**INFO 648 – Business Data Analytics**

**Team Project**

You have been hired by a music streaming company, *Spotify*, to help improve its efforts at recommending new songs to users using its streaming service. The company wants to predict which songs are likely to be popular in the future when they are released, and based on this, the company wants to make their recommendation system more effective. The file “songs\_utf.csv” contains a set of data collected on top songs in the U.S. Spotify chart during 1998-2020. Each record provides information on a song’s characteristics. For clarity, the variables in the dataset are defined below.

1. artist: Name of the Artist.
2. song: Name of the Track.
3. Song\_name\_len: Length of song name
4. duration\_ms: Duration of the track in milliseconds.
5. explicit: The lyrics or content of a song or a music video contain one or more of the criteria which could be considered offensive or unsuitable for children.
6. year: Release Year of the track.
7. popularity: The higher the value the more popular the song is.
8. 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.
9. energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.
10. 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. If no key was detected, the value is -1.
11. 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 typically range between -60 and 0 db.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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).
18. 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.
19. genre: Genre of the track.
20. pop: 1 if genre contains pop and 0 otherwise
21. rock: 1 if genre contains rock and 0 otherwise
22. hiphop: 1 if genre contains hip hop and 0 otherwise
23. dance: 1 if genre contains dance and 0 otherwise
24. folk: 1 if genre contains folk and 0 otherwise
25. rnb: 1 if genre contains R&B and 0 otherwise
26. latin: 1 if genre contains latin and 0 otherwise
27. hot: 1 popularity is greater than 75 and 0 otherwise. (Note: Be cautious when using this feature, as it is likely to exhibit high collinearity with the dependent variables.)

**Project**

**Q1 [50%].** The company has commissioned you to develop a predictive model to enhance its recommender systems by focusing on popular songs. The primary goal of this project is to predict whether a song’s popularity score exceeds 64 (>=65), enabling Spotify to prioritize songs that are more likely to engage users effectively.

To achieve this goal, you are required to develop and compare three different models to predict song popularity. The purpose of building multiple models is to identify and select the most suitable one based on its evaluation metrics. Once the models are developed and evaluated, you must justify your choice of the best model by explaining why it is the most suitable for predicting song popularity.

The project consists of the two phases to be completed:

1. Sample selection and data preprocessing [10%]:
   1. Download “songs\_utf.csv.”
   2. Apply an appropriate strategy for handling missing values. (e.g., removing all records with missing values)
   3. Select a set of attributes relevant for predicting popularity. Encode categorical attributes if necessary.
   4. Create the outcome variable: Set it as 1 if the popularity score is greater than 64 (>=65), and 0 otherwise.
2. Predictions [30%]:
   1. Develop and compare three different approaches with the goal of predicting whether a song’s popularity score will exceed 64 (>=65). If necessary, provide an appropriate fine-tuning strategy using 10-fold cross-validation. Report appropriate evaluation metrics
   2. Choose the best model based on accuracy. (You can choose the models using either holdout accuracy or cross-validation accuracy)
3. Notes on Documentation:
   1. You are free to explore and select features that provide the best prediction results and fine-tune the models as necessary. Ensure you justify your feature selection and fine-tuning processes clearly.
   2. Based on your results, provide appropriate managerial insights that can help guide decision-making. These insights should include, but are not limited to, suggestions and recommendations for future songs Spotify should consider including. Focus on attributes that are associated with higher engagement and popularity scores. By leveraging these insights, Spotify can optimize playlists and promote songs that are more likely to engage users and increase satisfaction.

**Q2 [30%].** The VP of sales has indicated that they would like to maximize the monetary value of the predictions (i.e., maximize revenue from correct predictions less the cost of prediction errors). The revenue and costs are tied to the confusion matrix. The revenue for correctly predicting a popular song is $1,000. The cost for incorrectly predicting a song is popular, when in fact it is not, is $700. The cost for incorrectly predicting a song is not popular, when in fact it is, is $900. There is no cost or revenue for correctly predicting a song is non-popular.

1. Redo the steps in Q1 and choose the best model based on the profit
2. Is the best model chosen based on profit (in Q2) the same as the one selected based on accuracy (in Q1)? Compare the best model selected in Q1 with the best model selected in Q2. Describe the new managerial insights you gained. Highlight the trade-offs that influenced the choice and explain how these insights might impact Spotify's recommendation strategy moving forward.

**Q3 [30%].** The company is also interested in understanding 1) the effects of valence of a song in predicting the song’s success and whether the effect differs across different types of music. 2) Which features or combinations of features are frequently associated with popularity of songs. This analysis would help improve their playlist recommendations by identifying similar songs based on these patterns.

1. Cluster all the songs in the datasets by danceability and energy (You may also include genre (pop, rock, hiphop, dance, folk, rnb, latin) in the clustering process, but this is optional)
2. For each cluster identified, examine the cluster characteristics and develop an appropriate model that examines the impact of valence on the song’s success.
3. Interpret the results and describe the managerial insights
4. Identify features or combinations of features that frequently appear together in songs with high or low popularity. Use these insights to suggest improvements to the company’s recommendation playlist.

**Final Report**

Each team will submit one report by May 6. The report will contain the following:

1. A Cover page that includes the report title, group name, the names of the group member and contributions to the project
2. A section that summarizes the data pre-processing (1 page).
3. A section that summarizes your results and insights for Q1
4. A section that summarizes your results and insights for Q2
5. A section that summarizes your results and insights for Q3
6. A link to your Google Colab code and list of references if you cite some articles in your reports
7. Appendix (5 pages allowed only)

**Note:** When drafting the final report, please **be professional and concise**. Do not copy and paste your Python code directly into the report. **Provide only appropriate and necessary plots in the main text**. If you wish to provide a complex plot (e.g., your large decision tree), please include it only in the appendix. Combine tables where necessary to save space and maintain clarity.

**Format:** Please use Times New Roman, 12-point font, with 1-inch margins on all sides, and 1.5 line spacing throughout the document.

**Presentation: (25 mins in total)**

Your presentation should follow the format below:

* A general introduction to the project questions and objectives

For each of Q1, Q2, and Q3:

* A brief summary of feature selection & data preprocessing
* A concise summary of the methods used to solve the problem (No Code)
* A clear explanation of the major findings
* Recommendations and managerial insights

Finally (Please manage your time wisely it is recommended to leave 3-5 min for Q&A)

• A Q&A session to address questions from the audience