Project Proposal
CS401
Date: 30/08/18

Shoaib Ahmed
Roll : 1501049

Problem Statement

Prediction of the release year of a song from audio features, using the YearPredictionMSD[1], a subset of the Million Songs Dataset[2]

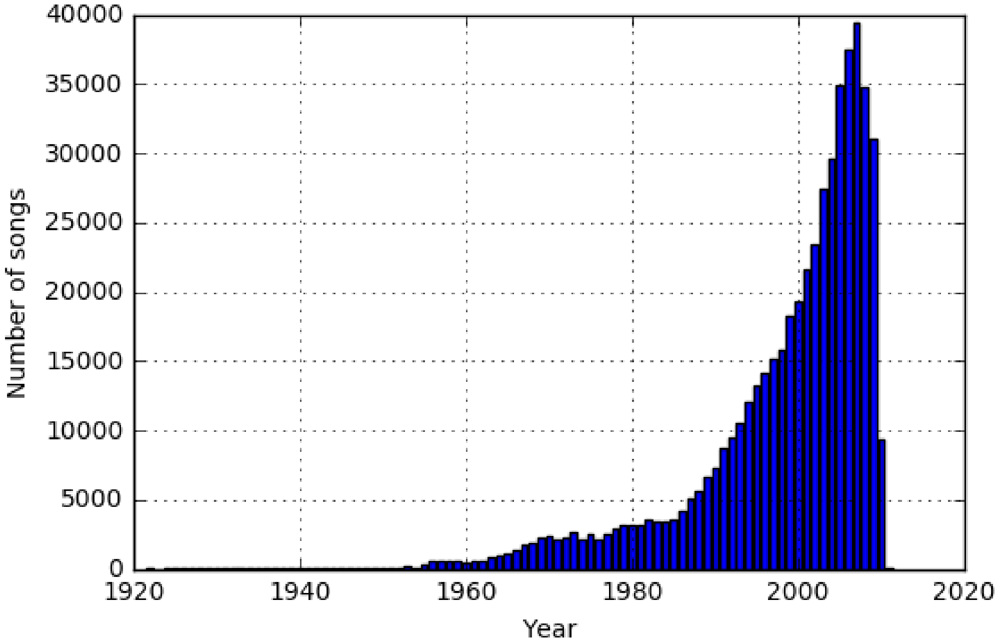
About the Problem

Year prediction is defined as estimating the year in which a song was released based on its audio features. Million Song Dataset, a famous data set with freely available collection of audio features and metadata for a million contemporary popular music tracks. There are many attractive features in the Million Song Database. Here, what we focus are *Timbre Average* and *Timbre Covariance* features, and then do the release year prediction based on these.

About the Dataset

The data set used for this project is UCI Machine Learning Respository [1] YearPredictionMSD which is a subset of the Million Song Dataset[2]. The songs here are mostly western, commercial tracks ranging from 1922 to 2011, with a peak in the year 2000s.

Figure 1: Trend of Data in the Dataset through the years.



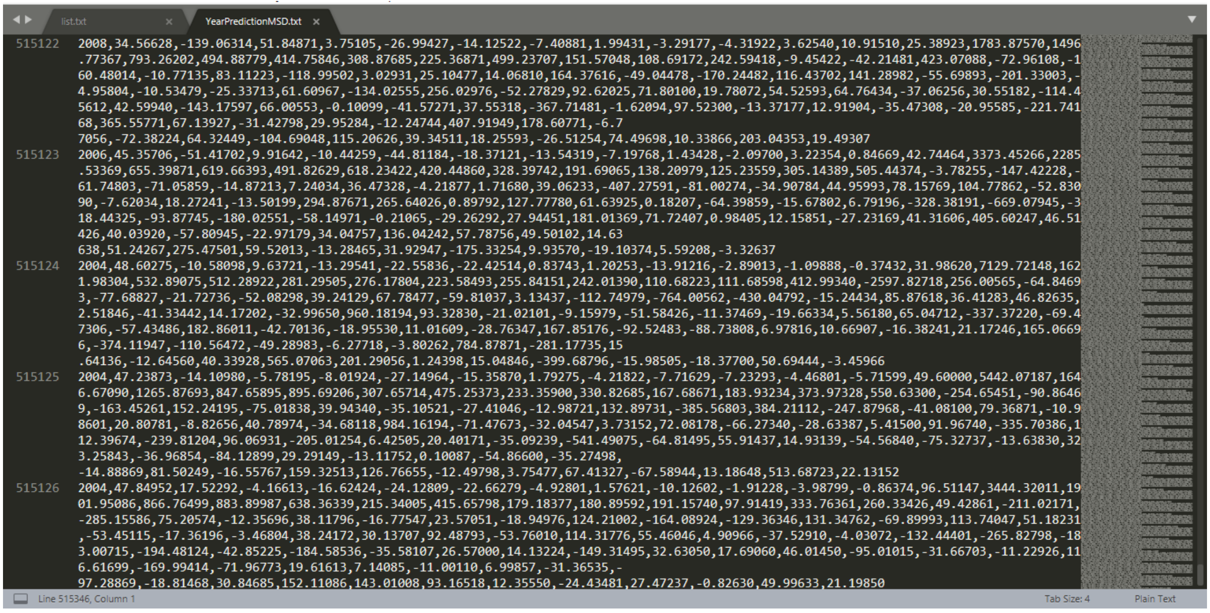
Data Collection

Specific characteristics of the dataset[1] used, are:

- Size: 460MB
- File Format: .txt
- Number of songs: 515345
- Attributes: 90 (12 = timbre average, 78 = timbre covariance)

The dataset contains features and labels for each song, where, label is the year in which the song was released. There are 12 timbre-average and 78 timbre-covariance features. Features have been extracted from the 'timbre'[4] features from The Echo Nest API[5]. We take the average and covariance over all 'segments', each segment being described by a 12-dimensional timbre vector. The API divides the song into multiple time segments. These segments were delimited by onsets and (or) other discontinuities in the signal

Figure 2: Screenshot of the data. Each line contains the Year followed by the 90 features.



```
2008,34.56628,-139.06314,51.84871,3.75105,-26.99427,-14.12522,-7.40881,1.99431,-3.29177,-4.31922,3.62540,10.91510,25.38923,1783.87570,1496
.77367,793.26202,494.88779,414.75846,308.87685,225.36871,499.23707,151.57048,108.69172,242.59418,-9.45422,-42.21481,423.07088,-72.96108,-1
60.48014,-10.77135,83.11223,-118.99502,3.02931,25.10477,14.06810,164.37616,-49.04478,-170.24482,116.43702,141.28982,-55.69893,-201.33003,-
4.95804,-10.53479,-25.33713,61.60967,-134.02555,256.02976,-52.27829,92.62025,71.80100,19.78072,54.52593,64.76434,-37.06256,30.55182,-114.4
5612,42.59940,-143.17597,66.00553,-0.10099,-41.57271,37.55318,-367.71481,-1.62094,97.52300,-13.37177,12.91904,-35.47308,-20.95585,-221.741
68,365.55771,67.13927,-31.42798,29.95284,-12.24744,407.91949,178.60771,-6.7
7056,-72.38224,64.32449,-104.69048,115.20626,39.34511,18.25593,-26.51254,74.49698,10.33866,203.04353,19.49307
2006,45.35706,-51.41702,9.91642,-10.44259,-44.81184,-18.37121,-13.54319,-7.19768,1.43428,-2.09700,3.22354,0.84669,42.74464,3373.45266,2285
.53369,655.39871,619.66393,491.82629,618.23422,420.44860,328.39742,191.69065,138.20979,125.23559,305.14389,505.44374,-3.78255,-147.42228,-
61.74803,-71.05859,-14.87213,7.24034,36.47328,-4.21877,1.71680,39.06233,-407.27591,-81.00274,-34.90784,44.95993,78.15769,104.77862,-52.830
90,-7.62034,18.27241,-13.50199,294.87671,265.64026,0.89792,127.77780,61.63925,0.18207,-64.39859,-15.67802,6.79196,-328.38191,-669.07945,-3
18.44325,-93.87745,-180.02551,-58.14971,-0.21065,-29.26292,27.94451,181.01369,71.72407,0.98405,12.15851,-27.23169,41.31606,405.60247,46.51
426,40.03920,-57.80945,-22.97179,34.04757,136.04242,57.78756,49.50102,14.63
638,51.24267,275.47501,59.52013,-13.28465,31.92947,-175.33254,9.93570,-19.10374,5.59208,-3.32637
2004,48.60275,-10.58098,9.63721,-13.29541,-22.55836,-22.42514,0.83743,1.20253,-13.91216,-2.89013,-1.09888,-0.37432,31.98620,7129.72148,162
1.98304,532.89075,512.28922,281.29505,276.17804,223.58493,255.84151,242.01390,110.68223,111.68598,412.99340,-2597.82718,256.00565,-64.8469
3,-77.68827,-21.72736,-52.08298,39.24129,67.78477,-59.81037,3.13437,-112.74979,-764.00562,-430.04792,-15.24434,85.87618,36.41283,46.82635,
2.51846,-41.33442,14.17202,-32.99650,960.18194,93.32830,-21.02101,-9.15979,-51.58426,-11.37469,-19.66334,5.56180,65.04712,-337.37220,-69.4
7306,-57.43486,182.86011,-42.70136,-18.95530,11.01609,-28.76347,167.85176,-92.52483,-88.73808,6.97816,10.66907,-16.38241,21.17246,165.0669
6,-374.11947,-110.56472,-49.28983,-6.27718,-3.80262,784.87871,-281.17735,15
.64136,-12.64560,40.33928,565.07063,201.29056,1.24398,15.04846,-399.68796,-15.98505,-18.37700,50.69444,-3.45966
2004,47.23873,-14.10980,-5.78195,-8.01924,-27.14964,-15.35870,1.79275,-4.21822,-7.71629,-7.23293,-4.46801,-5.71599,49.60000,5442.07187,164
6.67090,1265.87693,847.65895,895.69206,307.65714,475.25373,233.35900,330.82685,167.68671,183.93234,373.97328,550.63300,-254.65451,-90.8646
9,-163.45261,152.24195,-75.01838,39.94340,-35.10521,-27.41046,-12.98721,132.89731,-385.56803,384.21112,-247.87968,-41.08100,79.36871,-10.9
8601,20.80781,-8.82656,40.78974,-34.68118,984.16194,-71.47673,-32.04547,3.73152,72.08178,-66.27340,-28.63387,5.41500,91.96740,-335.70386,1
12.39674,-239.81204,96.06931,-205.01254,6.42505,20.40171,-35.09239,-541.49075,-64.81495,55.91437,14.93139,-54.56840,-75.32737,-13.63830,32
3.25843,-36.96854,-84.12899,29.29149,-13.11752,0.10087,-54.86600,-35.27498,
-14.88869,81.50249,-16.55767,159.32513,126.76655,-12.49798,3.75477,67.41327,-67.58944,13.18648,513.68723,22.13152
2004,47.84952,17.52292,-4.16613,-16.62424,-24.12809,-22.66279,-4.92801,1.57621,-10.12602,-1.91228,-3.98799,-0.86374,96.51147,3444.32011,19
01.95086,866.76499,883.89987,638.36339,215.34005,415.65798,179.18377,180.89592,191.15740,97.91419,333.76361,260.33426,49.42861,-211.02171,
-285.15586,75.20574,-12.35696,38.11796,-16.77547,23.57051,-18.94976,124.21002,-164.08924,-129.36346,131.34762,-69.89993,113.74047,51.18231
,-53.45115,-17.36196,-3.46804,38.24172,30.13707,92.48793,-53.76010,114.31776,55.46046,4.90966,-37.52910,-4.03072,-132.44401,-265.82798,-18
3.00715,-194.48124,-42.85225,-184.58536,-35.58107,26.57000,14.13224,-149.31495,32.63050,17.69060,46.01450,-95.01015,-31.66703,-11.22926,11
6.61699,-169.99414,-71.96773,19.61613,7.14085,-11.00110,6.99857,-31.36535,-
97.28869,-18.81468,30.84685,152.11086,143.01008,93.16518,12.35550,-24.43481,27.47237,-0.82630,49.99633,21.19850
```

Construction/Implementation

I am planning to implement this research paper[3]. It is implements four types of machine learning algorithms on Apache Spark Machine Learning library (MLlib). The algorithms applied on the dataset are Linear Regression, Random Forest and Gradient Boosted Trees. The data preprocessing step includes data rescaling and normalization as the dataset at our disposal is highly skewed towards latest sound tracks, i.e., the number of songs released after 1990 is much greater than the number of songs released before 1990, so, we divide the dataset into two parts - Before 1990 and After 1990

References

[1] *Source :* https://archive.ics.uci.edu/ml/datasets/yearpredictionmsd

[2] T. Bertin-Mahieux *Original Source :* http://labrosa.ee.columbia.edu/millionsong

[3] Mishra, P., Garg, R., Kumar, A.,2016 *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI).*
https://ieeexplore.ieee.org/document/7732275/

[4] *Timbre :* https://en.wikipedia.org/wiki/Timbre

[5] *EchoNest :* http://the.echonest.com/