DSC550 – Data Mining Final Term Project Write-Up

The basis for this project has evolved as the project has progressed over time. Ultimately the problem we will try to bring some light to is predicting North American video game sales of an upcoming release. The problem at hand would be worth exploring to any video game development company, such as Activision, Nintendo, or Amazon Games. The insights gained from a predictive analysis of a game release in North America would help the company decide how to market the game and possibly how much a game would need to be marketed to better perform in sales.

To pitch this project to a group of stakeholders, I would highlight and focus on the immediate and long-term benefits of the knowledge to be gained by working on the problem. The immediate benefit would be the prediction of how well the new game release would perform in the Noth American market. This would allow for a better understanding of how and when to market the new release. The long-term benefit of tackling this project is to fine tune metrics for analysis in the future. Once the game has been released and more data has been gathered the model can be adjusted to improve in the future.

The data for this project was sourced from Kaggle.com, it can be found with the following link: "https://www.kaggle.com/datasets/thedevastator/global-video-game-sales". I originally had used a different data set for this project and had a different topic but found justifying the exploration of the problem too difficult. The data set used for this project is one that was previously used in the course and one that I found interesting to explore.

Summary of Milestones One Through Three

Milestone One – Exploratory Data Analysis

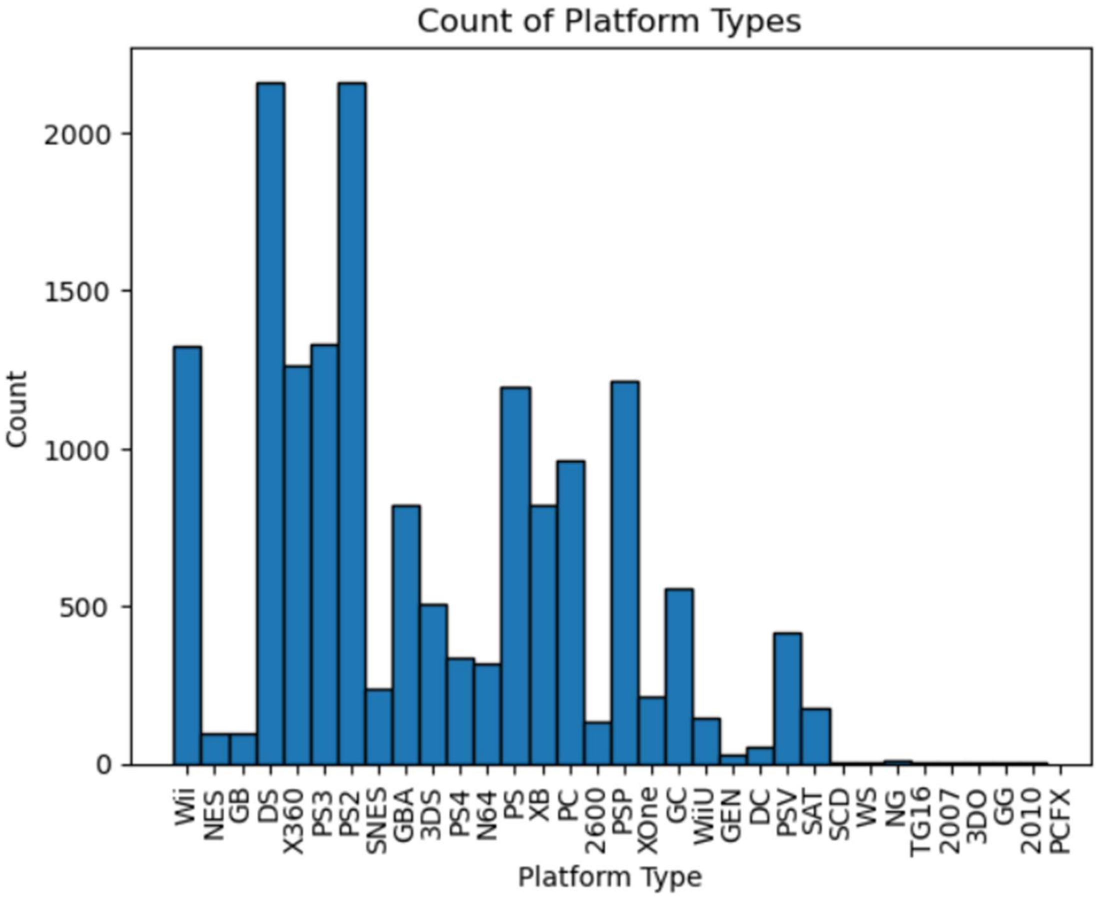
Milestone one was a mix of challenges for me. The biggest challenge in milestone one was the selection of the business problem for the project. A couple of factors played into the struggles I experienced for this portion of the project. One factor was the fact that I was forced to select a topic outside of my existing domain of knowledge. I currently (kind of) work for a large telecommunications company and have done so for the past 10 years. It would have been very easy for me to come up with multiple business problems to address and solve within this industry. However, this brings me to the second factor that created challenges for milestone one; having a business problem that also has enough accessible data. Even though I had plenty of business issues to explore in my existing domain of knowledge, I did not have enough accessible data to work with. This forced me to look outside my box and try to find a problem with data that existed.

After the business problem and data selection were completed, exploratory data analysis was done on the selected data set. I started by importing the downloaded .csv file into Jupyter Notebooks as a Pandas data frame. I reviewed the data using the ".head()" function and observed the values associated with each column. As part of the EDA process, I completed several visualizations on the data frame before making any changes to the data. The EDA process helps in initial discovery of how the features may or may not be related. EDA is also essential for viewing the state of the raw data set and reveals what changes may need to be made in the future.

The first visualization, Figure 1, created during the EDA process was a histogram of the feature "Platform Type". The histogram shows the counts of each platform type in the data set. The histogram can be interpreted as giving some indications of which platforms are the most popular to publish games for. The second visualization, Figure 2, created during the EDA process was a boxplot of the North American Sale by Platform. The boxplot shows the platforms along the x-axis and the percentage of North American sales on the y-axis. The boxplot reveals outlier data, indicated by the diamond icon in the figure as well as the quartiles and mean values.

Figure 1

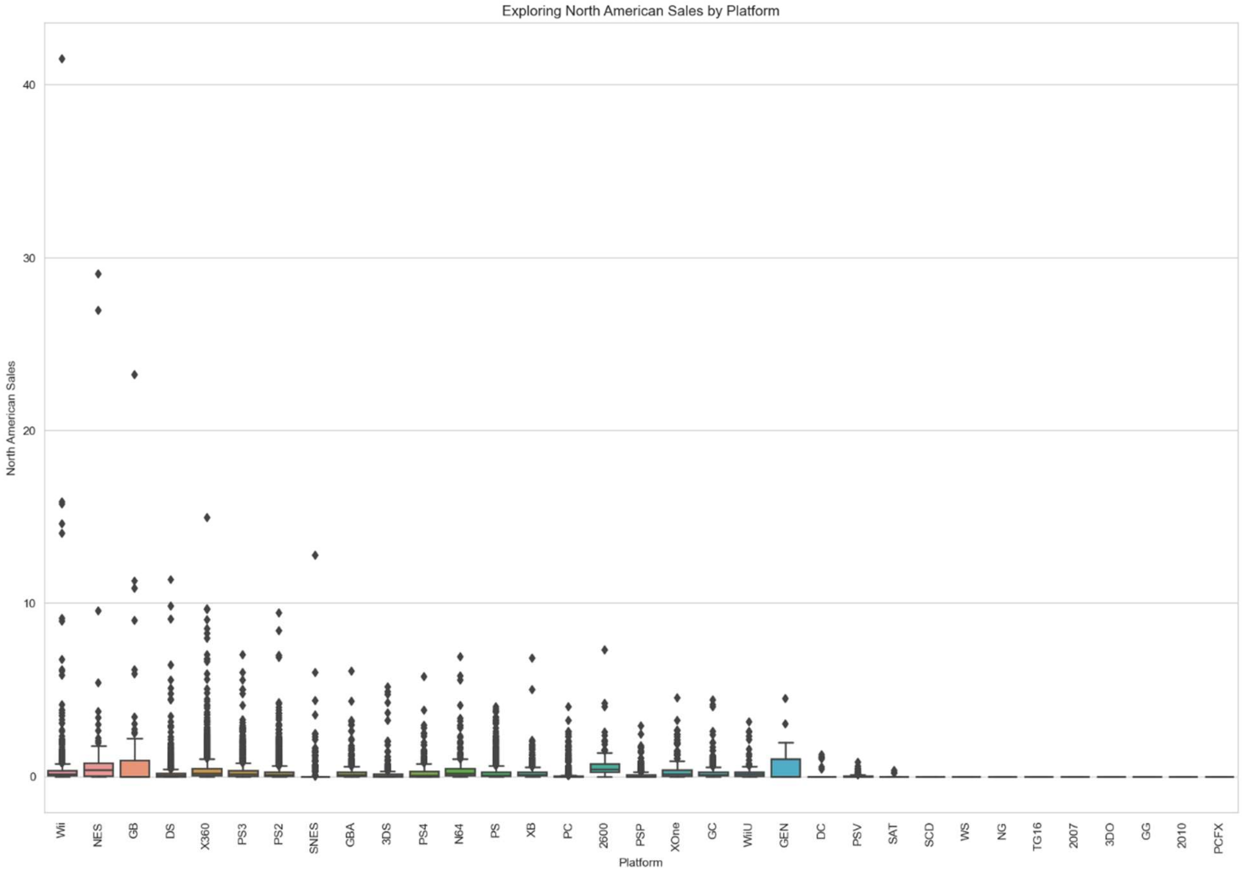
Histogram of Platform Types



Note: This figure illustrates the number of game titles by platform type. The count would be how many times the platform was represented by a video game title. No cleansing of the data set has taken place, this is the raw data set as downloaded.

Figure 2

Boxplot of North American Sales by Platform



Note: This figure illustrates the mean, quartiles, and outliers of North American Sales by Platform. As part of the EDA process the data has not been cleansed or changed. This is a representation of the raw data.

Milestone Two – Data Preparation

Milestone two of the project consisted of cleaning and preparing the data for future use in models. Cleaning and preparing the data for the model is one of the most pivotal steps when it comes to preparing data for the model. I found this step of the project the most time-consuming but also the most enjoyable. Something to note is that this step was revisited multiple times after the model evaluation and additional cleaning was done based on the feedback from future steps.

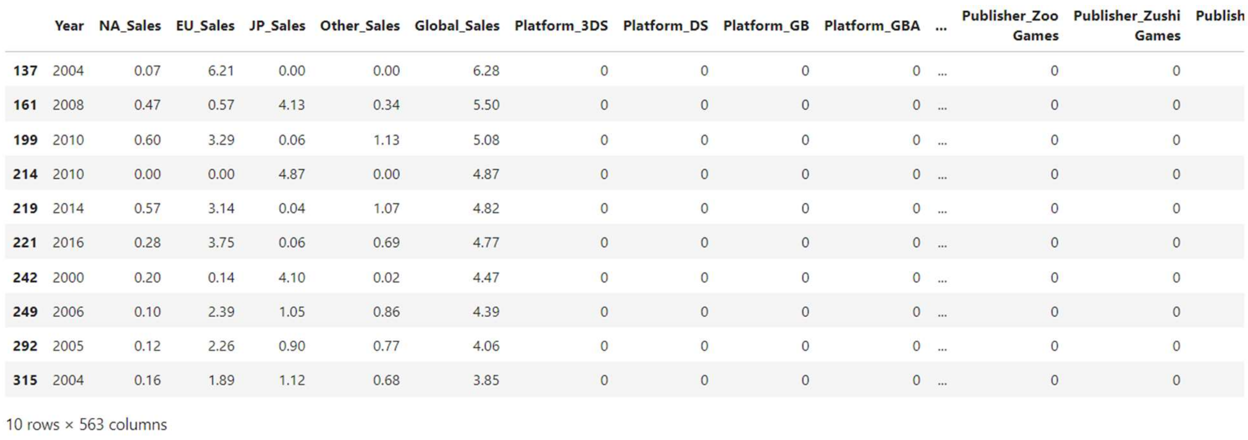
As part of the data preparation phase of the project, various steps were taken to clean the data. Based on the observations made as part of the EDA, certain types of platforms were dropped from the data frame. The platforms dropped did not have any significant number of games and would not have an impact on the target of North American sales. The "Name" and "Rank" columns were also dropped, these two features would have little impact, if any, on the data and model.

The next step taken in the data preparation process was to check for "NaN" or "Null" values in the data frame. 270 "NULL" values were found in the "Year" feature and 56 were found in the "Publisher" feature. It was determined that by dropping the rows with "NULL" values for the "Publisher" feature little impact to the model would occur. The "Null" values for year were replaced with the median "Year" value since the median is resistant to outliers. An additional cleansing step was added after completing the original Milestone two, which was removing North American Sales outliers based on quartile. This should scale the data better and create a better fitting model. The overall shape of the data frame after cleansing is 14006 rows and 9 columns.

The final original step in the data preparation process was to create dummy variables for the categorical features. This was done using the built-in Pandas function "get_dummies()" with a parameter of "dtype=int" to return an integer value of one or zero instead of a Boolean value. Creating the dummy variables with one-hot encoding made sense since the categorical values did not have any ordinal significance. After generating the dummy variables, the overall shape of the data frame increased to 14006 rows and 563 columns. A sample of the data frame can be seen in Figure 3.

Figure 3

Partial view of the Cleansed Data Frame



	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Platform_3DS	Platform_DS	Platform_GB	Platform_GBA	...	Publisher_Zoo Games	Publisher_Zushi Games	Publish
137	2004	0.07	6.21	0.00	0.00	6.28	0	0	0	0	...	0	0	
161	2008	0.47	0.57	4.13	0.34	5.50	0	0	0	0	...	0	0	
199	2010	0.60	3.29	0.06	1.13	5.08	0	0	0	0	...	0	0	
214	2010	0.00	0.00	4.87	0.00	4.87	0	0	0	0	...	0	0	
219	2014	0.57	3.14	0.04	1.07	4.82	0	0	0	0	...	0	0	
221	2016	0.28	3.75	0.06	0.69	4.77	0	0	0	0	...	0	0	
242	2000	0.20	0.14	4.10	0.02	4.47	0	0	0	0	...	0	0	
249	2006	0.10	2.39	1.05	0.86	4.39	0	0	0	0	...	0	0	
292	2005	0.12	2.26	0.90	0.77	4.06	0	0	0	0	...	0	0	
315	2004	0.16	1.89	1.12	0.68	3.85	0	0	0	0	...	0	0	

10 rows × 563 columns

Note: Figure 3 illustrates the cleaned data frame and the categorical features that have had dummy values substituted.

Milestone Three – Model Building and Evaluation

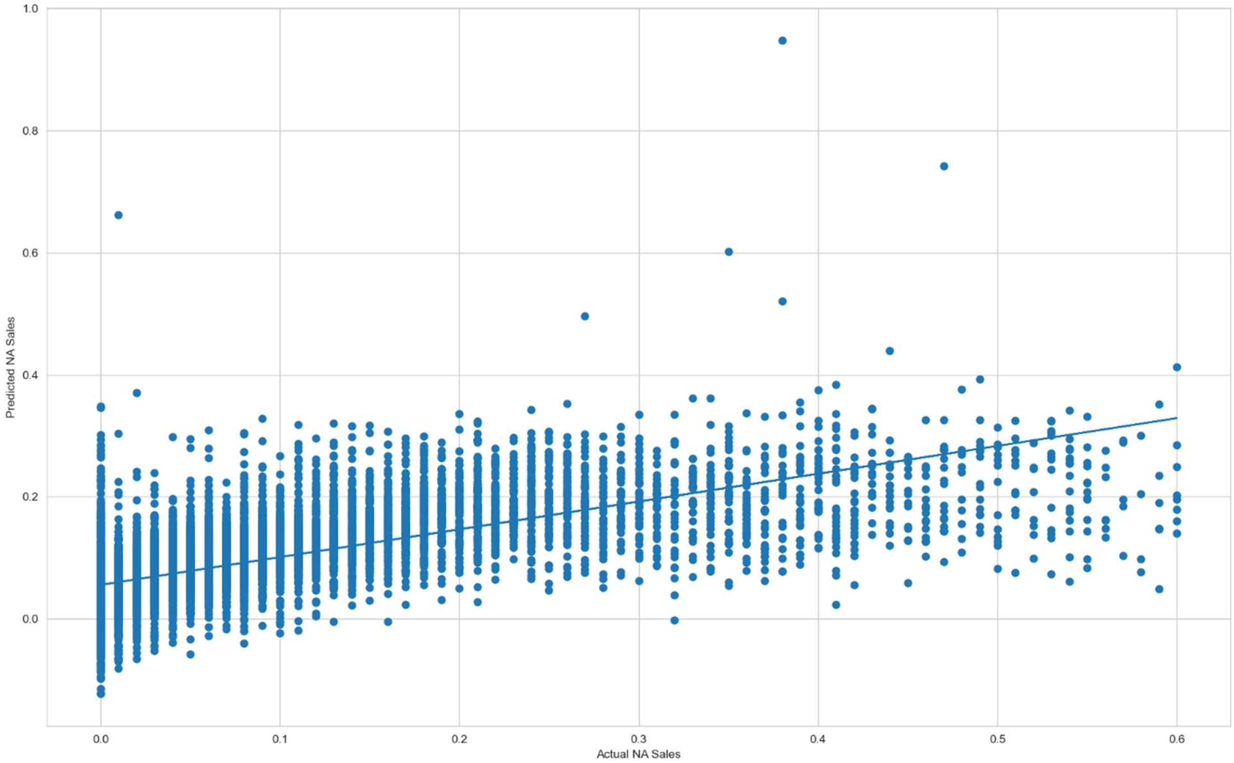
Milestone three consisted of the actual model selection and the evaluation of the model selected. The target for the model will be North American Sales, and the selected model for this is a linear regression model, specifically a multiple linear regression model. Multiple Linear regression was selected since I will be attempting to predict the North American Sales based on multiple features.

Based on all feedback provided, I adjusted the feature selections which resulted in additional features being dropped, the features dropped were “Global_Sales” and “Rank”. I left the remaining sales data from other regions since, in some cases, video games can be released in other markets before they are released in the North American market.

I used the sklearn library to split the final data set into test and training data with the “NA_Sales” columns being the target and all other columns being the features. I then created the instance of the linear regression model and saved it to the variable "model". I then fit the training data to the model and generated predictions. Figure 4 shows the resulting plot of the linear regression model on the training data.

Figure 4

Linear Regression Model of Training Data with NA Sales as the Target



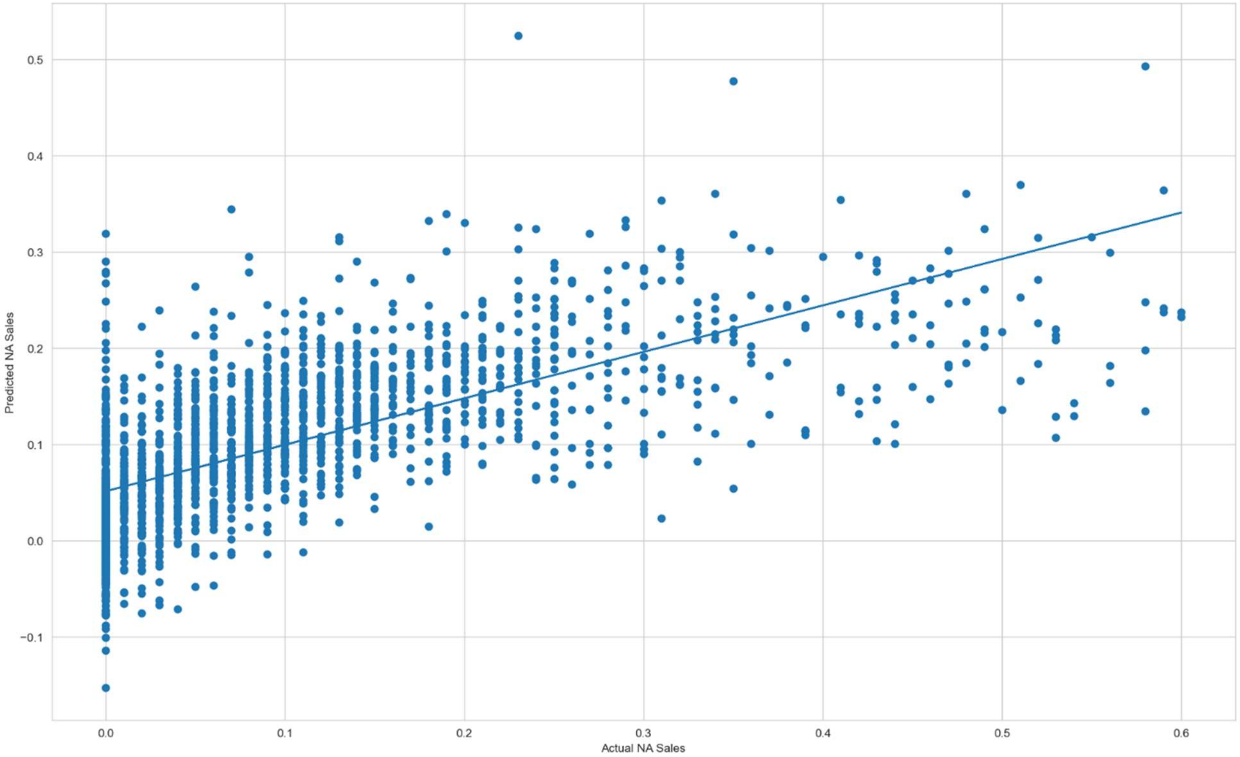
Note: Figure 4 illustrates the multiple linear regression model on the training data, with a line of best fit.

The next step to complete after fitting the model to the training data was to check the metrics to find if the model was a good fit. When reviewing the metrics the R-Squared and RMSE are the most common to use for a linear and multiple-linear regression model. R-Squared is the indicator the represents the variance between the target and features while the RMSE is the actual difference between the predicted values and the actual values. For the model fit to the training data, the r-squared values is 0.455 while the RMSE is 0.086. The results of the metrics appear to be conflicting; the r-squared value is lower than what is generally accepted as a good value, 0.7 or above. The RMSE of 0.086 indicates a low difference between actual and predicted, which is a sign of a good fit.

All that is left for the project is to run the model on the test set of data. The model was fitted to the test data and then the predictions were generated. The metrics for evaluation of the model were r-squared and RMSE again. The r-squared value of the test data was 0.471, a slight improvement over the training data, and the RMSE value returned was 0.084, again a slight improvement over the training data. Figure 5 shows the resulting model plotted. The

Figure 5

Multiple-Linear Regression Model Fit to Test Data



Note: Figure 5 illustrates the multiple-linear regression model fit to the test data set, with a line of best fit.

Conclusion

Taking in account the features used for the model, it tells us that the model would be a bad predictor of a North American video game sales. This model would not be ready for deployment in my opinion. I am not satisfied with results of the r-squared value returned by the model from the test data or training data. My recommendation for this model would be to incorporate more data and revisit the inclusion of the categorical features and explore other models, perhaps even re-phrasing or approaching the business question differently.

There is room for additional opportunities to be explored. The sales data pulled from the original data set could be better scaled. Currently it is a representation of percentage of sales, by scaling the sale data differently better result may have been achieved in the model. The categorical features could also be reviewed, and other methods explored to better format them for the multiple-linear regression model. There is also an opportunity to review other available models and find a better match. The multiple linear regression model seemed to make the most sense for the data, as some of the other models are focused on classification.