Steps to be followed –

EDA

1. Check the feature density and try to normalize it, using log function. After performing the log function, it was noticed that taking log of instrumentalness, increased the count of null in both train and test. So split the feature into 2 parts one for which log of instrumentalness didn’t yield NaNs and for the ones for which log of instrumentalness yielded NaNs, for them continue using instrumentalness value.
2. Check how the song\_popularity is related to each feature – Finding Relationship
3. Key to be treated as a Catagorical column.
4. Added a Boolean column for each column containing null value, so as to add more insights to our data, to be used by the models.
5. Try to create an Imputer function to impute each of the column containing missing data. Try to build a pipeline as the test data also has nans.

* Imputer tried –

KNN

Iterative Imputer

Model

1. After scaling and imputing the features, we can head to modeling the data.
2. The class is imbalanced, not greatly but 60-40 ratio is there between the 2 class of song popularity. So, try modeling by undersampling and oversampling.
3. Use GridSearchCV method to see which combination of hyperparameter works best for each of the model built.

Models to try out –

1. SVC
2. RandomForestRegressor
3. Stacking using Linear Regression and having base models are SVC RandomForest one more ensemble maybe

Evaluation

According to the competition evaluation page, they are saying ROC(AUC) metrics for evaluation, so we will have to use the same to see which all data are being misclassified and how to improve performance of the system.