STA 141A: Course Project Report

Ruhi Aggarwal, 920704800

Abstract

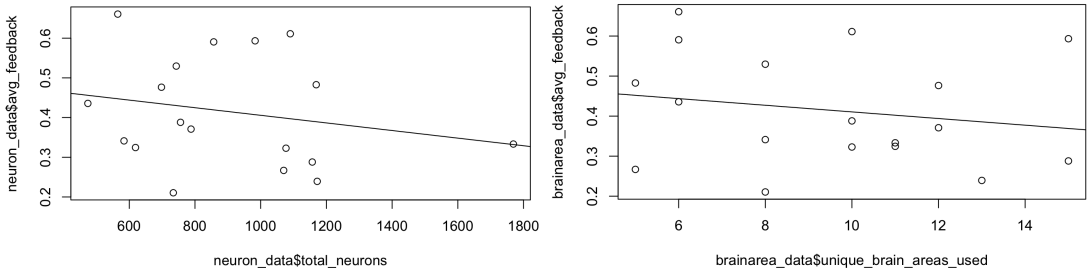
In this project, we explore a subset of data from a study by Steinmetz et al. (2019) focused on the decision-making process of mice when presented with visual stimuli of varying contrasts. The experiment involves trials that challenge the mice to choose between two visual stimuli. I completed initial exploratory data analysis to understand the dataset's structure, including the various attributes and relationships. Despite initial weak correlations between neuron count, unique brain areas, and success rate, further analysis proved certain predictors to be significant such as contrast differences and average spike rates. Additionally, visualizing neural spikes across sessions highlighted the heterogeneity in neurons across different brain areas and sessions. Logistic regression was chosen for predictive modeling since the outcome variable was binary. The model selected average spikes and specific contrast differences as significant predictors of success, aligning with the EDA findings. The final model demonstrated an accuracy of 71%, suggesting that the chosen features are significant predictors of success rate despite the variability across sessions. The following sections outline the procedures used to create the final predictive model in detail.

Introduction

For this project, I analyzed data collected by Steinmetz et al. (2019) regarding a study done with mice. In the study, a total of 10 mice were experimented on over 39 different sessions. In this project, only a subset of that data was used: results from 4 mice of 18 different sessions. Each session contained hundreds of trials. Each trial consisted of two sides of visual stimuli shown to the mice, with varying contrasts. The contrast levels have values in {0, 0.25, 0.5, 1}. The task for the mice was to turn a wheel in the direction of the visual stimuli with lesser contrast. If the mouse identified the correct stimuli, it resulted in a 1 (success) and if not, it resulted in a -1 (failure). If both contrasts are 0, a success occurs if the wheel is not turned, and if they both are non-zero but equal, the correct answer will be randomly chosen. The results of each trial are reflected in the “feedback_type” attribute of the trial. Along with the outcomes, the neural activity of the mice was recorded. Each trial has an associated matrix of neural spike data, taken at 40 different time intervals within 0.4 seconds of the visuals being shown. Additionally, each neuron recorded has an associated brain area. The task was to understand, explore, manipulate, and model the data to produce a method of predicting whether a trial will be a success or a failure.

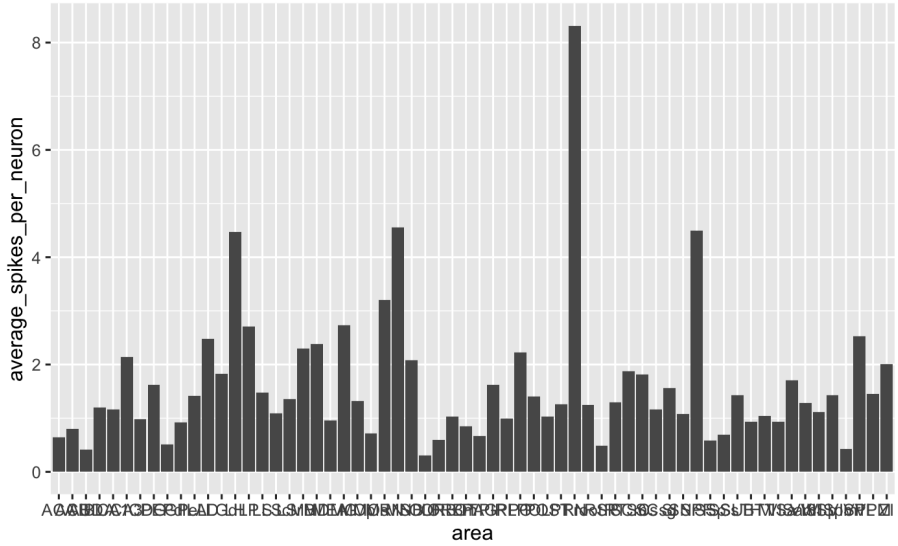
Exploratory Data Analysis

I began by attempting to understand the way the data was structured. Each of the 18 sessions contains hundreds of trials, which each contain 5 attributes: feedback_type, contrast_left, contrast_right, time, spks, and brain_area. After experimenting with printing certain sessions and attributes, I was able to understand that the "spks" attribute of the trials was a matrix of neural spikes, with each neuron represented by a row and each column representing a time interval. Additionally, each row (neuron) in the matrix had a designated brain area attached to it. I first began with very elementary analysis, understanding whether there was a correlation between the total neurons in each session and the average feedback for that session as well as the number of unique brain areas used and the average feedback in a session. The graph of these two relationships with regression lines is shown below.



Although it looks like there is a potential correlation in both figures, the p-values for the regression model between the two was very high: 0.38 and 0.44 respectively.

Next, I decided to explore the specific spike data, and created a data table that grouped by brain area and summarized the average spike per neuron in that brain area. Using this tibble, I created a barplot to visualize if there was a set of brain areas that significantly stood out.



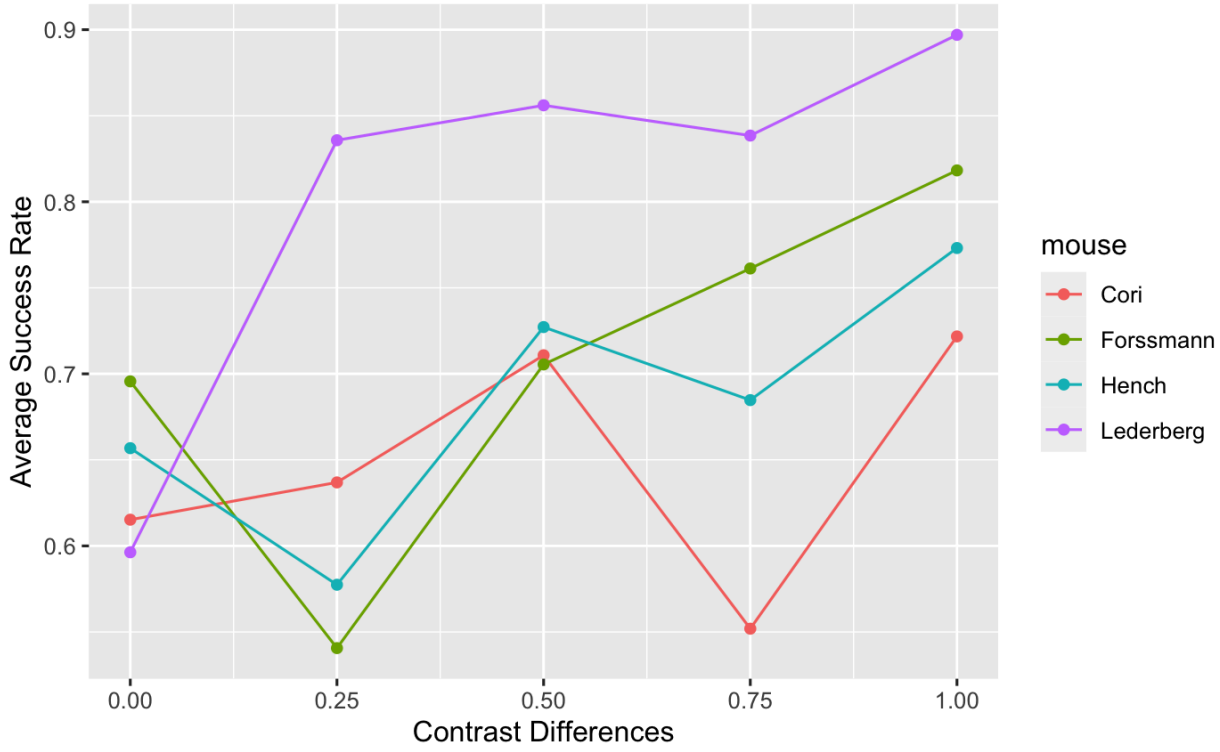
In the plot above, you can see that there is one brain area that has a significantly higher spike per neuron rate than the other areas. After sorting the data table, I found that it was RN. There

are three other brain areas with higher values of above 4 which were MS, SPF, and LH. The rest of the 58 brain areas didn't seem very significant in terms of their average spikes per neuron.

After exploring brain areas and spike rates, I decided to explore whether the left and right contrasts had an effect on the success rates of the mice. I started by creating a pivot table with the mouse name and the absolute differences in left and right contrasts.

mouse <chr>	0 <dbl>	0.25 <dbl>	0.5 <dbl>	0.75 <dbl>	1 <dbl>
Cori	0.6152397	0.6369048	0.7108025	0.5519048	0.7217442
Forssmann	0.6956295	0.5406204	0.7054660	0.7612237	0.8182299
Hench	0.6568078	0.5774370	0.7271879	0.6847389	0.7731267
Lederberg	0.5963047	0.8357568	0.8560565	0.8384680	0.8969955

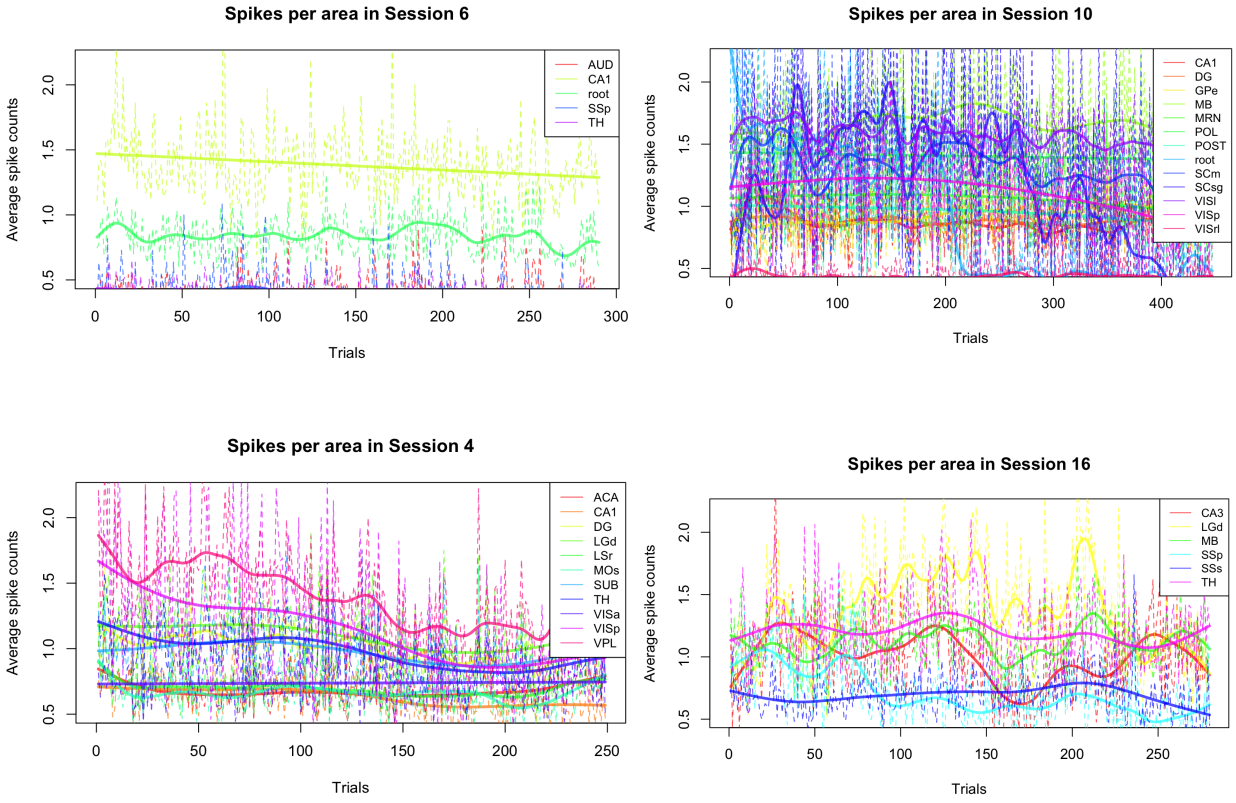
The values of the pivot table represent the success rate of that mouse within all the trials that had the contrast difference in the column label. For example, out of all the trials that Cori participated in where the contrast difference was 0.25, his success rate was 61.5%. Just by scanning over the values, it seems like greater contrast differences result in a higher success rate. To see this relationship more clearly, I graphed the data on a line graph below:



Here we can see that each mouse followed a similar pattern. All the mice had a drop in success rate at either 0.25 or 0.75 difference in contrast. Apart from that, it seems like there is a mostly positive correlation between contrast differences and average success rate. Since when there is a difference of 0, the success/failure outcome is randomly determined, the dip in success rate at 0.25 can be disregarded, and we can examine the graph starting at just 0.25. Given the random

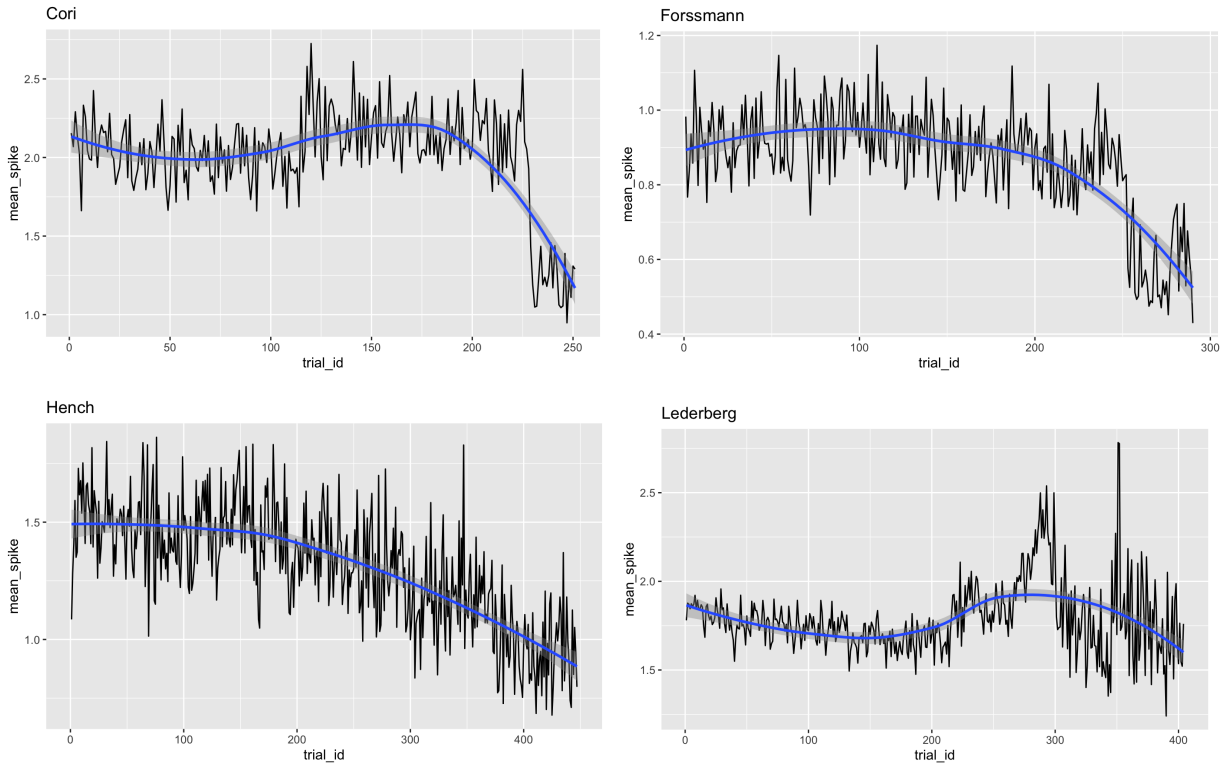
selection of success/failure at x=0, it makes sense that the success rates are centered around 0.5 when the contrast difference is 0. I then conducted a two way ANOVA test on this data and received a p-value of 0.0375 for the mouse variable and 0.0079 for the contrast difference variable. This analysis helped to confirm that contrast differences might be a significant variable in predicting success rates across all mice, since from this analysis, I was able to conclude that there is homogeneity in the behavior of the mice.

After analyzing the similarities and differences between the behaviors of the mice, I decided to explore whether there are any patterns in neural spikes across each of the sessions. I did this by creating one graph for each of the sessions depicting the average spike counts over the trials in the sessions, segmented by the brain areas used in that session. Four of the session graphs are shown below.



I have chosen the four graphs above to show the variability in the neural spikes across the sessions. For example, sessions 6 and 16 have a smaller set of brain areas used, versus sessions 4 and 10 which involve almost double the number of brain areas. Additionally, each of the brain areas in session 6 have visually distinct spike counts and are relatively stable across the trials. However, sessions 4, 10, and 16 have very overlapping spike counts and do not stay static. The heterogeneity across the different sessions became apparent through this analysis, and it seems that it is due to the different neurons/brain areas in each session.

The final feature I decided to explore was time. I plotted the mean number of spikes in each trial and segmented them by mouse to understand the behavior of each mouse over time.



Looking at the plots above, it seems like there is a common trend among all of the mice: as time goes on (number of trials increase), the average number of neural spikes per trial decreases. This affirms that the mice behave similarly to each other, and that average neural spikes could be a potentially strong indicator of success rate.

Data Integration

After conducting the exploratory analysis, I had an understanding of which variables I felt were most important in predicting the outcome, as well as how to format the data into one table. Given that the sessions varied from each other due to the differences in the neurons recorded in each session, it made the most sense to represent the spike data as averages across each trial. Additionally, since the contrast differences proved to be significant predictors through the ANOVA analysis, I decided to include that as a variable in the table as well. The final columns I chose to include in the data frame were: session_id, trial_id, avg_spikes, mouse, contrast_diff, and feedback as the target variable. The final table had a total of 4,064 rows, each one representing a single trial in a session. Next, I split the data table into a training and a test set to prepare for model building. I did this by taking a random sample of indices and indexing the data table using the sample, with 80% of the data in the training set and 20% in the test set. This process of taking a random sample of indices to split the data also ensured that the rows of

the data were shuffled, since before, the original data table was ordered by sessions and trials within each session. Below are the first 10 rows of the training set of the data frame.

session_id <int>	trial_id <chr>	avg_spikes <dbl>	mouse <chr>	contrast_diff <chr>	feedback <fctr>
9	326	1.2753807	Hench	0	-1
15	371	1.3876178	Lederberg	0.75	1
11	68	2.0536756	Hench	1	1
7	176	1.2003425	Forssmann	0.25	1
7	204	1.3082192	Forssmann	0	-1
8	245	1.1979257	Hench	0.75	-1
7	206	1.7157534	Forssmann	0.5	1
3	174	1.7915994	Cori	0	-1
14	67	1.1507937	Lederberg	0.25	1
16	15	1.0253165	Lederberg	0	1

There is a row for each combination of session and trial with the three main features I chose to include in the analysis portion, as well as the feedback column as the outcome variable. Setting the data up in this way made it much simpler to implement the prediction models in the next section.

Predictive Modeling

Since the sessions significantly varied with each other in terms of neural spike patterns and brain areas, it would not be advisable to only select certain sessions to be part of the training data. This would cause the model to be overfit to only the trends in those specific sessions and would not perform well on the test set. I also considered creating different prediction models for each session, however since the goal is to create a predictor that can perform on unseen data, these models would be obsolete since they will only be fit on those specific sessions. Because of this, I chose to include all sessions and trials in one data table to be then used to fit a model.

I decided to start with a logistic regression model to predict the success rate. Since the outcome variable is binary, I thought the model would be a good fit for this data. Below are the summary statistics from the regression model.

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.03160    0.11023  -0.287    0.774
avg_spikes         0.43133    0.07572   5.696 1.23e-08 ***
contrast_diff0.25  0.14052    0.10540   1.333    0.182
contrast_diff0.5   0.69948    0.10158   6.886 5.74e-12 ***
contrast_diff0.75  0.50448    0.11122   4.536 5.74e-06 ***
contrast_diff1     0.81494    0.11049   7.376 1.64e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4886.1  on 4063  degrees of freedom
Residual deviance: 4743.1  on 4058  degrees of freedom
AIC: 4755.1
```

The results of this logistic regression model suggest that avg_spikes, contrast_diff0.5, contrast_diff0.75, contrast_diff1 are significant predictors of feedback. This aligns with the findings in the graph from EDA because the success rate at 0.25 was the lowest in the graphs. There is a positive association between average spikes and feedback because for each one-unit increase in avg_spikes, the log-odds of feedback being 1 increases by 0.43133, indicating a positive association. The p-values of the significant variables are also very low, which are strong enough to indicate that the correlation is not due to random chance. After fitting the model, I used it to predict values for the test data, and achieved an accuracy of 0.706. This accuracy was surprisingly high given the variation among the sessions, and the model was not overly complex because there are only 4 significant variables.

The second model I tried was k-nearest neighbors with cross validation, using average spikes and contrast differences as predictors. After fitting the model on the test data I made a confusion matrix and produced that accuracy report as shown below.

```
Confusion Matrix and Statistics

          Reference
Prediction  -1   1
        -1  32  35
        1  267 683

               Accuracy : 0.703
                 95% CI : (0.6739, 0.731)
    No Information Rate : 0.706
    P-Value [Acc > NIR] : 0.5969

                  Kappa : 0.0753

 Mcnemar's Test P-Value : <2e-16
```

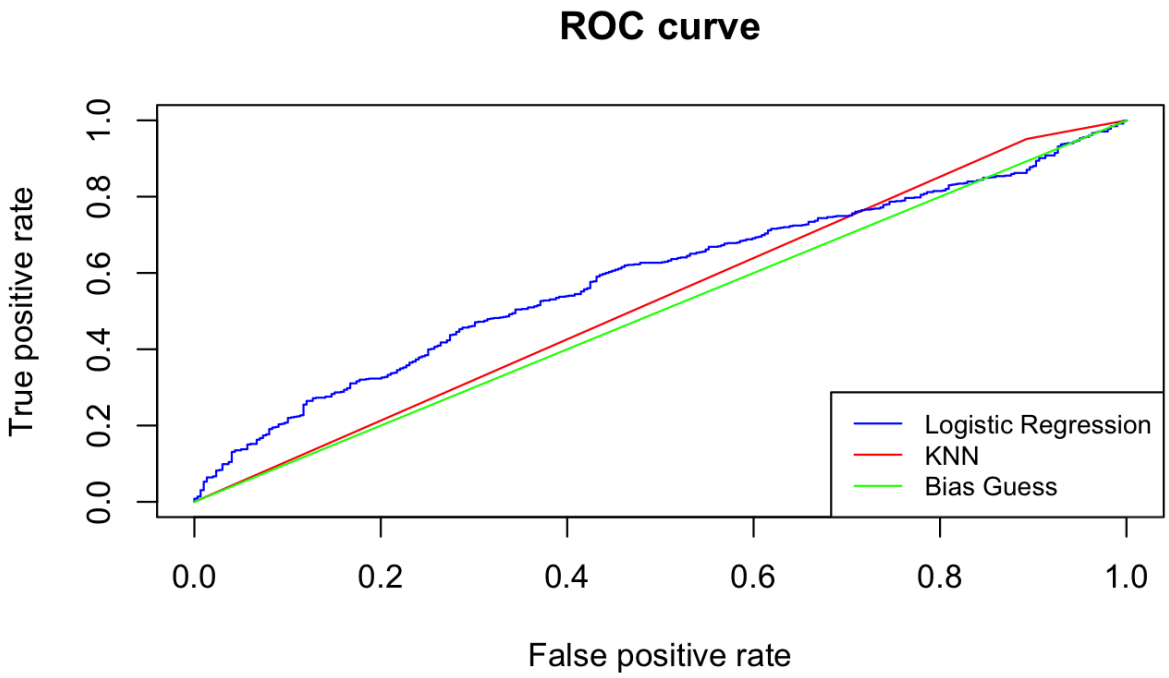
```
Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.10702
            Specificity : 0.95125
         Pos Pred Value : 0.47761
         Neg Pred Value : 0.71895
             Prevalence : 0.29400
         Detection Rate : 0.03147
   Detection Prevalence : 0.06588
      Balanced Accuracy : 0.52914

       'Positive' Class : -1
```

The accuracy from the KNN model was 0.703, which is slightly lower than the accuracy from the logistic regression model.

Lastly, to compare the two models, I plotted an ROC curve of the logistic regression and KNN models as well as the bias. The plot and AUC scores of the three are shown below.




```
"Log AUC: "         "0.586211233359108"
"KNN AUC: "         "0.529138446632694"
"Bias AUC: " "0.5"
```

The AUC value for the logistic regression is higher than the AUC for KNN, validating that logistic regression was the strongest model to predict the feedback type.

Prediction Performance on the Test Sets

After receiving the test sets, I applied the predict() function with the logistic regression model. After calculating the predicted outcome classes and comparing them to the actual outcome classes, I decided to use accuracy as the metric to evaluate the fit. I calculated the accuracy by summing up the number of predicted classes that matched the actual classes, then dividing by the total number of observations in the test set. I was able to achieve an accuracy of 0.625. Although this is lower than what was observed using the original data, it is still a decently high accuracy for a simple model like logistic regression.

I then tested the KNN model on the new test sets as well and received an accuracy of 0.641. Although during training the logistic regression model was performing slightly better, the KNN model performed better on the brand new data.

Discussion

Given that both of the models tested performed almost identically, I think it is safe to say that either model could be used as a predictor for feedback_type. Using the test data results as a metric for choosing one final model, I would say that KNN was the best model. It achieved an accuracy of 64.1% versus 62.5% in the logistic regression model. So, while the logistic regression slightly outperformed KNN in training by 0.03 in accuracy, the KNN model showed better results on unseen test data, proving its robustness in performing across varied datasets; especially because the test dataset only had data from sessions 1 and 18.

There were some limitations with this study due to the limited sessions we had access to as train data. To further improve this predictive model, it would be beneficial to train it on a wider range of data with more trials from more sessions, as well as more predictors that could increase the accuracy of the models. With an improved model, this study can help researchers further understand neural activity patterns in mice and maybe even extrapolate the findings to other species.