Sales Prediction On Steam Games

Apr 12, 2024

Team 87:
Xinmeng Wang (wang.xinm@northeastern.edu),
Wenbo Zhao (zhao.wenb@northeastern.edu)

Problem Statement and Background

Since the pandemic's spread in 2020, more people start getting used to in house entertainments, and video games become more popular in daily life. Steam, as a world leading online platform for the sale and distribution of PC games, has grown its market size by 60% from 2019 to 2020 and keeps an upward trend until now. In this project, we aim to explore the relationship between the total sale and the game details (tags, percentage of positive views, price, and time since the game's first release) that Steam presents on the game page, and try to train a model to predict the game sale based on these informations. With this prediction model, we may be able to help game companies to better arrange their game page and increase the potential sales in the future.

Introduction to Data

The data we used is scraping from Steam's World Best Selling Game (SWBG) page (https://store.steampowered.com/search/?filter=globaltopsellers&os=win). Based on the instructions posted by @DDDHL[1], we collected all the URLs from the SWBG page, but this instruction was posted in 2022 and Steam has changed its web design these years, so we have to make a huge modification based on @DDDHL's code. The information collected from the Best Selling List is not enough to build a predicted model, so we write a for-loop to iterate through each url to get more information from the game's detail page. Steam is famous for easy scraping and does not have a clear policy to ban web scrap, beyond that, there is no sensitive information being collected, so we believe our use for academic purposes is appropriate.

The features we are using are all collected from the Steam game details page:

'Current Price $': This field represents the normal price of each game in dollars. As opposed to its sale price, focusing only on the regular pricing of the games.

'Day_now': We collect the released dates for each game and convert these dates, which are hard to utilize in machine learning tools, into numerical data representing the totals days since the game was released to the date we collect the data, which is April 5, 2024.

'pct_positive_views': This information is directly collected from Steam, and indicates the percentage of positive views a game has received

'num_pop_tag': Steam shows up to 20 user-defined tags for each game, around 80% of the games in our dataset have exactly 20 tags. However, due to the highly skewed distribution, we decided to focus on the top 10% of the most frequent tags and count their occurrences across each game.

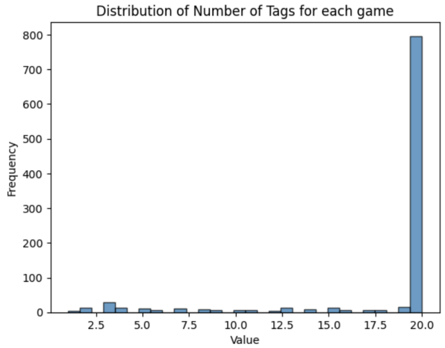
'total _reviews': this is our dependent variable. It's commonly known among Steam users that the total number of sales is confidential. However, to some extent the total number of reviews can serve as a proxy for sales. Clearly, using the total number of reviews as a substitute for actual sales data is not ideal, but it is the only feasible option given the absence of accessible sales data.

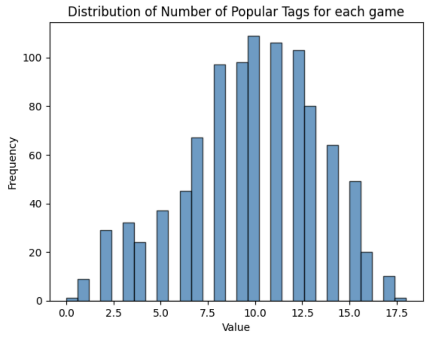
*Noticed: both 'pct_positive_views' and 'total _reviews' are based on the past 30 days.

While all our data is updated (collected at April.5.2024) and firsthand, we acknowledge that our data is biased. Steam lists 9,100 games on the SWBG page, but we have selected only the top 1,000 games to train our model. This sample size is too small to accurately catch the patterns across all games on steam Furthermore, given that global top selling games already have a good reputation, their number of sales is not affected by the information steam provides on the game page.

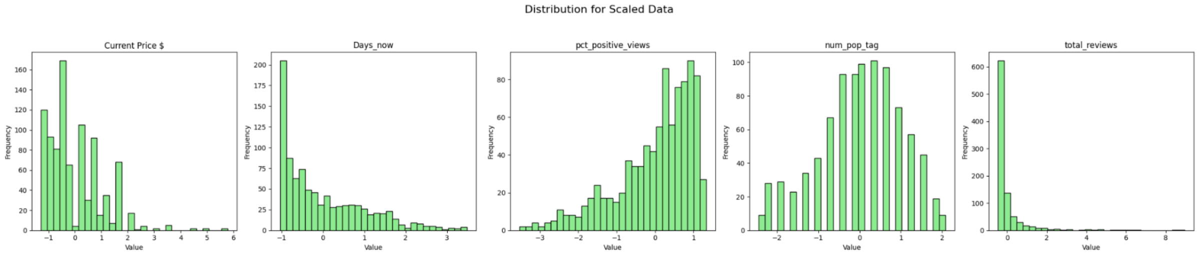
Data Science Approaches

For the data preprocessing, we primarily applied various functions to clean and organize the data. Initially, we addressed irrelevant text in the game names and release date columns using str.replace(), making sure only important information remained. Additionally, Steam's review data included both positive review percentage and total recent review counts, along with repeating text. We utilized the extract_data() function to isolate the essential data and create two new columns. Furthermore, we conducted a thorough analysis on the tag distribution among the games. While plotting a histogram, we found that approximately 80% of the games have 20 tags each, leading us to decide to focus on popular tag counts. By applying the count_elements() function and selecting the top 10% most popular tags, we quantified how many popular tags each game contains and stored the number in a new column 'num_pop_tag'. The visualization of this column revealed a compelling normal distribution, indicating a balanced representation of popular tags across the dataset.

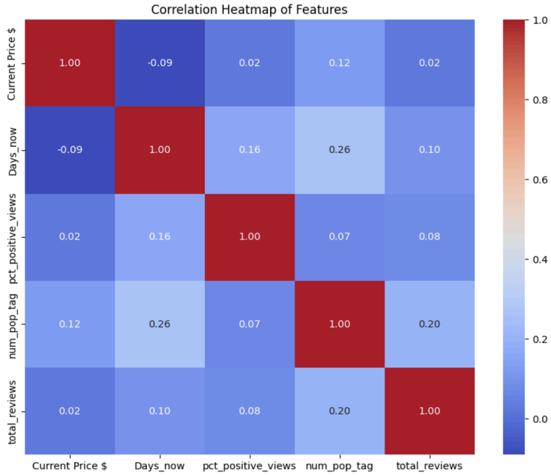




To further enhance the effectiveness of future research, we conducted an examination on the normality of the four features and alongside with total reviews. By plotting the histogram for each feature, it turned out that every variable has a strong skewness, except for the variable number of popular tags. Understanding the impact of skewed data on prediction accuracy, we decided to use the normalization method in order to lessen the deviation of the data and mitigate the impact of data size differences on prediction results. The method we chose is standard score normalization, which shows great effectiveness in scaling variables with a mean of zero and standard deviation of one. It also makes sure all the features contribute equally to the data, no matter the original size of the data. After standardization, the subplots for the five variables display a reduced skewness and a more uniformed range in values.

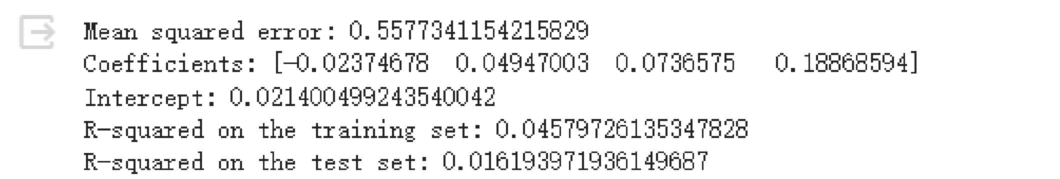


While the data are generally cleaned and normalized, we want to further study the correlation between the independent variables and total review numbers. We created a heat map and the actual correlation coefficients are pretty low, ranging from 0.02 to 0.20, indicating that there's a lack of strong statistical evidence to assert that the four chosen features significantly influence the total views over the last 30 days. Among these, the feature with the highest correlation to the dependent variable is the "total number of popular tags."



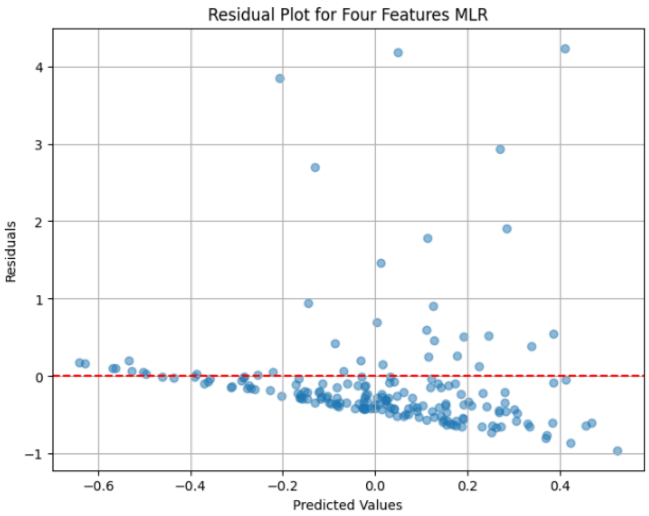
Results and Conclusions

Multi Linear Regression on Four Features



```
Mean squared error: 0.5577341154215829
Coefficients: [-0.02374678  0.04947003  0.0736575   0.18868594]
Intercept: 0.021400499243540042
R-squared on the training set: 0.04579726135347828
R-squared on the test set: 0.016193971936149687
```

R-squared is a statistical measure that represents the proportion of the variance for the dependent variable that's explained by the independent variables in the model. The R-squared for the training set is 4.6%, which means that the model explains about 4.6% of the variance in the dependent variable based on the training data. This underperformance is further illustrated by the residual plot, where most of our dependent variables' residuals fall within the range of -1 to 4, though some residuals are close to 4. This pattern suggests that there are instances where the predictions significantly deviate from the actual values. When this R-squared value drops even further to 1.5% on the testing set, it indicates an even weaker ability to predict or explain the variance in new, unseen data.

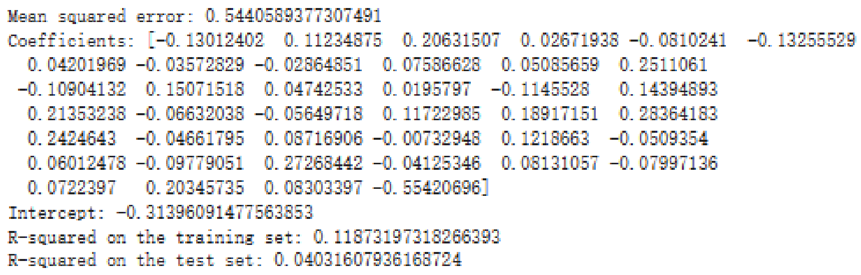


Overall, the R-squared values for the training and test sets indicate this model's limited ability to explain the variability in the dependent variable effectively.

Multi Linear Regression and OLS on Tags

Since the information Steam provides on each game page is limited, we cannot add more features into our prediction model after realizing our model is underperforming. We decided to move forward and focus on the tags, which have the highest correlation with the dependent variables among the four independent variables we considered. Because the data we collected contains more than 400 different tags, and too many features can disrupt the predictive results of the model, just like what we have done before, we only retained the top ten percent of the tags by frequency, we called it “popular tags”. For these popular tags, if a game contains a certain tag, the column named after that tag will be marked as 1, otherwise, it will be marked as 0. We converted the data into a binary format in this way to facilitate machine learning.

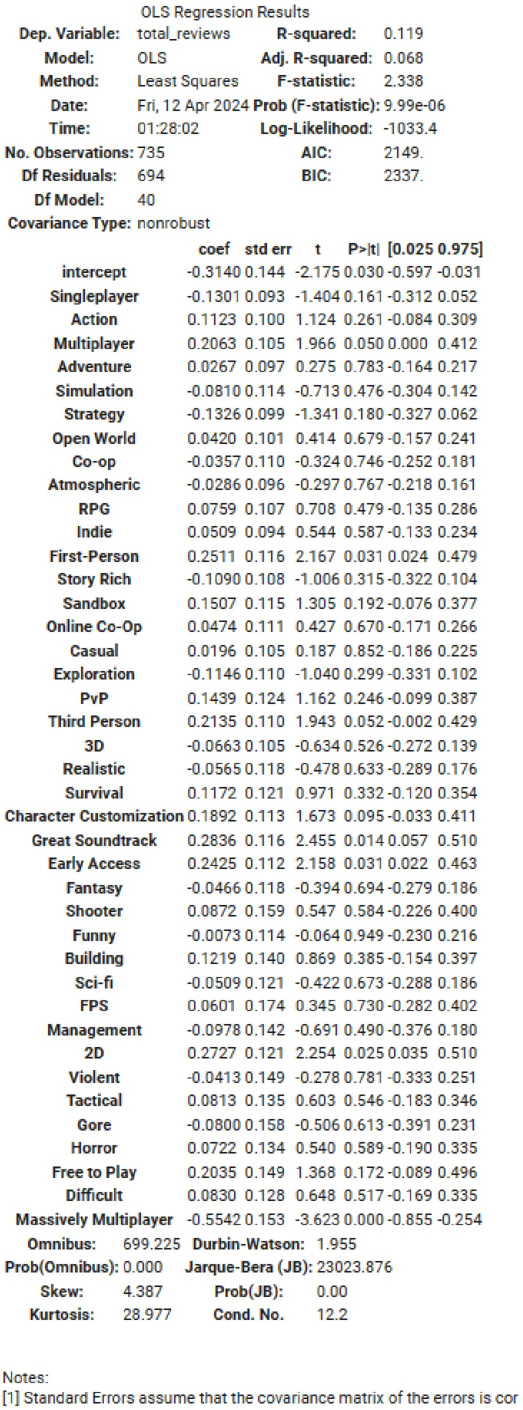
Instead of solely using the standard Multiple Linear Regression model, we have introduced another linear regression approach called Ordinary Least Squares (OLS). This method allows us to identify which features have a statistically significant impact on the prediction outcome.



```
Mean squared error: 0.5440589377307491
Coefficients: [-0.13012402  0.11234875  0.20631507  0.02671938 -0.0810241  -0.13255529
  0.04201969 -0.03572829 -0.02864851  0.07586628  0.05085659  0.2511061
 -0.10904132  0.15071518  0.04742533  0.0195797  -0.1145528   0.14394893
  0.21353238 -0.06632038 -0.05649718  0.11722985  0.18917151  0.28364183
  0.2424643  -0.04661795  0.08716906 -0.00732948  0.1218663  -0.0509354
  0.06012478 -0.09779051  0.27268442 -0.04125346  0.08131057 -0.07997136
  0.0722397   0.20345735  0.08303397 -0.55420696]
Intercept: -0.31396091477563853
R-squared on the training set: 0.11873197318266393
R-squared on the test set: 0.04031607936168724
```

The R-squared from the OLS is similar to the Multi Linear Regression’s R-squared on testing set, which is 11.9%, because both of them are regression models, which means that approximately 11.9% of the variance is explained by the training data. The R-squared on the test set is 4.03%, which is almost one third of that of the training set. This significant drop may indicate that our model lacks the ability to work on unseen data.

In the OLS output, significant variables are identified by their p-values (P>|t|). Typically, a variable is considered significant if its p-value is less than a chosen significance level (which is 0.05 here). The significant variables are: Multiplayer, Great Soundtrack, 2D, Massively Multiplayer... Other variables, despite having some level of association with the dependent variable, are not statistically significant at the common 0.05 threshold, as their p-values are higher. Their significance suggests they have a meaningful impact on the dependent variable.



```
                          OLS Regression Results
   Dep. Variable:   total_reviews        R-squared:      0.119
       Model:       OLS               Adj. R-squared:    0.068
      Method:       Least Squares        F-statistic:    2.338
        Date:       Fri, 12 Apr 2024 Prob (F-statistic): 9.99e-06
        Time:       01:28:02          Log-Likelihood:    -1033.4
No. Observations:   735                      AIC:        2149.
  Df Residuals:     694                      BIC:        2337.
    Df Model:       40
Covariance Type:    nonrobust
                           coef   std err    t     P>|t|  [0.025  0.975]
       intercept          -0.3140  0.144  -2.175  0.030  -0.597  -0.031
      Singleplayer        -0.1301  0.093  -1.404  0.161  -0.312   0.052
         Action            0.1123  0.100   1.124  0.261  -0.084   0.309
      Multiplayer          0.2063  0.105   1.966  0.050   0.000   0.412
       Adventure           0.0267  0.097   0.275  0.783  -0.164   0.217
       Simulation         -0.0810  0.114  -0.713  0.476  -0.304   0.142
        Strategy          -0.1326  0.099  -1.341  0.180  -0.327   0.062
       Open World          0.0420  0.101   0.414  0.679  -0.157   0.241
         Co-op            -0.0357  0.110  -0.324  0.746  -0.252   0.181
      Atmospheric         -0.0286  0.096  -0.297  0.767  -0.218   0.161
          RPG              0.0759  0.107   0.708  0.479  -0.135   0.286
         Indie             0.0509  0.094   0.544  0.587  -0.133   0.234
      First-Person         0.2511  0.116   2.167  0.031   0.024   0.479
       Story Rich         -0.1090  0.108  -1.006  0.315  -0.322   0.104
        Sandbox            0.1507  0.115   1.305  0.192  -0.076   0.377
      Online Co-Op         0.0474  0.111   0.427  0.670  -0.171   0.266
         Casual            0.0196  0.105   0.187  0.852  -0.186   0.225
      Exploration         -0.1146  0.110  -1.040  0.299  -0.331   0.102
          PvP              0.1439  0.124   1.162  0.246  -0.099   0.387
      Third Person         0.2135  0.110   1.943  0.052  -0.002   0.429
           3D             -0.0663  0.105  -0.634  0.526  -0.272   0.139
       Realistic          -0.0565  0.118  -0.478  0.633  -0.289   0.176
        Survival           0.1172  0.121   0.971  0.332  -0.120   0.354
Character Customization    0.1892  0.113   1.673  0.095  -0.033   0.411
    Great Soundtrack       0.2836  0.116   2.455  0.014   0.057   0.510
      Early Access         0.2425  0.112   2.158  0.031   0.022   0.463
        Fantasy           -0.0466  0.118  -0.394  0.694  -0.279   0.186
        Shooter            0.0872  0.159   0.547  0.584  -0.226   0.400
         Funny            -0.0073  0.114  -0.064  0.949  -0.230   0.216
        Building           0.1219  0.140   0.869  0.385  -0.154   0.397
         Sci-fi           -0.0509  0.121  -0.422  0.673  -0.288   0.186
          FPS              0.0601  0.174   0.345  0.730  -0.282   0.402
       Management         -0.0978  0.142  -0.691  0.490  -0.376   0.180
           2D              0.2727  0.121   2.254  0.025   0.035   0.510
        Violent           -0.0413  0.149  -0.278  0.781  -0.333   0.251
        Tactical           0.0813  0.135   0.603  0.546  -0.183   0.346
          Gore            -0.0800  0.158  -0.506  0.613  -0.391   0.231
         Horror            0.0722  0.134   0.540  0.589  -0.190   0.335
      Free to Play         0.2035  0.149   1.368  0.172  -0.089   0.496
       Difficult           0.0830  0.128   0.648  0.517  -0.169   0.335
  Massively Multiplayer   -0.5542  0.153  -3.623  0.000  -0.855  -0.254
     Omnibus:       699.225  Durbin-Watson:     1.955
Prob(Omnibus):      0.000   Jarque-Bera (JB):  23023.876
      Skew:         4.387        Prob(JB):      0.00
    Kurtosis:       28.977       Cond. No.      12.2


Notes:
[1] Standard Errors assume that the covariance matrix of the errors is cor
```

Even Though the R-squared is still low in this model, it has a significant increase compared with the four features' Multi Linear Regression model. The selection of the dependent variable and the very limited number of independent variables are likely the main reasons for the failure of our prediction model. The total number of reviews cannot perfectly substitute for total sales, and the four features extracted from the game pages have been calculated to have little significant correlation with the dependent variable. Due to the limited information provided on Steam pages, it is also difficult to add more features to increase prediction accuracy. Moreover, predicting game sales is essentially making predictions about human behavior and preferences, which are hard to predict. Based on these factors, I believe that the low R-squared of our model is reasonable and acceptable.

Future Work

To improve our model accuracy in the future, we identified several areas for improvement. Firstly, it will be crucial to expand our dataset. While our current dataset captures 1000 data points scraped from the Steam Global Top Seller webpage, there are in total 9100 games on the webpage and 400 million overall games selling on Steam platform. By gathering more comprehensive data, we can enrich our analysis and potentially uncover additional insights.

Additionally, we can broaden the scope of features utilized to improve our model's prediction ability. Since we only have four features due to the limited web page information, it will be a good choice to incorporate more features like cost, genre, DLC or game packages, and localization. These additional features have the potential to uncover stronger correlations with total reviews, as different game genres attract diverse player demographics and behaviors, influencing review counts uniquely. Moreover, obtaining a more informative dependent variable, such as gross profit or number of sales, could further enhance model accuracy. Since the number of reviews is heavily influenced by player behavior, studying the effects of more directly related features on game sales may find stronger correlations and improve predictive accuracy.

Lastly, exploring alternative machine learning techniques beyond linear regression, such as random forest or decision trees, might have strong potential for refining our model. These approaches offer different capabilities and may uncover patterns in the data that linear regression cannot capture alone. By implementing a variety of machine learning tools, we can optimize our model's performance and deliver more accurate predictions.

Reference

[1] DDDHL (2022) Steam Game Information

Scraping(https://blog.csdn.net/DDDHL_/article/details/111768725)