Introduction to Data Science Project Report

Project **D2**

Title : Top Spotify Songs in 73 Countries

Team members : Vladimir Makarenkov

Business understanding

Identifying your business goals

Background:

The music industry, driven by streaming platforms like Spotify, offers a wealth of data for analysis. This project aims to extract valuable insights from Spotify's dataset, focusing on identifying genre preferences across 73 countries, analyzing trends in song rankings over time, and predicting future song popularity using machine learning.

Business Goals:

- Develop a model to identify the most popular genres in each of the 73 countries. Artists, record labels, and streaming platforms seeking to adapt their content to regional preferences and interests.
- Implement algorithms to analyze temporal changes in song rankings, identifying emerging patterns. In music industry analysts and decision-makers are constantly looking to understand evolving consumer preferences, so it would help to understand them better.
- Build machine learning models predicting future song popularity based on historical data. Artists and labels interested in strategic planning and maximizing the impact of their releases.

Business Success Criteria:

- Accuracy of genre classification.
- Identification of significant trends.
- Precision and recall of the predictive model.

Assessing Your Situation

Inventory of Resources:

- Spotify dataset augmented with relevant data.
- Websites to check manually songs genres as its mostly pretty hard to describe song with one genre.
- People who have knowledge to work with data.
- Needed hardware to compute results

Requirements, Assumptions, and Constraints:

- Assuming accurate Spotify dataset.
- Potential lack of human resources, time.

Risks and Contingencies:

- Data Quality Issues: Regular audits and cleaning procedures.
- Model Overfitting: Cross-validation techniques and robust evaluation metrics.
- Regular changes in rankings, popularity and new releases.

Terminology:

- *Popularity Score:* Metric quantifying a song's popularity.
- *Explicit:* Boolean to check if song is explicit. (An explicit track is one that has curse words or language or art that is generally deemed sexual, violent or offensive in nature - basically how artist indents in to be heard)
- *Weekly movement:* The change in rankings compared to the previous week.

Costs and Benefits:

- *Costs:* Data acquisition and human time.
- *Benefits:* Informed decision-making for artists, record labels, and streaming platforms, potentially leading to increased revenue.

Defining Your Data-Mining Goals

Genre Identification:

- *Goal:* Develop a model to identify the most popular genres in each country.
- *Success Criteria:* High accuracy in genre classification. Will require manual internet research as its not provided in dataset and also not that easy always to identify songs genre.

Trend Analysis:

- *Goal:* Implement algorithms to analyze temporal changes in song rankings.
- *Success Criteria:* Successfully identify and interpret significant trends.

Predictive Modeling:

- *Goal:* Build machine learning models predicting song popularity.
- *Success Criteria:* Develop models with high predictive accuracy for future song popularity.

Data-Mining Success Criteria:

Genre Identification:

- High accuracy in classifying genres. Will lead to informed decision-making for content creators and distributors.
- Successful identification of significant trends. Will lead to proactive adaptation to changing market dynamics.
- High precision and recall of the predictive model.

In summary, this project aims to leverage data from Spotify to provide actionable insights for the music industry, enhancing decision-making processes for artists, labels, and streaming platforms. Through accurate genre identification, trend analysis, and predictive modeling, the project aligns with the specific needs of stakeholders, ensuring a direct and meaningful impact on their strategies and outcomes.

Data Understanding

Gathering Data

Outline Data Requirements:

Spotify Dataset: Obtain a comprehensive dataset from Spotify, including information on songs, artists, ranking, popularity, country, rankings movement, explicity.
Additional data: Augment the data with relevant genres information.

Verify Data Availability:

Spotify Dataset: Confirm the availability of a dataset covering a diverse range of songs, artists, ranking, popularity, country, rankings movement, explicity.
Additional data: Finding needed additional data in external datasets or finding overall needed information in Internet.

Define Selection Criteria:

Time Range: Select a suitable time range for the analysis, ensuring a balance between historical data and recent trends.
Genres: Select right genres for right songs based on research, in most cases genres are defined by people and written in public websites.

Describing Data:

Provided dataset contains needed information about releases, songs rankings, popularity, country etc. There are all needed cases provided, but needed to find additional information for those cases at least in terms of genre.

Exploring Data:

Spotify Dataset Overview:

- Explore the structure of the dataset, including fields such as song name, artist, daily rank, daily movement, weekly movement, country, date of case, explicity and popularity score.
- Identify any missing or incomplete data and decide on strategies for handling them during data preparation.

Verifying Data Quality:

Spotify Dataset:

Completeness:

- Identify and document any missing values in critical fields.
- Decide on strategies for handling missing data during data preparation.

Consistency:

- Check for inconsistencies in categorical fields such as artist names.
- Resolve any discrepancies to ensure accurate genre identification.

Accuracy:

- Validate the accuracy of numerical fields like popularity scores.
- Identify and rectify any outliers that may skew the analysis.

Currency:

- Ensure that the dataset is up-to-date by checking the latest release dates.
- Exclude outdated or irrelevant data to maintain the relevance of the analysis.
- Consider imputation or alternative strategies based on the extent of missing data.

Conclusion:

The data understanding phase of the CRISP-DM process has provided a solid foundation for the project. The Spotify dataset offers a wealth of information on songs and user preferences, while the additional demographic and cultural data enriches the analysis. By exploring, describing, and verifying the data quality, we have identified potential challenges and formulated strategies for data preparation. The next phase will involve cleaning the data,

handling missing values, and preparing it for the subsequent stages of analysis and modeling. The understanding gained during this phase ensures that the data used for further analysis is reliable, relevant, and aligned with the project's objectives.

Project Plan: Top Spotify Songs in 73 Countries

I. Data Preparation:

Cleaning Spotify Dataset:

- *Tasks:*
  - Handle missing values in fields like genre and artist.
  - Standardize inconsistent genre and artist names.
  - Remove outliers in popularity scores.
- *Hours:*
  - Overall around 10 hours.
- *Methods/Tools:*
  - Pandas for data cleaning.
  - Statistical analysis for outlier detection.

II. Exploratory Data Analysis (EDA):

Genre and Popularity Trends Analysis:

- *Tasks:*
  - Conduct EDA on song preferences across countries. After that find genres.
  - Analyze trends in popularity scores over time.
  - Visualize key insights for stakeholders.
- *Hours:*
  - Around 15 hours.
- *Methods/Tools:*
  - Matplotlib and Seaborn for visualization.
  - Time-series analysis for trend identification.

III. Machine Learning Model Development:

Feature Engineering:

- *Tasks:*
    - Create relevant features from existing data.
    - Engineer time-related features for predictive modeling.
- *Hours:*
    - Around 15 hours.
- *Methods/Tools:*
    - Scikit-learn for feature engineering.
    - Time-series modeling techniques.

Predictive Modeling:

- *Tasks:*
    - Build machine learning models for song popularity prediction.
    - Fine-tune hyperparameters for model optimization.
    - Evaluate model performance and adjust as needed.
- *Hours:*
    - Around 20 hours.
- *Methods/Tools:*
    - Scikit-learn for model development.
    - Cross-validation for performance evaluation.

Project Plan Summary:

This plan outlines four key tasks, assigning hours to each team member for optimal project management. The methods and tools specified align with the tasks.