



Steam Game Genre Analysis



Matt Porter and Grant Ballard



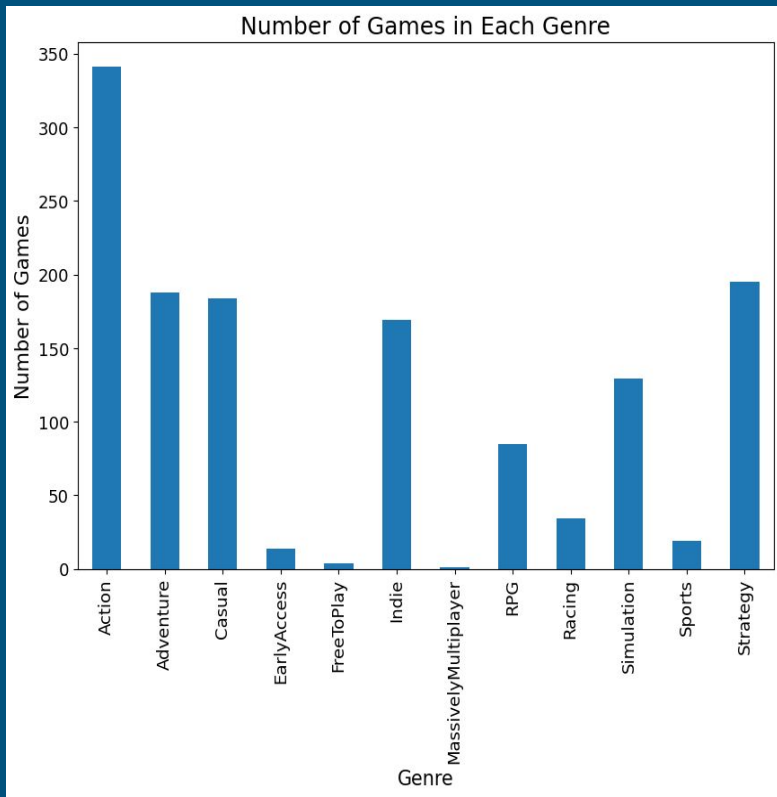
Introduction and Motivation

- We chose to use a dataset of Steam games due to interest and high familiarity
- Overall, we want to classify different genres and develop a relationship between the Metacritic score and user recommendation count
- We began by using linear regression, due to the relationship of the data in the Metacritic score and user recommendation count
- We switched to the Random Trees method, because it can account for small discrepancies in data

Dataset (Overview)

- The dataset we used is a collection of information about thousands of games on Steam
- The features are name of game, release date, Metacritic score, recommendation count, whether or not it is free, and genre(s)
- We had to do a lot of data manipulation before applying machine learning

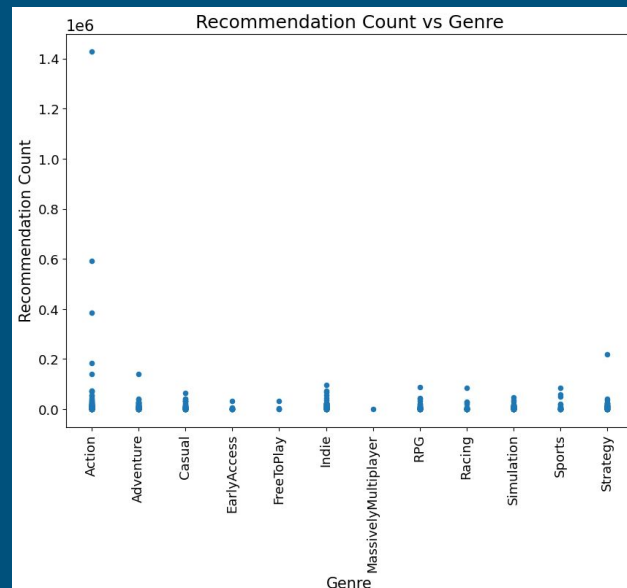
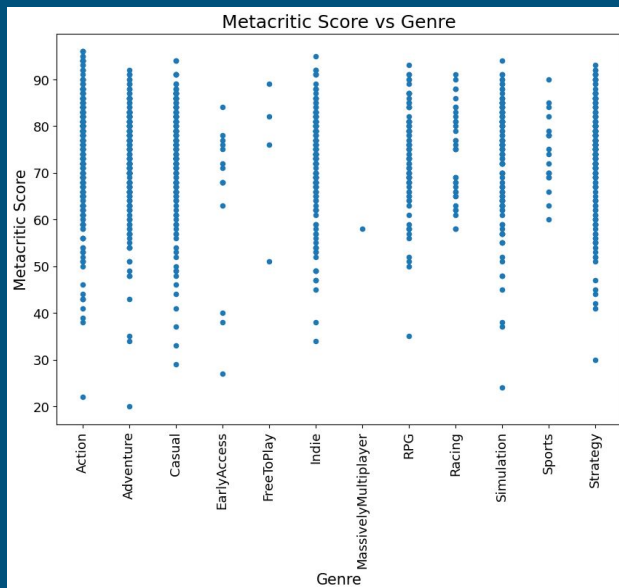
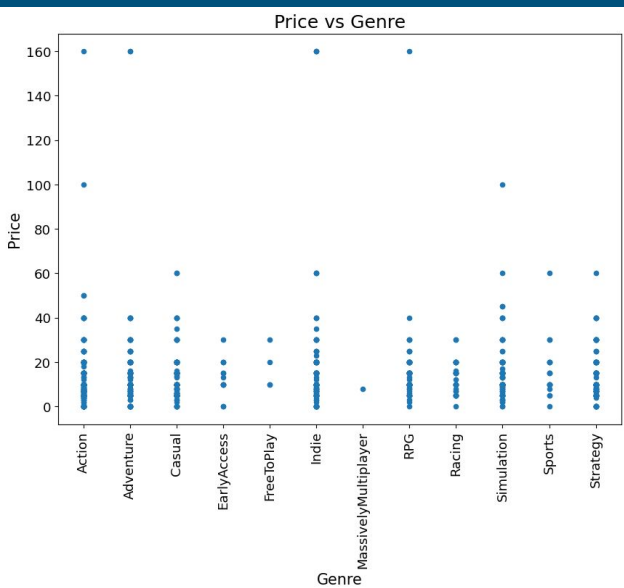
Dataset Visuals



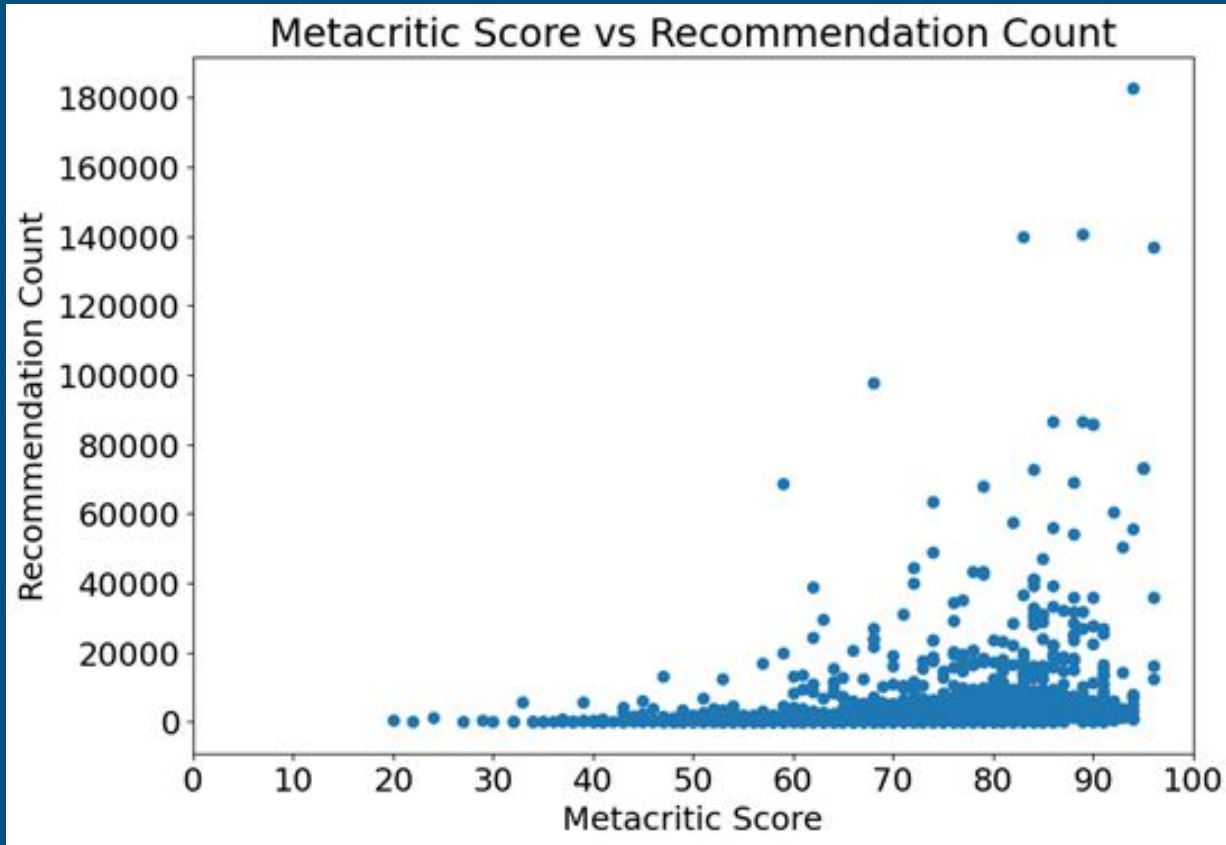
In order to get this graph, we had to manipulate the data to go from many genre features with a boolean value in to one genre feature with a text value stating the genre

Dataset Visuals (Random Forests)

- Below, we graphed what we considered to be the 'main' genres against the initial price, the Metacritic score, and the user recommendation count.
- As can be seen from the graphs, this does not indicate that we will get extremely good results



Dataset (Linear Regression)



ML Approaches and Methodologies

Random Forests:

- We chose to use this in comparing genres because how easily it fit the problem at hand and it would help us determine which features were important in distinguishing the genres
- Our success measure was the accuracy score of the Random Forest, and we would determine ourselves successful if our accuracy score was $(1/n * 100)\%$ or greater, with n being the number of genres we are classifying in the tree
- The hyperparameters we considered were number of estimators, max depth, and minimum samples split.
- We also used Adaboost on the random forest classifier, but did not explore the hyperparameters and instead arbitrarily used 100 estimators and the default learning rate

ML Approaches and Methodologies (cont.)

Linear Regression

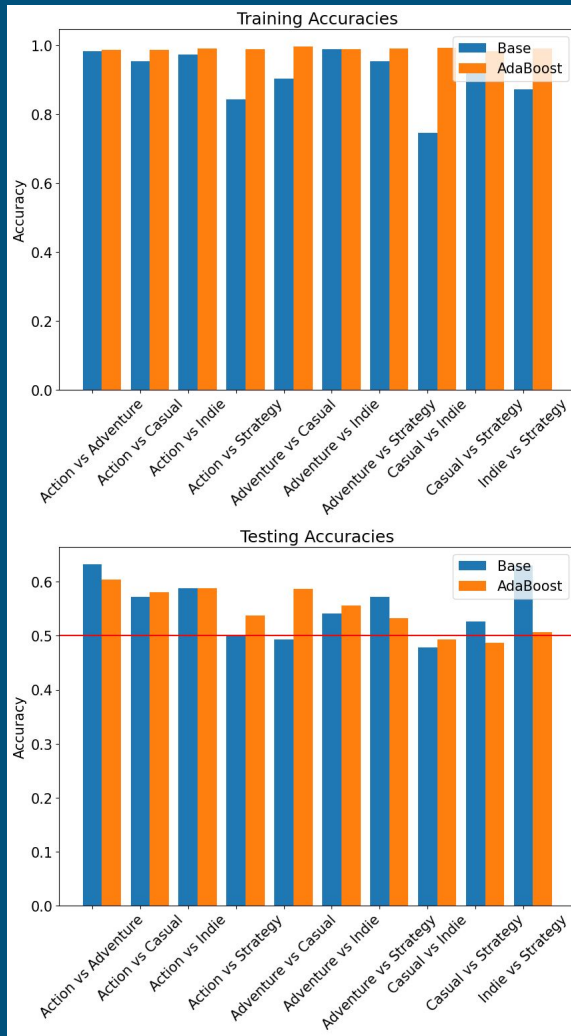
- We initially chose to use linear regression due to a semi-linear correlation between the Metacritic scores and user recommendation counts of all the games individually
- We compared both the ridge and lasso linear regressions on the data
- We also used KFold cross validation in order to determine the best value for the hyperparameter alpha

Results (Random Forest)

For the random forest accuracy scores, most of the genre comparisons had a training score at or close to 1, especially with Adaboost.

In contrast, the testing accuracies hovered between 0.5 and 0.6, and many accuracies actually decreased with adaboost.

This could be an indication of overfitting



Results

Interestingly enough, when we look at a few of the first decision trees of the random forests, we see that all four first split on the Metacritic score

This could be an indication of a connection between a game's genre and the Metacritic score it receives.

```
Action vs Adventure
```

```
|--- Metacritic <= 86.50  
|   |--- RecommendationCount <= 107.00  
|   |   |--- PriceInitial <= 2.49  
|   |   |   |--- class: 1.0
```

```
Action vs Casual
```

```
|--- Metacritic <= 65.50  
|   |--- RecommendationCount <= 170.00  
|   |   |--- PriceInitial <= 39.99
```

```
Adventure vs Strategy
```

```
|--- Metacritic <= 73.50  
|   |--- RecommendationCount <= 107.50  
|   |   |--- PriceInitial <= 27.49
```

```
Indie vs Strategy
```

```
|--- Metacritic <= 83.50  
|   |--- RecommendationCount <= 218.00  
|   |   |--- PriceInitial <= 2.49
```

Results

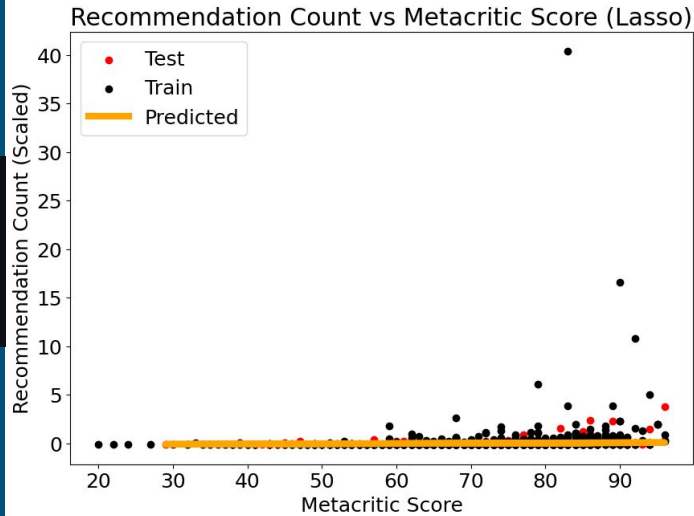
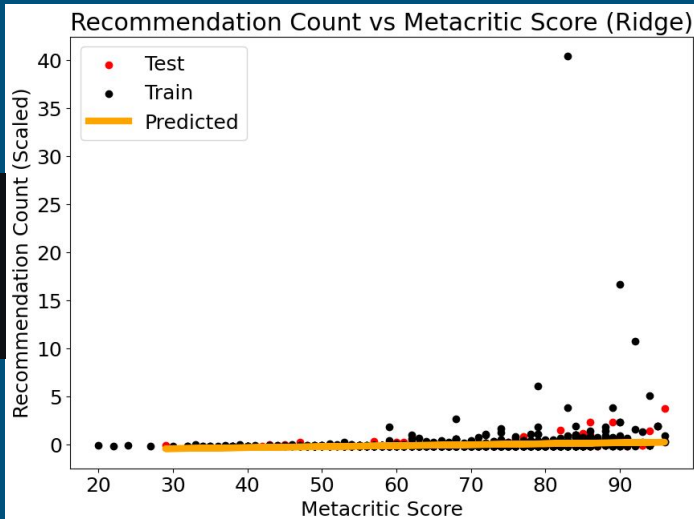
For both the lasso and ridge variations of linear regression, it can be seen that the results were extremely poor, due to the R^2 values being less than 0.1.

```
Training Data:
Mean squared error: 1.211018076554341
R^2: 0.01150468749732847

Testing Data:
Mean squared error: 0.09406610910353298
R^2: 0.06206191891962054
```

```
Training Data:
Mean squared error: 1.2190094587744245
R^2: 0.0049817098325808384

Testing Data:
Mean squared error: 0.0969146563135547
R^2: 0.03365890608644395
```



Discussion, Conclusion, and Future Work

Overall, our comparison between Metacritic score and user recommendation count did not work, and our comparison between genres was only semi-successful.

In the future, we would like to continue experimenting with Adaboost and other ensemble methods as they seemed promising. It also seems like a neural network might fit this dataset better.

We would also like to explore the question “Is there any correlation between the Metacritic scores and the recommendation counts within genres?”