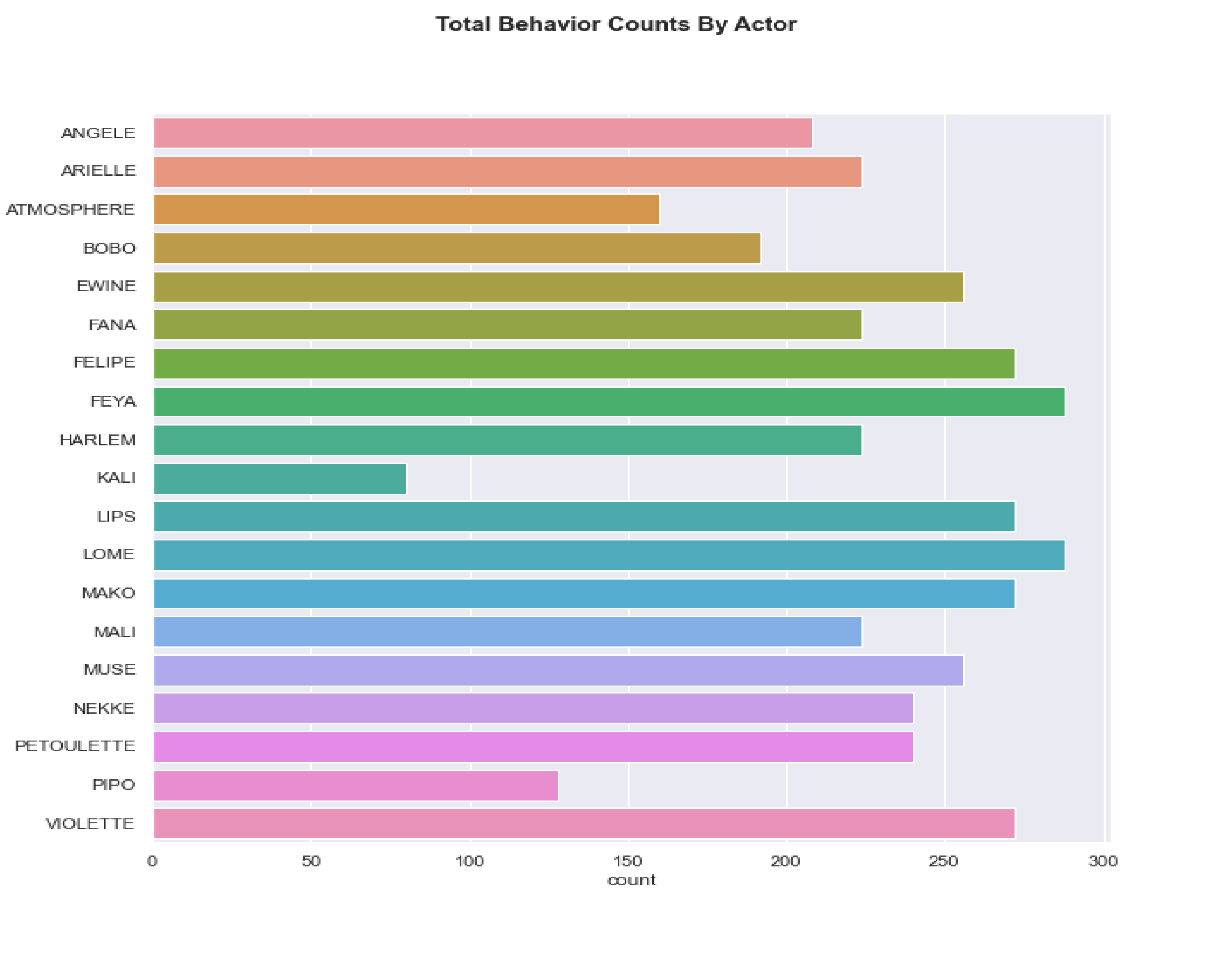
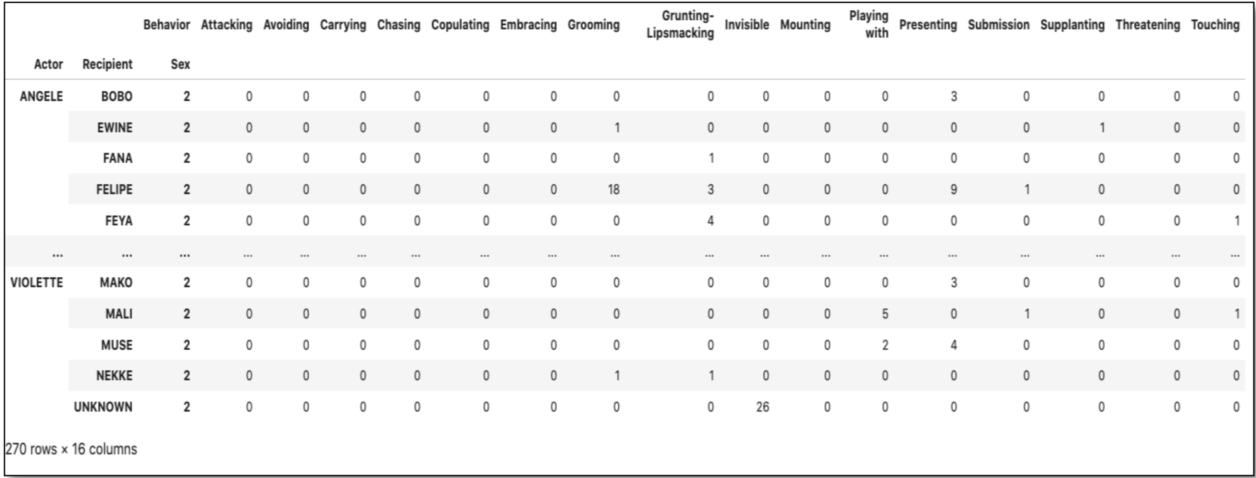


Figure 1 Behavior Counts by Actor



		Behavior	Attacking	Avoiding	Carrying	Chasing	Copulating	Embracing	Grooming	Grunting-Lipsmacking	Invisible	Mounting	Playing with	Presenting	Submission	Supplanting	Threatening	Touching
Actor	Recipient	Sex																
ANGELE	BOBO	2	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
	EWINE	2	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
	FANA	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	FELIPE	2	0	0	0	0	0	0	18	3	0	0	0	9	1	0	0	0
	FEYA	2	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
VIOLETTE	MAKO	2	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
	MALI	2	0	0	0	0	0	0	0	0	0	0	5	0	1	0	0	1
	MUSE	2	0	0	0	0	0	0	0	0	0	0	2	4	0	0	0	0
	NEKKE	2	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
	UNKNOWN	2	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0

270 rows × 16 columns

Figure 2 Full Dataset

Part 2. Feature Selection Phase

The correlation coefficients were computed for the behavior types to identify any highly correlated features. Additionally, the feature selection methods include the Chi Square and the RFEC methods. Each feature selection process uses a long format version of the full dataset. This dataframe consists of separate rows for each Actor/Recipient/Behavior type occurrence. It is the same dataframe shown previously in Figure 1 Full Dataset.

Correlation Coefficient Results

The results of the correlation coefficient computation show only a strong correlation for two pairwise sets of features. The remaining pairwise sets, the correlations are weak which suggests that the data does not have any linear relationships except for 2 sets of features. Figure 3 Correlation Heatmap illustrates the correlation coefficients.

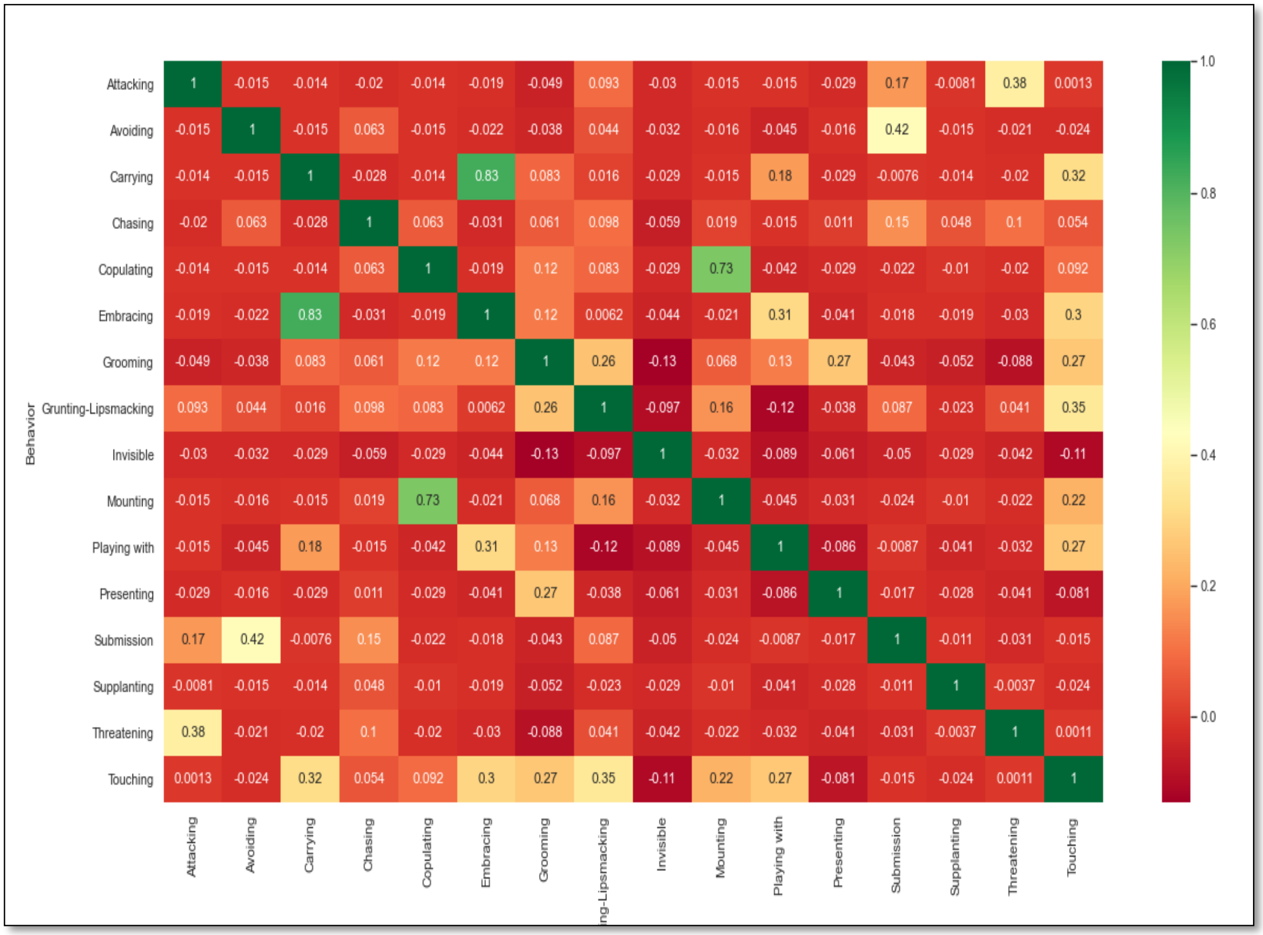


Figure 3 Correlation Heatmap

Chi Square Test

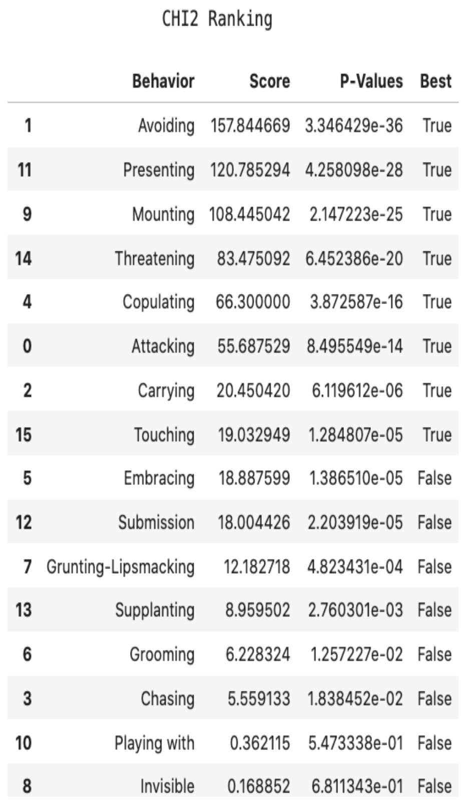
For feature selection, the objective is to find the features with the best or highest Chi-Square values. This indicates those features in the predictor set that are more dependent upon the response variable.

Recursive Feature Elimination Cross Validation (RFECV)

The objective of the RFECV algorithm is to iteratively rank features by importance, discard the least key features, and re-fit the model until the desired number of features remains. The RFECV extends the basic RFE algorithm to include the Cross-Validation loops.

Feature Selection Test Results

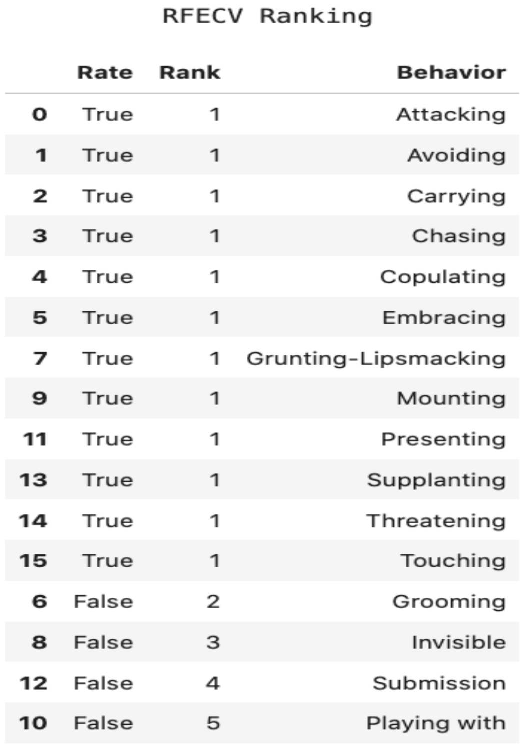
The results from the two feature selection algorithms identified two sets of features. The two sets of algorithms produced a common set of features. Although the Chi Square selected set has 8 features, compared to the RFECV results with 12 features, all features on the Chi square list were on the REFCV selected features. Figure 4 Chi2 Features and Figure 5 RFECV Features show the common feature sets.



CHI2 Ranking

	Behavior	Score	P-Values	Best
1	Avoiding	157.844669	3.346429e-36	True
11	Presenting	120.785294	4.258098e-28	True
9	Mounting	108.445042	2.147223e-25	True
14	Threatening	83.475092	6.452386e-20	True
4	Copulating	66.300000	3.872587e-16	True
0	Attacking	55.687529	8.495549e-14	True
2	Carrying	20.450420	6.119612e-06	True
15	Touching	19.032949	1.284807e-05	True
5	Embracing	18.887599	1.386510e-05	False
12	Submission	18.004426	2.203919e-05	False
7	Grunting-Lipsmacking	12.182718	4.823431e-04	False
13	Supplanting	8.959502	2.760301e-03	False
6	Grooming	6.228324	1.257227e-02	False
3	Chasing	5.559133	1.838452e-02	False
10	Playing with	0.362115	5.473338e-01	False
8	Invisible	0.168852	6.811343e-01	False

Figure 4 Chi2 Features



RFECV Ranking

	Rate	Rank	Behavior
0	True	1	Attacking
1	True	1	Avoiding
2	True	1	Carrying
3	True	1	Chasing
4	True	1	Copulating
5	True	1	Embracing
7	True	1	Grunting-Lipsmacking
9	True	1	Mounting
11	True	1	Presenting
13	True	1	Supplanting
14	True	1	Threatening
15	True	1	Touching
6	False	2	Grooming
8	False	3	Invisible
12	False	4	Submission
10	False	5	Playing with

Figure 5 RFECV Features

Part 3 Model Selection and Evaluation

Part 3 of the Case Study includes selecting machine learning algorithms, verifying the accuracy of the models, and conducting hyperparameter tuning. The objectives are to ascertain if the chosen models can predict the Sex of the Actors and predict the dominant male within the group based upon the Rank of the subject.

Model Selection and Execution

Prediction 1.

The Sex prediction is a classification problem with categorical variables. The models chosen for the test are the Logistics Regression, Support Vector Classification using the RBF kernel, and Random Forest Classification. The data input consists of three datasets split into training and test subsets. The datasets are:

- Dataset with the original 15 behavior features
- Dataset with the features identified by the RFECV feature selection algorithm
- Dataset with the features identified by the Chi Square feature selection algorithm.

Prior to executing the SVC and Rand Forest models a hyperparameter module was run to select the best parameters. The SVC model uses Scikit-Learn GridSearchCV module and Scikit-Learn HalvingRandomSearchCV module for the Random Forest Classifier. Each model test cycle began with using the default parameters for the Logistics Regression model and the best parameters for the other two models. Additionally, the initial run used the full data set for each model type. For subsequent iterations of all models, a brute-force method was used to modify model parameters until the next run resulted in a score equal to lower than the previous score. This process was repeated using the datasets consisting of the features identified by the Chi2 and the RFECV tests.

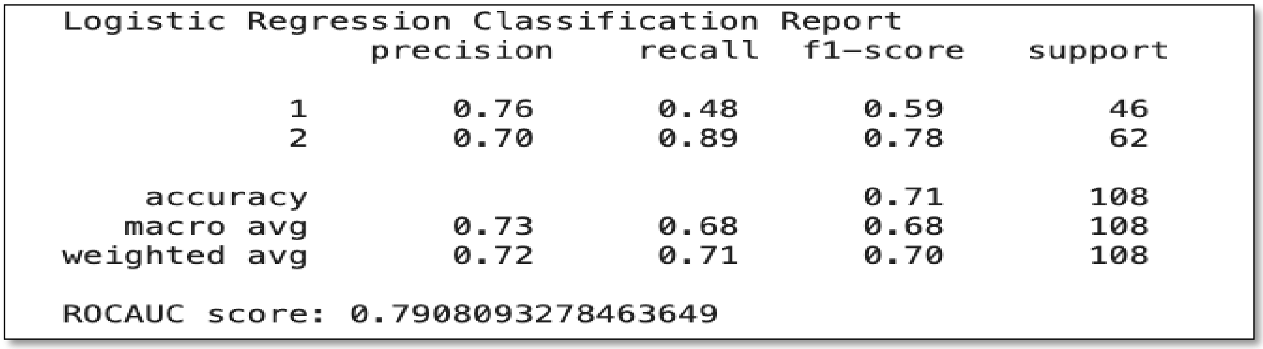
Model Evaluation

Model Evaluation consists of creating and displaying a Confusion Matrix and displaying the Classification Reports. The Confusion Matrix plot is only plotted for the best solution. The Classification Report is displayed for each model iteration and is used to compare results.

Results

Looking at the Classification Report as shown in Figure 6 Logistics Regression Classification Report; the Logistics Regression Classifier has an accuracy rate of 71 percent. This model shows an f1-score score of .59 for Class 1 (Male) and .78 for Class 2 (Female). These values reflect the imbalance in the classes. Out of 108 samples, 52% percent are Class 2 and 46% are Class 1. Because the sample size is so small, even though the percent values are close, the imbalance is reflected in the actual precision and recall values. The Confusion Matrix shown in Figure 7 Logistics

Regression Classification Confusion Matrix shows the break-down of the predictions. Out of 108 samples, there are 22 True Positives and 55 True Negatives. On the negative side, there are 24 False Negatives and 7 False Positives.



```
Logistic Regression Classification Report
              precision    recall  f1-score   support

           1       0.76      0.48      0.59        46
           2       0.70      0.89      0.78        62

    accuracy                           0.71       108
   macro avg       0.73      0.68      0.68       108
weighted avg       0.72      0.71      0.70       108

ROCAUC score: 0.7908093278463649
```

Figure 6 Logistics Regression Classification Report

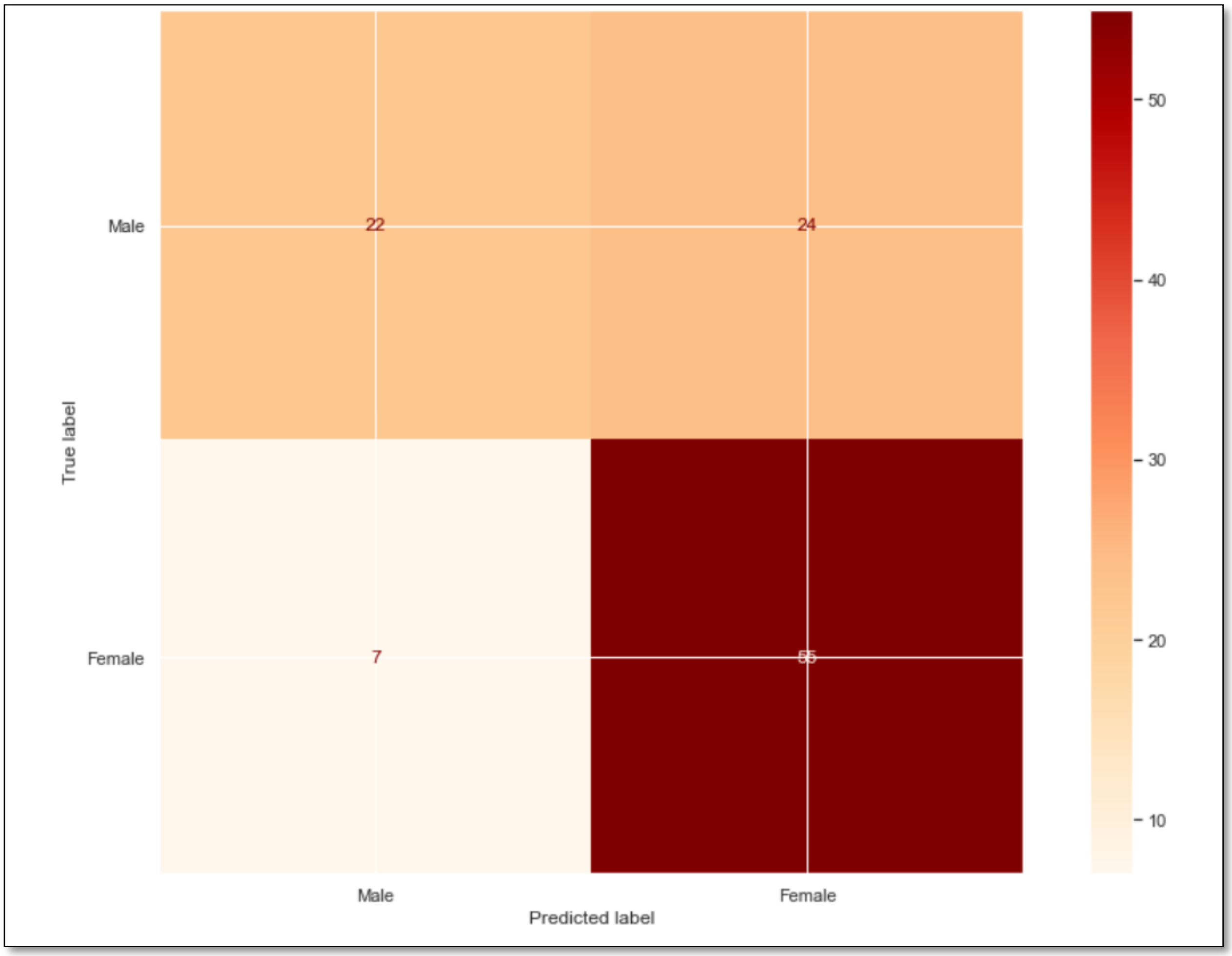


Figure 7 Logistics Regression Classification Confusion Matrix

Model Selection and Execution

Prediction 2.

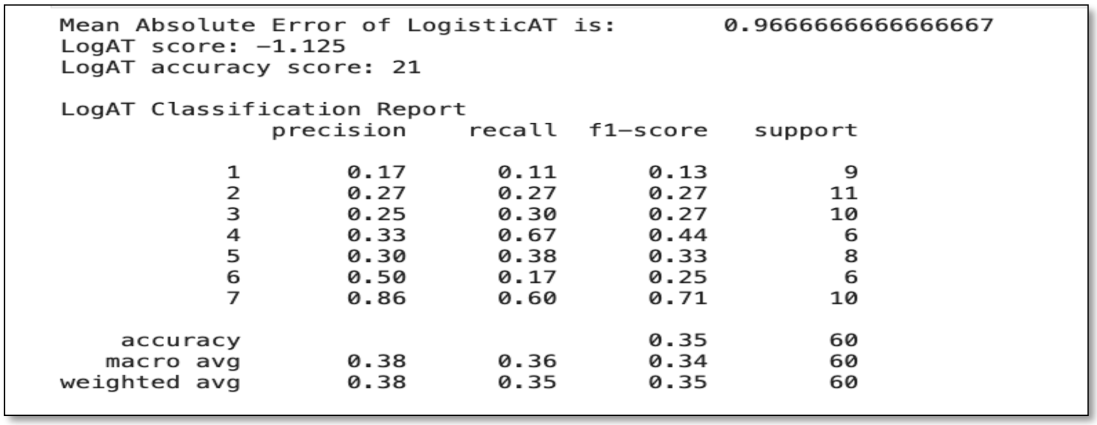
The Dominant Male test uses the Ordinal Regression Classification. The algorithm for this model is from the MORD library. This library contains algorithms that support several types of Ordinal Regression Models. The ones used for the test are the LogisticsIT, LogisticsAT, and Ordinal Ridge algorithms. For this prediction test, the column Rank was added to the Full Dataset. Each male subject was assigned a numerical rank ranging from 1-7. For this test, each model was run with the Full Dataset and then with the RFECV and Chi Square feature sets in similar fashion to the previous prediction test. This set of algorithms does not have any additional parameters with which to conduct hyperparameter modification tests.

Model Evaluation

Model Evaluation consists of creating and displaying a Confusion Matrix and displaying the Classification Reports. These are used for comparison and evaluation of the model's effectiveness.

Results

The accuracy shown by the Classification Report indicates that the three models have a prediction capability of less than 36 percent. Yet looking at the individual Precision/Recall/F1 scores, one or two of the results show an F1 score close to 71. An example is shown in Figure 8 LogisticsAT Classification Report and Metrics. Looking at the results of this model run, the model did identify correctly the highest-ranking individual which is #7. Upon further research regarding the interpretation of Ordinal Regression Classification results, one article suggested that this classification be thought of as multiple Binary Classifiers and to conduct Ordinal Regression through several Binary Classifiers. In consideration of this, each model's test and prediction sets were run through the Scikit-Learn multilabel_confusion_matrix module. This module treats each class as a binary classifier under a one-vs-rest transformation. The results are drastically different than those reported by the model's metrics. Unfortunately, there is little documentation regarding the algorithms used by the MORD models.



```
Mean Absolute Error of LogisticAT is:       0.9666666666666667
LogAT score: -1.125
LogAT accuracy score: 21

LogAT Classification Report
              precision    recall  f1-score   support

           1       0.17      0.11      0.13         9
           2       0.27      0.27      0.27        11
           3       0.25      0.30      0.27        10
           4       0.33      0.67      0.44         6
           5       0.30      0.38      0.33         8
           6       0.50      0.17      0.25         6
           7       0.86      0.60      0.71        10

    accuracy                           0.35        60
   macro avg       0.38      0.36      0.34        60
weighted avg       0.38      0.35      0.35        60
```

Figure 8 LogisticsAT Classification Report and Metrics

Conclusions

The Logistics Regression Classifier model produced the best results for Prediction Test 1, albeit at 71%. A contributing factor to the mediocre results could be that the data did not contain enough variations within the features to serve as a good sample set for training. The dataset may contain a poor balance of bias and variance. According to (Singh, 2018), bias is the difference between the average prediction of the model and the correct prediction value. Those models with high bias tend to result in high error rates on both the training and test datasets. Whereas variance shows the spread of the or variability within the dataset. Models with high variance tend to perform very well on training data but have high error rates on test data.

Another symptom of an imbalance of bias and variance is Underfitting and Overfitting. Underfitting is a symptom of the model not being able to capture the underlying pattern in the data. This occurs when the model has high bias and low variance. A cause for this is not having enough data with which to build an accurate model. Overfitting occurs when the model captures the noise along with the underlying pattern in data. These models have low bias and high variance.

References

Gelardi, V., Godard, J., Paleressompoulle, D., Claidiere, N., & Barrat, A. (2020). Measuring social networks in primates: wearable sensors versus direct observations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *476*(2236), 20190737. https://doi.org/10.1098/rspa.2019.0737

SocioPatterns. (2020, December 4). Baboons’ interactions « SocioPatterns.org. (C) 2008-2011 SocioPatterns.Org. http://www.sociopatterns.org/datasets/baboons-interactions/

Singh, S. (2018, October 9). Understanding the Bias-Variance Tradeoff - Towards Data Science. Medium. https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229