# Final Data Science Project

## Project Overview

SonicWave Productions is a growing company seeking to gain headway in the music industry. The CEO of SonicWave Productions is looking to understand what drives the current popularity of individual songs. In other words, they are trying to understand what factors are the most important in determining the popularity of a song. Similar to SonicWave Production's competitors, e.g. Spotify, they are also looking to implement a model capable of predicting the popularity of a song. Such a model would empower their team of music professionals to swiftly identify songs that are either undervalued or overvalued in the market, facilitating strategic decisions in song selection, promotion, and distribution.

Congratulations! You have recently been hired as a Data Science Consultant for SonicWave Productions. Your task, as outlined above, is to **predict the popularity of songs** from three different genres: rock, jazz, and pop. In order to tackle this problem, you have been given a set of music data containing metrics related to popularity as well as various song attributes.

Two datasets provided include: *train.csv* and *test.csv*. The training dataset consists of 1200 observations that include the popularity of the song and several other metrics (e.g. duration, loudness, time signature, etc.). The test dataset is the same except only contains 500 observations and does not include the popularity of each song. You will use the training data to build, implement and test models. You will then generate predictions based on the "best" model. The data dictionary on the last page of this document gives a description of each variable.

This project should be treated as a take-home exam and is to be completed independently. You may not consult with anyone about any aspect of the project other than the professor and TAs.

The Final Data Science project is due by **11:59 PM EST** on **Tuesday April 16.**

## Project Deliverables

1) **Predictions:** A single csv file with 500 test observations named "testing_predictions_LAST_FIRST_ID.csv" where LAST is the student's last name, FIRST is the student's first name, and ID is the student's e-mail ID (see example below). The file should contain two labeled columns in the following order:
    - id: song id provided in the test dataset.
    - popularity: predicted popularity of the song.

2) **Technical Report:** A pdf report that outlines your process from start to finish in technical detail. Please name it "technical_report_LAST_FIRST_ID.pdf" (see example below). The report should NOT include any code. You may use at most 4 figures or tables. Limit of 5 pages (double spaced). This should (at least) touch on the following:
    - Introduction and description of exploratory data analysis.
    - Identification of Data oddities e.g. missing data, extreme values, etc. and how you handled them.
    - Summary of all models considered.
    - How many models seemed to perform "best" in terms of predictive accuracy? How did you measure this?
    - What were the most important variables? How did you measure variable importance?
    - What were the most challenging aspects of this particular dataset? Were you able to mitigate these issues? Do you really trust your "best" model? If your job depended on this model, how worried would you be? Is there other information you may want in order to improve the final model/predictions/recommendations further?

3) **Final (non-technical) Report:** Discuss your findings with a non-technical decision maker, in this case the CEO and music professionals of SonicWave productions. Introduce your project and summarize some key findings that could be useful to understand the popularity of songs. Please name it "final_report_LAST_FIRST_ID.pdf" (see example below). Limit 1.5 pages (double-spaced). (Note: Non-technical decision maker means that they will not know phrases and concepts such as but not limited to: lasso, ridge regression, random forests, gradient boosted tree, tuning parameter, cross validation, mean squared error, bias, variance, overfitting, etc.)

4) **Code:** A single .R or .Rmd named "code_LAST_FIRST_ID.R" (see example below). Your code should be thoroughly commented and able to be run from another machine provided necessary packages and data are loaded.

*File names example: If the TA, Ryan Cecil, with Pitt e-mail RMC144@pitt.edu were to submit his files for submission for this project, they would be labeled as

1. "testing_predictions_Cecil_Ryan_RMC144.csv"
2. "technical_report_Cecil_Ryan_RMC144.pdf"
3. "final_report_Cecil_Ryan_RMC144.pdf"
4. "code_Cecil_Ryan_RMC144.R" or "code_Cecil_Ryan_RMC144.Rmd"

## Rubric

1) **Accuracy (10%):** Model predictions on the test dataset will be graded based on Mean Squared Error. This is largely an "all or nothing" category -- to earn full points, you simply need to have a model with lower test MSE than a pre-established base rate. You do not need the best possible model available, and you should not spend all your time trying *ad hoc* things in search of the lowest possible MSE. This project isn't a "predictive competition" like you might find on kaggle.com. The goal of this project is to find strong model(s) obtained by correct reasoning and to understand what those variables imply as well as the uncertainty surrounding them.
2) **Technical Report (50%):** The technical report will make up a significant chunk of your grade and should contain the guts of your process. The four main components of the technical report you will be graded on are:
   - *Introduction / EDA* – This should give an overview of the problem, general information of the data, identify data oddities, summary statistics, etc.
   - *Methods Overview/Details* - This should contain a summary of the methods explored and the various approaches that were considered.
   - *Summary of Results* - This should provide an overview of all of the results obtained. Comment on overall trends, contradictions between models, etc. You can include a table here if it helps summarize the findings. Include test error estimates from the best overall model(s) as well as from the model(s) you ultimately chose to rely on.
   - *Conclusions / Takeaways* - Based on the results described in the previous section ('Summary of Results'), describe what you feel can safely be concluded. If there are further tests/models that you think would be relevant to pursue given the overall results, note this.
3) **Final (non-technical) Report (30%):** How would you explain the results to someone interested in your findings that doesn't have a statistics background? Discuss your project and findings in a non-technical manner. Identify and summarize at least 2 or 3 specific key takeaways from your work. These can include any useful and potentially actionable findings and/or specific aspects of the work that decision makers should keep in mind.
4) **Quality of Code (10%):** Does code run from another machine provided necessary packages and data loaded? Is code "readable" and well commented.

## Data Dictionary

| Variable | Type | Description |
|---|---|---|
| id | Numeric | Unique song identifier |
| album_name | Character | Name of the album in which the song appears |
| track_name | Character | Name of the song |
| popularity | Numeric | The popularity of a song is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by an algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are |
| duration_ms | Numeric | The duration of the song in milliseconds |
| explicit | Character | Whether or not the song has explicit lyrics (TRUE is yes; FALSE is no) |
| danceability | Numeric | Danceability describes how suitable a song 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 | Numeric | Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic songs feel fast, loud, and noisy |
| key | Numeric | The key the song is in i.e. the major or minor scale that the song is based on. Each value maps to a specific pitch using standard Pitch Class notation e.g. 0 = C, 1 = C#/Db, etc. |
| loudness | Numeric | The overall loudness of the song measured in decibels (dB) |
| mode | Numeric | The modality (major or minor) of a song, the type of scale from which its melodic content is derived (1 is major, 0 is minor) |
| speechiness | Numeric | Speechiness detects the presence of spoken words in a song. The more exclusively speech-like the recording, the closer to 1.0 the attribute value. Values above 0.66 describe songs that are probably made entirely of spoken words |
| acousticness | Numeric | A confidence measure from 0.0 to 1.0 of whether the song is acoustic. 1.0 represents high confidence the song is acoustic |

| | | |
|---|---|---|
| instrumentalness | Numeric | Predicts whether a song contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word songs are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the song contains no vocal content |
| liveness | Numeric | Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the song was performed live. A value above 0.8 provides strong likelihood that the song is live |
| valence | Numeric | A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a song. Songs with high valence sound more positive (e.g. happy, cheerful, euphoric), while songs with low valence sound more negative (e.g. sad, depressed, angry) |
| tempo | Numeric | The overall estimated tempo of a song 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 |
| time_signature | Numeric | An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of 3/4 to 7/4 |
| track_genre | Character | The genre of the song. Either “rock”, “pop”, or “jazz” |