PREDICTING A SONG’S POPULARITY

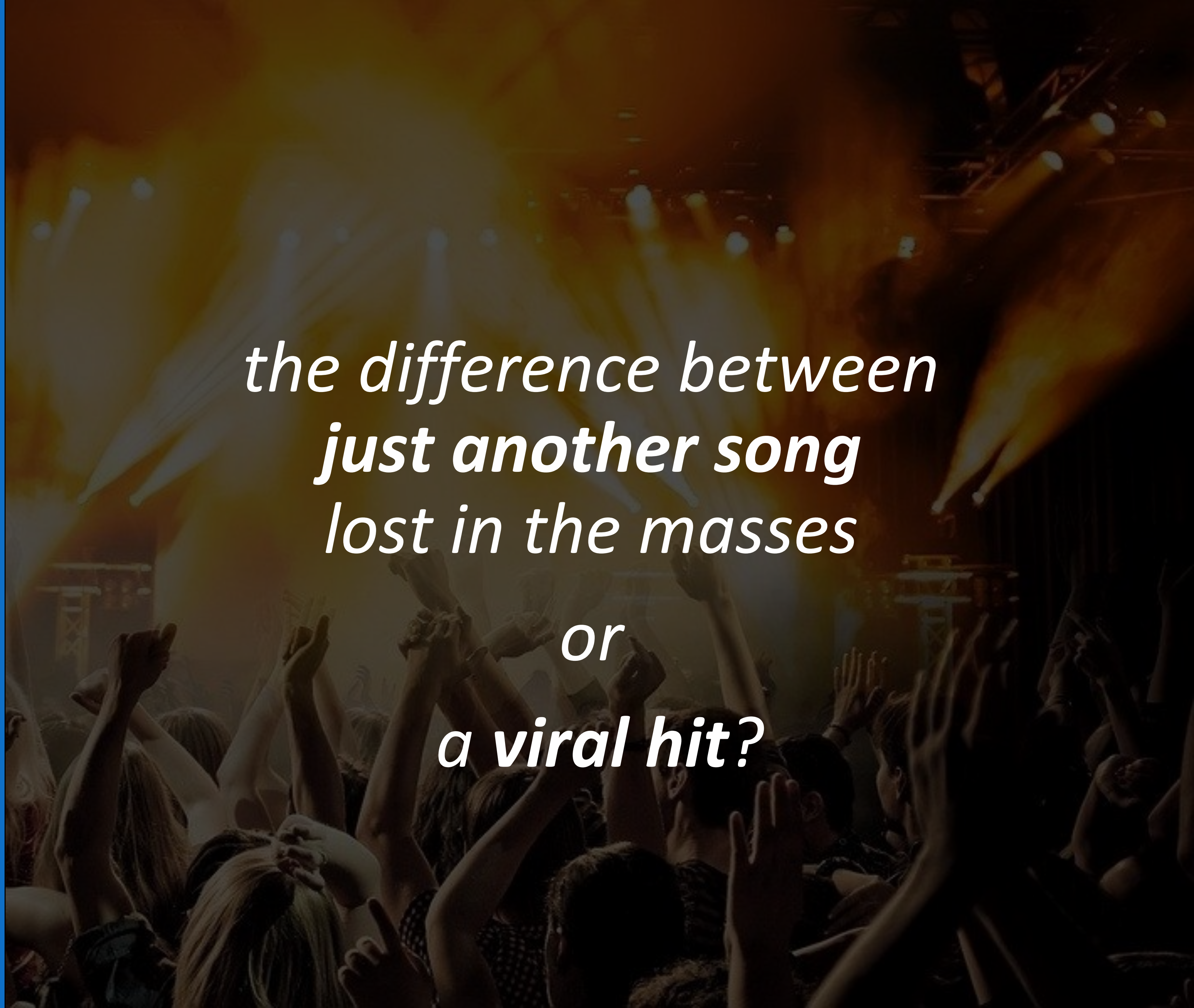
DAME JANKULOSKI - KRISTIN SKRITEK - NISHANTA KHANAL
VICTOR FONSECA - ZONAIR NADEEM

Famous:

Now as simple as a click, listen, and share

What if there was a way…

to give individuals, groups, and record labels a head start on what could be



the difference between ***just another song*** *lost in the masses*

or

a ***viral hit****?*

Agenda

1. Problem Statement
2. Dataset
3. Models
4. Results
5. Discussion

Problem Statement

Can the success of a song, based on making the cut for Billboard’s Top 100 Charts, be predicted?

DATASET CREATION AND DESCRIPTION

Spotify

Used Python library *Spotipy* to collect data on a selection of songs:

- Songs that placed on Billboard Top 100 (2010-2020)
- Songs that did not place on Billboard Top 100 (2010-2020)

Includes: Music features, Spotify engineered song features and musical genres

billboard

Used *BeautifulSoup* to scrape Billboard Top 100 charts in Wikipedia

Years: 2010-2020

Includes Ranking, song name, artist name

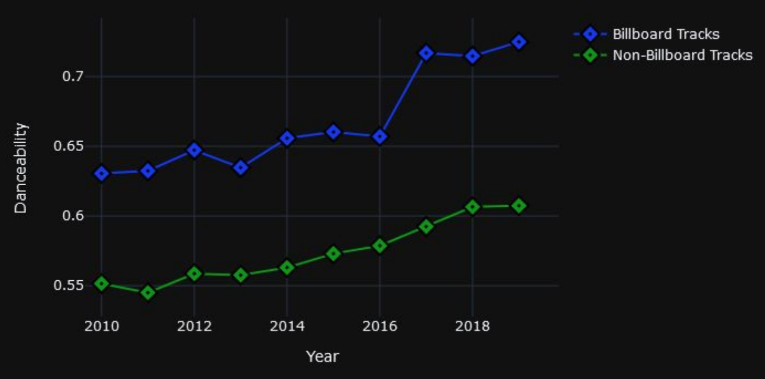
Final Table	
Rows	360,372
Features	46
Label	Billboard vs Not Billboard

Dataset Preparation

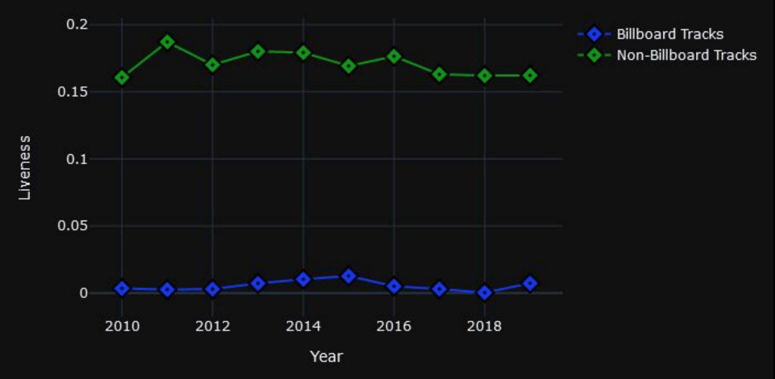
- Exclude tracks that did not correspond to songs
    - podcasts
    - readings
    - sound effects
    - environmental / background
    - drama
    - ASMR
- Exclude classical music
- Exclude songs with no features (missing values)
- Release date parts - day, month, year, week
- Flag - Artist has ever been on billboard
- Flags - 20 top musical genres

Data Exploration

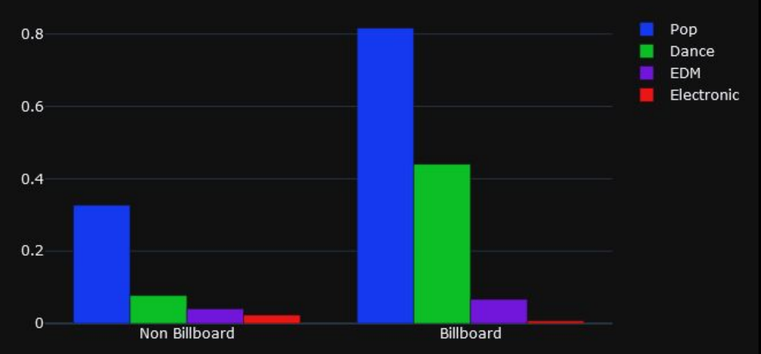
Danceability by Year



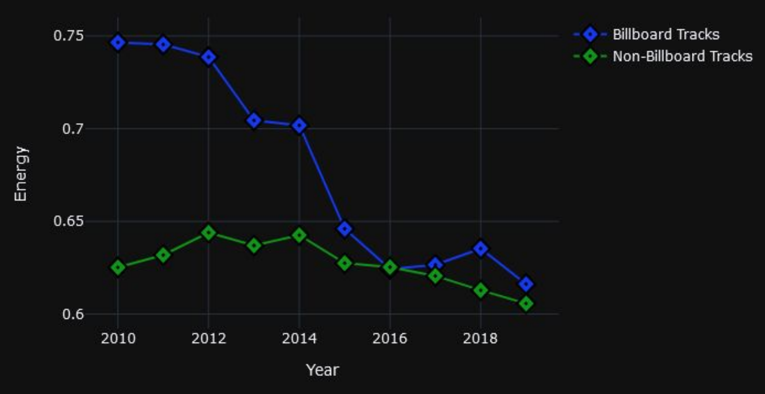
Instrumentalness by Year



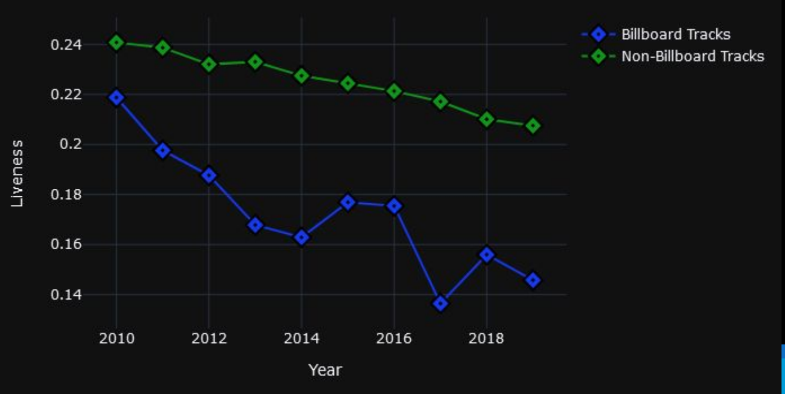
Percentage of Electronic and Pop Music on Billboard



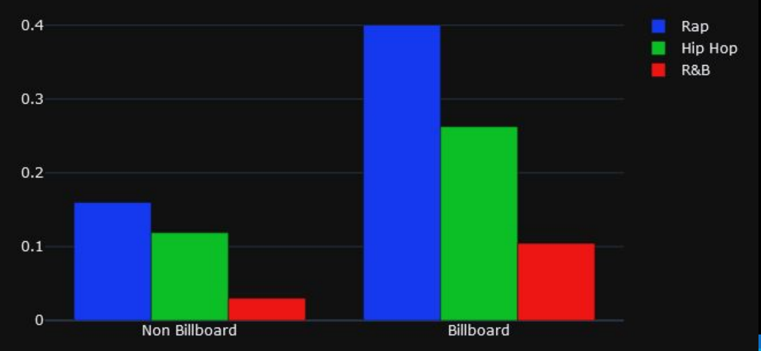
Energy by Year



Liveness by Year



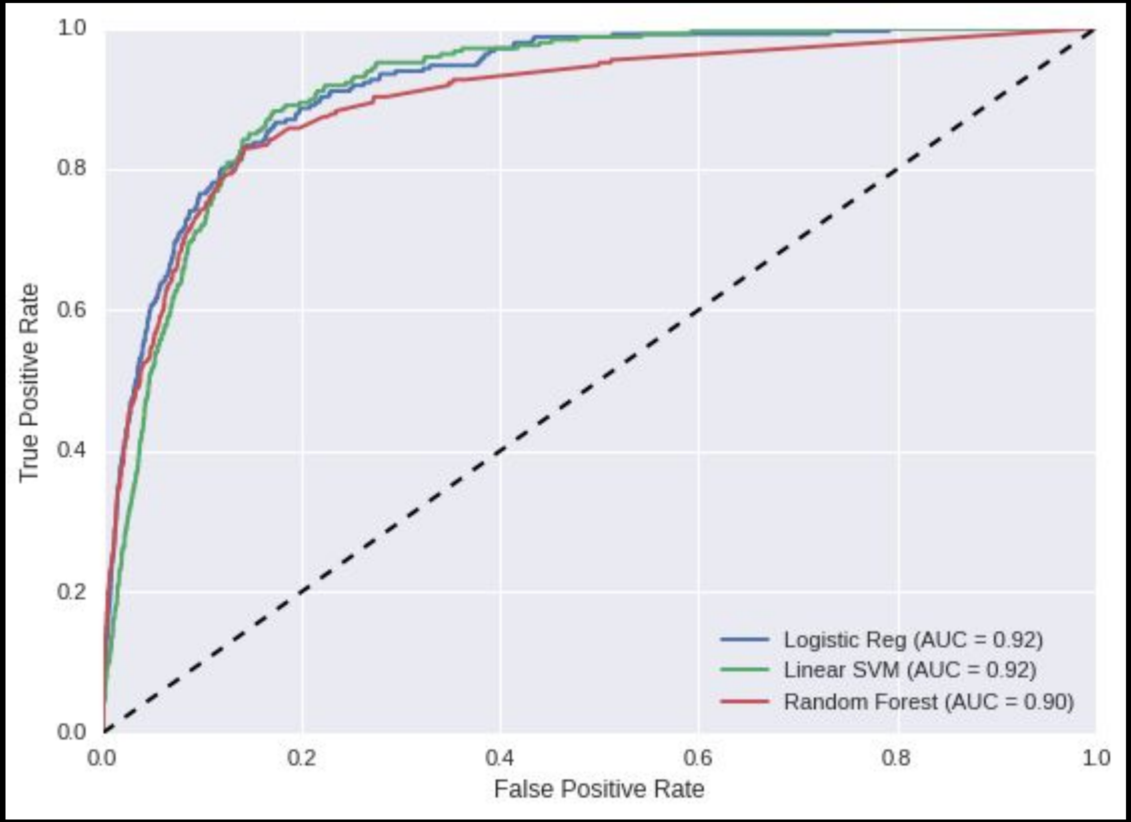
Percentage of Hip Hop, Rap and R&B on Billboard



Models Tested

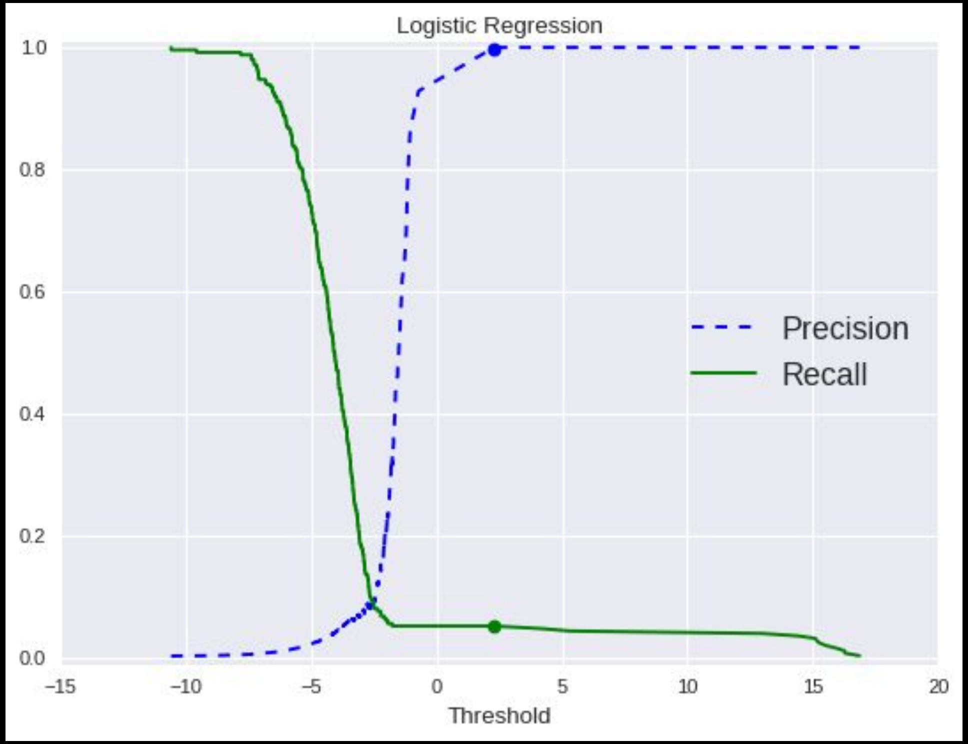
Model	GridSearchCV optimization	Hyper Parameters	ROC AUC	Accuracy	Precision	Recall	Confusion Matrix	
Logistic Regression	Accuracy	C - Inverse of reg, Class_weight, Penalty (l1, l2, elasticnet)	0.92	0.9997	1	0.05	86,104	0
							234	13
SVM	Recall	Kernel, Duality, Loss function, Penalty (l1, l2), C, Class_weight	0.92	0.7303	0.01	0.94	66,013	23,925
							15	232
Random Forest	Accuracy	n_estimators, max_depth, max_features, bootstrap, max_samples_leaf, max_samples_split	0.90	0.7286	0.01	0.96	62,284	23,280
							9	238

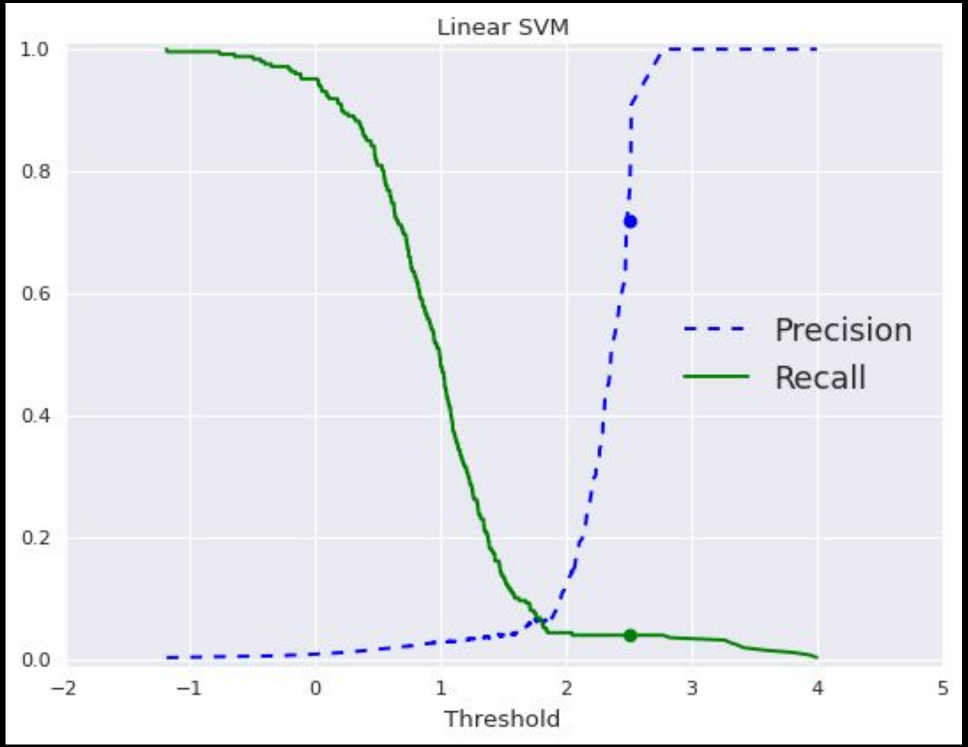
Model Comparisons



- Logistic Regression and Random Forest ROC curves grow faster (higher TPR with lower FPR)
- Linear SVM yields the highest AUC: best performance at distinguishing billboard songs from non billboard
- The three models perform well, way above the diagonal

Precision vs Recall





Discussion

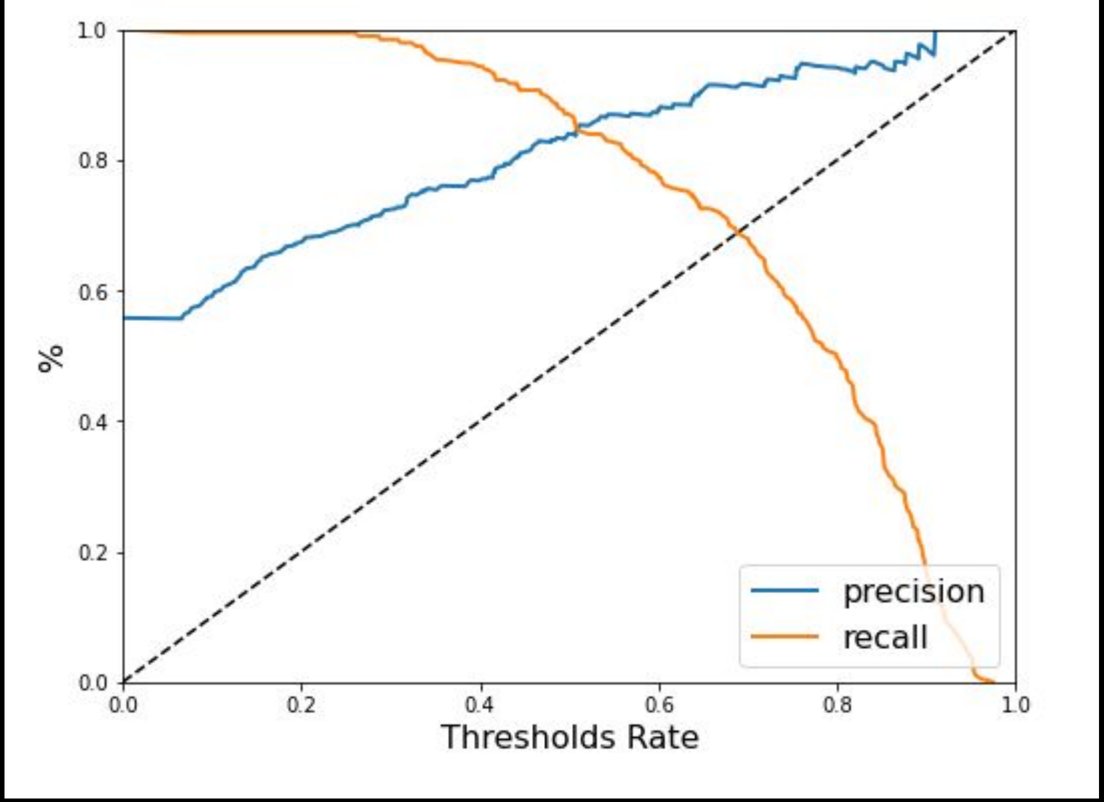
Points of attention

- How to measure and rank songs popularity
  - Binary vs Continuous variable
  - Billboard vs Spotify
- Balancing the dataset
  - Proportion of songs that go to billboard - rare event
- Trade-off between precision and recall

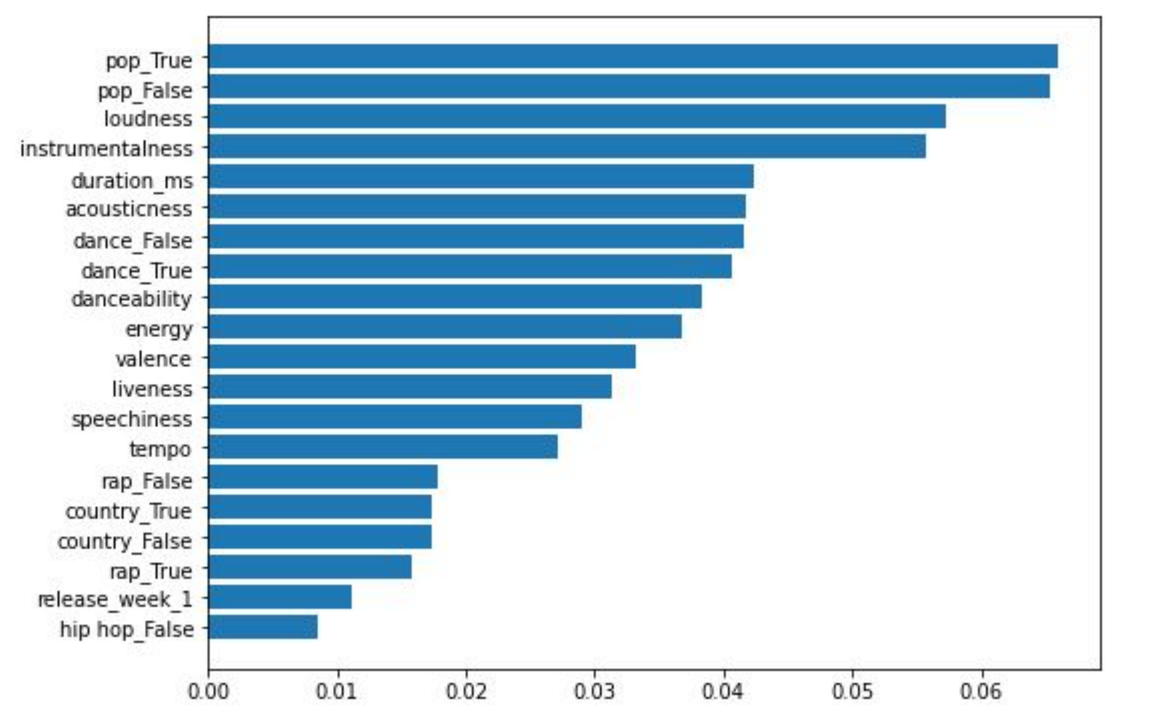
Questions?

Appendix

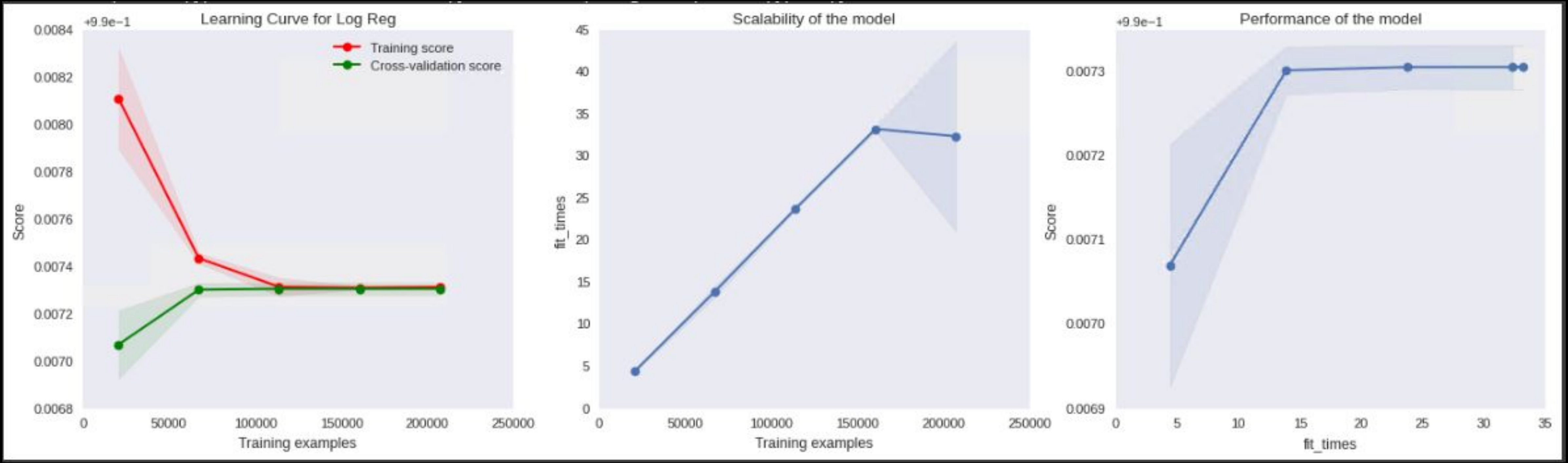
Balanced Dataset Training



Feature Importance Analysis



Learning Curve for LR


Learning Curve for Log Reg
+9.9e−1
Score
Training score
Cross-validation score
Training examples
Scalability of the model
fit_times
Training examples
Performance of the model
+9.9e−1
Score
fit_times