NC STATE UNIVERSITY
Department Of Computer Science

Hit Song Prediction

Shrey Anand, Ansul Atriek, Andrew Emerson, Amanul Haque,
Suvodeep Majumder, Vidhisha Jaswani

1. Introduction

1. What makes a song a hit?
2. Predict hit or not using the **acoustic properties** of the song.
3. **Data features**: acousticness, duration, danceability, energy, key.

2. Data Description

1. **Data Collection:** 3 Phases with 10,000 instances and **Spotify** API to collect features.

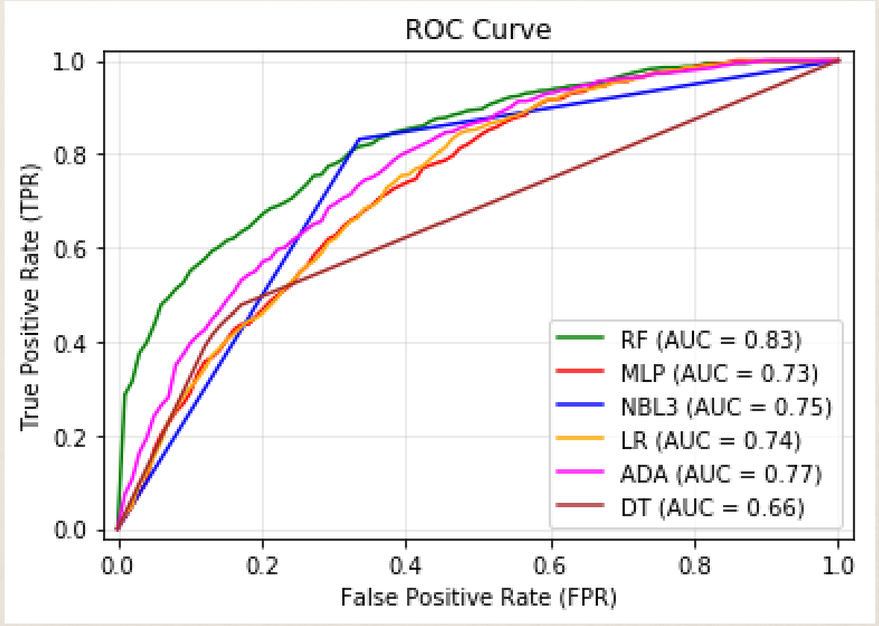
2. **Data Pre-processing:** Normalization, Imputation, removing duplicate data and feature selection.

3. Technical Section

The algorithm from built-in libraries are:
1. SVM
2. Naive Bayes
3. Logistic Regression
4. Decision Trees
5. Random Forests

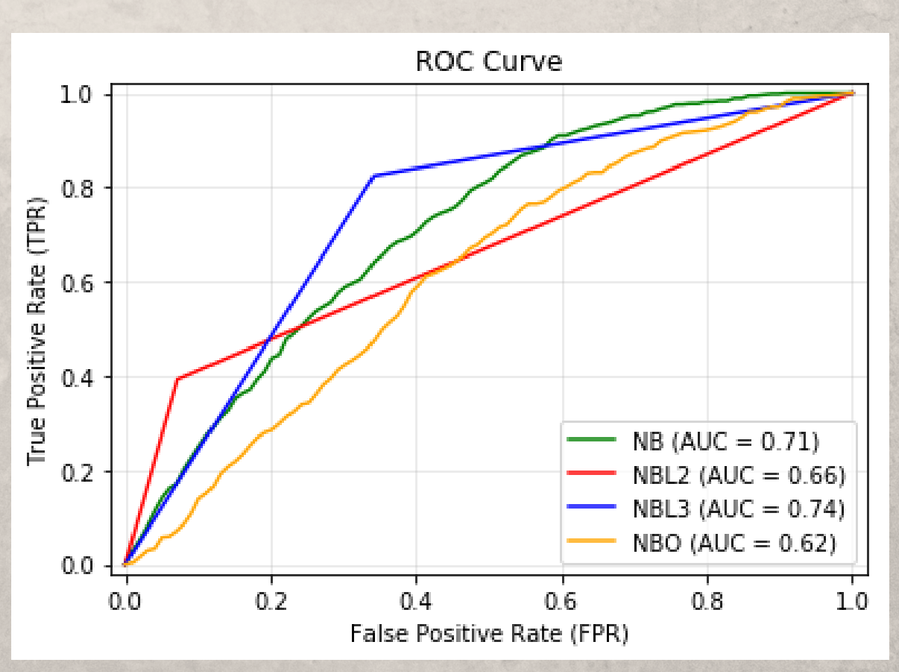
Implemented the following algorithms from scratch:
1. Neural Network
2. Naive Bayes
3. Multilayer tree based Naive Bayes using error correction.

4. Results



Learner	Accuracy	F-measure	AUC
Naïve Bayes	0.697	0.689	0.691
Naïve Bayes*+	0.575	0.621	0.617
2-Layer Naïve Bayes*	0.822	0.800	0.622
3-Layer Naïve Bayes*	0.693	0.727	0.744
Neural Network+	0.814	0.735	0.734
Neural Network*+	0.802	0.770	0.606
Decision Tree	0.767	0.770	0.628
SVM	0.816	0.733	0.552
Logistic Regression	0.816	0.733	0.707
AdaBoost	0.814	0.752	0.714
Random Forest+	**0.860**	**0.831**	**0.835**

*Implementation from scratch
+Implemented without Feature Selection.



5. Parameter Choices

Naive Bayes

1. priors: prior probabilities of the classes.

Neural Network

1. hidden_layer_sizes: ith element represents the number of neurons in ith hidden layer.
2. activation: activation function for the hidden layer.
3. learning_rate: learning rate schedule for the weight updates.
4. max_iter: maximum number of iterations.

6. Conclusions

1. The **feature selection** only marginally affected our results.
2. Our implementations of Naïve Bayes and Neural Network **performed fairly** against those implemented in sklearn package.
3. The **3-layer Naïve Bayes** method performs better than single-layer Naïve Bayes.
4. SMOTE was also used to solve the class imbalance problem, but was found to not affect the overall results.
5. Overall, **Random Forest** performs the best.

Reference: Herremans, Martens, and S¨orensen, Dance hit song prediction. Journal of New Music Research (2014)