Analysis of Predictors of Price and Popularity of Steam Games in 2016

Abstract

summarize your question, your methods, your results, and your main conclusions in a few hundred words or less.

Introduction

Introduce the question you're trying to answer at a reasonable level of detail. Give background and motivation for why it's important.

Our goal is to discover interesting trends that allow us to predict how a game is priced and it's popularity after release. The price is divided into price initial and price final. We are going to use price initial as our metric to determine price and SteamSpyOwners as a metric to predict popularity. As avid gamers, we have long been interested in the games industry and wish to uncover any insights behind what makes some games more successful than others. This kind of analysis is significant for any indie developer so that they can better understand what kinds of games garner player interest as well as how much to charge for them.

Methods

Describe your data set and the methods you will use to ana- lyze it.

This is a dataset on all steam games scraped from public Steam APIs and steamspy.com. The dataset comes from data.world at this url: https://data.world/craigkelly/steam-game-data. The dataset has 78 columns and includes things like price, descriptions and release date. According to the data on Github, the repository was created in Septembere 12, 2016 so we are assuming that the data also comes from the end of the year in 2016. The dataset has the following columns:

QueryID - (Integer) The original ID in idlist.csv

ResponseID - (Integer) The ID returned in the Steam response (should equal QueryID)

QueryName - (Text) The original name in idlist.csv

ResponseName - (Text) The name returned in the Steam response (should equal QueryName)

ReleaseDate - (Text) Appears to the be the initial release date for the game

RequiredAge - (Integer) list named required_age in JSON

DemoCount - (TextualCount) list named demos in JSON

DeveloperCount - (TextualCount) list named developers in JSON

DLCCount - (TextualCount) list named dlc in JSON

Metacritic - (Integer) numeric score from metacritic object in JSON

MovieCount - (TextualCount) list named movies in JSON (used object id for unique count)

PackageCount - (TextualCount) list named packages in JSON

RecommendationCount - (Integer) from recommendations.total in JSON

PublisherCount - (TextualCount) list named publishers in JSON

ScreenshotCount - (TextualCount) list named screenshots in JSON

AchievementCount - (Integer) achievements.total in JSON

AchievementHighlightedCount - (TextualCount) for achievements.highlighted in JSON

ControllerSupport - (Boolean) True if controller_support was full

IsFree - (Boolean) is_free in JSON

FreeVerAvail - (Boolean) True if is_free_license is True in package_groups list

PurchaseAvail - (Boolean) True if price_in_cents_with_discount greater than 0 in package_groups list

SubscriptionAvail - (Boolean) True if is_recurring_subscription is True in package_groups

PlatformWindows - (Boolean) True if platforms.windows is True

PlatformLinux - (Boolean) True if platforms.linux is True

PlatformMac - (Boolean) True if platforms.mac is True

PCReqsHaveMin - (Boolean) True if pc_requirements.minimum is non-empty string

PCReqsHaveRec - (Boolean) True if pc_requirements.recommended is non-empty string

LinuxReqsHaveMin - (Boolean) True if linux_requirements.minimum is non-empty string

LinuxReqsHaveRec - (Boolean) True if linux_requirements.recommended is non-empty string

MacReqsHaveMin - (Boolean) True if mac_requirements.minimum is non-empty string

MacReqsHaveRec - (Boolean) True if mac_requirements.recommended is non-empty string

CategorySinglePlayer - (Boolean) True if for any i, categories[i].description is “single-player”

CategoryMultiplayer - (Boolean) True if for any i, categories[i].description is one of: “cross-platform multi-player”, “local multi-player”, “multi-player”, “online multi-player”, “shared/split screen”

CategoryCoop - (Boolean) True if for any i, categories[i].description is one of: “co-op”, “local co-op”, “online co-op”

CategoryMMO - (Boolean) True if for any i, categories[i].description is “mmo”

CategoryInAppPurchase - (Boolean) True if for any i, categories[i].description is “in-app purchases”

CategoryIncludeSrcSDK - (Boolean) True if for any i, categories[i].description is “includes source sdk”

CategoryIncludeLevelEditor - (Boolean) True if for any i, categories[i].description is “includes level editor”

CategoryVRSupport - (Boolean) True if for any i, categories[i].description is “vr support”

GenreIsNonGame - (Boolean) True if for any i, genres[i].description is one of: “utilities”, “design & illustration”, “animation & modeling”, “software training”, “education”, “audio production”, “video production”, “web publishing”, “photo editing”, “accounting”

GenreIsIndie - (Boolean) True if for any i, genres[i].description is “indie”

GenreIsAction - (Boolean) True if for any i, genres[i].description is “action”

GenreIsAdventure - (Boolean) True if for any i, genres[i].description is “adventure”

GenreIsCasual - (Boolean) True if for any i, genres[i].description is “casual”

GenreIsStrategy - (Boolean) True if for any i, genres[i].description is “strategy”

GenreIsRPG - (Boolean) True if for any i, genres[i].description is “rpg”

GenreIsSimulation - (Boolean) True if for any i, genres[i].description is “simulation”

GenreIsEarlyAccess - (Boolean) True if for any i, genres[i].description is “early access”

GenreIsFreeToPlay - (Boolean) True if for any i, genres[i].description is “free to play”

GenreIsSports - (Boolean) True if for any i, genres[i].description is “sports”

GenreIsRacing - (Boolean) True if for any i, genres[i].description is “racing”

GenreIsMassivelyMultiplayer - (Boolean) True if for any i, genres[i].description is “massively multiplayer”

PriceCurrency - (Text) price_overview.currency in JSON

PriceInitial - (Float) price_overview.initial in JSON, divided by 100.0 to converts cents to currency

PriceFinal - (Float) price_overview.final in JSON, divided by 100.0 to converts cents to currency

SteamSpyOwners - (steamspy.com) total owners, which includes free weekend trials and other possibly spurious numbers.

SteamSpyOwnersVariance - (steamspy.com) total owners, which includes free weekend trials and other possibly spurious numbers. Note that this is not technically variance: according to steamspy.com, “the real number… lies somewhere on… [value +/- variance]”

SteamSpyPlayersEstimate - (steamspy.com) best estimate of total number of people who have played the game since March 2009

SteamSpyPlayersVariance - (steamspy.com) errors bounds on SteamSpyPlayersEstimate. Note that this is not technically variance: according to steamspy.com, “the real number… lies somewhere on… [value +/- variance]”

SupportEmail - (Textual) support_info.email in JSON

SupportURL - (Textual) support_info.url in JSON

AboutText - (Textual) about_the_game in JSON

Background - (Textual) background in JSON

ShortDescrip - (Textual) short_description in JSON

DetailedDescrip - (Textual) detailed_description in JSON

DRMNotice - (Textual) drm_notice in JSON

ExtUserAcctNotice - (Textual) ext_user_account_notice in JSON

HeaderImage - (Textual) header_image in JSON

LegalNotice - (Textual) legal_notice in JSON

Reviews - (Textual) reviews in JSON

SupportedLanguages - (Textual) supported_languages in JSON

Website - (Textual) website in JSON

PCMinReqsText - (Textual) text of pc_requirements.minimum

PCRecReqsText - (Textual) text of pc_requirements.recommended

LinuxMinReqsText - (Textual) text of linux_requirements.minimum

LinuxRecReqsText - (Textual) text of linux_requirements.recommended

MacMinReqsText - (Textual) text of mac_requirements.minimum

MacRecReqsText - (Textual) text of mac_requirements.recommended

From the above, we can see that there is a lot of information that is hard to use to predict price. One example is the description. Unless we can extract some kind of metric from this, we won't be able to use it to predict price. As such, we choose to remove the following columns: QueryID, ResponseID, PCMinReqsText, PCRecReqsText, MacMinReqsText, MacRecReqsText, LinuxMinReqsText, LinuxRecReqsText, Reviews, LegalNotice, HeaderImage, DRMNotice, ExtUserAcctNotice, ShortDescrip, Background, AboutText, PriceCurrency.

At this point, we need to apply the following modifications in order to extract useful information out of the other columns:

The ReleaseDate was transformed into the Date object type, with incorrect dates replaced with na.

Supported Languages was modified to instead contain a numeric value of the total number of supported languages

Full-length Descriptions was replaced with the total number of words in each description.

Support Email and Support URL were consolidated into a single binary variable called Support and marked TRUE if either URL or Email were provided.

Website was replaced with a binary variable indicating whether or not a website was provided.

All existing true/false columns were transformed into factors.

We decided on the following methods for data analysis:

Principal Component Analysis

Random Forests

Linear Regression

Results:

Tables, figures, and text that illustrate your findings. Keep the focus on the numbers here. You will interpret your results in the next section.

Principal Components Analysis

Using principal component analysis, we tried to reduce the dimensions of the data given to something that is easier to interpret. The components extracted using the PCA function are as follows:

```
##                FreeVerAvailTrue           ControllerSupportTrue
##                     0.074508528                     0.070456149
##                     ReleaseDate              GenreIsNonGameTrue
##                     0.045623039                     0.043416614
##             PlatformWindowsTrue           GenreIsFreeToPlayTrue
##                     0.004339572                    -0.003198739
##           SubscriptionAvailTrue               GenreIsSportsTrue
##                    -0.008387476                    -0.013410307
##           CategoryVRSupportTrue               GenreIsCasualTrue
##                    -0.014164153                    -0.014987983
##                 CategoryMMOTrue       CategoryInAppPurchaseTrue
##                    -0.017014156                    -0.017425834
##               GenreIsRacingTrue       CategoryIncludeSrcSDKTrue
##                    -0.019477385                    -0.019621476
## GenreIsMassivelyMultiplayerTrue          GenreIsEarlyAccessTrue
##                    -0.022229404                    -0.029657569
##                     RequiredAge                        DLCCount
##                    -0.035270958                    -0.041101655
##           GenreIsSimulationTrue                       DemoCount
##                    -0.045608914                    -0.058909342
##                  GenreIsRPGTrue          ControllerSupportFalse
##                    -0.069529578                    -0.070456149
##             RecommendationCount  CategoryIncludeLevelEditorTrue
##                    -0.072258215                    -0.075454092
##             GenreIsStrategyTrue            GenreIsAdventureTrue
##                    -0.076956068                    -0.082023865
##         SteamSpyPlayersEstimate                  SteamSpyOwners
##                    -0.083716051                    -0.087388876
##                CategoryCoopTrue               GenreIsActionTrue
##                    -0.094873905                    -0.099024698
##         CategoryMultiplayerTrue                      Metacritic
##                    -0.104781595                    -0.118210773
##                AchievementCount              SupportedLanguages
##                    -0.125752326                    -0.128166381
##               PCReqsHaveRecTrue                    PackageCount
##                    -0.138874371                    -0.149434076
##                GenreIsIndieTrue               PurchaseAvailTrue
##                    -0.155333942                    -0.164471179
##                 DetailedDescrip                  DeveloperCount
##                    -0.167448077                    -0.186315032
##                     WebsiteTRUE                      MovieCount
##                    -0.187502045                    -0.191829546
##                 ScreenshotCount            LinuxReqsHaveRecTrue
##                    -0.194203024                    -0.200554516
##     AchievementHighlightedCount               PlatformLinuxTrue
##                    -0.201618269                    -0.202732910
##              MacReqsHaveRecTrue        CategorySinglePlayerTrue
##                    -0.206223338                    -0.217801564
##                 PlatformMacTrue            LinuxReqsHaveMinTrue
##                    -0.220588387                    -0.238113499
##                     SupportTRUE                  PublisherCount
##                    -0.238153760                    -0.244145291
##              MacReqsHaveMinTrue               PCReqsHaveMinTrue
##                    -0.249825736                    -0.297218170
```

This first component explains the most variance out of all the components. The most significant coefficient indicates whether a Free version is avaiable. This makes sense because games that have a frere or trial version will get players hooked to purchase the full, more costly version. Release Date is another significant Coefficient, and suggests that newer games have higher prices.

```
##   ControllerSupportFalse        PCReqsHaveMinTrue           PublisherCount
##               0.26799885               0.23861955               0.19613841
##          ScreenshotCount          DetailedDescrip CategorySinglePlayerTrue
##               0.14893131               0.14595983               0.12773316
##           DeveloperCount             PackageCount        PurchaseAvailTrue
##               0.12171039               0.11257952               0.11030653
##  CategoryMultiplayerTrue
##               0.09904693
##  [ reached getOption("max.print") -- omitted 44 entries ]
```

The second component seems indicative of PC games due to the highest component being lack of controller support. Thus, we can conclude this category contains older AAA high budget games due to the low emphasis on release date as well as a high emphasis on number of Publishers and Developers.

```
## CategorySinglePlayerTrue        PurchaseAvailTrue        PCReqsHaveMinTrue
##               0.25858586               0.24076940               0.14102728
##        GenreIsCasualTrue         GenreIsIndieTrue   ControllerSupportFalse
##               0.13685124               0.12831029               0.11018263
##           PublisherCount     GenreIsAdventureTrue              SupportTRUE
##               0.10219650               0.09127584               0.08888365
##           DeveloperCount
##               0.06325220
##  [ reached getOption("max.print") -- omitted 44 entries ]
```

The third component describes niche indie games. We can conclude this because of the positive emphasis on the coefficients of Single Plater Games, Casual Games, and Indie Games as well as a negative emphasis on Number of owners and Number of Players.

```
## GenreIsMassivelyMultiplayerTrue                     ReleaseDate
##                       0.2963920                       0.2948865
##                 CategoryMMOTrue           GenreIsFreeToPlayTrue
##                       0.2843971                       0.2840576
##                     SupportTRUE          GenreIsEarlyAccessTrue
##                       0.2262849                       0.1853767
##          ControllerSupportFalse       CategoryInAppPurchaseTrue
##                       0.1646891                       0.1537688
##                GenreIsIndieTrue                  GenreIsRPGTrue
##                       0.1485697                       0.1083563
##  [ reached getOption("max.print") -- omitted 44 entries ]
```

The fourth component seems to describe newer multiplayer games. This is due to the high positive coefficients of Massive Multiplayer Genre, Release Date, and MMO Category.

We couldn't accurately determine the subset of the fifth component, so we decided to stop any further PCA analysis.

Tree

Initial Price

Metacritic

Linear Regression

Conclusion

Interpret what you found. What are the main lessons we should take away from your report?