Video Games Sales Analysis, Modeling, Predictions, and Recommendations

Created by: Marin Stoytchev
Date: Jan. 24, 2021

Executive Summary

Problem Statement

- Can we predict which game based on Genre, ESRB Rating, Platform and Publisher will have high probability of being a best-selling game in the next two years?

Recommendations Based on Data Analysis and Model Predictions

- Based on the predictions of two independently optimized models and statistical analysis our recommendations for future high-selling games are:
  - **Choice 1**: Genre – Shooter; ESRB Rating – Mature; Platform – current generation of PlayStation console; Publisher – Activision
  - **Choice 2**: Genre – Sport;, ESRB Rating – Everyone; Platform – current generation of PlayStation console; Publisher – EA Sports
  - **Alternative**: For the same two game choices the platform can be the current generation of XBox console
  - **Final decision**: Whether to develop a Shooter or a Sports game for PlayStation or Xbox must be decided based on development time, cost and team expertise.

I. Process Outline

Six-Step Process

1. **Problem Definition**
2. **Data Collection**
3. **Data Pre-processing/Cleaning**
4. **Data Exploration and Visualization**
5. **Data Modeling and Model Selection**
6. **Results Interpretation and Recommendations Based on Model**

II. Problem Identification/Definition

Problem Definition

Problem

- In the last year due to COVID-19 there has been a noticeable increase in the number of people staying at home and increase in the time spent playing video games. A small video games developer startup of excellent professionals wants to capitalize on this trend. However, due to limited resources, the team can develop only one game which can be released in the next two years. Therefore, it is of crucial importance that the team focuses on developing a game which will have high probability to be top-selling game.

Question

- Can we predict which combination of Genre, Platform, ESRB Rating and Publisher will result in a highly successful game?

Risk

- If our prediction is inaccurate, this could lead to the bankruptcy of the video game startup since they have the capacity of producing only one game in a two-year period and don't have additional resources to sustain themselves beyond that time period.

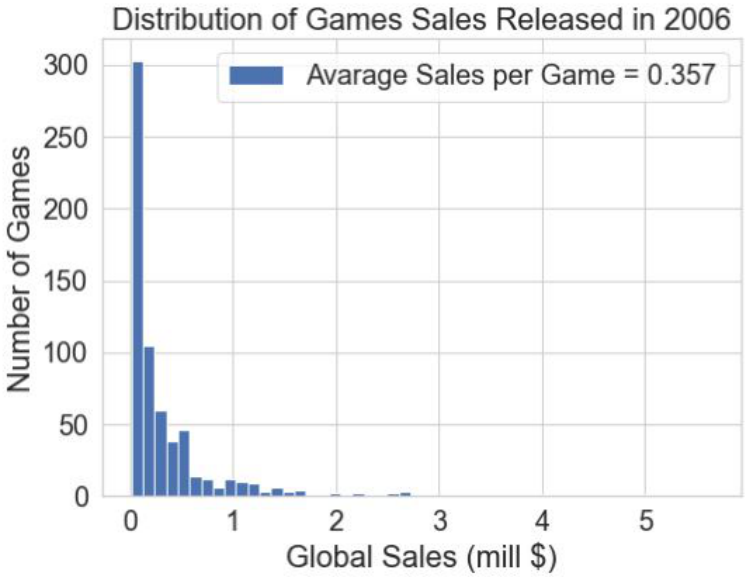
Task

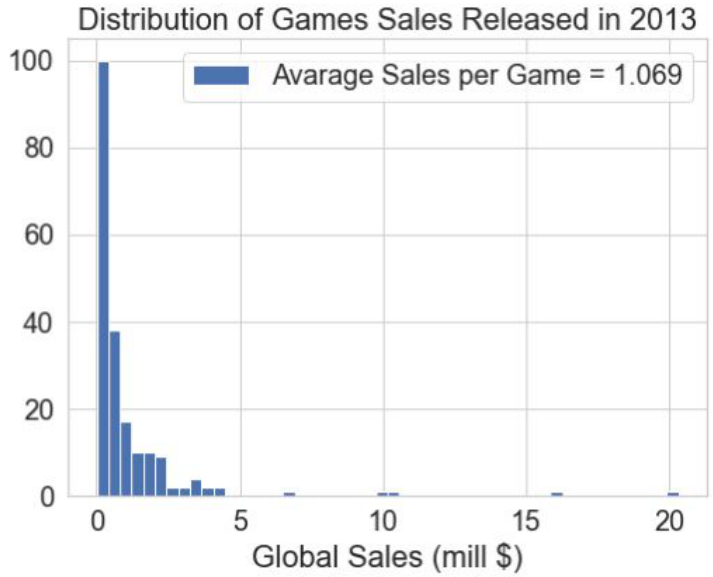
- Analyze available video games sales data and create a model that will predict with high confidence which video game will be successful in the next two years. Predictions to be delivered to client's executive team and the development team leader within 30 days.

III. Data Exploration

Distribution of Video Games Sales

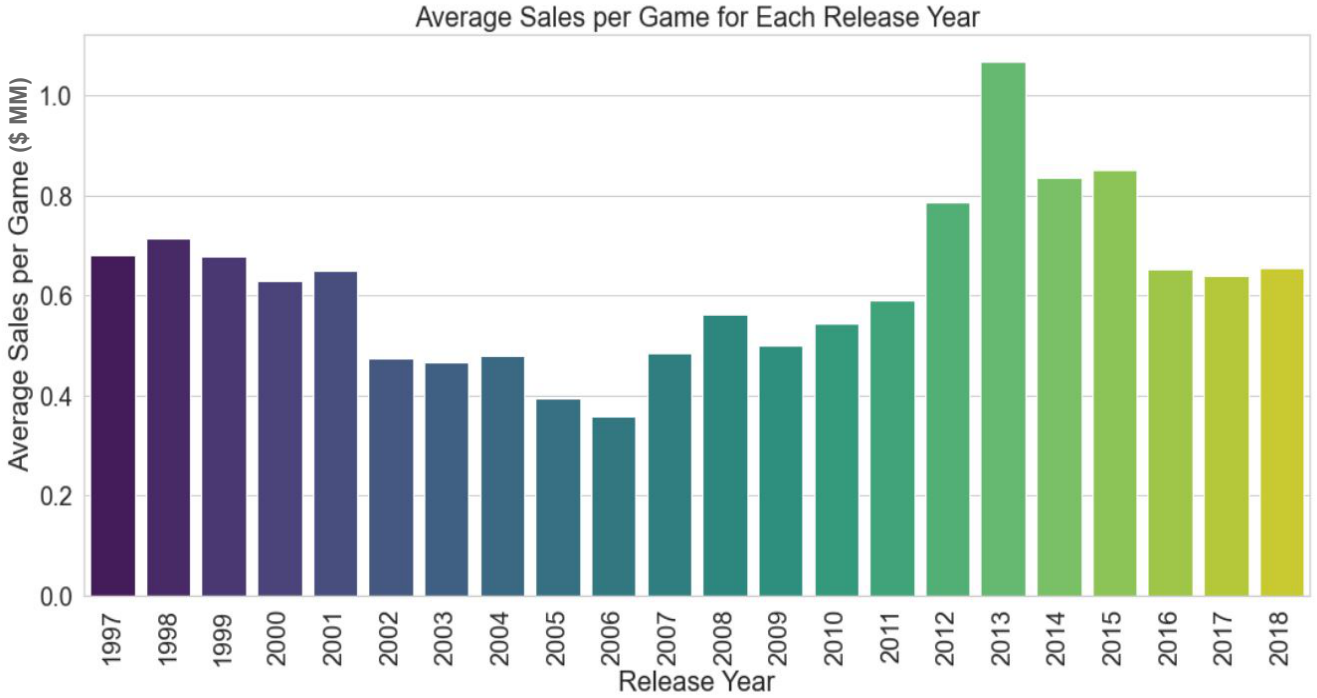
- The distributions of the sales of video games since their release are shown for two different years in our data
- The distributions are exponential and are characterized by:
  - Very small number of games with large sales since their release date
  - Large number of games with extremely small sales numbers
- That' why it is critical to be able to predict what attributes make the best-selling games



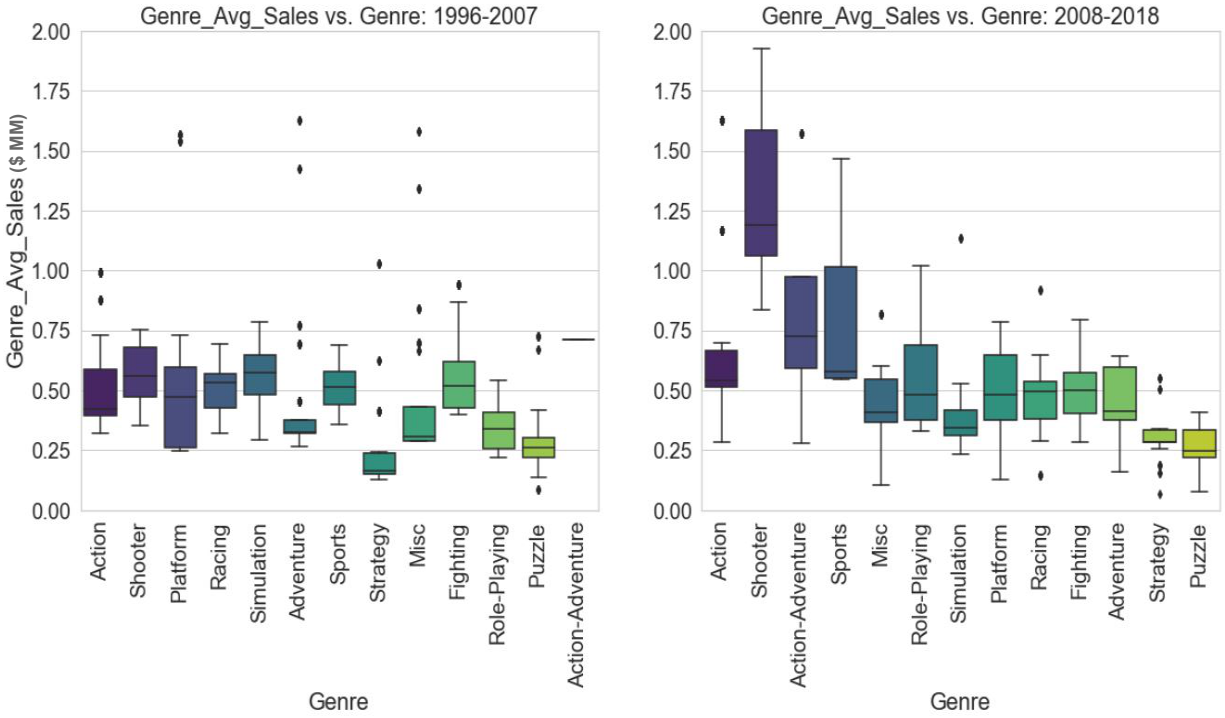


Video Games Average Sales vs. Release Year

- Average sales for games released in different years vary significantly
  - Smallest average sales value is observed for games released in 2006
  - Largest average sales are observed for games released in 2013

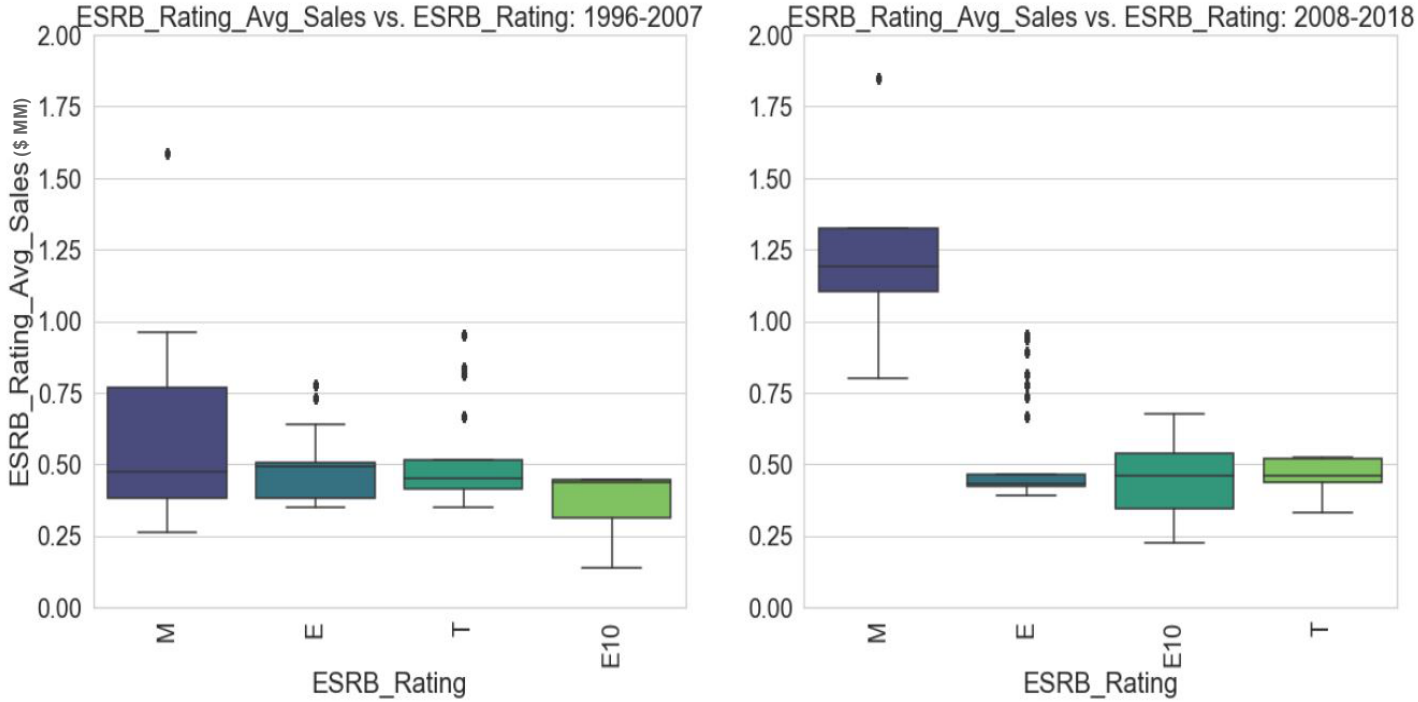


Average Sales per Genre



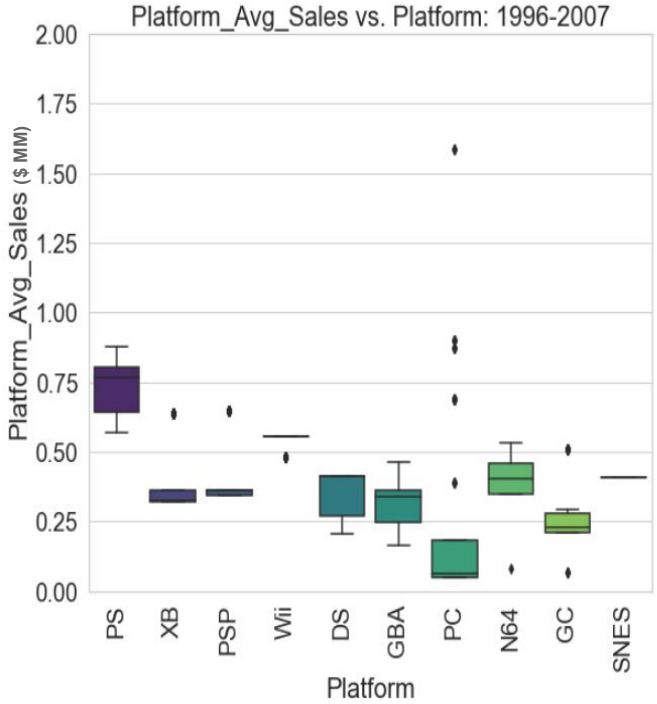
- Significant change in average sales by genre is observed between the two periods considered
  - 1996-2007: Sales for all genres fall within a close range without clear cut genre leaders
  - 2008-2018: Shooter genre approximately doubles in sales followed by Action-Adventure and Sports; the rest of the genres stay at approximately the same levels as in the previous period

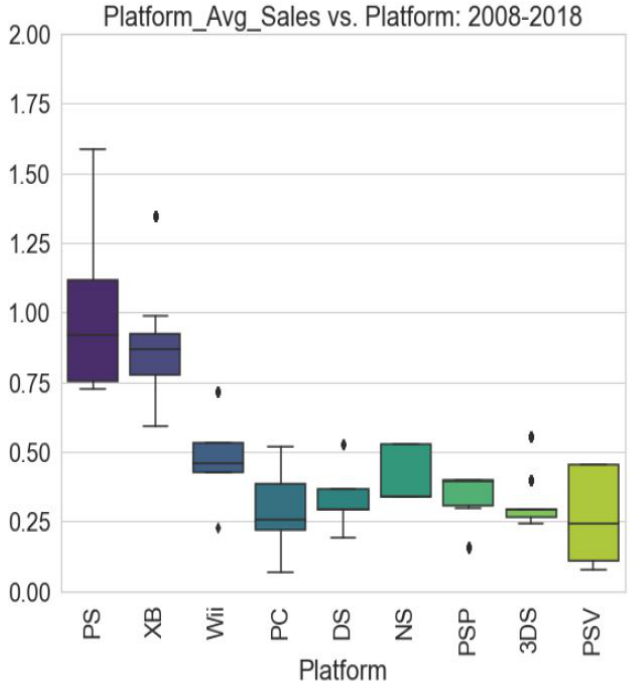
Average Sales per ESRB Rating



- Similar trend is observed here as well
  - 1996-2007: Sales for all ESRB ratings fall within a close range without clear-cut leader
  - 2008-2018: Games with "Mature" (M) rating double, while the rest stays at the same levels as before

Average Sales per Platform



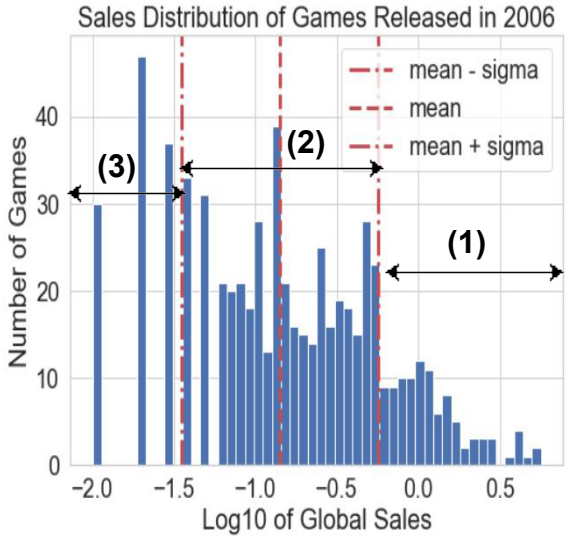


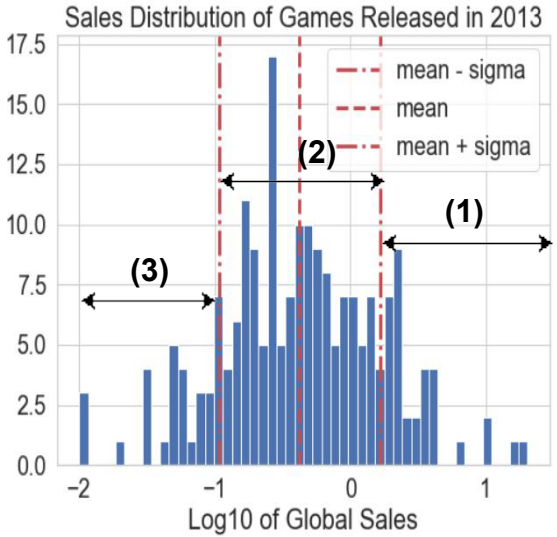
- Average sales by platform observations:
  - 1996-2007: Sony's PlayStation (PS) is the leading platform
  - 2008-2018: MicroSoft's XBox closes the gap and becomes the second leading platform

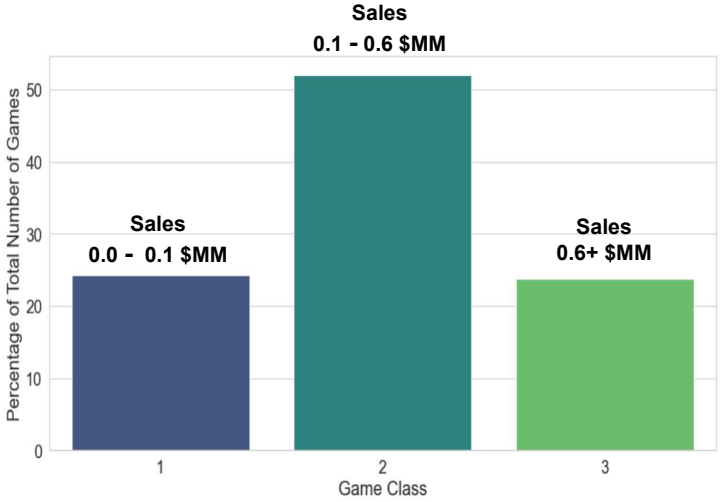
IV. Data Modeling and Predictions

Game Classes Based on Sales for Each Release Year

- The problem of predicting successful games in the future is a classification problem
- Accordingly games have been separated into three different classes
  - For most adequate partitioning Log10(Game Sales) is used resulting in: Class 1 - High-selling (24%); Class 2 - Midd-selling(52%); Class 3 - Low-selling (24%)
  - Please, note that the sales ranges for the game classes defined vary with release year. The ranges provided in the Game Class plot are the average ranges over all years







Model Predictions for Class 1 Games

- As the critical measure of model success we have chosen Class 1 Precision
- Using all available data we have tested two different models which provided the following results:
  - **RandomForest Classifier** without optimization - Class 1 Precision of 54%
  - **XGBoost Classifier** without optimization - Class 1 Precision of 67%
- Based on these results XGBoost Classifier has been selected for optimization to make future predictions

Results from RandomForest Classifier w/o Optimization

```
Confusion Matrix:
[[225 231  19]
 [165 656 166]
 [ 30 256 174]]
```

Classification Report:

	precision	recall	f1-score	support
1	0.54	0.47	0.50	475
2	0.57	0.66	0.62	987
3	0.48	0.38	0.42	460
accuracy			0.55	1922
macro avg	0.53	0.51	0.51	1922
weighted avg	0.54	0.55	0.54	1922

Results from XGBoost Classifier w/o Optimization

```
Confusion Matrix:
[[126 338  11]
 [ 53 878  56]
 [  8 350 102]]
```

Classification Report:

	precision	recall	f1-score	support
1	0.67	0.27	0.38	475
2	0.56	0.89	0.69	987
3	0.60	0.22	0.32	460
accuracy			0.58	1922
macro avg	0.61	0.46	0.46	1922
weighted avg	0.60	0.58	0.52	1922

Predicting Future High-Selling Games

- Predicting future high-selling games was performed using Precision and F1-score optimization methods of XGBoost Classifier
- Models were trained using data from 2014, 2015, and 2016
- Predictions were made for 2017 and 2018 and achieved the following Class 1 Precision marks
  - Optimization 1 model (metrics.precision_score) - Class 1 Precision of 72%
  - Optimization 2 model (metrics.precision_score) - Class 1 Precision of 71%

Results from Optimization 1 Model

```
Confusion Matrix:
[[ 48  41   8]
 [ 15 154  36]
 [  4  67  44]]
```

```
Classification Report:
              precision    recall  f1-score   support

           1       0.72      0.49      0.59        97
           2       0.59      0.75      0.66       205
           3       0.50      0.38      0.43       115

    accuracy                           0.59       417
   macro avg       0.60      0.54      0.56       417
weighted avg       0.59      0.59      0.58       417
```

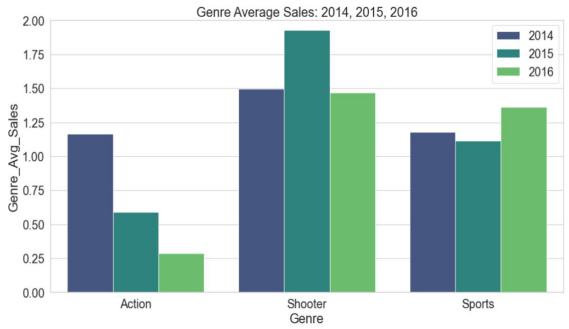
Results from Optimization 2 Model

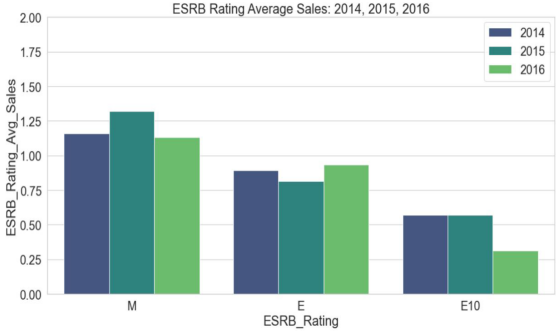
```
Confusion Matrix:
[[ 46  42   9]
 [ 15 156  34]
 [  4  68  43]]
```

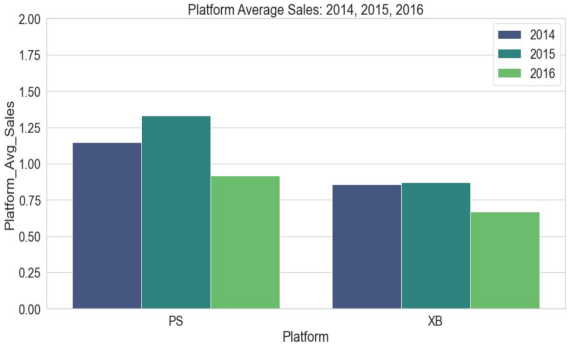
```
Classification Report:
              precision    recall  f1-score   support

           1       0.71      0.47      0.57        97
           2       0.59      0.76      0.66       205
           3       0.50      0.37      0.43       115

    accuracy                           0.59       417
   macro avg       0.60      0.54      0.55       417
weighted avg       0.59      0.59      0.58       417
```

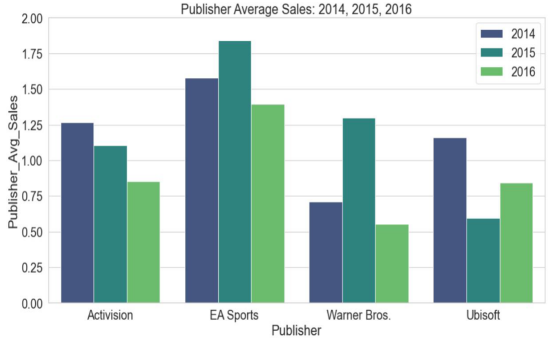
Attributes of Future Most Likely High-Selling Games

- The attributes of future most likely high-selling games were determined by selecting the feature vectors voted most often as Class 1 games
- The exact same seven feature vectors were provided by both models
- The most likely future high-selling games were reduced further down to two based on average sales for the preceding years shown to the right
- The top choices with 88% level of confidence are:
  - **Choice #1**: *Genre - Shooter; ESRB rating - Mature; Platform - PlayStation; Publisher - Activision*
  - **Choice #2**: *Genre - Sports; ESRB rating - Everyone; Platform - PlayStation; Publisher - EA Sports*









V. Summary

Summary

Summary of Performed Data Analysis and Modeling

- Video games sales data comprised of games released between 1997 and 2018 has been processed, explored and used with machine learning models to predict the attributes of future most likely high-selling games
- For classifications of the games as high-, mid-, and low-selling XGBoost Classifier has been optimized and used for predictions
- The top two choices for potentially high-selling games in the next couple of years have been determined with 88% confidence based on predictions from two independently optimized models and complimentary analysis of games sales by Genre, ESRB rating, Platform and Publisher
- These choices are presented in the Executive Summary

Potential for Improvement

- Critics and users games scores are practically missing in the current dataset and for this reason have not been used here. If these data were available, modeling and predictions would most certainly be improved.
- In addition to the current data which refers only to hard copies sold for paid games, it would be valuable to include data for soft copies sold and free-to-play games which currently are a significant part of the video game market.