CSE 4020 - Machine Learning

J Component

PROFESSOR

Dr. Bhargavi. R

STUDENTS

Srinath NS 20BCE1074
Aakash R 20BCE1003
Subramanian N 20BCE1019

Hit Song Prediction

MOTIVE

Goal: To determine whether the audio quality and lyrics of a song can accurately predict its popularity.

Dataset: The Spotify Hit Predictor Dataset (1960-2019) | Kaggle
The audio quality and lyrical notes of the song are captured in terms of multiple parameters. [41106 Instances x 22 Features]

Methodology:

- Remove Unwanted Features
- Fit Base Models
- Tune Hyper Parameters
- Ensemble Learning
- Conclude the most reliable model

Methodology

Step 1:

Run Different Models on complete dataset, with Hyper parameter tuning (using cross validation) [KNN, Naive Bayes, Logistic, SVM, DT, RF]

Step 2:

Perform Feature Selection using those models [RF, Boruta (RF), RFECV (Logistic)

Step 3:

Run the models used in **step 1** with reduced features and conclude the reliability using Cross validation.

Step 4:

Identify the best models using ROC and AUC metric

Step 5:

Using the top models identified from **step 4** , build ensemble classifiers (Voting [Hard, Soft] and Stacked [AdaBoost, GradientBoost] and XGBoost) .

Step 6:

Identify the Best performing Ensemble model and compare with the best performing Individual model and conclude the final model

Step 1 - Fitting Base Models without Feature Selection

Different Models viz., Logistic Regression, SVM, Decision Tree Classifier and Random forest Classifier were fit and tuned using Cross Validation

Logistic

Cross Validation Score**:** 0.728

Best Parameters: C=0.1, max_iter=10000, penalty='l1', random_state=42,solver='saga'

Decision Tree

Best CCP Value = 0.001

Cross Validation Score = 0.74

Best Parameters: random_state=42,ccp_alpha=0.001

SVM

Cross Validation Score: 0.7704

Best Parameters: C=10, gamma=0.1, kernel="rbf"

Random Forest

Train Score: 0.785

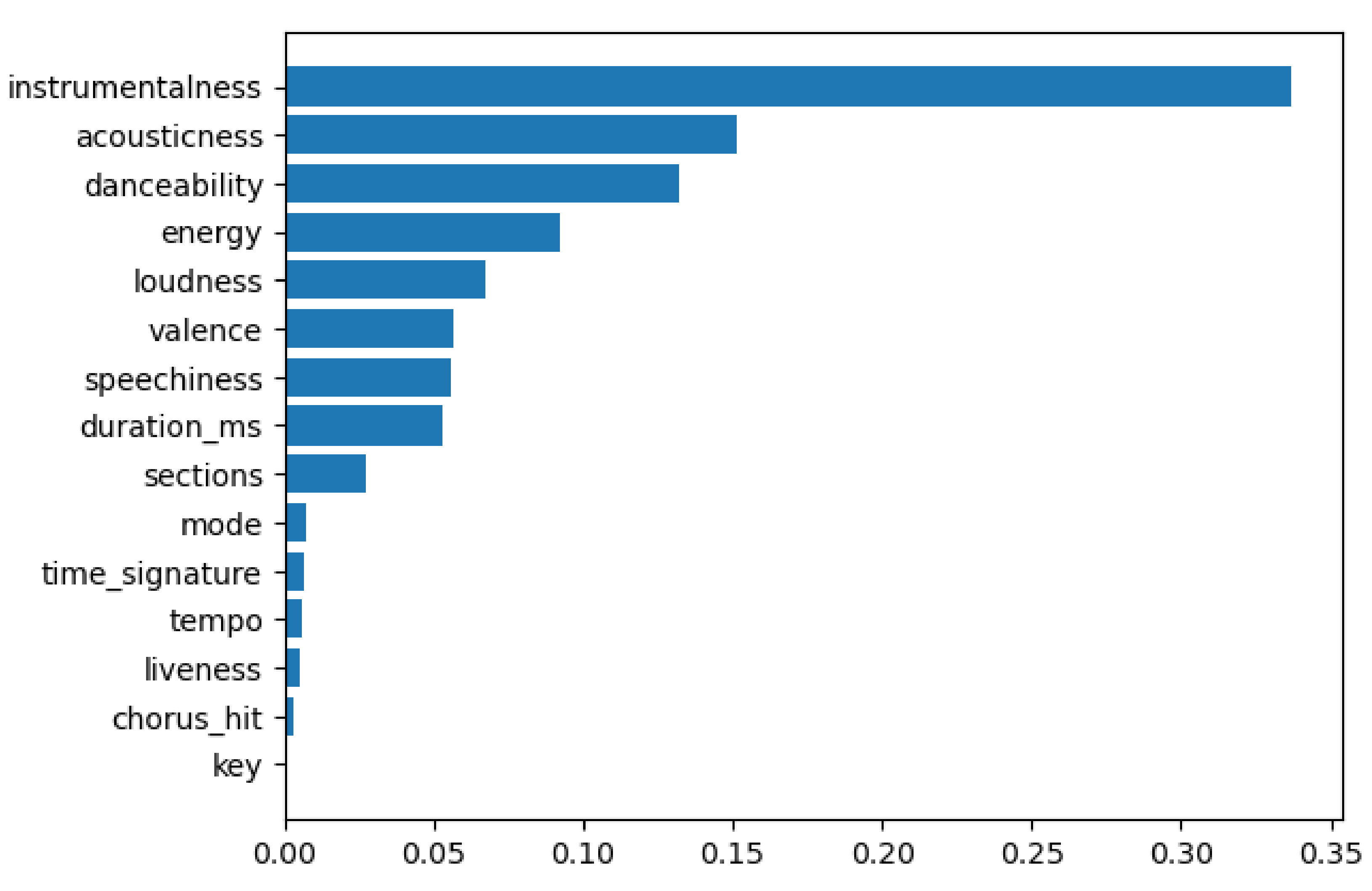
Test Score: 0.775

Best parameters: n_estimators=100, max_depth=100,max_leaf_nodes=500,min_samples_leaf=5,random_state=42,ccp_alpha=0.0004

Step 2 - Feature Selection

The Previously built Random Forest is used for Feature Selection. On top of that, Boruta Method and RFECV with the previosuly built Logistic Regression Model are used. The Best features are consolidated

Random Forest



Boruta Py

Feature suggested to be Removed: Key

RFECV with Logistic

Optimal Number of Features: 11

Features Suggested to be Removed: Speechiness, Liveness, Tempo, Duration_ms

Consolidation

Removed: key, chorus_hit, liveness, tempo, time_signature, model

(weightage given to RF owing to its accuracy)

Step 3 - Fitting Base Models with Feature Selection

K-NN

Best Parameters: 'algorithm': 'ball_tree', 'n_neighbors': 19, 'p': 1, 'weights': 'distance'

Cross Validation Score: 0.76

Naive Bayes:

Best Model: Gaussian

Best Parameters: 'var_smoothing': 1e-09

Cross Validation Score: 0.71

Logistic:

Best Parameters: C=0.1, max_iter=10000, random_state=42

Cross Validation Score: 0.72

Support Vector Machine:

Best Parameters: 'C': 1.5, 'class_weight': 'balanced', 'gamma': 'scale', 'kernel': 'rbf'

Cross Validation Score: 0.77

Decision Tree:

Best CCP Alpha Value: 0.001

Cross Validation Score: 0.74

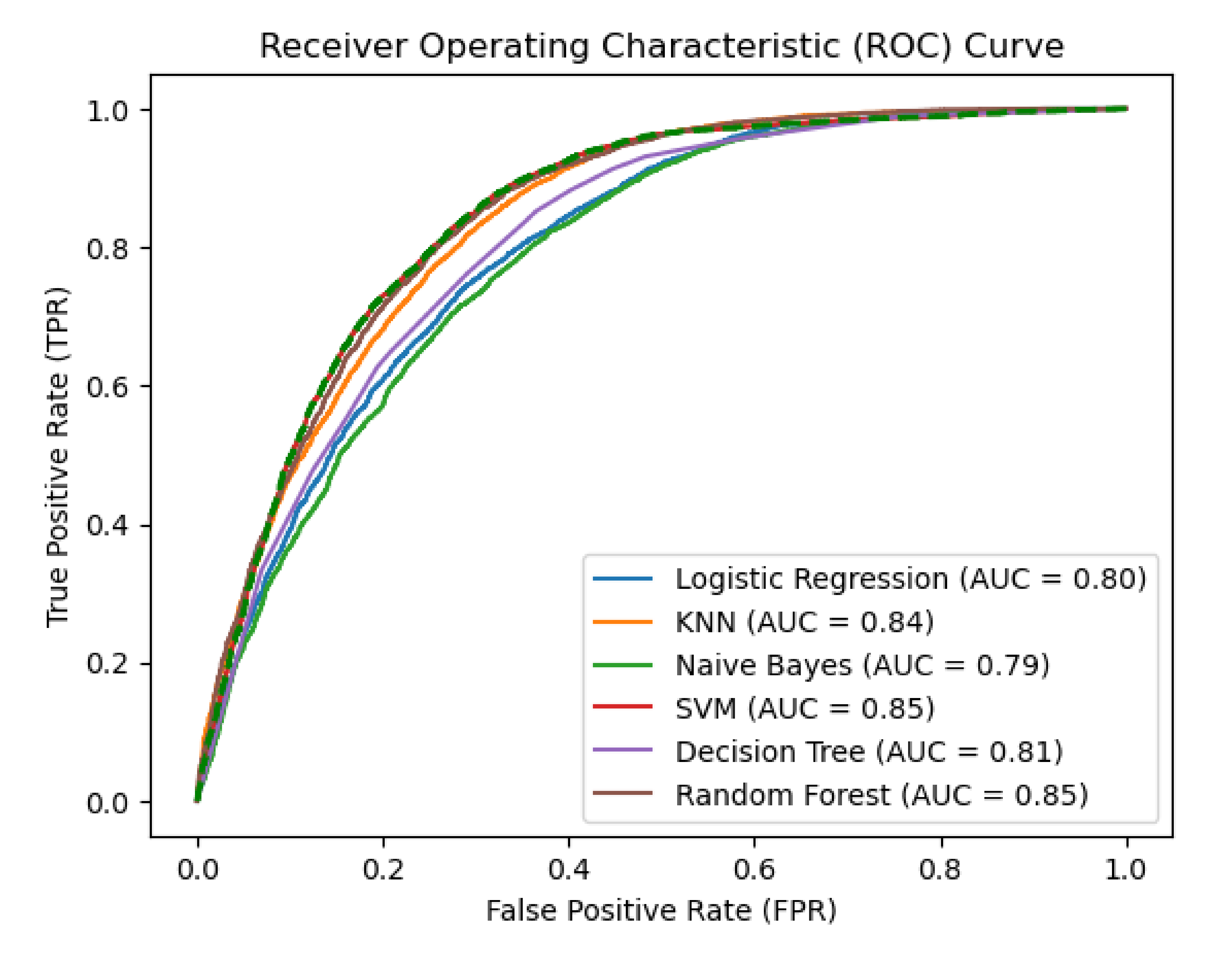
Random Forest:

Best Parameters (wrt OOB Score): max_depth=30, max_features='sqrt', n_estimators=300

Best CCP Alpha Value: 0.0004

OOB Score 0.77

Step 4 - Find Reliable models using ROC and AUC metric



Best Model(s):

A model with a higher AUC value is generally considered to be better at discriminating between positive and negative instances.

The top reliable models are:

- Random Forest
- SVM
- K-NN

These models will be used as base models for ensemble learning.

Step 5.1 - Ensemble Learning - Voting and Stacking

In a voting classifier, multiple base classifiers are trained independently on the same training data, and their predictions are combined using a majority voting scheme to make the final prediction. In a stacking classifier, the base classifiers are trained on the same training data, and their predictions are combined using a meta-classifier that learns to combine the base classifiers' predictions.

Voting - Hard Voting

- In hard voting, the predicted class label with the highest frequency is selected as the final prediction
- Cross Validation Score = 0.77

Voting - Soft Voting

- In soft voting, the predicted class label with the highest average probability across all the base classifiers is selected as the final prediction.
- Cross Validation Score = 0.78

Stacking w/ AdaBoost

Cross Validation Score = 0.78

Stacking w/ GradientBoost

Cross Validation Score = 0.70

Step 5.2 XGBoost

Best Parameters = {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 150}

Cross Validation Score = 0.77

I Step 6 - Concluding the best Model

Best Ensemble Model

Voting (soft) and Stacked (Ada Boost)

Cross Validation Score: 0.78

Best Base Model

SVM (RBF Kernel)

Cross Validation Score: 0.77

Best Overall Model

Voting (soft)

Preferred over adaboost owing to less computation cost

Cross Validation Score: 0.78

CSE 4020 - Machine Learning

J Component

PROFESSOR

Dr. Bhargavi. R

STUDENTS

Srinath NS 20BCE1074
Aakash R 20BCE1003
Subramanian N 20BCE1019

Thank you