- Due Jun 14, 2023 by 11:59pm
- Points 100
- Submitting a file upload
- Available until Jun 17, 2023 at 11:59pm

This assignment was locked Jun 17, 2023 at 11:59pm.

A4 Regression models, model fit and prediction errors

Instructions

Overview: The numeric prediction task in this assignment is created from a modified version of the video game sales data challenge in Kaggle - **https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings** **(https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings)** . The preprocessed data - **NA_sales_filtered.csv** **(https://utah.instructure.com/courses/882385/files/147060997/download?wrap=1)** **(https://utah.instructure.com/courses/882385/files/147060997/download?download_frd=1)** includes a single sales variable – *NA_Sales* and selected predictors that are more practical for inclusion in an assignment in an introductory data mining class.

Some missing values and outliers have been removed too.

Input file: NA_sales_filtered.csv **(https://utah.instructure.com/courses/882385/files/147060997/download?wrap=1)** **(https://utah.instructure.com/courses/882385/files/147060997/download?download_frd=1) .** **(https://utah.instructure.com/courses/882385/files/147060997/download?wrap=1)** **(https://utah.instructure.com/courses/882385/files/147060997/download?download_frd=1)** The target variable is ***NA_Sales***.

Below please find the data fields in NA_sales_filtered.csv and their description.

Target variable: NA_Sales

Other Variables:

Name: the video game name (It is included only for your information.) **Do not use the Name column in any model building.**

Platform: video game platform

Genre: the category of a video game

Rating: the **ESRB** **(https://www.esrb.org/)** player age and content ratings

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

Packages required: Install Rmarkdown, psych, rpart, RWeka, caret, rminer, matrixStats and knitr packages.

Submission: Submit two files – A4_yourLastName_yourFirstName.Rmd and A4_yourLastName_yourFirstName.html, generated from A4_yourLastName_yourFirstName.Rmd.

(Please do not compress these two files into one zip file. Submit them as separate files. Be sure that the HTML output file contains assignment title, author name – you, the file creation date, and table of content including section numbers.)

Note: If there is any discrepancy in the task descriptions in the assignment page and the short rubric descriptions, please follow the task descriptions in the assignment page.

Task I (95%):

Create A4_yourLastName_yourFirstName.rmd to meet the following requirements.

1. Code chunk 1 (20%)- Set up, data import, data exploration, data partitioning, and inspection code

A. Package loading, and data import.

Set the working directory to the directory where your rmarkdown program file resides in rstudio using getwd() and setwd(). For example,

```
mydir <- getwd()

setwd(mydir)
```

Load character strings as character fields. Show the overall structure and summary of the input data. Other than the *Name*, transform all other non-numeric fields to be factor variables.

B. Use pairs.panels to show distributions and correlations of all of the numeric variables.

C. Remove the Name variable from the data frame. All subsequent models should have this column excluded. 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. *Note that the purpose of this task is not to build a predictive model. Rather, it is often a good idea to explore a data set with white-box models like linear regression (for numeric target variable) or decision tree (for factor target variable).*

D. Partition the dataset for simple hold-out evaluation – 70% for training and the other 30% for testing.

E. Show the overall summaries of training and testing sets.

2. Code chunk 2 (20%)– lm, rpart and M5P model training and testing

A. Train three models using lm, rpart, and M5P on the training set (built in 1. D). Use the default settings of these methods throughout this assignment.

B. For each of the three models trained in 2.A, perform the following:

i) Show information about the model by specifying the model name, and summary(model name).

ii) Apply the model and generate the model-fit (R2) and prediction error metrics (MAE, MAPE, RAE, RMSE, RMSPE, RRSE) in both the testing and training sets.

3. Code chunk 3 (20%) – Cross-validation of lm, rpart, and M5P *NA_Sales* prediction models

A. Define a named function for cross-validation of numeric prediction models that generates a table of the model fit and error metrics specified in 2.B for each fold along with the means and standard deviations of the metrics over all of the folds.

B. Call the function in 3.A to generate 5-fold cross-validation results of lm, rpart and M5P models for *NA_sales.*

4. Code chunk 4 (20%) – Improve the models by adding a quadratic term of *Critic_Score*

A. Create and add the quadratic term of *Critic_Score*, e.g., *Critic_Score_Squared*, to the predictors for *NA_Sales* in the whole data set for this assignment.

B. Build an lm model using the whole data set that includes *Critic_Score_Squared* to predict *NA_Sales*. Show the summary of this lm model. This allows you to inspect if this squared term is significant or not.

C. Call the cross-validation function defined for 3.A to generate 5-fold cross-validation results of the lm, rpart and M5P models with *Critic_Score_Squared.*

5. Code chunk 5 (15%) – Improve the models with the log term of *User_Count*:

A. Create and add the natural log transformation of *User_Count*, e.g., *log_User_Count*, to the predictors for the target variable. The following is an excerpt of sample code in webinar's demo:

```
# Remove the original User_Count (7th column) and create a new data frame

df_log_User_Count <- sales[,-7]

# Create and add the natural log transformation of User_Count
 df_log_User_Count$log_User_Count <- log(sales$User_Count)
```

B. Build an lm model with the whole data set that includes *log_User_Count* and excludes *User_Count*. The input data should not include any quadratic terms created in the previous code chunk. Show the summary of this lm model. This allows you to inspect if this log term is significant or not.

C. Call the cross-validation function defined for 3.A to generate 5-fold cross-validation results of the lm, rpart, and M5P models with *log_User_Count* included and *User_Count* excluded.

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 A4_yourLastName_yourFirstName.Rmd.
- Feel free to add comment lines with the requirement item numbers (e.g., # 3.A or # 3.B) to your code cell to help TAs and instructors easily identify your code that addresses a particular requirement.
- Be sure to remove output or code that is not required for the tasks in this assignment.

Task II Reflections (5%):

What have you learned from building each of these models and the modeling impact of your adjustments to the hyperparameters or dataset? If you were explaining the results of these models to a supervisor what would you say about them? Attempt to do more than just state facts here, interpret the results. Coding is great, interpretation of output is even more important. Discuss each model. Write at least 150 words.

Use the RMD header to include assignment title, author name – you, and the file creation date. Also, include header specifications to generate a table of contents and section numbers of your code chunks.

Output:

Render A4_yourLastname_yourFirstname.Rmd to HTML output format. You can click on the “Knit HTML” button above the source code panel in RStudio.

Submit two files – A4_yourLastName_yourFirstName.Rmd and A4_yourLastName_yourFirstName.html, generated from A4_yourLastName_yourFirstName.Rmd.

(Please do not compress these two files into one zip file. Submit them as separate files.)

Assignment-3

Criteria	Ratings			Pts
1.A Package loading, and data import. Load character strings as character fields. Show the overall structure and summary of the input data. Other than the Name, transform all other non-numeric fields to be factor variables.	**6 to >5.5 pts** **Full Marks**	**5.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	6 pts
1.B Use pairs.panels for all of the numeric variables.	**3 to >2.5 pts** **Full Marks**	**2.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	3 pts
1.C.i Remove the Name variable from the data frame	**1 to >0.5 pts** **Full Marks**	**0.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	1 pts
1.C.ii Build a linear regression model. Show the summary of the model.	**5 to >4.5 pts** **Full Marks**	**4.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	5 pts
1.D Partition the dataset– 70% for training and the other 30% for testing.	**3 to >2.5 pts** **Full Marks**	**2.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	3 pts
1.E Show the overall summaries of training and testing sets.	**2 to >1.75 pts** **Full Marks**	**1.75 to >0.0 pts** **Partially**	**0 pts** **No Marks**	2 pts
2.A.i Train lm model on the training set (built in 1. D). Use the default settings.	**2 to >1.75 pts** **Full Marks**	**1.75 to >0.0 pts** **Partially**	**0 pts** **No Marks**	2 pts
2.A.ii Train rpart on the training set (built in 1. D). Use the default settings.	**3 to >2.5 pts** **Full Marks**	**2.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	3 pts
2.A.iii Train M5P model on the training set (built in 1. D). Use the default settings.	**3 to >2.5 pts** **Full Marks**	**2.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	3 pts
2.B.i Generate lm model’s explanatory evaluation metrics and predictive error metrics (Total: 7 metrics) in both the testing and training sets.	**4 to >3.5 pts** **Full Marks**	**3.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	4 pts
2.B.ii Generate rpart model’s explanatory evaluation metrics and predictive error metrics (Total: 7 metrics) in both the testing and training sets.	**4 to >3.5 pts** **Full Marks**	**3.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	4 pts
2.B.iii Generate M5P model’s explanatory evaluation metrics and predictive error metrics (Total: 7 metrics) in both the testing and training sets.	**4 to >3.5 pts** **Full Marks**	**3.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	4 pts
3.A Define a named function for cross validation of numeric prediction models that generates a table of the model fit and error metrics (7 total) for each fold along with the means and standard deviations of the metrics over all of the folds.	**11 to >10.0 pts** **Full Marks**	**10 to >0.0 pts** **Partially**	**0 pts** **No Marks**	11 pts

Criteria	Ratings			Pts
3.B.i Call the function in 3.A to generate 5-fold cross validation results of lm model for NA_sales.	**3 to >2.5 pts** **Full Marks**	**2.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	3 pts
3.B.ii Call the function in 3.A to generate 5-fold cross validation results of rpart model for NA_sales.	**3 to >2.5 pts** **Full Marks**	**2.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	3 pts
3.B.iii Call the function in 3.A to generate 5-fold cross validation results of M5P model for NA_sales.	**3 to >2.5 pts** **Full Marks**	**2.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	3 pts
4.A Create and add the quadratic term of Critic_Score.	**3 to >2.5 pts** **Full Marks**	**2.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	3 pts
4.B Build an lm model using the whole data set that includes the squared term of of Critic_Score, e.g., Critic_Score_Squared, to predict NA_Sales. Show the summary of this lm model.	**5 to >4.5 pts** **Full Marks**	**4.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	5 pts
4.C.i Call the cross-validation function defined for 3.A, to generate 5-fold cross-validation results of the lm model with Critic_Score_Squared	**4 to >3.5 pts** **Full Marks**	**3.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	4 pts
4.C.ii Call the cross-validation function defined for 3.A, to generate 5-fold cross-validation results of the rpart model with Critic_Score_Squared	**4 to >3.5 pts** **Full Marks**	**3.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	4 pts
4.C.iii Call the cross-validation function defined for 3.A, to generate 5-fold cross-validation results of the M5P model with User_Count_Squared	**4 to >3.5 pts** **Full Marks**	**3.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	4 pts
5.A Create and add the natural log of User_Count	**2 to >1.5 pts** **Full Marks**	**1.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	2 pts
5.B Build an lm model with the whole data set that includes log_User_Count and excludes User_Count. Show the summary of this lm model. Remove the previously added squared term.	**4 to >3.5 pts** **Full Marks**	**3.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	4 pts
5.C.i Call the cross-validation function defined for 3.A, to generate 5-fold cross-validation results of the lm model with log_User_Count included and User_Count excluded. Remove the previously added squared term.	**3 to >2.5 pts** **Full Marks**	**2.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	3 pts
5.C.ii Call the cross-validation function defined for 3.A, to generate 5-fold cross-validation results of the rpart model with log_User_Count included and User_Count excluded. Remove the previously added squared term.	**3 to >2.5 pts** **Full Marks**	**2.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	3 pts
5.C.iii Call the cross-validation function defined for 3.A, to generate 5-fold cross-validation results of the M5P model with log_User_Count included and User_Count excluded. Remove the added squared term. Remove the previously added squared term.	**3 to >2.5 pts** **Full Marks**	**2.5 to >0.0 pts** **Partially**	**0 pts** **No Marks**	3 pts

Criteria	Ratings			Pts
What have you learned from building each of these models and the modeling impact of your adjustments to the hyperparameters or dataset? If you were explaining the results of these models to a supervisor what would you say about them? Attempt to do more than just state facts here, interpret the results. Coding is great, interpretation of output is even more important. Discuss each model. Write at least 150 words.	**5 to >4.0 pts** **Full Marks**	**4 to >0.0 pts** **Partially**	**0 pts** **No Marks**	5 pts
Total Points: 100				