Video Game Sales Analysis

A presentation by:
Khan, Anumta
Mardosas, Tiara
Trinh, Lan

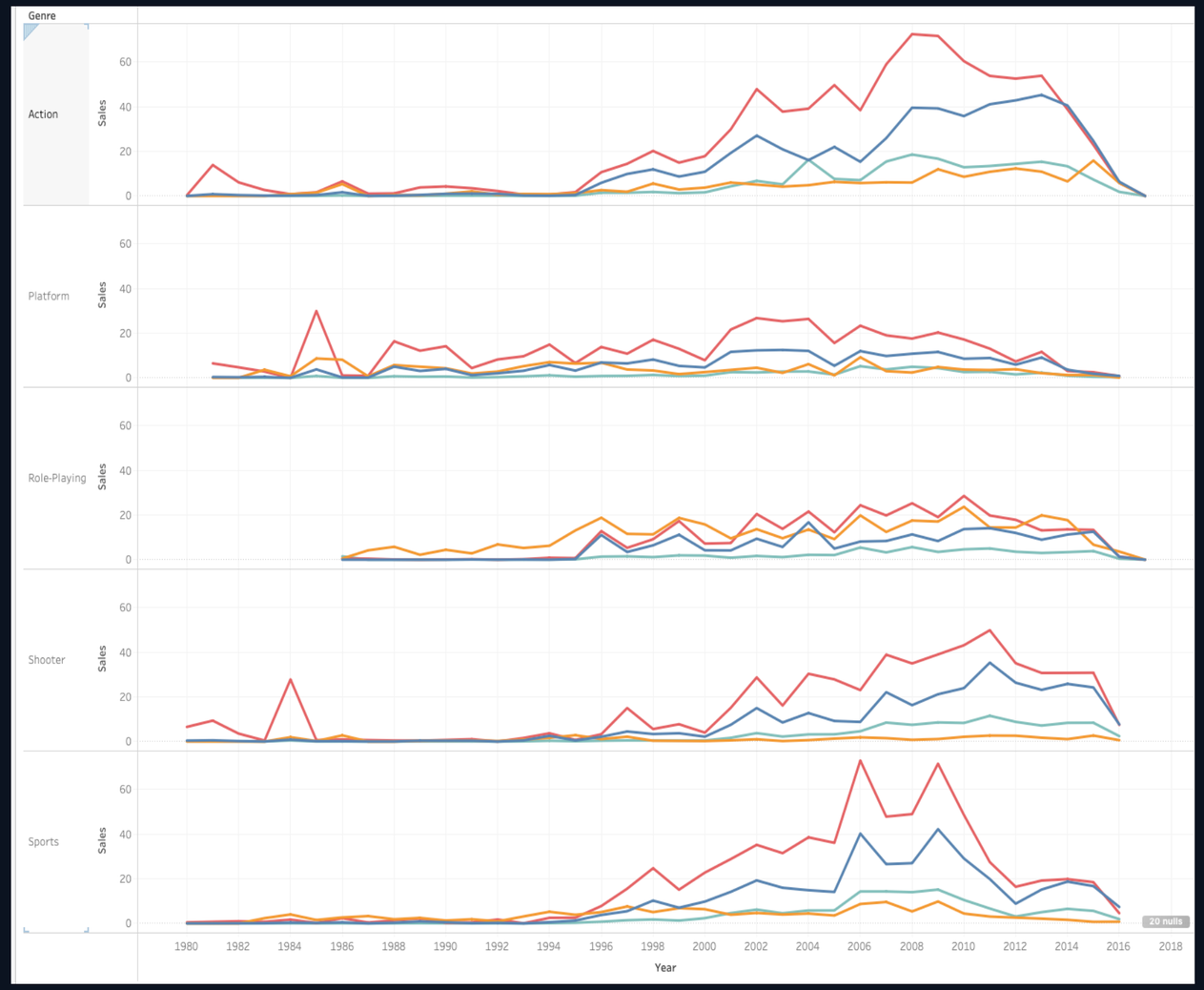
Agenda

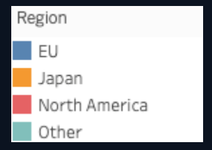
1. Dataset
2. Visualizations
3. Machine Learning (ML) Models
4. Analysis
5. Conclusion

Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
11493 unique values	DS 13% PS2 13% Other (12274) 74%	2009 9% 2008 9% Other (13739) 83%	Action 20% Sports 14% Other (10936) 66%	Electronic Arts 8% Activision 6% Other (14272) 86%	0 41.5	0 29	0 10.2	0 10.6	0.01 82.7
Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33
Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1	31.37
Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26
New Super Mario Bros.	DS	2006	Platform	Nintendo	11.38	9.23	6.5	2.9	30.01
Wii Play	Wii	2006	Misc	Nintendo	14.03	9.2	2.93	2.85	29.02
New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.59	7.06	4.7	2.26	28.62
Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31
Nintendogs	DS	2005	Simulation	Nintendo	9.07	11	1.93	2.75	24.76
Mario Kart DS	DS	2005	Racing	Nintendo	9.81	7.57	4.13	1.92	23.42
Pokemon Gold/Pokemon Silver	GB	1999	Role-Playing	Nintendo	9	6.18	7.2	0.71	23.1
Wii Fit	Wii	2007	Sports	Nintendo	8.94	8.03	3.6	2.15	22.72
Wii Fit Plus	Wii	2009	Sports	Nintendo	9.09	8.59	2.53	1.79	22
Kinect Adventures!	X360	2010	Misc	Microsoft Game Studios	14.97	4.94	0.24	1.67	21.82
Grand Theft Auto V	PS3	2013	Action	Take-Two Interactive	7.01	9.27	0.97	4.14	21.4

Dataset

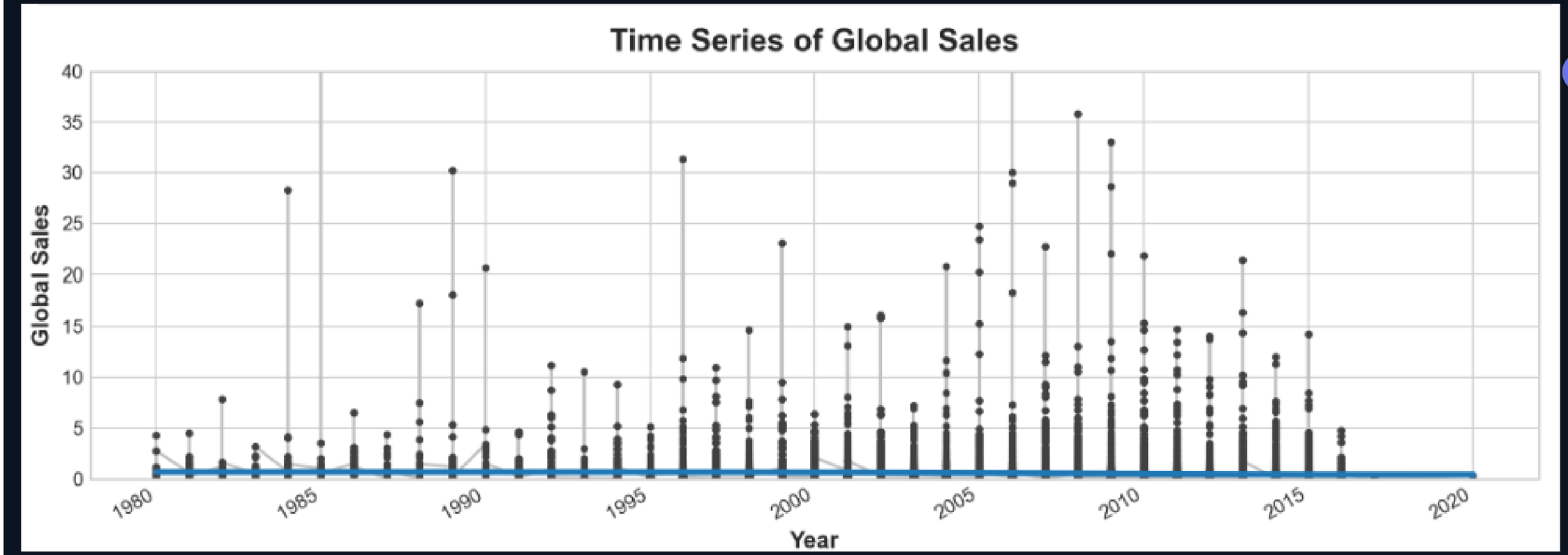
- The dataset used in this project was retrieved from Kaggle
- Year range from 1980 to 2020
- Features of focus: year, platform, publisher, region sales and global sales





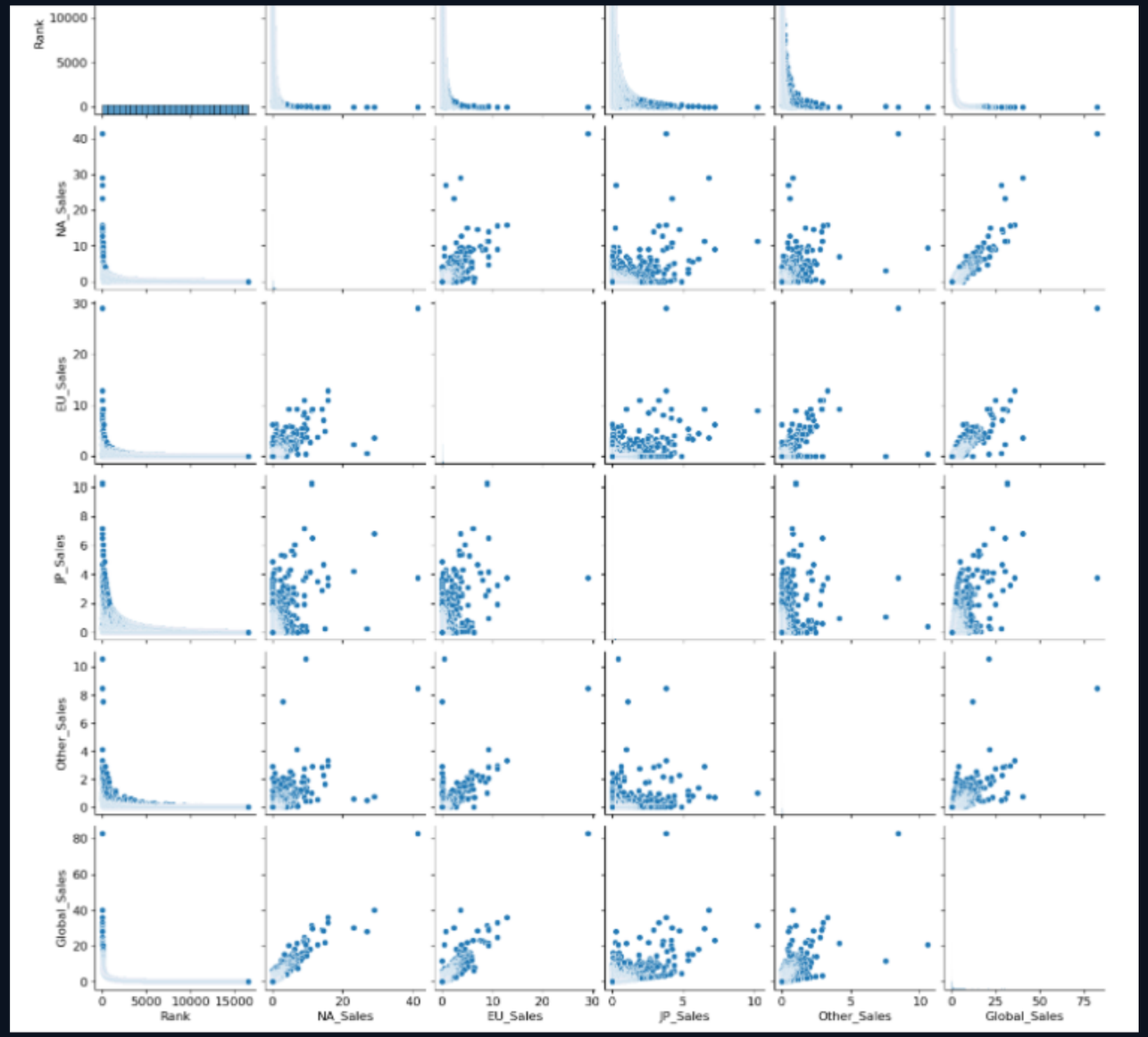
Visualizations: Sales by Genre

- The graph shows game sales of top 5 genres
- Top sales occur in North America
- 2002 has an increase of sales across genres



Visualizations: Time Series

- The graph shows Global Sales of all Genres over time.
- Majority of Sales are below 5M
- Outliers were considered when ML was deployed.



Analysis: Linear Regression

- Model yields an extremely low score of 15.5%
- This may be due to the dataset not having enough features for the model to make predictions on
- Lasso Regression was also trialed and the testing score was -3%

Imbalanced Testing Classification Report

	pre	rec	spe	f1	geo	iba	sup
0.0	0.92	0.97	0.96	0.94	0.96	0.93	1344
1.0	0.98	0.96	0.97	0.97	0.96	0.93	2729
avg / total	0.96	0.96	0.97	0.96	0.96	0.93	4073

RandomOverSampler Testing Classification Report

	precision	recall	f1-score	support
0.0	0.83	0.97	0.90	1344
1.0	0.98	0.91	0.94	2729
accuracy			0.93	4073
macro avg	0.91	0.94	0.92	4073
weighted avg	0.94	0.93	0.93	4073

Analysis: Logistic Regression

Hit Classification:

- 0 = Sales < $100k
- 1 = Sales >= $100k

- The Imbalanced Model yields an accuracy score of 96.2%

- The RandomOverSampler Model yields an accuracy score of 92.7%

Hyperparameter Optimization

```
{'activation': 'sigmoid',
 'first_units': 1,
 'num_layers': 3,
 'units_0': 1,
 'units_1': 16,
 'units_2': 26,
 'units_3': 26,
 'tuner/epochs': 50,
 'tuner/initial_epoch': 0,
 'tuner/bracket': 0,
 'tuner/round': 0}
```

Results

```
Best val_accuracy So Far: 0.9629265666007996
Total elapsed time: 00h 59m 06s
```

Analysis: Neural Networks

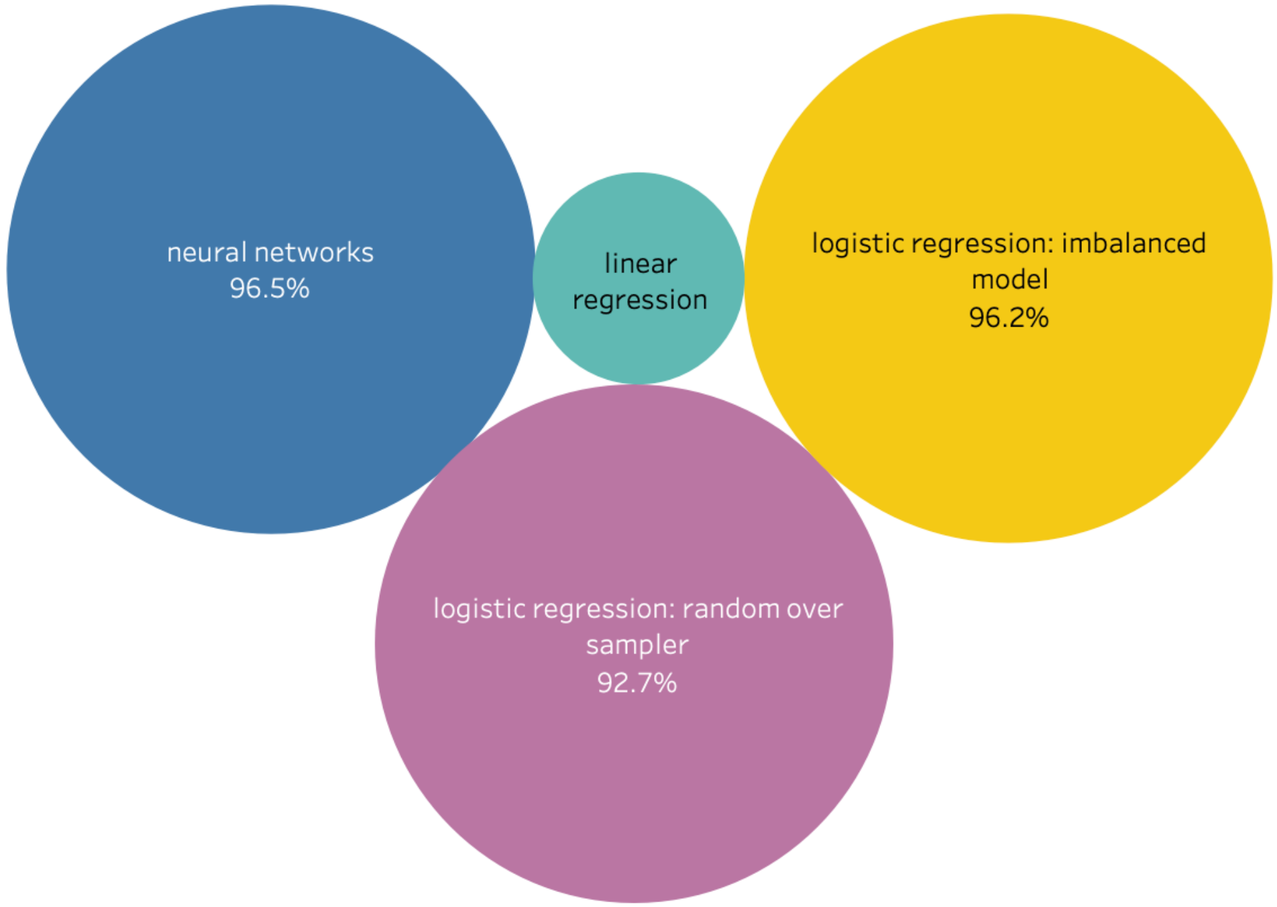
Hit Classification:

- 0 = Sales < $100k
- 1 = Sales >= $100k

- This model was used to compare traditional vs deep machine learning

- The best value accuracy score is 96.5%

Model Results Summary



1 negative

Conclusion

The required features were:

Year, genre, platform, publisher and global sales

The required classification:

0 = sales < $100k

1 = sales >= $100k

Best outcome was:

Classification with Logistic Regression

Alternate consideration: Neural Networks

Questions?