Ian Keller

Dr. Mensch

STAT 1361

16 April, 2024

Technical Report

Introduction and Exploratory Data Analysis

To begin our analysis, I started by looking to better understand the variable types and summary statistics to see what type of analysis and feature engineering needed to be done. Here, we can recognize first and foremost that the popularity variable is continuous and therefore we are dealing with a regression problem in which we are attempting to predict the "popularity" variable. Additionally, our dataset is comprised of mainly numeric values and integers, along with one logical data type. The dataset also contains 3 variables of character data types.

Next, looking into some data oddities. I started by checking if the data had any N/A values and it appeared to have none so the data is already pretty solid. Furthermore, I began to look into any outliers and extreme values by looping through boxplot and histogram graphs. The first thing I noticed is that popularity (dependent variable) had no outliers which is a good sign. On the other hand, a few variables had outliers such as duration, danceability, loudness, speechiness, instrumentalness, liveness, tempo, and time signature. Given industry knowledge and context, I decided to keep these outliers, for

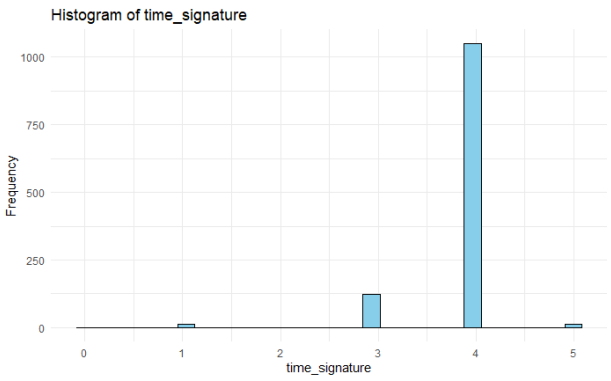

Figure 1: Histogram of Time Signature Variable

all categories except time signature, as I believed that they were valid data points that can help provide additional signal to our model. The time signature variable became of interest to me due to its highly skewed nature as shown in Figure 1 above. Here, we see that a large majority of time signature data points were equal to the value of 4. To increase the interpretability of our model, I transformed this continuous variable into a binary variable called time_sig_4 where 1's represented a time signature of 4 and 0's represented all other time signatures. This was done after analyzing the skewed histogram and constructing a table indicating that 1049 out of 1200 data points corresponded to a time signature of 4. Moreover, I conducted a t-test to demonstrate the significant disparity in mean popularity values between instances where the time signature equaled 4 and those where it did not. This comparison is shown significant by the output presented after line 97 in the RMD file.

Additionally, to incorporate all of the information from our dataset as possible I thought it was necessary to create dummy variables to represent our genre string variable. To do this I had to create n-1 new columns where I left the jazz genre as the base group. This left me with two new variables titled genre_pop and genre_rock, which were binary where 1 represented if a song fell under the given genre and 0 if otherwise. If both variables were equal to 0 for a given row, then we can conclude that the genre was Jazz. Finally, I subsetted the dataset to only encompass these numeric values, eliminating track and album names. Additionally, I excluded the 'id' column, along with the 'time_signature' and 'track_genre' columns, as they were modified for our analysis, leaving us with 16 independent variables to predict popularity.

Modeling

To start modeling, we split our training data into a 75/25 ratio for training and validation which yielded 901 observations for training and 299 for validation. We then created variables

X_train, y_train, X_test, and y_test for modeling with the aim to run regression models to minimize MSE and identify significant variables. First, I began by fitting several multiple linear regression models to the training dataset. These models included the full linear model, which used all 16 predictors. This was followed by 3 linear models selected by forward, backward, and stepwise model (variable) selection, minimizing BIC (Bayesian Information Criterion). With this dataset, the full linear model performed the best out of our linear models with an MSE of 934.70 although the stepwise and backward selection models (which selected the same predictors) were not far behind with an MSE of 944.66. In this scenario, I prefer the strong interpretability of the stepwise/backward model as it only used 5 highly significant predictors compared to the 16 used by the full model.

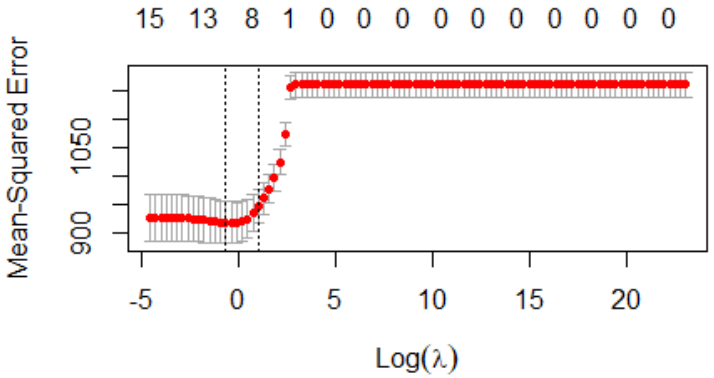

Figure 2: Mean Squared Error from different Lambda Values

Moving on, I utilized regularization models such as Naïve Bayes and Lasso regression. For both models, we had to begin by initializing the model and then tune the model using 100 different values of lambda from our grid variable. The results are shown above in Figure 2 from the Lasso Regression model. Here, a lambda of 0.498 minimizes our MSE through cross-validation. Despite the Naïve Bayes test MSE being lower (870.54), I prefer the Lasso model due to its interpretability as it shrunk 5 of our variable's coefficients down to 0, signifying their lack of importance in our model. Both Naïve Bayes and Lasso Regression performed better than Linear Regression.

Next, I began to fit several tree-based models to our training set. I started with a simple decision tree and a pruned decision tree model which resulted in a slightly better MSE than linear models. Moving to more in-depth tree models, I initiated the fitting process with a Bagging

model. Here, we have a substantial decrease in Mean MSE to 640.22 with the Bagging model. This benchmark was beaten slightly by the Random Forest model with 639.13 as the MSE. Random forest performing the best may be explained by its efficacy in dealing with real-world noisy datasets. Figure 3 shows the important features in the Random Forest model, which differs from the bagging model as shown by the plot on line 389 in the RMD. Lastly, I fit a Boosted Regression Tree and a Bayesian Additive Regression Tree model to the training set and saw strong performance but not as accurate as the Random Forest and Bagging models.

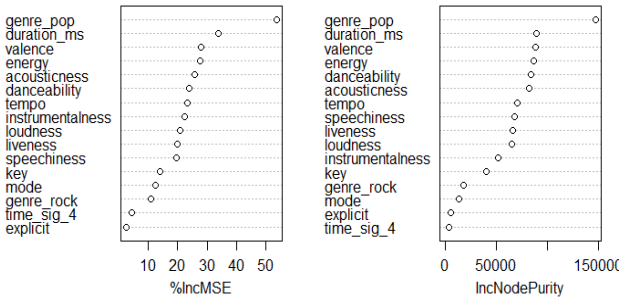

Figure 3: Random Forest Model Variable Importance

Conclusions

Overall, it is clear that the Random Forest and Bagging models performed the best on our data set in terms of predictive accuracy measured by MSE on the validation set. These models also give us some intuition behind variable importance using %IncMSE and IncNodePurity (Figure 3) which represent an increase in MSE from permutation and an increase in node purity by splitting on a particular variable, respectively. Combining these results with results from multiple a linear regression model we can better understand the influence of variable direction on popularity along with significance as shown in Figure 4. The p-values within the output tell us the variable importance (significance) and the "Estimate" represents a one-unit change in each variable's impact on the dependent variable of popularity. Additionally, it is best to use the stepwise/backward linear regression model in Figure 4, as opposed to the full model, as it induces a penalty for added variables, through BIC, which gives us a clear picture of what is truly important in predicting popularity. Figure 4 shows a curious

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.740e+00  5.635e+00  -0.486 0.626975
genre_pop    2.573e+01  2.311e+00  11.135  < 2e-16 ***
duration_ms  4.582e-05  1.821e-05   2.516 0.012031 *
valence     -2.547e+01  5.483e+00  -4.645 3.91e-06 ***
danceability 3.027e+01  8.248e+00   3.670 0.000257 ***
energy       1.423e+01  4.965e+00   2.867 0.004240 **
```

Figure 4: Stepwise/Backward Selection Model Summary

finding: while danceability positively correlates with popularity, valence has a negative coefficient. This suggests that both upbeat and sad songs gain popularity, which is unexpected.

Challenges and Limitations

One difficulty that I found with this dataset was regarding the number of outlier independent variables. I had debated transforming many of these variables since there were many outliers but decided against it due to the integrity of the data along with the emphasis on interpretability and being able to describe accurate results to the SonicWave CEO. The only variable I transformed was the time signature variable, which I believe adds to easy interpretation but consequently would hurt my model's performance due to loss of signal. Lastly, I think it could be beneficial to split our training set into three based on genres to see if variable importance changes within each genre, but in this context, I felt it was important to understand which genre as a whole will lead to an increase in popularity.

Overall, I have reasonable confidence in my best model's (Random Forest) ability to predict song popularity. The model's RMSE of 25.28 indicates a notable deviation from actual values, considering the popularity scale of 0 to 98. Additionally, our full linear regression model resulted in an R-squared value of 0.24, suggesting that only 24% of the variation is explained by our predictors. This underscores the importance of considering other factors, such as the artist's name and release date. Despite these limitations, my work is robust within the constraints of the dataset. While deep learning models might enhance prediction accuracy, they sacrifice interpretability, crucial in this context. Thus, employing multiple models—one for prediction and another for interpretability—seems essential when reporting to SonicWave Productions.