Finding the best generalized ML model to predict recall using scalp EEG data

Grace Fujinaga and Ziya Xu

Abstract

Thus far, much emphasis has been placed on developing subject specific machine learning models to predict recall. However, a successful generalized model for predicting recall could prove key for the development of brain stimulation devices. We have demonstrated that such a generalized model trained on data from the power spectral analysis of scalp EEG electrodes is possible, and the best performing models are logistic regression and Gaussian naive Bayes. However, model performance was limited by the low dimensionality of the dataset as many electrode recordings were removed to ensure consistent features across subjects. Future experiments that emphasize consistent electrode placements could result in improved performance of general recall prediction models.

Introduction

The ability to accurately predict recall during the encoding phase of free recall experiments has significant implications for understanding the mechanisms of memory and developing technologies that could enhance memory in individuals with memory impairments.

There is a variety of research that uses spectral powers to predict recall in free recall experiments using both intracranial data and scalp data during the encoding phase of a free recall experiment. The current studies explored a variety of ML models including logistic regression using L1 and L2 penalization, support vector machines approach (SVM), and recurrent neural networks (Kahana et al., Arora et al. 2018). The standard method used was a leave one list out cross validation scheme using z scored features. When using EEG data, the L2 penalization scheme resulted in the best classifier compared to L1 penalization, SVM, rf, xgboost (Kahana et al). When iEEG data was used, the SVM classifier outperformed logistic regression (Arora et al. 2018).

There has also been research that suggests that using the retrieval process has greater specificity than using the encoding model, but this report will focus on the encoding phase of free recall (Kragel et al. 2017). It is also known that certain parts of the brain are more involved in memory than others. Specifically, we see that the lateral and medial temporal cortices and certain hippocampal regions play a considerable role in memory (Kragel et al. 2017).

With this information in mind, there has been a gap in the research that looks at using scalp data to create a generalizable classifier. The studies above used within subject modeling with spectral powers as the features using a leave one list out cross validation scheme. Being able to have a model that predicts recall across multiple subjects has a variety of benefits including advancing brain implant technology and understanding which neural features may or

may not be more important in memory moving forward. In this report, we explore the performance of various machine learning models, including logistic regression with L1 and L2 penalization, K-nearest neighbors, Gaussian naive Bayes, and decision trees, to identify the best generalized model for predicting recall using scalp EEG data. By achieving this, we can advance our understanding of the neural features that are important for memory and pave the way for the development of novel memory-enhancing technologies.

Methods

Data Used

We used the FR1 and CatFR1 datasets which contain scalp EEG readings for free recall and categorical free recall. These datasets allow us to investigate whether a general model for recall prediction is possible without the effects of brain stimulation.

Feature Generation

Since our goal is to create a generalized prediction model, we must train the model across subjects. In order to ensure that features are consistent across subjects, we included a set number of electrodes from each brain region. 48 subjects had at least 16 electrodes on their temporal lobe, 8 electrodes on their frontal lobe, and 2 electrodes on their hippocampal region. For each of those subjects, we performed power spectral analysis for their 26 selected electrodes (16 from the temporal lobe, 8 from the frontal lobe, and 2 from their hippocampal location) and z-scored the powers. Then, we split the 48 total subjects into 24 training subjects and 24 testing subjects and concatenated them into an overall training set and an overall testing set.

Models

In order to find the best model for predicting recall, we selected a variety of models and also used GridSearchCV (a cross validation technique) to find the optimal hyperparameters for those models using the training data. Then, we used the optimized model to predict the testing data. Note that our training and testing datasets are completely independent and contain no overlap, and cross validation validates on folds held out from the training data. Thus, none of the testing data is used to train the model. Below, we will discuss the motivations and methods for each model before reviewing the results.

Logistic Regression

Logistic regression uses a logistic function to classify objects into binary states. It is one of the simplest machine learning algorithms, so the computational overhead is relatively low. This is critical for the development of brain stimulation devices, and logistic regression has proven to be one of the most accurate models for predicting recall (Kahana et al.).

In logistic regression, we are choosing the penalty scheme as well as tuning the regularization strength. The penalty scheme can be L1, L2, or elastic net which is a combination

of them. L1 adds information about the complexity of the feature weights, and it can make some coefficients zero which results in a sparse model. This is less beneficial for our dataset because we have already dropped many of the electrodes. However, L2 will not have any zero coefficients and will likely be the optimal penalty scheme for this dataset. L2 has also been the most accurate model for other recall prediction and brain stimulation models (Kahana et al.). Elastic net is a combination of L1 and L2. The regularization strength affects how much weight is given to the training data compared to the complexity penalty. The goal of choosing the most optimal penalty scheme and regularization strength is to prevent overfitting.

Decision Tree

This model generates a decision tree with internal nodes as conditions, leaves as classifications, and edges and probabilities. Unlike logistic regression, decision trees are not limited to binary classification (although there are only two outcomes in our dataset—recalled and not recalled). For decision trees, we tuned the hyperparameters for the maximum depth of the tree and the minimum number of samples required to be at a leaf node.

K-Nearest Neighbor (KNN)

KNNs classify objects based on the classifications of their k-nearest neighbors. Like decision trees, KNNs are not limited to binary classification. We are using cross validation to determine the optimal value for k: the number of neighbors that will determine the classification of an input. If k is 1, then an input will be assigned the classification of its closest neighbor.

Gaussian Naive Bayes

Naive Bayes uses Bayes' Theorem and assumes independence between features (which is likely not true for our dataset). However, naive Bayes is powerful when the dimensionality of data is high. Gaussian naive Bayes specifically supports continuous data that is assumed to be normally distributed (it gives more weight to inputs that are closer to the mean). In this model, we are tuning the smoothing variable which affects the width of the curve of the distribution and controls the weights of inputs that are farther from the mean. A higher smoothing value increases the width of the curve and increases the weights of inputs that are farther from the mean.

Analysis

For each different model, there were a variety of analyses done. The first is a general analysis that tests the model on all of the test data pooled together. The ROC curve and AUC analyses are conducted. A histogram of the prediction probabilities and a confusion matrix are too. It is worth noting that we were not able to conduct a t-test using this method because we grouped all of the test data and got one resulting ROC curve and AUC value. The second set of analyses addresses this problem by randomly ordering the test data and splitting it into 24 folds. Note that every encoding event was used exactly once and NOT randomly selected from a pool of data. Any reference to random folds refers to this technique. The model was tested on each fold and the resulting ROC curves and AUC values were used to run t-tests. Finally, the third set of analyses were conducted by testing the model on test data for each subject. The subjects

that were tested on were not in the train data set. We tested the model on each subject separately to test if the model could be generalized from the test data to a specific subject. This directly addresses if a model trained on different subjects could correctly classify recall in other subjects.

Logistic Regression

As predicted, the optimal penalty scheme was L2 after performing grid search cross validation. The optimal value of C was 7.743e-12 which gives more weight to the complexity penalty, implying that the data is not as important.

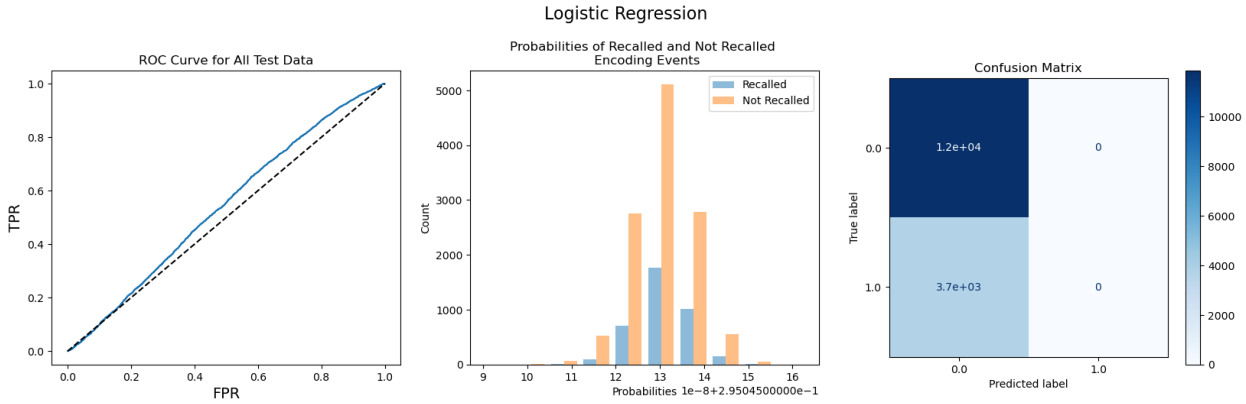

Figure 1. Analysis of Logistic Regression with L2 Penalization on Grouped Test Data

The ROC curve shows that the classifier performs slightly better than chance. The AUC is 0.5393 and the accuracy of the model across all of the test data is 75.98%. Note that here there is no statistical test for significance because all of the test data is grouped together. Statistical tests were conducted using other techniques. It is important to note that because around 75% of the encoded words were not recalled, the model clearly predicts not recalled more than recalled, so AUC is a better measure of performance.

The center histogram displays the predicted probability of recalled and not recalled encoding events. Predicted probabilities close to 1 indicate that the classifier is confident that these events are recalled events while the predicted probabilities close to 0 mean that the classifier is confident that these events are not recalled. In a perfect classifier, we would see separation between recalled and not-recalled events. Not recalled events would be predicted with a probability close to 0 and recalled events would be predicted with a probability close to 1. Clearly that is not the case with both recalled and not-recalled events organized in a similar shape and clustered at the same low predicted probability. This indicates a high level of noise in the data.

The confusion matrix on the right displays the classifiers predictions where 1 indicates a recalled event and 0 indicates a non-recalled event. The matrix clearly shows that the model never predicts a recalled event despite having an AUC value that is a bit above chance. This demonstrates the higher proportion of non-recalled events compared to recalled events in the

data. The F1 score of the model for non-recalled events and recalled events is 0.86 and 0 respectively. The F1 score is the harmonic mean between precision and recall, and measures the model's accuracy. This again shows the skew toward non-recalled events.

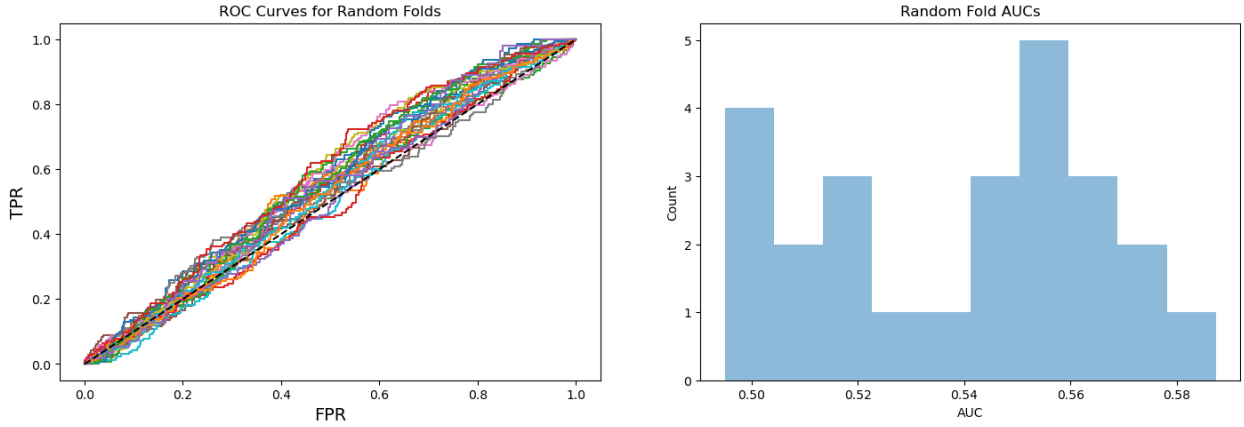

Figure 2. Analysis of Logistic Regression with L2 Penalization on Individual Subjects

The second set of analyses were conducted on random folds of test data as discussed above. The ROC curves are mostly above chance with a few underperforming chance. The histogram on the right shows that most AUC values are above 0.5. We performed a one sample t-test to determine if the model classifies better than chance the p-value was 7.671e-07, which is clearly less than 0.05, and shows that the model classifies random folds of data better than chance.

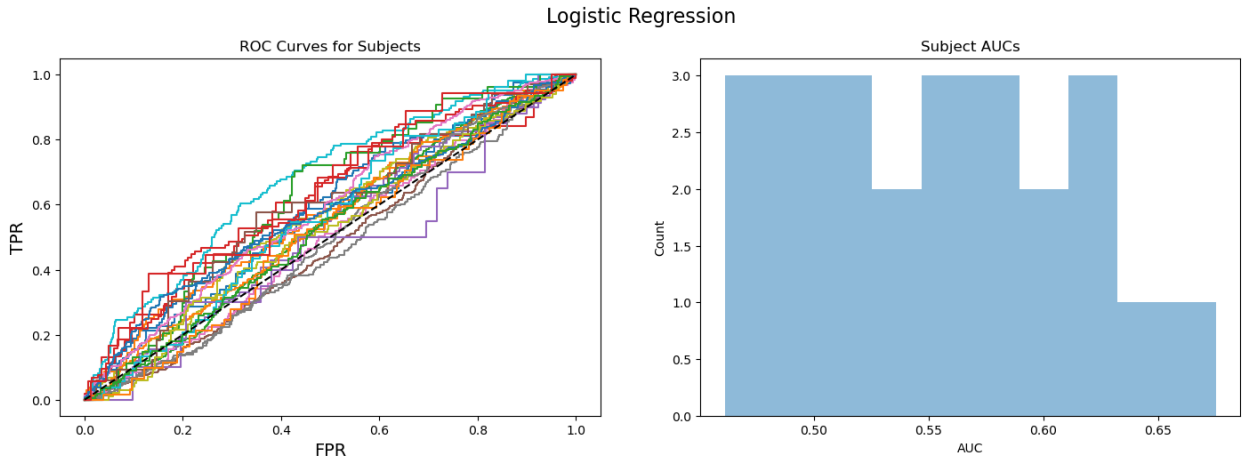

Figure 3. Analysis of Logistic Regression with L2 Penalization on Individual Subjects

Figure 3 displays the results of the model on classifying recalled and not-recalled events for each test subject. As we expected, the model performed better for some subjects than for others. In the histogram on the right, you can see that the majority of the subjects AUC is higher than 0.5. It is also clear that there is a wider range across the AUC values compared to the random folds of data with some being higher than 0.65 and some being less than 0.5. Statistical analyses were also performed across the subjects comparing the subject AUC to chance, 0.5.

The resulting p-value was 7.495 e-05, which is far less than 0.05, which shows that the model predicts recall significantly better than chance.

Decision Tree

The optimal parameters were a maximum depth of 2 and a minimum sample count at each leaf of 1. These parameters make sense because there are only 2 possible classifications, and the majority of classification in the dataset are “not recalled” (around 80%).

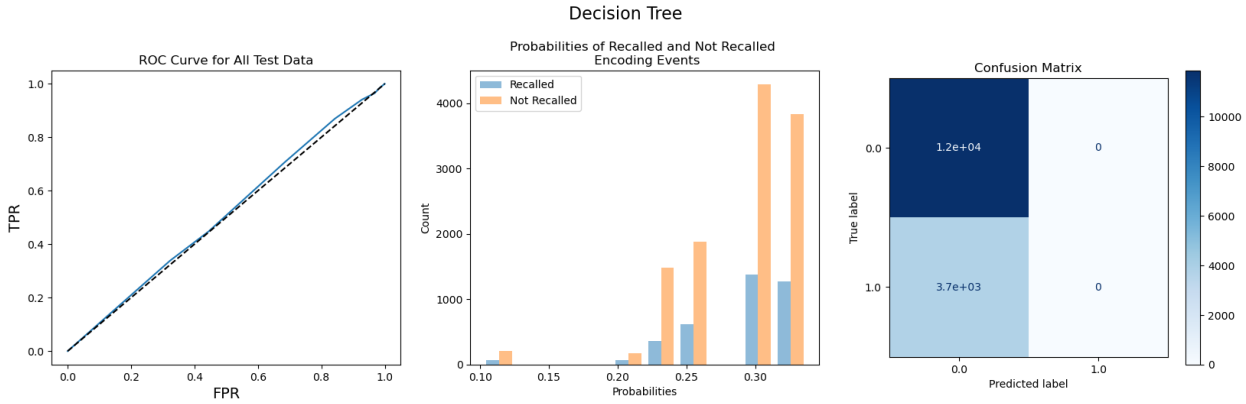

Figure 4. Analysis of Decision Tree on grouped test data

The ROC curve resulting from the decision tree model does not appear to be above chance. The AUC is 0.5123 and the accuracy of the model is 75.98%. This is the same accuracy as the logistic regression model because the decision tree model, like the logistic regression model, also never predicts a recalled event.

The histogram displaying predicted probabilities also does not show a trend of separation. There is not a significant difference in the shape of the distribution of recalled and not recalled prediction probabilities, again indicating high levels of noise across neural features.

The confusion matrix also bears considerable significance to that of logistic regression because it does not predict any recalls. The F1 score for recalled and not-recalled events is 0 and 0.86 respectively.

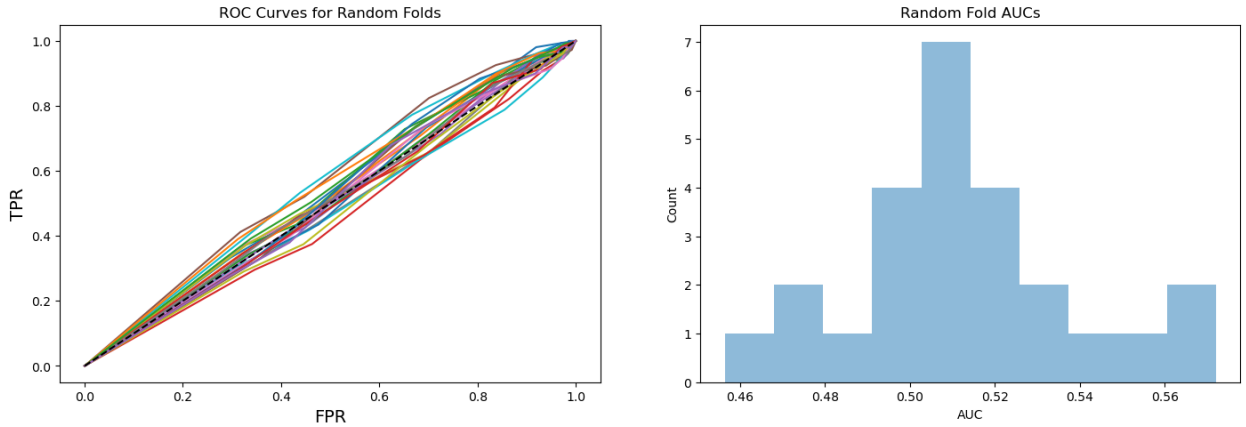

Figure 5. Analysis of Decision Tree model on random folds of test data

The second set of analyses conducted on random folds of test data show that the classifier did not perform consistently better or worse than chance. The ROC curves in figure 5 display this and the histogram of the resulting AUC values confirm that some ROC curves did better than chance with an AUC value of over 0.56 while others did much worse with AUC values less than 0.46. The one sample t-test that was conducted yielded a p-value of 0.03880 While this is less than 0.05, it did not perform as well as logistic regression.

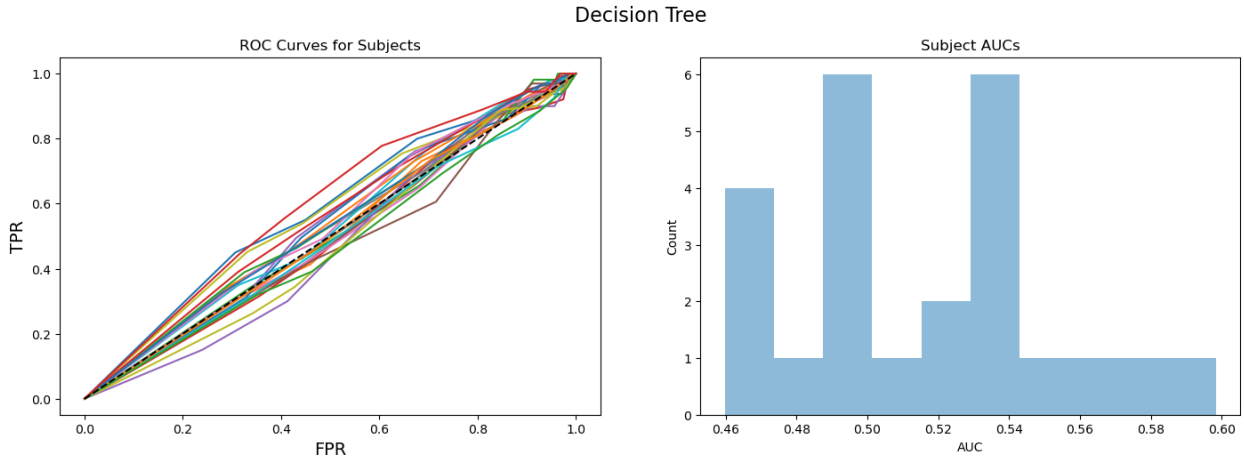

Figure 6. Analysis of Decision Tree model on Individual Subjects

Figure 6 displays the results of the model on classifying recalled and not-recalled events for each test subject. Again, the model performed better for some subjects than for others, with the highest AUC value being just under 0.6. The one sample t-test yielded a p-value of 0.04932. This is less than 0.05 but just barely, and shows that logistic regression performs better.

K-Nearest Neighbors

The optimal number of neighbors was 7, meaning that the classifications of the 7 nearest samples were used to determine the classification of an input.

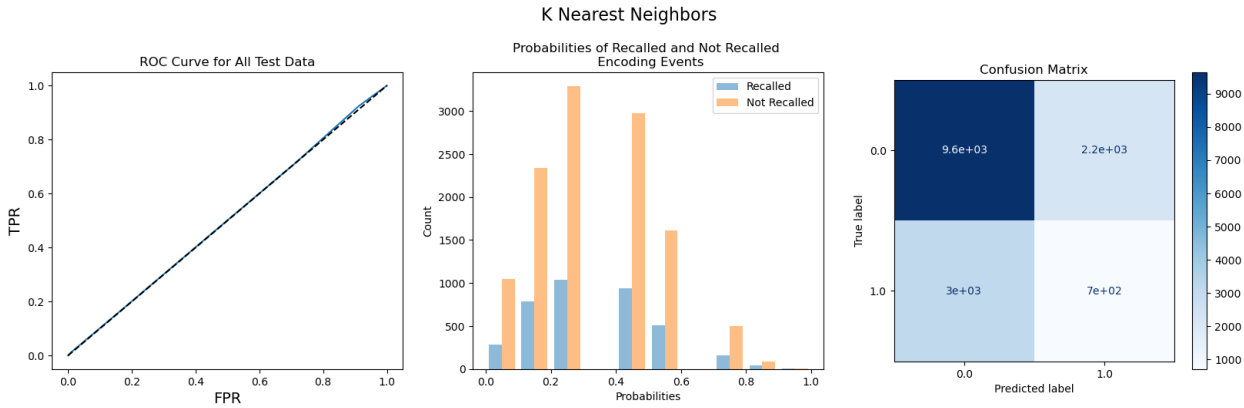

Figure 7. Analysis of KNN model across all test data

The ROC curve resulting from the KNN model does not appear to be above chance. The AUC is 0.5025 and the accuracy is 66.33%. The histogram displaying predicted probabilities also does not show a trend of separation and there is not a significant difference in the shape of distribution between the predicted probabilities for recalled and not-recalled events. The confusion matrix shows that the KNN model does actually predict recalls unlike logistic regression and the decision tree model. The F1 score for recalled and not-recalled events is 0.19 and 0.81 respectively. This shows that the model is much better at predicting not-recalled events, which is consistent with the other models.

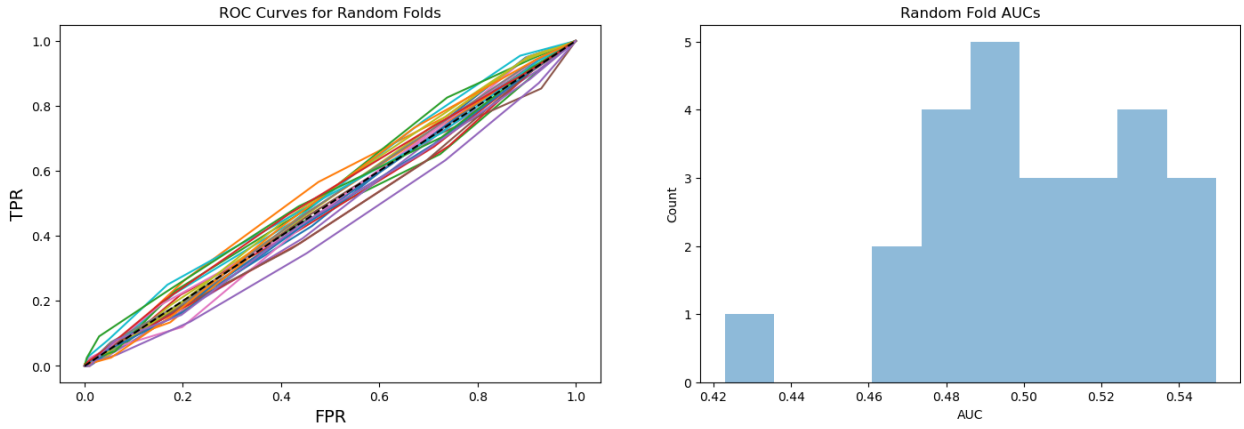

Figure 8. Analysis of K-Nearest Neighbors on randomly folded test data

Figure 8 shows that the ROC curves appear to closely follow chance. The histogram displaying AUC values shows this as well. The one sample t-test yielded a p-value of 0.7618 which confirms what appears to be true. The KNN model does not predict recalled and not-recalled events better than chance.

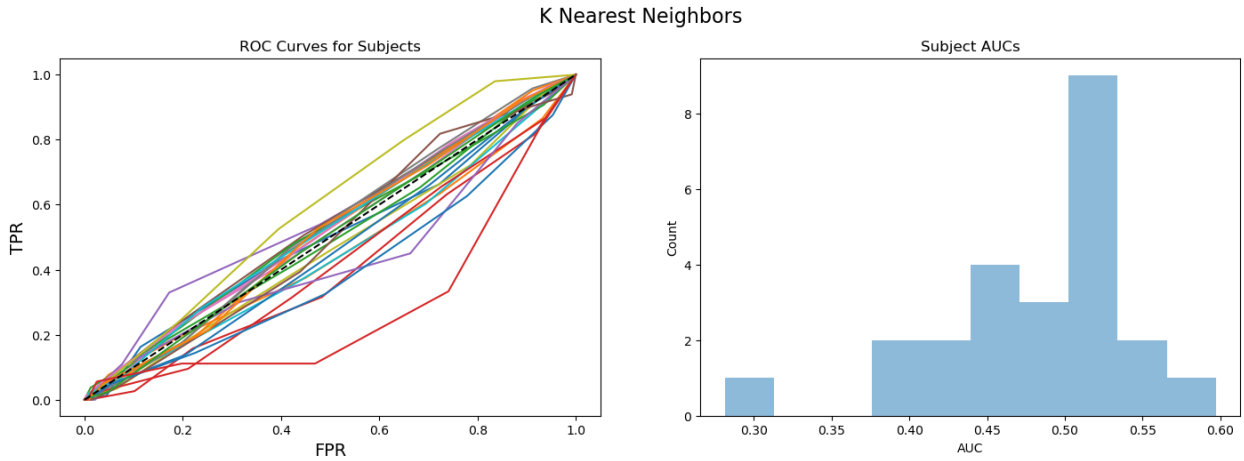

Figure 9. Analysis of K-Nearest Neighbors on individual subjects

Figure 9 shows that the ROC curves vary with some underperforming chance significantly and others doing much better than chance. The AUC histogram also displays this

with one AUC value lower than 0.3 and others greater than 0.55. The one sample t-test yielded a p-value of 0.2413, which is greater than 0.5. This analysis confirms the analysis from the randomly folded data that the KNN model is not the best model to predict recall.

Gaussian Naive Bayes

The optimal smoothing value was 5.857e-4 which is higher than the default value of 1e-9. This means that the width of the curve was more narrow and samples farther from the mean were not given much weight.

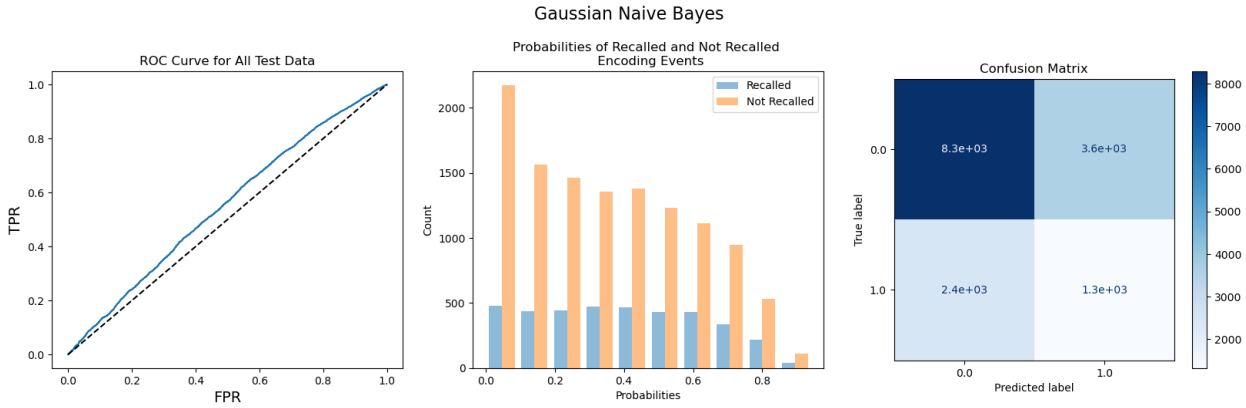

Figure 10. Analysis of a Gaussian Naive Bayes model on all test data

The ROC curve resulting from the Guassian Naive Bayes model appears to perform better than chance at first glance. The AUC is 0.5487 and the accuracy is 61.63%. The histogram displaying predicted probabilities does not show separation. This model did result in predicted recalls in addition to non-recalls as shown by the decision matrix. The F1 score for recalled and not-recalled events is 0.35 and 0.70 respectively. The better F1 score for not-recalled events compared to recalled events is consistent with other models.

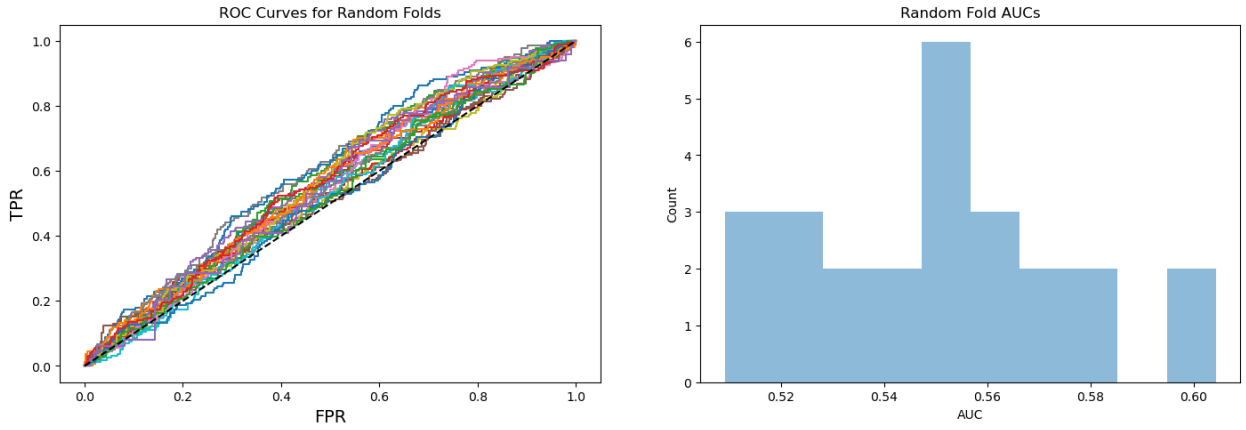

Figure 11. Analysis of Gaussian Naive Bayes on randomly folded test data

Almost all AUC values are above 0.5 and the ROC curves also appear to be above chance. The one sample t-test yielded a p-value 5.645e-10, which is much less than 0.05, which shows that this classifier performs better than chance on random folds of data.

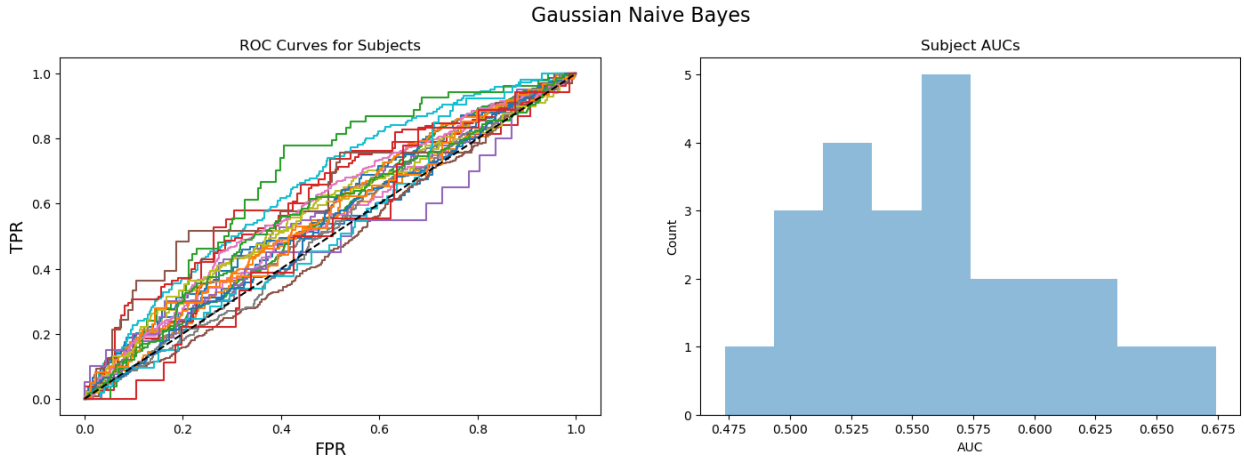

Figure 12. Analysis of Gaussian Naive Bayes on individual subjects

There is greater variability across AUC values as expected for the analyses across subjects. The majority of AUC values are still above 0.5. The one sample t-test yielded a p-value of 2.737e-06 which is much less than 0.05.

Results

Model	**AUC (across all test data)**	**Subject p-value**	**Random fold p-value**
Logistic Regression	0.5392	7.495e-05	7.670e-07
Decision Tree	0.5123	.04932	0.03880
K-Nearest Neighbors	0.5025	.2413	0.7618
Gaussian Naive Bayes	0.5487	2.738e-06	5.646e-10

The table above shows the consolidated results from the four models we used. The decision tree model and KNN model performed no better than chance while logistic regression with L2 penalization and Guassian Naive Bayes both performed better than chance. Overall, across all four models it was clear that there was variation of the performance of the model across subjects. It was also clear that the neural features are noisy as shown in the histograms of predicted probabilities.

Further Investigation

Include More Electrodes

Even though the AUC values for Logistic Regression and Gaussian Naive Bayes were above .5 and statistically significant, all of the models' AUC values were quite low. Given that we had optimized the model parameters, performance improvements could likely only come from improving our dataset features. We initially had 48 subjects and included 16 electrodes from their temporal lobe, 8 from their frontal lobe, and 2 from their hippocampal location. However, given that most subjects have far more than 26 electrodes taking scalp readings during these free recall experiments, our training (and testing) data did not include many potentially key features/EEG readings.

To address this, we wanted to include more electrodes from each brain region. We performed the same investigation on subjects that had at least 24 electrodes on their temporal lobe, 16 on their frontal lobe, and 4 on their hippocampal location. This left us with 14 total subjects, and we used the specified minimum number of electrodes from each region to perform another power spectral analysis. Like before, the powers were z-scored for each subject. 7 of the subjects were concatenated into a training dataset and the other 7 were concatenated into a testing dataset.

All of the other analyses performed were the same as above, although only 10 random folds were created rather than 24 because we were working with a smaller dataset.

Logistic Regression

Like before, the optimal penalty scheme was L2. However, the optimal value of C was 1.292e-4 which was higher than the initial dataset, meaning that this model gave more weight to the training data rather than the complexity penalty.

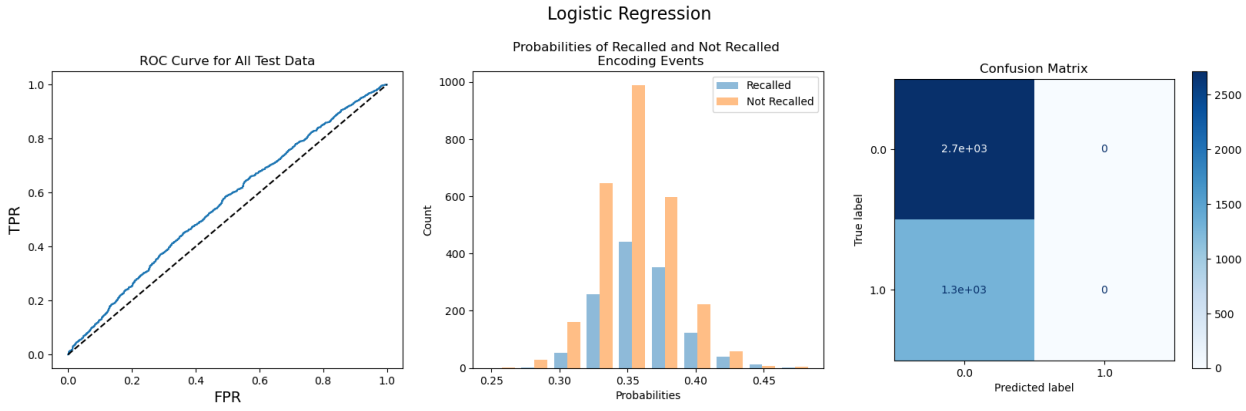

Figure 13. Analysis of Logistic Regression with more electrodes

The AUC for logistic regression was 0.5560, which is above chance. The histogram does not show separation and shows that the model is not confident. The model notably does not predict any recalls. The F1 score for recalled and not-recalled events is 0 and 0.81 respectively.

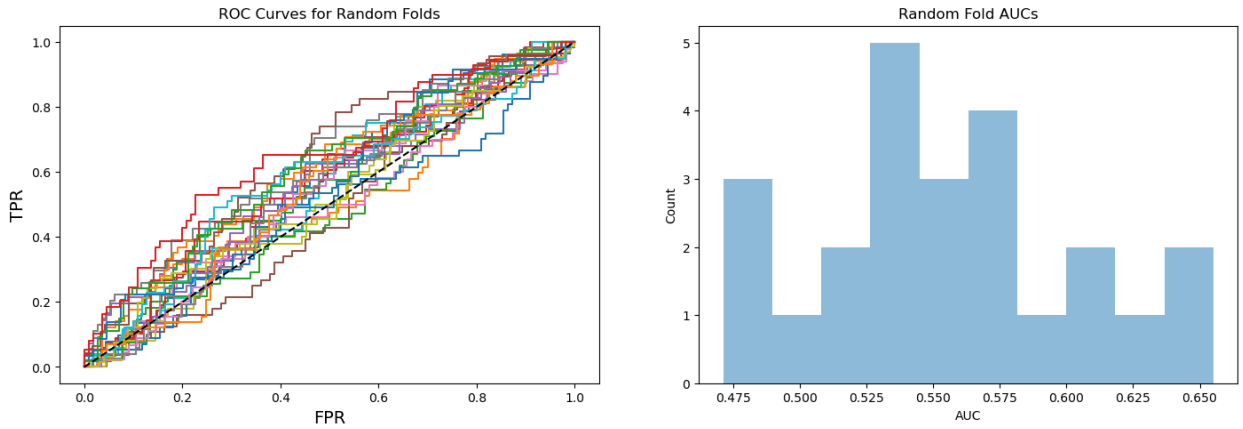

Figure 14. Analysis of Logistic Regression with more electrodes on random folds of test data

The majority of the AUC values are above chance and the one sample t-test yielded a p-value of 9.379e-06, which is considerably less than 0.05 and shows statistical significance.

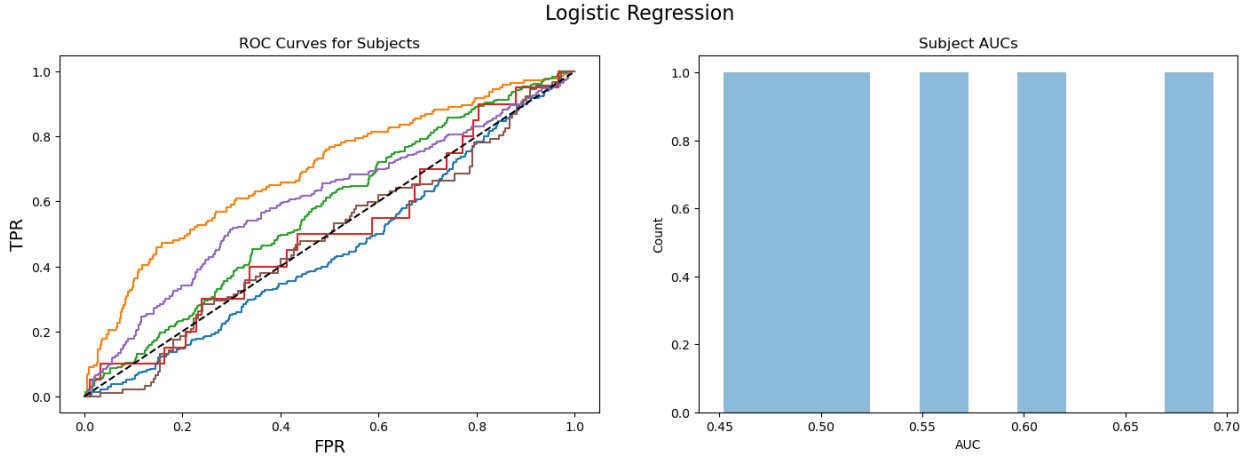

Figure 15. Analysis of Logistic Regression with more electrodes on individual subjects

There were some AUC values that were much higher and close to 0.7 and some AUC values that were close to 0.45. The t-test yielded a p-value of 0.2252, which is not less than 0.05. This does not confirm the results on the random folds of data. This could be because of the higher variability across subjects and the fewer number of subjects involved in this analysis. The model performed very well on a few subjects but also poorly on others.

Decision Tree

Like before, the optimal parameters were a maximum depth of 2 and a minimum sample count at each leaf of 1.

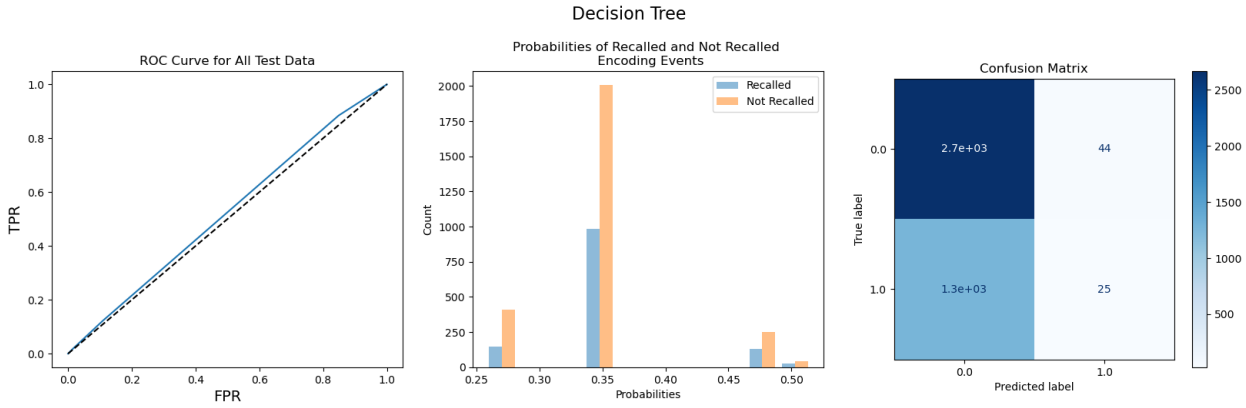

Figure 16. Analysis of decision tree with more electrodes

The AUC for the decision tree model was 0.5216 which is above chance. The histogram does not show separation and also shows that the model is not confident. The model does predict recalls but not well. The F1 score for recalled and not recalled words is 0.04 and 0.80.

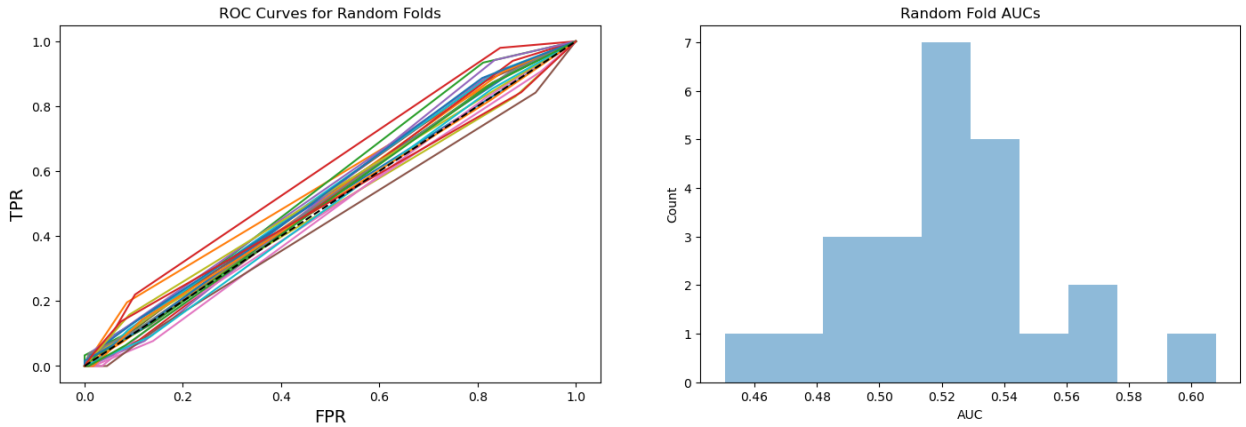

Figure 17. Analysis of decision tree with more electrodes on random folds of data

The t-test on the decision tree model yielded a p-value of 0.003454. This is below 0.05 and is statistically significant.

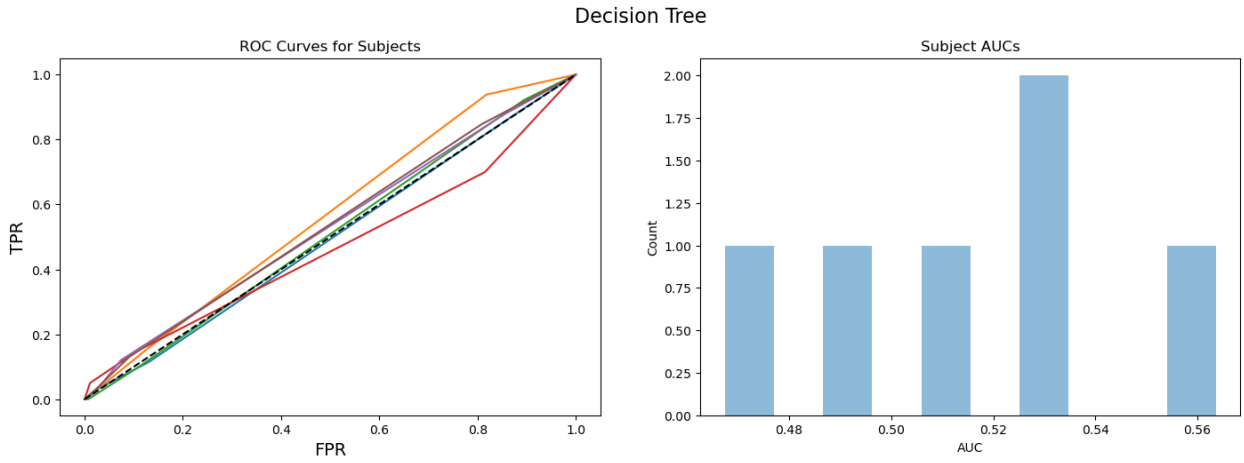

Figure 18. Analysis of decision tree with more electrodes on individual subjects

The t-test on the decision tree model yielded a p-value of 0.2992. This is above 0.05 and not statistically significant and shows the variability across subjects.

KNN

The optimal number of neighbors was 10, meaning that the classifications of the 10 nearest samples were used to determine the classification of an input.
Accuracy: 0.67

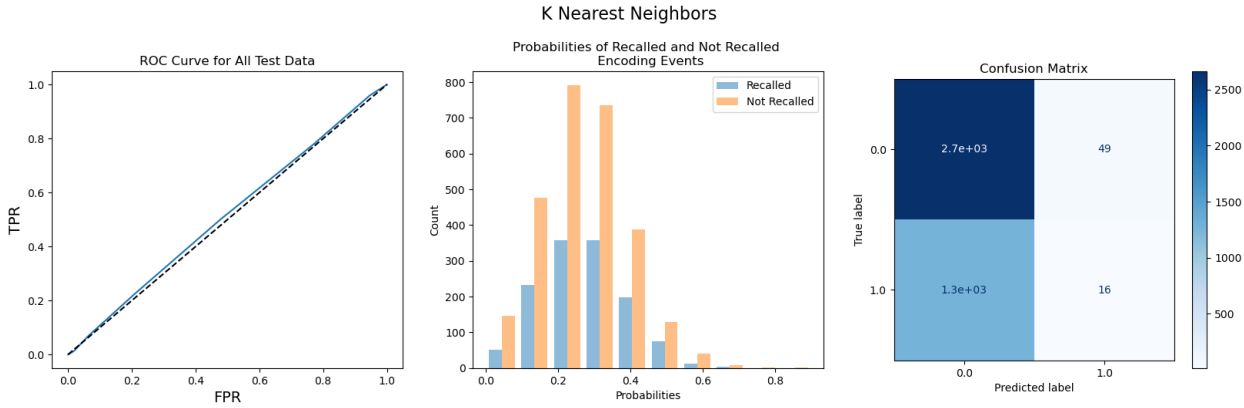

Figure 19. Analysis of KNN with more electrodes

The AUC for KNN was 0.5134 which is above chance. The histogram does not show separation and also shows that the model is not confident. The model does predict recalls but not well. The F1 score for recalled and not recalled words is 0.01 and 0.98.

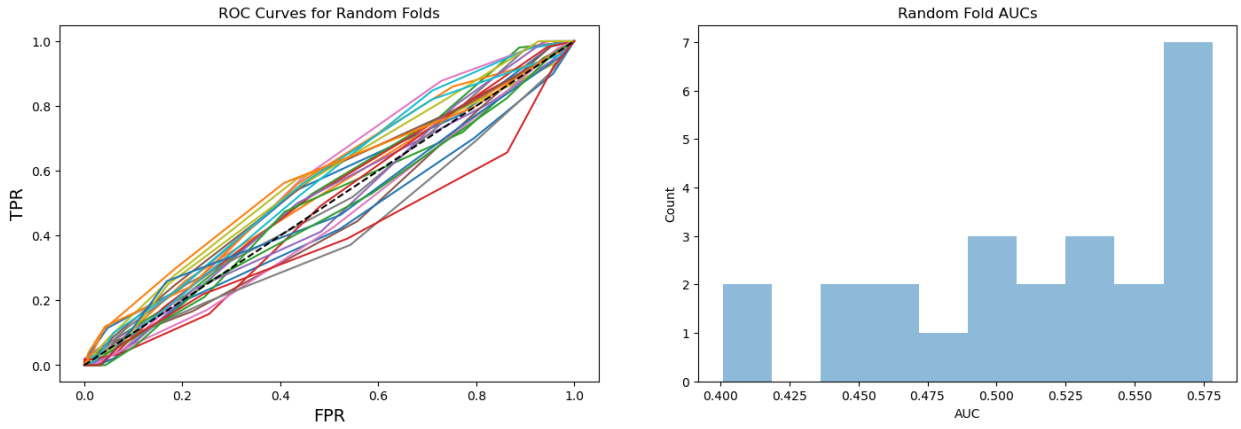

Figure 20. Analysis of KNN with more electrodes on random folds of test data

The p-value was 0.2193, which is greater than 0.05 which shows that the KNN model even using more electrodes does not predict recall better than chance.

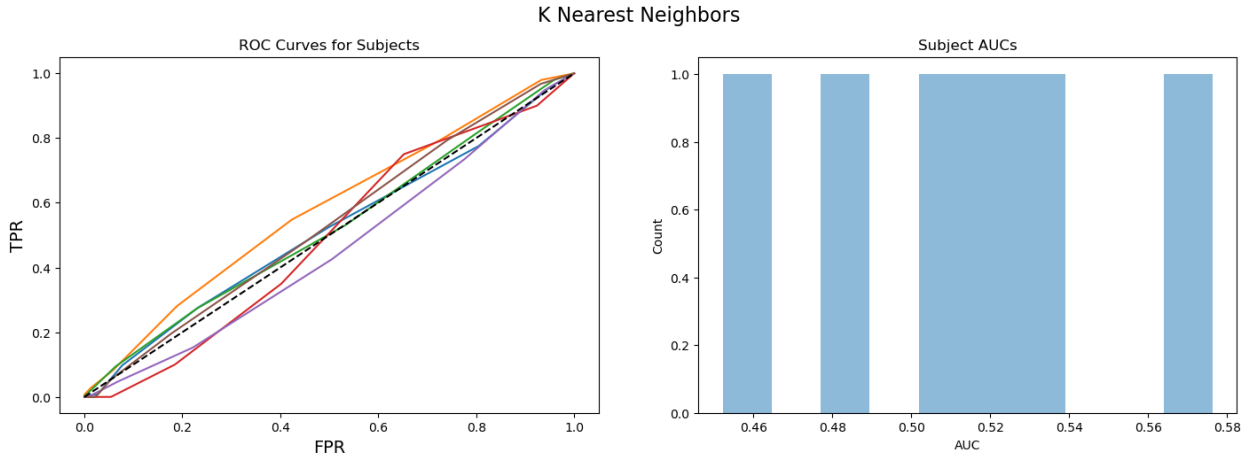

Figure 21. Analysis of KNN with more electrodes on individual subjects

The p-value was 0.4840, which is greater than 0.05 which shows that the KNN model even using more electrodes does not predict recall better than chance for individual subjects.

Gaussian Naive Bayes

The optimal smoothing value was 2.983e-3 which is higher than the default value of 1e-9 and higher than the previous dataset's optimal value of 5.857e-4. This means that the width of the curve was more narrow and samples farther from the mean were not given much weight.

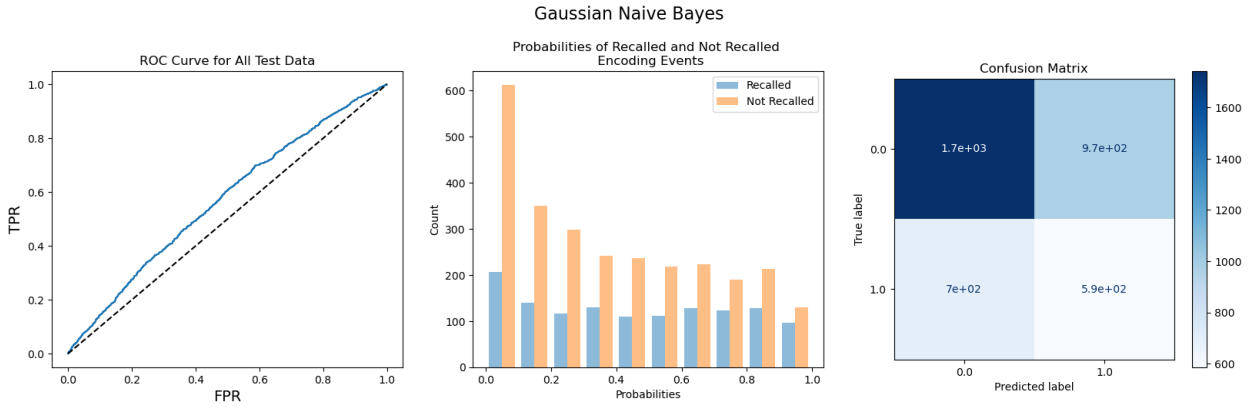

Figure 22. Analysis of Gaussian Naive Bayes with more electrodes

The AUC value is 0.5696 and the histogram of predicted probabilities does not show any separation or confidence. The confusion matrix does show that the model predicts recalls and non-recalls. The F1 score for recalled and not-recalled words is respectively 0.41 and 0.68.

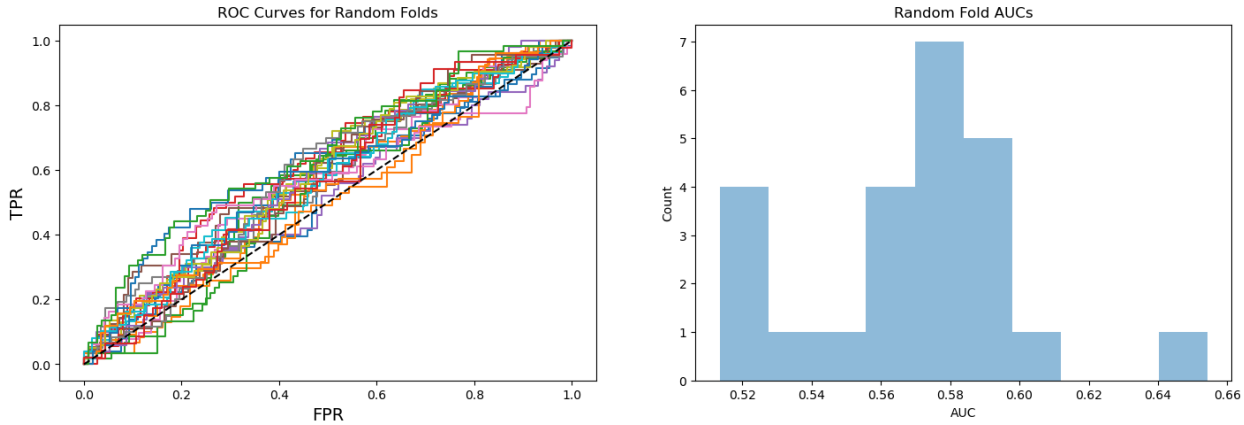

Figure 23. Analysis of Gaussian Naive Bayes with more electrodes on random folds of test data

The AUC values all appear higher than chance. The p-value from the one sample t-test was 2.276e-10, which is much less than 0.05 and shows statistical significance.

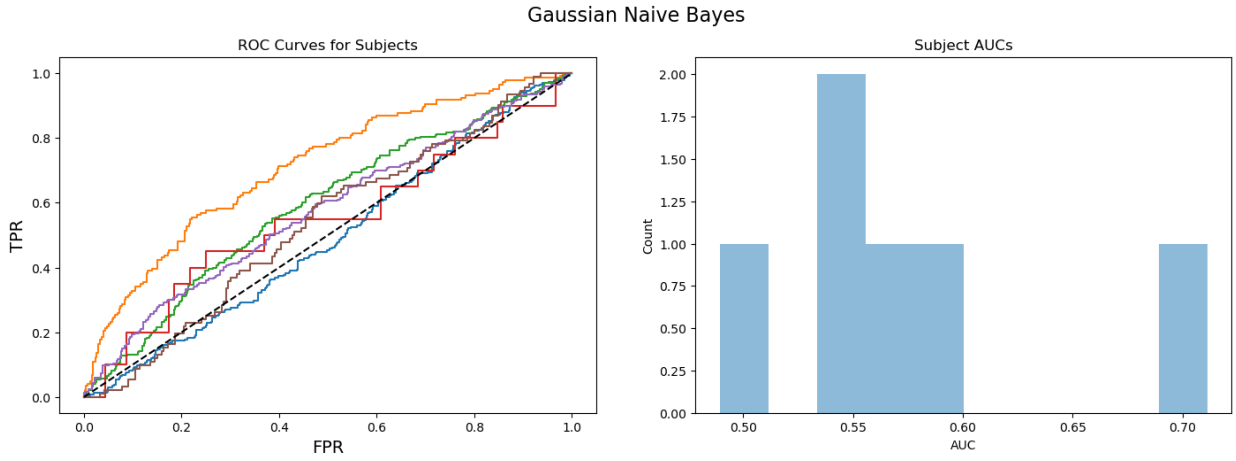

Figure 24. Analysis of Gaussian Naive Bayes with more electrodes on individual subjects

The AUC values all appear higher than chance for almost all subjects except for one. The p-value from the one sample t-test was 0.05996, which is greater than 0.05 and does not show statistical significance. With that said, the model performed very well on a few subjects and shows promise.

Results

Model	**AUC**	**Subject p-value**	**Random fold p-value**
Logistic Regression	.5560	0.2252	9.379e-06
Decision Tree	.5216	.2992	.003455
K-Nearest Neighbors	.5135	0.4841	.2193
Gaussian Naive Bayes	.5696	.05996	2.276e-10

The performance of all models marginally improved (AUCs increased by a couple hundredths) as more features were added to the dataset. Like before, logistic regression and Gaussian naive Bayes were the two dominant values with AUCs that were statistically significant. While the AUC value of the decision tree model was .5216, that number was also statistically significant meaning that it did perform better than chance.

Comparison to Subject-Specific Models

We can compare the performance of these general models trained on various subjects to the performance of a subject-specific model. Since models that were trained on more features (24 temporal electrodes, 16 frontal electrodes, and 4 hippocampal electrodes vs. 16 temporal electrodes, 8 frontal electrodes, and 2 hippocampal electrodes) generally have higher

AUC scores, we will look at the logistic regression analysis of the 3 subjects with the most rows from the dataset with more features.

The optimal penalty scheme was again L2 for all three subjects, and the optimal C values were similar to the optimized general logistic regression model.

R1468J

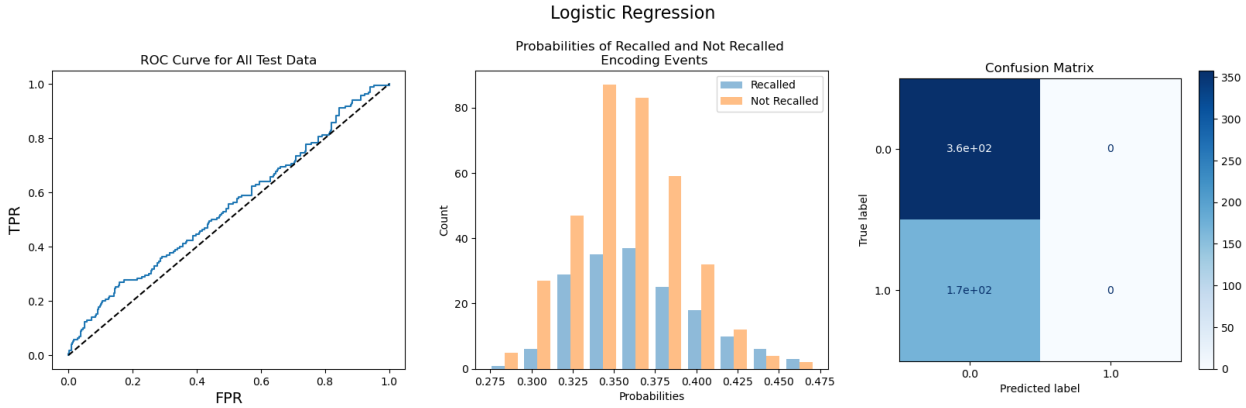

Figure 25. Analysis of logistic regression on subject R1468J

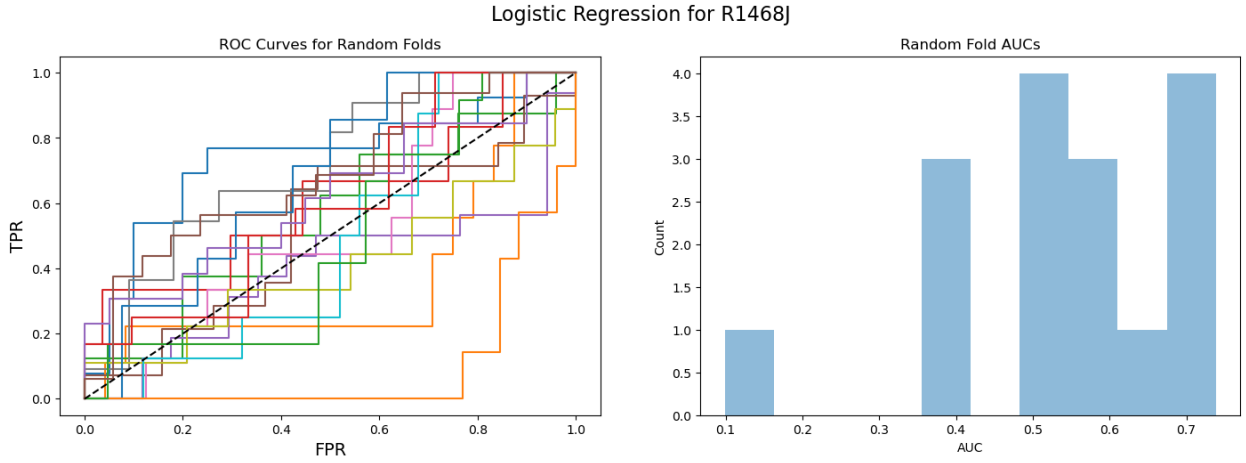

Figure 26. Analysis of logistic regression on subject R1468J AUC and ROC

The AUC is 0.5445 and the probability histogram again does not show separation. Consistent with the other logistic regression analyses conducted, the model does not predict any recalls. The p-value from the t-test is 0.4599, which is obviously greater than 0.05 and not statistically significant. There was high variability in AUC across different folds.

R1157C

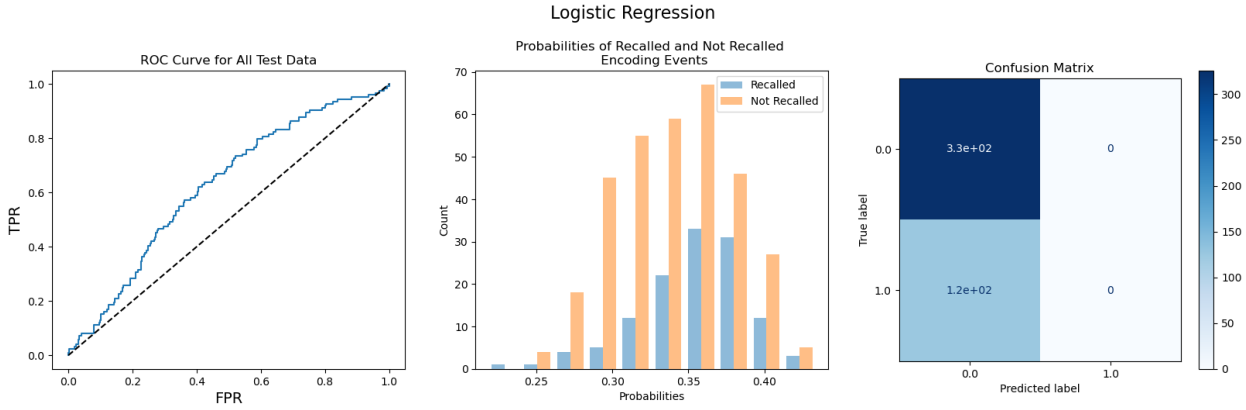

Figure 27. Analysis of logistic regression on subject R1157C

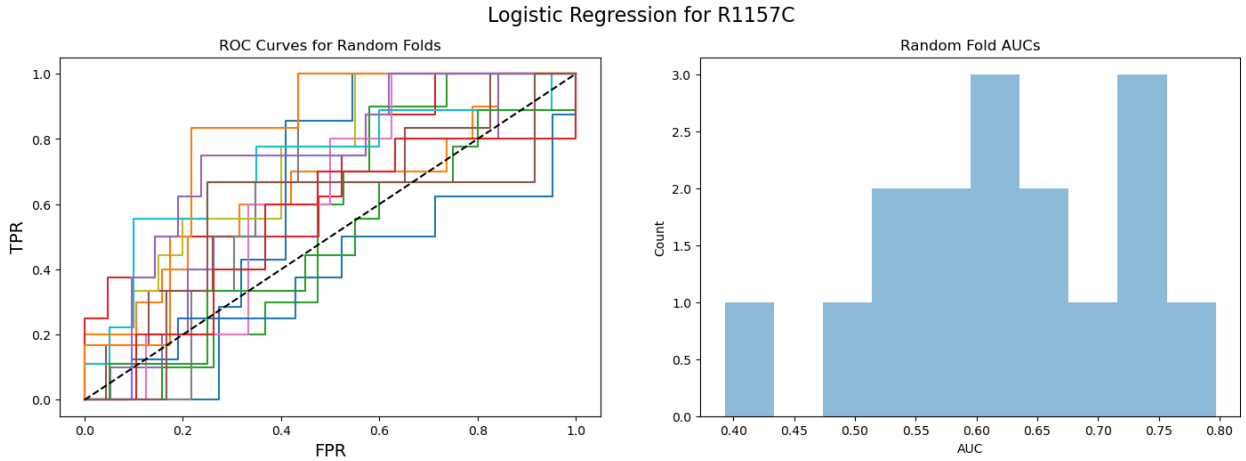

Figure 28. Analysis of logistic regression on subject R1157C AUC and ROC

The logistic regression model performed well. The AUC was 0.6223 when plotted using all of the test data. The p-value from the t-test using random folds of data from the subject was 0.0003166 which is less than 0.05 and statistically significant.

R1364C

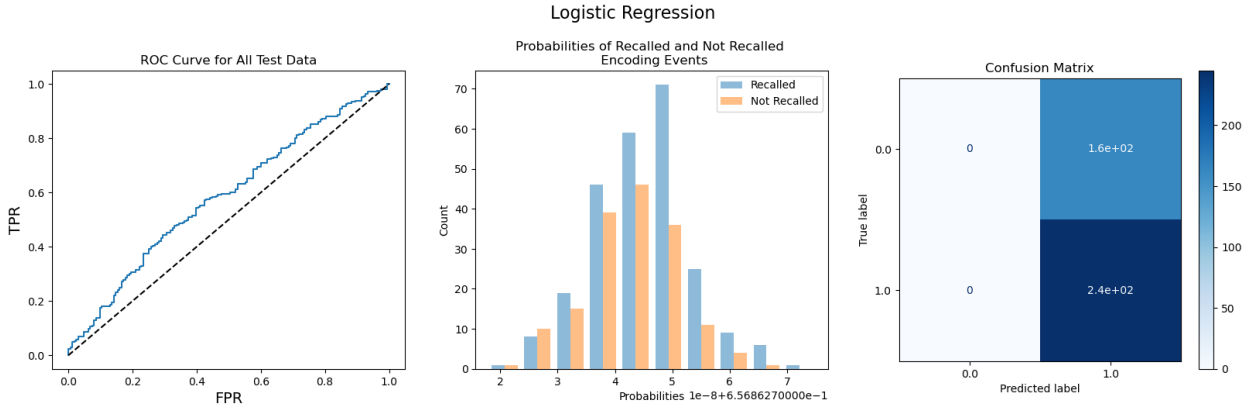

Figure 29. Analysis of logistic regression on subject R1364C

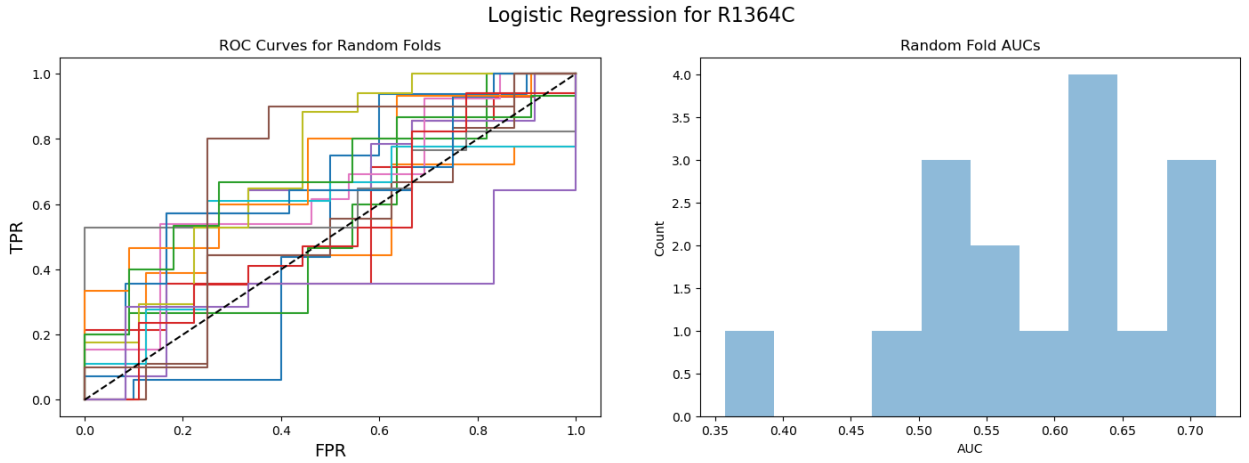

Figure 30. Analysis of logistic regression on subject R1364C AUC and ROC

The AUC value for the subject across all test data is 0.5845. When using the random folds, the one sample t-test yielded a p-value of 0.002639, which is less than 0.05. Notably, for the subject, there was never a predicted non-recall as shown in the confusion matrix in figure 29, which is not consistent with the way this model classified in other cases.

Consolidated Results for Subject Specific Models

Subject	**AUC**	**p-value**
R1468J	.5445	.4599
R1157C	.6223	3.165e-4

R1364C	.5845	.002639

The performance of models on individual subjects can vary, likely due to the low dimensionality of the dataset. For some subjects like R1157C, we have a higher AUC score of .6223 that is statistically significant, but for some subjects like R1468J, we have an AUC score similar to those of the general models that is not statistically significant.

Conclusions

Generalized models for predicting recall do perform above chance, and sometimes even as well as subject specific models (i.e. models that are trained on the subject they aim to predict recall for). Logistic regression, as expected, performed well relative to the other models. Surprisingly, Gaussian naive Bayes also performed well. This initial investigation bodes well for future research into creating generalized models for recall prediction with brain stimulation. If a general model can be created, the computing power for brain implants that stimulate the brain to improve memory would be vastly reduced since pre-trained models could be utilized.

Furthermore, increasing the dimensionality of the data by including more electrodes for each region of the brain increased model performance. The higher p-values for the dataset with more electrodes are likely inflated due to the fewer number of subjects. In future data collecting experiments, more emphasis could be placed on the location of scalp electrodes. If the location of the electrodes could be more consistent across subjects (and also across sessions for a subject), we would be able to utilize electrode readings when creating our dataset and thus increase the performance of recall prediction models. Specifically, adding more electrodes in the hippocampal region could improve model performance as the hippocampus plays a considerable role in memory (Kragel et al. 2017). Since naive Bayesian models do well with high dimensionality datasets, there is a possibility that this future work may result in Gaussian naive Bayes significantly outperforming logistic regression.

Citations

Arora, A., Lin, J., Gasperian, A., Maldjian, J., Stein, J., Kahana, M., et al. (2018). Comparison of logistic regression, support vector machines, and deep learning classifiers for predicting memory encoding success using human intracranial EEG recordings. Journal of Neural Engineering, 15(6), 066028.

Kahana, M. J., Ezzyat, Y., Phan, T.D., Weidemann, C. T., Jacobs, J. EHM chapter 5 Electrophysiology of Human Memory. Unpublished. (Accessed April 24, 2023).

Kragel, J. E., Ezzyat, Y., Sperling, M. R., Gorniak, R., Worrell, G. A., Berry, B. M., et al. (2017). Similar patterns of neural activity predict memory function during encoding and retrieval. NeuroImage, 155, 60–71.

Github Repository