Analysing and Predicting Xbox Game Sales using Machine Learning

Josh Cox

University Of Exeter

2023

Research Question

- Can we use machine learning to predict the sales of an Xbox game in different parts of the world?
- This is useful for upcoming games:
  - Change the level of marketing in different areas
  - Different pricing of the game depending on the area
  - Some games might only release in certain areas

Dataset

- 613 Xbox games from 2013 to 2020 [1]
- Columns we will be using:
  - Genre - The genre of the game
  - Publisher - Publisher that released the game
  - North America - Sales in North America
  - Europe - Sales in Europe
  - Japan - Sales in Japan
  - Rest of World - Sales in other areas

	Pos	Game	Year	Genre	Publisher	North America	Europe	Japan	Rest of World	Global
0	1	Grand Theft Auto V	2014.0	Action	Rockstar Games	4.70	3.25	0.01	0.76	8.72
1	2	Call of Duty: Black Ops 3	2015.0	Shooter	Activision	4.63	2.04	0.02	0.68	7.37
2	3	Call of Duty: WWII	2017.0	Shooter	Activision	3.75	1.91	0.00	0.57	6.23
3	4	Red Dead Redemption 2	2018.0	Action-Adventure	Rockstar Games	3.76	1.47	0.00	0.54	5.77
4	5	MineCraft	2014.0	Misc	Microsoft Studios	3.23	1.71	0.00	0.49	5.43

Figure: Xbox game dataset

Preprocessing

- Removed unwanted columns
  - Game name, Year of Release, Global Sales, Position
- Formatted data
  - Changed column names and converted to lowercase
  - Encoded categorical data
  - Scaled data

	na_sales	eu_sales	jp_sales	other_sales	genre_n	publisher_n
2	3.75	1.91	0.00	0.57	11	5
3	3.76	1.47	0.00	0.54	1	69
4	3.23	1.71	0.00	0.49	5	50
5	3.25	1.49	0.01	0.48	11	5
6	3.37	1.26	0.02	0.48	11	24

Figure: Preprocessed dataset

Check for Outliers

- Checked for outliers
  - Using Interquartile Range (IQR) isn't applicable
  - Used visualisation of data

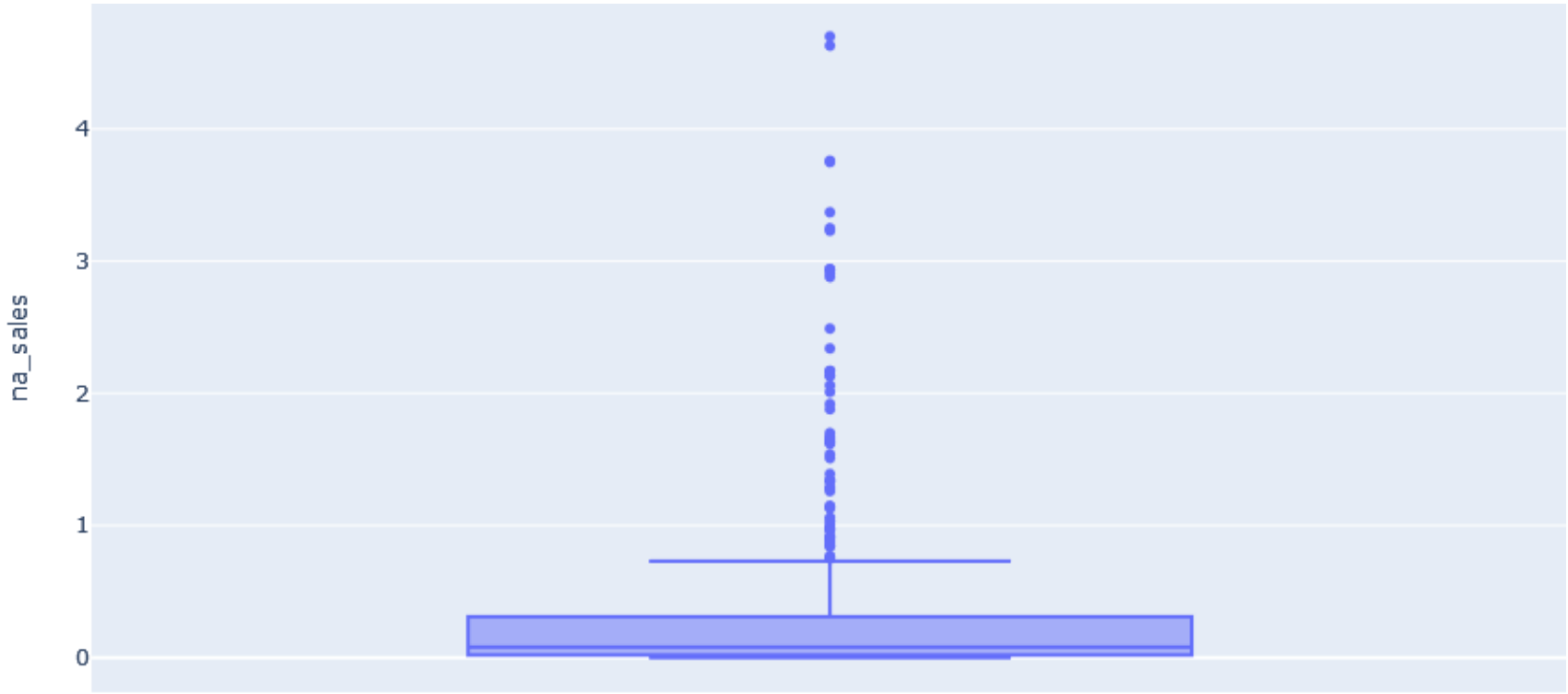

Figure: Box plot of data

Approach

- Train regression models to predict sales in each area based on:
  - Genre of the game
  - Publisher of the game
- Cluster the data into groups based on sales in each area
  - For example one group of games did better in Europe than North America
- Train a classifier to organise data points into created clusters

Regression

- Trained two regressors:
  - Random Forest (RF)
  - Gradient Boosting (GB)
- Results:
  - GB outperformed RF for Europe
  - RF outperformed GB for North America and Other
- Decide to drop Japan as not enough data points

	Random Forest		Gradient Boosting	
	R^2 Score	MSE	R^2 Score	MSE
North America	0.495	0.226	0.479	0.233
Europe	0.339	0.092	0.343	0.091
Japan	-0.110	3.141	-0.161	3.284
Other	0.476	0.005	0.461	0.005

Table: Results for RF and GB Regressors

Clustering

- Tested 3 clustering algorithms:
  - K-Means
  - Density-based spatial clustering of applications with noise (DBSCAN)
  - Gaussian Mixture Model (GMM)

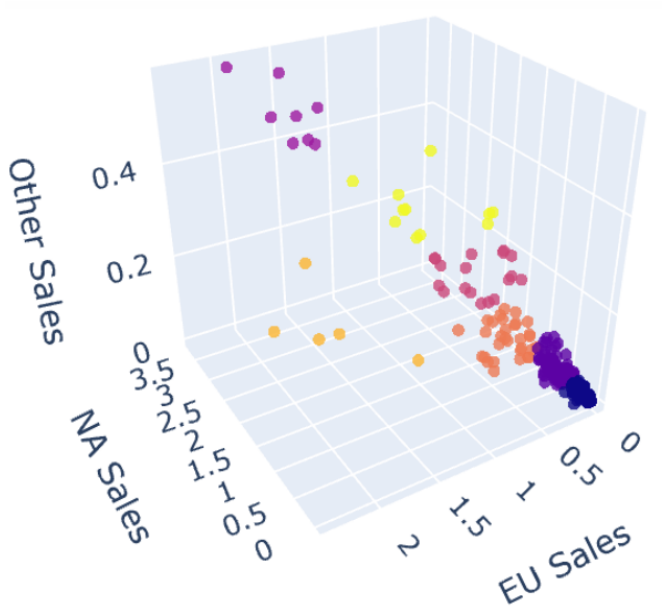

Figure: GMM Clusters

- Used elbow plot and average silhouette scores to find $K = 7$
- Silhouette Scores:
  - K-Means: 0.961
  - DBSCAN: 0.937
  - GMM: 0.961

Dimensionality Reduction

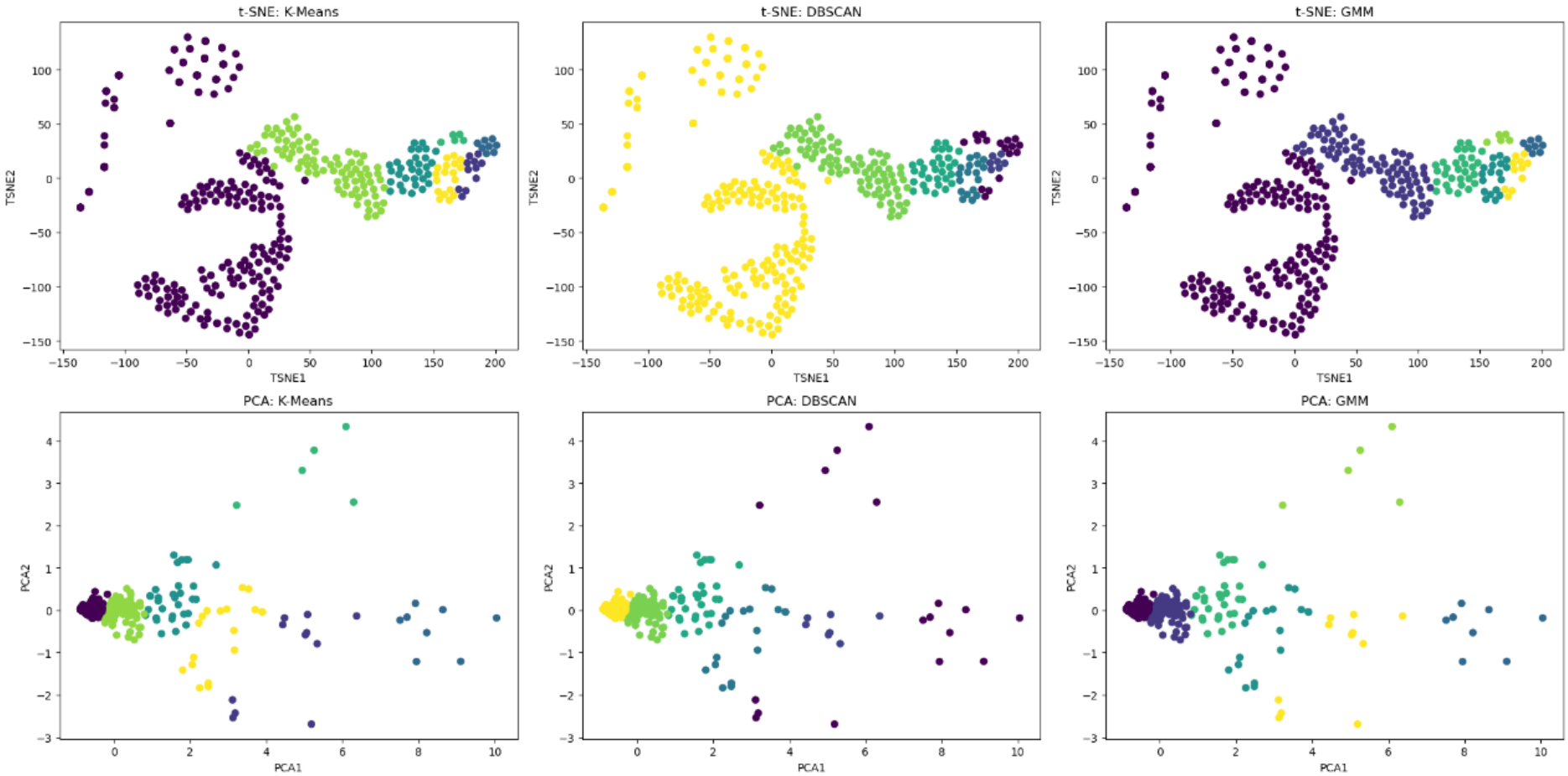

Figure: t-SNE and PCA Scatter Plots

Classification

- Tested RF and GB
- F1 Score:
  - RF: 0.973
  - GB: 0.957
- Mean Squared Error:
  - RF: 0.119
  - GB: 0.205
- Skewed results due to number of 0 values, but still accurate

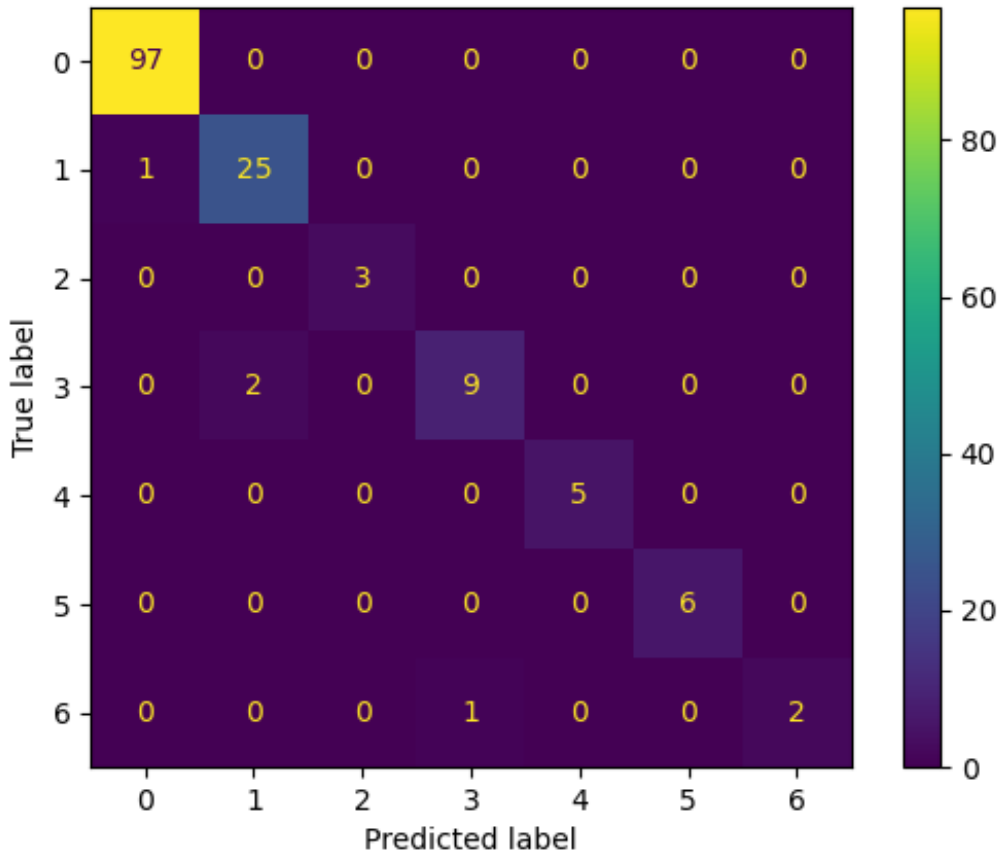

Figure: RF Classifier Confusion Matrix

Conclusion

- Predictor
  - Accuracy was lacking in regressors
  - Likely due to only using genre and publisher
- Clustering
  - Resulted in accurate groupings
  - Useful for organising and visualising sales of games
- Classifier
  - Skewed due to data
  - Provided accurate classifying of data
  - Useful for visualising sales of a game

Further Improvements

- Using a more complex machine learning model
  - For example tuning a Neural Network
  - Testing different architectures (number of hidden layers etc)
- Using more complex features
  - Genre and publisher don't define a game's sales
  - Considering factors such as graphics, sounds, online capability etc

References

[1] S. TWR, "Video Games Sales Dataset," *Kaggle*, Accessed: Nov. 16, 2023. [Online]. Available: https://www.kaggle.com/datasets/sidtwr/videogames-sales-dataset