Cave Johnson Approved: Prediction of Steam User Reviews from Metacritic Data

DSI-11 Capstone Project - Adam Cohen

Steam and the Games Market: At a Glance

VIDEO GAMES MARKET:

- $33B in sales/distribution revenue in 2020
- $1.1B in e-sports revenue in 2020
- Consistently shatters sales records for entertainment property opening gross values.

STEAM:

- 75% Market Share
- $4B in Annual Revenue
- 20MM concurrent users in Q1 2020.
- Publisher, distribution platform, and hardware developer. Best-in-class in almost all consumer platform features.

Problem Statement

We at Catalyst Consulting, as part of our full-spectrum services to developers, publishers, and investors, want to provide prediction services for user reviews on the distribution platform Steam. We'll use a Regression model, and based on a game's metacritic scores, an aggregate value of professional reviewers' opinions, we will predict Steam user scores.

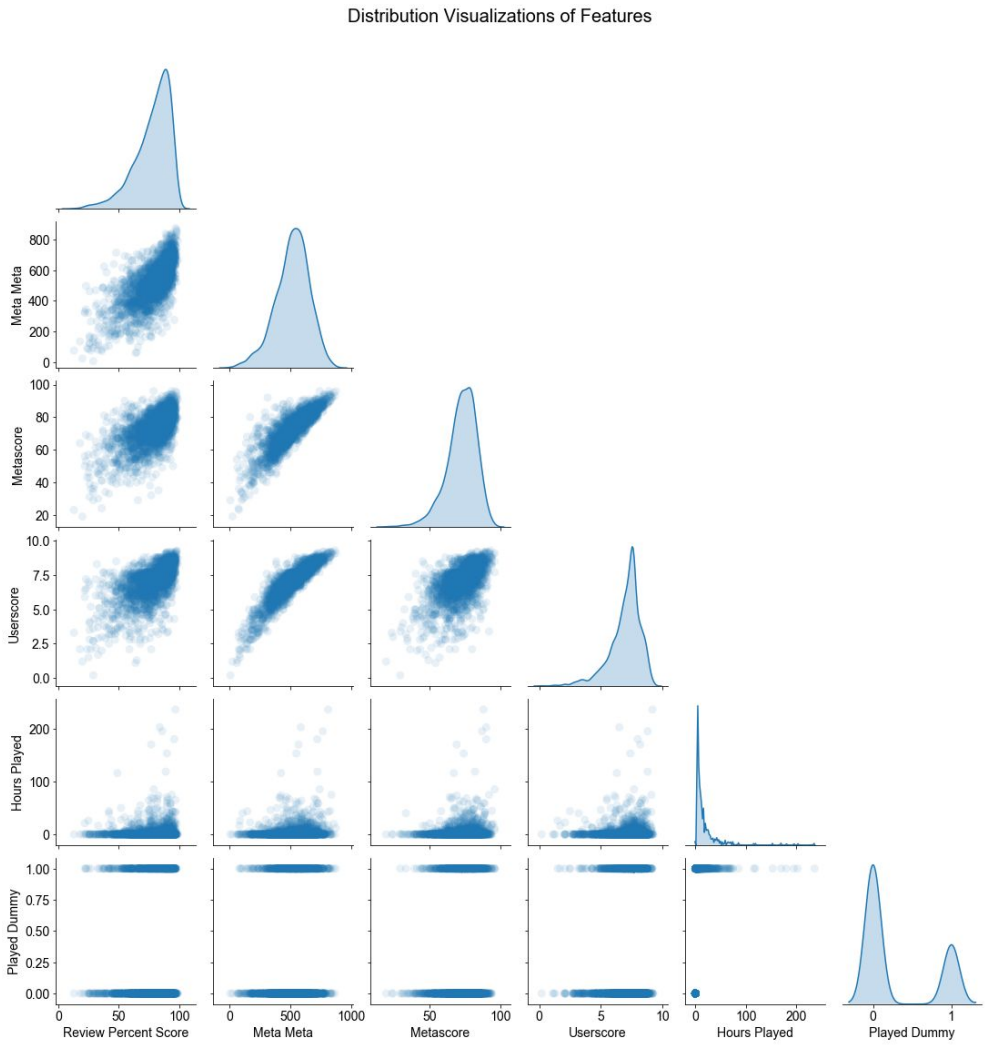
Our predictions will allow developers to appropriately budget for advertising and predict revenue models. Our metric of success will be RMSE value, due to its ease of interpretation in relation to our parameter of interest.

Data Collection & EDA

- Three datasets condensed to 3790 rows.
- 5% data loss during manual data collection from online databases.
- Split 60:40 Training:Testing along whether or not the game was listed on Steam
- Significantly left-skewed

EDA

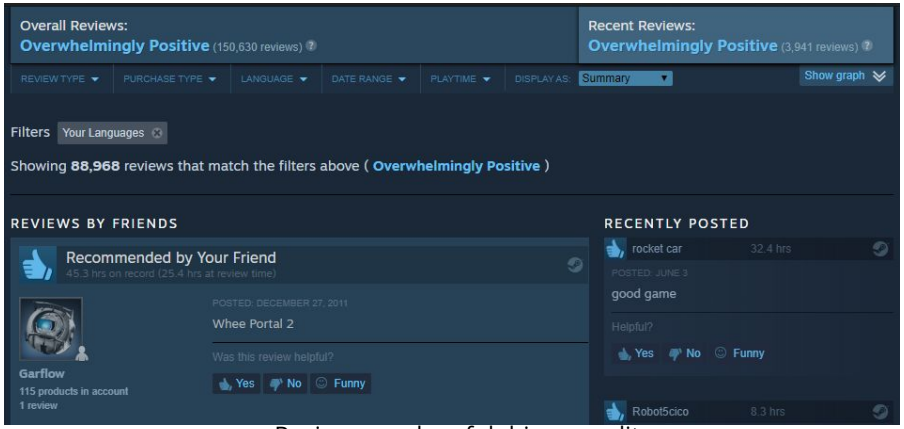
- Linear Trends in data
- “Blob” Shapes between our parameter of interest (Review Percent Score) and other metrics.
- Extreme low correlation between Hours Played and parameter of interest
- Almost all data left-skewed, time factors extremely right-skewed.



A Study in Review Percent Score

- Ratio of Positive:Negative reviews
- Does not account for number of reviews for a game
- *Giants: Citizen Kabuto* and *Portal 2* scored comparably. The former has 58 reviews, the latter has over 150,000
- Review Percent Score is a back-end metric

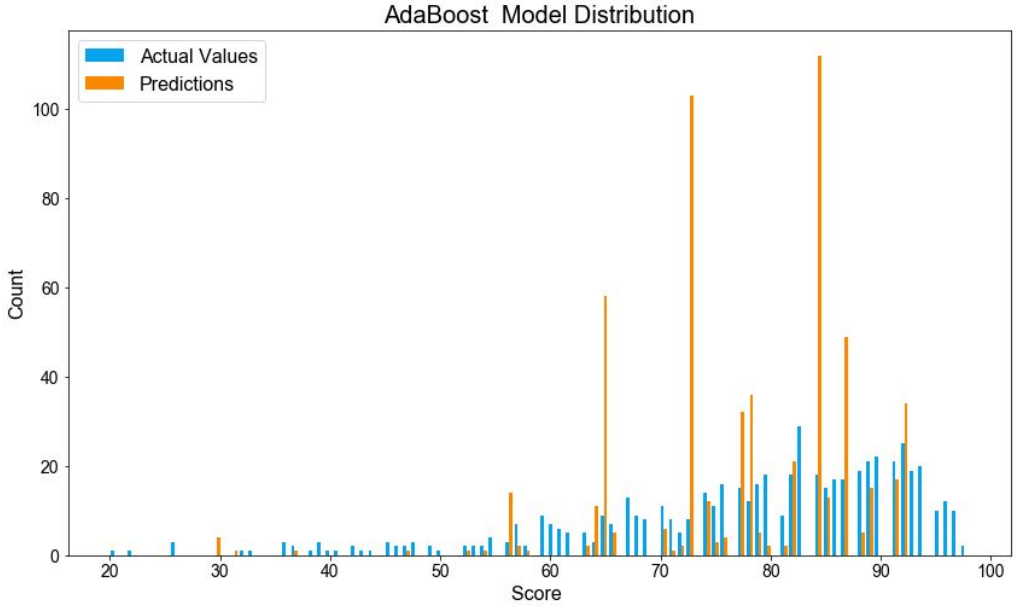
Cutting edge graphics (for the year 2000)



Reviews can be of dubious quality.

Modeling Data

- 5 Features used
- 5 models tried:
  - Dummy Regressor
  - Linear Regressor
  - AdaBoost Regressor
  - GradientBoost Regressor
  - Random Forest Regressor
- RMSE to beat was 14.7, target RMSE score was 5.
- Almost all models scored around 10.5.



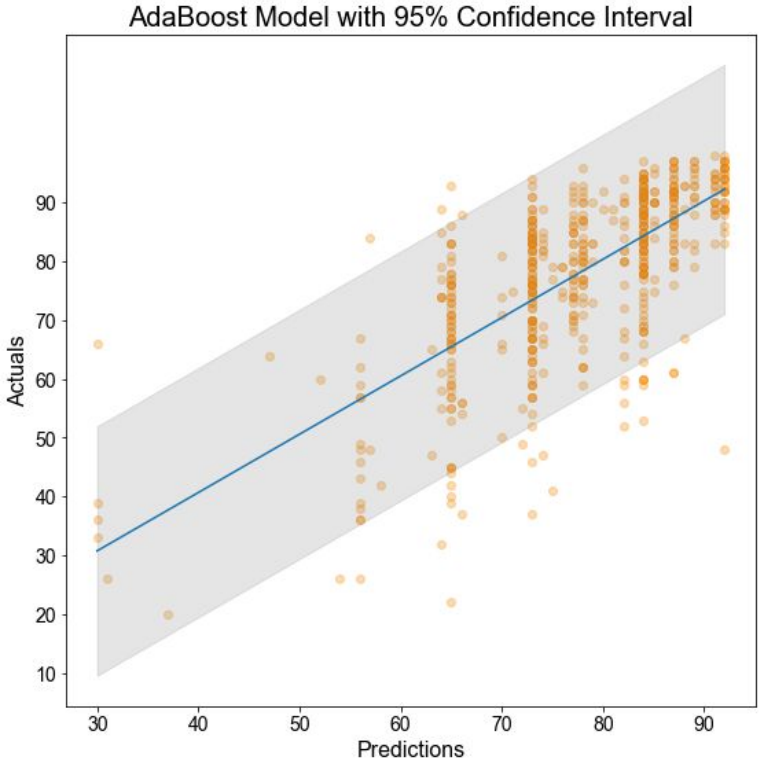
Despite good RMSE scores, some models were really bad predictors.

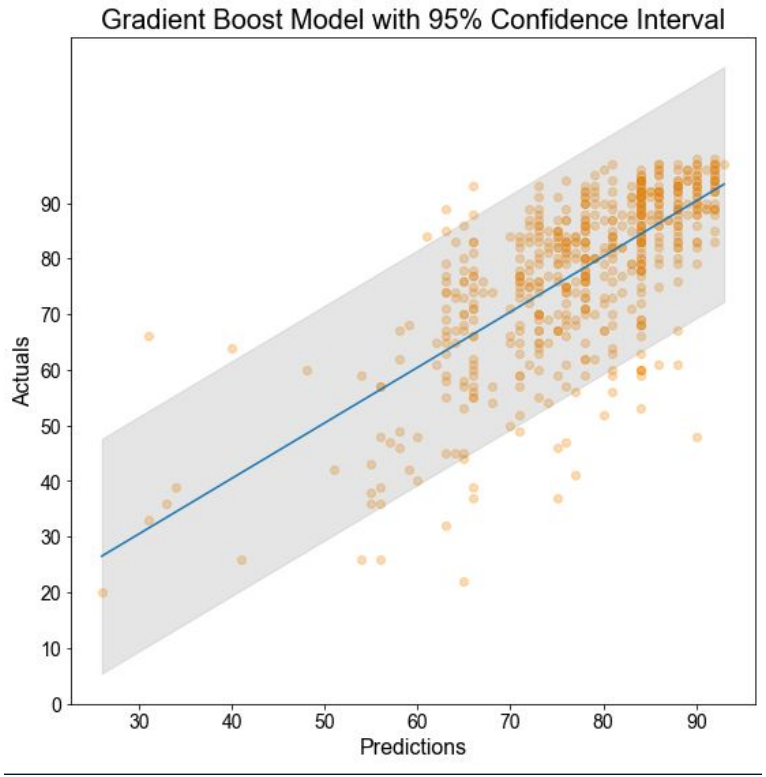
Model Selection

- Almost all models aside from Dummy predicted a 21-point window
- Variation down to tenths of a point, negligible value in terms of our parameter of interest
- Train/Test scores similar between all models: visible underfit.
- Selecting Linear Regressor

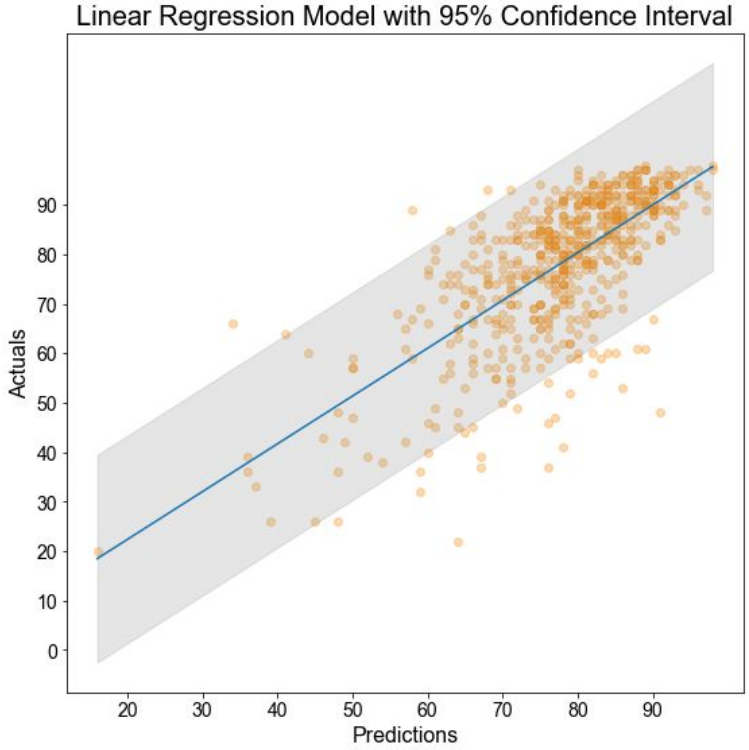
Model Name	Train	Test	RMSE
Dummy Regressor	0	-0	14.7
Linear Regressor	.47	.49	10.47
AdaBoost Regressor	.50	.48	10.62
GradientBoost Regressor	.52	.49	10.55
Random Forest Regressor	.51	.48	10.54

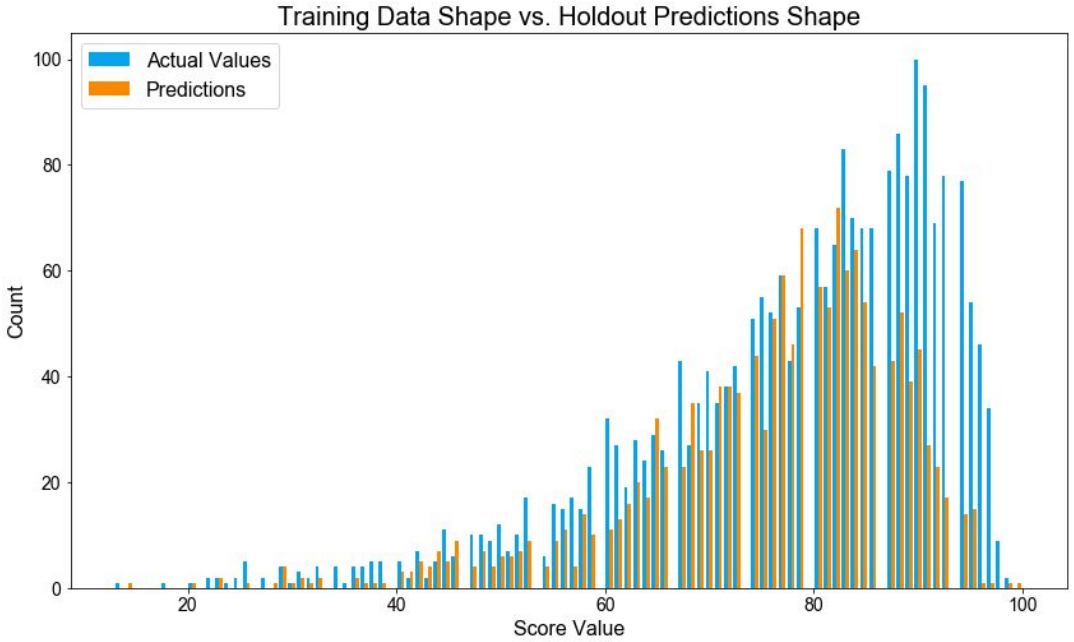
Model Selection Pt. 2





Model Evaluation (Pt. 1)





Model Evaluation (Pt. 2)

- Predictions were largely accurate: high-scoring games sold millions of copies
- Top three games considered genre-defining
- Bottom five games were commercial flops.
- Who even remembers that movie *R.I.P.D*?

Game Title (Top 5)	Review % Score
The Orange Box	100
Baldur's Gate II	99
Warcraft III: Reign of Chaos	97
The Operative: No One Lives Forever	96
Homeworld: Cataclysm	95

Game Title (Bottom 5)	Review % Score
Ride to Hell: Retribution	14
Infestation: Survivor Stories	20
R.I.P.D: The Game	23
Leisure Suit Larry: Box Office Bust	23
Terrawars: New York Invasion	26

Conclusions

- Model not a successful predictor of user sentiment
- Review Percent Score not a successful predictor of commercial success.
- Additional data required in pre-launch window to more successfully gauge user sentiment

	metascore	name	console	userscore	release_date	review_percent_score	hours_played	meta_meta	played_dummy
475	91	Call of Duty	PC	8.5	2003-10-29	93.0	2.966667	773.5	1
476	86	Call of Duty 2	PC	8.4	2005-10-25	91.0	5.187500	722.4	1
477	92	Call of Duty 4: Modern Warfare	PC	8.5	2007-11-05	92.0	0.000000	782.0	0
479	81	Call of Duty: Black Ops	PC	5.3	2010-11-09	89.0	0.000000	429.3	0
481	74	Call of Duty: Black Ops II	PC	4.3	2012-11-12	85.0	0.000000	318.2	0
482	73	Call of Duty: Black Ops III	PC	3.0	2015-11-05	69.0	0.000000	219.0	0
483	68	Call of Duty: Ghosts	PC	2.1	2013-11-04	59.0	0.000000	142.8	0
484	73	Call of Duty: Infinite Warfare	PC	3.4	2016-11-04	46.0	0.000000	248.2	0
485	86	Call of Duty: Modern Warfare 2	PC	4.3	2009-11-10	92.0	0.000000	369.8	0
487	87	Call of Duty: United Offensive	PC	8.2	2004-09-14	86.0	0.000000	713.4	0
488	73	Call of Duty: WWII	PC	3.3	2017-11-03	60.0	0.000000	240.9	0
489	83	Call of Duty: World at War	PC	7.5	2008-11-10	93.0	0.000000	622.5	0

By 2015 these titles had made $11 Billion combined.

	metascore	name	console	userscore	release_date	review_percent_score	hours_played	meta_meta	played_dummy
1353	93	Grand Theft Auto III	PC	8.3	2002-05-20	86.0	3.661765	771.9	1
1354	90	Grand Theft Auto IV	PC	6.7	2008-12-02	69.0	20.735841	603.0	1
1355	96	Grand Theft Auto V	PC	7.7	2015-04-14	79.0	85.340892	739.2	1
1356	93	Grand Theft Auto: San Andreas	PC	8.9	2005-06-07	90.0	0.000000	827.7	0
1357	94	Grand Theft Auto: Vice City	PC	8.8	2003-05-12	92.0	0.000000	827.2	0

One of these titles sold 90 million units. It is not the top-scoring one.

REFERENCES

https://www.statista.com/statistics/292460/video-game-consumer-market-value-worldwide-platform/#:~:text=This%20statistic%20shows%20the%20value,11.2%20billion%20a%20year%20earlier.

https://newzoo.com/insights/trend-reports/newzoo-global-esports-market-report-2020-light-version/#:~:text=Highlights%3A,from%20%24950.6%20million%20in%202019.

https://www.titlemax.com/discovery-center/money-finance/the-25-highest-grossing-media-franchises-of-all-time/

https://www.statista.com/statistics/308330/number-stream-users/#:~:text=Steam%20added%20approximately%2062.76%20million,steadily%20fallen%20in%20recent%20years.

https://www.gamesindustry.biz/articles/2018-04-09-gta-v-is-the-most-profitable-entertainment-product-of-all-time#:~:text=The%20site%20reports%20that%20with,media%20title%20of%20all%20time%22.