MSCI446 Introduction to Machine Learning
Project Proposal

Christian Chan 20844761
Kevin Jiang 20832404
Jingshi Liao 20822390

March 7, 2022

Business Problem

Games have been around since 1958[1] and since then they have become a multi-billion dollar industry[2], serving games to millions of people around the world.

Data analytics and machine learning are primarily used within the game industry to benefit the game companies in terms of game quality, monetization, marketing, or behavior data. This project seeks to explore the use of machine learning on game data to help users save money in the current expensive industry.

Through the advancement of technology, games have also made tremendous leaps in game design using things like facial and voice recognition, virtual reality, high definition graphics, and much more. However, these advancements have caused games to become more expensive to develop, while also requiring a higher demand in computing power. All these factors have heavily increased the price of games over the last four years. The average price of the top anticipated or played games have risen to the range of $70 - $90[3], some even more expensive. This has made the gaming community extremely hesitant to break their bank to play a few games. That's why this project focuses on helping the gaming community, to answer the question if gaming industry data can help game enthusiasts save money or to make games more affordable.

To explore this problem, we will be focusing on Steam, the largest game distributor on PC. They hold the largest market share and have been around since 2003 and are currently distributing more than 100000 games globally. Games frequently go on sale on the Steam store but the timeframe of the sale and their discount prices are always unknown. There are also external sites that sell games for discounted prices, for example: Green Man Gaming, Daily Game Deals, Fanatical, and Humble Bundle are all sites where games can go and purchase games for a cheaper price. They may also go to Is There Any Deal (ITAD) to view deals on many different sites. For players, it would be helpful for them to have a prediction of when these sales would occur so they can save money by waiting for a sale they know is likely to occur instead of paying for games at full price. Therefore, we aim to use data to help gamers gain insight into possible discounts for games they are interested in as well as when they may go on sale on Steam.

Motivation

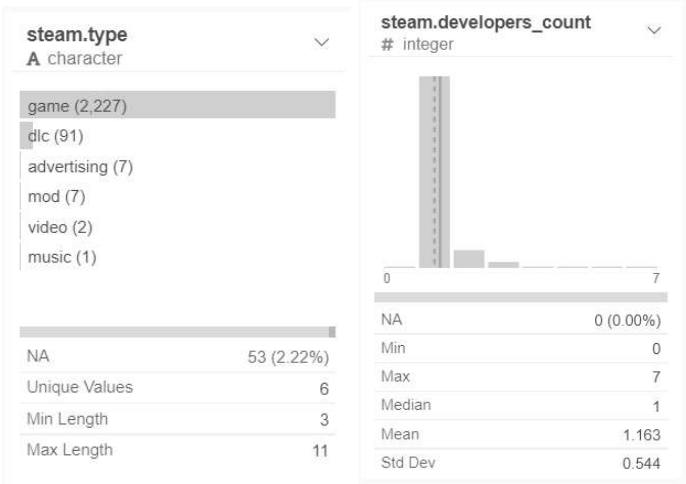
As technology continues to develop and the video game industry continues to thrive (and is one of the most revenue-generating industries globally), corporations in the industry have put in a lot of resources to better understand the market trends, study customers' purchasing patterns, predict sales, etc. Most of the research, studies and analytics have been done to aid corporations to maximize their profits. However, very few analystics were carried out to aid customers to make the best purchasing decision.

The cost to purchase video games has significantly increased over the years. Purchasing games directly from the publishers (e.g. Steam) has become less and less affordable for some customers. Voices have raised about unreasonable pricing to purchase a digital game, as well as the different pricing based on country/region. For the same reason, the 'grey market' (which is made up of ecommerce sites that sell activation keys for games, including newer

titles) have expanded and become more and more popular. The grey market, as its name suggests, is a place to buy video games that is not illegal but might not be totally ethical. Purchasing activation keys from the grey market also comes with various risks like financial information leak, possible malware and scams, etc[4].
Therefore, being able to predict when a game will go on sale, and how much it would go on sale by, is going to help customers find opportunities to save money by purchasing games in a smart way, or deciding when is the best time for them to purchase a game. In addition, it is likely that less people will resort to the grey market sites for cheaper but insecure purchases. This could lead to a better balance between meeting the customers' price expectations and protecting the publishers' work.

Dataset Description

The data was created by scraping the Steam API for game information, and ITAD for historical prices. A preliminary version of the data can be found at https://github.com/kevin51jiang/msci-446/blob/master/report/rNormalized.json, but the full one could not be hosted online due to its size of 1.8 GB. At the time of scraping, there was no knowledge of what factors may affect the prediction, so all data was downloaded in order to save time. While games for ITAD popularity ranks of 1-2500 were targeted, some pages experienced issues, and others turned out not to be games. These types are filtered and ignored for the rest of the data exploration.



The unfiltered types that made it into the dataset, the number of developers

Sample Rows:

	itadPlain	**itad.position**	**itad.rank**	**itad.sales.timestamp**
0	nierautomata	1	1	145600574900

0	nierautomata	1	1	1458665932000
1	stardewvalley	2	2	1456005749000

itad.sales.amount	steam.steam_appid	steam.required_age	steam.is_free	steam.dlc
19.99	524220	0	FALSE	1
19.99	524220	0	FALSE	1
19.99	413150	0	FALSE	1

steam.detailed_description	steam.about_the_game	steam.short_description	steam.supported_languages
<h1>NieR:Automata™ Game of the YoRHa Edition</...	<a href="https://store.steampowered.com/app/11...	NieR: Automata tells the story of androids 2B,...	english french italian german spanish spain j...
<h1>NieR:Automata™ Game of the YoRHa Edition</...	<a href="https://store.steampowered.com/app/11...	NieR: Automata tells the story of androids 2B,...	english french italian german spanish spain j...
Stardew Valley is an open-ended country-life R...	Stardew Valley is an open-ended country-life R...	You've inherited your grandfather's old farm p...	english german spanish spain japanese portugu...

steam.reviews	steam.pc_requirements	steam.mac_requirements	steam.linux_requirements	steam.developers
TRUE	TRUE	FALSE	FALSE	[Square Enix, PlatinumGames Inc.]

TRUE	TRUE	FALSE	FALSE	[Square Enix, PlatinumGames Inc.]
TRUE	TRUE	TRUE	TRUE	[ConcernedApe]

steam.categories_description	**steam.genres_description**	**steam.controller_support**	**steam.metacritic.score**
[Single-player, Steam Achievements, Steam Trad...	[Action, RPG]	FALSE	0.866
[Single-player, Steam Achievements, Steam Trad...	[Action, RPG]	FALSE	0.866
[Single-player, Multi-player, Co-op, Online Co...	[Indie, RPG, Simulation]	TRUE	0.9175

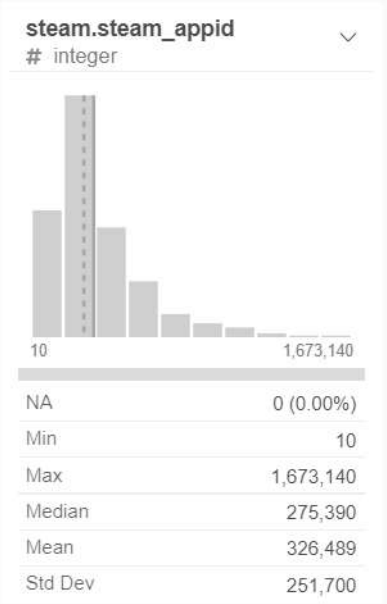
steam.drm_consolidated	**steam.ext_user_account_notice**
NaN	NaN
NaN	NaN
NaN	NaN

There are 2388 individual games, with 18212359 rows, and 22 columns.

The “itadPlain” column is the ID for a game given by ITAD, and it is a string. The “itad.position” column is for the ranked popularity of a game. There can be ties, so it goes from 1 to 1770, with 1 being the most popular and 1770 being the least popular. The “itad.rank” column is the trending level of a given game on ITAD. This one cannot tie, and so it goes from 1 to 2499, again with 1 being most popular and 1770 being the least popular.

The “steam.steam_appid” is the unique ID for a game given by Steam. While it should be considered a string, it can also be used as a proxy for how old the game is, since ITAD only

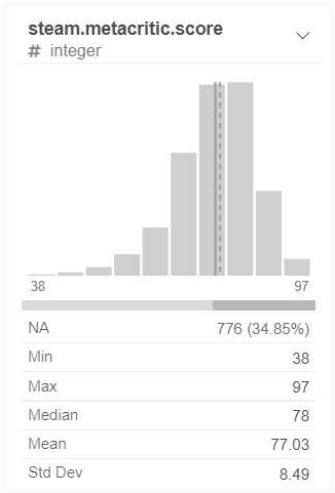
started scraping prices in March 2016. It goes from 10 to 1673140. An interesting observation is that a disproportionate amount of the most popular 2500 games are from small (i.e. old) app IDs.



Distribution of Steam AppIds in the dataset

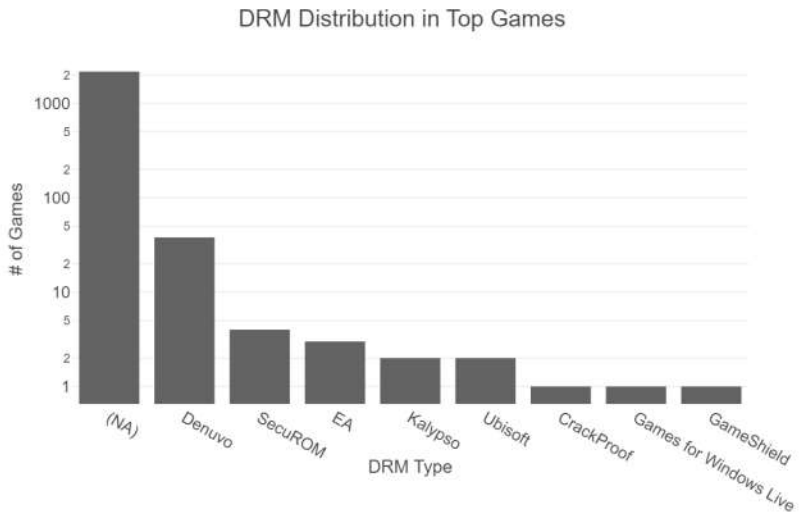
Next, there is the “steam.required_age” column, which indicates the minimum age required to play a game. It goes from 0, being no minimum requirement, to 18 years of age. It has a mean of 0.12, with the vast majority of games having no minimum requirement.
The “steam.is_free” column has a boolean type, which indicates whether or not a game is free. There is roughly a 12:17 ratio for free games to non-free games.
The “steam.dlc” column is an integer type, which lists the amount of downloadable content (DLC) that is available for the game after purchase. It has a median of 1, with a mean of 4.09. There is one outlier with 1555 pieces of DLC, which should be removed before analysis.
The “steam.detailed_description” column contains the HTML code for showing the detailed description on the Steam store page. It has a length of 50 to 19556 characters, but its utility in this analysis may be low. The “steam.about_the_game” and “steam.short_description” columns have similar content, with ranges of 50-15702 characters and 17-342 characters respectively.
The “steam.supported_languages” column is a vector of a categorical variable. It contains the languages supported by each game, in regular English format, for example “english french spanish”. There are 39 total types of languages that are supported.
The “steam.reviews” column is a boolean variable showing if their store page featured a review blurb or not. There is roughly a 5:4 ratio for games that do show a review to games that do not show a review. The reason this is used rather than text is because these reviews are highly likely to be favorable to the game developer, and so running sentiment analysis on this would not be useful.

The “steam.pc_requirements”, “steam.mac_requirements”, and “steam.linux_requirements” columns denote if a requirement string is displayed for the user. If any sort of requirement exists, we consider it to be true, otherwise it is false. Over 99.96% of games support Windows, around 71% of games support Mac, and roughly 61% of games support Linux.
The “steam.developers” column is a vector of strings, with each string showing a developer that has worked on that game. Over half of the games have 1 developer, with some having up to 7.
The “steam.controller_support” column is of a boolean type. If it is true, then the game supports using a handheld controller to play the game, otherwise it is false. Only 57% of games support controllers.



Metacritic score distribution in top games

The “steam.metacritic.score” column shows integers that can go from 0 up to 100, though in the data it goes from 38 to 97. Metacritic is a site that aggregates reviews from many sources, and so is likely to be a good measure for the quality of the game. Roughly 35% of games do not have a metacritic score.



DRM (Digital Rights Management) software present in games

The “steam.drm_consolidated” column includes categorical variables, showing which types of DRM (Digital Rights Management) are present in the game. DRM is used to prevent software from running on unauthorized systems. Since Steam already includes a layer of its own DRM, many games opt out of including their own. Of the 2388 games, over 2000 of them do not have any extra DRM, while the next most popular is Denuvo, a type of DRM known for being almost uncrackable.

The “steam.ext_user_account_notice” column is a categorical type of variable, alerting the player if a non-Steam account is required to be able to play the game. There are 32 unique external accounts that are needed overall.

Frequencies of categories in “steam.categories_description”. Darker means higher ranked.

The “steam.categories_description” is a column containing vectors of categories. It includes game features like “Captions”, “VR.Support”, and “Co.op”. Single-player is the most frequent category that is present, potentially due to the ease of developing for single-player games. There are 34 types of categories.

Frequencies of categories in “steam.genre_description”. Darker means higher ranked.

The “steam.genre_description” column is another column containing a vector of categories. This column shows the games genres, such as the ever-popular Action and Indie genres. There are 24 different genres.

The final two columns, “itad.sales.timestamp”, and “itad.sales.amount” are linked. For a given point in time, there will be an “itad.sales.timestamp” showing the time in UTC Unix time, and a corresponding price of the game at that point at $“itad.sales.amount” CAD. The first time was on February 20, 2016, which is when ITAD started keeping track of sales. The latest was on March 11, 2022 at 00:00 GMT, which is the previous midnight at the GMT time zone when the data was scraped.

The variables that are expected to be the most important are the time of year, the game’s popularity, and the game’s age. These are found in the dataset as timestamps of the price over time, “itadRank” which is a rank of the game’s popularity, and the game’s age which is determined by how long it has been since the first time a price has been reported on ITAD.

When forecasting, there are two important features that a plot has. It has a trend, which can usually be approximated as a linear equation, and it has seasonality, which can be approximated by a sinusoidal equation. The trend can be used to predict future values, while the seasonality can be used to predict when a certain event will happen again.

Methodology

Supervised Learning

In this project, we aim to use supervised learning to predict how much a game could go on sale for during a given day. Within this prediction we would need to consider a multitude of variables; for example, DRM type would be one string categorical variable to consider as this refers to the copy protection for video games, meaning there will be no redistribution of games from third party sites with significant discounts. Another string categorical variable would be genre, there are genres of games that are more likely to go on sale after a certain period of time. Small indie games would go on sale for more than an A-list highly anticipated video game. In addition, using the games current numerical rank with the Steam store would also help predict the amount discounted because games that are lower rank would tend to have higher discounts due to the lack of players. The last numerical variable to be used are the past prices for an individual game, if a game has already had one discount in the past, the new price may be more discounted than before. With a mix of numerical and categorical variables, one hot encoding must be employed so that the categorical variables can become numerical and be used within linear regression or even time series prediction.

Unsupervised Learning

This project aims to use unsupervised learning to predict if a sale would occur on a given day. Since all variables would become numerical through one hot encoding, clustering would be the appropriate method to complete our prediction. The variables we will cluster with will be previous sales dates, price, and genre. Previous sales dates will help us identify seasonal trends as it is common for game sales to occur during holidays or changes in the season. The price variable will give insight into the pricing trend, if it has been decreasing over time and is due to be put on sale. Lastly, the genre variable will be a good complement to previous sales dates as it can be possible that certain genres of games have sales of their own during some point in the year.

Prior work

1) ***Game Sales Prediction: ML in R* by Jiawei Xia[5]**
   Xia explored ‘what would be key factors to determine the sales of a game’ through linear regression and LASSO calculation. The linear regression model determined that User Score and Critic Score are the ultimate determinants. However, the linear regression model generated many significant factors, and possibly suffered from overfitting. The LASSO method determined the significant predictors are Publisher and Platform and is believed to be more efficient and accurate than the linear regression model in determining significant variables. To predict ‘how many percentages of newly developed games could be successful’, Xia implemented the logistic regression method and the KNN method. The outcomes of the 2 models predict that the successful game rate is between 10% - 20%, while the KNN model has a higher accuracy rate.

2) ***Sales Prediction on Video Games Using Machine Learning*** **by Bodduru Keerthana[6]**
   Keerthana applied 4 prediction modeling algorithms, which are linear regression, support vector regression, random forest and decision tree, for the best results. Dataset was split into a training dataset (on which the different algorithms apply), and a testing dataset (to test the algorithms' accuracy). Comparison indicates that the random forest model gives the most accurate result with the lowest error rate.

3) ***Sales Forecasting for Retails Chains*** **by Ankur Jain, Manghat Nitish Menon & Saurabh Chandra[7]**
   The authors used the Extreme Gradient Boosting algorithm to predict sales for retail outlets of a major European Pharmacy retailing company. The performances of the XGBoost predictor were compared with Linear Regression and Random Forest Regression. By comparing the Root Mean Square Percentage Error of the three models, the XGBoost was observed to perform the best at prediction.

4) ***Common Time Series Data Analysis Methods and Forecasting Models in Python*** **by Yuefeng Zhang[8]**
   Zhang explored a basic data analysis pipeline on how to structure data for time series forecasting, and some common algorithms that may work well, using a Kaggle dataset for global warming. He uses forward fill to fill in any missing data. When using ARIMA (AutoRegressive Integrated Moving Average), he preprocesses the data to separate the trend and the seasonality. When using LSTM (Long Short-Term Memory), it required minimal preprocessing, other than filling in nonexistent data. The LSTM model performed better than the ARIMA model. Unlike this project, it only uses time and previous price history as an explanatory variable. It also only considers one series, rather than the 2000+ this project will.

5) ***Promotional Analysis and Forecasting for Demand Planning: A Practical Time Series Approach*** **by Michael Leonard[9]**
   Leonard provided practical advice on how to apply the time series model and the intervention model on promotional analysis. Techniques were introduced to identify either the underlying time series model (e.g. stationarity analysis) or the intervention response specification (e.g. analyzing the historical data). Promotion can be forcased once the combined model is identified, fitted to the historical data and checked for adequacy. Leonard concluded that historical data can be decomposed into two parts (the underlying time series process and the intervention effect) for past promotional analysis, and forecasts can be adjusted based on proposed promotions using analysis of past promotions.

References

[1]Tretkoff, E. (n.d.). *October 1958: Physicist invents first video game*. American Physical Society. https://www.aps.org/publications/apsnews/200810/physicshistory.cfm#:~:text=In%20October%201958%2C%20Physicist%20William,Brookhaven%20National%20Laboratory%20open%20house.

[2]Clement, J. (2021, November 30). *Global Video Games Market Value 2021*. Statista. https://www.statista.com/statistics/246888/value-of-the-global-video-game-market/#:~:text=This%20timeline%20presents%20a%20forecast,surpass%20138%20billion%20by%20202

[3]Bennett, B. (2021, August 25). *It looks like next-gen video games will cost $90 in Canada*. MobileSyrup. https://mobilesyrup.com/2020/07/03/next-gen-video-games-cost-90-canada/#:~:text=Currently%2C%20new%20video%20games%20in,this%20is%20a%20%2410%20increase

[4]*The Grey World of key selling: Grey market sites come with many risks*. Official Site. (n.d.). https://au.norton.com/internetsecurity-kids-safety-grey-world-of-key-selling.html#:~:text=What%20is%20the%20grey%20market,for%20games%2C%20including%20newer%20titles.

[5]Xia, J. (2019, Jan 8). Game Sales Prediction: ML in R. https://rstudio-pubs-static.s3.amazonaws.com/548644_ab7f7c6b917442cd9c263e4f0675ce12.html#classification

[6]Keerthana B. (June 2019). Sales Prediction On Video Games Using Machine Learning . https://www.jetir.org/papers/JETIR1907H50.pdf

[7]Jain, A, Menon, M, N, Chandra S. Sales Forecasting for Retail Chains. https://cseweb.ucsd.edu/classes/wi15/cse255-a/reports/fa15/004.pdf

[8]Zhang, Y.. *Common time series data analysis methods and forecasting models in Python*. (2020, June 24) https://towardsdatascience.com/common-time-series-data-analysis-methods-and-forecasting-models-in-python-f0565b68a3d8

[9]Leonard, M.J. (2001). Promotional Analysis and Forecasting for Demand Planning : A Practical Time Series Approach. https://www.semanticscholar.org/paper/Promotional-Analysis-and-Forecasting-for-Demand-%3A-A-Leonard/70011dd1e1fd189a3cc273939cfa750e3b55f364