Project Phase 4: Data Mining
Video Game Sales

University of Ottawa
School of Electrical Engineering and Computer Science
CSI4142 Fundamentals of Data Science
Professor Yazan Otoum

uOttawa

Yara Elmasry 300157512
Eric Germond 0300164005
Jonathan Treuil 300178100
Sam Mulvey 300201795
Group#46

Part A. Data summarization, data preprocessing and feature selections

One-Page Summary :

In preparing the dataset for analysis, several preprocessing steps were undertaken over our three datasets (Video Game Sales, Video Game Usage, Backlogs) to ensure data quality and compatibility with modeling techniques. Firstly, we identified and handled missing values across all features. Techniques such as imputation and removal of rows and columns with missing values we applied. In addition to that, we detected and removed duplicated records to avoid redundancy and ensure the integrity of the dataset.

Secondly, we derived new features from existing ones to capture additional insights and improve model performance. This involved operations such as one-hot encoding for categorical variables. We then scaled numerical features to a standard range to prevent bias towards certain features during model training. Techniques such as Min-Max scaling were applied on the distribution of the data. Moreover, we converted categorical variables into numerical format using one-hot encoding to enable modeling algorithms to interpret the data correctly.

Publisher_Zushi Games	Publisher_bitComposer Games	Publisher_dramatic create	Publisher_fonfun	Publisher_iWin	Publisher_id Software	Publisher_
False	False	False	False	False	False	
False	False	False	False	False	False	
False	False	False	False	False	False	
False	False	False	False	False	False	
False	False	False	False	False	False	

Then, we ensured that numerical features are on a similar scale to prevent certain features from dominating other features. Normalization techniques such as Min-Max scaling were employed.

NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Total_Sales	Average_Sales_Game	Sales_Per_Platform	Average_Sales_Per_Year
1.000000	1.000000	0.368885	0.800378	1.000000	1.000000	0.500000	1.000000	1.000000
0.700892	0.123363	0.666341	0.072848	0.486281	0.486343	0.500000	0.486343	0.486281
0.382020	0.443832	0.370841	0.313150	0.432854	0.433043	0.500140	0.433043	0.432854
0.379610	0.379394	0.320939	0.280038	0.398767	0.398840	0.500000	0.398840	0.398767
0.271632	0.306340	1.000000	0.094607	0.379064	0.379260	0.500159	0.379260	0.379064

Part B. Classification (Supervised Learning)

Summary Table

	Decision Tree	Gradient Boosting	Random Forest
Accuracy	0.1905492482356551	0.2396440625958883	0.209266646210494
Precision	0.1780954004330541	0.2295515874663211	0.1924347430621395
Recall	0.1905492482356551	0.2396440625958883	0.209266646210494

200 - 300 Word Summary:

After evaluating the Decision Tree, Gradient Boosting, and Random Forest Algorithms on our dataset, several actionable insights were uncovered by us. Firstly, when comparing the accuracy metrics across the 3 algorithms, it's evident that Gradient Boosting performed the best with an accuracy score of approximately 0.24, followed by Random Forest with around 0.21, and Decision Tree with lowest accuracy around 0.19. This suggests that Gradient Boosting and Random Forest are more effective in making accurate predictions compared to Decision Trees.

Secondly, analyzing the precision scores, which represent the ratio of correctly predicted positive observations to the total predicted positives, we found that Gradient Boosting has the highest precision approximately 0.23, followed by Random Forest with about 0.19, and Decision Tree with the lowest around 0.18. This indicated that Gradient Boosting has the best ability to make precise predictions.

Lastly, examining the recall scores, which denote the ratio of correctly predicted positive observations to all actual positives, we observed similar trends to accuracy and precision metrics. Gradient Boosting achieved the highest recall score about 0.24, followed by Random Forest with approximately 0.21, and Decision Tree with the lowest recall of around 0.19. This implies that Gradient Boosting is most effective in capturing true positive cases.

In summary, based on our analysis, Gradient Boosting appears to be the most promising algorithm for our Video Games Dataset (Main Dataset), as it consistently outperforms Decision Tree and Random Forest in terms of accuracy, precision, and recall. These insights provide valuable guidance for selecting the most suitable machine learning algorithm for predictive modeling tasks.

Part C. Detecting Outliers

200-300 Word Summary on Outliers

To identify outliers in our vgsales dataset, we employed the One-Class SVM algorithm, which is effective in detecting anomalies in data. After applying the algorithm, we obtained a list of outlier indices. Upon examining these outliers, we observed that they exhibit unusual sales patterns compared to the majority of the data points. For instance, some outliers represent games with exceptionally high sales figures across different regions, while others correspond to games with extremely low sales or even zero sales. These outliers could offer valuable insights into various aspects of the video game industry. For example, games with unusually high sales could be indicative of popular titles or releases. On the other hand, games with extremely low sales may highlight instances of failure. Understanding such characteristics of these outliers can provide valuable insights for stakeholders in the gaming industry, including game developers, publishers, and marketers. Moreover, analyzing outliers can help identify potential data quality issues or errors in the dataset. It allows for a deeper understanding of the distribution and characteristics of the data, facilitating better decision making and strategy.

Important Links:

https://colab.research.google.com/drive/1beAZTT1jGoauABGCt3C_HAplYg8bkqom?usp=sharing

https://github.com/Yaraelmasry/CSI4142_Project_Group46