**DSC1 5330 Project Proposal:**

**Decoding What makes a Chartbuster:**

**Predictive Modeling of Song Popularity based on**

**Spotify Audio Features**

**Proposed By:**

**Yusuf Saadu**

**Executive Summary**

This project seeks to build a predictive model that calculates the popularity score of a song using its audio features. This predictive model is to be used in the music industry by labels to select the best songs to invest in marketing wise to release to appeal to the masses. This is a very important decision as the correct choice not only results in mass revenue through sales, streaming, touring, licensing, sponsorships, sampling royalties, etc, it also prevents a waste in finite resources that could go to other artistes and their songs. Similar to the stock market, placing your money behind the right song is a great investment.

The dataset being used is sourced from Kaggle where it was scraped from Spotify using their API. The dataset is fairly huge with at least a million rows and 20 columns. Out of these columns, only those having to do with audio features are being utilized to build the model. Variables like key, tempo, duration, danceability, energy, etc. are to be used for this predictive model. The algorithm used to build this model will have to be a regression since we’re trying to find a numerical result. Linear, Random Forest, Support Vector and Gradient Boosting are all on the table. A final selection will be made based on how they each perform with the lowest MAE and RMSE. The data will be split into three for training, validation and testing. The result should be a predictive model that can accurately predict the popularity score of a track who quantified audio features have been put in.

This project and the resulting model don’t just save money for the music labels but it also provides a blueprint for the artistes who are trying to make their dreams come true and have their art be heard. The model won’t be perfect due to the constant evolution and change in music taste but I feel this is a a great first step towards guaranteeing a form of success in the competitive market of music and entertainment.

**Introduction**

When it comes to the music industry, the value of a song comes in many forms. Careers are made and broken by singular tracks with One hit Wonders able to coast and grow generational fanbases off one song. Revenue is generated from hard copy sales, digital sales, streaming, touring, licensing and even sampling. One iconic and hugely popular track can keep an artiste fed and the music label that signs them paid for decades. However, it isn’t easy to get a song to the level of fame or popularity required to reap such heavy long-lasting rewards. While sometimes, songs can pop off and get hot randomly or in other words organically, a lot of the times the biggest songs that dominate the airwaves and our collective consciousness are put there through large marketing budgets and clever strategies. Music labels and their marketing departments do and spend a lot to put songs and artistes in as many ears and in front of as many eyes as possible. They use their resources and influence to place their songs in the most convenient place to catch attention and hopefully favor. This can be achieved in a myriad of ways but doing so is not cheap and resources are finite.

Therefore, record labels need a way to be sure of the star potential of the song they’re putting money behind. A prodigious hardworking artiste creates several songs each session and in the creation of an album could create up to a hundred songs not to mention how many unreleased they would have over the span of their career. Chris Brown the platinum selling artiste reportedly has 15,000 songs unreleased in his vault. Normally the song choice decision is left to the producer in conjunction with the artiste as faith is put in them by the label to determine songs for release and especially single rollout that would really connect with the audience but what if there was another way. What if there was a data driven and analytical method to determine the popularity potential of a song created by an artiste prior to release? Therefore, the label could confidently put their money behind songs bound to blow up rather than sinking money into flops that irritate the masses when they come on the radio or playlist.

**Problem Statement, Objective & Scope**

The problem of artiste song selection whether for a simple album release or more importantly a single push is one that can be likened to the selection of the right stock to invest in. Models have been made hoping to tame the beast that is the stock market and mitigate as much risk of failure as possible. Why not then apply this same principle to the music industry where resources are finite, and the value of a hit extends well beyond the industry. From purchases to movie/tv show/advertising licensing not to mention live performances, sampling and streaming royalties. All that and more are what a label and artiste stand to benefit by picking the right song to push to the masses. Failure to do so negatively affects the credibility, brand and profile of the artiste while wasting the resources of the label that put a lot into not just making the music but also advertising it.

This is the issue I am attempting to tackle in this project that I am proposing. Artists create so many tracks of various quality, so how are they meant to select the right song to reach the masses? By training a predictive model that utilizes Spotify audio features to produce a popularity score. I hope to use this resulting metric as a measure of potential for Music Labels to bet on in terms of resource and investment. The objective of this project is building a model that predicts the potential popularity of a song based on various audio features that come in numerical form. By using audio features found in the Spotify database such as speechiness, acousticness, instrumentalness, liveness, valence and tempo. With Spotify being one of, if not the largest and most popular music streaming services in the world, I feel like their data would give an accurate measure of the popularity and commercial potential of a track. Utilizing historical Spotify data gathered through API, the goal is to train the model to a point where artistes’ track can be analyzed and their quantified audio features can be loaded into the model to produce a popularity score and the highest scoring song will be released to the masses and push by an advertising budget. That way the artiste and label are playing their best hand in the competitive music industry where day by day the barrier to entry gets lower and it becomes harder to stand out from the crowd. Labels have the power and resources to make their artistes’ releases stand out, the only issue is whether the song will be received well or not. This model will work to make sure that when a song is selected to be set at the front of the pack by labels, it will have the best fighting chance in winning over the hearts and minds of the masses.

Of course, this does not mean that the model will be an absolute solution. The popularity variable merely tracks how many people are playing a track and how recently they’re playing it. The model will utilize data from songs spanning multiple genres, artistes and eras. It must be noted that there are numerous factors that go into the blowing up of a song aside from its audio features or how good it is. Numerous old songs have come back into the spotlight and charted simply for being in the right place at the right time. Take the example of Kate Bush’s song “Running Up That Hill (A Deal with God)” released in 1985 that resurged in popularity simply due to it being featured in the hit Netflix tv Show “Stranger Things”. Songs can gain popularity for a myriad of reasons aside from how they sound, it could be the artiste or the story around it. Also, general music tastes change with time as songs and sounds that were popular in their day may not be popular now though one cannot discount the nostalgia effect. I believe the project to be interesting as it seeks to map out what kinds of songs people enjoy listening to.

**Literature Review**

I haven’t done extensive literature review on this topic, yet I have managed to gain access to some research articles on the topic of song popularity prediction model and machine learning using audio features, but they seemed to use a wholly different types of dataset and more complex algorithms. The articles will be found in the Reference Literature section. I plan to revisit them at a later date for more intensive analysis. I feel like I could learn a lot from these research articles.

**Methodology**

This project utilizes a dataset sourced from Kaggle that was compiled utilizing Spotify API. The dataset has more than a million rows and 20 columns. The dependent/predicted variable will be Popularity which measures how many plays a track gets as well as how recently people listened to it. It ranges from 0 to 100 with the higher numerical value being the more popular and vice versa. Meanwhile the independent variables will include:

* Genre: Serves as a categorical variable that delineates the musical style or category of the track.
* Danceability: A numeric variable indicating the suitability of a track for dancing based on various musical elements, ranging from 0.0 (least danceable) to 1.0 (most danceable).
* Energy: A numeric variable measuring the intensity and activity level of a track, ranging from 0.0 (least energetic) to 1.0 (most energetic). Energetic tracks are typically fast, loud, and noisy.
* Key: Categorizes the key to the track using integers that map to pitches in standard Pitch Class notation, e.g., 0 = C, 1 = C♯/D♭, 2 = D, etc.
* Loudness: Quantifies the overall loudness of a track in decibels (dB), averaged across the entire track, typically ranging between -60 and 0 dB. It is useful for comparing the relative loudness of tracks.
* Mode: A binary variable indicating the modality of a track, where 1 represents major, and 0 represents minor.
* Speechiness: Refers to the proportion of spoken words in a track. Values closer to 1.0 indicate a higher prevalence of speech-like recording.
* Acousticness: A confidence measure ranging from 0.0 to 1.0, indicating the likelihood that the track is acoustic. A value of 1.0 denotes high confidence in the track being acoustic.
* Instrumentalness: Denotes the likelihood that a track contains no vocals, with values closer to 1.0 indicating a higher probability of being instrumental.
* Liveness: Indicates the presence of an audience in the recording. A value above 0.8 suggests a high probability that the track is a live recording.
* Valence: Depicts the musical positiveness conveyed by a track. Higher values indicate more positive-sounding tracks, e.g., happy, cheerful, euphoric, while lower values suggest more negative-sounding tracks, e.g., sad, depressed, angry.
* Tempo: Represents the estimated overall tempo of a track in beats per minute (BPM), reflecting the speed or pace of a piece of music.
* Duration\_Ms: Specifies the duration of the track in milliseconds.
* Time\_Signature: Denotes an estimated overall time signature of a track, specifying how many beats are in each bar (or measure) following musical notation conventions.

**Data Preprocessing & Analysis**

Given that the dataset being used is so huge, care needs to be taken to not miss out on any errors in the data. Before any analysis or model training can be done. The data needs to check for missing and null values. Upon identifying them, the plan is to impute the ones that can be figured out and drop the ones that can’t. I noticed that some entries had symbols for artiste and track name. While I will try to complete them as best I can, I’m not really using artiste names and track titles in the model so rectifying the errors there is unnecessary. I do also plan to check for duplicate values and outliers, but I doubt those will be present. The dataset seems to be pretty balanced in terms of popular and non-popular songs. I plan to encode the categorical variables as well as the dummy variables. Data transformation might not be necessary for this project in terms of normalization or standardization as the dataset seems to be pretty simplified in data structure and range.

I intend to analyze and visualize the data using Tableau. The plan is to track the yearly trend of each variable and their correlation to each other to suss out multicollinearity. I also plan to build scatter plots of each variable against popularity to parse the relationship overall and year by year. This is simply to identify the change in trends over the years because as most know different kinds of sounds become popular in different time periods. Given that the dataset is so huge, analysis using tableau hasn’t been carried out yet as process request times tend to stretch. Once a solution has been figured out to get around that issue, analysis can begin. There’s been thought of utilizing a different smaller dataset, but we’ll see. Upon a cursory visual analysis, the dataset seems to be extremely diverse in genres, years, artistes, sounds and popularity which I feel will serve as a huge strength to the prediction of a song’s popularity.

**Model Selection, Training & Evaluation**

When it comes to the type of algorithm, I intend to use to build my predictive model, since the result I’m trying to produce comes in the form of a numerical value (Popularity), I made the decision to use regressions. I might use Python or R. So far, I haven’t decided exactly what kind of algorithm to use but I will probably just train 4-5 models using different models and choose the one that performs best. The algorithms I plan to use include Linear Regression, Random Forest Regressor, Gradient Boosting Regressor and Support Vector Regressor. I chose Linear for its simplicity and the rest for their ability to take on more complex relationships. I was thinking about using Neural Network, but I don’t hold confidence in my ability to execute it, but we’ll see.

The plan is to split the dataset randomly into 3 subsets, one for training, one for testing and one for validation. The idea is to do a 70-15-15 split randomly. The models will then be trained on the training subset before being validated and compared against each other using the validation subset with the final model selected being tested using the test set. There’s the option to use K-fold cross validation but I intend to hold off on that till I see the results of testing. The final model will be evaluated on various metrics but primarily on the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). I intend to choose the model that performed with the least RMSE and MAE as that would give me a good measure of the models’ prediction accuracy. The smaller the difference between the predicted values and actual values, the better the model seems to perform at accurately producing the right popularity score based on the input audio features

**Conclusion and Expected Outcomes**

At the end of this project, I hope to have a competently accurate working predictive model that takes in a bunch of data on a track’s or list of tracks’ audio features and produces a popularity score that gives a popularity score that can serve as a good measure potential for the song’s mass appeal. Of course, the model and prediction won’t be perfect. Popularity of a track ebbs and flows. We are in the microwave era where songs come and go. They blow up and vanish like dust but that is something that can be built on. This project aims to identify a hit, a chart buster, a billboard topper and a viral ear worm. By the end, I want to have a working model that can predict this.

By doing this I feel like incorporating real time data would be much easier. I’ll have to look more into scalability but by using Spotify’s audio analyzing algorithm to quantify the audio features of any track, music labels can easily find the gems in their collection that can be invested in and pushed to the masses. This is important for the industry as a whole as it prevents a waste of resources and customer dissatisfaction from being bombarded constantly with a song they hate. Similar to the stock market, this model hopes to allow music labels to make quality data driven investments in their artistes and their art. That way, the returns on investment can be invested back into new artistes and the industry continues to thrive.

**Research Literature References**

* Araujo, C. S., Cristo, M., & Giusti, R. (2019). Predicting Music Popularity on Streaming Platforms. *Anais Do Simpósio Brasileiro de Computação Musical (SBCM 2019)*. <https://doi.org/10.5753/sbcm.2019.10436>
* Yutong, G., Yutong, S., & Jiaqian, W. (2021, May 26). Popularity Prediction of Music Based on Factor Extraction and Model Blending. Ieeexplore.ieee.org. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9434816>
* Beniwal, R., Gupta, S., Shekhar, S., & Batra, S. (2023, August 7). Hindi Hit Songs Prediction Using Machine Learning Algorithms. Ieeexplore.ieee.org. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10205626>
* Kamal, J., Priya, P., Anala, M. R., & Smitha, G. R. (2021). A Classification Based Approach to the Prediction of Song Popularity. 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 2021 International Conference On, 1–5. <https://doi.org/10.1109/ICSES52305.2021.9633884>
* Martin-Guiterrez, D., Hernández-Peñaloza, G., Belmonte-Hernandez, A., & Álvarez García, F. (2020, February 24). A Multimodal End-to-End Deep Learning Architecture for Music Popularity Prediction. Ieeexplore.ieee.org. <https://ieeexplore.ieee.org/abstract/document/9007339/citations#citations>
* Pham, J., & Kyauk, E. (2015). Predicting Song Popularity. <https://cs229.stanford.edu/proj2015/140_report.pdf>
* Essa, Y., Usman, A., Garg, T., & Singh, K. (2022). Predicting the Song Popularity Using Machine Learning Algorithm. <https://ijsret.com/wp-content/uploads/2022/05/IJSRET_V8_issue2_281.pdf>