IST 5420 – Business Analytics and Data Science, Spring 2017, Chen

**Project Evaluation Form – Milestone 2: Data Analysis I**

**Submission Due Mar 5, 11:59 PM**

**Instruction:**

1. Analyze data and report the results. The project report should include:

* Introduction (refined from M1)
* Data Source and Collection (refined from M1)
* Data manipulation (newly developed)
* Data visualization and summarization (newly developed)
* Predictive modeling (newly developed)

Read the evaluation criteria carefully on the next page for the detail.

1. Use R Markdown to write your project report. You need to use proper Markdown syntax to format your report.
2. Please submit the following documents into Canvas:

* The project report written in R Markdown .Rmd file;
* The Word/HTML/PDF report that is directly generated from your Markdown file;
* The Evaluation form with project team information (see below table).

**Project Team Information (filled in by students)**

|  |  |  |
| --- | --- | --- |
| **Member name** | **Percent contribution** | **Activities completed by the member** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Evaluation Summary – M2 (filled in by instructor)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Criteria** | **Target %** | **Comments** | **Evaluation** |
| * Manipulate your dataset properly * Properly detect outliers and deal with missing values in your dataset. | 10 | * Better to provide a table of variable description to explain the meaning for all variables and possible values for each categorical variable. * So the Occupation variable should be a factor? * As the zip code has so many levels, may try to use the first 5 digits to represent zip code in order to reduce the number of predictors (zip code will be converted as dummies). You can try both ways and evaluate the performance of your models. * Check the range of length for the zip code variable. Find if there are some invalid values in your dataset. * You need to interpret those “X” columns in the movie\_info object. Better to rename those automatically named columns by R. * There is no “All\_Genres” column in your movie\_info object. * The timestamp might be useful. Try to convert it to a timestamp object [use the as.POSIXct() method, for example]. The timestamp may help to show if there is some time effect. You can get the movie release year from the movie title and draw to graph to show if the average user rating for each movie changes over time. * Why some users has the age as 1? Are they valid observation? Some websites may have some QA accounts for testing purpose. * The statement “For our model, user location is irrelevant so we will drop the ‘Zip Code’ variable” is problematic. How did you know location information does not matter? You know this only after you get some evidence from your dataset. Don’t drop them at this stage, you just need to cleanse your data. In this case, remove all zip code values that do not have the correct format. Do not use the remove NAs function as it will drop the whole observation. Just keep missing values in your dataset. * For the same reason, don’t drop occupation variable. You always don’t want to lose some potentially useful information for your predictive modeling. * Your logic for loop is seriously flawed. From the result of **head**(df), you can easily find that the same user ID has multiple age values. Use the join functions in the dplyr package to combine the two datasets. (-1) * Why did you generate a user-movie matrix? The user and movie have a n:m relationship and such relationship is perfectly stored in the rating\_info object. The matrix seems only record one rating score (there are multiple rating scores for each user-movie combination) * Why did you have a “rate.csv” data file? The data structure is very different with other objects (it does not contain all genre dummies). If you need to do sampling, then use R code to sample and keep the same data structure. Don’t hide any logic that is outside of your R markdown. Or else, the other people won’t be able to replicate your predictive modeling. (-1) * The same problem for “uservsmov.csv”. * Also you replace the old df object by reading the “rate.csv” file? Why? * As page 8 shows, you have 1000146 observations (not the number you specified). Use sample() method won’t have this problem. | 8 |
| * Summarize and visualize data by using appropriate methods. * At least 10 professional graphs with detailed and proper interpretations are required. | 40 | * Provide a summary statistics for your dataset (use the stargazer package). * A better way to visualize the rating across genre is to create a new data frame that shows the average user rating for each genre per gender (use all dataset, no sampling), then plot this special data frame on a grouped bar chart. The point is not to directly use the large dataset for plotting, instead aggregate the dataset into higher levels to create a small dataset that is used for plotting. This can significantly reduce the time consumed for plotting and the size of your report. Now your pdf report is more than 300M, which is unusual.(-2) * Follow this similar logic, you can easily show a count of rating for each genre per gender. This is for your whole dataset, so there is no bias due to sampling. * Redo your visualization in M3. | 38 |
| * Properly use predictive analytics to solve some classification and prediction problems on your dataset. * The predictive analytics should include model building and tuning to get a “best” predictive model for your dataset. * At least use 4 prediction and classification models, properly visualize these models and compare their performance. | 40 | * Essentially, you just use genres to predict rating. This model is too simple and misses so many important predictors. As a result, you won’t be able to accurately predict rating. You need to include all available and potentially information (user characteristics and movie attributes) in your models. As commented in the data manipulation sections, you need to consider adding additional information such as location and some time factors. (-5) * You need to interpret each step of your predictive modeling. Why you did so? Why what you did are legitimate? Why you choose to transform some variables? Why you include some variables as predictors? What potential problem they may have? You need to tell an interesting story along with your data analyses. (-2) * For example, in your regression model, what conclusions can you get from the data analysis? How would you like to interpret the effect? * Compare the performance of your model. (-1) * Netflix uses ensemble method (refer to <https://amba-bigdata.wikispaces.com/file/view/Netflix_general.pdf>). You can try this method to see if you can improve the predictive performance. To learn more about the ensemble method, refer to the article [Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research, 11, 169-198.] | 32 |
| * Format your project report in a professional way. * Write your project report by using appropriate Markdown syntax. | 10 | * The level of headings in your R markdown is not consistent across the whole report. You can easily check the headings in the pdf report by looking at the bookmarks (see below).      * You also need to remove all unnecessary content that is automatically generate by the R markdown template (check the last two paragraphs). (-1) | 9 |
| The report satisfies all of the following criteria:   * It tells a very interesting story; * The graphs and tables as well as their interpretations are very professional; * The whole document is well written | 10 bonus | Some editing issues:   * Page 1: “…to rate atleast five movies…” * Page 2: “Each user rated atleast 20 different movies” | 10 |
| **Total** | **100** |  | **97** |