Billboard Top 200 Binary Classifier

By Ankit Prasad and Prateik Mahendra

Abstract:

The primary objective of this undertaking is to create a binary classifier for whether a given album will get ranked on the Billboard Top 200. We attempted four different types of classifiers: support vector machines and random forests. Between these two, random forests yielded the most promising results with an F-measure of 73% after parameter optimization.

Introduction:

The global music industry is a goliath in not only its breadth of reach to consumers/listeners but the money involved in all facets of its value chain. In 2014, the Investing in Music report suggest that record labels in just the United States spend $4.3 billion on artist recruitment/development and marketing[1]. Given how flush this industry is with money spent by key players, it would be of great business value if these organization had some means of determining what are some key drivers in an album's success. We deemed success to be whether an album gets ranked on the Billboard Top 200 because this ranking pertains to albums and the rich history of Billboard (its first music magazine was published in 1936[2]). The idea of using an empirical data-driven approach in assisting the marketing operation of a entertainment industry has been implemented by other researchers. One such example was when Zhao & Lash applied a feature engineering methodology based on various meta details about movies to help predict a given movie's profitability[3]. Using some analogous features (for the music industry vs movies) from this study, we applied the generated features to our own modeling methodology.

Data Collection:

The first step was gathering historical data and metadata on as many albums and artists as possible. For historical album data http://www.allmusic.com was used and for information on miscellaneous metadata, Wikipedia was used. Upon completion of web scraping phase, the extracted data was stored on SQL cloud. We used MySQL Work Bench to store and access the scraped data. As the data came from different sources, we created more than one tables. The master table called "AllMusicAlbum" contained data of all albums with primary key as their album ids. Another tables had albums that got ranked on billboard.

Feature Engineering:

After scraping and gathering the data, the next important step was to come up with a feature matrix which could be used to run machine learning classification algorithms to achieve the goal of the research and

[1] http://www.ifpi.org/news/record-labels-invest-us-4-3-billion-in-AR-and-marketing

[2] *Sale, Jonathan (January 4, 1996). "Sixty years of hits, from Sinatra to ... Sinatra". The Independent. Retrieved January 3, 2017*

[3] https://www.biz.uiowa.edu/faculty/kangzhao/pub/JMIS_2016.pdf

classify which albums that would make it to billboard. All features were created using a Python Script which we ran on the university servers. We used Python pyMySQL which allows you to connect to relational databases that use MySQL to make any manipulations.

For sake of organization and better understanding we decided to create features under three different categories.

1. Star Power Features
2. Genre Features
3. Record Label Features

1. *Star Power Features:*

Simply defined, star power features are statistic measures to estimate the power of the artist that is in question. Human brain can understand if an artist is more popular than another but to make a machine learn about this, a quantitative measure must be calculated which would be compared to that of another artist to decide on the popularity of an artist. There are multiple approaches to this. While we can come up with one such statistic, we can also take many measures and use them separately to measure a cumulative power. We took the later approach and created 5 such features.

i. *Number of albums:*
It is the statistic that measures the number of albums released by an artist before the release date of the album in question. For instance, if the snoop dog has 10 albums released, the "Number of albums" measure for his 11th album would be 10.

ii. *Number of ranked albums:*
It is the statistic that measures the number of ranked albums released by an artist before the release date of the album in question. For instance, if snoop dog has 10 albums, 6 of which got ranked, the "Number of ranked albums" measure for his 10th album would be 6.
But, this measure is not enough in itself as rank set 1, 2, 3 and rank set 10, 20, 30 give the same answer for this feature which is 3. To consider the power of the rank we take the next feature.

iii. *Average of the rank of the ranked albums:*
It is the statistic that measures the average rank of ranked albums by an artist before the release date of the album in question. While we have the average, we still do not know the distribution of the ranks. To consider how scattered the ranks are, we take the next feature.

iv. *Standard deviation of the rank of ranked albums:*
It is the statistic that measures how scattered are the rank of the ranked albums by an artist before release date of the album in question.

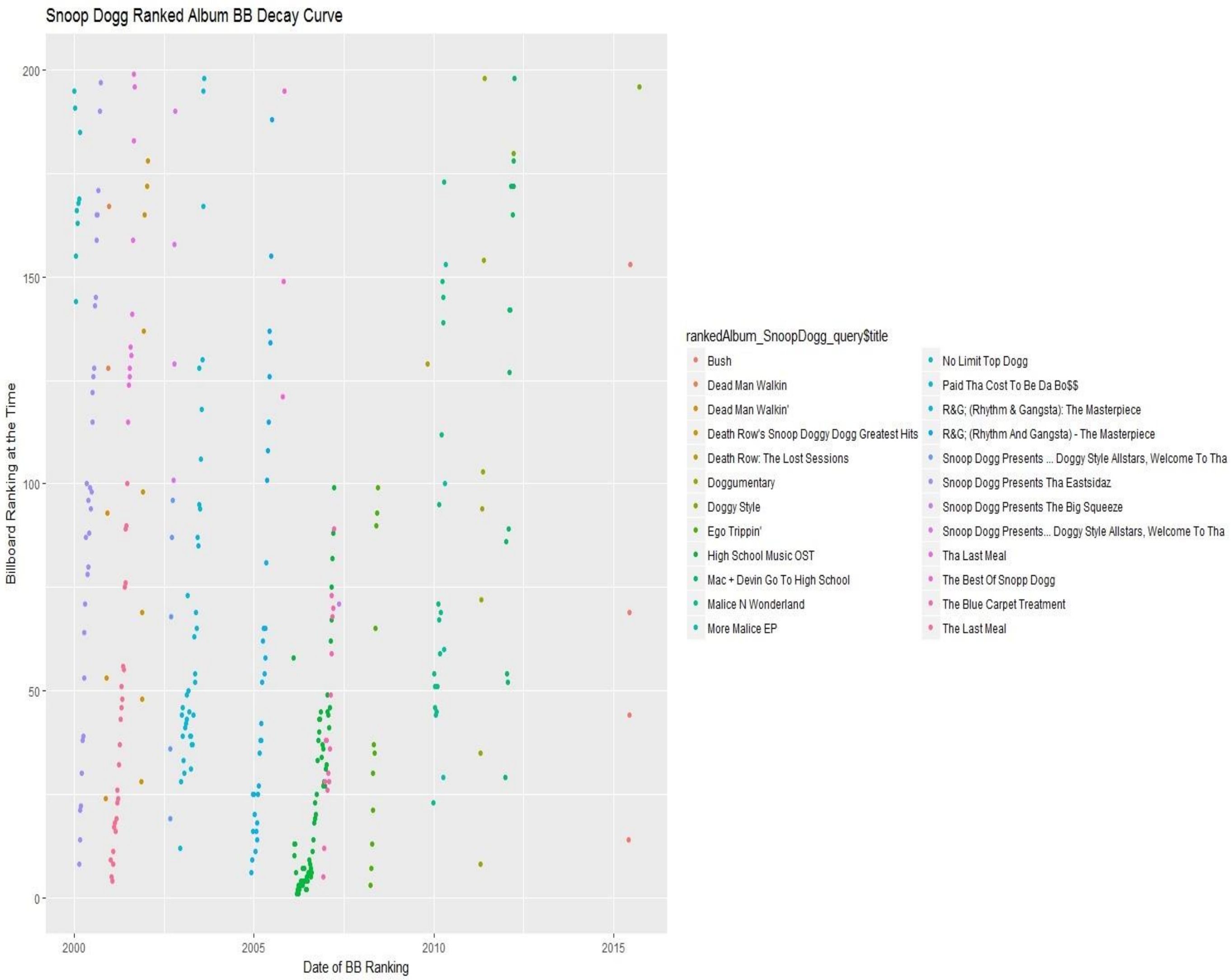
v. *Artist Tenure:*
It is the number of years the artist has been in the industry.

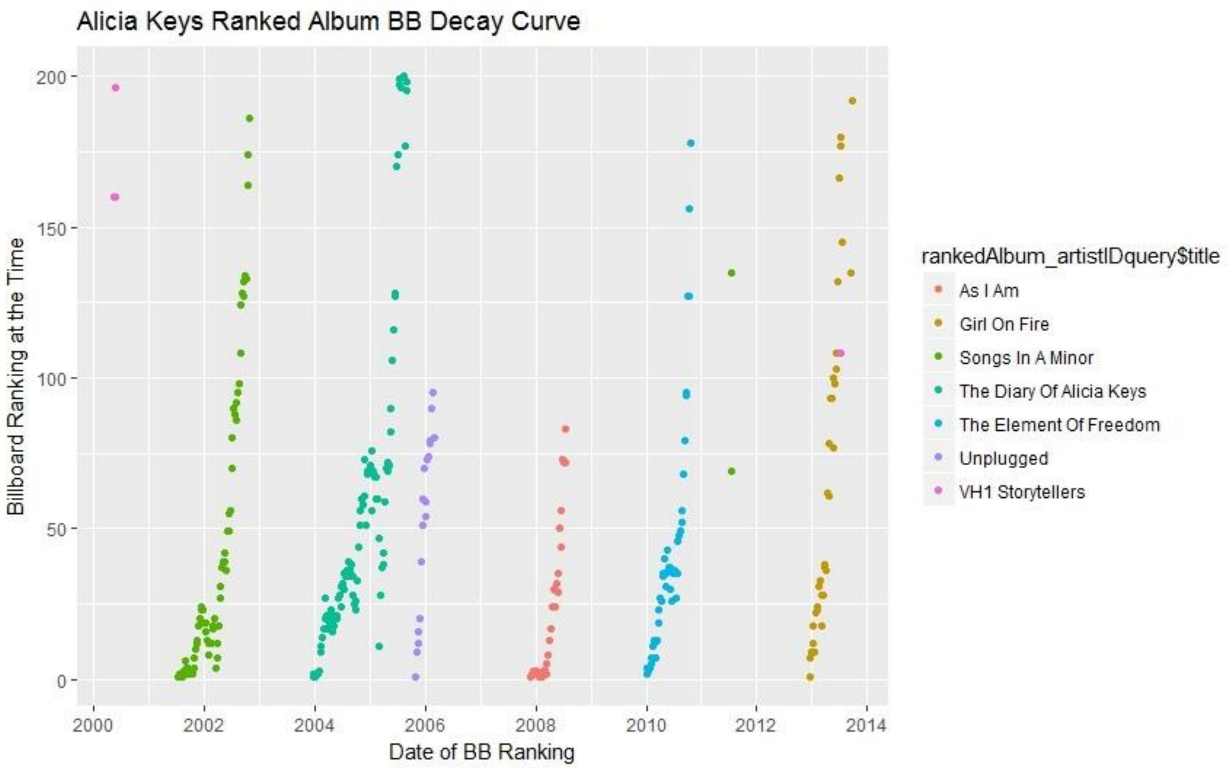
In the end, we had 5 Star Power features.

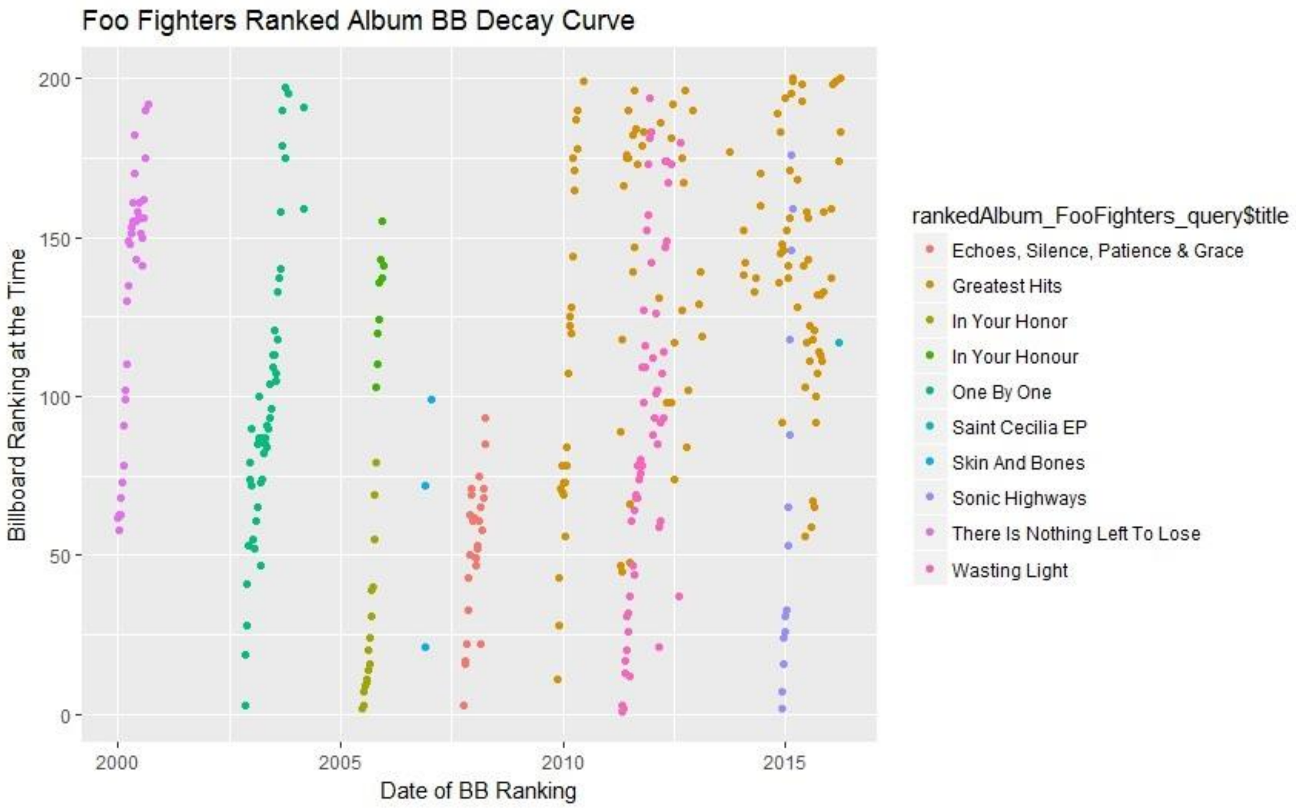
2. *Genre features:*

We did some exploratory analysis to create the genre features. We used R’s ggplot function to 3 plots of artists across different genres. The y-axis of the plots has years and the x-axis represents the billboard rank at that time.
The first plot shows Hip Hop genre through Snoop Dog’s albums, second plot shows R & B through Alicia keys albums and the third plot shows Rock genre through Foo Fighters albums.

While Hip Hop/ Rap decay out of the Billboard top 200 almost perfectly linearly with a steep slope almost equally distributed, R & B are more concentrated in the first 50 ranks. Rock genre also has some details that are different.




Alicia Keys Ranked Album BB Decay Curve
Billboard Ranking at the Time
200
150
100
50
0
2000
2002
2004
2006
2008
2010
2012
2014
Date of BB Ranking
rankedAlbum_artistIDquery$title
As I Am
Girl On Fire
Songs In A Minor
The Diary Of Alicia Keys
The Element Of Freedom
Unplugged
VH1 Storytellers


Foo Fighters Ranked Album BB Decay Curve
Billboard Ranking at the Time
200
150
100
50
0
2000
2005
2010
2015
Date of BB Ranking
rankedAlbum_FooFighters_query$title
Echoes, Silence, Patience & Grace
Greatest Hits
In Your Honor
In Your Honour
One By One
Saint Cecilia EP
Skin And Bones
Sonic Highways
There Is Nothing Left To Lose
Wasting Light

While it can be argued that these differences may not be genre specific as we have taken one artist, it is important to understand that there would be differences in the genre power. To study that we created the genre feature and we categorize them as follows:

i. *Number of albums in genre:*
It is the statistic that measures the number of albums released by an artist, in the genre and before the release date of the album in question. For instance, if the snoop dog has 10 albums released in Hip Hop, the "Number of albums" measure for his 11th album would be 10.

ii. *Number of ranked albums in genre:*
It is the statistic that measures the number of ranked albums released by an artist, in the genre and before the release date of the album in question. For instance, if snoop dog has 10 albums in Hip Hop, 6 of which got ranked, the "Number of ranked albums in genre" measure for his 10th album would be 6.
But, this measure is not enough in itself as rank set 1, 2, 3 and rank set 10, 20, 30 give the same answer for this feature which is 3. To consider the power of the rank we take the next feature.

iii. *Average of the rank of the ranked albums in genre:*
It is the statistic that measures the average rank of ranked albums released by an artist, in the genre and before the release date of the album in question. While we have the average, we still do not know the distribution of the ranks. To consider how scattered the ranks are, we take the next feature.

iv. *Standard deviation of the rank of ranked albums in genre:*
It is the statistic that measures how scattered are the ranks of the ranked albums by an artist, in the genre and before the release date of the album in question.

Music industry evolves and so does the popular genres. Let us consider that Hip Hop has become more popular now or Rock is not that popular. To take time into account, we created 4 sub features for each genre feature that we defined above. Each feature was sub categorized into past 1 year, past 3 years, past 5 years and complete past.

In the end, we had 16 Genre features.

3. *Record Label Features:*

Record labels are the most powerful factors as they can lure in the money and take care of the production cost. Having association with a well-known record label can have a massive influence on the rankings. Just like genre we categorized the record label features as follows:

i. *Number of albums by the record label:*
It is the statistic that measures the number of albums released by an artist, by the record label and before the release date of the album in question. For instance, if the snoop dog has 10 albums released in Hip Hop, the "Number of albums" measure for his 11th album would be 10.

ii. *Number of ranked albums by the record label:*

It is the statistic that measures the number of ranked albums released by an artist, by the record label and before the release date of the album in question. For instance, if snoop dog has 10 albums in Hip Hop, 6 of which got ranked, the “Number of ranked albums in genre” measure for his 10th album would be 6.

But, this measure is not enough in itself as rank set 1, 2, 3 and rank set 10, 20, 30 give the same answer for this feature which is 3. To consider the power of the rank we take the next feature.

iii. *Average of the rank of the ranked albums by the record label:*

It is the statistic that measures the average rank of ranked albums released by an artist, by the record label and before the release date of the album in question. While we have the average, we still do not know the distribution of the ranks. To consider how scattered the ranks are, we take the next feature.

iv. *Standard deviation of the rank of ranked albums by the record label:*

It is the statistic that measures how scattered are the ranks of the ranked albums by an artist, by the record label and before the release date of the album in question.

Maybe not as much as genre, but record labels can also become more popular with time or may lose popularity. Too incorporate this behavior we divided the record label features into 4 sub-categories as in past 1 year, past 3 years, past 5 years and the complete past.

In the end, we have 16 record label features.

We finally had our feature matrix which looked like this:

Star Power
Genre Features
Label Features
Albums

We ended up with a biased dataset which had 88% non-ranked albums and 12% ranked albums.

<u>Modeling:</u>

Now that we had our feature matrix ready, we had to run a binary classifier which could predict whether an album would make it to billboard top 200.

We choose to run two algorithms – Support Vector Machines and Random Forest.

Support Vector Machines

Support vector machines is a binary classification machine learning algorithm that performs well on non-linear data.

We chose python to run the algorithm. We used the following packages and libraries:

- *numpy and pandas:* for data manipulation
- *sklearn:* to use machine learning
- *SVM*: for running support vector machines (from sklearn)
- *cross_validation*: for cross validation (from sklearn)
- *grid_search*: for grid search for parameters used in support vector machines (from sklearn)
- *Pipleline:* To pipe the parameters (from sklearn)
- *Precision_recall_fcore_support:* for accuracy measures (from sklearn)

Support vector machines have 3 main tuning parameters – Kernel, C and Gamma.
While Kernel is rbf and regular depending on whether the data is non-linear or not, C is the misclassification error, gamma is rbf tuning parameter.
We used the following 6 values for C and 6 values for gamma with Kernel as rbf as data is non- linear and 10-fold cross validation.

C = 10e-5 to 10e5
Gamma = 10e-6 to 10e6

The top 2 results are as follows:

Parameter	Precision	Recall	F measure
C= 10, gamma=1	73%	31%	45%
C = 1, gamma=1000	70%	40%	50%

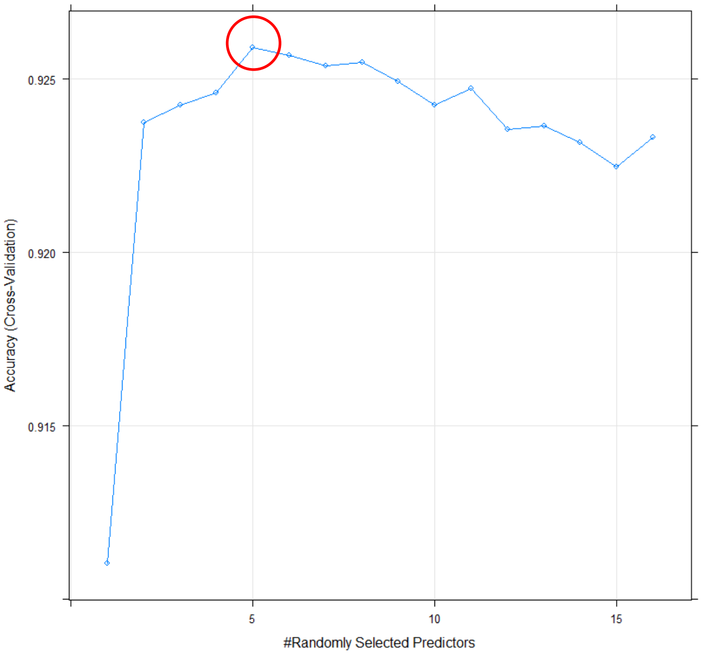
Random Forest

The next algorithm we tried was random forest. Random forest is a binary classifier which an ensemble technique which combines many decision trees.

We used R for running Random Forest and used the following packages:

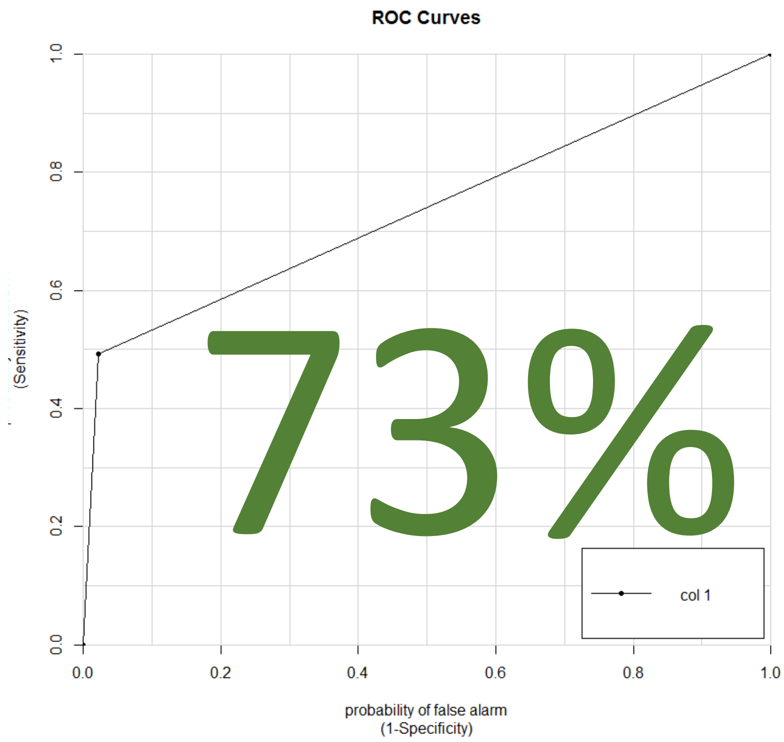
- *Caret:* Used for running random forest
- *Mlbench:* For random forest and features
- *Catools:* For calculating area under the curve accuracy measure

Random forest has one parameter to tune called “mtry”.
“Mtry” is the number of variables to be used at each split. We used the tune grid control parameter in caret package to input different values of “mtry” to find the best one. After using values 1 to 18 for mtry we found the best model was when mtry = 5. This can be found out by a graph that plots the model.



From the above graph, we see that the model is best when mtry = 5 as the accuracy peaks at 5.

After using the “catools” package we see that the area under the curve is 73% for it.



Conclusion:

From the above we conclude that our model performs best with mtry = 5 and that random forest outperforms support vector machines.

Assumptions:

We made a few assumptions for the sake of this project that helped us perform the modeling to the best. Our first assumption was to select only solo artists.
The second assumption was to take album data of only the albums that were released post 2000. The reason for this was that there was no billboard before 2000.

Contribution:

Data collection - We already had the data
Data cleaning - Ankit (50%) and Prateik (50%)
Modeling - Ankit (50%) and Prateik (50%)
Programming - Ankit (50%) and Prateik (50%)
Experiments - Ankit (50%) and Prateik (50%)
Paper writing - Ankit (50%) and Prateik (50%)