Fall 2016 OMS DAVA Project - Part II

Submission Deadline: 2016-11-30 (see T-Square for specific time)

- You may use R packages to fit and evaluate a regression model (no need to implement regression models yourself). Please stick to linear regression, however.
- We will continue to work with the movies dataset from part I. It is OK to use code and data that you created in Part I, but if you do please mention it and include that code in this submission.
- The task is to predict the profit associated with a movie, defined as gross revenue minus budget.
- In all the assignments below, please omit all rows in which gross and budget are not available.
- Since dollar amounts depend on the year they are reported in (due to inflation), remove all movies released prior to 2000. This is only a partial remedy, but it will be sufficient for our purposes.
- For all of the assignments below please use the following evaluation strategy:
  - Discard rows as described above
  - Randomly divide the rows into two sets of sizes 5% and 95%
  - Use the first set for training and the second for testing. Compute the MSE on the train and test sets (normalize the MSE by the number of samples).
  - Repeat the above data partition and model training and evaluation 10 times and average the MSE results so the results stabilize.
  - Repeat the above steps for different proportions of train and test sizes: 10%-90%, …, 95%-5%.
  - Generate a graph of the averaged train and test MSE as a function of the train set size
  - The instructor will give a (modest) award to the student whose model in Q5 performs the best on the test data.

Assignments:

1. Use linear regression to predict profit based on all available numeric variables. Graph the train and test MSE as a function of the train set size (averaged over 10 random data partitions as described above)?
2. Try to improve the prediction quality in (1) as much as possible by adding feature transformations of the numeric variables. Explore both numeric transformations such as power transforms and non-numeric transformations of the numeric variables like binning (e.g., is_budget_greater_than_3M). Explain which transformations you used and why you chose them. Graph the train and test MSE as a function of the train set size (averaged over 10 random data partitions as described above)?
3. Write code that featurizes genre (can use code from Part-I), actors, directors, and other categorical variables. Explain how you encoded the variables into features.
4. Use linear regression to predict profit based on all available non-numeric variables (using the transformations in (3). Graph the train and test MSE as a function of the train set size (averaged over 10 random data partitions as described above)?
5. Try to improve the prediction quality in (1) as much as possible by using both numeric and non-numeric variables as well as creating additional transformed features including interaction features (for example is_genre_comedy x is_budget_greater_than_3M). Explain which transformations you used and why you chose them. Graph the train and test MSE as a function of the train set size (averaged over 10 random data partitions as described above)?