Region Based Video Game Sales Prediction with Regression

Ediz Tekok, Furkan Kılıçaslan

Abstract

In this experiment, decision tree regressor and k-neighbors regressor were used to determine possible future sales potential in a certain region. Using a set of features (rank, genre, platform, year and publisher) a predictive model has been built to estimate the value of an outcome of interest, which is either Europe, Japan, America or other region sales. We achieved better results with decision tree regressor with r2 values between 0.54-0.65, whereas k-neighbors regressor performed with r2 values between 0.45 - 0.55. In conclusion decision tree regressors perform better than k-neighbors regressor under these circumstances for predicting future video game sales.

Keywords: decision tree regressor, k-neighbors regressor, data mining, sales prediction, video games

1. Introduction

Video Games are increasingly getting more and more popular every day and more developers are getting into game development.. However, there is always this issue that most game developers run into: how are they going to make a profit from the games release? More importantly, how are they going to keep profiting annually? This report intends to tackle this problem by looking at a video game sales dataset that contains information of video games that have sold more than 100,000 copies since 1980 to 2015[1].

This dataset consists of 16,600 records with information on ranking of overall sales, the games name, the platform of the games release (such as XBox, Playstation et cetera), the year of the game's release, genre of the game, publisher of the game, sales in North America, Europe, Japan and the rest of the world and the total worldwide sales. There were some game sales information whose year was not included, so those samples were imputed by removal from the dataset. After this preprocessing step two different regression methods were tested to predict region based video game sales.

2. Experiment Goal

In this experiment the video game sales dataset provided by kaggle [1] is used to build a predictive regression model with the aim of finding insight over regional future sales of a video game based on 'Rank', 'Genre', 'Platform', 'Year', 'Publisher' to estimate the value of an outcome of interest. We pick a region to predict the outcome and use the other regions to compare to the chosen region. This can provide insight to game publishers in various ways such as which market to target for what genre of games or which platforms are more lucrative in which markets et cetera. These insights are crucial to companies for critical decision making.

Our aim in this project is to utilize machine learning and data mining methods to prevent possible money and time losses for companies publishing games. The results of this experiment are not decisive nor it should be seen as facts. This experiment is performed from data obtained from the past which might not be statistically significant for the future and it might not capture all the elements of entrepreneurial success.

3. Data Preparation

We first load the data on our program using pandas, a library for analyzing and manipulating data using data structures[2]. The dataset input is in comma separated values (CSV) format so we parse this data using pandas. Below is an example dataset we have using first and last five datas inside:

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37
16593	16596	Woody Woodpecker in Crazy Castle 5	GBA	2002.0	Platform	Kemco	0.01	0.00	0.00	0.00	0.01
16594	16597	Men in Black II: Alien Escape	GC	2003.0	Shooter	Infogrames	0.01	0.00	0.00	0.00	0.01
16595	16598	SCORE International Baja 1000: The Official Game	PS2	2008.0	Racing	Activision	0.00	0.00	0.00	0.00	0.01
16596	16599	Know How 2	DS	2010.0	Puzzle	7G//AMES	0.00	0.01	0.00	0.00	0.01
16597	16600	Spirits & Spells	GBA	2003.0	Platform	Wanadoo	0.01	0.00	0.00	0.00	0.01

Figure 3.1) Pandas DataFrame for Video Game Sales Data [2]

After loading the data we have imputed the rows with any missing values by removal. The different sales figures can be represented with a single dimension which is what we obtained using “Principal Component Analysis”. For this purpose we have implemented the PCA function of Scikit-learn[3] library. Working with non-numerical data can be problematic for most machine learning algorithms. For this reason we have converted all non-numericals to numerics. Lastly the data is split into training and test sets with an 80/20 split using “train_test_split” function of Scikit-learn [3].

4. **Exploratory Data Analysis**

Here we will be plotting some graphs to uncover some trends and information about the data we have. Firstly we are looking at the total game sales in $ millions per year. The trend shows that the early 2000s were the booming era of game industry and it has been in decline since 2008. It is interesting that this coincides with the financial crisis of United States in 2008. This may be due to people spending more time with video games as lack of employment increased.

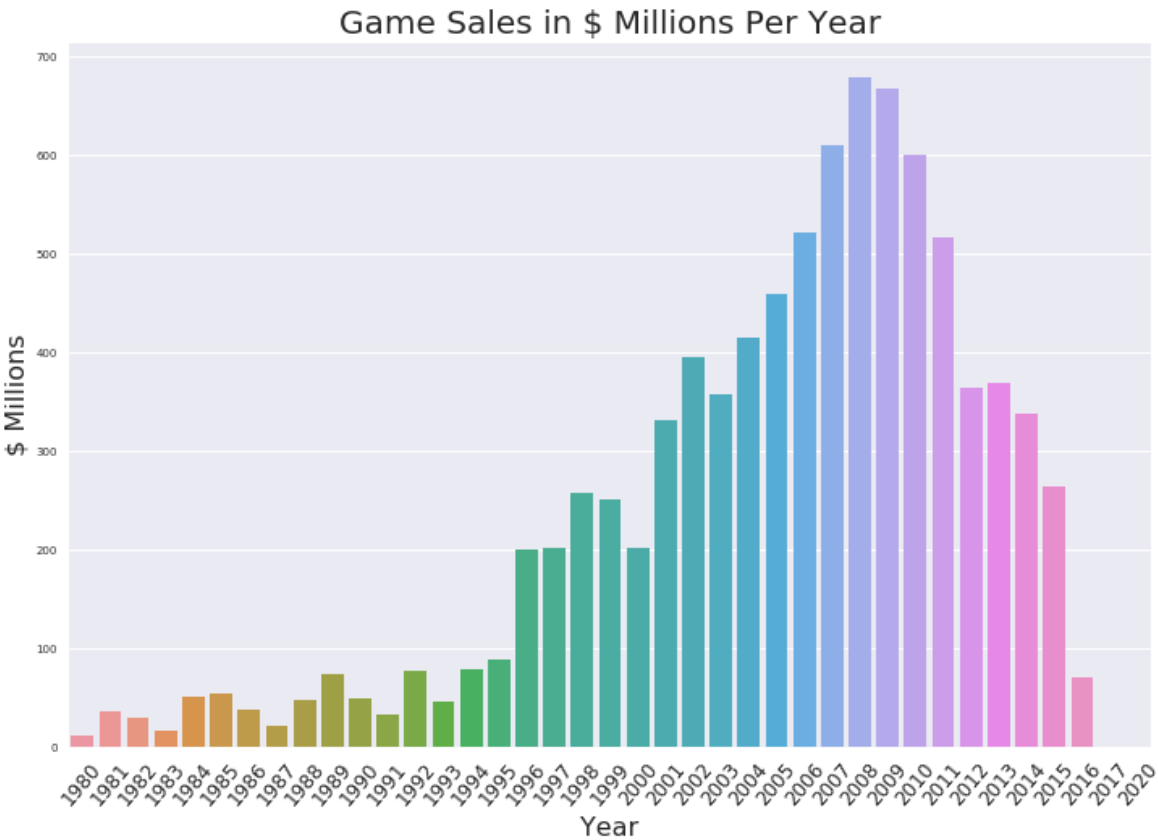

Figure 4.1) Game Sales in $ Millions Per Year [5]

Another figure we can look at is the percentage comparison of video game sales in regards to their genre and region. This heatmap uncovers interesting information about the consumer behaviour in different regions. For example shooter games dominate North American market against Japanese market while the sales are on a very similar level for role-playing games. Also it is seen that the European market is robust to genre differences and sales are predictable.

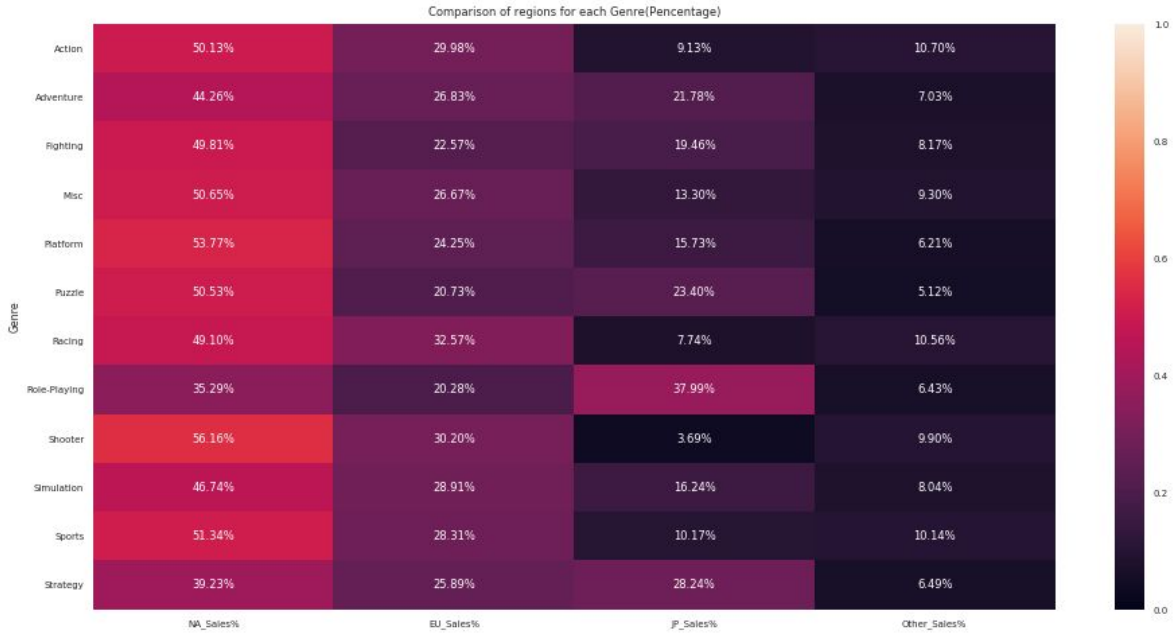

Figure 4.2) Game Sales in $ Millions Per Year [5]

Another interesting figure we can look at is the total revenue per region graph. North American market have had the lion share of video game sales between 2000 and 2011 while the Japanese market have been following a stable trend. This supports our earlier claim about American video game market booming due to the financial crisis. Although European region has a similar sales trend with North America which might be due to the financial crisis as well as other reasons.

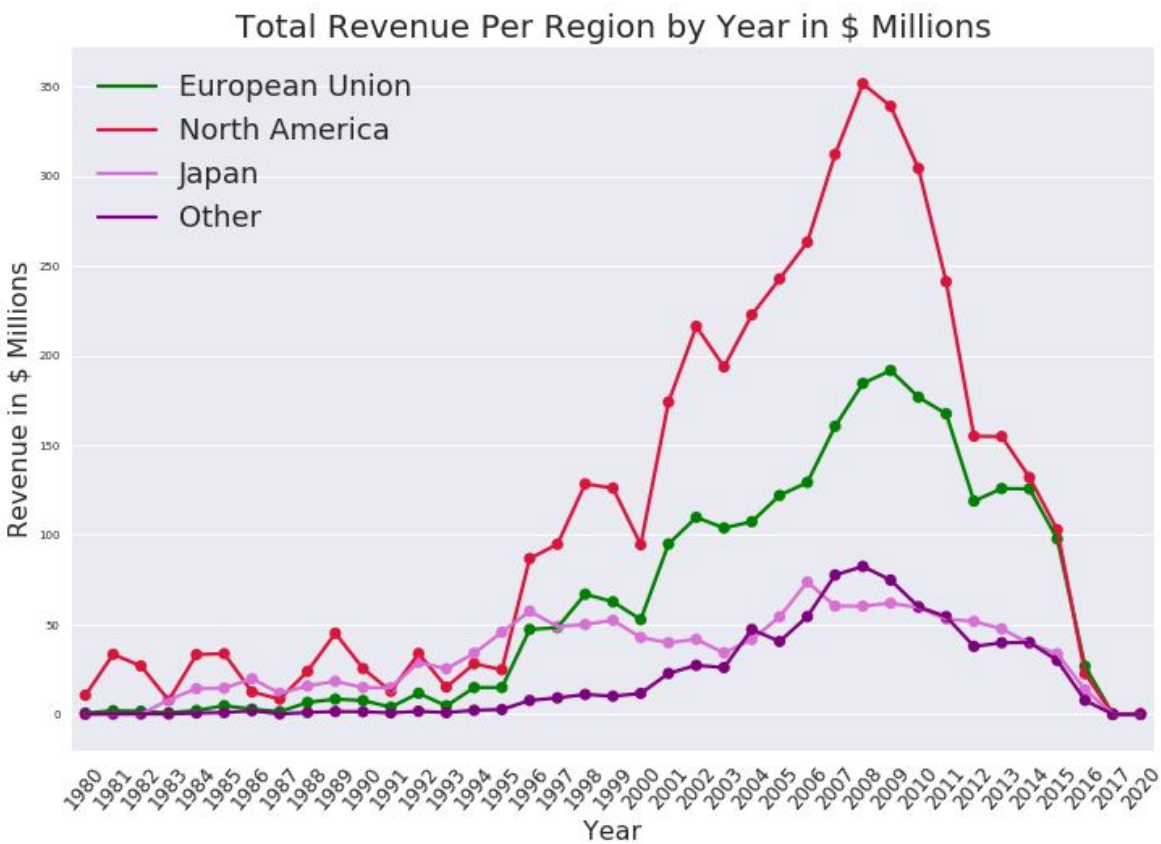

Figure 4.3) Total Revenue Per Region by Year in $ Millions [5]

5. **Model Selection**

For this regression problem we believed a decision tree regression or a K-Neighbors regression algorithm would fit the data well. So we tested both on the data using R-squared score as our evaluation metric. We have implemented these two regression algorithms for NA_Sales, EU_Sales, JP_Sales and Other_Sales individually. These algorithms are also utilized from the Scikit-learn [3] library. Below are the R-squared scores for both algorithms and the scatter plots for the Decision Tree Regression predictions for each region :

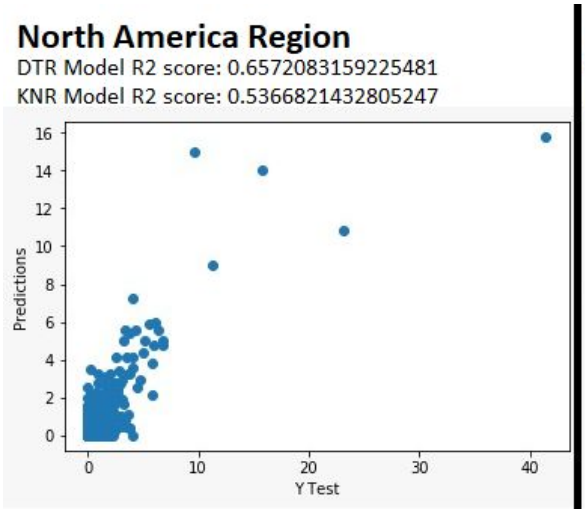

Figure 5.1) Scatter plot of error distribution for NA

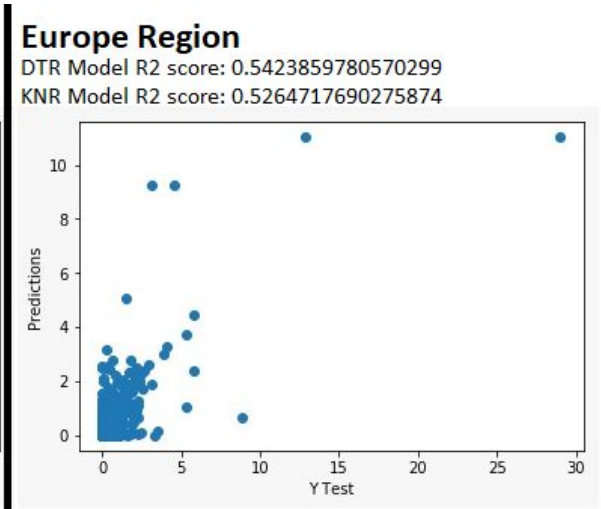

Figure 5.2) Scatter plot of error distribution for EU

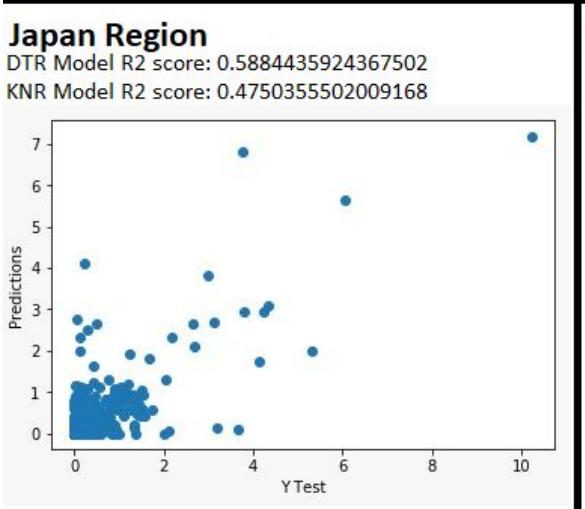

Figure 5.3) Scatter plot of error distribution for JP

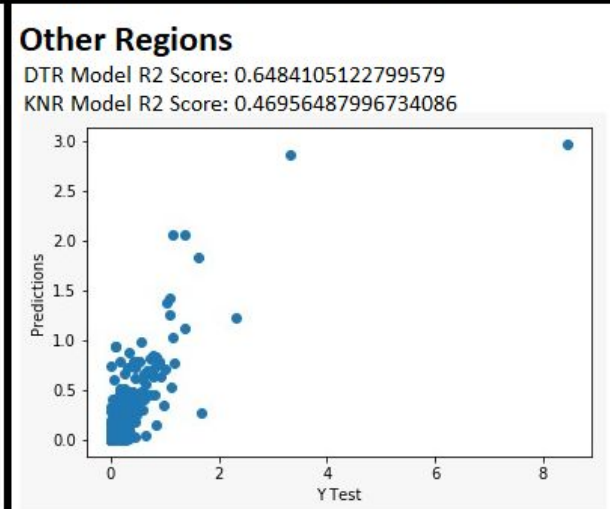

Figure 5.4) Scatter plot of error distribution for other regions

6. Conclusion

In conclusion, having such a vast amount of data about video game sales from 1980 to 2016 enabled us to capture some trends in these sales based on market and genre. Using “Decision Tree Regression” we were able to learn some of the underlying structure about a markets’ sales potential. This can help video game companies decide on what market to target for what genre of games.

This study can further be expanded by using different optimization methods such as scaling. Also implementing various other machine learning algorithms such as linear regression, deep neural networks etc. can produce different, potentially better results for sales forecasting.

References

1) Gregory Smith,Video Game Sales Data, Retrieved from:https://www.kaggle.com/gregorut/videogamesalesü
2) McKinney, Wes (2018, May 1) pandas library. Retrieved from https://pandas.pydata.org/about.html
3) Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
4) Jones E, Oliphant E, Peterson P, et al. SciPy: Open Source Scientific Tools for Python, 2001-, http://www.scipy.org/
5) Hunter, J. D.,Matplotlib: A 2D graphics environment,10.1109/MCSE.2007.55,2007