Predicting Music Genre: Multinomial Logistic Regression vs Random Forest

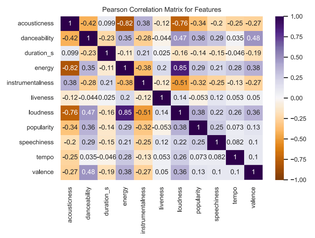
Benny Collins Student ID: 210036917

1 Description & Motivation

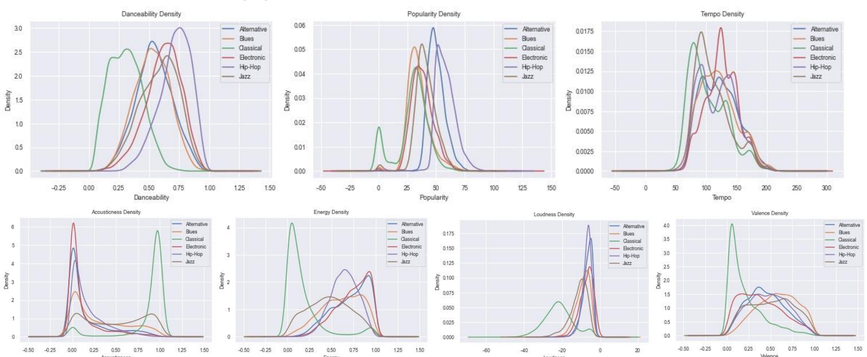
We are aiming to solve the multinomial classification problem of assigning songs to music genre classes based on a set of up to 12 musical features acting as predictors. The algorithms tasked with finding a solution, **multinomial logistic regression (MLR)** and **random forest (RF)**, will then be compared using performance metrics calculated for each algorithm.

2 Pre-Processing & Initial Analysis

- The dataset used is the Prediction of Music Genre dataset from Kaggle, consisting of 50005 songs with 17 features and the genre target class, all of which are observations taken from the Spotify API (gaoyuan, 2021)
- All columns containing data that was unhelpful or didn't fit our model selection, such as artist name, were removed and the binary variable, mode, was numerically encoded with 0 and 1 indicating major and minor modes, respectively..
- Rows containing missing values or invalid values, such as negative times for duration, were then removed.
- The target variable is a set of nominal classes of music genres, with a few that are very similar in terms of musical features; Rap and Hip-Hop, for example. These similarities had the potential to present complications in our model fitting, so we took a random sample of 4000 observations from each of the 6 most popular genres that were considered to be adequately distinct from each other: Alternative, Blues, Classical, Electronic, Hip-Hop and Jazz.



- The Pearson correlation coefficient matrix calculated for the 12 features in X shows a very high level of correlation between the pairs of attributes: energy and loudness, loudness and acousticness, and energy and acousticness. This suggests a degree of multicollinearity between the attributes.
- Our dataset, before removing features based on calculations of variance inflation factor (VIF), contains 24000 observations, evenly split over our 6 target variable genre classes, as well as 10 continuous, quantitative features and 1 binary feature.
- We plot the Kernel Density Estimation graphs, and remove the duration feature, as there is very little distinction in the measurements across the different classes.
  - The 3 graphs along the top suggest that the 'danceability', 'popularity' and 'tempo' predictors could have a reasonable level of importance in our models, as there are distinctions between the different genres.
  - The 4 in the bottom row display a large distinction between attribute measurements for the Classical genre class and the other 5 genre classes. This implies that our models will classify Classical observations more successfully than observations belonging to other classes.



3 Multiple Multinomial Logistic Regression (MLR): Summary

- Logistic regression is a supervised machine learning classification algorithm for predicting the value of a binary dependent target variable, Y, based on a set of independent attributes, X.
- MLR is an extension of this for multinomial classification, where Y is nominal and contains more than one class.
- The objective of logistic regression modelling is to model the posterior probabilities of N classes "via linear functions of the attributes in X, [with each attribute being assigned a weight], while ensuring that these probabilities sum to one and remain in [0, 1]" (Hastie, Tibshirani and Friedman, 2017, p. 119)
- Each observation the model is tested on will return a set of N probabilities, where the i^{th} probability is the posterior probability that the observation belongs to class i. We obtain our final classifications by assigning each observation to the class for which it has the highest posterior probability.
- We select a baseline class, against which the other classes are compared in a pairwise manner. The final model is specified in terms of N-1 log-odds ratios, where the denominator is the baseline class, and the numerators correspond to the other N-1 classes. The choice of baseline class is arbitrary, as the log odds ratios estimates are equivariant under this choice (Hastie, Tibshirani and Friedman, 2017, p. 119).The log-odds are then mapped into posterior probabilities through the use of the sigmoid function.
- Feature weights in logistic regression models are usually fit through the use of Maximum Likelihood Estimation, applied to the conditional likelihood of Y given X (Hastie, Tibshirani and Friedman, 2017, p. 120). This method involves maximising the likelihood or, more often, log-likelihood function.
- Two other methods that are used are gradient descent and the Newton-Raphson method, however maximum likelihood "is extremely flexible" (Eliason, 1993, p. 1) and, thus, more generalisable, often meaning it's the preferred choice.

References

Belsley, D. A., Kuh, E. and Elsch, R. E. (1980) *Regression diagnostics: Identifying influential data and sources of collinearity.* Nashville, TN: John Wiley & Sons.

Breiman, L. (2001) "Random forests," *Machine learning*, 45(1), pp. 5-32. doi: 10.1023/a:1010933404324.

Couronné, R., Probst, P. and Boulesteix, A. (2018) "Random forest versus logistic regression: a large-scale benchmark experiment," *BMC bioinformatics*, 19(1). doi: 10.1186/s12859-018-2264-5.

Eliason, S. R. (1993) "Maximum Likelihood Estimation: Logic and Practice," *Sage University Paper series on Quantitative Applications in the Social Sciences*, (07-096)

Gaoyuan (2021) Prediction of music genre, Kaggle. Available at: https://www.kaggle.com/vicsuperman/prediction-of-music-genre (Accessed: December 10, 2021)

Hastie, T., Tibshirani, R. and Friedman, J. (2017) *The elements of statistical learning: Data mining, inference, and prediction.* 2nd ed. New York, NY: Springer.

Jeune, W. *et al.* (2018) "Multinomial logistic regression and random forest classifiers in digital mapping of soil classes in western Haiti," *Revista brasileira de ciencia do solo*, 42(0). doi: 10.1590/18069657rbcs20170133.

Kirasich, K., Smith, T. and Sadler, B. (2018) "Random forest vs logistic regression: Binary classification for heterogeneous datasets," *SMU Data Science Review*, 1(3), p. 9. Available at: https://scholar.smu.edu/datasciencerevew/vol1/iss3/9 (Accessed: December 11, 2021)

3.1 Advantages:

- Easy to interpret, as it models posterior probabilities.
- Variables can be discrete or continuous.
- No assumptions made about distribution of classes or attributes.
- Doesn't require a linear relationship between attributes and target classes.
- Attribute coefficients provide insight to attribute prediction importance, as well as identifying a positive or negative association of a variable to the two classes involved in each individual binary logistic regression.
- Overfitting can be addressed with Lasso, Ridge or Elastic Net regularisation, in order to remove redundant attributes as well as reduce the coefficients of overemphasised attributes.

3.2 Disadvantages:

- Having less observations than features can lead to overfitting.
- High-dimensional datasets may lead to overfitting.
- Multicollinearity between predictors can cause major issues in training the model.
- Can only be used to predict categorical data.
- Hard to model complex relationships, as it's a linear model.

4 Random Forest (RF): Summary

- RF is a supervised machine learning algorithm for either classification or regression tasks, and has grown into a standard classification approach competing with logistic regression in many innovation-friendly scientific fields (Couronné, Probst and Boulesteix, 2018). In this project the algorithm is used for classification of a multinomial dependent target variable, Y, based on a set of independent attributes, X.
- RF is an extension of the decision tree algorithm that has displayed significant improvements in classification accuracy, by means of "growing an ensemble of trees and letting them vote" for the predicted class of a test observation (Breiman, 2001).
- Each of the trees are fit on different bootstrapped datasets that are the same size as the original training set, before the trees are combined using bootstrap aggregation. The number of trees in the ensemble acts as a hyperparameter for the algorithm.
- Subsets of variables are used to determine each split in a tree. The size of a subset of variables for each split is another hyperparameter for RFs.

4.1 Advantages:

- Variables can be discrete or continuous.
- No assumptions made about distribution of classes or attributes.
- Can be employed for classification as well as regression.
- Decision trees, the building blocks of RFs, "can capture complex interaction structures in the data, and if grown sufficiently deep, have relatively low bias" (Hastie, Tibshirani and Friedman, 2017, p. 587-588).
- Minimises overfitting problem often encountered in decision trees.
- The classification algorithm can handle missing values without losing too much accuracy.
- Out-of-bag permuted predictor importance estimates provide insight to attribute prediction importance.

4.2 Disadvantages:

- Builds many trees and, as a result, is very computationally expensive for both training and testing.
- Not easily interpretable, due to the large number of decision trees.

5 Hypothesis Statement

- I hypothesise that both functions will predict target classes with moderate misclassification rates. This is because determining which music genre a composition belongs to is a task that is often subjective in nature.
- I further assert that our MLR algorithm will perform better than the RF algorithm. This is because, "in general, logistic regression performs better when the number of noise variables is less than or equal to the number of explanatory variables", whereas "RF has a higher true and false positive rate as the number of explanatory variables increases in a dataset" (Kirasich, Smith and Sadler, 2018, p. 21), meaning RF's misclassification rate is unlikely to be greatly improved by adding more explanatory variables.
- On the other hand, RF has advantages compared to MLR, and other linear models, due to its ability to model non-linear relationships between predictors and the response variable. This allows it to handle noisy data more accurately (Jeune *et al.*, 2018), which will likely be very useful for classifying music genres due to their relative subjectivity.

6 Methodology

- We did a 7:3 ratio split of our dataset for the train and test sets respectively, resulting in a train set containing 16800 observations and a test set containing 7200 observations, with the proportions of the dependent variable maintained across all sets, due to the split being stratified.
- The training set is also subject to a stratified 10-fold split. 9 of these folds will be used to train a model, with the remaining fold acting as a validation set for testing and calculating error. This will be repeated with each of our 10 folds taking a turn as the validation set. We then acquire the k-fold cross-validation error by averaging over the 10 misclassification errors we have calculated. This reduces bias in our training of the model and, thus, provides us with a less overfit model to test on and a greater understanding of how appropriate the algorithms, MLR or RF, are for this task.
- Classification errors, cross-validation errors, training errors, precision, recall, F1, and confusion matrices will all be utilised in order to compare the models.

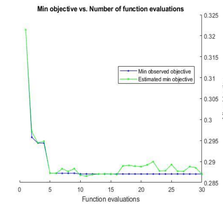
7 Parameter Choice & Experimental Results

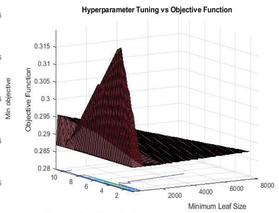
7.1 Multinomial Logistic Regression

- To reduce multicollinearity between our variables, we calculated the variance inflation factor (VIF) for each variable then removed the variable with the highest VIF, if it was over a certain threshold. Each time a variable was removed, we repeated the entire process until there were no more attributes with a VIF over the specified threshold.
- We used the thresholds of 2.5 and 5, however our logistic regression model achieved its best performance when VIF was not considered, which resulted in no features being removed..
- This supports the assertion that one of the key weaknesses of VIF is its lack of a "meaningful boundary to distinguish between values of VIF that can be considered high and those that can be considered low" (Belsley, Kuh and Welsch, 1980).

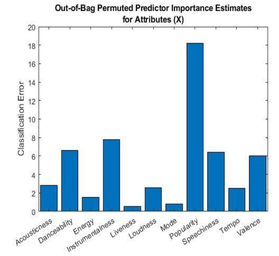
7.2 Random Forest

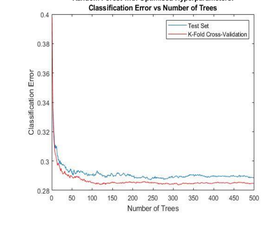
- Hyperparameters were optimised by having the *fitcensemble* function try different combinations of hyperparameters for minimum leaf size, the number of variables to sample to determine each split and the number of trees to fit.
- The optimised hyperparameters:
  - Minimum leaf size of 6.
  - 5 variables sampled for each split.
  - 499 trees.





- The out-of-bag predictor importance estimations were calculated. This is done for a particular variable by calculating the increase in the out-of-bag (OOB) error for the model with the relevant variable's values permuted between different observations. The mean and standard deviation of the difference in OOB error for a variable is calculated from the separate errors for each tree. The predictor importance estimator for that variable is then equal to the mean divided by the standard deviation.





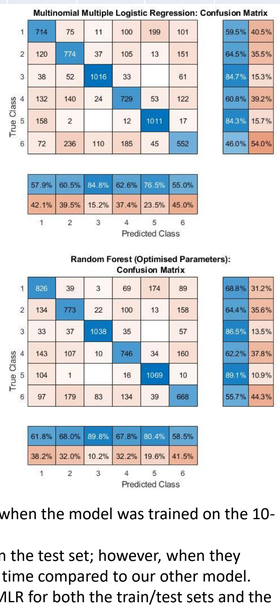
- We see very low estimated predictor importance for the attributes 'mode' and 'liveness', which is unsurprising considering mode is a binary variable where both options likely occur a similar number of times in each genre. Despite this, the performance of the model with these two features removed was less accurate, so we have kept them in the final model.

8 Analysis & Evaluation of Results

- Generalisation error estimates (test error & 10-fold cross-validation error) for both models range from 28.47% to 33.39%, which is tolerable considering some of the fairly broad genre classes, such as 'Electronic'. This class primarily contains songs from different sub-genres of dance music, all of which are distinguishable from each other on the basis of tempo, such as 160bpm for jungle, contrasted with 130bpm for house/techno. This, combined with tempo's low predictor importance estimate for the RF model, would suggest that attributes in the model would have an increase in predictive importance if the dataset contained more specific genre classes.
- The RF algorithm had lower rates in both the test set classification error and the 10-fold cross-validation error, by 4.5% and 3.7% respectively, which advocates RF as a more accurate, and generalisable, modelling algorithm for classifying music genre for this dataset.

	Train Error	Test Error	Val. Error	Prec. (Macro)	Recall (Macro)	F1 (Macro)
MLR	32.10%	33.39%	32.17%	66.22%	66.61%	66.42%
RF	10.14%	28.89%	28.47%	71.06%	71.11%	71.09%

- In contrast to the preferable generalisation error estimates, the RF's large difference in training and testing errors could imply a degree of overfitting. The MLR model achieves more consistent performance on the train and test sets.
- Confusion matrices demonstrate that the genre classes Classical (3) and Hip-Hop (5), were successfully predicted at a significantly higher rate than the other genre classes. On top of this, these two genres showed the most distinction between each other, as no Classical songs were misclassified as Hip-Hop songs, and vice versa, in either of the models.
- The genre class Jazz (6) is seen to have been the least successfully recognised genre in both models' confusion matrices. The MLR model successfully classified less than half, 46%, of songs from this genre class, while the RF model successfully classified only 55.7%, of songs from this genre class.



8.2 Precision, Recall and F1 Score

- The RF model displayed better performance in all 3 performance metrics.
- Micro averages of all 3 measures were 66.61% for RF and 71.11% for MLR.
- Given the context, recall is most likely of higher importance than precision. When using online music distribution services such as Spotify, filtering songs by genre in order to build playlists and discover more music is an important feature and missing songs would be more of a complication than scrolling through some songs that were misclassified as positive for a particular genre. The macro-averages for both model's precision and recall scores suggest that both models behave in a way suited to this contextual obstacle.

8.3 Computation (CPU) Times

- CPU times for training the MLR model were faster than the times for RF, particularly when the model was trained on the 10-fold cross-validation partitions.
- CPU times for testing the MLR model were shorter when both models were tested on the test set; however, when they were both tested on the 10-fold partitions, the RF model took less than half the CPU time compared to our other model.
- The combined CPU times of the training and testing of the models were quicker for MLR for both the train/test sets and the 10-fold partitions. On top of this, the hyperparameter optimisation for the RF model had a very long CPU time as it involved running many RF models.

9 Conclusions

- RF is the more accurate of the two models, reflected by superior scores in generalisation error, precision, recall, and F1 score.
- Broad target classes reduce the predictive capabilities of both models in this context.
- Undefined borders on VIF thresholds means that variables may be removed that diminish the predictive powers of a logistic regression model, thus it is advisable to train the model after filtering according to a number of different VIF thresholds.

9.1 Ideas for the Future

- Ridge or lasso regularisation could be applied to the logistic regression model, the latter of which would remove redundant attributes with little predictive power.
- The RF model can be optimised on the 10-fold cross-validation partitions.
- Choose a dataset containing songs split into more specific genre classes, or even sub-genre classes. This would increase the level of distinction in musical features between each class.
- Attempting to minimise different performance scores, such as recall or F1 score, could yield improved results.
- Trying to maintain a balance between accuracy and computation time may also be worthwhile, due to the rate at which new songs are added to platforms, such as Spotify, and require genre classification. The models could also be retrained as new songs are added, increasing the size of the training set and, thus, possibly improving accuracy.
- The higher levels of accuracy for the RF model suggests that it may be worth trying to fit other non-linear models to this dataset, such as naive Bayes or k-nearest neighbours.