# Final Project

Wilson Huang

11/16/2020

---

## Summary

This report mainly addresses three questions. The first is whether different game genre receive different community review. The second is whether the original price affects community review and whether such effect varies across other game features. The last is finding other factors that affect odds of receiving positive review. Multiple logistic regression is used to model how different games features affect community review since the response variable is collapsed into binary format. And stepwise selection from AIC perspective and anova chi-square test are utilized to determine variables in the model.

This study reveals that RPG games are likely to have higher odds of positive review than Action games. Under 90% confidence level, the estimated odds ratio is within [1.04, 2.06]. Also, original_price does affect the odds and such effect varies across different game genres. Among them, Simulation and RPG games are seeing larger decrease in odds. In addition, being offered free is likely to undermine community review and different languages supported like Russian are likely to impose different effects on community review.

## Introduction

STEAM is a video game digital distribution service by Valve. It contains all kinds of community services and among them, the most important one is the review on games. By studying these reviews, manufacturers can know what to improve, generate sales overview based on pre-sale reviews, and decide DLC or music tracks to make profits. In addition, users can learn whether it is worthwhile to purchase specific games. One important benchmark for classifying community review is the proportion of positive review. Intuitively, assuming the game needs to be purchased, if it is favored by a larger proportion of users, it is likely to bring more positive review and get popular with users.

## Data

### Data Preprocessing

The original dataset contains over 30k observations about STEAM games before June 2019. Around 2000 observations have all the column values missing since they are mostly DLC (Downloadable Contents), duplicated bundle, and music tracks. Since the goal here is to explore factors affecting the odds of positive community review on games, these rows are simply dropped. Also, there are several outliers in achievements and original_price. For example, Casino Simulator has 5000 duplicated achievements and Euro Truck Simulator Bundle is priced at $3000. These observations are also dropped in order to eliminate unwanted bias in regression. In addition, all the games collected support English so this variable is dropped. Furthermore,

the original categories of response variable, review, include several levels such as very negative (13% of users vote positive) and mixed (61% of users vote positive). Given that STEAM labels labels "Mixed" to any games whose positive review proportion is within [30%, 70%], most of the game reviews are clustered in group "Mixed" and the number of observations in different levels is highly unbalanced. In order to reduce the imbalance, review categories are collpased into two: positive = 1 and negative = 0 using the benchmark 50% and the number becomes closer to each other.

Cleaned data contains predictors such as game genre, original_price, languages supported, and cloud services.

## Explanatory Data Analysis

To understand the baseline probability of positive review, a table with probability of different levels in review is constructed. The result shows that the baseline probability of positive review is around 88%. Conditional probability tables are also constructed.

### review vs cloud

There might be association between review and cloud since as a game switch from not having cloud backups to having cloud backups, the conditional probability of having positive review increases. The chi-square independent test afterwards verifies such association. However, it still needs to be examined in the final regression model. (Please refer to the appendix for detailed tables)

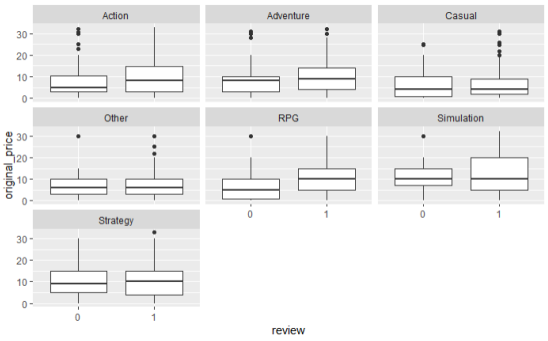
### review vs other predictors

Fro other categorical variables, there appears to be association with the response variable in game genre, mature_content, free, controller, Spanish, and Portuguese. Among those game genres, Adventure and RPG games tend to have higher probability of receiving positive review. Also, being offered free might decrease the odds of positive review. For continuous variable, original_price, achievements, and totalLan display possible association with the response variable. As original price increases, the odds of positive review might increase. More languages supported might increase the odds of positive review.

However, these association are not certain and need further examination in the model part.

### interaction effect

It appears that the association between original price and review differs by game genre. It corresponds to intuition since game genres like Adventure are the top choice for 3A games. Changes in original price for different games might bring different result.




Also, there is possible interaction between totalLan and genre. intuitively, more languages supported in games such as Strategy and RPG might make the game more popular since such game genre requires heavy

reading and users might be more satisfied if they do not need to find third-party translation mod to facilitate gaming. To generate a more solid conclusion about interaction terms, these potential interaction terms are examined in the model selection.

# Model

## Model Selection

Since I am looking for association and inference about probability, multiple logistic regression model is appropriate here. The model selection follows two methodologies: AIC stepwise selection and ANOVA chi-square test.

The results drop variables such as achievements and online. To make sure that they are indeed negligible, anova chi-square test is performed and I incrementally test each element against the baseline AIC model.

Those individual predictors dropped by the stepwise selection such as achievements (0.1629) and mature_content have poor performance (0.8159) on ANOVA test (p-value $> 0.05$), which indicates they can be excluded from the model. Even though total number of languages vs genre appeared significant in EDA, assessment found the term negligible (p-value = 0.478). Original_price vs genre is indeed significant with p-value = 0.0097. Therefore, it is included.

## Final Model

According to VIF, other than interaction term, there is suspicious multicollinearity if totalLan is included (4.169). Therefore, totalLan is dropped since the variables already include each individual language.

Here is the final logistic model:

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 genre + \beta_2 original\_price + \beta_3 totalComments + \beta_4 free + \beta_5 discount + \beta_6 trading\_cards + \beta_7 controller$$

$$+ \beta_8 cloud + \beta_9 Russian + \beta_{10} Spanish + \beta_{11} German + \beta_{12} original\_price : genre$$

At 0.1 significance level, the significant predictors in the final model are genre, original_price, free, discount, trading_cards, controller, cloud, Russian, Spanish, German, and the interaction term original_price:genre.

|   | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| **(Intercept)** | 1.844 | 0.09998 | 18.45 | 5.359e-76 |
| **genreAdventure** | 0.2785 | 0.06527 | 4.266 | 1.986e-05 |
| **genreCasual** | 0.2289 | 0.07711 | 2.968 | 0.002994 |
| **genreOther** | 0.1036 | 0.09661 | 1.072 | 0.2838 |
| **genreRPG** | 0.2786 | 0.1829 | 1.524 | 0.1276 |
| **genreSimulation** | -0.1955 | 0.09103 | -2.148 | 0.03172 |
| **genreStrategy** | 0.02189 | 0.07565 | 0.2893 | 0.7724 |
| **original_price** | 5.211e-06 | 3.071e-05 | 0.1697 | 0.8653 |
| **totalComments** | 8.587e-06 | 5.408e-06 | 1.588 | 0.1123 |
| **freeTrue** | -0.1522 | 0.063 | -2.416 | 0.01569 |
| **discountLow** | 0.1462 | 0.148 | 0.988 | 0.3232 |
| **discountZero** | 0.314 | 0.09639 | 3.257 | 0.001124 |
| **trading_cardsTrue** | -0.1993 | 0.04879 | -4.084 | 4.421e-05 |
| **controllerTrue** | 0.2785 | 0.04951 | 5.625 | 1.853e-08 |
| **cloudTrue** | 0.358 | 0.05398 | 6.632 | 3.309e-11 |
| **RussianTrue** | -0.2046 | 0.05894 | -3.471 | 0.000519 |
| **SpanishTrue** | 0.3206 | 0.07904 | 4.057 | 4.975e-05 |
| **GermanTrue** | -0.1691 | 0.0769 | -2.2 | 0.02784 |
| **original_price:genreAdventure** | 0.001565 | 0.001149 | 1.362 | 0.1733 |
| **original_price:genreCasual** | 0.00185 | 0.002029 | 0.9117 | 0.3619 |
| **original_price:genreOther** | 0.003432 | 0.002989 | 1.148 | 0.2509 |
| **original_price:genreRPG** | 0.01367 | 0.01098 | 1.245 | 0.2132 |
| **original_price:genreSimulation** | -0.0007954 | 0.002708 | -0.2937 | 0.769 |
| **original_price:genreStrategy** | 0.006187 | 0.003051 | 2.028 | 0.04259 |

Holding other factors constant and under 90% confidence level, compared with Action games, odds of positive review is expected to be

1) 23% higher for Adventure games 2) 39% higher for Casual games 3) 45% higher for RPG games 4) 21% higher for Strategy games

one unit increase in original price is expected to increase the odds of positive review by 0.9%.

free games is expected to have 13% lower odds of positive review than charged games.

compared with not supporting, odds of positive review is expected to be

1) 30% higher for controller 2) 20% lower for Russian 3) 37% higher for Spanish 4) 16% lower for German

Interaction: compared with Action games, one unit increase in original price is expected to

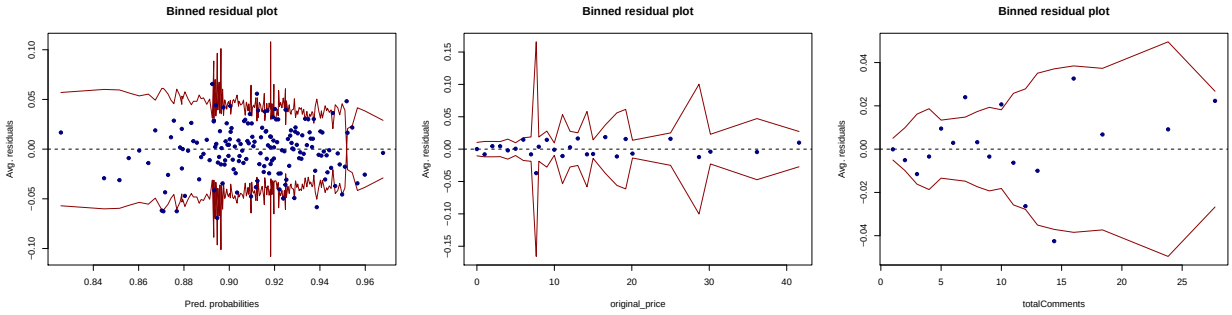
1) decrease by 1.57% for Simulation games 2) decrease by 1.59% for Strategy games

Therefore, in the context,

1. Game genre does affect the odds of positive review. RPG games are expected to have the highest odds which is 48% higher while Action games have the lowest odds. The odds ratio for RPG and Action games is expected to be within [1.04, 2.06] under 90% confidence level.
2. There is association between original price and community review and one unit increase in original price is expected to increase the odds of positive review by 0.9%. Such association differs by game genre. Compared with Action games, one unit increase in original price of Simulation games is expected to decrease the odds by 1.57% while for Strategy games, the odds is expected to decrease by 1.59%.
3. Being offered free (13% lower) is likely to undermine people's review about the game. Supporting controller is likely to increase the odds of positive review by 30%. In addition, supporting Russian is likely to undermine the odds of positive review by 20%.

## Model Assessent

From the overall binned residual plot, there is no discernible pattern and more than 95% of data points are within SE bands. The same conclusion applies to individual binned residual plots for original_price and totalComments.



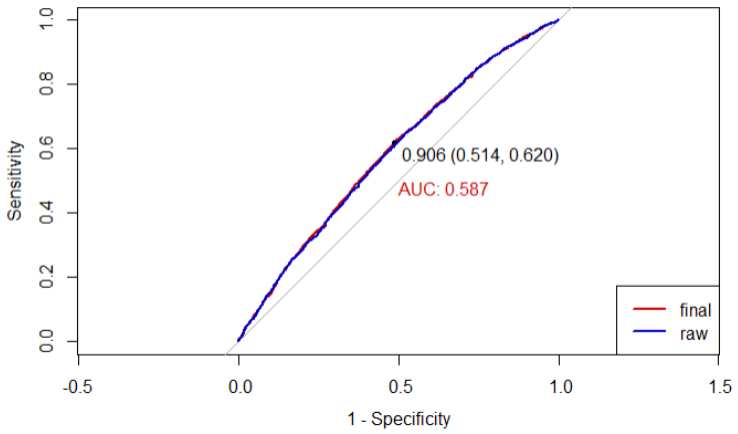

Also, there is no risk of multicollinearity since the vif results are now all under 5.

| | GVIF | Df | GVIF^(1/(2*Df)) | | | | |
|---|---|---|---|---|---|---|---|
| achievements | 1.109942 | 1 | 1.053538 | trading_cards | 1.201969 | 1 | 1.096343 |
| genre | 1.332534 | 6 | 1.024212 | online | 1.603545 | 1 | 1.266312 |
| mature_content | 1.027228 | 1 | 1.013523 | controller | 1.294788 | 1 | 1.137887 |
| original_price | 1.257759 | 1 | 1.121498 | cloud | 1.255992 | 1 | 1.120711 |
| totalComments | 1.030436 | 1 | 1.015104 | Chinese | 1.335080 | 1 | 1.155457 |
| free | 1.058344 | 1 | 1.028758 | Spanish | 3.114591 | 1 | 1.764820 |
| discount | 1.194168 | 2 | 1.045361 | Portuguese | 1.654335 | 1 | 1.286210 |
| multi_player | 1.708540 | 1 | 1.307111 | German | 2.848675 | 1 | 1.687802 |

### Model Validation

The accuracy is 0.682 and the sensitivity is 0.708.

### ROC curve




At optimal decision threshold, sensitivity increases to 0.514 and 1-specificity reduces to 0.620 and AUC is 0.587. The final model is slightly better than the raw model. Among all positive cases classified, 51.4 percent are indeed positive. For all negative cases classified, 62 percent are falsely classified. Compared to the benchmark of 0.5 (no diagnostic ability), AUC at 0.587 indicates some diagnostic ability but the performance is not ideal.

## Conclusion

This report reveals that game genre does affect the odds of positive review and Adventure and RPG games have the highest odds. Also, the association between original price and positive review differs by game genre. To be specific, Simulation and Strategy games see higher decrease in odds as price goes up since these game genres are used by small/medium games more. In addition, being offered free is expected to undermine people's review about the game, which is verified by Mtenga's research that people tend to value less on free goods. Furthermore, odds of positive review for games supporting Russian is generally lower maybe because Russia is known for being a region where average price is lower and piracy is rampant. Users holding copyrighted games give negative review to alarm producers about these issues.

There are mainly two limitations. One is that with almost every column value missing, much information about INDIE games (games produced by independent studios) is excluded. As a matter of fact, INDIE games are important component of STEAM so it is limited to large/3A games. Another limitation is that sometimes, negative review is given to online games whose server periodically shuts down. Also, people simply give negative review when steam restricts them from purchasing. Such phenomena is not modeled.

---