Predicting Hit Songs using Spotify and Billboard Data

॥ सा विद्या या विमुक्तये ॥

भारतीय प्रौद्योगिकी संस्थान धारवाड

Indian Institute of Technology Dharwad

Adhokshaja V Madhwaraj, 160010032
Raaj Tambe, 160010034
Venkata Kowsik T, 160010035

ABSTRACT

The Hit Song Science (HSS) problem aims at using Machine Learning and Predictive Analysis, to help artists and producers predict a song's success. This problem was approached using Spotify Web API and the Billboard 100 charts' historical data to predict whether a song is a hit or not. A dataset of 73,000 songs was constructed from the above-mentioned sources (since 2010 only). Five classification algorithms were tested with subtle modifications and best-performing algorithms were SVM-rbf (77.69%) and Random Forest (86.68%).

INTRODUCTION

Every week, Billboard releases the Top 100 chart, and this chart is one of the most definitive ways to measure the success of a song. We have used Machine Learning classification algorithms to label songs as hits or not, using audio feature data obtained from Spotify Web API. The dataset has 73,000 songs with 8,000 positive samples, so every iteration of an algorithm involved randomly sampling an equal number of negative samples to that of positive ones, to prevent skew. We have used feature importance and selection in order to give producers and artists key insights on the most important features that lead to featuring on the chart. Producers and labels can concentrate their capital on the marketing and publicity of the songs that are predicted as hits.

MODEL

PREPROCESSING

- The entire Spotify dataset was obtained from the Spotify Web API. Song titles were obtained from Billboard Hot 100 chart, and tagged based on this. Data consists of 11 features. An equal number of +/- examples are chosen from both buckets to form sample data.
- Features with highest correlation selected based on Correlation Heatmap, conjunctive feature creation by product is used on these features to get more features. Highly correlated features were Energy, Valence, Danceability, Loudness, and Speechiness. A total of 7 features were added to the primitive features to get a total of 18 features
- Extremely random trees are used with multiple features as root nodes. Algorithm for feature importance iterates through the data to learn weights corresponding to the importance of the feature (Artist popularity – most important feature)

TRAINING/TESTING

- Sampled Data split into training/testing sets. Models (Logistic Regression, SVM-rbf, SVM-poly, Naïve Bayes, Random Forest) were used for classifying.
- Accuracy and f1 Scores were calculated for each model and the results were averaged over 20 iterations

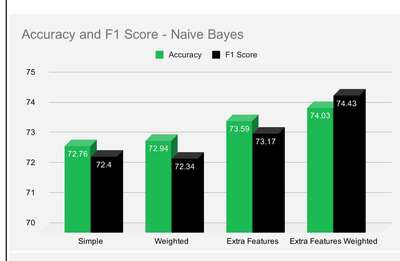
$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

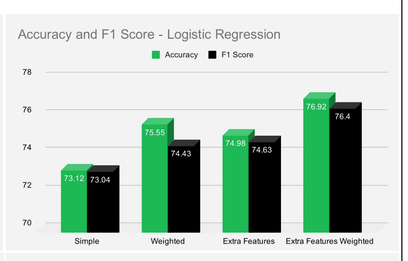
$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

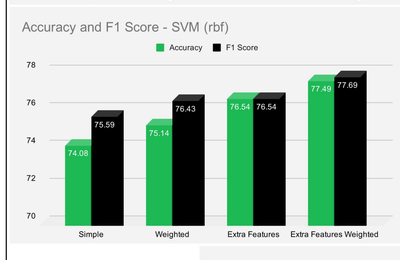
RESULTS

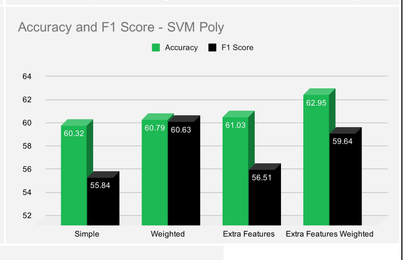
		Accuracy	F1 Score
Logistic Reg.	Simple	73.12	73.04
	Weighted	75.55	74.43
	Extra Features	74.98	74.63
	Extra Ft Wt	76.92	76.4
SVM rbf	Simple	74.08	75.59
	Weighted	75.14	76.43
	Extra Features	76.54	76.54
	Extra Ft Wt	77.49	77.69
SVM poly	Simple	60.32	55.84
	Weighted	60.79	60.63
	Extra Features	61.03	56.51
	Extra Ft Wt	62.95	59.64
Naive Bayes	Simple	72.76	72.4
	Weighted	72.94	72.34
	Extra Features	73.59	73.17
	Extra Ft Wt	74.03	74.43
Random Forest	Simple	85.33	85.33
	Weighted	85.43	85.43
	Extra Features	86.58	86.57
	Extra Ft Wt	86.68	86.68

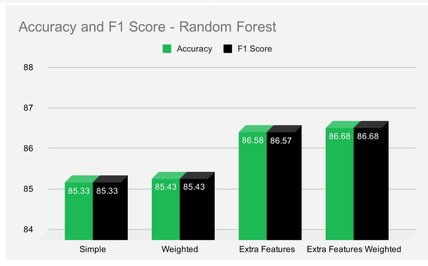
GRAPHS





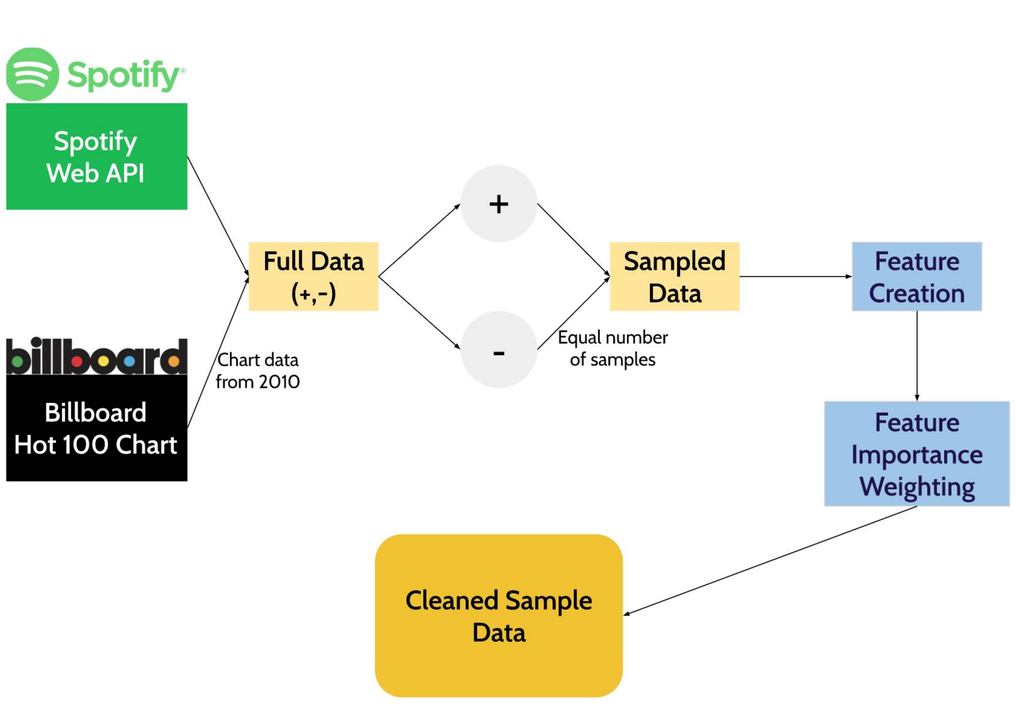


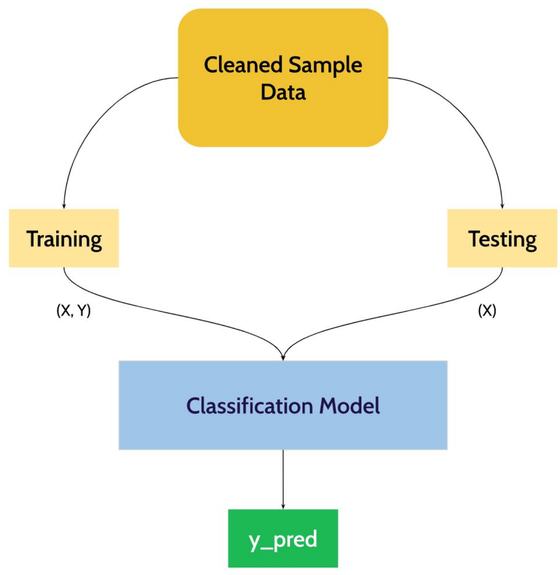




BLOCK DIAGRAM

DATA PREPROCESSING & MODEL





CONCLUSION

High accuracy in the prediction of hit songs was achieved by the system - 86.68% (Random Forest) and 77.49% (SVM rbf). On adding additional modification of extra feature creation, we observed an increase in accuracy. Also, by using Feature Importance weighting, an increase in accuracy was observed. This observation is supported by the fact that irrelevant or partially relevant features negatively impact model performance.

REFERENCES

[3] Efficient Classification With Conjunctive Features
-www.datascience.com
-www.kaggle.com
-www.developer.spotify.com

We would like to heartily thank Prof. Mahadeva Prasanna for mentoring us throughout this project. We are also grateful to Prof Bharath B.N and Prof. Naveen M.B for giving us significant insights.