Identifying Hit Songs using Machine Learning

By: Prathap Rajaraman

Spotify

Background

- Motivation
  - Hit songs can make an artist’s career and make millions for record labels.
- Objective
  - A data driven classification approach to determine what makes a song popular
  - Explore relationship between audio features and popularity

Spotify

Data & Methodology

- Data
  - X variables: Audio Features for over 34,000 songs on Spotify dating back to 1985.
  - Y variable: Binary indicator of whether or not song ever appeared on Spotify 100 or Billboard 100

- Methodology
  - Classification metrics ranging from Logistic Regression to Ensembling Algorithms considered

Modeling

- F1 and ROC-AUC analyzed
- Classification Algorithms Modeled:
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - ADA Boost
  - XG Boost
- GridSearchCV and SkLearn used to tune hyperparameters and probability thresholds

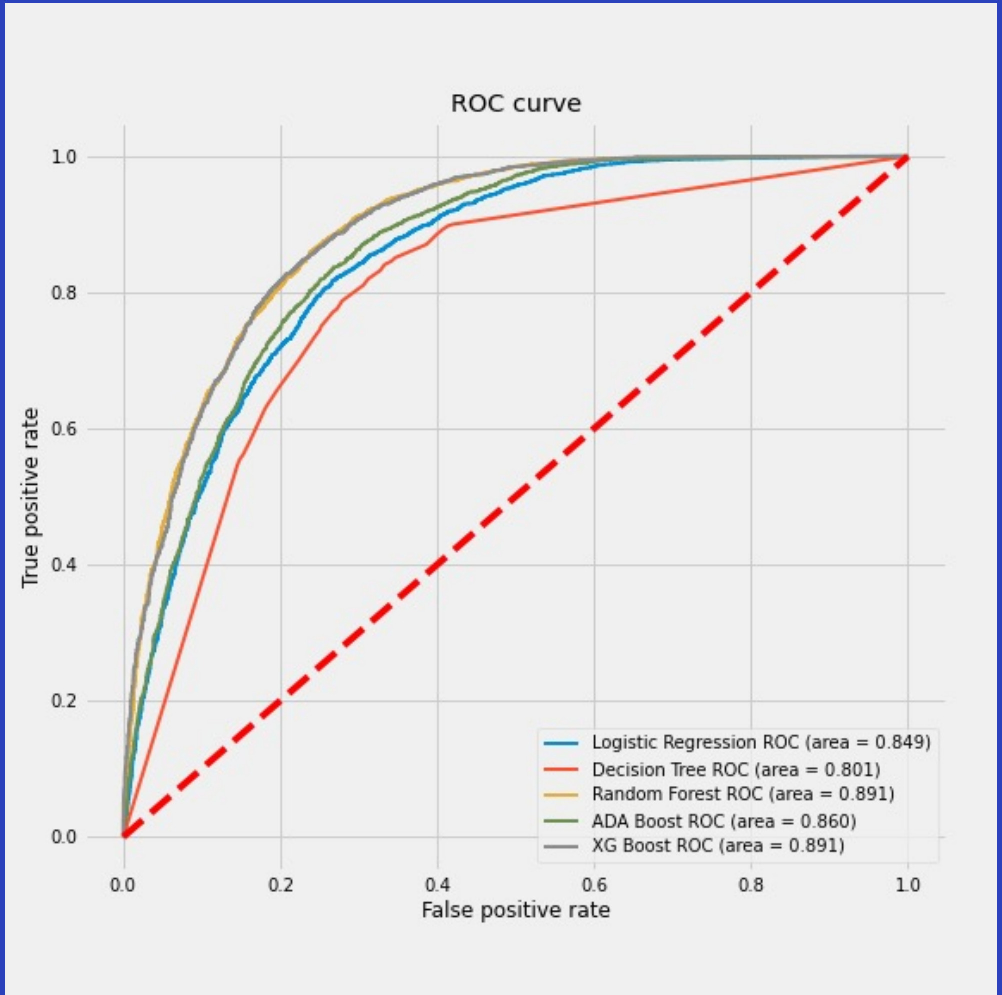
Feature Engineering & Variable Selection

- Audio Features*
  - Duration (in m)
  - Energy
  - Acousticness
  - Instrumentalness
  - Valence
  - Speechiness
  - Loudness (in dB)
- Feature Engineered Variables
  - 4 Beats per Measure Indicator
  - Loudness below -10 dB Indicator
  - Energy x Dance interaction Variable

*Values range from 0 to 1 unless specified otherwise

Results

- XG Boost Won
- Similar scores as Random Forest but without overfitting
- 4% F1 and 4.5% ROC-AUC Boost over Logistic Regression



Model	Train F1	Test F1
Logistic	0.792	0.788
Decision Tree	0.868	0.777
Random Forest	**0.919**	**0.826**
ADA Boost	0.805	0.803
XG Boost	**0.845**	**0.824**

Algorithm in Action

Artist	Track	Energy	Valence	EnergyxDance	Prediction	Actual
Travis Scott	Goosebumps	0.593	0.808	0.499	Hit	Hit
Drake	Controlla	0.476	0.347	0.289	Non-Hit	Hit

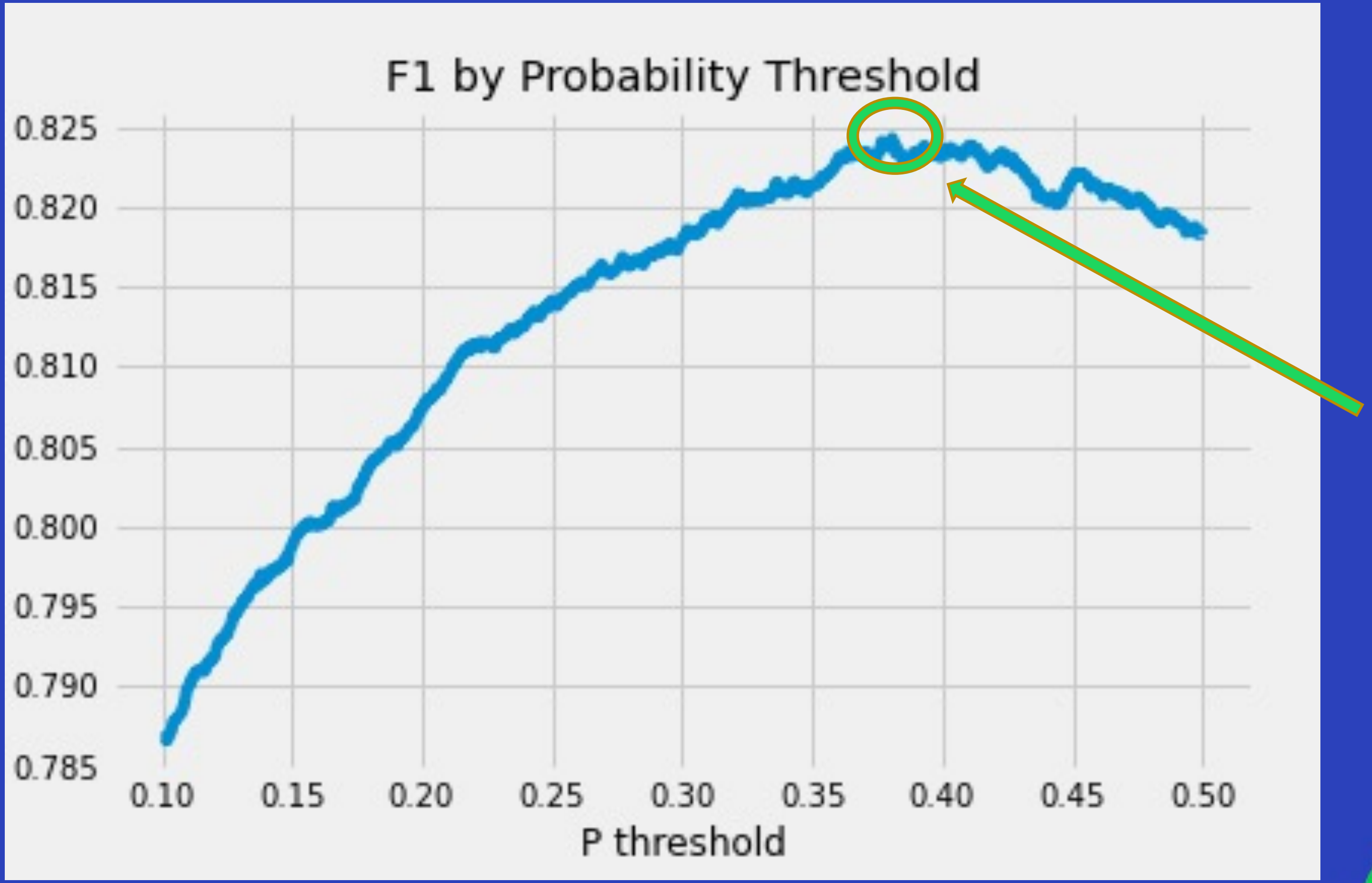
The algorithm correctly applies audio features,
but misses out on artist reputation

Conclusion

- The XGBoost algorithm is the best performing model
- With classification modeling, labels and producers are better equipped to know what constitutes a hit song
- Future Work:
  - Extract dataset that represents the true distribution of popular songs
  - Incorporate social media presence and artist popularity as an additional variable

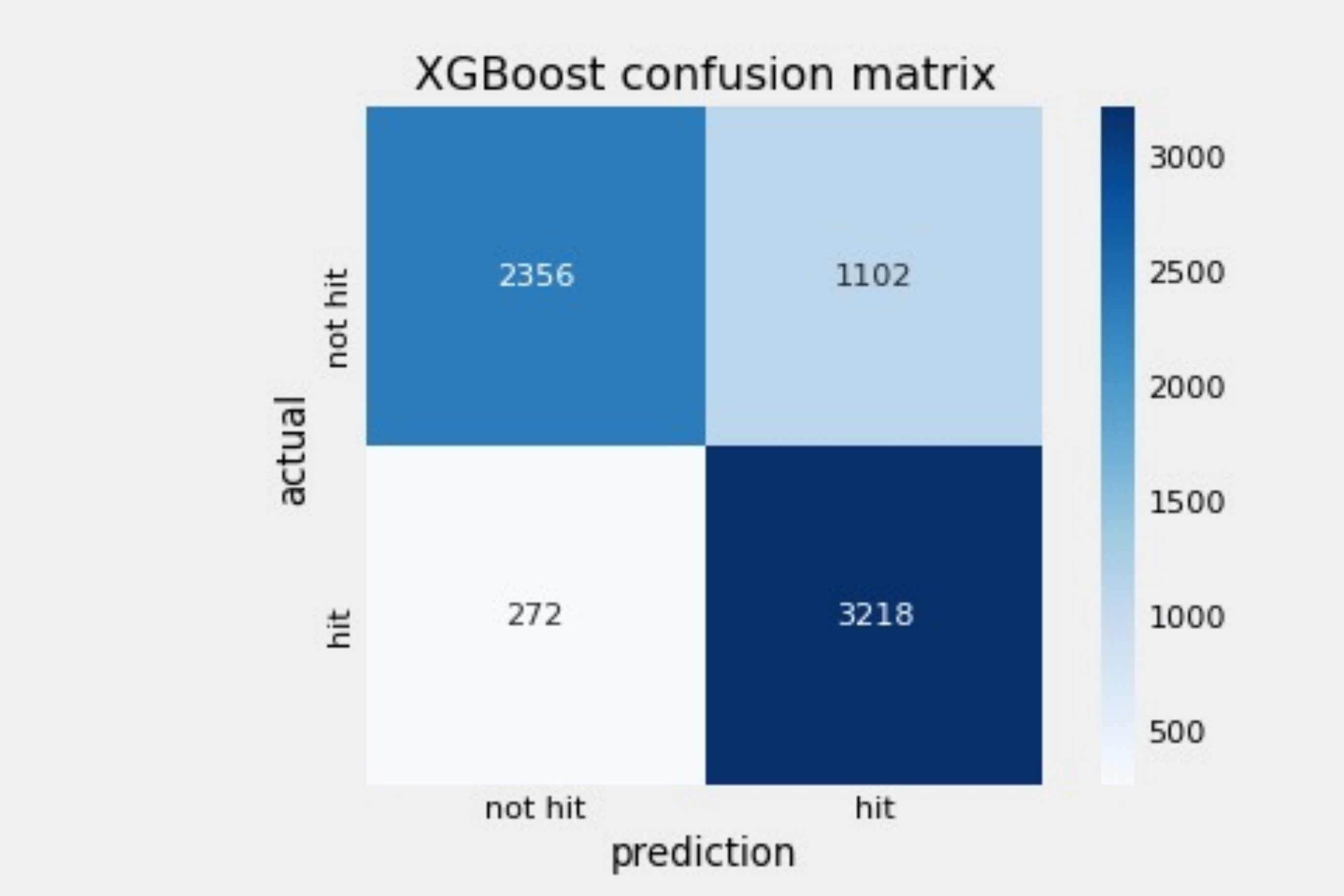
Appendix

Probability Threshold



A threshold of 0.381 yields the highest F1 score for XGBoost

Confusion Matrix


XGBoost confusion matrix
actual
not hit
hit
prediction
not hit
hit