TITLE: Genre Classification of songs using Spotify dataset

Coursework: INM431 Machine Learning Name: Shreenika Aldur Krishnegowda Email: shreenika.aldur-krishnegowda@city.ac.uk

Description and Motivation:

- Classification of Music into various genre based on its underlying features
- Find the best suited classification Method between Logistic Regression and Random Forest on identifying correct genre

Initial Analysis of Dataset:

- The dataset is taken from Kaggle, and it contains 232725 rows divided among 18 columns
- The original dataset contains 18 columns, out of which one is the target(genre) which is a categorical value and 9 are predictors
- The remaining columns are dropped off.
- Figure 1 show the boxplot of the predictors, and we can see ,there are many outliers in the plots for loudness and tempo. Having too many outliers will hinder the performance of the algorithm and cause inconsistency in the accuracy
- The figure 2, Pie Chart shows the top 10 genres with highest number of data. The top 4 amongst them were choose which dominated the pie chart.
- Figure 3, Correlation matrix investigates the relationship between the variables i.e., predictors. There are not many with noticeably higher correlation. Energy and loudness have a strong linear relationship. But it was decided to keep all the predictors purely for learning purpose.

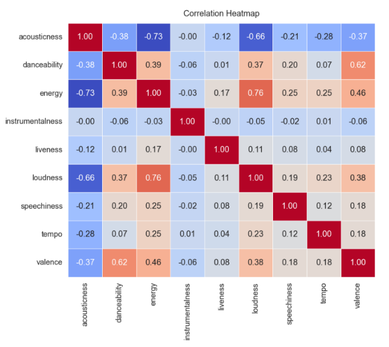

Figure 3

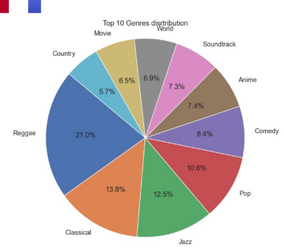

Figure 2

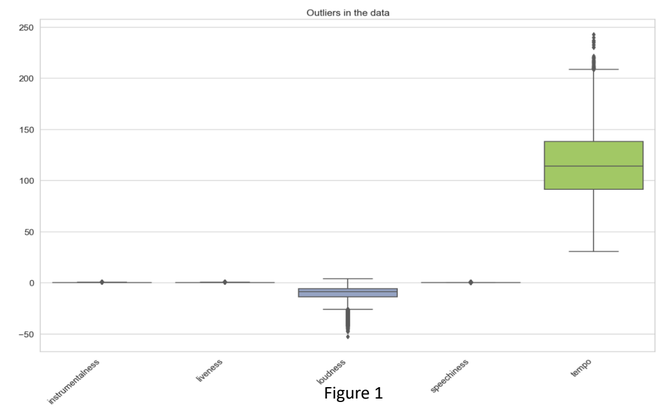

Figure 1

Logistic Regression:

- MLR is applied when we are dealing with multiclass problems
- LR predicts the probability that a instance belongs to a particular class.
- Since logistic regression is a supervised machine learning algorithm and requires a binary target (0,1),it faces problems during multiclass operations.
- LR works effectively against instances in which the target is categorical. In the context of this coursework, LR is used for multiclass operations since this is music genre classification. The model is trained to perform Predictions of each genre for a given collection of features

Pros

- Designed specifically for multiclass classification problem
- Training is more efficient when compared to complex model such as neural networks

Cons

- Similar to BLR it assumes there exist a linear relationship between features.
- It is sensitive to outliers, affecting its performance

Random forest:

- RF is ensemble learning method and it constructs multitude of decision tree during its Training phase
- Each tree makes a prediction based on a random subset of data
- Random forest model is trained using bagging algorithm
- Random predictor selections are made at each split.
- RF is a bagging technique where multiple DT are trained independently on various subsets of the Training set and their predictions are Voted on.

Pros

- RF often provides high accuracy for multiclass classification
- Its nature makes it less prone to overfitting
- It captures the relationship between features and target variables

Cons

- Training and prediction is expensive especially for large datasets
- it can be memory intensive when large number of trees are involved

Hypothesis Statement:

1. Acoustic features such as energy, Instrumentalness and danceability plays a significant role in distinguishing music genre.
2. Random forest takes significantly longer time to classify the genre compared to logistic regression.

Methodology:

- The dataset is divided into Testing set(2/10) and Training set (8/10) .
- Cross validation technique is implied to find best parameters and estimate generalization error on train set.
- Evaluating the models on training and validation set after obtaining the best hyper parameter.
- Best model evaluated against test set and selected
- The model used is adopted for multiclass classification using Multinominal Logistic Regression and “one-vs-all“ coding is used to handle this.
- “fitceoc” function is employed to train multinominal logistic model and handle multiclass classification using svm.
- BO is used to determine the best hyper parameters
- Random forest will specifically optimize 2 parameters ,which are number of trees provided (50-to 150) and minimum number of observations to form a leaf(1-10)
- Best parameter selection is determined by Bayesian optimization.

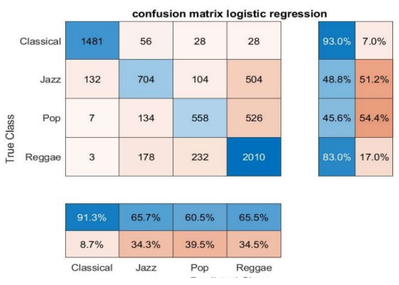

Figure 4

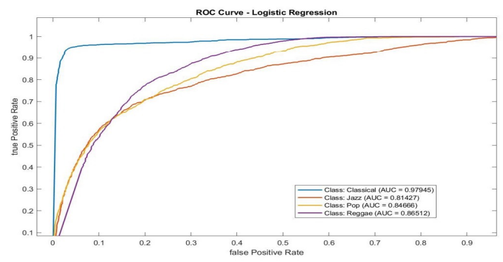

Figure 5

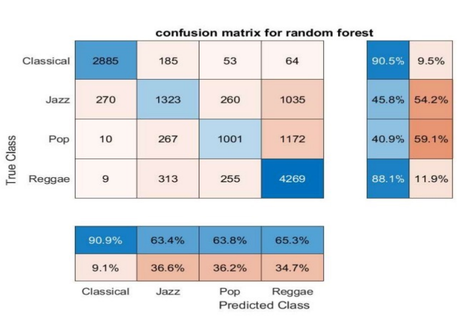

Figure 6

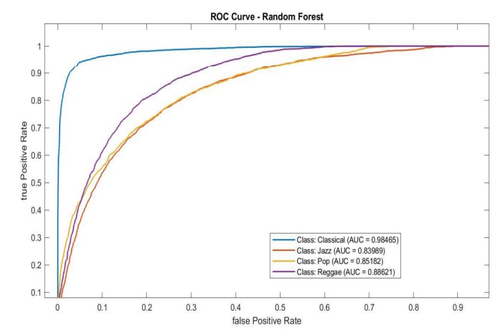

Figure 7

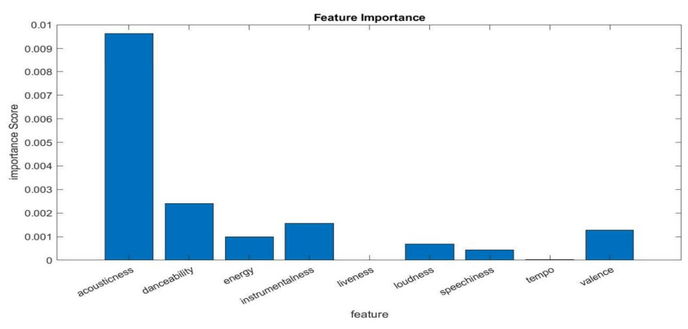

Figure 8

Analysis and critical evaluation of the result:

- Random forest algorithm produces the higher accuracy (6%) more than logistic regression but takes a lot of time of (76.3 minutes)
- Logistic regression is a simple algorithm that works on linear relationships ,whereas random forest is an ensemble of decision trees that captures complex interactions.
- The target variable is taken as a categorical value to fit multinominal logistic regression. Bayesian optimization is implemented to perform auto hyperparameter optimization on multiclass logistic regression model to evaluate the performance and find the best performing model
- The accuracy of training phase was (0.69832) and the model was evaluated on test set yielding the accuracy (0.71099)
- Confusion matrix (figure 4) provides significant insight to the performance of our logistic regression model. As we can see the model has a significant higher level of classification for “classical genre” and performs relatively lower while classifying reggae.
- ROC (figure 5) shows us the AUC values and provides additional insights on the final performance of our logistic regression model. AUC value of classical genre indicates excellent model performance, our model can strongly distinguish between positive and negative instances of classical music. The model faces some challenges in jazz genre but showed good ability to classify pop music. ROC (figure 5) shows Reggae genre performing slightly lower compared to classical but better than Jazz and Pop. however, we can’t ignore the miss classification observed in confusion matrix, this might be due class imbalance.
- Random forest is an ensemble algorithm hence it produces higher accuracy.
- Bayesian optimization is implemented to carry out hyperparameter optimization on total number trees and total number of leaves
- The best hyperparameters where found at iteration number 33.
- Confusion matrix (figure 6) shows that the trained model performs exceptionally well for classical music and struggles a bit in reggae music. it produces a mix of correct classification and mix classification for Jazz and Pop
- ROC (figure 7), as indicated the random forest model shows the exceptional performance in distinguishing positive and negative instances of classical music. The AUC value of Reggae genre shows the second highest performance but as indicated by the Confusion matrix (figure 6) there seems to be a large disparity between correct classification and miss classification, this must be due to data imbalance between the classes.
- Feature importance (figure 8) shows that Accousticness, Danceability and Energy along with Instrumentalness has the higher influence while making predictions. Out of which Accousticness shows the highest influence over genre classification. Hence proving our hypothesis number1.

Lessons learned and future work:

- Data had miss labelling issues which resulted in lower accuracy and improper training ; must perform generalization on multiple labels to reduce the target
- Random forest has a very high number of hyper parameters ,so should be careful while determining the number of trees .
- Bayesian optimization works faster than grid search
- Accuracy of the model doesn’t show the entire picture ,other indications such as Confusion Matrix, ROC, F1 score and Feature Importance Score should be evaluated

Future work

- As seen by feature importance (figure 8) classification highly depends on Accousticness, Danceability and Energy along with Instrumentalness with Accousticness having the most significant impact on classification
- Reducing feature dependency and implementing feature engineering might produce interesting results
- Additional algorithms can be implemented and few more classification data should be collected to improve classification and reduce outliers

References

Li, T., Ogihara, M., & Li, Q. (2003). A comparative study on content-based music genre classification. Multimedia systems, 9(4), 359-366.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction.

McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.