Predicting Video Game Sales Performance Levels

Daniel Rios

Can we predict the performance level of a video game?

The Dataset

- From Kaggle
  - Scraped from vgchartz.com

The Dataset

- From Kaggle
  - Scraped from vgchartz.com
  - Video games with sales greater than 100,000 copies

The Dataset

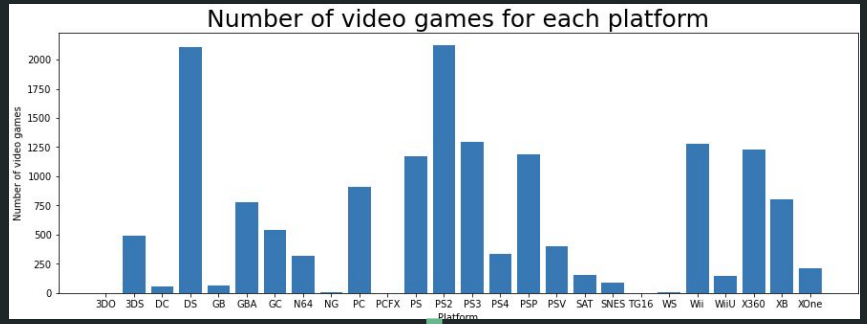
- From Kaggle
  - Scraped from vgchartz.com
  - Video games with sales greater than 100,000 copies
  - 16,598 total records

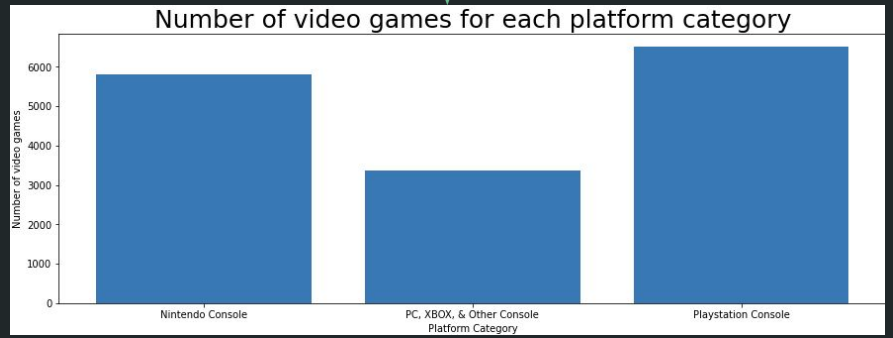
The Dataset

- From Kaggle
  - Scraped from vgchartz.com
  - Video games with sales greater than 100,000 copies
  - 16,598 total records
- Columns used:
  - Platform
  - Year
  - Genre
  - Publisher
  - Global Sales

Feature and Target Engineering

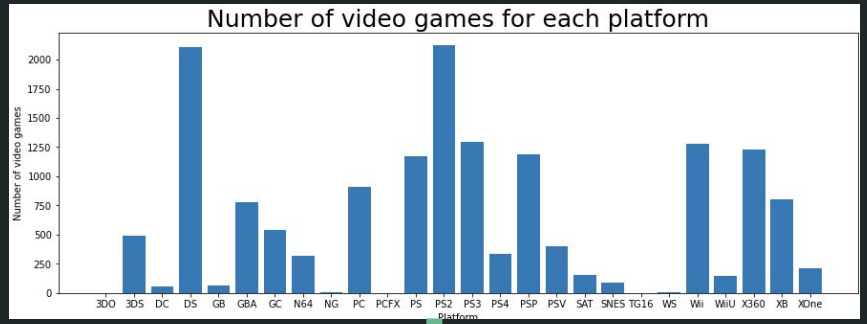
- 3 platform categories and publisher categories

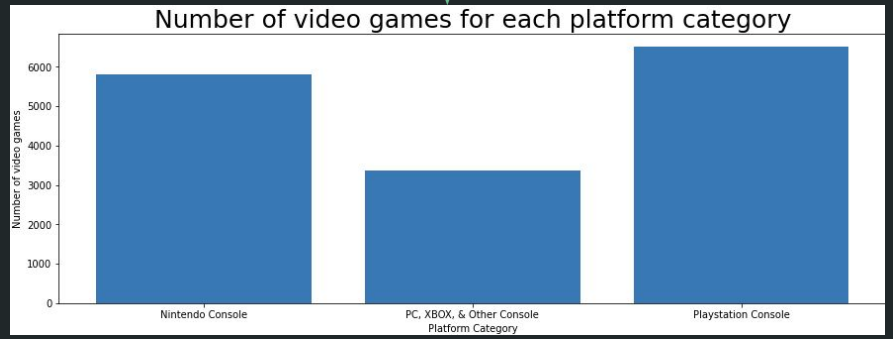




Feature and Target Engineering

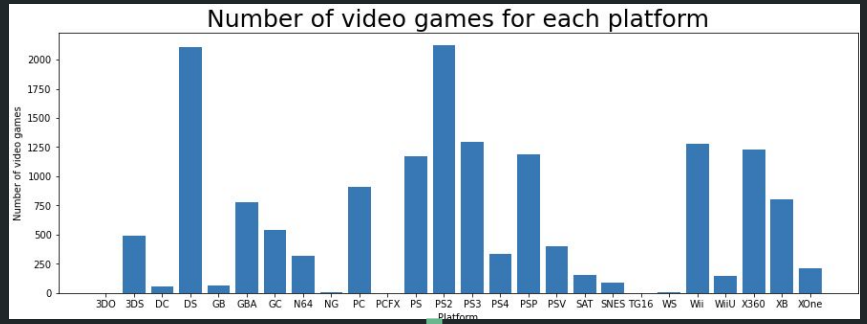
- 3 platform categories and publisher categories
- Genre left imbalanced for explanatory and predictive power

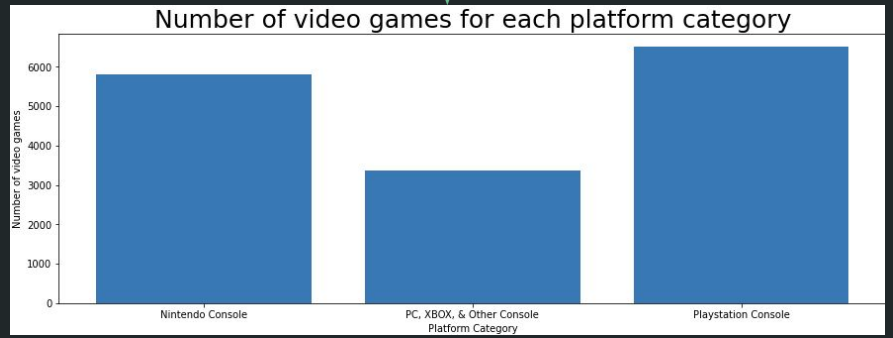




Feature and Target Engineering

- 3 platform categories and publisher categories
- Genre left imbalanced for explanatory and predictive power
- Global sales converted into 3 performance categories





Why would we want to know a video game's performance level?

- To help determine which game(s) a video game outlet could promote for the upcoming year as a method for maximizing sales

Why would we want to know a video game's performance level?

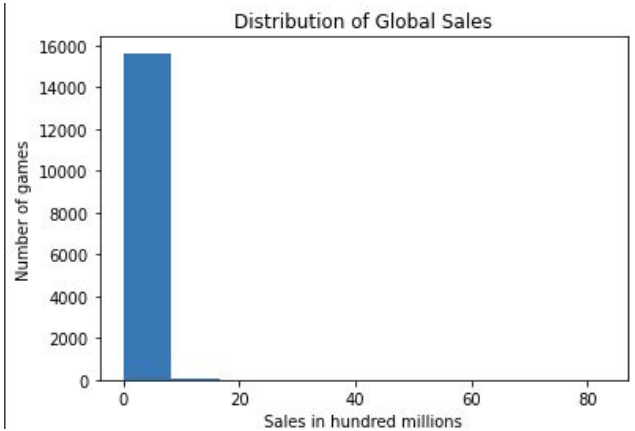
- To help determine which game(s) a video game outlet could promote for the upcoming year as a method for maximizing sales
- A method to consider which video games to invest stock in

Why would we want to know a video game's performance level?

- To help determine which game(s) a video game outlet could promote for the upcoming year as a method for maximizing sales
- A method to consider which video games to invest stock in
- Performance level vs. global sales count
  - Relative to other games
  - Predictive ability of the engineered features

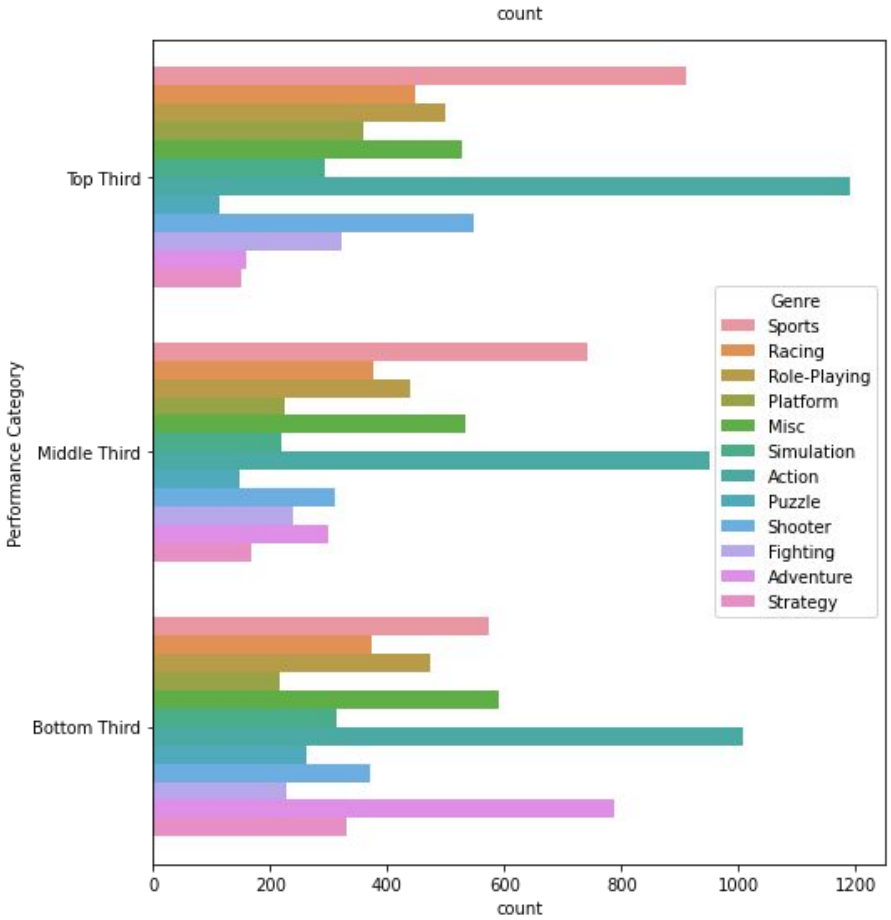
Why would we want to know a video game's performance level?

- To help determine which game(s) a video game outlet could promote for the upcoming year as a method for maximizing sales
- A method to consider which video games to invest stock in
- Performance level vs. global sales count
  - Relative to other games
  - Predictive ability of the engineered features
- Distribution of the global sales variable



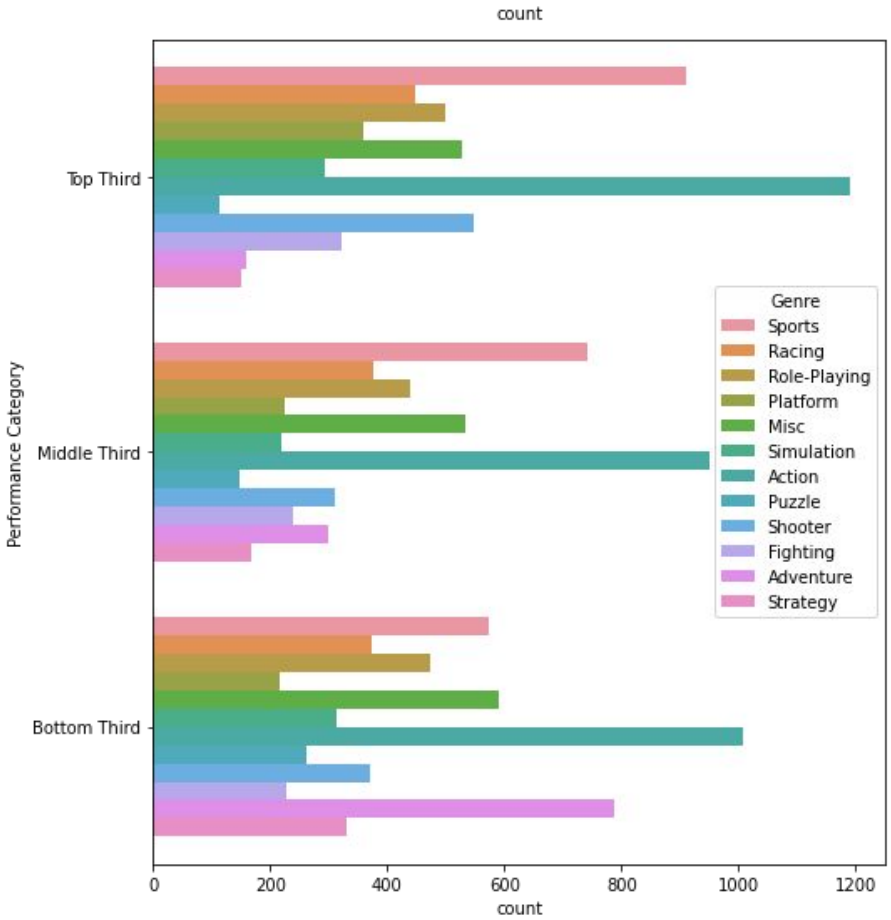
Data Exploration

- Number of games by genre in each performance category
  - Adventure games in top third vs. bottom third



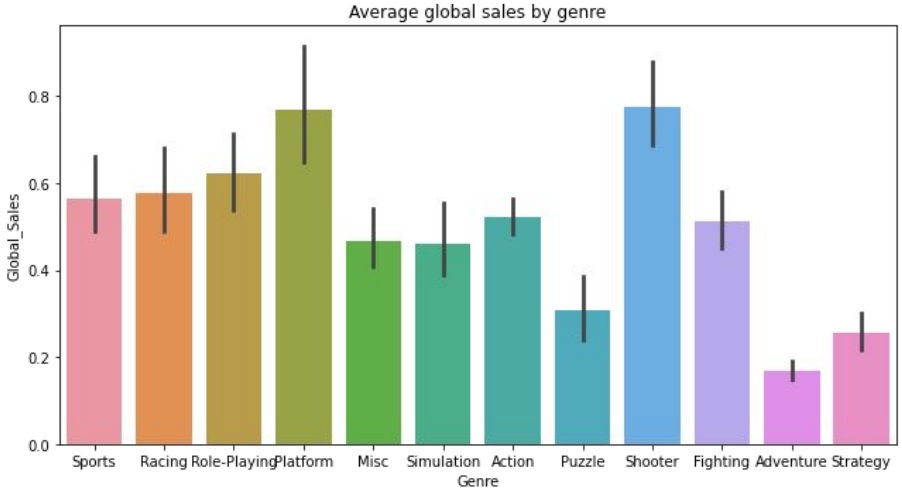
Data Exploration

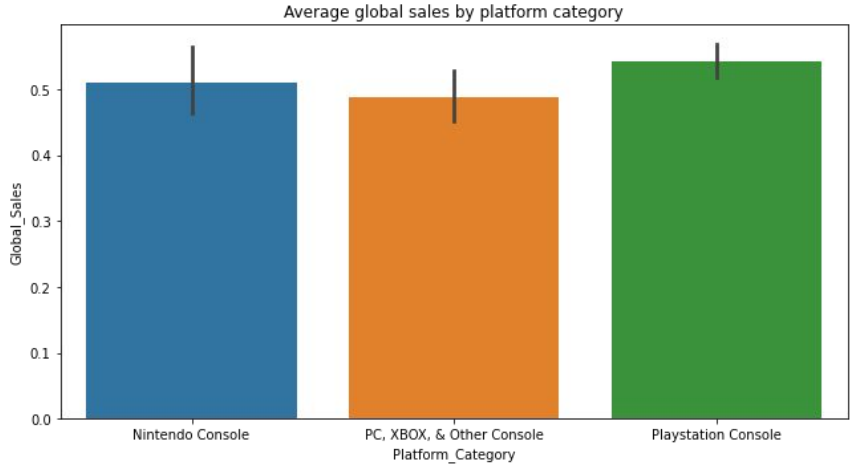
- Number of games by genre in each performance category
  - Adventure games in top third vs. bottom third
  - Action and sports tied with the variability of Playstation



Data Exploration

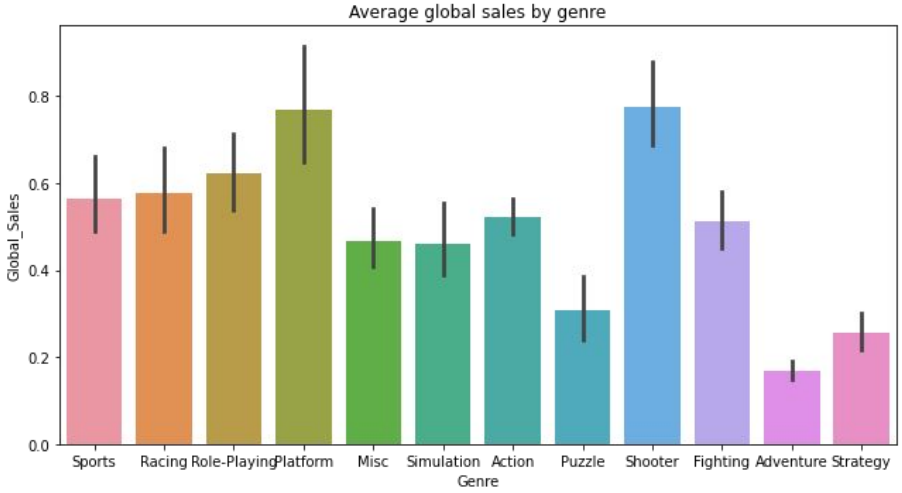
- Global sales by genre and platform category
  - Adventure games have lowest average global sales

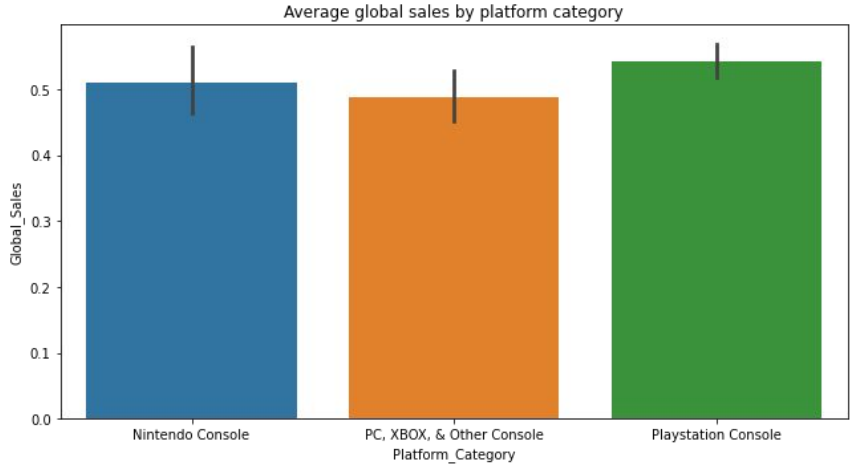




Data Exploration

- Global sales by genre and platform category
  - Adventure games have lowest average global sales
  - Global sales does not vary much due to the console





Training and Testing Models

- Using categorical variables to predict another categorical variable

Training and Testing Models

- Using categorical variables to predict another categorical variable
- Multi-class problem (Top third, Middle third, Bottom third)

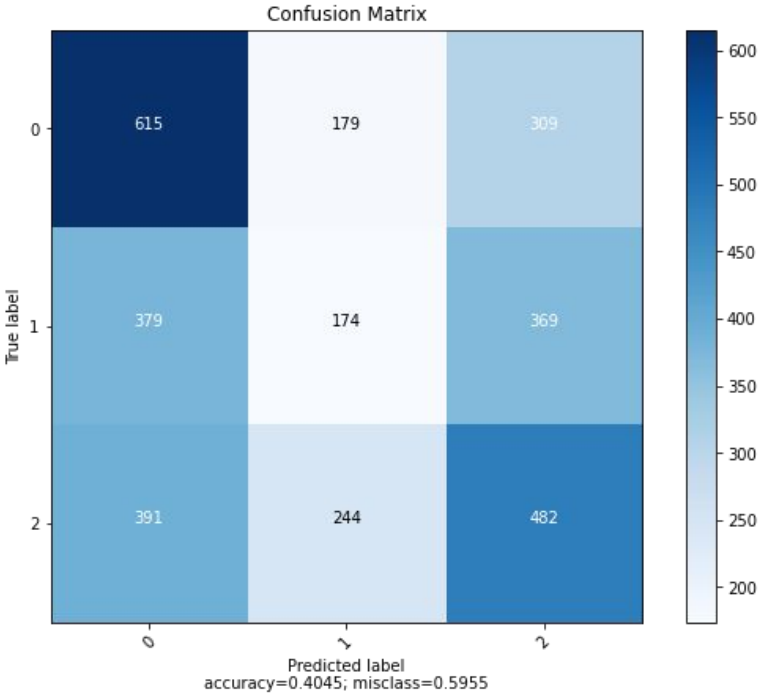
Training and Testing Models

- Using categorical variables to predict another categorical variable
- Multi-class problem (Top third, Middle third, Bottom third)

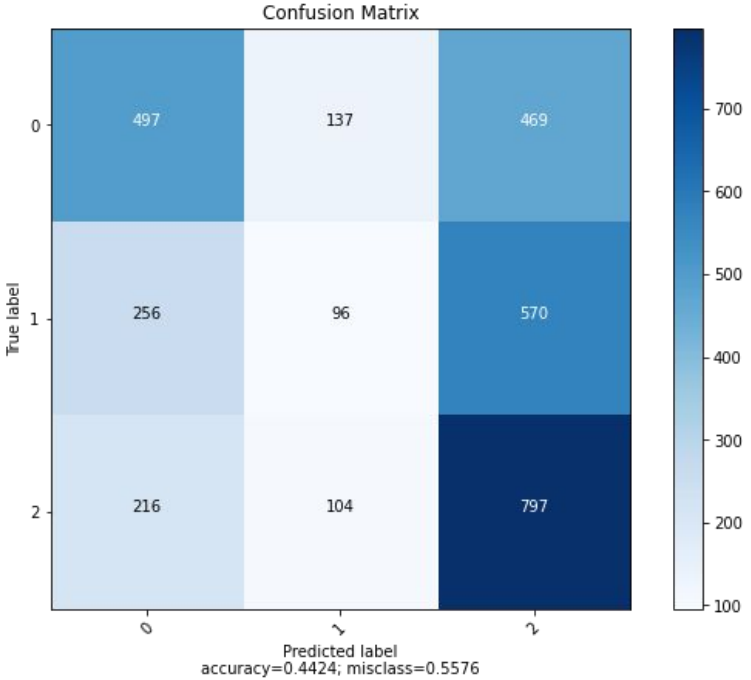
Model Type	Range of Accuracies (%)
Logistic Regression	40 – 48
9 Nearest Neighbors (Optimized KNN)	40 – 45
Random Forest Classifier	40 – 47
Support Vector Classifier	42 – 44
Gradient Boosting Classifier	43 – 45

Confusion Matrices

KNN

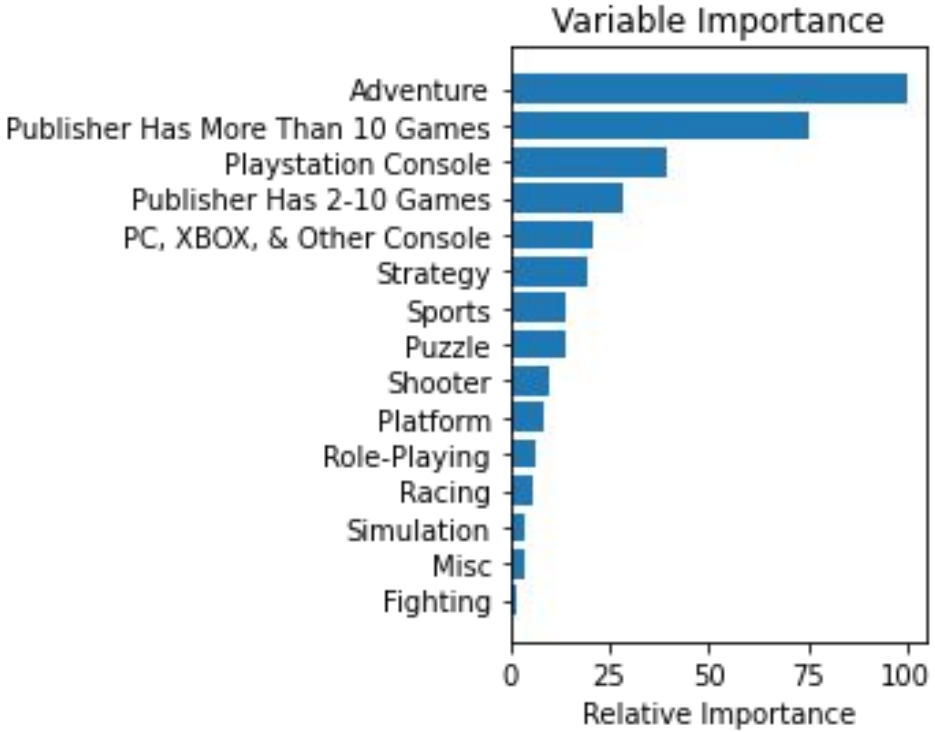


Random Forest



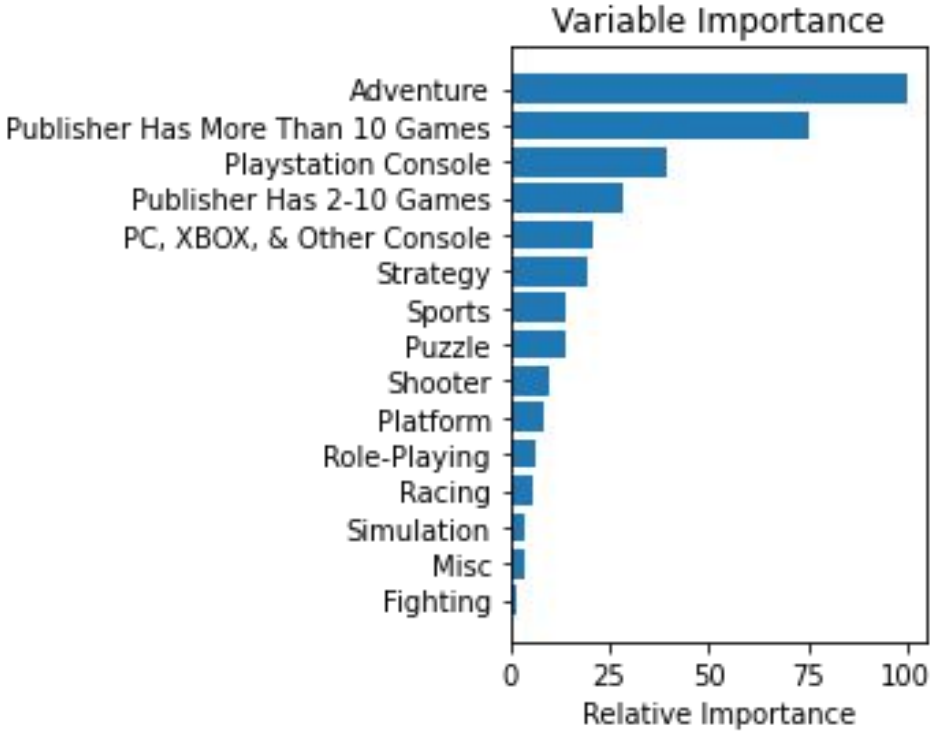
Feature Importance

- Playstation has the most variability in performance category



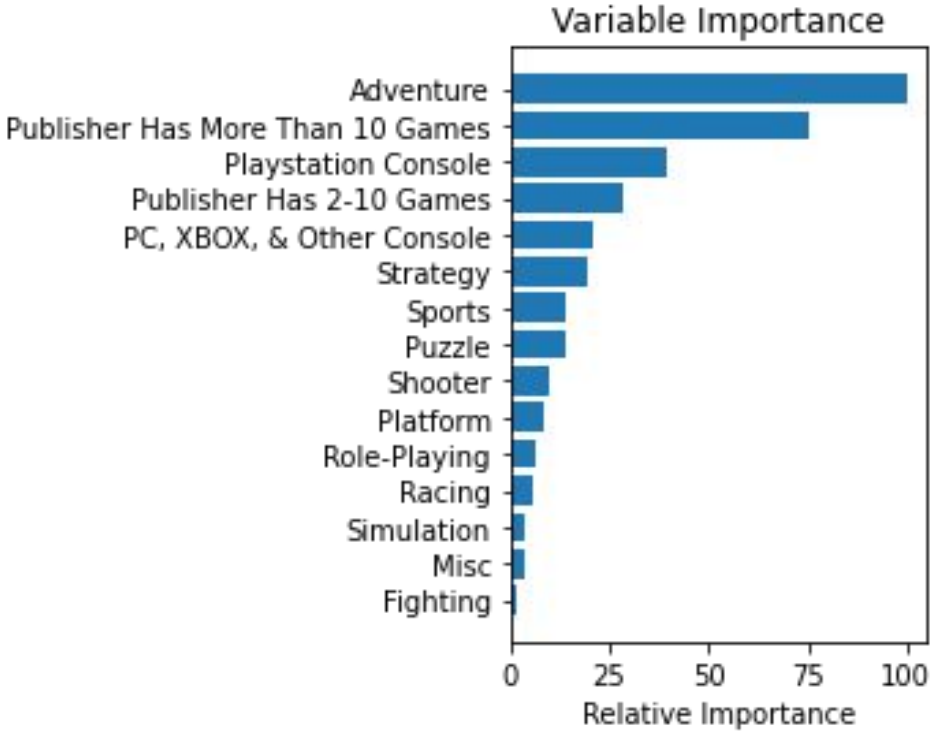
Feature Importance

- Playstation has the most variability in performance category
- Games that are made by publishers with prior experience have more variance in performance



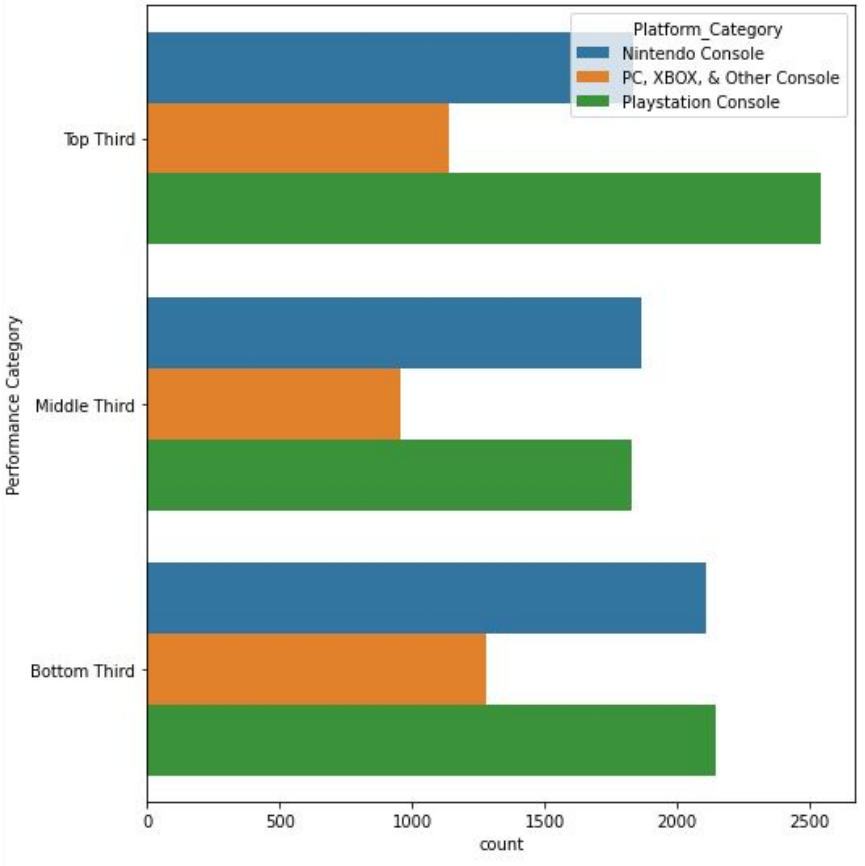
Feature Importance

- Playstation has the most variability in performance category
- Games that are made by publishers with prior experience have more variance in performance
- Platform is one of the only genres that explains some variance *and* performs better in regards to global sales



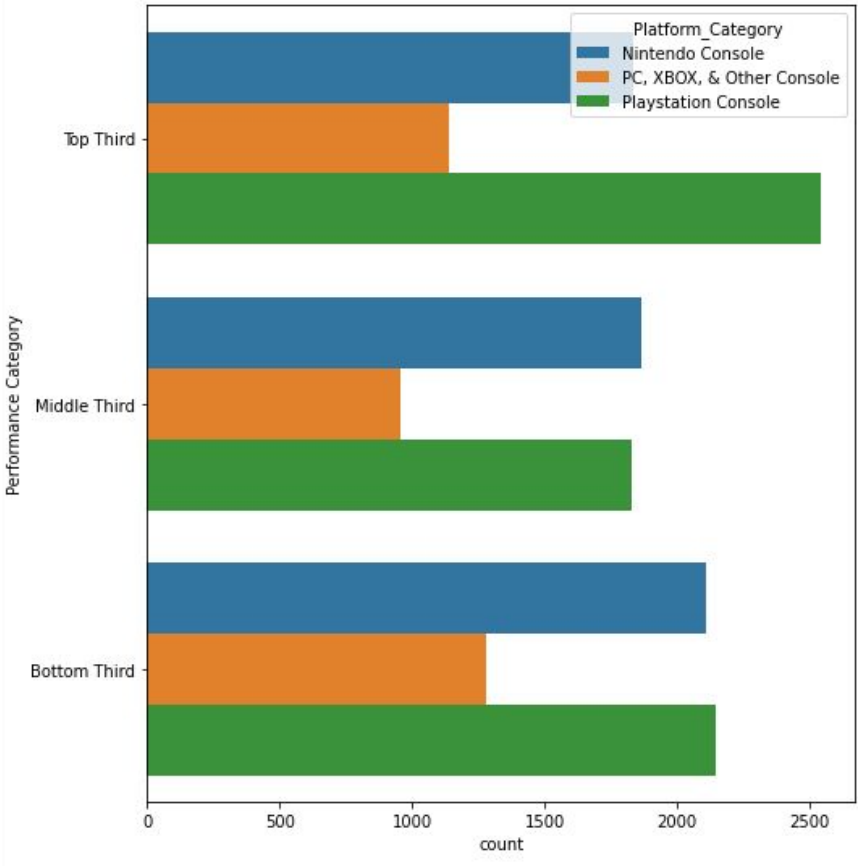
Shortcomings

- Severe underprediction of a certain class in the target variable



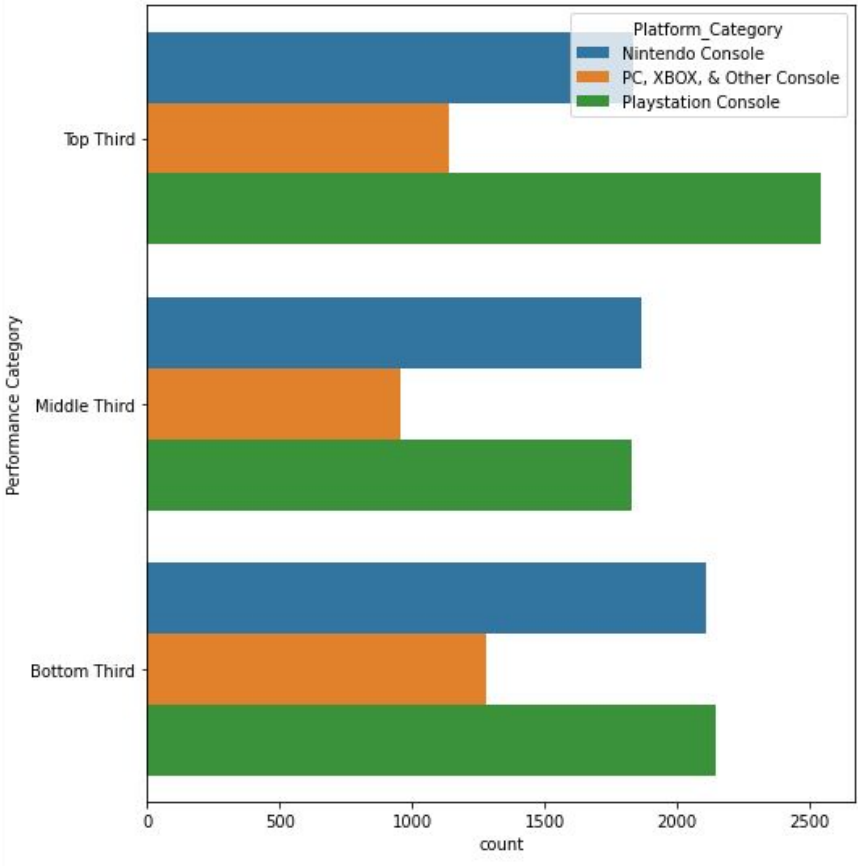
Shortcomings

- Severe underprediction of a certain class in the target variable
  - Adjusting the number of classes



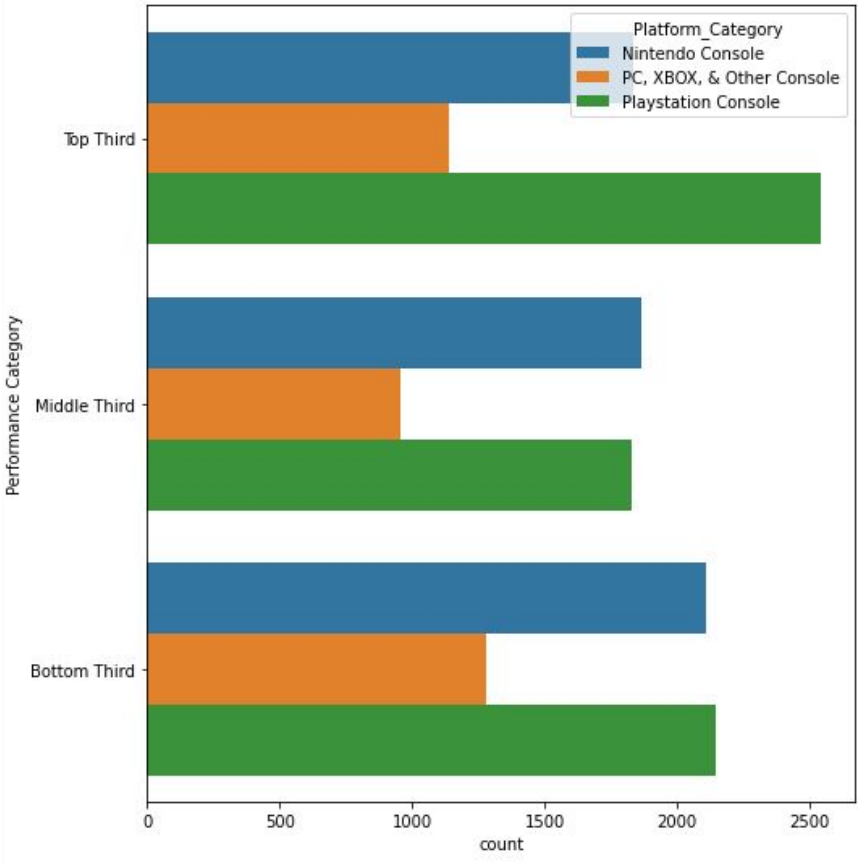
Shortcomings

- Severe underprediction of a certain class in the target variable
  - Adjusting the number of classes
  - Explanation of variance in data from features



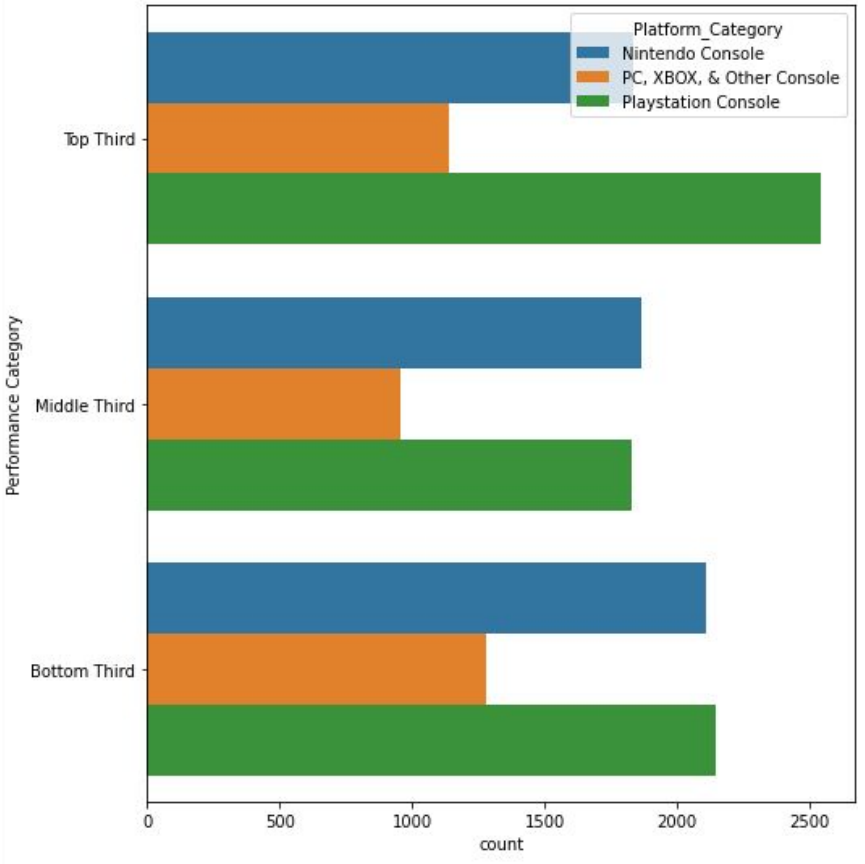
Shortcomings

- Severe underprediction of a certain class in the target variable
  - Adjusting the number of classes
  - Explanation of variance in data from features
    - PCA/MCA did not help address this issue



Shortcomings

- Severe underprediction of a certain class in the target variable
  - Adjusting the number of classes
  - Explanation of variance in data from features
    - PCA/MCA did not help address this issue
- Could use further parameter tuning in GBC



Model Comparison and Evaluation

- Models that predicted all 3 classes
  - Logistic Regression
  - 9-Nearest Neighbors
  - Random Forest Classifier
  - GB Classifier

Model Comparison and Evaluation

- Models that predicted all 3 classes
  - Logistic Regression
  - 9-Nearest Neighbors
  - Random Forest Classifier
  - GB Classifier
- 9-Nearest Neighbors was more accurate with the under represented class
  - Range of accuracies 40% – 45%
  - 597 for the “middle third” class

Final Notes

- 9-Nearest Neighbors would be the optimal model

Final Notes

- 9-Nearest Neighbors would be the optimal model
- If the model predicts a video game to be in the top third, take this with extreme caution.

Final Notes

- 9-Nearest Neighbors would be the optimal model
- If the model predicts a video game to be in the top third, take this with extreme caution.
- If the model predicts a video game to be in the bottom third, you can expect the game to perform badly.

Final Notes

- 9-Nearest Neighbors would be the optimal model
- If the model predicts a video game to be in the top third, take this with extreme caution.
- If the model predicts a video game to be in the bottom third, you can expect the game to perform badly.
- Adventure games tend to not do well

Final Notes

- 9-Nearest Neighbors would be the optimal model
- If the model predicts a video game to be in the top third, take this with extreme caution.
- If the model predicts a video game to be in the bottom third, you can expect the game to perform badly.
- Adventure games tend to not do well
- Shooter, Platform, and Role-Playing games tend to outperform the others

Thank you! Questions?

Daniel Rios