Data Scientist Analytics Test

Summary

There are two different publically available data sets provided for this test, and you are welcome to choose either one to work with. The two data sets are:

1. **Meteorite Landings** – A NASA data set that comprises 45,000 entries for meteorites that have fallen to earth.
2. **20 Years of Games** – A data set taken from IGN of 18,000+ game reviews over 20 years.

Regardless of which data set you decide to work with, your task is to produce a report on the data, including the scripts, code and visualised findings.

You are free to use either R or Python for this task and please return your finished work as a PDF exported from R Markdown, Jupyter or equivalent.

Your Report

The following should be included in your report:

1. Describe and summarise the fields available. This should include how much missing data there is, how the data can be grouped, how unique the data is, any errors in the data and what should be done with those errors, and visualisation of important features.
2. A description of what new features could be created using the data provided as a starting point, with examples.
3. Describe any trends that can be found in the data that could potentially be useful to help predict future events.
4. Splitting your data into test and training sets, build 2 baseline models to see how predictive the data is for the following:
    a. **Meteorites** – Given all other fields, the likely mass range of the object.
    b. **Games** - Given all other fields, the likelihood that the game was made an editor's choice

    Include the following in your analysis:
    1. Your rationale for test and training set splits.
    2. What do the models tell us about features and their importance?
    3. Compare the performance of the two models and describe what could be done, and what expectations should be, for future development of them.

The Data – Meteorite Landings

- **name -** the name of the meteorite (typically a location, often modified with a number, year, composition, etc)
- **id -** a unique identifier for the meteorite
- **nametype** - one of:
  - valid: a typical meteorite
  - relict: a meteorite that has been highly degraded by weather on Earth
- **recclass** - the class of the meteorite; one of a large number of classes based on physical, chemical, and other characteristics (a primer on meteor classifications is at https://en.wikipedia.org/wiki/Meteorite_classification for those interested)
- **mass** - the mass of the meteorite, in grams
- **fall** - whether the meteorite was seen falling, or was discovered after its impact; one of:
  - Fell: the meteorite's fall was observed
  - Found: the meteorite's fall was not observed
- **year** - the year the meteorite fell, or the year it was found (depending on the value of fell)
- **reclat** - the latitude of the meteorite's landing
- **reclong** - the longitude of the meteorite's landing
- **GeoLocation** - a parentheses-enclose, comma-separated tuple that combines reclat and reclong

The Data – 20 Years of Games

- **score_phrase** – A descriptive phrase relating to the review score
- **title** – Game title
- **url** – The original URL on the IGN website
- **platform –** The gaming platform the release was on
- **score –** The review score
- **genre** – Type of game, e.g. Platformer, Fighting etc.
- **editors_choice** – Y/N flag describing if the game was made an editor's choice on the site
- **release_year** – Year the game was released
- **release_month** – Month the game was released
- **release_day** – Day of the month the game was released