Breast Cancer Diagnosis Classification Analysis

DANNY's TEMPLATE

December 9th, 2022

Contents

1 Background (Should we do an Abstract?)

Insert text here

2 Data

Insert text here

2.1 Variable Descriptions

Insert text here

3 Explatory Data Analysis

Insert text here

4 Austin's Classification Proccess and Analysis

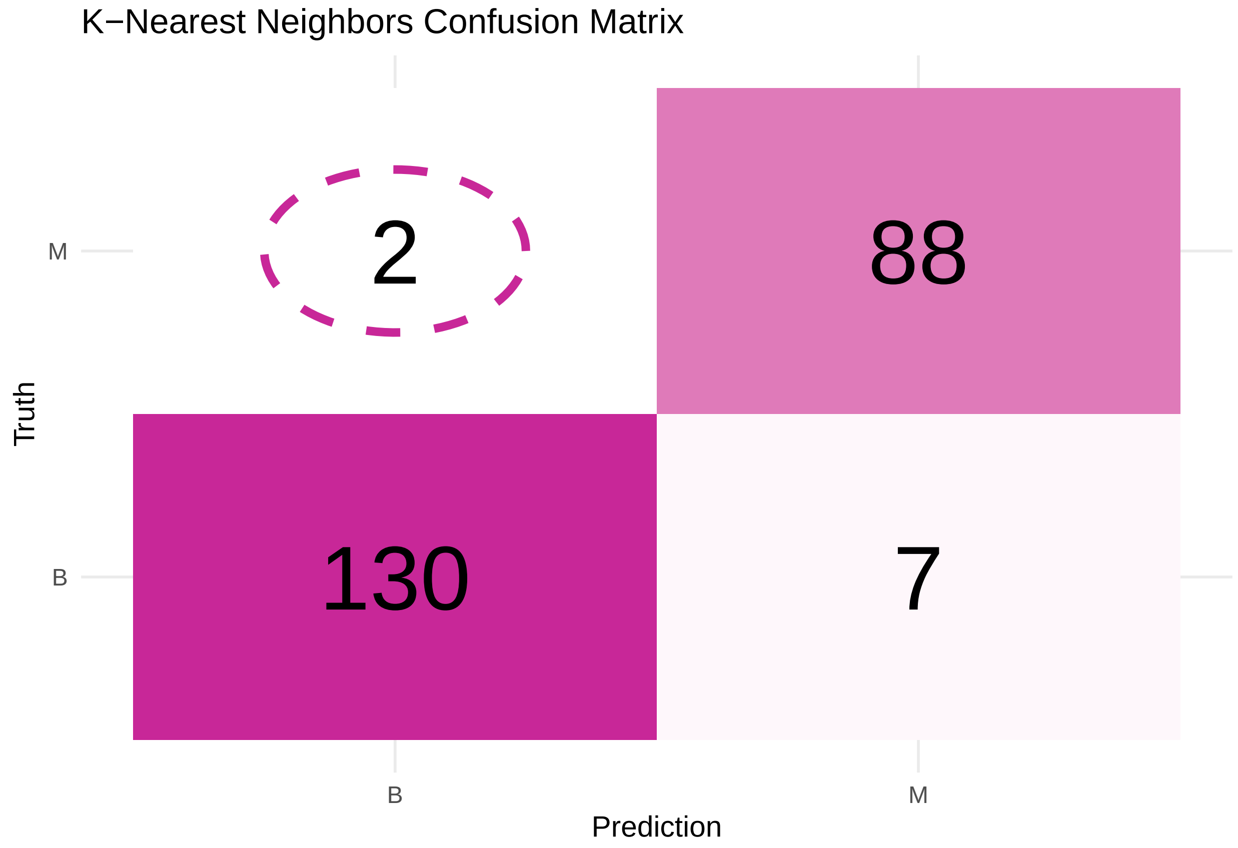
Insert text here

5 Dannys's Classification Proccess and Analysis

As the main purpose of this analysis is to reduce Type II error, we wanted to try different ways of recuding this within similar models. The following models all use classification to classify a tumor as benign or malignant however they have a lower prediction cutoff of 25% malignant to bias our results towards more malignant. For example, if a tumor is only 30% likely to be malignant, we still would classify that tumor as malignant to avoid false negatives. While this does lead to higher Type 1 error and sometimes lower overall accuracy, we aren't as worried about the false-positives. We fit a number of models on the training data using this approach, including Logistic Regression, K-Nearest Neighbors, Quadratic Discriminant Analysis (QDA), Random Forest, xgBoost, and lastly a support vector classifier. While all the most models performed quite decently, there were a few that stood out.

5.0.1 K-Nearest Neighbors

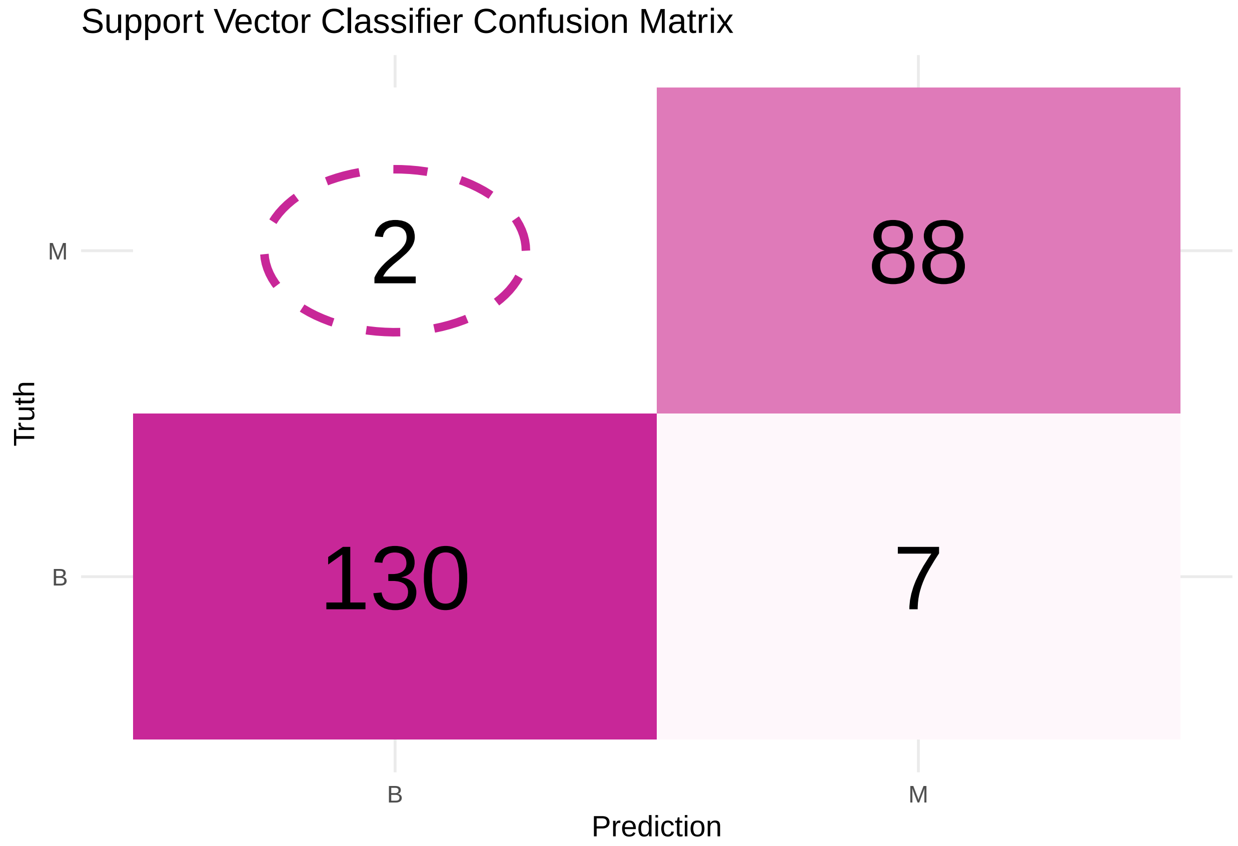
K-Nearest Neighbors is a classifying method that estimates the bayes classifier using the closest training points to the test point that is being classified. It uses a neighbor parameter that is a big factor in determining the classifier. For this model, we used 10-fold cross-validation to tune the neighbor parameter based on the highest accuracy. Once the neighbors parameter was tuned to 7, the model was fit and prediction ensued. The confusion matrix showcasing the classification prediction of this model is below.



As shown in the confusion matrix, this KNN model only predicted two false-negatives out of 223 test observations and 90 true test malignant tumors. The KNN model predicted with an overall accuracy of 96.04. While this model did perform quite well, we still want to improve these results.

5.0.2 Support Vector Classifier

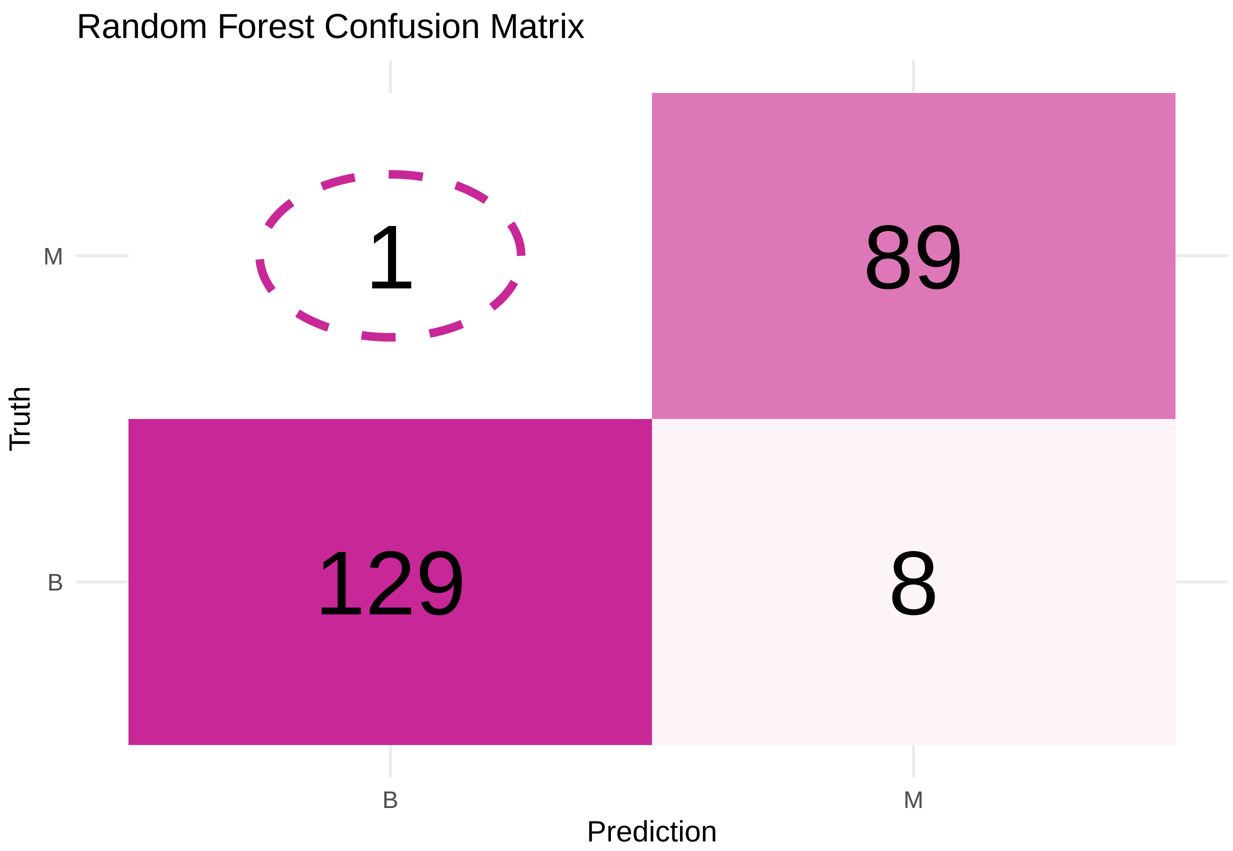
A support vector classifier uses a softened margin classifier to classify the observations but with less sensitivity than a maximal margin classifier. The support vector classifier will not necessarily classify all training points accurately, but this allows for better test prediction. We tuned our support vector classifier to take a cost value of 1 using 10-fold cross-validation. The confusion matrix showcasing the prediciton accuracy of the support vector classifier model can be seen below.


Support Vector Classifier Confusion Matrix

As shown above, the support vector classifier predicted the exact same as the KNN model. With only 2 false negatives and an overall accuracy of 96.04, this model was not an improvement from the KNN model, but it is reassuring to see consistent high accuracy amongst different models. The last model we will look at is a Random Forest model.

5.0.3 Random Forest

Random Forest creates many decision trees from bootstrapped sample training points and uses a subset of predictors for each tree. The main difference between Random Forest and bagging decision trees is this feature subset that helps decorrelate the trees. For this Random Forest model, a default of 500 trees was used as tuning did not improve the model. The confusion matrix showcasing the prediction accuracy of the Random Forest model is shown below.



As shown above, the Random Forest model predicted the best out of all our models according to the standards we set. With only one false-negative, this is the lowest out of all the models and with an overall accuracy of 96.04 it holds up against the other models and does not gain too much Type I error. This confusion matrix portrays the predicition cutoff very well. Because the model biases towards classifying as malignant, there is much higher Type I error than Type II error, which is exactly what we were aiming for.

6 Ethans's Regression Proccess and Analysis

Insert text here

7 Analysis Summary

Insert text here

8 Potential Improvements

Insert text here

9 Works Cited

- Breast Cancer Wisconsin (Diagnostic) Data Set:
  - https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?datasetId=180&sortBy=voteCount
- Definition of Features (Variables):

- https://www.causeweb.org/usproc/sites/default/files/usclap/2017-2/Evaluating_Benign_and_Malignant_Breast_Cancer_Cells_from_Fine-Needle_Aspirates.pdf