**Predicting Song Popularity on the Spotify Platform**

**Project Proposal**

| **Member 1** | Purna Srivatsa Boggavarapu | purnasrivatsa@vt.edu |
| --- | --- | --- |
| **Member 2** | Marti McElreath | martimc@vt.edu |
| **Member 3** | Nihitha Veeramachaneni | nihithav@vt.edu |

**Introduction**

Hit Song Science, or the practice of predicting music popularity through machine learning, is a well-known and well-studied issue in artificial intelligence. Knowing what songs will become hits has obvious commercial and cultural significance, allowing song-writers and producers to pool resources into songs that are most likely to sell big and giving researchers insight into popular musical trends. In an increasingly global world where popular media is shared across national borders, it is also invaluable to predict what markets a song or artist will be most successful in. By looking at Spotify music data, we aim to train a model that will be able to accurately predict the countries in which songs are (or are destined to be) hits.

**Project Problem Statement**

The question of what makes a song popular is important to those that make music, to those who are interested in what affects music listenership, and to those who earn millions from making the songs and artists under their management big. In 2020 it is estimated that the global music industry’s annual revenue was almost $60 billion, a figure that has grown year-on-year with no end in sight. Consequently, the value of knowing what songs will become best-sellers in a given place is immeasurable.

Machine learning is the ideal tool for teasing out exactly what it is about a song or artist that made it so popular in certain geographic locations, an insight that can be easily translated into actionable changes for budding artists and established record companies alike.Prior work in this area has been conducted using music and artist features provided by Echo Nest’s Million Song Dataset to train classification models such as Logistic Regression and SVM (Stanford University and University of Antwerp), and additional research has focused on related issues such as predicting what songs a specific user will listen to next (Matthew Moocar).

We plan on tackling the adjacent problem of location-based popularity predictions by using a Spotify dataset with both artist and music metadata to identify the features most highly correlated with song popularity and draw conclusions about what drives music popularity across space and time. We believe Spotify in particular is a very promising target, as its Spotify Popularity index is used heavily by both artists and listeners to rank songs against one another. Spotify is also the most popular music streaming service on the market today, giving its data unique weight when measuring listener sentiment.

**Data Set**

We are planning to use the [Spotify Huge dataset](https://www.kaggle.com/datasets/pepepython/spotify-huge-database-daily-charts-over-3-years?select=Final+database.csv) which is publicly available on Kaggle for our study. This dataset has 170663 records of songs with their characteristics and popularity with respect to 36 countries. This data is retrieved from Spotify using their API. We aim to predict the popularity of a new song in a specific country with the help of this data by building a supervised machine learning model. There are about 30 features provided that include details about the songs like Genre, Danceability, Energy, Loudness etc. and these can be used to train a model to learn the patterns that tend to make a specific type of song popular in a country. These features are a mix of categorical and numerical values. The categorical values have been converted to numerical values using one hot encoding. With this there are about 150 columns of data that can be used to train the model.

The target variable is popularity which is a numerical quantity such that a higher score means higher popularity.

**Preprocessing steps**

We will be doing some preprocessing steps over our data so that it can be ready to be used for machine learning algorithms. We will be checking for any missing values and populating them with the most likely values or removing the records completely if they do not have much value. We will also specifically check if a song’s data is missing for any of the 36 countries. We will also check if there are any outliers and handle them accordingly. Finally since all the features have different ranges, we will be normalizing the data so that all the features have similar ranges.

**Methods and Models**

The machine learning problem for this dataset is “Regression”. We would be predicting the popularity of the song which is a numerical value. The songs have positions or rankings assigned to them on a daily basis in each of the countries that they are released in. The rankings are used to calculate the popularity score of a song in a particular country. We plan to build regression models that can predict the popularity of a song in a given country. We then intend to use these models to predict the popularity scores of a song when it is introduced into a new country. This is also useful in predicting the global appeal of a song in countries other than where it has been introduced.

**What models you are using. List them clearly.**

1. Linear Regression -

* Lasso
* Ridge

2. Decision Tree regression

3. Random Forest regression

4. Support Vector regression

5. K Nearest neighbors

**Results**

Use tables, and appropriate metrics.

| **Model** | Mean Absolute Error (MAE) | Root mean squared error(RMSE) | R2(Coefficient of determination/Goodness of fit) |
| --- | --- | --- | --- |
| Linear Regression - Lasso |  |  |  |
| Linear Regression - Ridge |  |  |  |
| Decision Tree regression |  |  |  |
| Random forest regression |  |  |  |
| Support vector regression |  |  |  |
| K Nearest Neighbors |  |  |  |

**Conclusion**

The major takeaways would be -

* Understand capability of song audio characteristics in predicting song popularity
* Compare various regression models listed in the “models” section to understand how each of the models are performing
* Compare model scores across countries to determine which countries are relatively easier to predict and which are not
* Obtain feature importances to understand which features of a song are more important in determining the song popularity

In conclusion, the impact of this study will be to allow users of the end model to predict how well a song would perform were it to be introduced in a new country, as well as help identify the countries where there is a good potential for the song to gain popularity. This is invaluable knowledge to those in the music industry looking to find the best markets for their music.