HarvardX Data Science Program:

PH125.9x - Capstone - Personal Project

Travis Horesh

November 2022

Introduction

The goal of this project was to further build on the modeling and prediction techniques practiced during the movielens project portion of the course. Again, using the knowledge and techniques developed over the course of the HarvardX Data Science program on edX. Primarily, this focused on the machine learning topics discussed during the preceding course (HarvardX PH125.8x).

For my personal project, I utilized one of the Kaggle datasets recommended in the course literature. Specifically, this was the Video Game Sales with Ratings dataset. As noted in the Kaggle documentation, the dataset is combination of two existing datasets which were originally obtained by separately scraping VGChartz and Metacritic.

The dataset used for this project contains a total of 16,719 observations with 16 variables. However, there are significant gaps in the data. After eliminating all records with null values in any variable, only 6,826 observations remain. In contrast to the Movielens project, this project required more cleaning and transforming of the data before it could be analyzed.

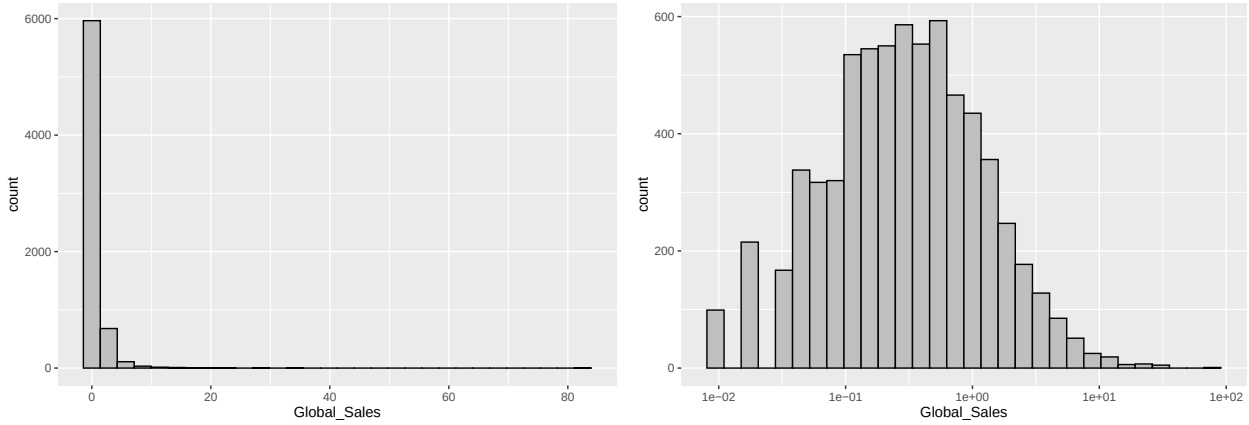
In summary, the project consisted of ingesting the data, an iterative process of exploring, cleaning, and transforming the raw dataset, creating the train, test, and validation subsets, and developing and refining a model to predict global sales.

Analysis

After exploring the dataset and examining the correlation between numerous variables and global sales, my analysis focused on using simple linear models which utilized the most correlated variables. However, prior to fitting models, more significant data preparation was required. Changes to the dataset are noted below:

- Formatting Year_of_Release as numeric
- Formatting User_Score as numeric and aligning its scale and precision with Critic_Score
- Removing NAs
- Adding a column for the count of platforms on which each game was released
- Relabel games with AO, K-A, and RP rating, which had only one record each, to “Other”
- Create columns to capture whether games were produced by a top publisher or developer

Plotting the distribution of sales revealed little spread in the data. Therefore, subsequent analysis and model development used both unmodified sales values and those same values on a logarithmic scale



Datasets

Once cleaned and explored, the main dataset was split 90/10, with 90% retained for use in model development and the remaining 10% set aside as a hold-out or validation dataset. The 90% retained for model development was further split into training and test sets using a test index and the createDataPartition() function, with 90% of the observations allocated to the training set and the remaining 10% to the test set. One key step in data preparation had to occur at this point - ensuring all three datasets (train, test, and validation) contained the same levels as the main dataset. This was achieved using the rbind function, a simple for loop function, and the levels function.

```
total_data_split <- rbind(data_split, validation)
for (f in 1:length(names(total_data_split))) {
  levels(data_split[, f]) <- levels(total_data_split[, f])
}
```

The entire process of eliminating NAs consisted of simply piping data to a filter function using is.na

```
# Remove all NAs
data_clean <- data %>% filter_all(~!is.na(.))
```

Considering the small size of the dataset, I relied on the lm() function to fit a linear model. I chose to create three total variations, which are detailed below in the bulleted lists:

Main Model

- Critic Score
- User Score
- Genre
- Year of Release
- Number of Platforms
- Critic Count
- User Count
- Rating
- Whether it was produced by a top publisher or developer

Logarithmic Model

- Same features as the Main Model
- Used logarithmic values of Global Sales instead

Minimalist Model

- Reduced features to Critic Score, Genre, Year of Release, Platform, Rating, and whether it was produced by a top publisher or developer

Additionally, a naive model was created using the overall average of Global Sales. The models were first run on the test set to judge initial performance. The more complete model using logarithmic sales data performed the best based on RMSE, with both the non-logarithmic complete model and the "minimalist" model beating the naive model on RMSE.

Next, the Main Model and the Logarithmic Model were run on the validation set. The Main Model using logarithmic sales data performed better, even better than on the test set data, based on RMSE, while the non-log complete model actually performed worse.

Results

Method	RMSE
Naive Model	1.962523
lm_fit_test	1.803190
lm_fit_log_test	1.149829
lm_fit_min_test	1.855923
lm_fit_log_val	1.100874
lm_fit_val	2.290683

Conclusion

I succeeded in developing a model that performs better than simply guessing the average, but there is significant room for improvement. Several derived features or combinations of features were not explored, as well as other model types (such as knn or random forest) and other more advanced techniques.

So, linear regression as a method of predicting video game sales is viable but may not necessarily be the preferred method.

Overall, the project provided another excellent opportunity to test and solidify knowledge gained in the latter portions of the HarvardX Data Science program. I look forward to feedback from my peers and further improving my skills and knowledge in this area!