An Applied Data Science Project on

Predicting Song Success

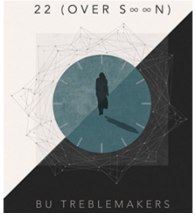
Abstract

Predicting the success of a song using data scraped from Spotify and Billboard websites.

Authored by:
Ameya Gokhale (adgokhal)
Sharath Kashyap (sharathk)

Advised by:
Prof. Artur Dubrawski

Special thanks to:
BU Trebblemakers and Midland Sparks (Indie Artists)
Jieshi Chen and Karen Chen (ADS Teaching Assistants)



(Images sourced from google search)

Introduction and Motivation

Every year, a number of individual music artists start their career by releasing fresh new albums on song aggregation platforms such as Spotify. The success of a song depends on various factors, but if there was a mechanism in place which would notify the right stakeholders the probability of success of a particular song, this could be beneficial to many indie artists to launch their careers. Moreover, knowing the probability of success is good information to have for both record labels as well as Spotify.

We started exploring the possibility of this project by talking to an Applied Data Scientist at Amazon, Dr. Saket Joshi, who also holds a PHD in artificial intelligence. After couple of talks we understood how we could help multiple stakeholders at the same time and combine good business and philanthropic work in one applied data science project.

To solidify our understanding of the musical features and what could potentially help upcoming artists we started collaborating with 2 upcoming artists, The BU Treblemakers and Midland Sparks. We got better intuition about what could possibly help upcoming artists and the problems they face in showcasing their talent in today's competitive environment. After many talks with these artists, we came up with 2 different use cases of our solution.

The main aim of this project is to help ease the process for upcoming artists to showcase their music and give an idea to the appropriate stakeholders about the potential of a song and an artist. We have tried to build our business case in such a way that all the stakeholders find it equally compelling to participate in the process and gains equally in terms of what they most desire from this product.

Executive Summary

This project aims to classify whether a song, given its musical and artist features will ever appear on billboard or not. Billboard is an entertainment brand owned by the Billboard-Hollywood Reporter Media Group. It publishes a variety of articles but for our project we concentrated on the Billboard top 50 music chart which the platform publishes every week. For a song to be published on Billboard is a huge success for any artist as the platform draws a very big fan base and it showcases the artist better.

Spotify has musical feature data on every song ever uploaded on Spotify; also, the platform stores data about artist followers on Spotify and the overall artist popularity score. Our model works per genre as after consulting various artists we understood that we need to make genre specific models which would enable us to only compare songs from the same genre. We decided to go with genre specific models as music evolves after some time duration, if 90's was all about Rock and Roll the early part of the 20th century was taken over by the electronic dance music also popularly known as EDM. To get rid of this discrepancy we decided the genre specific approach.

There were 2 main parts of this project - data engineering and model building. We did not have any readily available dataset on which we could directly start our analysis. We had to build a data set by combining API's of Spotify and billboard. The attributes were broadly classified into musical attributes, artist attributes and the target being whether the song has ever appeared on billboard or not.

We started building models for 2 different use cases:

- Does the artist have data on Spotify?
    - Artist features such as artist popularity and follower count were included as predictors

- Does the artist have no prior data on Spotify? (1st every Spotify release)
    - Artist features were not used as predictors

Most of our models performed well. But we are targeting high precision because we felt it was a better success metrics than accuracy. Of all the songs that our model predicted as a “hit”, we wanted very few of them to be false positives as it would destroy the trust aspect of our model. The trade-off is that the model let some true positives slip under the radar. However, this was a less sever business implication than predicting too many poor songs as hits.

One of the most predictive models gave us a precision of 92% for a target number of true positives. We have tried many classification models and experimented with a variety of hyper-parameters. After we built our models, we concluded that musical and artist features are an important set of predictors for whether a song will be published on Billboard of not.

Framework and Business case: The Stamp of Approval model

Stakeholders: Record labels, Individual artists, Spotify Ltd.

The stamp of approval model is a framework where Spotify acts like a 3rd party approval agency much like a notary. It vouches for whether a song given its features come on billboard top 50?

Stakeholder: Artist

If the artist wishes, he/she can get the Spotify stamp for success in terms of whether the song will be on billboard or not. Spotify will run the machine learning algorithm for that particular genre and let the artist know if Spotify thinks that the song has the potential to be on Billboard top 50. The artist can use this stamp to negotiate with record labels to sign them. They can use the certificate however they wish. The most important aspect of the certificate is that it has the brand name of Spotify on it which is a well-respected song aggregation website.

Stakeholder: Record Labels

Record labels are constantly scouring for talent and want to sign artists who will go on to produce blockbusters. Record labels want to identify artists before they become famous as they are in a better position to sign a deal before the artists become famous. Record labels can make use of these certificates by paying Spotify if they want some predictive information about an artist before they sign them. This way the record labels can decide whom to sign and how to write up their contract.

Stakeholder: Spotify

Through the predictive ability of the model Spotify will be able to predict if a song will be on billboard or not. We are setting an 80% threshold for the model to qualify a song as a “hit”. This prediction gives a sense of how well a song resonates with currently popular songs. Most songs that have come on billboard have earned their artists and related record labels millions of dollars. Spotify will not only make money of this model but will have bragging rights for identifying good talent in the industry and tomorrow can say, “we predicted the next star before he became one!”.

If an artist or record label opts for getting the certificate and if the song actually does appear on billboard then the concerned party will have to pay a pre-decided amount to Spotify for every week the song appears on billboard. This amount can be decided by Spotify with respect to the artist followers and the artist popularity.

Spotify could earn a considerable amount of money using the data that it already has while providing a valuable service to both artists and record labels. It’s a viable revenue stream.

Methodology and Data

Spotify comes up with a “popularity” score for songs. They are calculated by taking many factors into regard and fuel Spotify’s legendary recommendation engine. The algorithm used to calculate these scores is proprietary. However, from scouring the internet, we have found out that these scores attribute to the general likeability of music. These scores are important features.

At first, we were predicting this number, making the problem a regression problem. Here we were only considering the likeability of a song; i.e. the musical sophistication of the song, but after talks with industry experts, peers and artists, we understood that it takes more than musical sophistication to make a song popular.

Popularity of a song depends on the likeability if a song and the virality of the artist. The features of an artist are equally important to predict whether a song will be popular or not. To take into consideration the virality of a song we decided to include two more attributes which Spotify comes up with; but for artists. These are popularity of an artist and the number of followers an artist has on Spotify as a part of the predictor variables.

After including the artist features, we decided to change our target variable which would take into consideration likeability and the virality of a song. Billboard appearance is one of the best target variables as it takes into consideration both of these aspects. Hence, we decided to use this binary attribute as our target variable and started building classification models.

We at first decided to verify the proof of concept by limiting our models to work on Rock songs. We scraped Billboard to get the weekly top 50 rock songs for around 2 years. Then, using the song name and artist name of the scraped data, we sent HTTP requests to the Spotify web API with pagination to retrieve the details of all the songs from the album of that particular song and artist (billboard and non-billboard songs). To avoid any model bias resulting from collinearity between songs from the same album/artist, we scraped Spotify using the “Spotipy” python wrapper for additional random rock songs that were released in the same time period and not come on billboard. Finally, we ended up with a dataset of around 6000 tuples. This is a case of fusion of evidence by fusing features obtained from different sources.

The final list of attributes in our data set that we used for predictive modeling are listed below. We calculated the days since release from the Date of release field.

Musical Features:

- Danceability
- Acouticness
- Energy
- Instrumentalness
- Key
- Liveness
- Loudness
- Mode
- Speechiness
- Tempo
- Time Signature
- Valence

Basic Features:

- Duration of the song
- Days since release
- Explicit (y/n)

Billboard target:

- If the song appeared on Billboard top 50 (1/0)

Figure 1

Our hold-out validation dataset constitutes 20% of the overall data and represents the overall source distribution well. Out of the 1194 validation data points, 217 of the data points belong to the positive class.

Assumptions and Risks

Billboard releases a top-50 list every week and we found out from our research that Spotify also updates its popularity scores every week. Hence, we are dealing with a snapshot of the data to build model. From exploring the billboard song data that we retrieved, we discovered that around 60% of the songs repeat from the previous week. Hence, our model has a predictive utility of approximately a fortnight. However, ideally, the model must be updated every week with the latest data. Our data was scraped on Nov 10, 2018.

We would like to have worked with time-series data essentially. However, gathering the data is very resource and time intensive as it took around 3 seconds to retrieve a single complete record of the data. Furthermore, Spotify doesn't let us query its continuously updated details by the date. This limited our capacity to gather data every week and analyze trends over time.

We have tried to much of the "song popularity" related noise into consideration. Different songs have different marketing strategies, social media strategies, record label investments, music videos, etc. We have proceeded under the assumption that Billboard and Spotify data present an accurate representation of the aspects of a song's genre that are perceived as popular.

Furthermore, several externalities influence Song popularity. For instance, several songs of the band Queen popped up in last week's billboard top 50. This can be attributed to the release of the movie, "Bohemian Rhapsody" during the previous week, which is a biopic of the band. This would have led to a large number of people reminiscing about the old songs from this legendary band which led to it being talked about more on social media and blogs and thereby creating a listening hype for the songs. Our model does not take these externalities into consideration.

Building the Models

For the classification models our target variable was whether a song will appear on Billboard or not. The independent variables had two broad feature sets; music features + artist attributes for established artists having prior Spotify data and the second set only had musical attributes for artists having no prior data on Spotify.

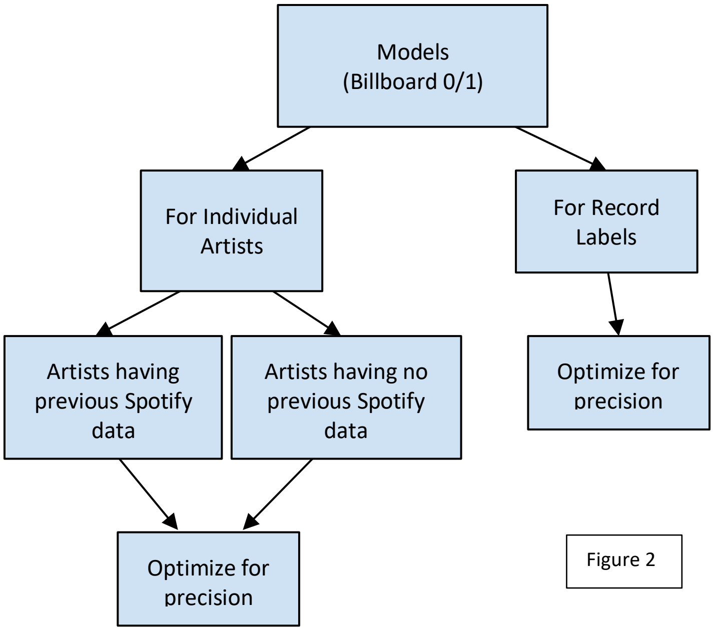


Figure 2

1. Artist Followers - 0.3808
2. Explicit (1/0) - 0.0683
3. Danceability - 0.0652
4. Instrumentalness - 0.0642
5. Artist_popularity 0.0589
6. Acousticness - 0.0589
7. Liveness - 0.0563
8. Duration - 0.0493
9. Tempo - 0.0486
10. Speechiness - 0.0463
11. Loudness - 0.0462
12. Energy - 0.0232
13. Valence - 0.0139

Figure 3

Our dataset was imbalanced in that only 21.75% of it had rows of the positive class. We used SMOTE (Synthetic Minority Over-Sampling Technique) to reduce this imbalance and use a 1:1 ratio to train the model. The thresholds for the two different use cases are different as we are optimizing different success metrics. All the models have been optimized for best hyperparameters using grid search with cross validation and refit. The models are validated using a hold-out validation set of the original data.

Also, we wanted to know which attributes were important for predicting whether the song appeared on Billboard or not. Following is a list of features ranked from most to least important. Top 13 features have been shown below in figure 2.

The Classification models for artists having previous Spotify data

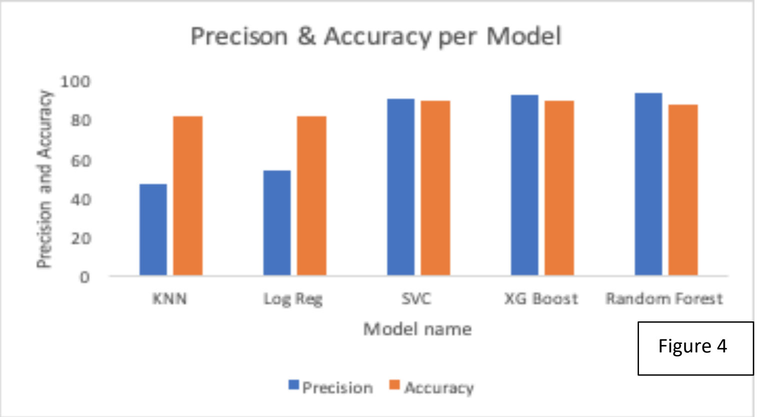


Figure 4

From the adjacent diagram, we see that the Support Vector Classifier, Random Forest Classifier and Gradient Boosting Classifier (XGBoost) have the highest accuracy and precision for the model built to make predictions for semi-popular artists with previous Spotify data.

We selected the best model by studying the ROC curves and AUC scores for our specific use-case.

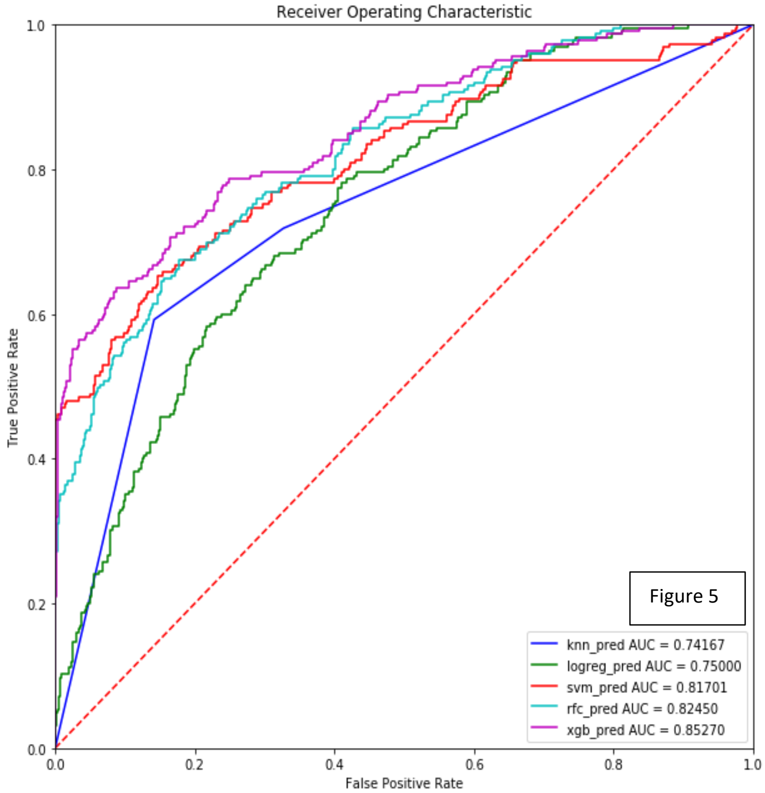
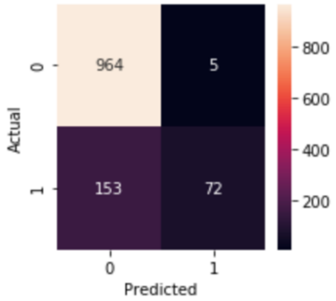
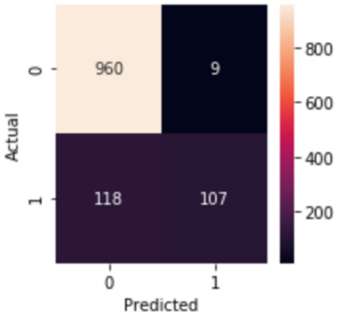


Figure 5

Random Forest



XGBoost



XGBoost is clearly a better performer except for a very minute portion in which Random Forest performs better.

We studied the confusion matrix to choose the best model. We found that the Random Forest Classifier is not able to hit our target number of true positives (100). XGBoost had 107 True Positives and 9 False Positives whereas Random Forest had 72True Positives and 5 False Positives.

Hence, we decided that the XGBoost classifier a better model for our use case and target requirement of 100 True Positives at a threshold of 0.8.

Below are specific metrics that describe the performance of our models at the preset threshold.

Model Name	**Accuracy**	**Precision**	**True Positive rate**	**False positive rate**	**F-Score**	**Threshold**
KNN	81.23%	46.56%	59.22%	14.17%	0.52	0.8
Logistic Regression	81.57%	53.24%	18.22%	3.7%	0.27	0.8
SVC	89.11%	90.59%	47.11%	1.13%	0.61	0.8
XG Boost	89.36%	92.24%	47.55%	0.9%	0.62	0.8
Random Forest	86.76%	93.50%	32%	0.5%	0.47	0.8

As we can see, for our chosen precision target, there are many misclassifications in terms of false negatives. However, these are less severe than false positives. False negatives represent a lost opportunity for Spotify in terms of not being able to spot a good artist. However, it is still better than labelling a poor artist as a good artist as this will cause distrust among the stakeholder's about Spotify's value proposition through this model. A similar trend is observed in the following model which is used for predicting popularity for new artists.

The Classification models for artists with no previous Spotify data

A similar methodology is followed in this section. However, we are excluding the artist features like Artist popularity and number of followers the artist has because a new artist won't have these features.

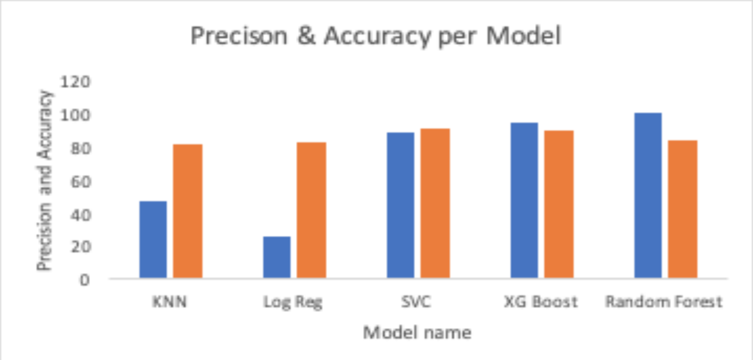


Figure 6

From the adjacent diagram, we see that, the Random Forest Classifier and Gradient Boosting Classifier (XGBoost) have the highest accuracy and precision for the model built to make predictions for new artists with no previous Spotify data.

We selected the best model by studying the ROC curves and AUC scores for our specific use-case.

We noted that our models performed significantly better than the default models which predict every song in the validation set as a "not-popular" song and had a Recall score of 0. This can be attributed to the class imbalance as around 80% of the tracks in our dataset did not show up on Billboard.

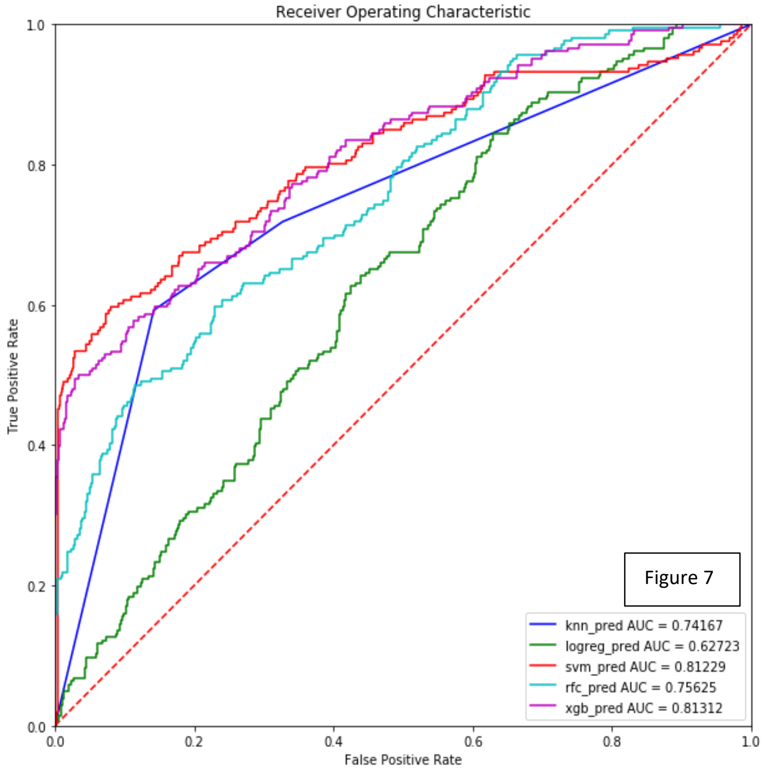
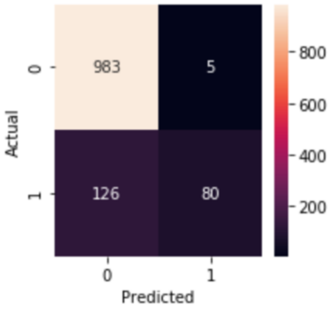
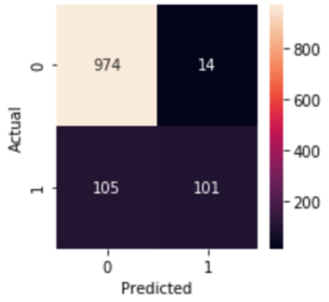


Figure 7

XGBoost



SVC



From the ROC, it is clear that the in some areas, XGBoost classifier is a better performer and in others, it's the Support Vector Classifier.

We studied the confusion matrix to choose the best model. We found that the XGBoost Classifier is not able to hit our target number of true positives (100). SVC had 101 True Positives and 11 False Positives whereas XGBoost had 80 True Positives and 5 False Positives.

Hence, we decided that the SVC classifier a better model for our use case and target requirement of 100 True Positives at a threshold of 0.8.

Below are specific metrics that describe the performance of our models at the preset threshold.

Model Name	**Accuracy**	**Precision**	**True Positive rate**	**False positive rate**	**F-Score**	**Threshold**
KNN	81.23%	46.56%	59.22%	14.17%	0.52	0.8
Logistic Regression	82.57%	25%	0.4%	0.3%	0.009	0.8
SVC	90.03%	87.82%	49.02%	1.4%	0.62	0.8
XG Boost	89.02%	94.11%	38.83%	0.50%	0.54	0.8
Random Forest	83.41%	100%	3.8%	0%	0.074	0.7

Through the results, we can conclude that musical features are very important in predicting the popularity of a song. Certain musical features correlate well with the possibility of a song appearing on billboard.

The Longshot

We are always encouraged to try and work on new ideas to solve our data science problems. In light of this, we tested another idea in our analysis. We experimented with Principal Component Analysis and used it to decompose the Artist Popularity, Track Popularity and the Number of times the song has appeared on Billboard into a singular value via the first principal component; our "Song Popularity Index".

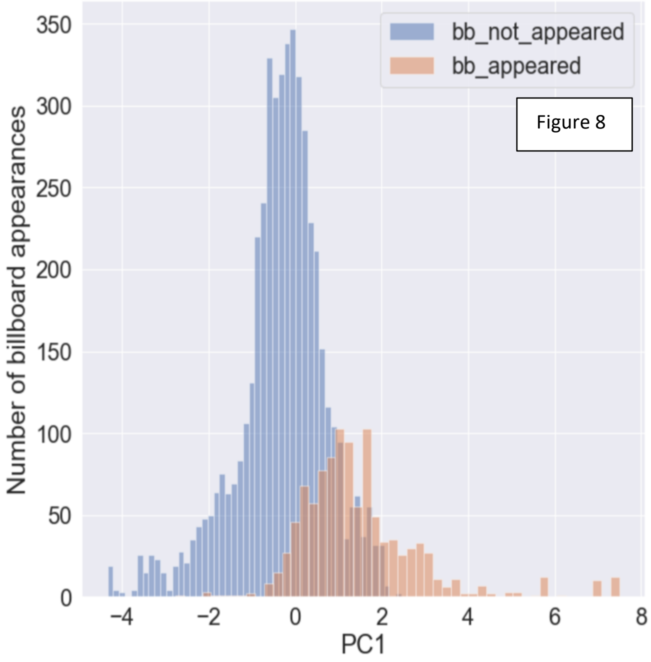


Figure 8

The PC1 captured around 60% of the overall variance among the three attributes. From observing the loadings, we saw that this index has loadings of around 0.4, 0.5 and 0.25 for track popularity, artist popularity and number of times the song has appeared on billboard. The hypothesis was that songs that score highly on our created index correlate well with songs produced by artists who are popular, have created songs that are popular and have songs that have appeared on billboard one or many times.

We normalized this value to fit it between 0 and 1 and tried some regression models to predict this index. To understand the model performance, we used the threshold as a measure of the probability that a song would come on billboard. We drew up ROC curves and evaluated the resulting confusion matrix and got surprisingly good results. Figure 8 shows how our index separates songs that have and have not appeared on billboard.

We tried Ridge, Support Vector and Decision Tree Regression to predict this popularity index. All models gave similar results with the ridge regression model performing slightly better than the rest. The root mean squared error was lowest for Ridge. It gave a surprisingly high accuracy and recall score although the precision was low.

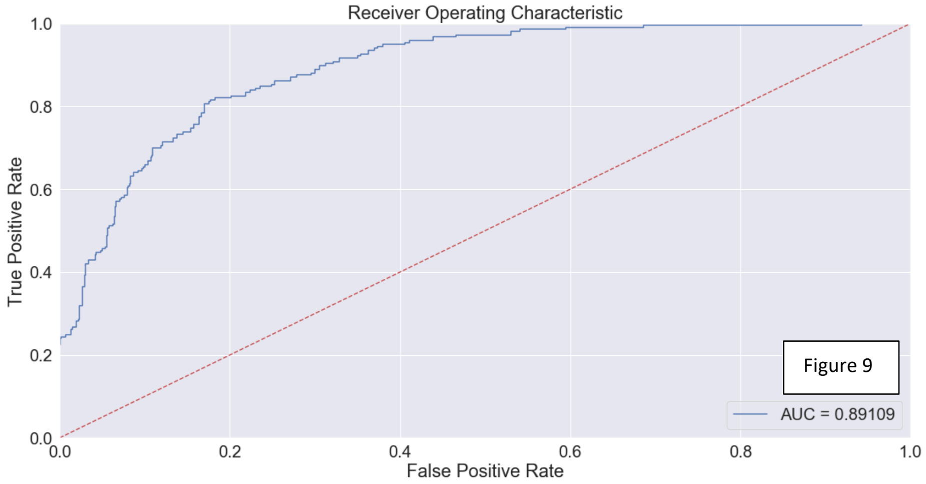
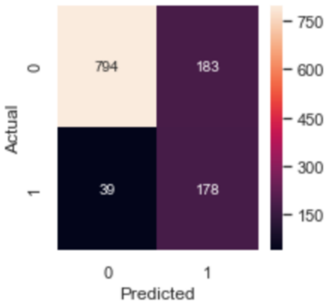


Figure 9

TPR	0.82
FPR	0.18
Accuracy	0.82
Precision	0.5
F score	0.62



This process was an experiment to understand the dynamics of how a song's features correlate with a single score that captures the popularity scores from Spotify and number of billboard appearances. It is too soon to tell if this model has any serious predictive utility. It is tough to explain quantitively what the predicted popularity index represents. The reconstruction error of around 40% makes it difficult accurately get back the original values.

Future Work and Conclusion

There is quite a bit of work which can be done to improve the work done by us in this project. We saw the problem statement from more of a proof of concept perspective, where we were testing the hypothesis whether we can predict if a song will appear of Billboard or not given its musical and basic features. According to our research and analysis through this project we have come to a conclusion that we can predict the success of a song given its musical and basic features. Music sophistication is not the only criteria when it comes to predicting success of a song. Virality of the song and popularity of the artist are equally important features.

After this system is deployed, we would like to improve it further using the following ideas:

- Our goal is to deploy this model as a mainstream revenue earner for Spotify. Currently, we are setting a very high standard to declare a song as popular (0.8). This high standard is letting a lot of potentially good songs to slip under the radar (low recall) which is basically a lost opportunity to earn revenue. However, this is necessary to increase the awareness and recognition of the model as a strong and reliable predictor of popularity.

- Once this model has gained recognition, efforts have to be made to increase the overall accuracy of the model. Collaborations with Billboard and other websites that aggregate song data could potentially provide a number of relevant features that our proof of concept doesn't include. Furthermore, based on these features, a number of new and innovative models could be built to predict targets such as the number of times a song is expected to appear on billboard and whether a song is going to win a Grammy award or not. These could be other stamp of approval certificates that Spotify could issue. Efforts could be made to include social media buzz surrounding artists and songs as inputs to the model

- In this project, we have focused on building models to predict popularity of rock songs. Future work would need to include models to be built for predicting the success of songs from other genres as well. An approach that we have thought of is to use leader clustering to ascertain the song's genre through its features and then predict its future popularity/success metrics using the models that have worked best for songs that share similar features to this input song. Different genre models would encourage more artists to take part in gaining the certificate and understanding what the probability of success for their song is.

- As mentioned, we were constrained by limited bandwidth and processing power and had to work with a snapshot of the actual data. However, we feel that the model would benefit from time-series data which Spotify has easier access to as the source. Models must be trained and redeployed weekly to ensure that they are sensitive to the most current trends and are predicting metrics based on the latest trends.

- Being able to come up with a recommendation system on how artists could improve their songs to bring them more in line with songs that are currently a success is another important step in the pipeline. Again, introducing this could help Spotify provide another service out of which they can make more money.

This model would be beneficial to Spotify as there is not much of investment to be put into this project. Spotify already has all the data and the billboard data is freely available to scrape. This model could add two constant revenue streams. Spotify will have to work on spreading the word about its certificate extensively so that it is accepted in the musical community.

Spotify will have to actively engage with new and popular artists to understand any reservations they might have with the certificate. Spotify could crowdsource the information and understand what features could be added to the model to further increase its utility in the musical community.

As for the target variable we used, if a new artist's song is predicted as a true positive by the model, it can be interpreted that the song correlates well with the features of past songs that have shown up on Billboard. It gives record labels or artists a hit that it may be a track or an artist that has potential and may be worth checking out and pursuing. This is where our model for Spotify finds the most business utility. Without this model, record labels would have no indicator or helper to scout out fresh talent that they would want to get into business with and artists would have no frame of reference against which they can compare their track to understand where they stand with their song.

In conclusion, we see our Stamp of Approval model as an easy way for Spotify to earn revenue with freely available data while providing a valuable service to artists and record labels.

Appendix

Keyword	**Description**
acousticness	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
danceability	Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
energy	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
instrumentalness	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
liveness	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
speechiness	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
tempo	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
valence	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

(Source: https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/)