Final Project Presentation

Dane Polchin • 05.23.2025

Dataset Overview

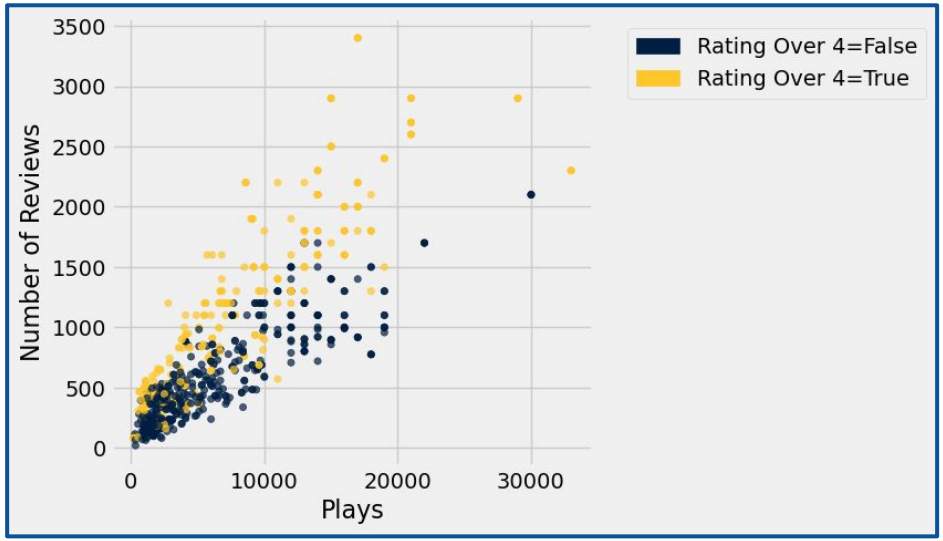
- Video game statistics - Dataset found on Kaggle
- Originally gathered from scraping a website called Backloggd
  - Allows users to see what games they've played and for how long
  - Also has ratings, reviews, play counts, etc.

Variables

Initially looked at total number of plays vs. number of reviews for different video games:

Also comparing whether or not a game has a rating of 4 stars or above (out of 5 stars).

This is the yellow (rating above or equal to 4) and blue (rating below 4) dots.



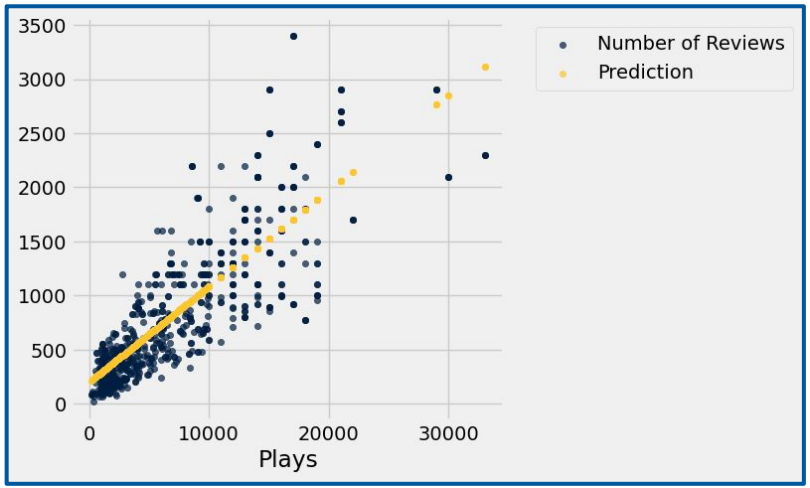
Variables

95% Correlation Confidence Interval (rounded):
[0.79, 0.84]
Since 0 is not in the interval, there is likely a correlation

I used 12000 plays as the value of x for prediction
99% Prediction Confidence Interval (rounded):
[1193, 1337] (predicted reviews)

Potential bias = Users on Backloggd are predominantly frequent gamers

Linear Regression Line:



Features and Classifier

Features were the same as numerical variables (Number of plays vs. Number of reviews)

Used the **rating** as a classifier (as a measure of whether a game was "good" or not):

- Rating above or equal to 4 = Is a "good" game
- Rating below 4 = Is not a "good" game

Used three nearest-neighbors to classify prediction points:

- If more neighbors were rated well, the prediction would also be rated well
- If more neighbors were rated poorly, the prediction would also be rated poorly

Accuracy of classifier after splitting the dataset into a test and train set and testing predictions for the points in the test set: **~0.84** after standardizing units (0.82 w/o standardizing) or **84% accuracy**

To **improve** my classifier, I tried increasing the number of nearest neighbors used. I also tried looking at several other features, but none of these ended up creating a more accurate classifier than my initial one.

Conclusion

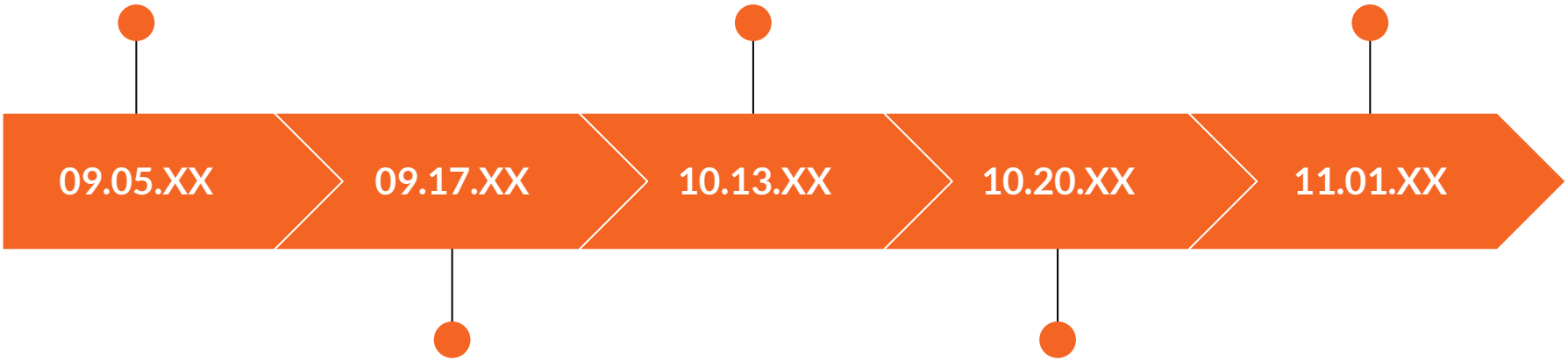
- Wanted to see if plays, reviews, and rating are correlated
- Learned dataset may not be the best for this
- Learned a lot of high-rated games have fewer plays and more reviews relative to others
  - Shows subtle pattern in dataset
- Not great for prediction and classification
  - 84% accuracy = almost 1/5 of predictions would be misclassified
  - Not practical – Probably easier to just look up the rating of a game

Schedule

Lorem ipsum dolor sit amet, consectetur adipiscing elit

Lorem ipsum dolor sit amet, consectetur adipiscing elit

Lorem ipsum dolor sit amet, consectetur adipiscing elit



Lorem ipsum dolor sit amet, consectetur adipiscing elit

Lorem ipsum dolor sit amet, consectetur adipiscing elit

Next steps

Assignment 1

Lorem ipsum dolor sit amet
Consectetur adipiscing elit, sed do eiusmod tempor

Assignment 2

Lorem ipsum dolor sit amet

Assignment 3

Consectetur adipiscing elit, sed do eiusmod tempor

Goals for next meeting

1. Lorem ipsum dolor sit amet, consectetur adipiscing elit
2. Sed do eiusmod tempor incididunt ut labore
3. Ut enim ad minim veniam, quis nostrud exercitation