OSGD Associate Project Brief 2021

Introduction

The following project serves as an introduction into the role of a data science consultant and your submission will be our primary criterion for determining which associates will be promoted to senior consultants next term.

The project consists of two sections, a more guided section focused on exploratory data analysis and an open ended section where you can show off your machine learning skills.

If you have any questions please get in contact with us directly or post on the Facebook page. The use of reference material (eg. docs, stackexchange) is allowed, and encouraged, though the work you submit must be completely of your own merit.

Enjoy,

Oscar & Jacob
OSGD Training Committee

Submission Details (Deadline: Sunday Week 8 - 7th March)

How to submit: A google forms link will be posted to the facebook page for you to submit to.

Section A

Submit a “Section A - name.ipynb” file as well as typing your answers to the questions in the google form.

Section B

Submit a “Section B - name.ipynb” file - use markdown cells to add commentary

Report

Along with your code, please submit a (three page max) pdf report summarising your findings to the client, feel free to include any graphs or tables as you see fit.

Presentation

A few associates will be selected to present their work at OSGD’s end of term presentation - a great chance to see what the senior consultants have been up to. If selected, please prepare a short (5/10 min) presentation to share which walks us through your findings.

Data

You will have been provided with a data.csv file containing information about 175,000 songs from Spotify. A data dictionary has been appended at the end of this document to give you a description of the features in the data.

We have also included a “missing_data_functions.py” file which will be used for section B.

Client Briefing
Your client is G&H Sounds, a medium sized record label who are currently managing 30 artists from a wide range of genres. The bulk of their revenue comes from live concert ticket sales, where they also sell cds and merchandise. G&H Sounds does not currently have a presence on any streaming service.

They are unfamiliar with the user base for streaming services and have so sought your expertise to help them decide what music they should release on Spotify.

Detailed below are some metrics about their best performing artists under management. They would like you to help them choose which artist is likely to gain the most popularity.They would also like you to advise them on what type of artist performs best on the platform and what qualities they should keep an eye out for when scouting for new talent.

Avg Characteristics	Elliot Tempest	Amy Apollo	Pocket Rockets
acousticness	0.15	0.3	0.24
danceability	0.4	0.55	0.47
duration_ms	259000	230000	244000
energy	0.8	0.57	0.7
explicit	1	0	0
instrumentalness	0.03	0.05	0.2
key	6	5	4
liveness	0.36	0.13	0.27
loudness	-5	-7	-10
mode	1	1	1
speechiness	0.06	0.05	0.1
tempo	127	123	128
valence	0.5	0.4	0.6

Note: Potential song names and release dates haven't been decided upon yet, this is something you could advise on if you think it is important to the artist's success.

Questions

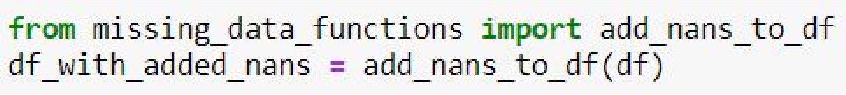
Section A

0. Import the data into a dataframe and ensure each column is of the correct data type.
1. How many unique artists are in the dataset?
2. What is the most common word in a song title (ignoring stop words)?
3. What characteristics does the average 'Frank Ocean’ song have?
4. Which artist saw the biggest (absolute) jump in popularity between consecutive song releases in the dataset?
5. Genre labels have been removed from this dataset - using kmeans clustering, what do you think would be a good number of ‘genres’ for us to group the dataset by? Then add these labels to the dataset.
6. Produce some graphs and your own insights into what trends music on the platform has seen in the last 5 years - include these in your report.

Section B

Dealing with Missing Data

In the real world, dealing with missing data is an essential skill. To give you some practice we have written a function which will artificially add some nans to the numerical columns in the dataframe. The code below shows you how to import and use our function.



```
from missing_data_functions import add_nans_to_df
df_with_added_nans = add_nans_to_df(df)
```

You should write code to ‘deal’ with the missing values that are created.

[What we mean by ‘deal’ with is up to you]

Model Building

Build a model to predict a song’s popularity, then use your model to predict how popular the record label’s artists will be.

[Really show off your machine learning skills!]

Ambiguous Question from Client

The record label is considering ways to boost their artist’s popularity. They have decided that they would like to explore the idea of getting one of their artists to feature on a more popular artists upcoming album.

G&H sounds would like you to recommend an artist who would be a good fit for the artist you recommended in the model building section to work with. Note that G&H sounds will have to pay a licencing fee to arrange the collaboration with the other artist’s record label, here is a formula that you can use to estimate this cost;

$$\text{Cost of a feature with artist x} = \frac{\textit{avg popularity of artist x over last 2 years}}{(\textit{avg number of artists working on each song})1.3}$$

[Think what might be a way to measure a ‘good fit’]

Report

In your report (3pg max) please explore the following ideas;

- Give a summary of the broad trends the music industry has seen over the last 5 years.
- What are your answers to the record label’s questions?
- Feature importance - what qualities should G&H keep an eye out for when scouting for new talent.

Along with any other thoughts or recommendations that you might have.

Data Dictionary

Feature	Description	Type
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.	Float
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.	Float
duration_ms	The duration of the track in milliseconds.	Int
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.	Float
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.	Float
key	The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭ , 2 = D, and so on.	Int
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.	Float
loudness	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.	Float
mode	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.	Int
popularity	The popularity of the track. The value will be between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are	Float
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.	Float
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.	Float
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).	Float