Introduction to Data Science – DS UA 112

Capstone project

The purpose of this capstone project is to tie everything we learned in this class together. To make things more straightforward, we will use the same dataset we have used throughout the class, and with which you should be well familiar with, by now. The cover story is that you are working for a major entertainment corporation that is trying to get a better handle on both what makes a good movie as well as a better understanding of the viewers. Historically, this process was mostly guided by intuition, but is increasingly infused by data based decision making. This is where you – a budding data scientist – come in. Can you do better, to justify your rather high salary?

Mission command preamble: As in general, we won't tell you *how* to do something. That is up to you and your creative problem solving skills. However, we will pose the questions that you should answer by interrogating the data. Importantly, we do expect you to do this work yourself, so it reflects your intellectual contribution – not that of third parties. By doing this assignment, you certify that it indeed reflects your individual intellectual work.

Dataset description: This dataset features ratings data of 400 movies from 1097 research participants and is contained in the file "movieReplicationSet.csv". It is organized as follows:
1st row: **Headers** (Movie titles/questions) – note that the indexing in this list is from 1
Row 2-1098: Responses from individual **participants**
Columns 1-400: These columns contain the ratings for the 400 **movies** (0 to 4, and missing)
Columns 401-420: These columns contain self-assessments on **sensation seeking** behaviors (1-5)
Columns 421-464: These columns contain responses to **personality** questions (1-5)
Columns 465-474: These columns contain self-reported **movie experience** ratings (1-5)
Column 475: **Gender identity** (1 = female, 2 = male, 3 = self-described)
Column 476: **Only child** (1 = yes, 0 = no, -1 = no response)
Column 477: **Social viewing preference** – "movies are best enjoyed alone" (1 = y, 0 = n, -1 = nr)

Note that we did most of the data munging for you already (e.g. Python interprets commas in a csv file as separators, so we removed all commas from movie titles), but you still need to handle missing data in some way.

Format: The project is comprised of your answers to 10 (equally-weighed, grade-wise) questions. Each answer should ideally include some paragraph of text (describing what you did and what you found), a figure that illustrates the findings and some numbers (e.g. test statistics, confidence intervals or p-values). Please save it as a word, pdf or pages document. This document should be 4-6 pages long (arbitrary font size and margins). About half a page per question is reasonable. In addition, open your document with a brief statement as to how you handled dimension reduction, data cleaning and data transformation, as this will apply to all answers. Make sure to include your name.

Academic integrity: You are expected to do this project by yourself, individually, so that we are able to determine a grade for you. There are enough degrees of freedom (e.g. how to clean the data, what variables to compare, aesthetic choices in the figures, etc.) that no two reports will be alike. We'll be on the lookout for suspicious similarities, so please refrain from collaborating.

Questions corporate would like you to answer:

1) What is the relationship between sensation seeking and movie experience?
2) Is there evidence of personality types based on the data of these research participants? If so, characterize these types both quantitatively and narratively.
3) Are movies that are more popular rated higher than movies that are less popular?
4) Is enjoyment of 'Shrek (2001)' gendered, i.e. do male and female viewers rate it differently?
5) Do people who are only children enjoy 'The Lion King (1994)' more than people with siblings?
6) Do people who like to watch movies socially enjoy 'The Wolf of Wall Street (2013)' more than those who prefer to watch them alone?
7) There are ratings on movies from several franchises (['Star Wars', 'Harry Potter', 'The Matrix', 'Indiana Jones', 'Jurassic Park', 'Pirates of the Caribbean', 'Toy Story', 'Batman']) in this dataset. How many of these are of inconsistent quality, as experienced by viewers?
8) Build a prediction model of your choice (regression or supervised learning) to predict movie ratings (for all 400 movies) from personality factors only. Make sure to use cross-validation methods to avoid overfitting and characterize the accuracy of your model.
9) Build a prediction model of your choice (regression or supervised learning) to predict movie ratings (for all 400 movies) from gender identity, sibship status and social viewing preferences (columns 475-477) only. Make sure to use cross-validation methods to avoid overfitting and characterize the accuracy of your model.
10) Build a prediction model of your choice (regression or supervised learning) to predict movie ratings (for all 400 movies) from all available factors that are not movie ratings (columns 401-477). Make sure to use cross-validation methods to avoid overfitting and characterize the accuracy of your model.

Extra credit: Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions.

Hints:

*In order to do some analyses, you might have to apply a dimension reduction method first. For instance, "sensation seeking" and "movie experience" are characterized by 10-20 variables each. You'll need to distill their essence into much fewer factors first, before you can answer questions about the relationship between them. The same is true for personality, which is characterized by even more variables.
*In order to do some analyses, you will have to clean the data first, either by removing or imputing missing data (either is fine, but explain and justify what you did)
*If you encounter skewed data, you might want to transform the data before doing anything, e.g. z-scoring, using log-transforms or the like.
*For some hypothesis tests, you will have to discretize the data, i.e. transform numerical values into categories so you can use them as categorical variables, e.g. by doing a median-split.
*You can interpret "types" as clusters here.
*You can operationalize the popularity of a movie by how many ratings it has received.
*Avoid overfitting with cross-validation
*The accuracy of your model can be stated with RMSE, R^2 or AUC.
*You can use conventional choices of alpha (e.g. 0.05) or confidence intervals (e.g. 95%) throughout.