Project Phase 2 CNIT 570

Landan D. Perry
Purdue University Student
West Lafayette, IN
ldperry@purdue.edu

***Abstract*— This paper is a report on the findings of Predictive Analytics used on a dataset containing a wide variety of information on Video Games. Both linear regression and logistic regression were used. Outlier detection to find and consider statistical anomalies is also present in the findings. Using Linear regression, it was found that as the average game review scores increased so did the average number of sales. Then using logistic regression, I tried to determine if you could predict whether a game was multiplayer or not based off the sales and review score of a game. The results were inconclusive, and it doesn't appear that you can make accurate predictions with just those data points using logistic regression.**

***Index Terms*— Linear regression, logistic regression, R coding language, statistical anomalies.**

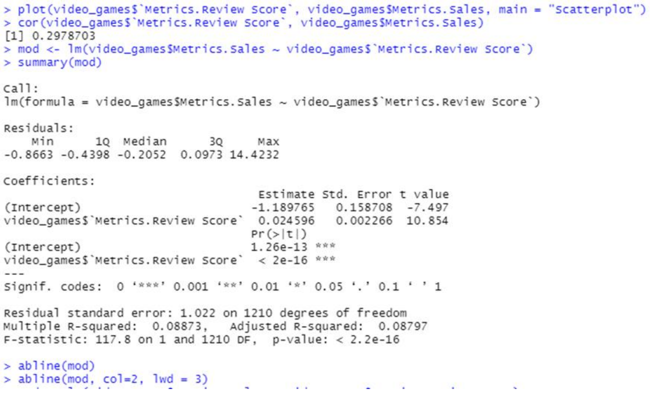
I. Introduction

This paper is a report and comparisons of Predictive analytics performed on a database from the "CORGIS Dataset Project". It is a collection of data that contains a variety of information on thousands of video games released between 2004 and 2010. The data points are things such as playtime, sales, game types, ratings, etc. The analysis primarily focuses on the data points of over all sales, which is the average sales in Millions of dollars, Review Score, which is the average review score of the given game, and max players, which is the number of players that are allowed to play the game at once. The software used to analyses the dataset was a combination of R and excel. Both were capable of performing the analysis that done but through use of multiple tools I was able to broaden my experience and see what was most useful for the project. It was a mixed bag when it came to results and this project and its analysis can be improved significantly as more experience is gained through practicing data analysis and the tools that are used in it. On top of the analysis performed this paper will also report on the comparison found between the analysis performed in Project Phase 2 and the analysis performed in Project Phase 1. Results were less clear in project phase 2 but regardless there are similarities and differences in the resulting information and our new insight of the dataset.

II. Findings

A. Linear Regression

R was used for the Linear Regression analysis of the data. A linear regression line was created using the average game review scores and the average game sales. The R code and model with the regression line can be seen in figures 1 and 2 below. The linear regression analysis showed clearly that as the average review score increased the average number if sakes did as well. This shows there is a direct correlation at least when averaged with games sales and how games are rated.



```
> plot(video_games$`Metrics.Review Score`, video_games$Metrics.Sales, main = "Scatterplot")
> cor(video_games$`Metrics.Review Score`, video_games$Metrics.Sales)
[1] 0.2978703
> mod <- lm(video_games$Metrics.Sales ~ video_games$`Metrics.Review Score`)
> summary(mod)

Call:
lm(formula = video_games$Metrics.Sales ~ video_games$`Metrics.Review Score`)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.8663 -0.4398 -0.2052  0.0973 14.4232 

Coefficients:
                                    Estimate Std. Error t value
(Intercept)                        -1.189765   0.158708  -7.497
video_games$`Metrics.Review Score`  0.024596   0.002266  10.854
                                   Pr(>|t|)    
(Intercept)                        1.26e-13 ***
video_games$`Metrics.Review Score`  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.022 on 1210 degrees of freedom
Multiple R-squared:  0.08873,	Adjusted R-squared:  0.08797 
F-statistic: 117.8 on 1 and 1210 DF,  p-value: < 2.2e-16

> abline(mod)
> abline(mod, col=2, lwd = 3)
```

Fig. 1 R code for Linear Regression

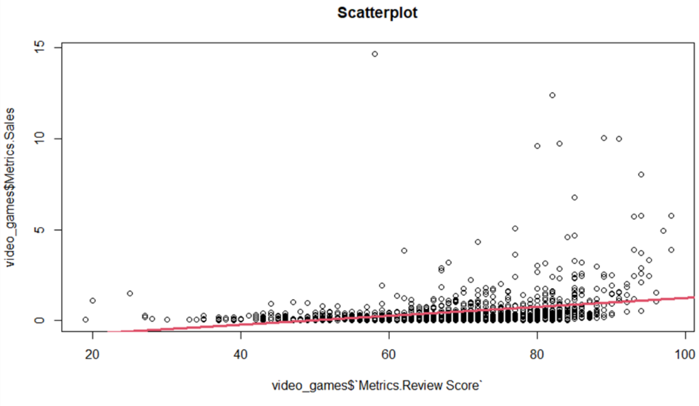


Fig. 2 Scatter Plot for Linear Regression

B. Logistic Regression

The Logistic Regression was done using Excel and an add on called the "Real Statistics Resource Pack". This pack creates a command in excel that automatically performs logistic regression analysis as long as you have the correct data types for it to be able to use. The linear regression was performed on the average review score, the average sales, and a value called multiplayer that recorded 1 for multiplayer games and a 0 for non-multiplayer games. The goal of the analysis was to predict whether a game was multiplayer or not based on its review score and sales. The graph of the results can be seen below in figure 3 as well as a table of predictions and accuracy in figure 4. The findings of the analysis is inconclusive and their seems to be data being reported that conflicts with each other. I am also skeptical of the results simply due to how accurate it reports to be. The only way I can really throw it off is if I make the cutoff 1 or 0 in their given directions. It seems that it is either extremely accurate or extremely inaccurate and either way more in-depth analysis needs to be done.

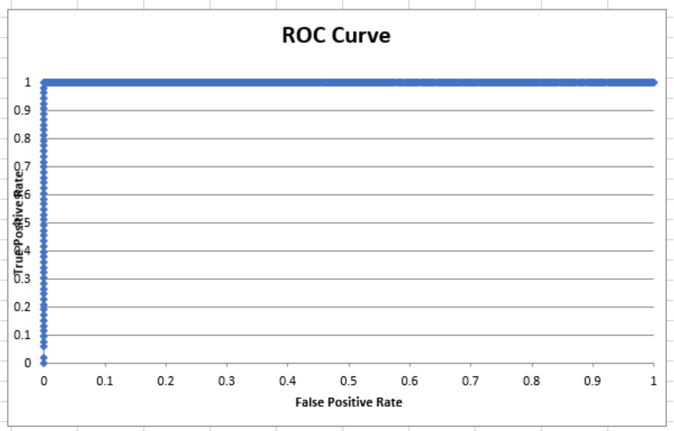
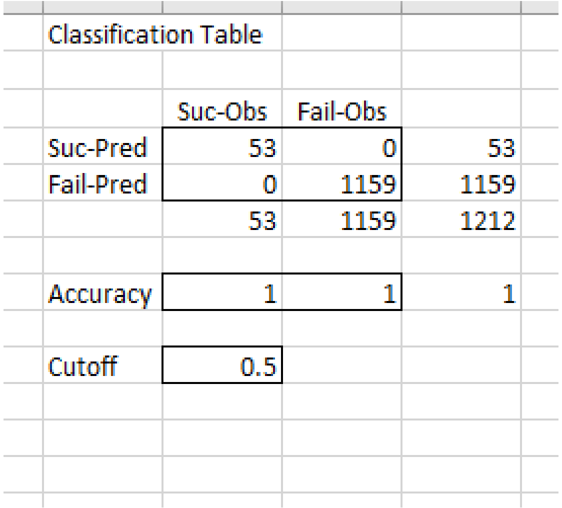


Fig. 3 Logistic Regression Graph



Classification Table			
	Suc-Obs	Fail-Obs	
Suc-Pred	53	0	53
Fail-Pred	0	1159	1159
	53	1159	1212
Accuracy	1	1	1
Cutoff	0.5		

Fig. 4 Logistic Regression Table

C. *Statistical Anomalies*

Statistical Anomalies were checked using R as well. This was done through outlier detection which allowed the data to be processed so that the analysis could be performed more accurately. The outlier detection was performed on Games sales specifically and it was found that values beyond 0.61 million dollars were considered outlier within the dataset. These outliers were then removed as seen in figures 5 and 6 below. The R code can be seen in figure 7.

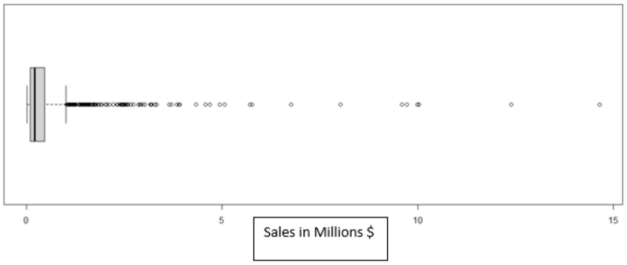


Fig. 5 Before Outliers Removed

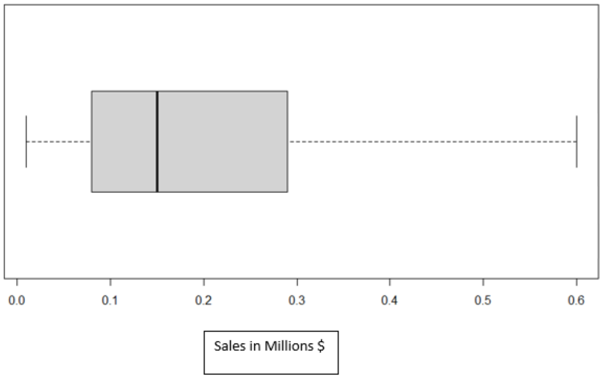
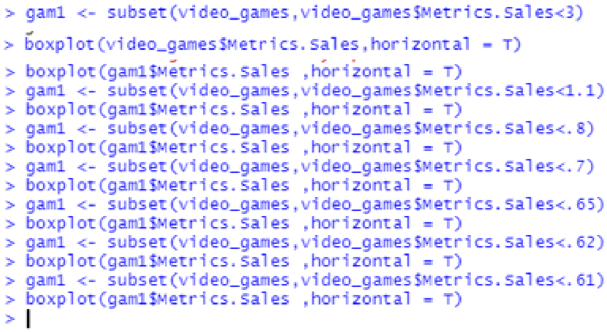


Fig. 6 Post Outliers Removed



```
> gam1 <- subset(video_games,video_games$Metrics.Sales<3)
> boxplot(video_games$Metrics.Sales,horizontal = T)
> boxplot(gam1$Metrics.Sales ,horizontal = T)
> gam1 <- subset(video_games,video_games$Metrics.Sales<1.1)
> boxplot(gam1$Metrics.Sales ,horizontal = T)
> gam1 <- subset(video_games,video_games$Metrics.Sales<.8)
> boxplot(gam1$Metrics.Sales ,horizontal = T)
> gam1 <- subset(video_games,video_games$Metrics.Sales<.7)
> boxplot(gam1$Metrics.Sales ,horizontal = T)
> gam1 <- subset(video_games,video_games$Metrics.Sales<.65)
> boxplot(gam1$Metrics.Sales ,horizontal = T)
> gam1 <- subset(video_games,video_games$Metrics.Sales<.62)
> boxplot(gam1$Metrics.Sales ,horizontal = T)
> gam1 <- subset(video_games,video_games$Metrics.Sales<.61)
> boxplot(gam1$Metrics.Sales ,horizontal = T)
> |
```

Fig. 7 R Code to Detect and Remove Outliers

III. COMPARISONS

A. *Statistical Anomalies Comparison*

For statistical anomalies the addition of logistic regression did not change any of the previous findings from phase 1 of the project. It did however show that games sales may not be the only data point that needs to be analyzed for anomalies. It is somewhat clear now that there are other outliers or data points that could be significantly influencing results. More data cleaning and processing on top of finding statistical anomalies would most likely have improved the results of Project Phase 2. If a phase 3 were to every come to reality it is clear more processing would need to be implemented to continue to improve on phase 2. This was not clear in phase 1.

B. *Linear Regression (Phase 1) and Logistic Regression (Phase 2) Compared*

When comparing these two phases the biggest difference if the confidence in the analysis results. During Phase 1 it was easy to understand and visualize the results. They made sense and could even easily be improved on using some processing and data clean up. The linear regression results were straight forward, and we could see exactly how game ratings and game sales coincided with each other. In the logistic regression analysis, the goal was to see if using game sales and game ratings logistic regression could predict if a game would be multiplayer or not. The analysis results seem to give us no true decision. The accuracy is recorded as being perfectly 1 to 1 so it seems to be that one can predict perfectly whether a game is multiplayer or not. However, when you look at the graph accompanying the report and the finer details it shows that many of the guessing were simply wrong. The errors may be coming from either data being skewed and their may be a need for improved processing. It may also be that when performing the analysis their may have been some user errors. This is far different from Phase 1 and is frustrating as this is the final report. Luckily during phase 2 I was able to confirm and test the Linear regression once again to confirm my phase 1 findings, so all was not lost. Another comparison that was made between Linear and Logistic regression was that what may seem like the most useful data points may not be the case. For phase 1 Sales and Rating were the perfect data points to run linear regression. However, when it came to logistic regression a Boolean data type with a value of 1 or 0 was needed. This required some changes to be made for my analysis. The additional use of the multiplayer data point allowed me to use that Boolean point of if the game was multiplayer and gain a level of insight into what encourages people to buy games other than their rating of how good it is. That added layer is another major different between phase 1 analysis and phase 2 analysis.

References

[1] G. O. Young, “Synthetic structure of industrial plastics (Book style with paper title and editor),” in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2] N/A, “Real Statistics Resource Pack”. Real Statistics Using Excel. https://www.real-statistics.com/free-download/real-statistics-resource-pack/ (accessed October 19, 2020).

[3] Data Analysis Videos, “Logistic Analysis Using Excel”. YouTube. https://www.youtube.com/watch?v=EKRjDurXau0 (accessed October 19, 2020)

[4] N/A, “Research And Citation/IEEE Style”. PurdueOWL. https://owl.purdue.edu/owl/research_and_citation/ieee_style/reference_list.html (accessed October 19, 2020).

[5] B, N. “Dealing with Outliers in R”. YouTube. https://www.youtube.com/watch?v=tOAJi9-qDm0 (accessed October 19, 2020).

[6] Bart, A & Cox, J, “Video Games CSV File” GitHub. https://corgis-edu.github.io/corgis/csv/video_games/ (accessed October 19, 2020).

[7] StatsLecturers, M, “Simple Linear Regression in R” YouTube. from https://www.youtube.com/watch?v=66z_MRwtFJM (accessed October 19, 2020).

AUTHORS BIO

Landan D. Perry is a current undergraduate level Purdue University student that is planning to graduate in December of 2020. He a computer and Information technology major. Outside of class work he is also involved in his fraternity as well as various intermural and other school/community-based activities.

He is currently looking for full time employment for after graduation. In the past he has worked for a variety of companies and in a variety of different positions. Most recently he worked from May 2020 to July 2020 as a software engineering intern for FP Complete remotely. In the fall of 2018, he also worked the Walt Disney Company as a college program intern in Orlando, Florida. His professional career began in 2015 as a senior in high school working for a company called Miller’s Consulting Group where he was a Design engineering intern.

Landan Perry is a member of Sigma Alpha Epsilon fraternity and served on the fraternity standards board as well as the scholarship committee. He has also received awards for his academic achievement in the classroom for the last 2 semesters earning semester honors within Purdue Polytechnic.