Team Sigma: Galvin Li, Dominic Matriccino, & Kevin Nguyen

Regression Problem: We began by creating a regression tree. Cross-validation found the optimal cost parameter to be 0. When the decision tree was applied to test data, the MSPE was found to be 2201.03 (Appendix B). The decision tree is shown in Appendix C. Because the optimal cost parameter was 0, no pruning was performed. Our random forest model (mtry = 2) and bagging (mtry = 7) performed similarly, with MSPEs of 2100.803 (Appendix B). Finally, we applied gradient boosting to our regression problem. The optimal number of trees was found to be 7221 through 10-fold cross-validation (Appendix I), and interaction depth was set to 1. The MSPE for boosting was found to be 2299.528 (Appendix B). Of the four methods, boosting surprisingly performed the worst while random forests/bagging performed the best. Boosting's poor performance may be because a single split per tree might be too simple to explain the data properly, even with 7221 trees built sequentially. Random forests/bagging performed better than the regression tree, and this may be because the regression tree overfit the training data and because random forests/bagging allowed better tree splits to occur that would not have been possible in the regular decision tree due to greedy splitting. Appendices F, G, and J show the variable importance plots, and Appendix E shows the decision tree. Interestingly, all methods agreed that passenger class was the single most important variable in predicting fare, and the number of parents and children (parch) was a distant second in most models.

Classification Problem: Our first nonlinear method is the basic decision tree. After training the model and using cross validation to find the best complexity parameter, our final classification model had a complexity parameter of 0.022. After the best complexity parameter and retraining the model, our misclassification rate using the test data was 19.03%. Our second nonlinear method is random forest. The tuning parameter for random forest is the argument mtry, which specifies the number of predictors that are candidates for each split. Typically, the value for this tuning parameter is the square root of the number of predictors, rounded down. Since there are seven predictors, the best value for mtry is two. After performing random forest with this value for the tuning parameter, our misclassification rate using the test data is 18.28%. Our third nonlinear method is bagging. The tuning parameter for bagging is the same as random forest, mtry. However, for bagging, the value for this tuning parameter is the number of predictors, which is seven. After performing bagging with this value for the tuning parameter, our misclassification rate using the test data is 21.27%. Our final nonlinear method is boosting. The tuning parameter is n.trees, the number of trees to build. We used 10-fold cross-validation to choose the best number of trees to build. After performing boosting with the optimal number of trees, our misclassification rate using the test data is 21.64%. Therefore, random forests performed the best.

Comparison of Performance Between Nonlinear and Linear Methods: Appendices C and D compare nonlinear methods with our linear methods. We can see that, overall, our tree based models performed better. Our best performing model was our random forest model. Our worst performing model was our logistic regression model. Between the best and worst models, there is some slight overlap, with bagging obtaining the same error rate as naive bayes and boosting obtaining the same error rate as LDA. The reason why our linear methods struggle with predicting may have to do with the lack of linearity of our response variable. Linear boundaries such as what can be found in LDA and logistic regression, do not perform as well whereas non linear decision boundaries such as decision trees and QDA perform better because they may be better able to accurately pick up on nonlinearities in the feature space.

Team Sigma: Galvin Li, Dominic Matriccino, & Kevin Nguyen

Appendix A: Comparison of Misclassification Rate (Classification)

Decision Tree	Random Forest	Bagging	Boosting
0.1902985	0.1828358	0.2126866	0.2164179

Appendix B: Comparison of MSPE (Regression)

Decision Tree	Random Forest	Bagging	Boosting
2201.0300	2100.8030	2100.8030	2299.528

Appendix C: Comparison of Nonlinear and Linear Methods' Misclassification Rate (Classification)

Decision Tree	Random Forest	Bagging	Boosting
0.1902985	0.1828358	0.2126866	0.2164179

Naive Bayes	Logistic Regression	LDA	QDA
0.2126866	0.2574627	0.2164179	0.2052239

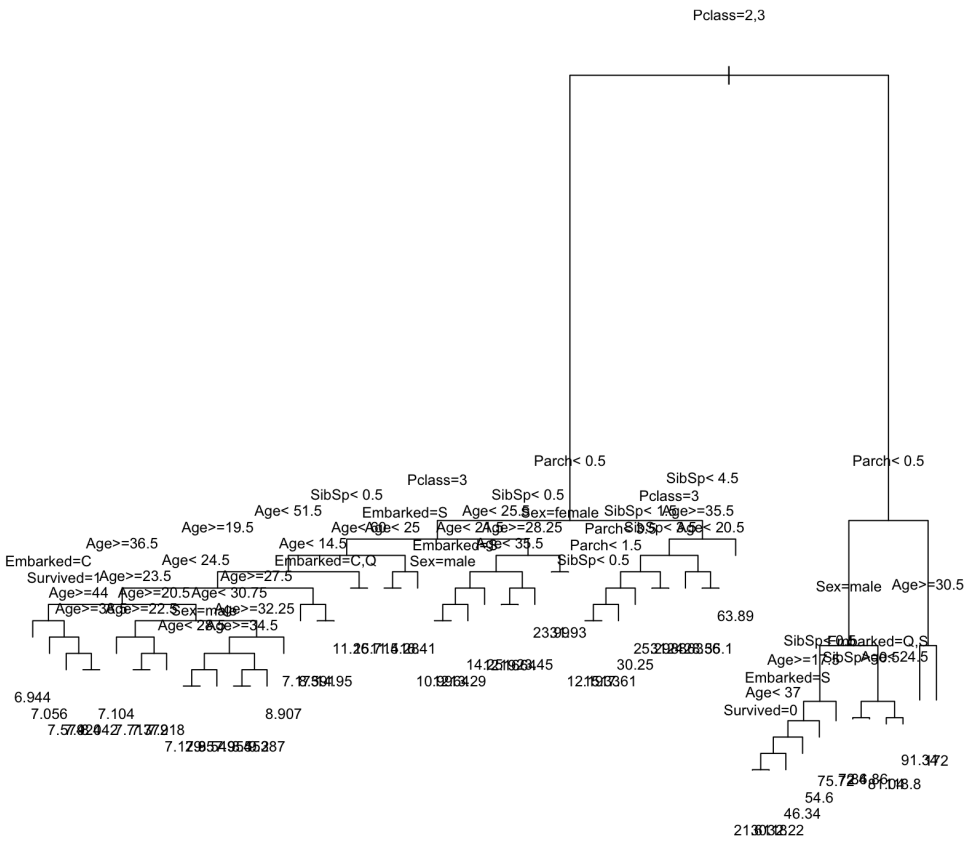
Appendix D: Comparison of Nonlinear and Linear Methods' MSPE (Regression)

Decision Tree	Random Forest	Bagging	Boosting
2201.0300	2100.8030	2100.8030	2299.528

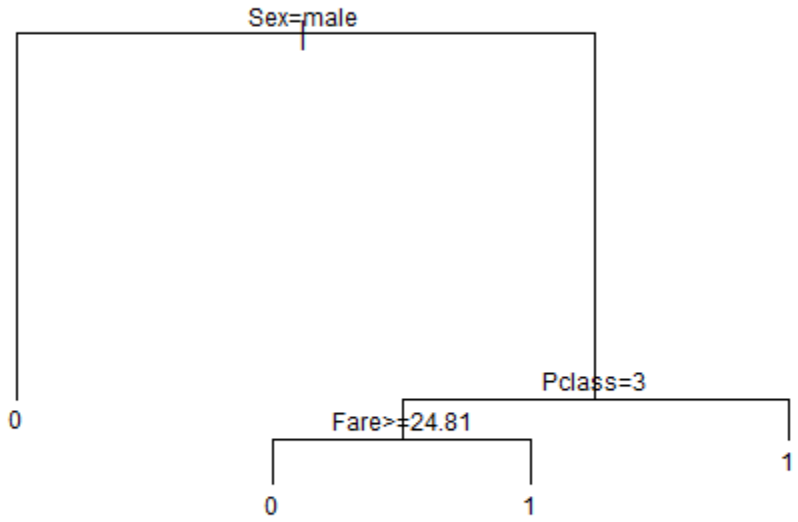
Naive	OLS	Ridge	Lasso
3580.933	2249.013	2277.664	2251.337

Team Sigma: Galvin Li, Dominic Matriccino, & Kevin Nguyen

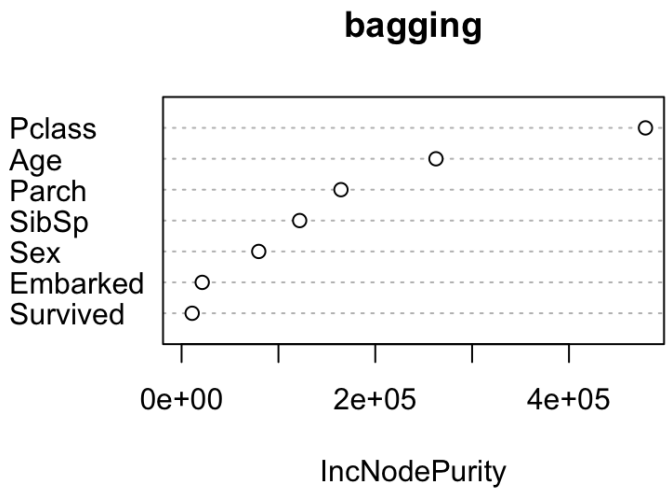
Appendix E: Decision Tree with Complexity Parameter = 0.0 (Regression)



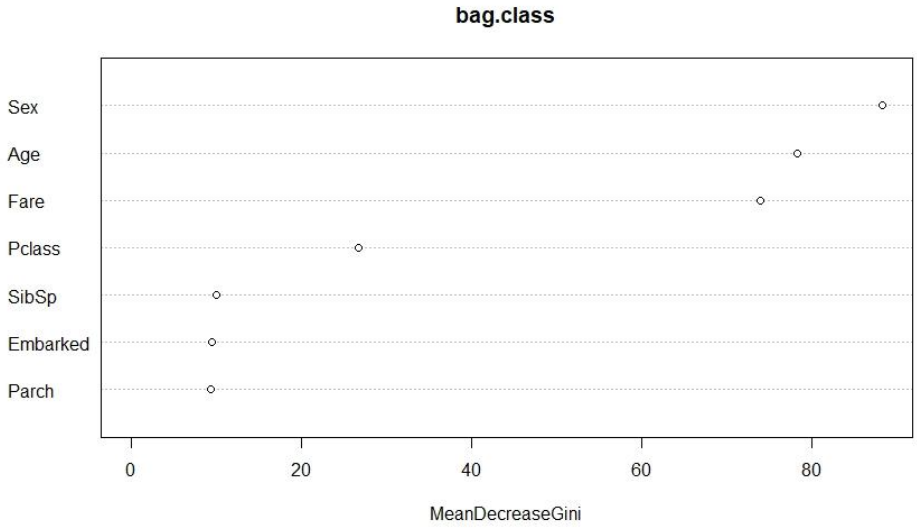
Appendix F: Decision Tree with Complexity Parameter = 0.022 (Classification)



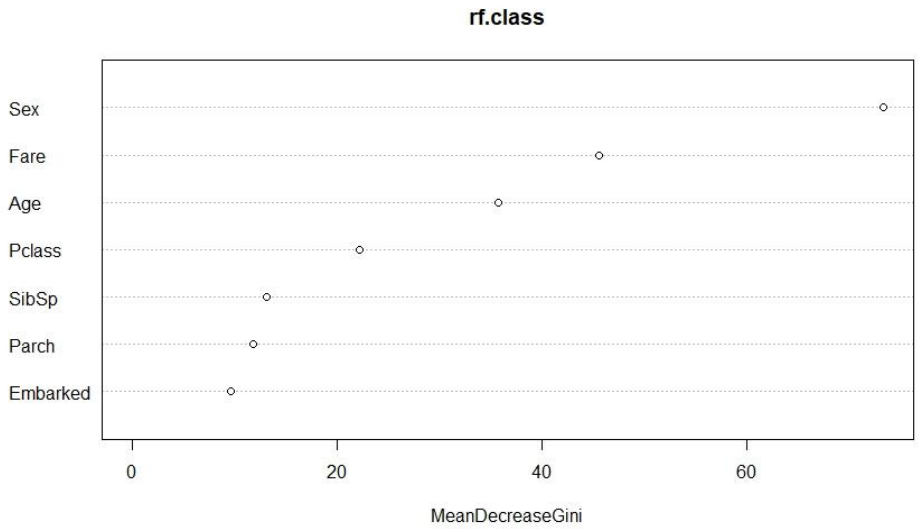
Appendix G: Bagging Variable Importance Plot for Regression Problem



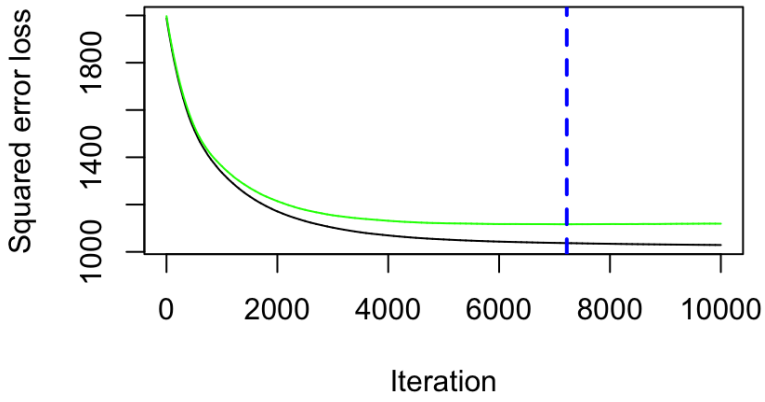
Appendix H: Bagging Variable Importance Plot for Classification Problem



Appendix H: Random Forest Importance Plot for Classification Problem



Appendix I: 10-fold Cross Validation for Boosting to Select the Optimal Number of Trees (Regression)

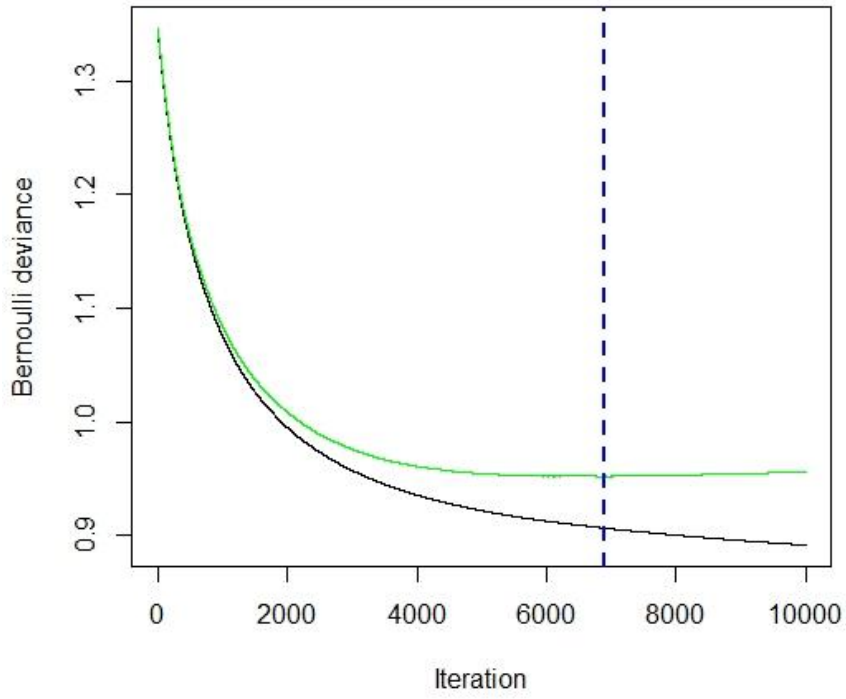



```
A gradient boosted model with gaussian loss function.
10000 iterations were performed.
The best cross-validation iteration was 7221.
There were 7 predictors of which 7 had non-zero influence.
```

Appendix J: Variable Importance for Boosting (Regression)

	var	rel.inf
Pclass	Pclass	71.8107541
Parch	Parch	14.4676809
SibSp	SibSp	6.6545295
Age	Age	2.8911491
Sex	Sex	2.0452823
Embarked	Embarked	1.5557048
Survived	Survived	0.5748991

Appendix K: 10-fold Cross Validation for Boosting to Select the Optimal Number of Trees (Classification)




```
A gradient boosted model with bernoulli loss function.
10000 iterations were performed.
The best cross-validation iteration was 6890.
There were 7 predictors of which 7 had non-zero influence.
```

Appendix L: Variable Importance for Boosting (Classification)

	rel.inf
Sex	57.163351
Pclass	15.902001
Fare	9.274288
Age	6.612239
Embarked	4.917296
SibSp	4.872435
Parch	1.258390