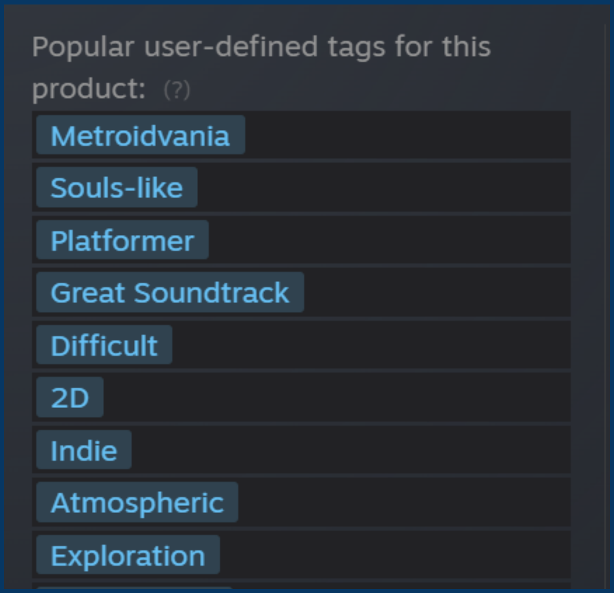
STEAM STORE: USING GAME FEATURES TO PREDICT REVENUE

CONTEXT:

Steam is a digital platform centered around PC games. Users of the platform can purchase and play games, interact with friends, and participate in a larger community of gaming enthusiasts.

Games on the Steam store are categorized using a user-defined tagging system, which allows for greater flexibility in game categorization and recommendations.

These tags, however, can refer to a wide variety of characteristics, such as genre, playstyle, or Steam-specific features, such as Steam trading cards or achievements, which allow the user to "level up" within the platform.



Popular user-defined tags for this product: (?)

Naturally, some tags or game features may make a game more appealing to a potential buyer. The aim of this project was to develop a model that can adequately predict estimated revenue (calculated as price * number of purchases).

DATA:

The original data was sourced from Nik Davis on kaggle. Two datasets, one consisting of basic information on each game, and the other consisting of extensive tag information per game, were merged. The final dataset consisted of 26,355 games and 43 columns. Tag-based columns were one-hot encoded into binary values. Season of release was extracted from release date and one-hot encoded as well. The percentage of positive reviews was calculated using the number of positive and negative reviews. Revenue, the target feature, was calculated using the price and average owners columns.

One major data entry error was noticed, in which 66 games were given a 'free_to_play' tag when they had a listed price, and 719 were missing a listed price when they were not in fact free. For the first set, the 'free_to_play' tag was removed, and they were grouped with the paid games. The second set was removed from the final dataset, as there were too many to manually input the missing prices. The 1,841 free-to-play games were separated from the 24,515 paid games prior to data visualization and analysis.

The final list of features used to train our model were:

Genre

- indie
- action
- adventure
- casual
- strategy
- simulation
- rpg
- puzzle
- 2d
- great_soundtrack
- atmospheric
- vr
- difficult
- story_rich
- free_to_play
- anime
- horror
- platformer
- pixel_graphics
- violent

Gameplay

- Multiplayer
- Co-op
- Controller Support
- Steam Achievements
- Steam Trading Cards
- Steam Cloud

Other Features

- Years since release
- Released in winter
- Released in spring
- Released in summer
- Released in fall
- Number of achievements
- Median Playtime
- Total Ratings
- Percentage of positive ratings
- Number of tags per game

MODELING:

Three types of models were trained and tested: 1) Linear Regression, 2) Random Forest Regressor, 3) Gradient Boost Regressor. Hyperparameter tuning was also performed using Grid Search or Randomized Search cross-validation.

The linear regression underfit the data and performed poorly when cross-validated. The random forest regressor and gradient boost regressor performed at about the same level, though the random forest regressor fine-tuned with randomized search cross validation was selected as the best model. This model used a robust scaling technique and selected a best k of 31 (out of 35 total features).

Model metrics can be compared in the table below.

Model	**R^2 (Train)**	**R^2 (Test)**	**RMSE (Train)**	**RMSE (Test)**	**MAE (Train)**	**MAE (Test)**
Linear Regression	0.55	0.53	5,360,643.06	5,422,963.47	1,410,534.18	1,422,660.63
Linear Regression (Grid Search CV)	0.55	0.53	5,360,649.75	5,422,949.54	1,410,635.20	1,422,649.47
Random Forest	0.94	0.67	1,874,654.06	4,576,871.91	257,098.75	672,476.00
Random Forest (Randomized Search CV)	**0.95**	**0.68**	**1,834,738.39**	**4,530,477.82**	**254,405.53**	**666,622.53**
Gradient Boost	0.92	0.67	2,237,714.42	4,551,346.36	467,590.29	656,420.03
Gradient Boost (Randomized Search CV)	1.00	0.65	114,906.85	4,688,391.94	60,094.85	680,286.72

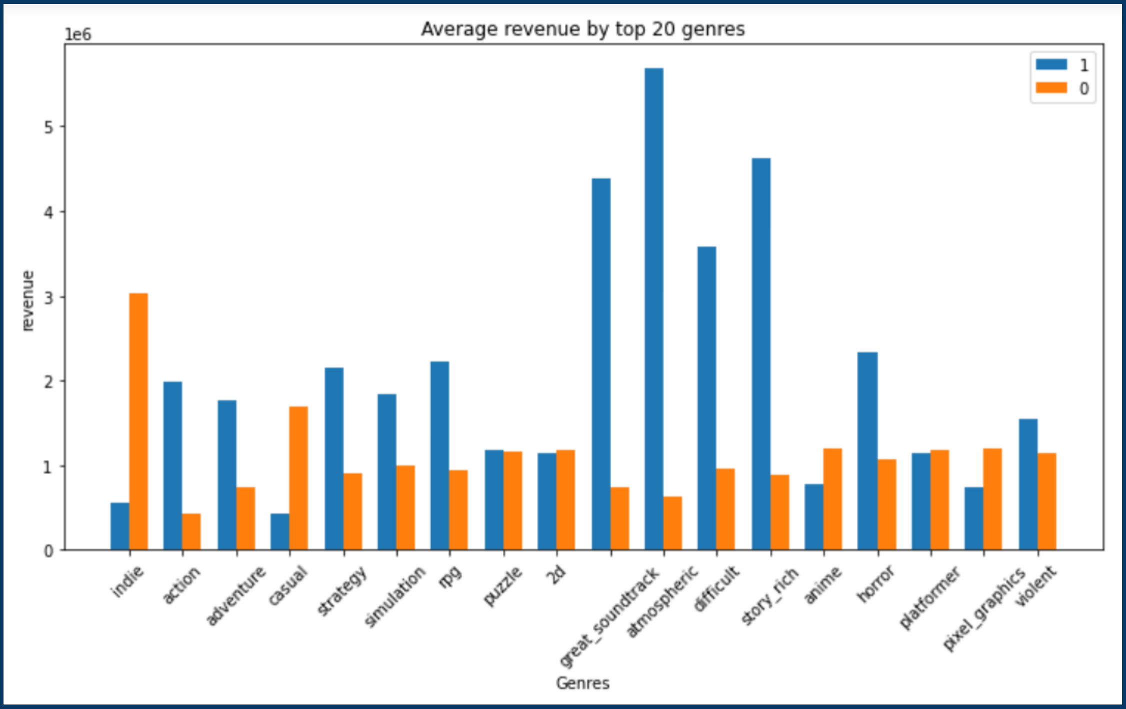
KEY FINDINGS:

Using our optimized random forest model, these game-related features were found to have the most impact on revenue:

- Games with Steam achievements
- Indie games
- RPGs (Role-Playing Games)
- Co-op (Cooperative) games
- Games released in summer
- The number of tags listed in a game

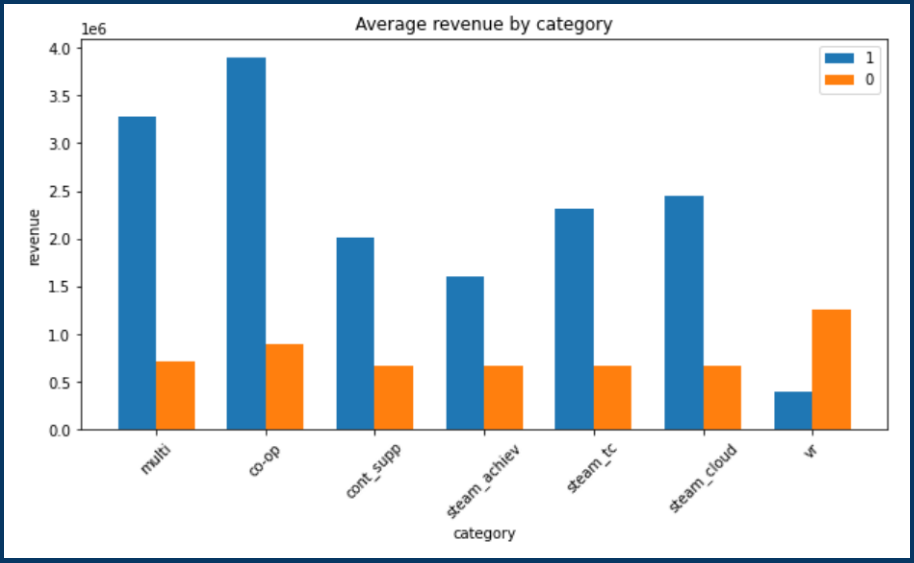
Due to the ‘black box’ nature of the Random Forest model, directionality of this influence is not explicitly stated. However, when visualizing the mean revenue across these categories, some inferences can be made.

For the most part, games that possessed one of the top 20 most popular genre tags earned on average higher revenue than those without. The few exceptions were indie, casual, pixel graphic, and anime games. Based on this, it seems likely that a game tagged as indie will earn less revenue than a game without the tag. This makes sense, as a game developed by an independent studio is likely to have a lower budget and marketing capacities.



Based on this graph, atmospheric, story rich, difficult, and games with great soundtracks have the highest average revenue. However, the RPG tag is more influential in predicting revenue.

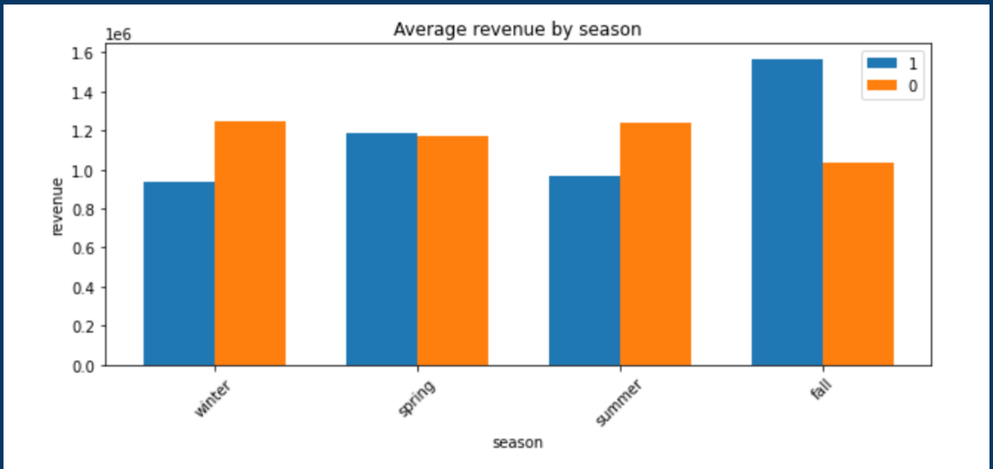
Across gameplay categories, all types (multi-player, co-op, steam achievements, steam trading cards, etc.), games with the feature earned more revenue on average than games without the feature. The one exception was VR (virtual reality) games, which generated less average revenue than non-VR games.



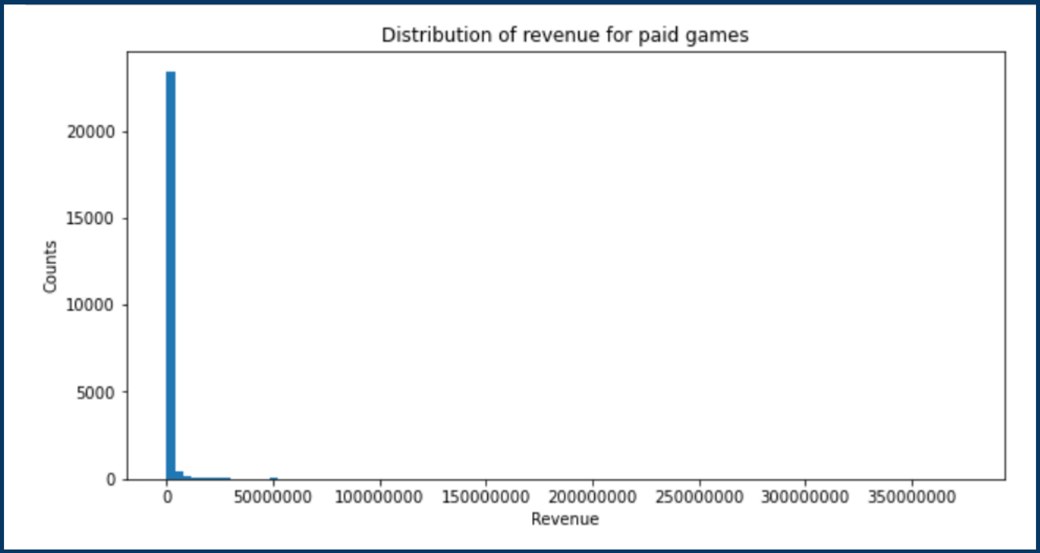
Of these gameplay categories, the most influential in predicting revenue was **co-op**.

Regarding the season of release, games released in the fall had the highest average revenue, while games released in the summer and winter had a lower average revenue. This is interesting, as overall there are slightly fewer releases in summer and winter (6,286 and 6,520 respectively) compared to fall and spring (7,189 and 7,080 respectively), so it's unlikely that a larger influx of games is bringing down the average price.

The two major Steam sales are the summer and winter sale. Perhaps games released during these time periods compete with a large number of older releases that are currently on sale, and therefore don't perform as well. Another possibility is that games released during these seasons are released with a lower price in order to compete with games on sale.



One important point to note is that most games do not generate much revenue. The distribution of revenue amongst paid games was very skewed, with most games earning a fairly low amount (**median = $69,650**) and a few earning very high amounts (**mean ≈ $1,259,000**). One major outlier was found at a revenue of around $2 billion. As the next highest value was only around $375 million, this game was dropped. However, as seen in the following figure, the distribution was still noticeably skewed to the right.



CONCLUSION:

In summary, our best model was able to predict revenue to a certain degree, though the current level of accuracy may not be sufficient for implementation. A more extensive analysis, perhaps factoring in less popular tags, pre-existing franchises, or game developers, may yield a stronger model. We did, however, identify a noticeable influence of indie, RPG, and co-op games on revenue, as well as summer releases.

Based on these results, we can suggest that new games schedule release dates around spring or fall, away from big sale seasons. Furthermore, the platform can do more to promote co-op or RPG games, as they tend to generate more revenue.

Some limitations of this project include the fact that our target variable was extremely skewed, making prediction difficult. Furthermore, the analysis was limited to the top 20 genres with no method of hierarchy. For example, a game with action as its main tag might be quite a different gameplay experience from a game where action is lower on the tag list. Future studies may be able to take this into, perhaps limiting each game to its top three genres. Finally, each game in this dataset was evaluated independently, and does not account for video game series, larger franchises, or well-established game studios.