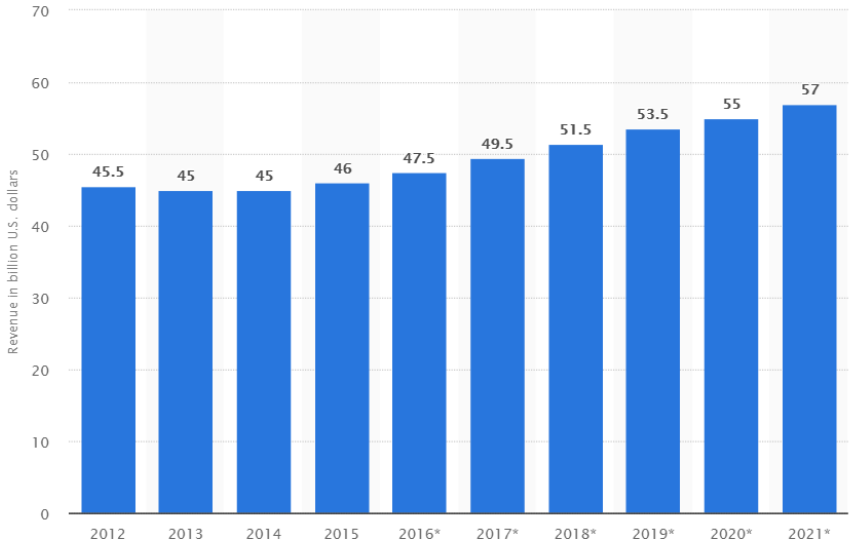
Predicting Song Popularity

By : Amninder Dhillon

Background

- Music integral part of our daily lives
- 50 Billion dollar industry
- Many stack holders who would like to predict song popularity
- Examples:
  - Venues
  - Music distributors



Hypothesis and Desired Outcomes

- Using song features, build a model which can predict if a given song will be above or below the mean of song popularity scores of the data (0.487166)

- Understand which features of a song are important to predict the song popularity

Data

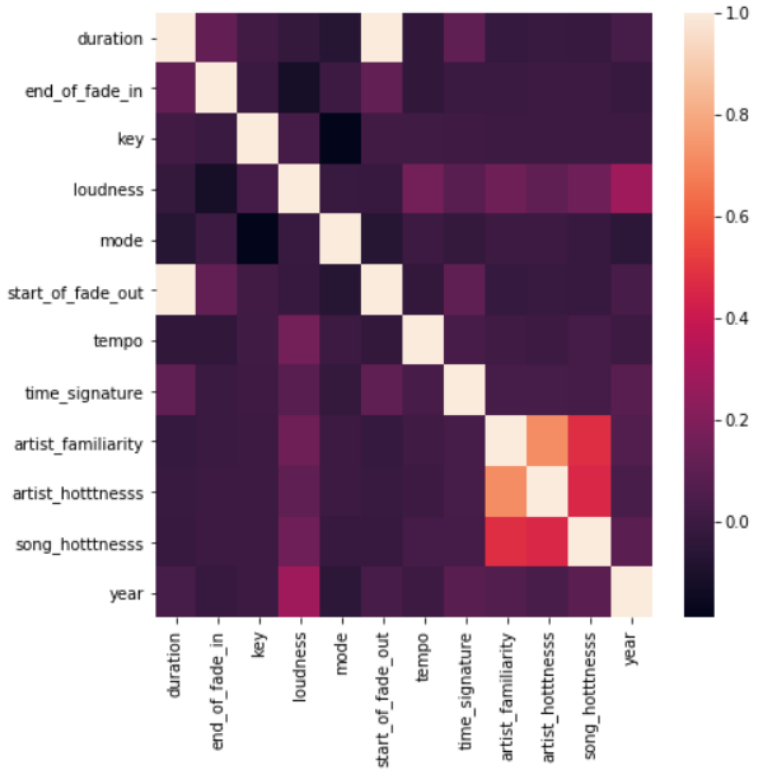
- Million Song Dataset (MSD) is freely available collection of audio features and metadata for a million contemporary tracks
- From The Echo Nest (now owned by Spotify)
- Size of dataset
  - 300GBs
  - One Million Songs
  - Over 50 features
- All tracks are stored in HDFS file system

Data Processing

- Downloaded the entire dataset on AWS EC2 node
- Extracted all the metadata using Python and created a CSV summary file
- Each song has 53 features such as Mode, Tempo, Duration, Title, End of Fade In etc
- Traditional data cleansing
  - Remove NaNs
  - Remove Songs that are missing year or song popularity metric

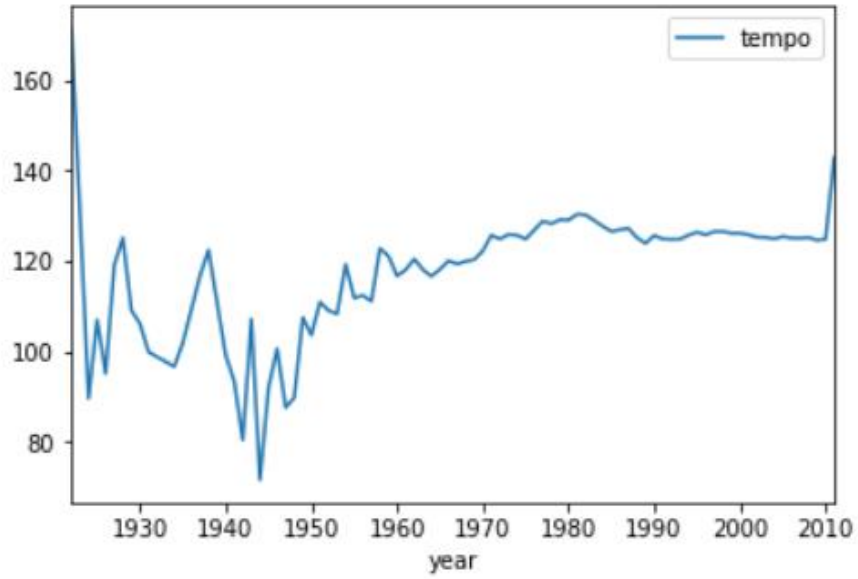
Data Exploration

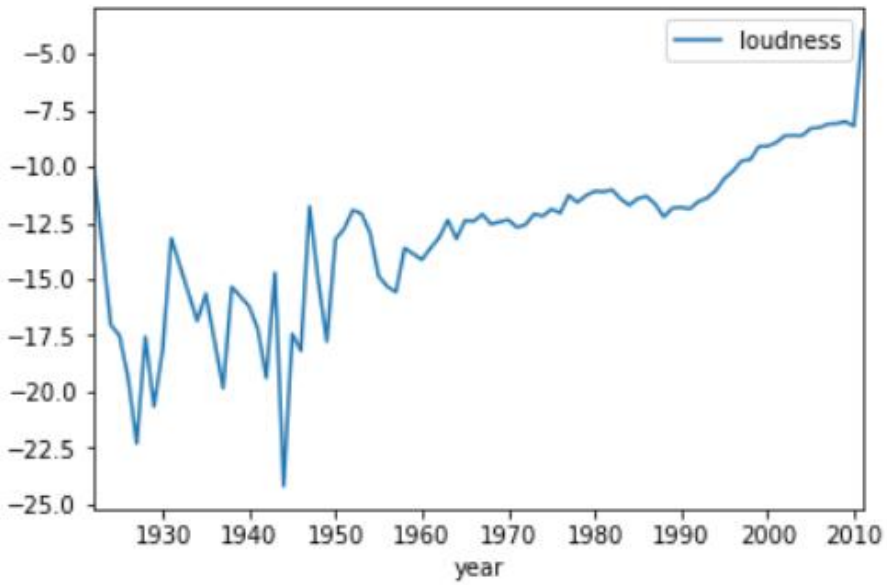
- After data cleansing, left with 306298 rows
- 19 Features (Mistake during data extraction)
- Correlation Matrix



Data Exploration

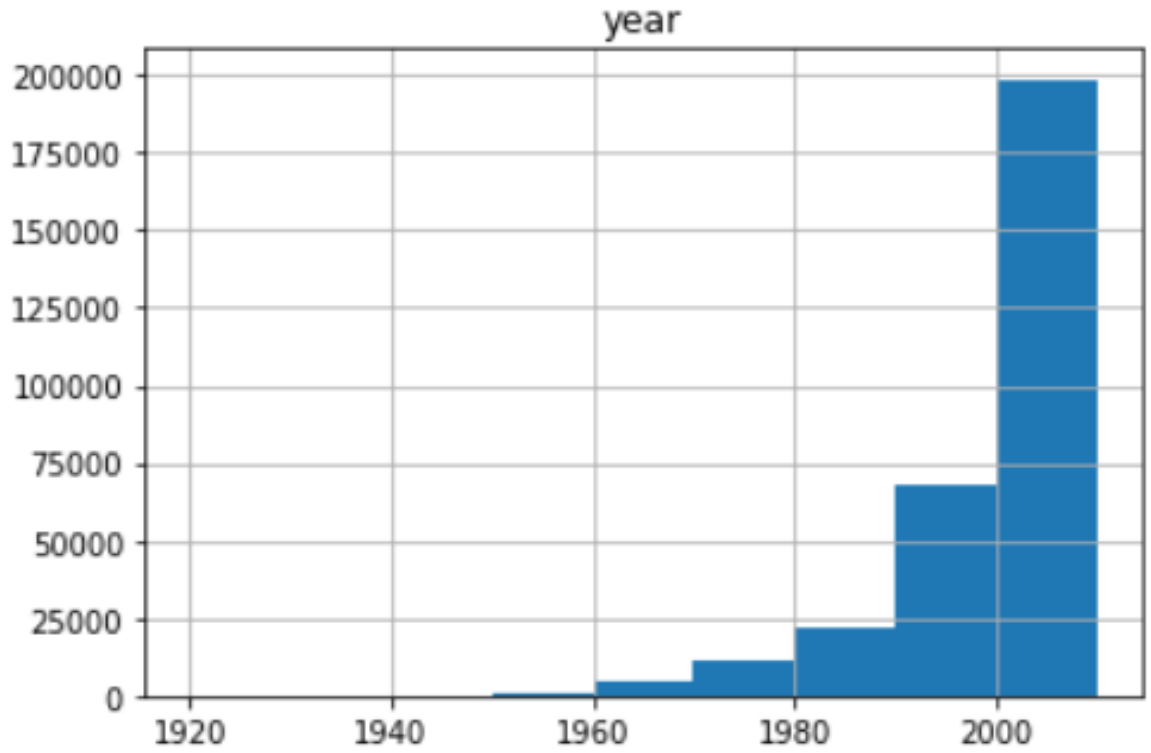
- Data Trends over the years





Data Exploration

- Number of Songs Over the years



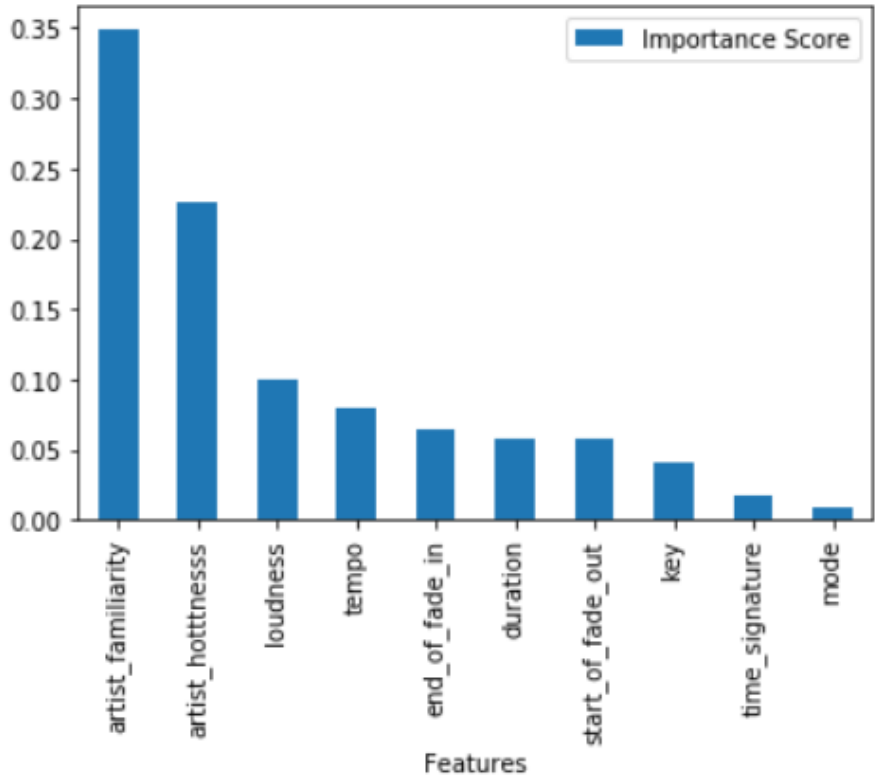
Data Exploration

- Most Popular songs sorted by song popularity and year

	title	artist_name	release	song_hotttnesss	year
69427	If We Ever Meet Again	Timbaland / Katy Perry	If We Ever Meet Again (Featuring Katy Perry)	1.0	2010
78108	Alice	Avril Lavigne	Almost Alice	1.0	2010
85836	Cooler Than Me	Mike Posner	Cooler Than Me	1.0	2010
196742	Somebody To Love	Justin Bieber	My Worlds	1.0	2010
332746	Holiday	Vampire Weekend	Contra	1.0	2010
424919	Nothin' On You [feat. Bruno Mars] (Album Version)	B.o.B	B.o.B Presents: The Adventures of Bobby Ray	1.0	2010
711682	Somebody To Love	Justin Bieber	My World 2.0	1.0	2010
712993	Odessa	Caribou	Odessa	1.0	2010
746866	Alice	Avril Lavigne	Alice	1.0	2010
828866	Tighten Up	The Black Keys	Tighten Up	1.0	2010

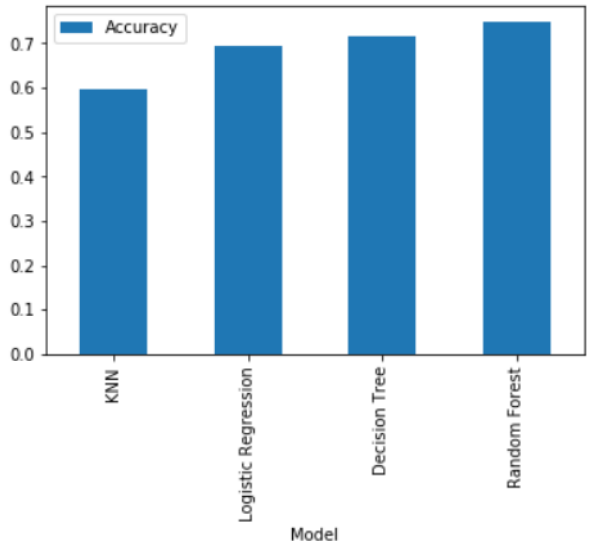
Feature Importance and Top Features

- Random Forest to determine feature importance score
- N_estimators = 10
- Max_depth = 40



Model Selection

Model	Accuracy
KNN	0.58966
Logistic Regression	0.69487
Decision Tree	0.71610
Random Forest	0.74869



Model Evaluation

- Random Forest gives the best accuracy of all the models tried
- It also had the best ROC and F1 scores (0.82192 and 0.74658)
- Precision and Recall
  - Precision is the probability that retrieved object is relevant
  - Recall is the probability that relevant object is retrieved
- Type 1 Error = FP = 39604
- Type 2 Error = FN = 37628

	precision	recall	f1-score	support
0	0.75	0.74	0.75	154233
1	0.74	0.75	0.75	152065
avg / total	0.75	0.75	0.75	306298

	Predicted No	Predicted Yes
Actual No	TN = 114629	FP = 39604
Actual Yes	FN = 37628	TP = 114437

Conclusion, Future and Spin-offs

- Predicted song popularity based on song features
- Artist familiarity, Artist Hotness, Loudness, Tempo, End of Fade In are top five features
- Random Forest gave the best results
- Future
  - Include song genres, most frequent words in lyrics as features
- Spin-off Projects
  - Use similar tracks and similar artists information to create inter-relationship map for artists and songs