Music Recommender and Genre Classification System

(CSCI 5502)

Team Overview

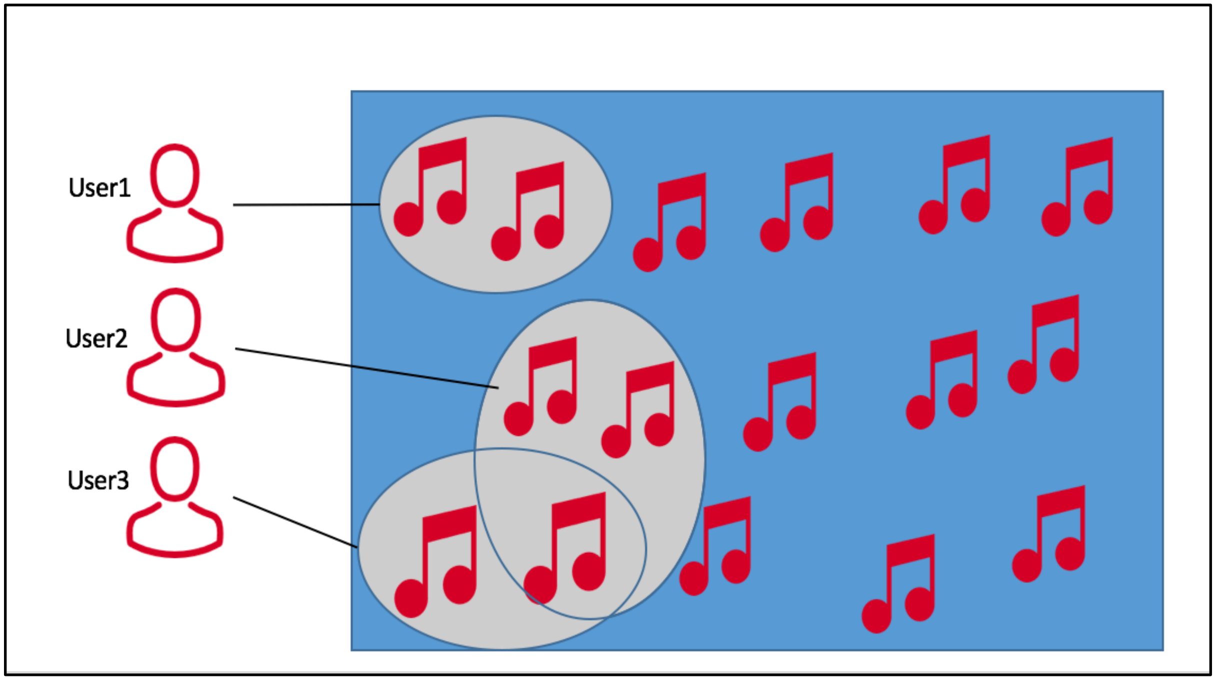
- Suresh Kumar (suku3476)
- Sunil Baliganahalli Narayana Murthy (suba5417)
- Praveen Kumar Devaraj (prde1873)

Problem

How to recommend music to Users ?

Given

- Set of Songs
- Set Users
- User profile



Previous Works

- Recommendation Systems produce recommendation in two ways - through collaborative or content-based filtering.

Collaborative filtering

- Depends on Community of Users
- Based on User Similarity
- “Cold Start” Problem

Content Based filtering

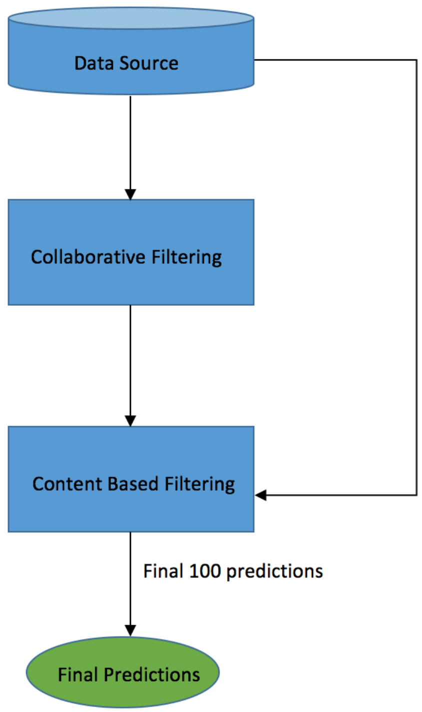
- Based on Song Similarity
- Depends on Song Features

Collaborative

Content Based

Proposed Work

- Combine the advantages of collaborative filter and content-based filter to form a hybrid recommendation system
- Genre Classification to categorize music
- Recommendation based on time of the day and User mood



Hybrid Model for Recommendation

Libraries / Frameworks Used

- Python
- Apache Spark
- Sklearn
- Theano
- Tensorflow google
- Numpy
- Pandas

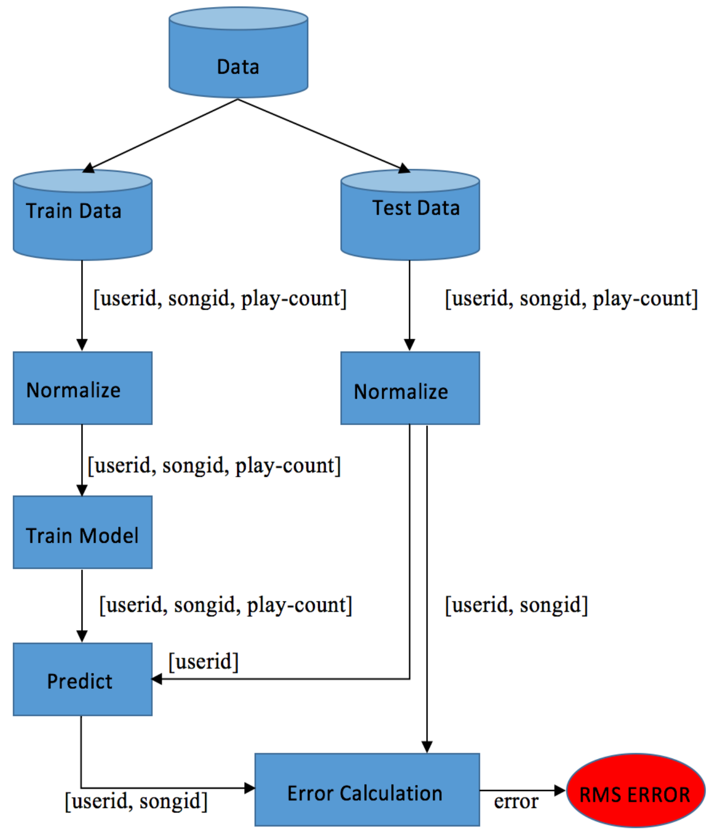
Datasets

- Million Song dataset (http://labrosa.ee.columbia.edu/millionsong/) - Open Source
- Available Datasets
  - SecondHandSongs dataset -> cover songs
  - musiXmatch dataset -> lyrics
  - Last.fm dataset -> song-level tags and similarity
  - Taste Profile subset -> user data
  - thisismyjam-to-MSD mapping -> more user data
  - tagtraum genre annotations -> genre labels
  - Top MAGD dataset -> more genre labels

- The raw data consists of listening history of a million users in HD5 format.
- Only a subset (1 million entries) of the data is used for testing.
- Full dataset (46 million entries) – AWS (~4hr) Collaborative Filtering
- K-Fold Cross-Validation (4:1)

Collaborative Filter(CF)

- Memory based CF:
  - Based on user rating to compute the similarity between the users
  - Performance decreases as the data gets sparse
- Model Based CF:
  - Based on machine learning algorithms to find patterns on training data
  - The users and items are a small set of latent factors used to predict missing entries
  - Performance is better in case of sparse data and is scalable
  - Used Alternating Least square algorithm to learn the latent factors
  - Play count of a song is used as rating

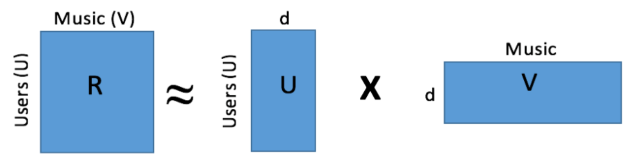


Collaborative Filter(CF)

- Used Apache Spark's Mlib - Model based collaborative – Uses Alternating Least square Algorithm.

Algorithm 1 ALS Algorithm

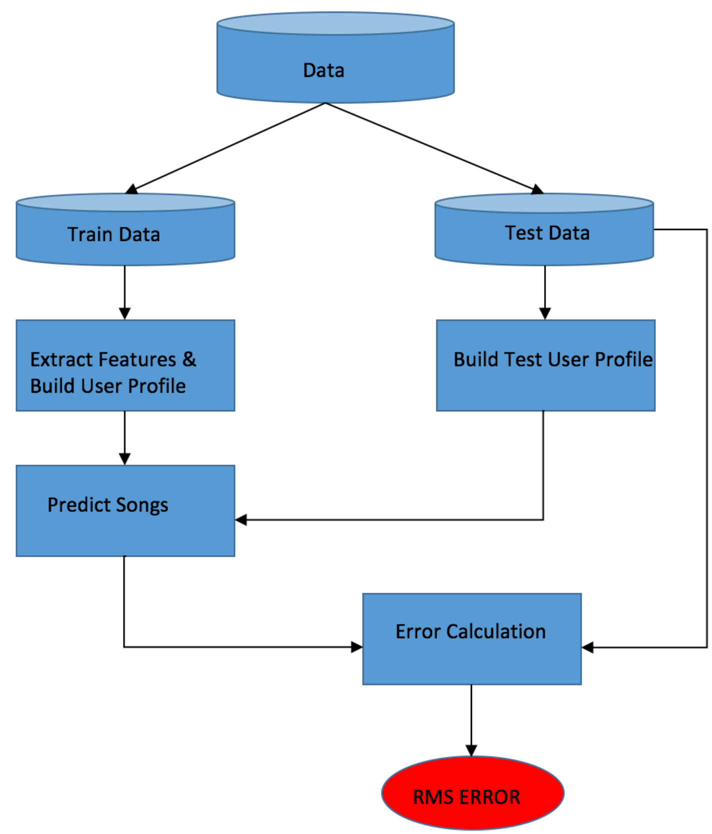
Set up vectors yUsers, yItems, xq, and xp
Initialize μ, a_i, b_u, y_{ui}
while MSE not converged and iteration $\neq$ max limit **do**
    Create q_i vectors
    Solve for p_u vectors
    Create p_u vectors
    Solve for q_i vectors
    Rescale each p_u and q_i vector
    Update μ, a_i, b_u, y_{ui}
    Find MSE
end while



Low Rank Matrix Factorization

Proposed Work – Content-Based Filtering

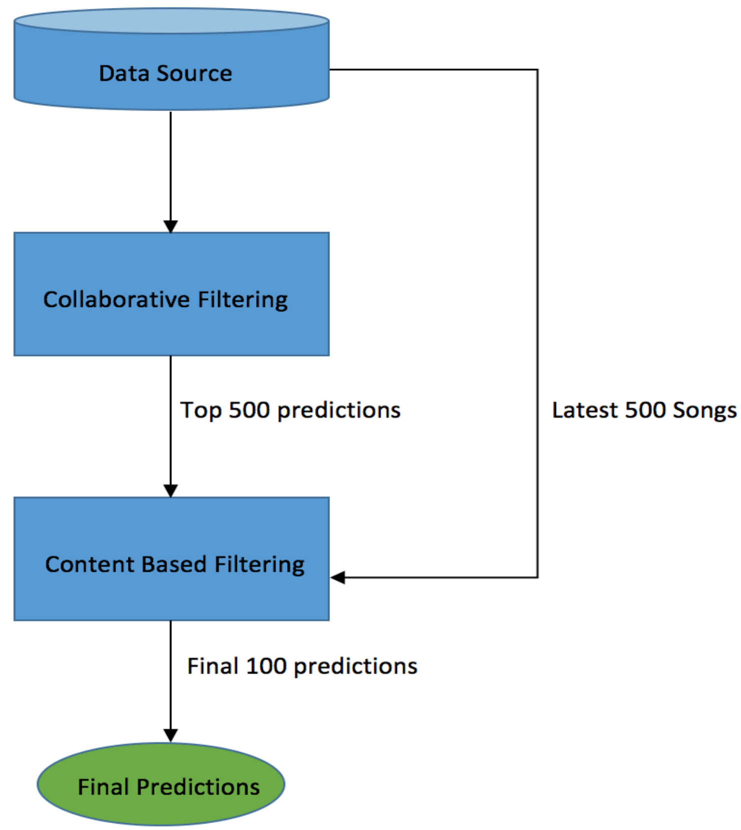
- Depends on the attributes and features of the songs
- Song attributes considered include danceability, tempo, energy, timbre, key, loudness
- User profile is based on a weighted vector of songs This weights denote importance of each feature.
- Machine learning algorithms like Bayesian Classifiers, cluster analysis, decision trees, artificial neural networks, etc., were used



Hybrid Model = Collaborative Filter + Content-based Filter

Two approaches are pipelined – Output of Collaborative Filtering is fed as input for Content based Filtering

Feed latest songs for content based filtering to recommend the latest songs also – to overcome the disadvantage of Collaborative Filtering.



Genre Classifier

- Automatic Genre Classification – Categorize music based on genre
- Data mining techniques and Machine learning algorithms on existing data.
- Feature vector include
  - Loudness,
  - Tempo
  - Danceability
  - Key
  - Energy
  - Duration
  - Artist Name etc
- Model is constructed using above feature vector to predict genre.
- Modelling techniques –
  - Support Vector Machines (SVM), Nearest-Neighbor (NN) classifiers
  - Gaussian Mixture Models, Linear Discriminant Analysis (LDA), Random Forest

Completed Tasks

- **Collaborative Filtering**
  - Apache Spark model-based collaborative filter using Alternating Least Squares (ALS) algorithm.
- **Content Based Filtering**
  - Logistic Regression from Apache Spark
  - SVM
- **Hybrid Model**
- **Genre classifier:**
  - Constructed feature vector using attributes like “Loudness", “Tempo", “Time-Signature", “Mode", “Key", “Average Timbre", “Duration" and “Variable Timbre”.
  - Logistic Regression
  - Naïve Bayes
  - Random Forest
  - Decision Tree
- **Evaluation Metrics**

Performance Evaluation

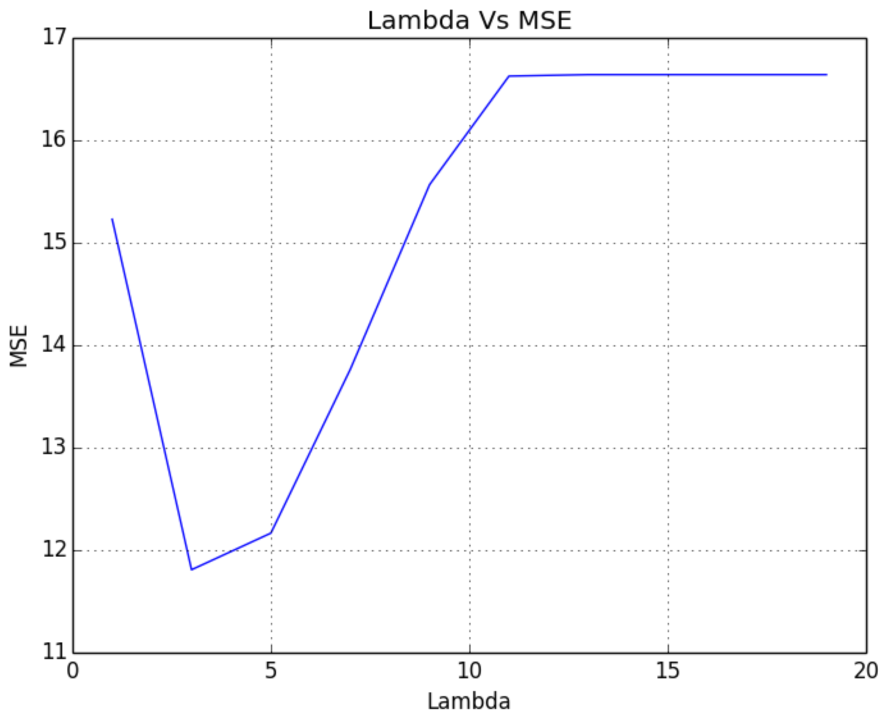
- Mean Squared Error and Root Mean Squared Error.
- Precision and recall or DCG to assess the quality of the recommendation system.
- K Fold Cross Validation in the ratio of 4:1 (Split the input data into training set and test set)
- RMSE formula :

$$RMSE = \sqrt{\frac{1}{n}\sum_{p,q}|P_{u,v} - R_{u,v}|^2}$$

- Evaluation of Genre Classifier – Predicted genre for songs of known genre and compare the predicted genre with the actual genre of the song.

Key Results

- Collaborative Filtering Results



Rank = 8 Number of Iterations = 10

Lambda – Regularization Parameter	Mean Squared Error
1	15.23
3	11.80
5	12.16
7	13.76
9	15.56
11	16.62
13	16.64
:	:
20	16.64

Key Results

- Collaborative Filtering Results

-------------------- Songs Recommended by Collaborative Filter(ALS) --------------------

song_id	title	artist_name	duration	year	track_id
SONZXZH12A6D4F7C62	Drop In The Ocean	Andreas Johnson	186.4616	2005	TROMFYO128F1473C56
SOMAGUJ12AB0190062	My Sanctuary	First State	458.26566	2010	TROQTMN12903CEC501
SOPEMAZ12A6D4F957C	(Part One) The Mi...	Monty Python	207.43791	1983	TRSAYJF128F149805A
SOJSXJY12A8C13E32E	Clara meets Slope...	Clara Hill	266.44852	2006	TRHRTYE128F427EA7D
SOETMEE12A67AD86A7	Star For A Week	Steve Harley & Co...	326.1122	0	TRDCOWJ128EF345FAC
SOVNNBW12A8C136A6D	Nur Ein Wort (Dem...	Wir Sind Helden	195.57832	2005	TRFVBYF128F42632EF
SOTKSXQ12A8C143608	Self Control	Paffendorf	208.27383	2009	TREIIEB128F93263A9
SOBIQZK12AB017F9D5	Castallion Springs	Lonesome Standard...	249.15546	0	TRCYCYA128F92F25CA
SOOJJTV12A6D4F7930	Love Is All We Need	Mel Carter	107.38893	1996	TRNLCWP128F1473836
SOJQHWS12A6701F823	Clever Kicks	The Hiss	143.59465	2003	TRMUKES128E0792241
SOSVFGM12AB0182373	Without Bill the ...	Rolfe Kent	53.21098	0	TREEXSF128F931F7DE
SOASIRM12A8C13C69F	Problem Child	Doyle Bramhall II...	350.4322	0	TRVGDNS128F42911EE
SOLEDTI12AB0185675	Hölle	Stahlhammer	250.4616	2002	TRJTFPM128F9320721
SOCJDCM12A8C1396D7	El Sr. Durito Y Yo	León Gieco	277.4722	1997	TRZYMBQ128F42804D5
SOFAEOZ12AF729BEBB	Through With Love	Destiny's Child	215.48363	2004	TRVIFOA128F92C2E2E
SORFXUP12A8C142CDC	Just A Few Sweet ...	Slim Whitman	153.80853	0	TRYICYQ128F92D2E97
SOCWUOO12AC392B5DA	Hollywood_ I'm Co...	Twiztid	166.922	0	TRALXKT12903CE6AB6
SOEQQAA12A6D4F851C	Tyyssija	Mokoma	287.32037	1999	TRLGCNE128F14740B1
SOUEKDP12A58A79B60	In Sickness And H...	The Gathering	419.00363	0	TRVFQJD128F93324CF
SOBXJRG12A8C13C5DD	At Point Break	The Unseen	116.40118	2007	TRHNORJ128F42AB3B4

only showing top 20 rows

-------------------- Start Training LogisticRegressionWithLBFGS using ALS predicted songs ----------------

Key Results

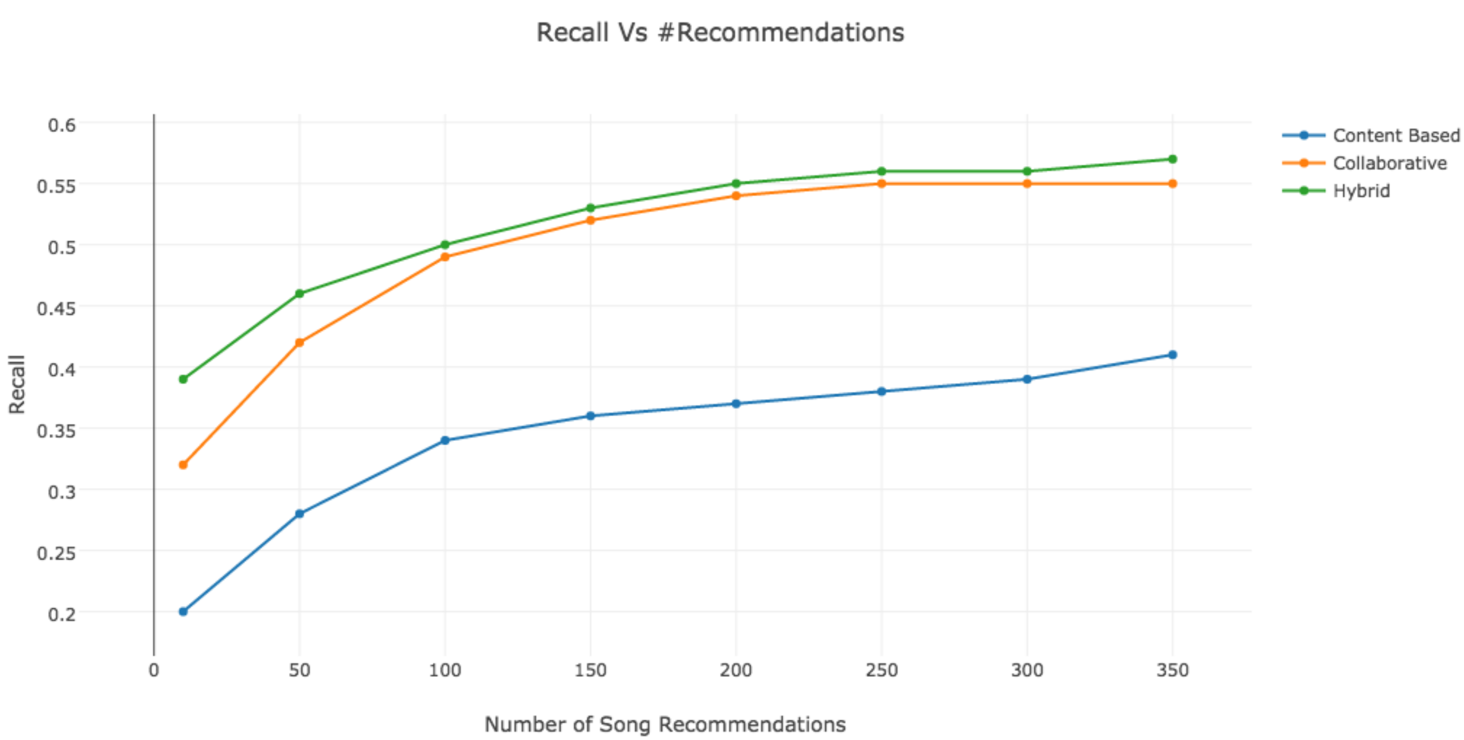
- Hybrid Recommendation Model - Predictions

song_id	title	artist_name	duration	year	track_id
SOWSQIA12A58A78B5A	Dancing With Tear...	Ke$ha	209.24036	2010	TRGQZTB128F92F8767
SODGPRV12AB0186F51	Shameless	Groove Armada	285.88364	2010	TRHEMWO12903CBFA47
SOXHYRU12AC4688759	The Echoing Green	Erland And The Ca...	218.56609	2010	TRHMVLA12903D07644
SOGOTTK12AB0185F95	A Far Cry	We Were Promised ...	292.10077	2010	TRKNNPV12903CCBF1C
SOZNAXT12AB0188B07	The White Queen	Danny Elfman	36.64934	2010	TRKUPDA12903CBDAF2
SOZZQLZ12AB018B9D2	Kiss Me When I'm ...	Gary Allan	232.5677	2010	TRLQCRG12903CB9648
SOIYYHP12A8C141468	World Is Mine	Showtek	366.44525	2010	TRLUHMO128F92F3E2C
SORRIDE12AC468E8FA	If I Had My Way (...	Patty Griffin Fea...	203.83302	2010	TRNRMHC12903D075CB
SOUQQWB12AB01828AB	Dreams-Come-True-...	Cass McCombs feat...	322.06323	2010	TRRBOOC128F92EAA53
SOCSGGX12AB018B6D4	Sing	Four Tet	840.17584	2010	TRSNCHY12903CA72DA
SOMRXGA12AB01867E2	Me And The Devil	Gil Scott-Heron	218.33098	2010	TRTKXNY12903CB661C
SOJOAEN12AC3A4DF25	Meet Me On The Da...	Melissa Auf der Maur	246.88281	2010	TRTVFJJ12903CE6962
SOTLEYY12AB018E02D	Godless (Hard Dan...	Unter Null	327.1571	2010	TRUREGR12903CE38DE
SODDIUK12AC468C0BD	Can't Let You Go	Adam Lambert	254.1971	2010	TRVOCNU12903D13774
SODSDKE12AB018BC91	Greenland	Emancipator	190.9024	2010	TRWRHSX12903CD22D2
SOKGAKY12AB01883E1	Oh You (Christmas...	LCD Soundsystem	230.0077	2010	TRYTXXJ12903CD1947
SOWQHCQ12AB01851AE	Protection Racket	Fionn Regan	159.242	2010	TRZDNGG128F934B160
SOMXPRX12AB0183570	It'll Be Alright	The Infamous Stri...	201.0379	2010	TRASQOM12903CA8E5E
SOVTJDL12AC4687C21	Glass Arrows (Alb...	Circa Survive	252.21179	2010	TRCDPTV12903D140B0
SOBPLZP12AC3DF66CF	Desert Sand	Beach Fossils	240.0126	2010	TRCFFYV12903CE4DD8

only showing top 20 rows

Key Results

- Hybrid Filter model

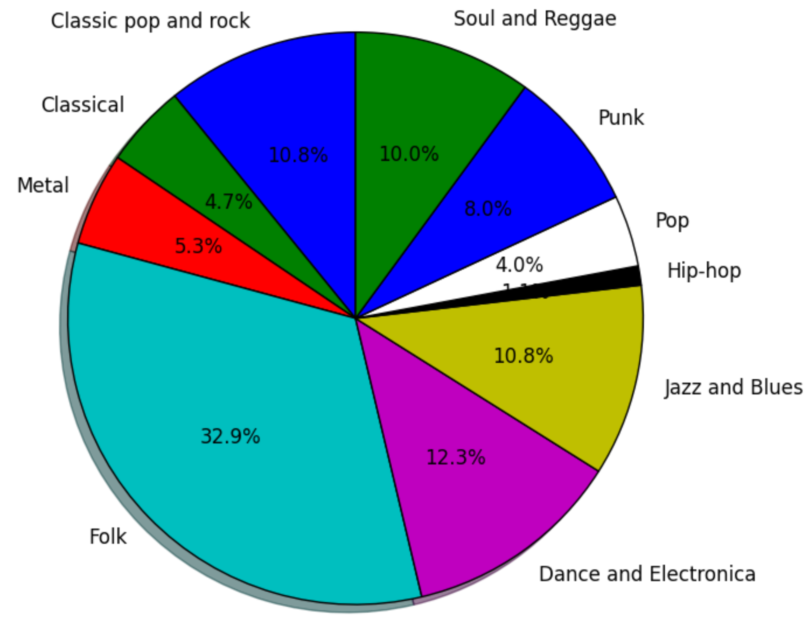


Key Results

- Genre Classifier

Classifier	Accuracy
Logistic Regression	53.76
Decision Tree	47.58
Naïve Bayes	24.52
Random Forest	58.37
SVM	41.73
SVM + Boosting	49.64

Classifier Vs Accuracy

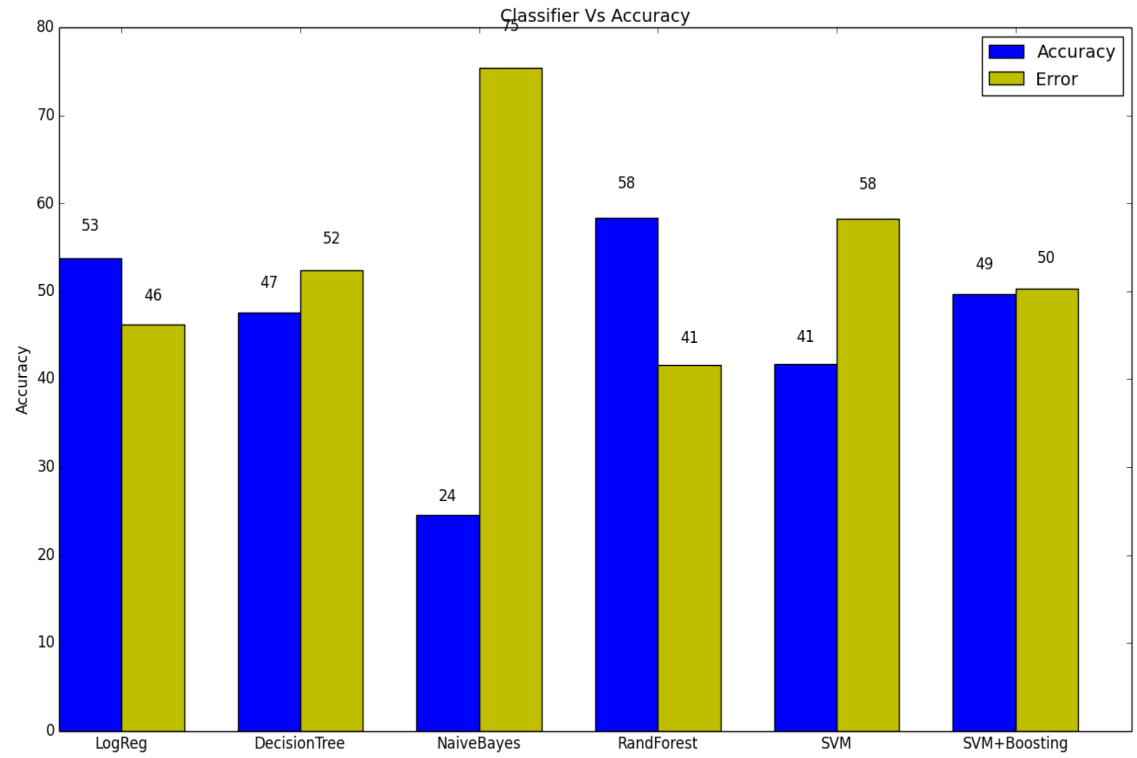


Distribution of Genre

Key Results

Genre Classifier

- Random Forest – Highest Accuracy – Ensemble method



Showing Top 20 rows in Genre Classification using Random Forest

genre	track_id	features	indexedLabel	rawPrediction	probability	prediction	predictedLabel
classic pop and rock	TRAAAGR128F425B14B	[-7.322,123.989,4...	0.0	[0.30337078651685...	[0.15168539325842...	2.0	dance and electro...
classic pop and rock	TRAACPE128F421C1B9	[-11.939,110.189,...	0.0	[0.87855449075979...	[0.43927724537989...	0.0	classic pop and rock
classic pop and rock	TRAACQW128F428854F	[-7.032,47.271,4....	0.0	[0.98946869070208...	[0.49473434535104...	0.0	classic pop and rock
classic pop and rock	TRAAGJV128F1464090	[-9.091,129.122,1...	0.0	[0.90697350069735...	[0.45348675034867...	0.0	classic pop and rock
classic pop and rock	TRAAGNL128F4299BF1	[-11.335,137.29,4...	0.0	[1.09313479623824...	[0.54656739811912...	0.0	classic pop and rock
classic pop and rock	TRAAIAN12903CFF16D	[-11.479,151.911,...	0.0	[0.96922310797627...	[0.48461155398813...	0.0	classic pop and rock
classic pop and rock	TRAAMUY128F4283222	[-7.947,108.94,4....	0.0	[0.57220342053330...	[0.28610171026665...	1.0	folk
classic pop and rock	TRAAOAU12903D0154B	[-15.146,77.27,4....	0.0	[0.88663745892661...	[0.44331872946330...	0.0	classic pop and rock
classic pop and rock	TRAAUSW128F426646E	[-10.173,102.055,...	0.0	[0.84805175267037...	[0.42402587633518...	0.0	classic pop and rock
classic pop and rock	TRAAXRS128F932F05D	[-14.26,121.93,4....	0.0	[1.38208708930594...	[0.69104354465297...	0.0	classic pop and rock
classic pop and rock	TRABEKP128E078C123	[-17.772,116.267,...	0.0	[0.64463445530220...	[0.32231722765110...	1.0	folk
classic pop and rock	TRABJYG128F92EB8DC	[-8.482,120.809,4...	0.0	[1.28955547736335...	[0.64477773868167...	0.0	classic pop and rock
classic pop and rock	TRABLSO128F14A4707	[-13.134,152.434,...	0.0	[0.31194968553459...	[0.15597484276729...	1.0	folk
classic pop and rock	TRABMTM12903D083D2	[-4.942,92.014,1....	0.0	[1.15726233421274...	[0.57863116710637...	0.0	classic pop and rock
classic pop and rock	TRABTFI128F14905F6	[-18.264,151.477,...	0.0	[1.00660429279031...	[0.50330214639515...	0.0	classic pop and rock
classic pop and rock	TRABTYR128F9304934	[-16.172,127.207,...	0.0	[0.93363338820870...	[0.46681669410435...	0.0	classic pop and rock
classic pop and rock	TRABWPT128F42B06CC	[-10.548,89.966,4...	0.0	[0.57273482959268...	[0.28636741479634...	1.0	folk
classic pop and rock	TRABXHU128F147EDE9	[-14.369,157.219,...	0.0	[0.93995765275257...	[0.46997882637628...	0.0	classic pop and rock
classic pop and rock	TRACRBQ128F4263964	[-14.293,111.492,...	0.0	[0.96624803767660...	[0.48312401883830...	0.0	classic pop and rock
classic pop and rock	TRACTIQ128F4288A7C	[-8.45,130.031,4....	0.0	[1.18783945428682...	[0.59391972714341...	0.0	classic pop and rock

Future Scope

- Use Deep learning techniques.
- Use Probabilistic Topic Modelling with Collaborative Filtering
- Recommend songs based on time of the day (Morning, Noon, Evening).
- Recommendation based on mood tags sad, happy, joyful, romantic, etc.
- Integration with Spotify/I-tunes etc