Data analysis project 2:

Applying machine learning methods to movie ratings data

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 tell you what we would like you to do. One exception: We do expect you to do this work yourself, so it reflects your intellectual contribution.

Purpose: In this project, you will demonstrate essential machine learning skills. We revisit the same dataset you already used in project 1. This will highlight what machine learning methods can and cannot do for you, compared to inferential methods. You also already know the dataset. Please write a report (1-3 pages, as needed) that answers the questions below. Use figures as needed to make your case.

Dataset description: This dataset features ratings data of 400 movies from 1097 research participants.
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: Movies are best enjoyed alone (1 = yes, 0 = no, -1 = no response)
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.

Data handling suggestions: To answer the questions properly, you'll have to do some kind of imputation of missing ratings (nans). Given the scope of this class, replacing them with a blend (50/50 is ok) of the arithmetic mean of each column and each row might be most suitable. Don't get used to this – there are many problems with this approach. But for now, this is ok - you'll learn more sophisticated methods later. But let's say that the rating of user 350 for movie 200 is missing and that the average rating of this user for other movies is 4 and the average rating (by other users) for this movie is 3, the to-be-imputed rating would be 3.5, using this method.

What we would like you to do: (each question is worth 20% of the grade score):

1) For each of the 400 movies, use a simple linear regression model to predict the ratings. Use the ratings of the *other* 399 movies in the dataset to predict the ratings of each movie (that means you'll have to build 399 models for each of the 400 movies). For each of the 400 movies, find the movie that predicts ratings the best. Then report the average COD of those 400 simple linear regression models. Please include a histogram of these 400 COD values and a table with the 10 movies that are most easily predicted from the ratings of a single other movie and the 10 movies that are hardest to predict from the ratings of a single other movie (and their associated COD values, as well as which movie ratings are the best predictor, so this table should have 3 columns).
2) For the 10 movies that are best and least well predicted from the ratings of a single other movie (so 20 in total), build multiple regression models that include gender identity (column 475), sibship status (column 476) and social viewing preferences (column 477) as additional predictors (in addition to the best predicting movie from question 1). Comment on how R^2 has changed relative to the answers in question 1. Please include a figure with a scatterplot where the old COD (for the simple linear regression models from the previous question) is on the x-axis and the new R^2 (for the new multiple regression models) is on the y-axis.

3) Pick 30 movies in the middle of the COD range, as identified by question 1 (that were not used in question 2). Now build a regularized regression model with the ratings from 10 other movies (picked randomly, or deliberately by you) as an input. Please use ridge regression, and make sure to do suitable hyperparameter tuning. Also make sure to report the RMSE for each of these 30 movies in a table, after doing an 80/20 train/test split. Comment on the hyperparameters you use and betas you find by doing so.
4) Repeat question 3) with LASSO regression. Again, make sure to comment on the hyperparameters you use and betas you find by doing so.
5) Compute the average movie enjoyment for each user (using only real, non-imputed data). Use these averages as the predictor variable X in a logistic regression model. Sort the movies order of increasing rating (also using only real, non-imputed data). Now pick the 4 movies in the middle of the score range as your target movie. For each of them, do a media split (now using the imputed data) of ratings to code movies above the median rating with the Y label 1 (= enjoyed) and movies below the median with the label 0 (= not enjoyed). For each of these movies, build a logistic regression model (using X to predict Y), show figures with the outcomes and report the betas as well as the AUC values. Comment on the quality of your models. Make sure to use cross-validation methods to avoid overfitting.

Extra Credit: Use machine learning methods of your choice to tell us something interesting and true about the movies in this dataset that is not already covered by the questions above [for an additional 5% of the grade score].

Note: *Please answer these questions in accordance with the AFYD guidance that was already posted for Data Analysis project 1.*