# DS3000 Final Project: Predicting Steam Games Discount Probabilities

Alex Yu, Meng Un, Andrew Jin, & Ian Lau

## ABSTRACT

This project aims to determine the probablility of a game on Steam going on sale. This data is derived from the Steam DB, which contains data pertaining to price and Kaggle Steam Database, which contains information and values about the user ratings, release date, release price, discount price, and final price. The data spans from 2008 to the present and we have decided to use the XGBoost model to predict whether or not a game will go on sale. The performance of the model was also evaluated.

## INTRO

Steam is a popular platform for video game distribution, but it can be challenging for gamers to predict when a game will go on sale and for game developers to make informed pricing decisions.

**Defining the Problem:**

This project aims to use machine learning to predict whether a Steam game will go on sale based on historical data and game ratings. By doing so, we aim to help gamers buy games at lower prices and assist game developers in making informed pricing decisions for their products.

**Motivation for the topic:**

As gamers ourselves, we know the frustration of buying a game at full price only to see it go on sale shortly after. We also understand the importance of data-driven decision making for game developers, as they need to make time-crucial decisions in order to maintain popularity, support, and success in their products to provide the monetary support in order to keep the product up and running.

**Goals & Objectives:**

Our main goal is to use a machine learning model to predict whether a Steam game will go on sale with high accuracy. Our objectives include analyzing factors such as game rating, pricing, and release date to understand the variables that influence game sales, as well as cleaning such data to remove any possible insignificant data that could affect our results, such as games being free, much older games, etc.

## METHODOLOGY

In order to start with this project, we gathered data from the Kaggle Steam Database starting with games from 2008. We loaded it into a dataframe and performed an EDA. Then we started cleaning, we to drop games that were free since that was not in the scope of our project and well as filtered out games that have been released prior to 2019. The EDA showed that games from earlier skewed the results severely in terms of average pricing and discounts. We also checked for blank values within columns, though no such invalid values were found.

**Features:**

We selected ratings, positive ratio, final price, original price, and discount amount as the features in our model for predicting whether or not a Steam game goes on sale. While deciding on which model to use, one of the models in considerations was the random forest regression, and we noted that the features should only be numerical values. Thus, in order to get rid of biases, all the features that were analyzed throughout the entirety of the project in all the different models remained consistently numerical.

**Model Testing:**

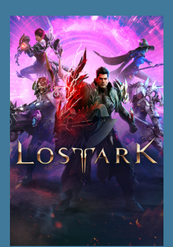
The models that we tested and built were logistic regression, random forest regression, and XGBoost. We built the initial model, tuned the hyperparameters, implemented the hyperparameters and cross validated for all of the algorithms. The first two algorithms however ran into the same issue where they do not accept nan values natively for cross validation. Thus, not allow us to evaluate the performance of the model. Both algorithms also took for than 2 minutes for the randomized search and to fit to process.

**XGBoost:**

The XGBoost algorithm in comparison was much more effiecient in comparison, the randomized search cv was much more efficient for the XGBoost algorithm. To train the XGBoost model, we gave it the features mentioned above and the discount amount as the target. The data was partitioned into a training and test set, where the training set is 80% of the data and the test set is 20% of the data. Once partitioned, we normalized the numerical features in the training and test sets. We hypertuned parameters such as objective, colsample_bytree (random selection of a fraction of the features for more accuracy per tree, exploring each combination of features), learning rate (prevent overfitting), max depth (manual decrease to prevent overfitting), alpha (high dimension), and n estimator. We fit the grid using the x and y training data and displayed the best possible score.

**Model Performance Evaluation:**

For evaluating the performance of the model, we used the mean R-squared score value and its standard deviation. As well as cross validated the model.



Add to your wishlist

## RESULTS AND EVALUATION

The best parameters which corresponds to the best score achieved in our XGBoost model with an accuracy of 98.86%.

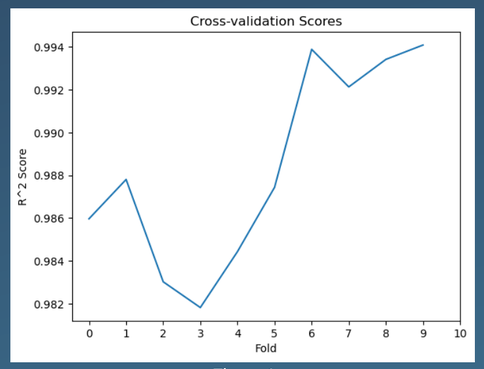



*Figure 1*

The cross validation scores line chart (Figure 1) is hovering above a 98% R^2 score. We can make the inference that this model is performing well, especially since the mean R-squared score is also very high at 98.8% and the standard deviation being relatively low at 0.004. This low standard deviation also means that there is less variability in the data and suggests that the R^2 scores are consistent and close to the mean score as well as the data points being grouped closely together.

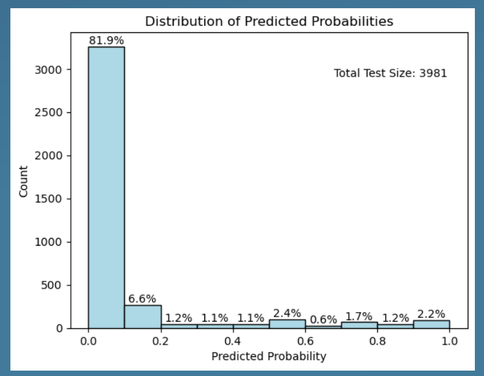



*Figure 2*

As we can see in Figure 2, almost 82%, 3,260 games, of the test size of games reside in the first bin between 0 and 0.1. What this tells us is that most games will either have a 0 or 10% chance of going on sale. The rest of the distribution is split into 9 more bins with only about 18% of the test set left. This visualization shows us a very good representation on the grand scheme of things, however, we want to dive deeper and see more about how many games could have a higher chance of going on sale. Thus, we made another visualization of probabilities above 50% (Figure 3).

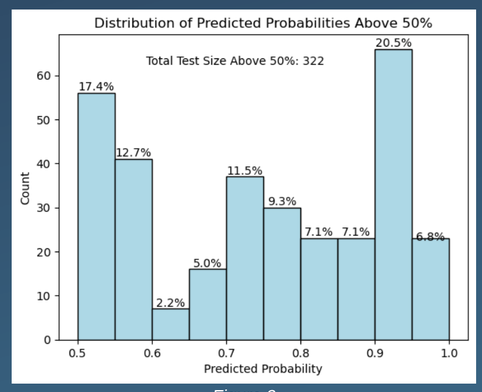




*Figure 3*

Figure 3 shows the distribution of probabilities above 50% with only 322 test data points. We can see that many games could go on sale with the biggest bin being between 0.9 and 0.95 at 20.5% of the test data points. There is a less defined trend compared to the general overview for the full range of test data points above. Since 322 is about 1% of the total test data size we can make the observation that most games will probably not go on sale.

## IMPACTS

With the adoption of our model, both parties (gamers and game developers) benefit. Our model can give a better understanding of the likelihood of a game going on sale, to ensure gamers don't end up overpaying for a game. Overall, this will provide more information about buying games in general, which will make the game-buying experience much better and create a positive impact on the gaming community.

## CONCLUSION

Overall, in terms of addressing our prompt, I think we found the right model to help us best predict the general likelihood of Steam games going on sale. As mentioned earlier, using featurers such as ratings, positive ratio, final price, original price, and discount amount we utilized our knowledge of different models and tested accordingly. When we attempted with a linear regression model and a random forest regression model, we ran into an error with NaN values being rejected natively during cross-validation, and the XGBoost model proved most accurate with a 98.86% accuracy rate.

As we have concluded from our results section, most games will not go on sale. Based on our XGBoost model, we can make that strong inference that since only about 1% of games from the test data size had a 50% chance or higher of going on sale, most games in the overall data will most likely not go on sale either. Essentially our answer is of the mindset that it is extremely unlikely a game goes on sale based on ratings alone. There is definitely more at play than simply looking at ratings as one of the main contributing factors to the probability of games going on sale.

Things we could have done differently would include involving more data for the analysis. We now know that rating is just a small factor in the equation as to how gaming companies will establish discounts. Also if possible in the future, we would like to explore different genres to see if there are trends amongst specific genres for discounts.