PREDICTING STEAM GAME PLAY TIME

Dominika Jones, General Assembly

TABLE OF CONTENTS

STEAM®

LIBRARY UPDATE

Introduction

Steam platform is currently the largest PC game online distributor in the world and has accumulated a vast amount of user and game data.

One metric is particularly important to the developers who plan the game's support cycle - expected play time.

01

PROBLEM STATEMENT

Problem Statement

Customer engagement is a measurement of a user's response to the product. Average playtime metric is a major factor influencing the player's buying decision. By predicting the average total playtime through various regression models, I hope to improve user experience and customer engagement.

02
DATA

DATA ACQUISITION

Steam Web API

- Public access to player and game data.

Steam Spy API

- 3rd party site that aggregates statistics on Steam games.-

Data Issues Along the Way

- Data gaps due to users setting accounts to private.
- Distribution of playtimes is very skewed.
- Difficult to choose a good cost function.
- Daily issues with steamspy API

Hi Sergey

My significant other (CCed) is attempting to use your steamspy API to get some initial steam data for a data science class. She was trying to reproduce the code written here but is having issues. She is trying to use the request all type to retrieve a bunch of data but now is getting empty jsons. It was not defined in the API document but does an empty json entry mean that we attempted to poll the site too quickly. It's extra strange because appdetail requests always return data so the behavior is not consistent. If we do start to get empty APIs how long do we need to abstain from pulling data from the site before we can retrieve data gain?

STEAMTIME.INFO

840 hrs →


Zig6
STEAM ID: 76561197969355314
Offline
0h
0 days / 0h per year
WASTED!
Games
58
Friends
16
years ago created
Open in Steam
View Profile


PLAYERUNKNOWN'S
BATTLEGROUNDS

Feature Processing

One-hot encoded

CATEGORY

GENRE

DEVELOPERS

Raw numerical input

PRICE

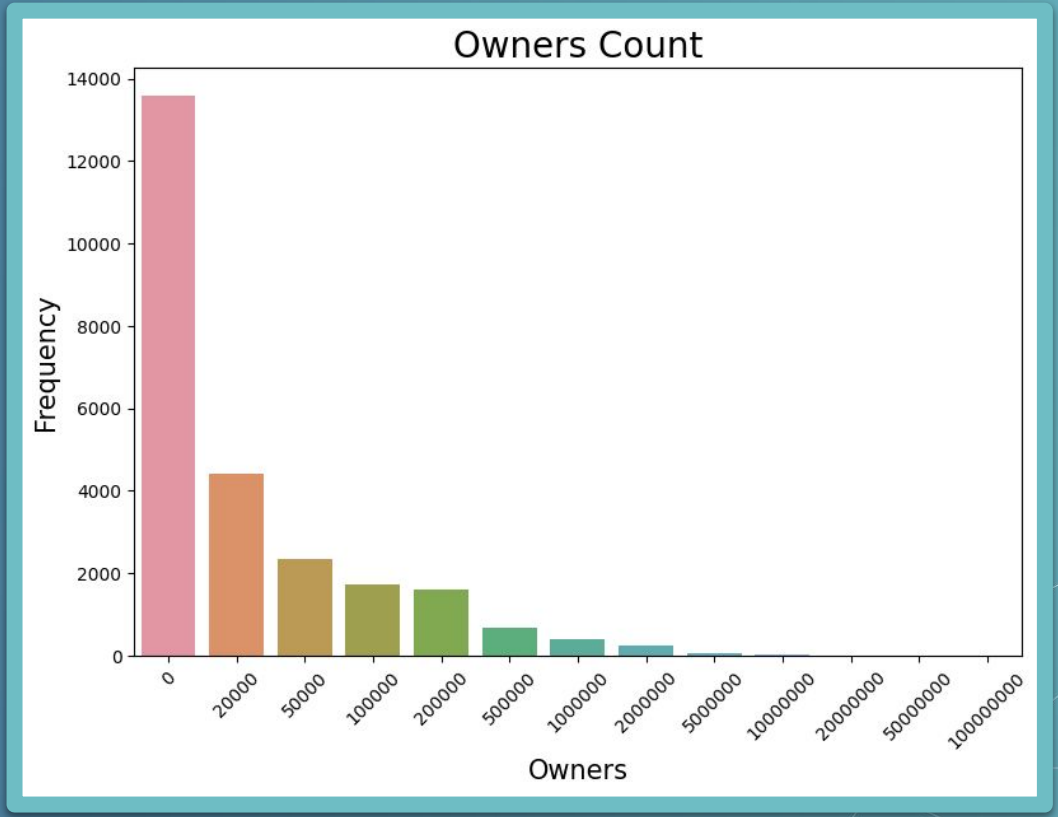
RATINGS

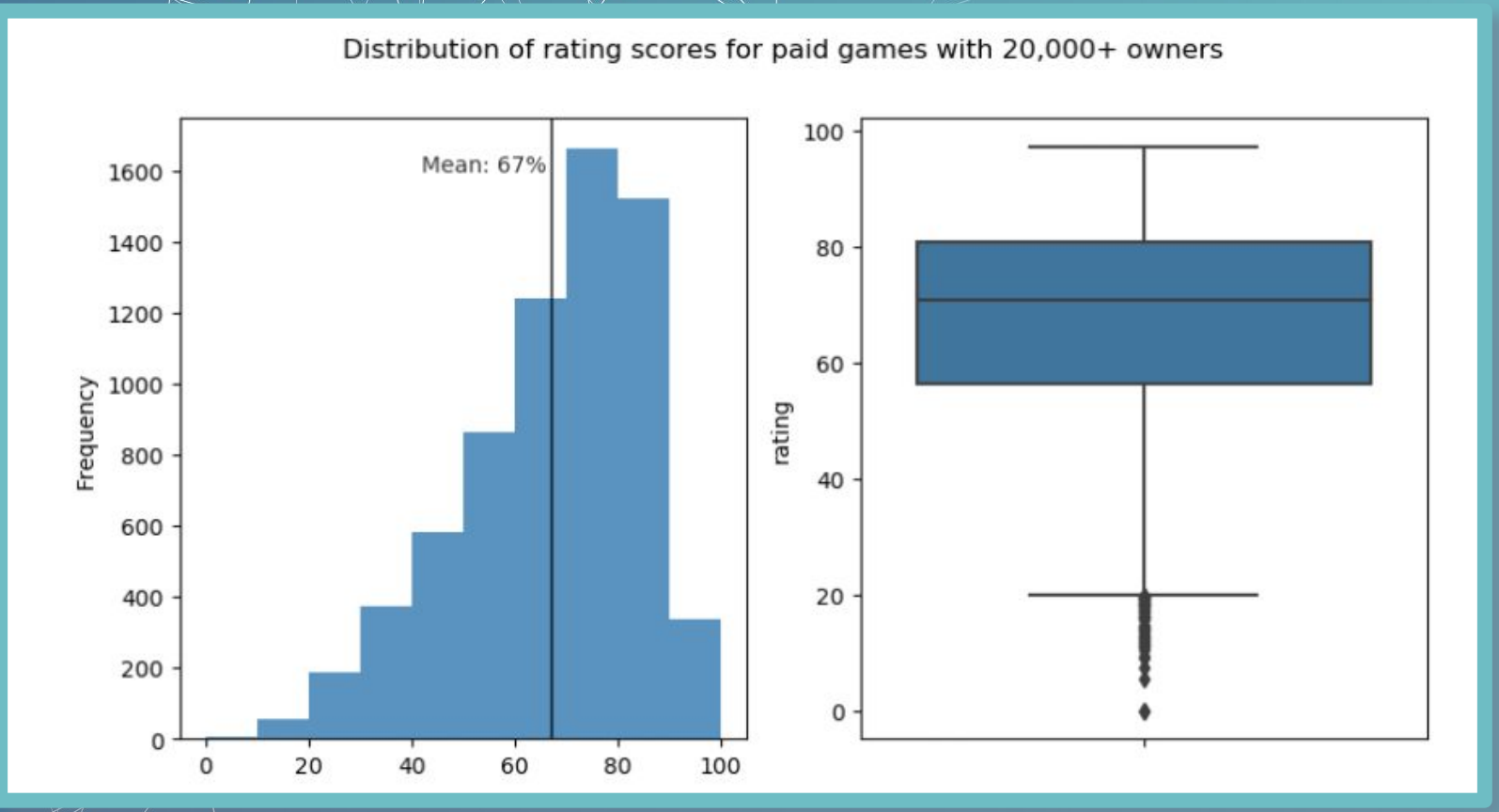
OWNERS

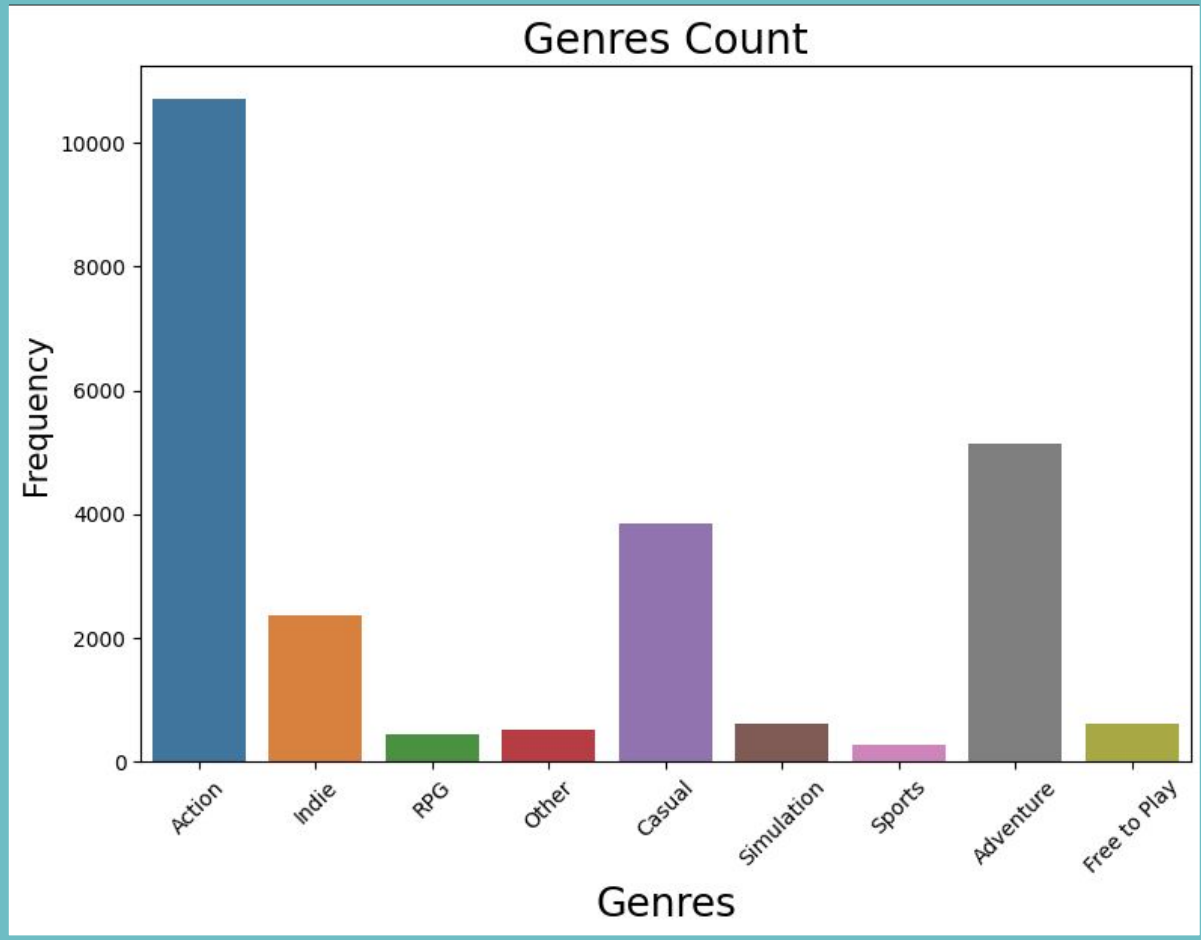
03

Exploratory Data Analysis

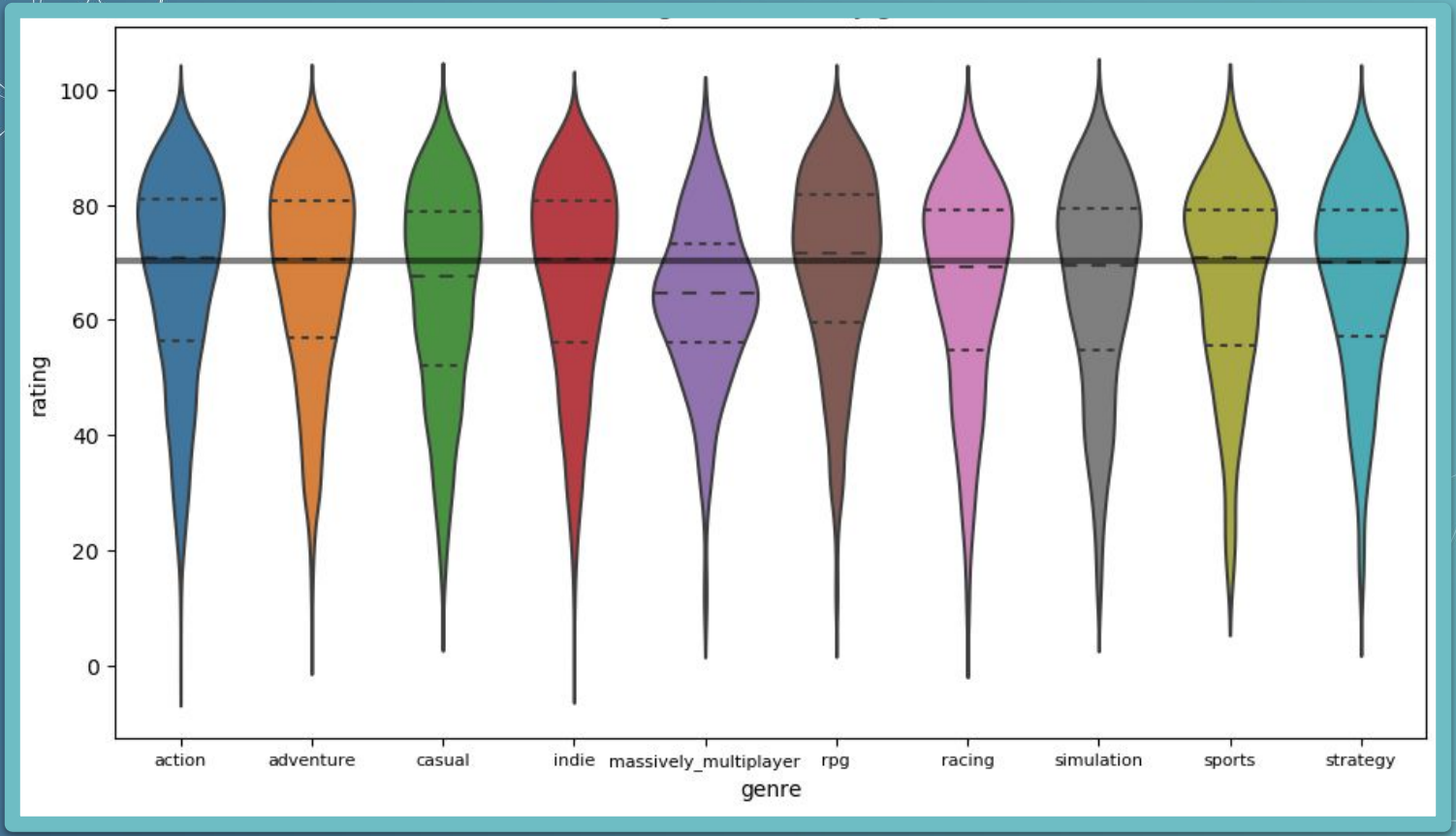
OWNERS AND RATINGS

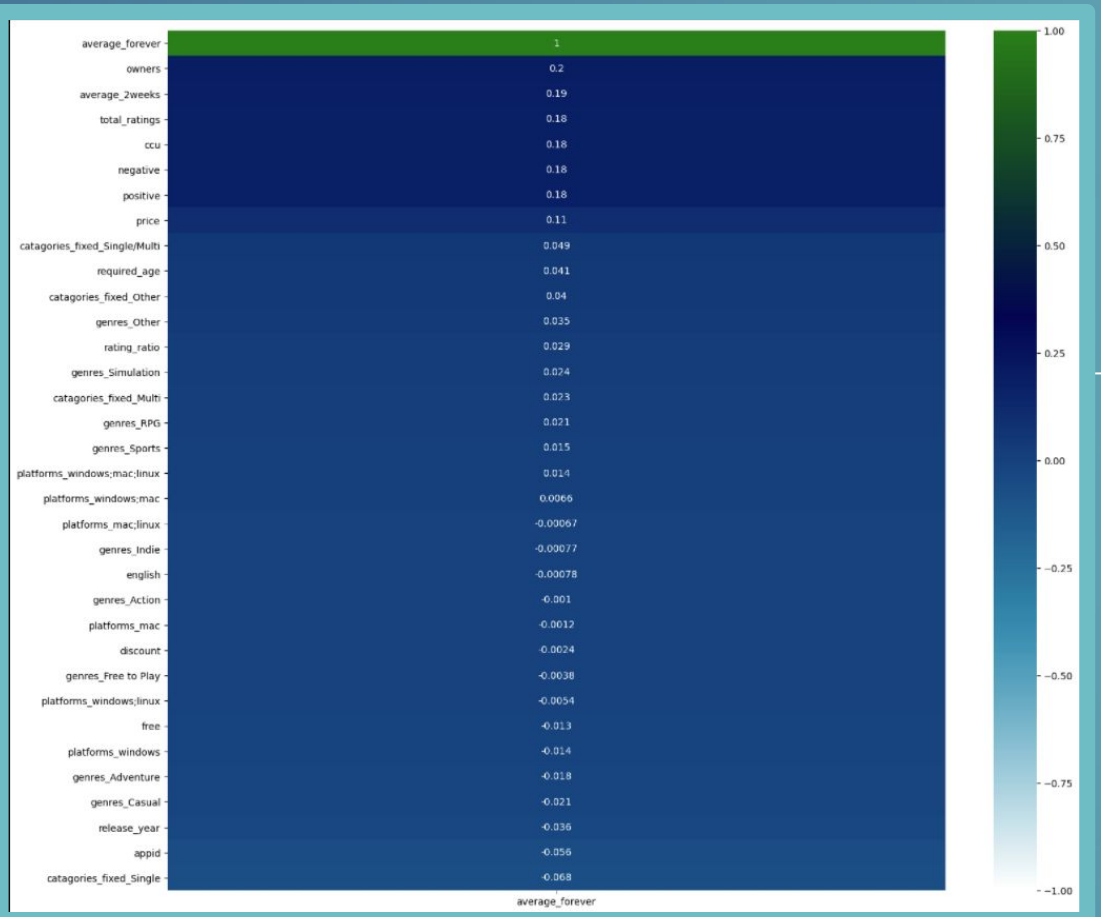



Distribution of rating scores for paid games with 20,000+ owners
Mean: 67%
Frequency
0
200
400
600
800
1000
1200
1400
1600
0
20
40
60
80
100
rating
0
20
40
60
80
100


Genres Count
Frequency
10000
8000
6000
4000
2000
0
Action
Indie
RPG
Other
Casual
Simulation
Sports
Adventure
Free to Play
Genres

Rating Distribution by Genre



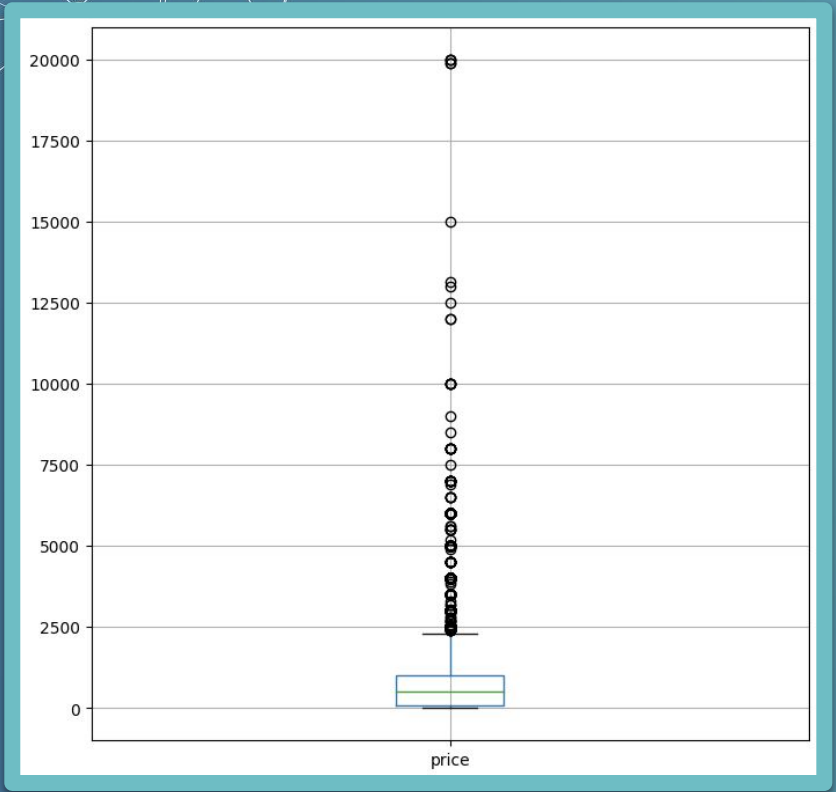


Correlations

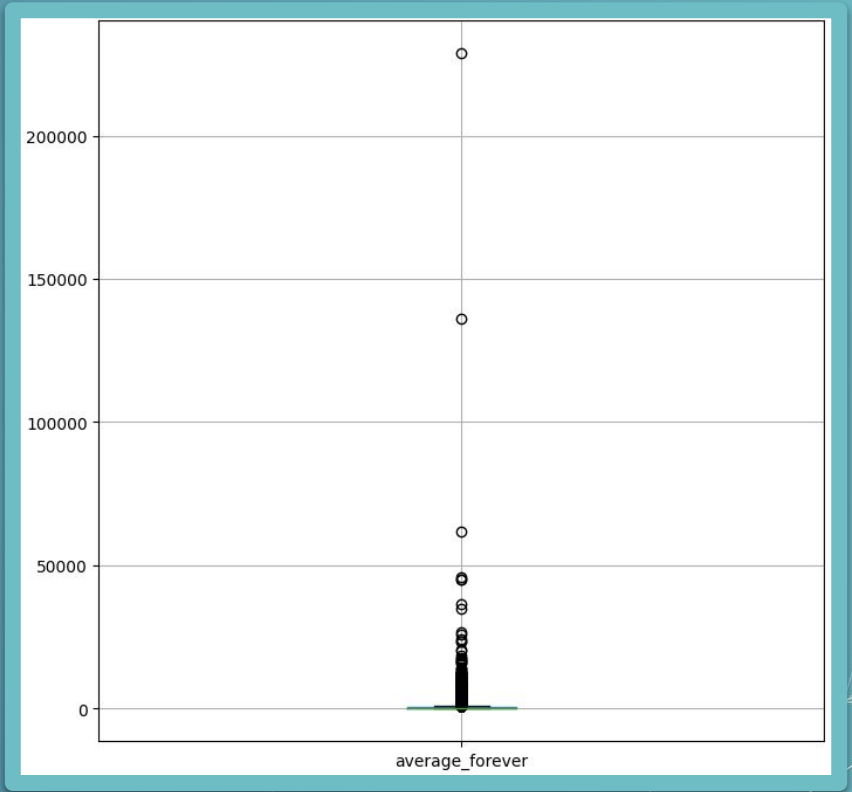
Price, Owners and Ratings have a positive correlation on average playtime (average_forever)

Release year surprisingly has a negative effect or little correlation on the average playtime.

Price Distribution



Playtime Distribution



04
MODELS

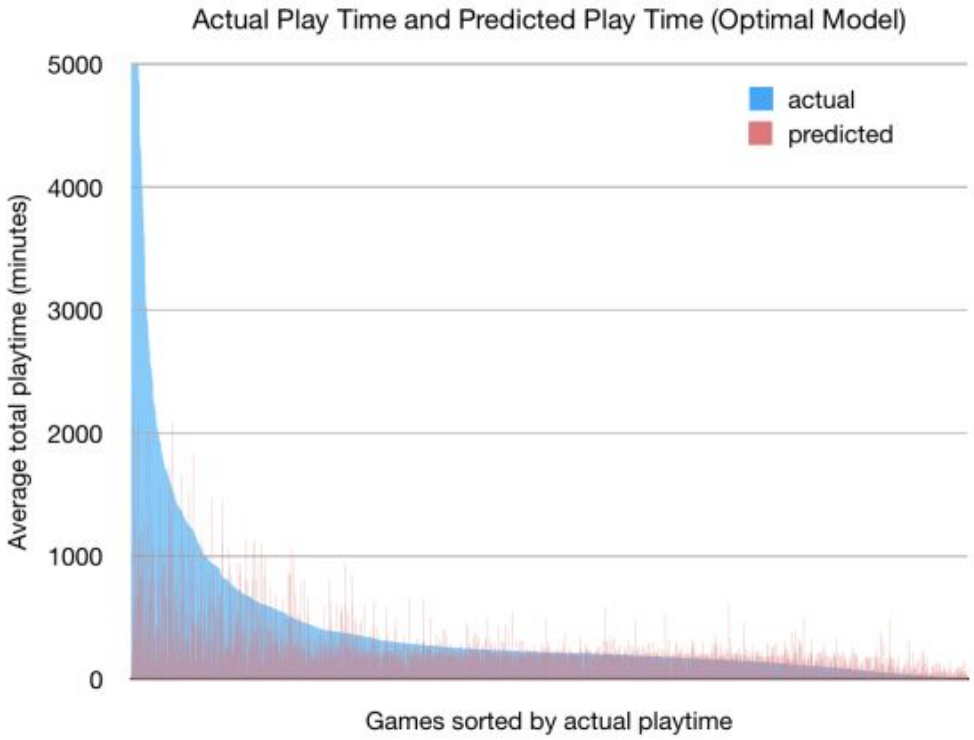
THE MODELS

- Linear Regression
- Random Forest Regressor
- Gradient Boosting Tree
- GridSearchCV

MODEL RESULTS

	MSE Train/Test	RMSE Train/Test	R2	ACC. Score
Linear Regression	2.5/3.3	1.5/1.8	0.1	7%
Random Forest Regression	0.26/2.05	0.5/1.4	0.9	33%
Gradient Boosting	2.5/2.6	1.6/1.7	0.14	14%

Predicted Play Time



05

CONCLUSION

CONCLUSION AND FUTURE WORK

- Positive influence features:
  - Positive Ratings, Number of Owners, Price
- Negative influence features:
  - Total Ratings, Category, Genre
- The model does not capture the skewed distribution of playtime.
  - Underestimates the higher end
  - Overestimates the lower end
- More informative features
- More flexible model (tuning gradient boost tree)

THANKS

RESOURCES

- http://store.steampowered.com/
- https://developer.valvesoftware.com/wiki/Steam_Web_API
- https://steamspy.com/about
- https://steamtime.info/