Project report

PROJECT D7: EUROVISION-RESULTS
Analysing the results of the Eurovision Song Contest
Sander Soodla, Teele Tars, Maria Küüsvek
Project repository: https://github.com/MariaKuusvek/Datascience-STM-Project

Business understanding

Identifying our business goals

Background:
We as fans of the Eurovision Song Contest (ESC) seek to better understand the factors which influence success in the contest and to predict the results of the contest more accurately. In our personal experience, we have identified a number of factors which we believe to indicate an entry's potential success or failure in the competition; these include the country being represented and the language of the song, among others. While we have good reason to believe that these factors have an influence on the outcome of the contest, their effects have not been quantified. We are also interested in the lyrical contents of ESC songs and we want to determine if there are any interesting patterns involving the lyrics of these songs.

Business goals:
Our primary goal is to be able to predict the results, including the winner of ESC, more accurately than what has previously been possible. Our secondary goal is to analyse the lyrics of ESC songs.

Business success criteria:
We consider the project to be successful if we manage to develop a method for predicting the results of ESC that is more accurate than just random guessing, and if we have produced any interesting findings related to the lyrics of the songs (or concluded that there are none).

Assessing our situation

Inventory of resources:
The resources available for this project include

- Ourselves i.e three team members who act as data miners as well as experts on the business problem, instructors of the IDS2021 course as consultants on the methods used in our project
- Our dataset (https://github.com/Spijkervet/eurovision-dataset/releases)
- GitHub and Jupyter Notebook for manipulating and mining the data

- Our personal computers

Requirements, assumptions, and constraints:
The requirements for this project include

- The project must be finished by December 16th, 2021
- The dataset must be correct and have sufficient amounts of consistent data
- The resulting prediction model must be able to predict the success of an entry (and in doing so, the results of the contest) based on the attributes of the entry, such as the country being represented, the language of the song etc.

Risks and contingencies:
There are no foreseeable risks that could delay the completion of the project aside from us, the team members, not doing the work that needs to be done on time. The contingency for this risk is to make a detailed schedule for completing the project on time and working in order to follow that schedule.

Terminology:

- entry/contestant - an act consisting of a song and its performers, which represents a country in ESC in a particular year

Costs and benefits:
We find that it is not relevant in our project to determine costs and benefits in terms of monetary units, since we don't intend to profit from the results of this project financially.

Defining our data-mining goals

Data-mining goals:
The data-mining deliverables include several different models for predicting the success of an entry along with the accuracy reports for each model and a comparison of these models based on their success. For our secondary goal, our goal is to try different algorithms for mining frequent patterns and measure the interestingness of the patterns we find, if we find any.

Data-mining success criteria:
The resulting model for predicting the success of entries (and the results of the contest based on that) must show significant predictive improvement compared to random guessing. We must also reach a conclusion whether there are any interesting patterns associated with the lyrics of ESC songs, and what those patterns are.

Data understanding

Gathering data

Data requirements:
In order to address the data mining goals of this project we need data about a lot of Eurovision entries. The data needed for each contestant and song should include at least:

- the country represented by the contestant
- title, lyrics and language of the song
- place and number of points in the final
- performing order in the final and semi-final

Additionally:

- audio features (like rhythm and tonal properties) of the songs.

Data availability:
There are a few easy to find Eurovision datasets on the internet. However some of these only include the voting data and no information about the songs. Luckily we found one dataset that has most of what we need. One thing that's missing is the language of the song. This could be solved by detecting the language with some existing tool, using song lyrics that are included in this dataset.

Gathered data:
Dataset: https://github.com/Spijkervet/eurovision-dataset (data from the eurovisionworld.com fansite)
The author of the dataset has also made it easy to extract audio features using Essentia music extractor and music from youtube links.

Describing data

The dataset we plan to use in this project is a freely available Eurovision Song Contest dataset from the above-mentioned github page. For our purposes, we only need the dataset about the contestants (contestants.csv) and not the specific voting data. The contestants dataset has information about all songs that have participated in Eurovision contests from the very first ESC (1956) to 2019. In terms of size, this dataset should cover our needs, as it includes 1562 songs. This dataset also meets the set requirements for features, apart from one.
Features in the dataset:

- year - contest year
- to_country_id - country id of contestant
- to_country - country name of contestant
- performer

- song - title of the contestant’s song
- sf_num - which semi-final the contestant participated in
- running_final - order in the broadcast of the contest's final
- running_sf - order in the broadcast of the contest's semi-final
- place_final - place in the final
- points_final - points in the final
- place_sf - place in the semi-final
- points_sf - points in the semi-final
- points_tele_final - televoting points in the contest's final
- points_jury_final - jury points in the contest's final
- points_tele_sf - televoting points in the contest's semi-final
- points_jury_sf - jury points in the contest's semi-final
- composers
- lyricists - writers of the lyrics
- lyrics - lyrics of the song
- youtube_url - url to video on YouTube

The lack of song language information can be solved, as song lyrics are present in the dataset.

Exploring data

At first glance there are a lot of non-existent values in the dataset, but this is not a case of missing data. This is because of how Eurovision’s rules have changed over the years. For example there were no semi-finals before 2004. Also the voting has changed quite a bit, including the amounts of points given and the fact that until quite recently there was no televoting in Eurovision.
The data does indeed include song and ranking info for all 1562 entries from years 1956-2019. It also has song information for the year 2020, when the competition was cancelled (we could try to predict the results of it). Importantly, the dataset has no obvious errors. For example no contestant has placed in a position below first or above the highest number of participants.

Data quality

The data we have should be good enough to support our goals. As mentioned previously, some data is only missing due to different rules of the competition and the most important data is there. Interestingly, we found 3 songs that didn’t have a title, but this ended up being the case only when the song title was the same as the performer’s name, so it wasn’t a quality issue after all. In conclusion, there doesn’t seem to be any data quality issues that would prevent us from reaching our goals. However, the data quality of audio features extracted from the songs on youtube is yet to be determined.

Project plan

List of Tasks

Tasks	Hours	Team Member
Data preparation	1	Maria
Extract audio features of songs and add them to the dataset.	2	Sander
Split Dataset into 70% training, 15% test, and 15% validation data	1	Teele
Train machine learning models with different algorithms - Linear Regression - Lasso Regression - Ridge Regression - KNN - Random Forest	2	Maria
Test accuracy of the models	2	Teele
Using the most accurate model, predict the results of an example year at Eurovision	1	Sander
Using NLP methods, mine song lyrics for interesting findings, including - Word frequency - Correlations between words	2	Teele
Create Project Poster	3	Combined effort between all 3 team members

List of Methods

We plan to use machine learning methods to train our model to correctly predict the results of test data. Every entry must have a weight and this will determine their position on the leaderboard, which the model will then attempt to predict.

Models to use:

- Linear Regression
- Lasso Regression
- Ridge Regression
- KNN
- Random Forest

For the secondary goal of lyric analysis, we aim to use NLP methods in order to find words that appear most frequently, as well as pairwise correlations between words and clusters of words. For that, we plan to use

- Tokenization
- Word counts
- Phi coefficient