Introduction

For this project I decided to go over video game sales in different places around the world mostly out of curiosity but also to try and see if anything sticks out. More specifically I wanted to see if there are any correlations between the genre of game, what platform it’s on, who published it and where in the world it sold the best. There are a lot of people who are in the gaming space who adamantly believe that certain places in the world only like certain types of games. For example, many people around the world think that Americans prefer shooters for some very crass and also some good reasons as well. Many people also believe that Korean gamers prefer MMORPG’s mostly because so many MMORPGs are made in Korea. So, the long story short is I want to see if any of these beliefs have any credence or not.

Data

The data I used to try and find this is found on Kaggle [here.](https://www.kaggle.com/datasets/gregorut/videogamesales) This data set is pretty simple compared to many others I have used this semester but it also includes pretty much everything I need to do at least the bare minimum for predictive models. The dataset includes the rank of the game based on total global sales, next is the name of the game in question, then the platform that the game was released on, next is the year it came out, then the genre of the game such as shooters RPGs sports games etc. It also has information on who published it which there are a lot of publisher’s way more than I expected initially. Lastly the dataset also has information on sales in certain places of the world such as the US, Japan, Europe, Other and total global sales as well.

Methods

Ok there is a lot to talk about here so to start the methods I planned on using for this project were linear regression and decision trees. The regression would be done first to see if there really is any correlation with the genre, platform, publisher, and year with global sales in general. Then if that came back with good results, I would then move to decision trees to try and figure out the more nuanced causes for popularity for the sales in specific areas of the world. This did not go entirely to plan but for the most part it worked out code wise so let’s get into that.

The first thing I did was actually to just jump right into it and try and make a linear regression model without altering the data to much other than finding and removing rows with null values. That plan very quickly fell apart as most things did with this project because I did not know how to do linear regression with categorical data. Thinking back on project 3 I just got rid of all non-numeric data instead of converting it so first I had to pivot and spend hours and hours figuring out how to do that properly so this project could even start. Once I did that I started messing with the code to try and convert all the categorical data into something that the linear regression model could use and an idea hit me. Categorical data is just a pointer that we can read right? Well, why not just change it to a pointer the computer can read and just label them for us primitive humans. Which is what I did with this code here:

A screen shot of a computer screen

Description automatically generated

I then realized that all the other data were floats so I had to parse this new integer data to float for consistency which I did with this code here:

A screen shot of a computer

Description automatically generated

Its very simple code but it actually took me a while to figure out how to do this because it kept throwing an error when I tried to use another method we used in class to do it so I used this instead. I did this for all the categorical data types the only exception to this was the publisher column. This is because as you could see above I hard coded each entry manually which works fine if you only have a small amount of data to work with or if the number on unique values is also limited at was the case for platform and genre. In the case of publishers, however, there were a ludicrous amount of them over 100 if I remember correctly so I did something else instead. Instead of hard coding all of them I decided to only use the publishers that had over 100 unique entries in the data set which came out to 29 of them. I did this with this code here:

A screenshot of a computer program

Description automatically generated

This code let me see all the unique data in the column and how many times they occurred. With this information I then hard coded the top entries and then parsed it to a float like the rest of the columns. This was not easy to do however because I only changed some data in the column not all of it so I had to have code that would get rid of all the other data that was not changed which I did with this code here:

A screen shot of a computer

Description automatically generated

Next I moved to making the actual regression model after dropping more than a few columns such as rank, name and every sales category except the global sales column. The code for the regression model is here:

A screenshot of a computer

Description automatically generated

A screen shot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

The code above is just about everything that I did for the regression model other than some basic scatterplots that did not tell me much honestly so I will omit them. I will speak about the actual results of the model later after going over the decision trees as well.

The trees.

Next up are the decision trees which while I was only planning on doing a couple of them I ended up doing four of them but the process for each was very similar to one another just with different variables going into the tree and what they were trying to predict. Starting with trying to predict the genre of the game based on the sales data. For the most part I did the same steps as the regression model only I didn’t need to convert any data only drop the necessary columns and the first entry in the data set because Wii sports was an extreme outlier. Otherwise, the code that I used for all the trees was as follows:

A black screen with orange text

Description automatically generated

A screenshot of a computer program

Description automatically generated

A black screen with white text

Description automatically generated

All things considered the code for each tree is very simple as all the others follow this same routine generally. The only differences are that the second one just tests publisher instead of genre and the third one tests platform. The fourth one tests platform again, but it uses all the data in the data set instead of only sales data. All things considered the trees were simple to build but I don’t think any of the pair plots that I made were any sort of helpful at all.

**Evaluation**

The models that were produced for this project were less than accurate in most cases but there is one exception to this rule. That being the regression model which I would argue was fairly successful. Which was surprising to me because I thought it would be the worst model for the questions I was asking and because of all the extra work I had to do to even get it to function in the first place. Here are the results for the linear regression model:

A screenshot of a computer program

Description automatically generated

The main metric I want to look at here is the MSE value as the other two are less relevant to the questions that I asked. The MSE value is 4.5 relatively which translates to a margin of error of around 4.5 million global copies sold for each game. Now for some games that could be really bad but considering the dataset went all the way up to 80 million copies sold for Wii sports I think 4.5 million is not that bad all things considered. It also helped answer one of my questions by proving that there is at least some correlation between all the categorical data and total sales for each game.

The results from the decision trees on the other hand were very bad so much so that I didn’t bother to even make visualizations of them as the information they would show is so bad that it would be just as bad as lying to your face. The results for each tree are as follows:

A computer screen shot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated

A screen shot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

As you can see the first two decision trees were effectively useless as the accuracy of both of them is below 20 percent, which is downright awful. No matter what I did to try and make these scores better nothing helped. In fact, some of what I did made them even worse. The only two trees that were not catastrophic failures were the last two which were still failures just not as bad. The last two were both predicting the platform the game was on. The scores were 44 and 50 respectively which does prove that the platform the game is played on seems to have much more of an impact than the other categories. As for the questions from the introduction I was not able to fully answer them, but I did get some information but not everything I wanted. I found that there is at least some correlation between all the categorical data and the global sale count of the game based on the linear regression. I also found that the individual sale data per country is most heavily related to the platform the game is on rather than anything else.

Story/Conclusion

When I started working on this project, I initially wanted to find the correlation between Genre Platform and Publisher and how they affect sales in certain places in the world. I also wanted to use linear regression to help with this specifically because I did not fully understand it earlier in the semester, so I wanted to fix that. While trying to use linear regression again I had to figure out how to properly change data to use it which solved the problem of understanding linear regression. Like I said before I was not fully able to answer all my initial questions, but I did find some answers as stated before and if I had more time then I would most likely use decision forests and logistic regression, if possible, to try and find more information and hopefully answer the other questions. So, to summarize I used logistic regression and decision trees to try and find answers to the questions posed at the beginning. The decision trees were less than helpful, and I tried many different things to try and fix that with no luck. The information I did gather points to Platform being the most important aspect of regional sales which makes sense thinking about it. Xbox is a US platform so it would make sense if it sold games better in the US and Nintendo is mostly a Japanese company, so it sells games better in Japan etc. So that makes sense at the very least, but I was surprised that Genre had so little correlation. Publisher also had very little correlation which also surprised me because you would think a primarily Japanese publisher would correlate with sales in Japan but that’s just me. It was interesting to see the results, but they were unexpected from what I initially thought going into the project.

Impact section

The impact of this project overall I believe is not very big. The main thing I could see coming from this project is the misinformation of the decision trees being as bad as they are currently. The trees show little to no correlation with both Publisher and Genre, which I still believe too not be entirely accurate to be honest. The other possible impact would be for people to take it the right way and realize that you can make any type of game as long as it’s fun and with a little bit of luck and it can blow up to be a hit all over the place. Otherwise, the project as is, is not complete and what can be gleaned from it is rather negligible at first glance, which would lead to a smaller impact, I think. As a side note a funnier impact may be that Microsoft and Sony may realize that packaging a game with the console my lead to massive “sales” numbers like Nintendo did with Wii sports because good lord Wii sports was a massive outlier considering it sold 80 million copies and second place was at around 40 million. Which was also a Nintendo product take notes people they are clearly doing something right.

References:

<https://stackoverflow.com/questions/42850280/pandas-drop-non-integer-data>

<https://www.kaggle.com/datasets/gregorut/videogamesales>

code: