ST344 Group Coursework

October 2019

Deliverables (from each group/team):

1. **Plan of work.** By Tuesday of Week 5 (29 October), 12 noon. Counts for 5% of ST344 module marks.
2. **Group report.** By Thursday of Week 9 (28 November), 12 noon. Counts for 35% of ST344 module marks.
3. **Group presentation.** In Week 10 (starting Monday 2 December). Presentation slides to be submitted by Friday 29 November, 12 noon. Counts for 10% of ST344 module marks.

In addition, also by Friday 29 November at 12 noon, each group should submit its jointly agreed statement — just a short paragraph — about which team members contributed to which parts of the work, and the suggested allocation of marks among the team members (see details below).

Mark Scheme

Plan of work: Should give a realistic timeline for the work the deliverables 2 and 3, with appropriate work-milestones and meeting dates identified **(3 marks)** and provisional allocation of tasks to group members **(2 marks)**. Maximum 1 side of A4, submitted electronically in PDF.

Group report: A formal report, to professional standards. The report should have two parts, both submitted electronically:

1. PDF document (maximum 8 typeset A4 sides including graphs, tables and references; minimum body-text font size 11pt, and minimum 2cm margins all round) for the report itself. This document should be written for intelligent readers who do not necessarily have advanced statistical training.
2. A 'technical appendix' in the form of a .zip archive, which contains a .Rmd file giving full details of the analysis that is reported, along with any other files that are needed in order to allow the analysis to be reproduced.

Marks will be allocated as follows:

- Report structure and presentation: **10 marks**
- Appropriate and well explained exploratory data analysis and conclusions: **20 marks**
- Clear, well documented and functioning code in the .Rmd document: **5 marks**

Group presentation: A professional oral presentation of the analysis and main findings, for an audience of fellow statisticians. Maximum length 12 minutes. Marks allocated as follows:

- Structure and pace: **5 marks**
- Clarity of presentation: **5 marks**

All group members should contribute roughly equally to delivery of the presentation.

Each team should decide how to distribute the group mark by allocating to each team member a share of $n \times 100\%$ where n is the number of students in the team. This will act as a weighting factor to convert the group mark into an individual mark. For example, suppose the group mark is 70% and a team of 4 students decides to allocate 100% to each team member, then each member receives the mark of 70%. On the other hand if the team decides to allocate 106% to one team member and 98% to the other three members, then the former receives a mark of 74.2% and the other three team members receive the mark 68.6%. The maximum weighting factor that can be awarded is 110%, the minimum weighting factor is 90%. The module leaders reserve the right to moderate the weighting factors, impose equal weighting factors and/or request further evidence.

An outline of the task

After graduating from Warwick, you have started a job as a Data Scientist at a record company in the music industry. Roughly speaking, a record company invests money into recording, distributing and promoting artists that are "signed" to their company. They usually make more profit from artists whose music is more popular. In the past decade or so the business model of a record company has changed dramatically due to the influence of iTunes and other companies that distribute music as mp3s, and more recently, Spotify and other streaming services.

The leader of your team wishes to use data to help the company with two challenges:

1. To predict what type of music will be popular, in order to help the company invest in artists that will help them make a profit.

2. To produce a system for recommending new music based on the tastes of a listener.

The data

Your team leader has put together a data set by merging data collected from Spotify and the Official Charts Company (which keeps records of the "charts", lists of the best selling albums, in the UK). The filename of this dataset is '`edited_spotify.xlsx`'. He selected 35 albums from each decade from the 1960s to the 2010s. From Spotify, he obtained: the album's release date and a measure of its popularity; the artist that recorded the album, a description of the genres of music associated with that artist, and a measure of the artist's popularity; a number of descriptive variables relating to each track (described in an accompanying text document). From the Official Charts Company, he obtained the maximum chart position of each album (if it reached the top 100 albums), the number of weeks the album was in the top 100, and the number of weeks the album was at the top of the charts.

The first meeting with your team leader (who, to place him demographically, is a 40-49 year old White British male with a university degree) about the data and the task reveals the following additional information:

- The data was retrieved from Spotify on 25th September 2019.

- The albums were not chosen in any systematic way. They were chosen in an attempt to represent the genres of music that the team leader perceives to have been popular at the time, but an attempt was also made to choose a sample of albums that had varying degrees of success (there are some albums that topped the charts, but others that did not enter the chart). He admits that some of the albums were his personal favourites, but that he did not personally like all of the albums he had chosen.

- Some artists have a number of albums in the sample. The team leader wasn't very clear in articulating his reason for doing this, but suggests that one motivation for doing this could be to examine whether the tracks produced by a particular artist tend to vary over time in their characteristics.

- Album data for some very famous artists (including, for example, The Rolling Stones) was not available on Spotify. Hence, these artists are not represented in the data.

- The team leader did some ad hoc manual filtering of non-musical tracks he found in the data (e.g. the short "skit" tracks commonly found on hip hop albums, which can consist solely of conversation rather than music).

- The data is saved in Excel (.xlsx) format. The team leader edited the data in Excel, and compiled the data set in a non-reproducible way. He therefore cannot provide you with the scripts he used to construct the data.

- He suspects that to do the prediction (challenge 1), you might need to estimate a model, then use the model for the prediction.

- To simplify the recommendation problem, he says that the first version of the system should base recommendations **only** on the track that the listener is currently listening to, rather than the entire history of their listening choices.

- He welcomes any constructive criticism of the approach the team is taking. It is important to the company to tackle these challenges, and this is intended as an initial short project to help the team understand what type of approaches they might consider, and also the limitations of this way of approaching the problem. Suggestions, based on your experience, of how these challenges might be approached differently, would be very welcome.

Variation between groups

Each group will receive the full data set on which to perform the analysis. However, each group will be randomly assigned two albums from each decade to use as a "test" set for both tasks. The rest of the albums should be used as "training" data. For the first challenge, the test set should be used to test the accuracy of predictions of popularity. For the second challenge, you should produce recommendations for one or more tracks in the test set. Note that at no point do you need to listen to any of the music that you are working on: both challenges should make use solely of the data provided.

Output

You report should document the approaches you used (and the reasoning behind these choices), results found when applying your approaches, and your conclusions about the effectiveness (or otherwise) of the methods you chose. Based on your analysis and experience on this project, you should also include recommendations about what data might be useful in the future, and about how one might best tackle these challenges if there were no constraints on time or budget.

In the course of doing this work you might come across published work by other people, in this same topic area. If you do get ideas from someone else's work, or if you want to mention someone else's findings in your report, that is fine: just be sure to credit the prior publication, and to cite the source accurately. The team might well be interested to know of other such work on the topic. (But the main interest will be in your group's own, new exploration of this data.) Your report will be scanned by the University's Turnitin software system — for general guidance on how to avoid plagiarism, see the one-hour online tutorial at http://warwick.ac.uk/plagiarwise .

Disclaimer

The inclusion of an artist in this data set does not imply an endorsement, by the lecturers, the department or the university, of the views or behaviour of the artist!