MATHS 7107 Data Taming Assignment Final Report

Due date: 5pm, Wednesday 26th April 2023.

Scenario

You are employed as a data scientist at Spotify, an audio streaming and media services provider with over 365 million monthly active users, including 165 million paying subscribers. As Spotify is the world's largest music streaming service provider, the founders are interested in trying to predict which genre a song belongs to in order to better enhance their customer's experience and ultimately remain the best music streaming service available. Their idea is that if they can predict what genre a song belongs to, they can better recommend / advertise songs to customers and more effectively update compilation playlists. They are interested in how the following factors can predict a song's genre:

- the year the song was released;
- how "speechy" the song is;
- how danceable the song is; and
- the tempo of the song.

The founders have asked the web services division to collate an extensive dataset for you to consider which will need to be cleaned first and reduced to 1,000 songs per genre due to computing power.

During your analysis, the Spotify founders would like you to specifically explore the following questions:

- Does the popularity of songs differ between genres?
- Is there a difference in `speechiness` for each genre?
- How does track popularity change over time?

Following this, the main consideration is to build a model to predict genre based on appropriate variables in the provided dataset. This may be all variables in the dataset or a subset of variables if not all are useful. After consultation with an expert statistician, they have decided they would like you to compare the following three models:

- A linear discriminant analysis;
- A K-nearest neighbours model with a range of 1 to 100 and 20 levels; and
- A random forest with 100 trees and 5 levels.

You are required to determine and explain which of the above is the best model. You may want to consider metrics such as AUC, sensitivity, specificity. It is important to provide an explanation to the founders as to why you made the choices you did, to fully justify your decision.

Once you have determined the best model, you will need to test it in an appropriate manner. That is, you need to predict the genre of the song, based on the variables and compare this to the actual genre of the song.

Hint: During preprocessing the data consider using the following:

- `step_date`: This allows you to manipulate date data if you have extracted the year from the date variable.
- `step_rm`: This will remove a predictor from our data. If you have already extracted what you need from the date variable, we no longer need to keep hold of it.

Also whenever you use the command `set.seed()` to get the same result, you should use your index number as the seed. As an example if your index number is a1234567 you should use `set.seed(1234567)`.

These are not the only steps that you should consider.

Formatting

The founders of Spotify have asked if you could provide a report in the following format:

- An executive summary, with key results outlined in plain English;
- A methods section, outlining what data was analysed, steps that were taken, and stating the software used for your analysis;
- A results section, including an exploratory data analysis, and interpreting and evaluating your models in language the founders will understand;
- A discussion section, in which you discuss your models' outcomes and predictions, with specific relevance to the founders' objectives;
- A conclusion, in which you summarise your findings and recommendations to the founders, written in plain English; and
- An appendix, including all R code you used to perform your analysis, output where required to show how the processes you implement are working as well as any technical material beyond what you might expect the founders to understand.

Some more rules about your report:

- **You must complete this assignment using R Markdown**;
- Your report must be submitted as **pdf only** on MyUni together with your Rmarkdown file.;
- You must include **units** when providing results;
- Include any working when providing solutions;
- Provide all numerical answers to **3 decimal places**;
- Make sure you include both your code and R output / plots in your appendix;
- Make sure any tables or plots included have captions;
- Do not write directly on the question sheet;
- You can submit more than once if you find errors and your latest submission will be marked;
- Make sure you only upload one document for your final submission. If you submit multiple pages (i.e. one per question) you will be deducted 10% per page submitted;
- Penalties for late submission - within 24 hours 40% of final mark. After 24 hours, assignment is not marked and you get zero; and
- Finally, make sure you check your submitted assignment is the correct one, as we cannot accept other submissions after the due date.

[Assignment total: 114 marks]