**IS 6482 001 - Spring 2017  
Assignment 2 – Video Game Sales Prediction**

**Due: 11:59 pm, February 20 (Monday)**

**Overview:** The numeric prediction tasks in this assignment is created from a modified version of the video game sales with rating data in Kaggle - <https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings>. The slightly cleaner version of the data set includes a single sales variable – Global\_Sales and selected predictors that are more practical to including in an assignment in introductory data mining class. Some missing values and outliers have been removed too.

**Input file and data dictionary:** sales\_filtered.csv. The target variable is Global\_Sales.

Self-explanatory data fields from <http://www.vgchartz.com/> include Name, Platform, Year\_of\_Release, Genre, Publisher, NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales, and Global\_Sales, out of which sales\_filtered.csv includes only Name, Platform, Genre, and Global\_Sales. Do not use the Name field in any model building. It is included only for your information.

Data fields from [www.metacritic.com](http://www.metacritic.com) include:

Critic\_score - Aggregate score compiled by Metacritic staff

Critic\_count - The number of critics used in coming up with the Critic\_score

User\_score - Score by Metacritic's subscribers

User\_count - Number of users who gave the user\_score

Developer - Party responsible for creating the game

Rating - The [ESRB](https://www.esrb.org/) player age and content ratings

**Packages required:** Install Rmarkdown, rpart, rweka, caret, rminer, matrixStats and knitr packages.

**Submission:** Submit two files – A2\_your\_initials.Rmd which is an R code file together with comments and text required by this assignment’s tasks, and A2\_your\_initials.html, (or another output format) generated from rendering (or knitting) A2\_your\_initials.Rmd.

**Task I:** Create A2\_your initials.rmd to meet the following requirements:

Use beginning text (meta fields) to include assignment title, author name – you, and the file creation date. Set output to an output format of your choice.

Create code chunks to meet the following requirements. You do not need to write down nor submit answers to 1.C, 1.E, 3.C, 4.C, 4.E, 5.C, and 5.E. These are questions intended to guide you to examine the model outputs.

1. Code chunk 1 or code chunk group 1 (20%)- Set up, data import, data exploration, data partitioning and inspection code for the following:
   1. Package loading, and data import. Show the overall structure and summary of the input data. Other than Name, all other non-numeric fields should be factor variables.
   2. Include commands to explore numeric variables’ distributions and their correlations, as well as commands to explore factor variables and their relationships to the target variable.
   3. *Think about how 1.B help you assess the potential strong predictors and potentially helpful data transformations, and anticipate the potential model fit performance and the necessary improvements.*
   4. Use the whole data set without the Name field, build a linear regression model. Show the summary of the model to understand the significance and coefficients of the predictors in the model and the overall model fit.
   5. *Think about the evidence for why your thoughts for task I.C are confirmed or disconfirmed.*
   6. Partition the data set for simple hold-out evaluation – 70% for training and the other 30% for testing.
   7. Show the overall summaries of train and test sets.
2. Code chunk 2 (20%)– Simple lm, rpart and M5P model training and testing:
   1. Train three models using lm, rpart, and M5P. Use the default settings of these methods throughout this assignment.
   2. Generate and this model’s explanatory evaluation metrics and predictive error metrics (as in Week 4 tutorials) in both the testing and training sets.
3. Code chunk 3 (20%) – Cross-validation of simple lm, rpart, and M5P models:
   1. Define a named function for cross-validation of numeric prediction models that generates a table of the model fit and error metrics used in Week 4 tutorials for each fold along with the means and standard deviations of the metrics over all folds.
   2. Call the function in 3.A to generate 10-fold cross validation results of the simple lm, rpart and M5P models.
   3. *Are the performance results from 2.B and 3.B similar or not? What do the comparisons suggest? Are you concerned with applying some or all of the models to predict the target variable when new data that has similar patterns, is available? Why or why not?*
4. Code chunk 4 (20%)– Improve the models with the quadratic term of User\_Count:
   1. Create and add the quadratic term of User\_Count, e.g., User\_Count\_Squared, to the predictors for the target variable.
   2. Build a lm model with User\_Count\_Squared included. Show the summary of this lm model.
   3. *Has the model fit improved over that from 1.D? What are the changes in the coefficients and significance of predictors compared to those of the model in 1.D? Are these changes meaningful and actionable? Would you suggest some other quadratic terms, interaction terms or predictor transformations that might further improve the model fit?*
   4. Call the cross validation function defined for 3.A, to generate 10-fold cross validation results of the simple lm, rpart and M5P models with User\_Count\_Squared included
   5. *Are the performance results from 4.B and 4.D similar or not? What do the comparisons suggest? Are you concerned with applying some or all of the models in 4.D to predict the target variable when new data that has similar patterns, is available? Why or why not?*
5. Code chunk 5 (15%) – Improve the models with the log term of User\_Count:
   1. Create and add the natural log transformation of User\_Count, e.g., log\_User\_Count, to the predictors for the target variable.
   2. Build a lm model with log\_User\_Count included and User\_Count excluded. Show the summary of this lm model.
   3. *Has the model fit improved over that from 4.B? What are the changes in the coefficients and significance of predictors compared to those of the model in 4.D? Are these changes meaningful and actionable? Would you suggest some other quadratic terms, interaction terms or predictor transformations that might further improve the model fit?*
   4. Call the cross validation function defined for 3.A, to generate 10-fold cross validation results of the simple lm, rpart and M5P models with log\_User\_Count included and User\_Count excluded.
   5. *Are the performance results from 5.B and 5.D similar or not? What do the comparisons suggest? Are you concerned with applying some or all of the models in 5.D to predict the target variable when new data that has similar patterns, is available? Why or why not?*

For each chunk:

* Add some simple descriptive text in the text area before the code chunk.
* Add a name or description of each code chunk in {r}. Be sure that you allow code and output from executing code to be included in the file from rendering A2\_your\_intials.Rmd.
* Add comment lines for each code requirement item.

**Task II (5%):**

Render A2\_your\_initials.Rmd to HTML output format. You can click on the “Knit HTML” button above the source code panel in RStudio.