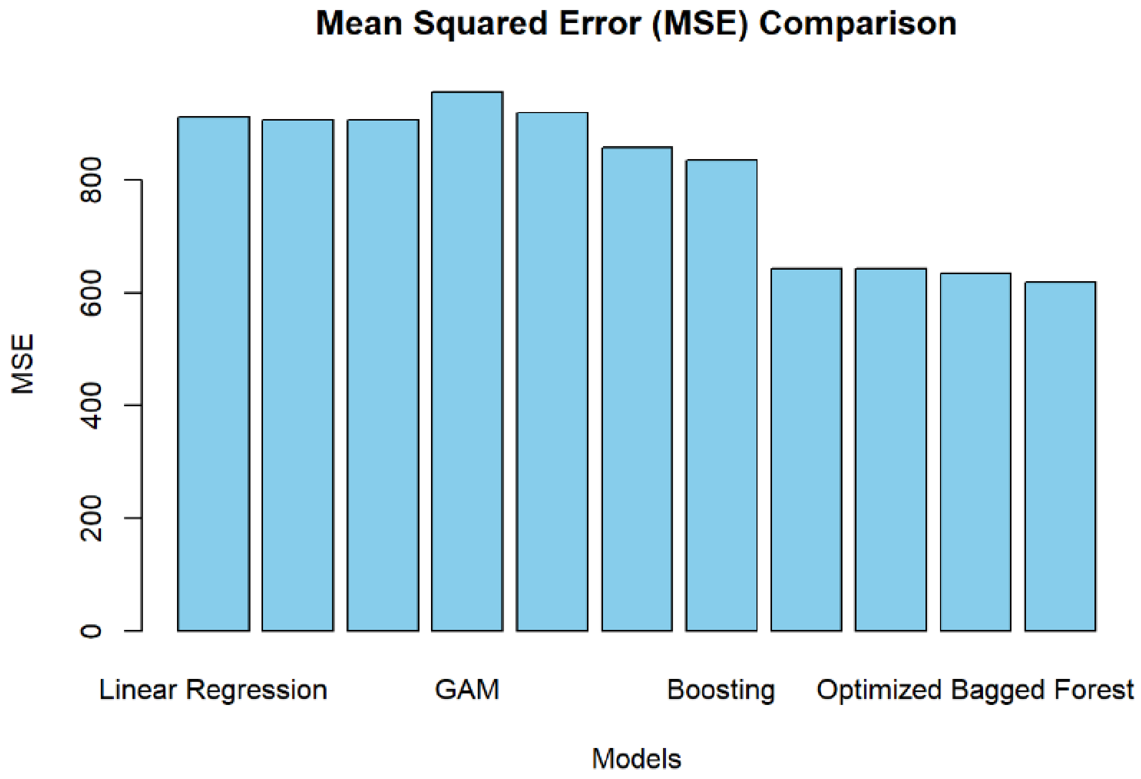
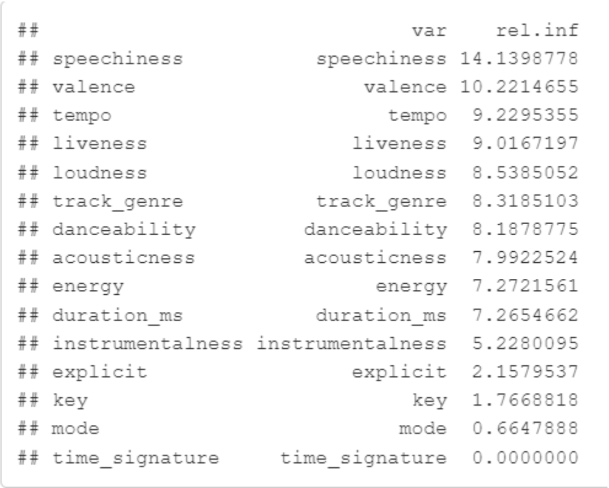
SonicWave Productions approached us with the goal of gaining headway in the music industry. Being able to understand what makes a song popular and to predict the popularity of a new song would be a very valuable asset for the company. Thus, we were given a dataset with 1,200 songs that had a variety of features such as length, genre, tempo, etc. Our goal from our client is to take this dataset and create a model to predict how popular a new song would be. Steps were taken to prepare this dataset for analysis. First, the dataset had to be split into a training set and test set. This is standard practice for creating machine learning models and essential in future evaluation of our created models. Next, handling categorical variables. Categorical variables can cause issues when trying to apply them within models, however after mapping genre, and explicit to a unique numeric value they were able to be used within models to be able to analyze the data. Finally, id, album name, and track name were removed as they were not useful in predicting the popularity of a song. After all of these changes were made, we could start working with the data.

I created a variety of different types of models with the goal of having a large sample size of models to choose from. The first batch of models I created were linear models. This included a simple linear regression model, a lasso model, a ridge regression model, and a best subset selection model. I thought out of all of the models I would create, the simple linear regression model would be the worst and would therefore be a great baseline to compare the rest of the models I created to. Ridge and Lasso are very similar to simple linear regression however their regularization methods make them generally perform better then normal linear regression when applied to the test set. I did cross validate the lasso and ridge models. My next model method was best subset selection, where I used Mallows Cp to evaluate the models and choose the best one. To test for exponential relationships, I created a generalized additive model with splines to the second degree. My next batch of models were tree based models. I started with a normal tree using cross validation, and then pruned it. I followed the basic and pruned tree with a boosted tree. Since boosted trees are highly dependent on their tuning parameter, alpha, I tried a variety of alpha values and used the one that resulted in the minimum amount of test error. I followed my boosted model by a bagged, and random forest. This selection of models includes a mix of robust regularized and non regularized linear and tree based models.

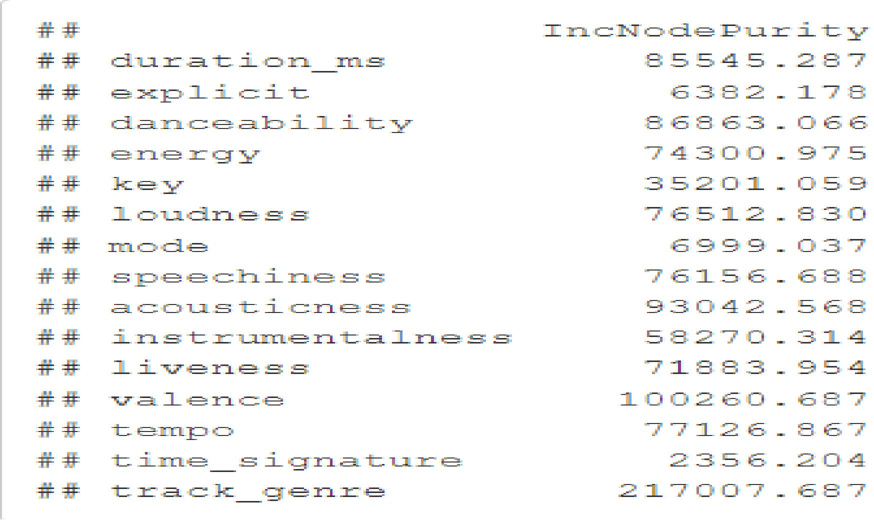


The table above summarizes the results of all of the models we made. From left to right the errors go as follows: Simple Linear 911.2, Ridge 907.9, Lasso 907.9,GAM 956.7, Tree 919.6, Pruned Tree 857.4, Boosted Tree 835.9, Bagged Forest 642.2, Random Forest 642.2, Optimized Bagged Forest 635, and Optimized Random Forest 619.7. This generally follows one's expectations of where each model should place compared to one another. The linear models performed the worst, with the exception of the GAM model which was done just to test which variables might have a nonlinear relationship with the response. I am surprised the Lasso and Ridge regression models did not perform significantly better then the simple linear regression model. Normally, by regularizing linear regression coefficients it leads to better performance in the test set, but this was not really the case here. All of the linear models and

GAM, had test errors in the 900s. This is pretty bad and reflects a pretty inaccurate model. I am not surprised the linear models were not as accurate as the regularized tree based models. The normal tree was worse than the linear models but was made significantly better after it was pruned. By pruning the tree we decreased some of the variance, acting as a better baseline for the tree based models. The pruned tree was beat by boosting, which was tested with a variety of shrinkage values and the alpha value with the lowest error was used. By tuning the alpha value we were able to beat the normal tree and pruned tree in test mse. The last, and best models were the bagged and random forest. The random forest value was used with a fixed mtry value of 3. These were the only models with a test mse in the 600s, and no other model came close to their predictive capability. The overall trend being displayed here is that tree based models perform better than linear based models. Furthermore, regularization was not shown justice within the linear models, however, regularizing models does generally make test performance better. This is true here in our tree models, as the normal, and even pruned tree displayed too much variance within data. Pruning a tree is a form of regularization, but not enough. After introducing randomness into our trees, and therefore regularizing the models, they became the most powerful models out of our sample. Initially there was not a significant difference between the bagged forest and random forest so either one could have been used. However, after optimizing the bagged forest by trying different numbers of trees, and optimizing the random forest with different pairings of number of trees and mtry values, the random forest performed the best having the lowest test MSE. This lead me to decide I will be using the random forest on the test set. Other important considerations were the different model’s analysis of which variables were important:



```
##                                 var    rel.inf
## speechiness             speechiness 14.1398778
## valence                     valence 10.2214655
## tempo                         tempo  9.2295355
## liveness                   liveness  9.0167197
## loudness                   loudness  8.5385052
## track_genre             track_genre  8.3185103
## danceability           danceability  8.1878775
## acousticness           acousticness  7.9922524
## energy                       energy  7.2721561
## duration_ms             duration_ms  7.2654662
## instrumentalness instrumentalness  5.2280095
## explicit                   explicit  2.1579537
## key                             key  1.7668818
## mode                           mode  0.6647888
## time_signature       time_signature  0.0000000
```



```
##                       IncNodePurity
## duration_ms               85545.287
## explicit                   6382.178
## danceability              86863.066
## energy                    74300.975
## key                       35201.059
## loudness                  76512.830
## mode                       6999.037
## speechiness               76156.688
## acousticness              93042.568
## instrumentalness          58270.314
## liveness                  71883.954
## valence                  100260.687
## tempo                     77126.867
## time_signature             2356.204
## track_genre              217007.687
```

A perk of tree based models is they offer a way to measure variable importance. Using the boosted tree model and the out of bag variable importance from the random forest model we are able to use the above tables to understand the most important variables in predicting popularity. The variables with the highest variable influence in the above left table and the variables with the highest node purity(Gini Index) in the above right table will be the most significant. In summary, both tables agree valence is highly significant, and track genre, tempo and speechiness are highly likely to be significant.

The meaningful conclusions here are tree based models are better than linear models, and regularization is a powerful tool that produces the best models. Considering all tree models besides the normal tree were better than the linear models, and the degree of difference between bagged and random forest models and linear models test MSE, it is safe to say that tree models are significantly better than linear models in predictive capability. My second major takeaway is despite the impact of regularization being less apparent in the linear models, it was quite profound in the tree models. I was disappointed that despite regularization being such a powerful tool, there was not a significant difference between the regularized linear models and simple linear model. In the future I would like to develop a method to tune the ridge and lasso regression shrinkage coefficient on my own in a similar way for the boosted model’s alpha was tuned to find the optimal level. I feel cross validating the models was an acceptable substitute, but I am curious to see if tuning the alpha myself would create a significant difference. I am confident that random forests and bagged forests would still be significantly better than these new tuned regularized linear models leading me to conclude trees are generally better. The basic tree did not perform well and it was actually worse than the linear model. However by pruning the tree, it performed better than all of the linear models, and by introducing randomness and using the bagged and random forest, the best model was created. This goes to show regularizing highly flexible models creates highly accurate predictive models. The last thing I would like to investigate, the data. I am curious if there is a connection to the ridge and lasso models performing similar to the simple linear regression model and the bagged forest and random forest being so similar. Regularized regression methods are supposed to perform better than simple linear regression.

This is made more interesting by the bagged and random forest performing so similar. Bagged forests have been proven to perform better with high snr ratio data, and random forests perform better with low snr data. This makes me question the data, where 2 separate groups of models that theoretically were supposed to perform differently, performed almost the same.

This study produced meaningful conclusions. The comparison of tree based to linear models and the analysis of the impact of regularization led to the creation of models that had incredibly low test MSE relative to simple linear regression. We have concluded random, and bagged forest models are the best, and the only differentiator is the ability to slightly optimize the random forest by mtry and number of tree values. More generally, regularizing highly flexible models produces the lowest test MSE.