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

Abstract

The objective of this paper is to determine the most significant predictors of a game's price and sales. Using Principle Component Analysis(PCA) and Random Forests, we were able to identify characteristics of games that predicted either a high price or greater sales. We discovered that having a good marketing strategy and catering to those who want a casual single player game will help boost sales and allow for a higher initial price.

Introduction

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

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.

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:

Due to the high dimensionality of the dataset, we decided that it would be helpful to reduce it to its most significant components so that we could observe any trends that would help to describe the initial price.

Random Forests

Linear Regression

Results:

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:

Initial Price

Table 1

```
##      FreeVerAvailTrue ControllerSupportTrue           ReleaseDate
##           0.074508528           0.070456149           0.045623039
##    GenreIsNonGameTrue   PlatformWindowsTrue GenreIsFreeToPlayTrue
##           0.043416614           0.004339572          -0.003198739
## SubscriptionAvailTrue     GenreIsSportsTrue CategoryVRSupportTrue
##          -0.008387476          -0.013410307          -0.014164153
##     GenreIsCasualTrue
##          -0.014987983
##  [ reached getOption("max.print") -- omitted 44 entries ]
```

Table 2

```
##   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 ]
```

Table 3

```
## 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 ]
```

Table 4

```
## 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 ]
```

Sales

Table 5

```
##           PCReqsHaveMinTrue              PublisherCount
##                   0.3014106                   0.2475819
##          MacReqsHaveMinTrue                 SupportTRUE
##                   0.2449544                   0.2406077
##        LinuxReqsHaveMinTrue    CategorySinglePlayerTrue
##                   0.2327935                   0.2208328
##             PlatformMacTrue          MacReqsHaveRecTrue
##                   0.2150283                   0.2041172
## AchievementHighlightedCount        LinuxReqsHaveRecTrue
##                   0.2011133                   0.1982281
##  [ reached getOption("max.print") -- omitted 44 entries ]
```

Table 6

```
##     PlatformLinuxTrue  LinuxReqsHaveMinTrue       PlatformMacTrue
##             0.35228092            0.33061232            0.32858360
##    MacReqsHaveMinTrue ControllerSupportTrue  LinuxReqsHaveRecTrue
##             0.30190186            0.26014042            0.23926442
##    MacReqsHaveRecTrue           ReleaseDate      GenreIsIndieTrue
##             0.21823165            0.16998588            0.11573785
##      FreeVerAvailTrue
##             0.09483667
##  [ reached getOption("max.print") -- omitted 44 entries ]
```

Table 7

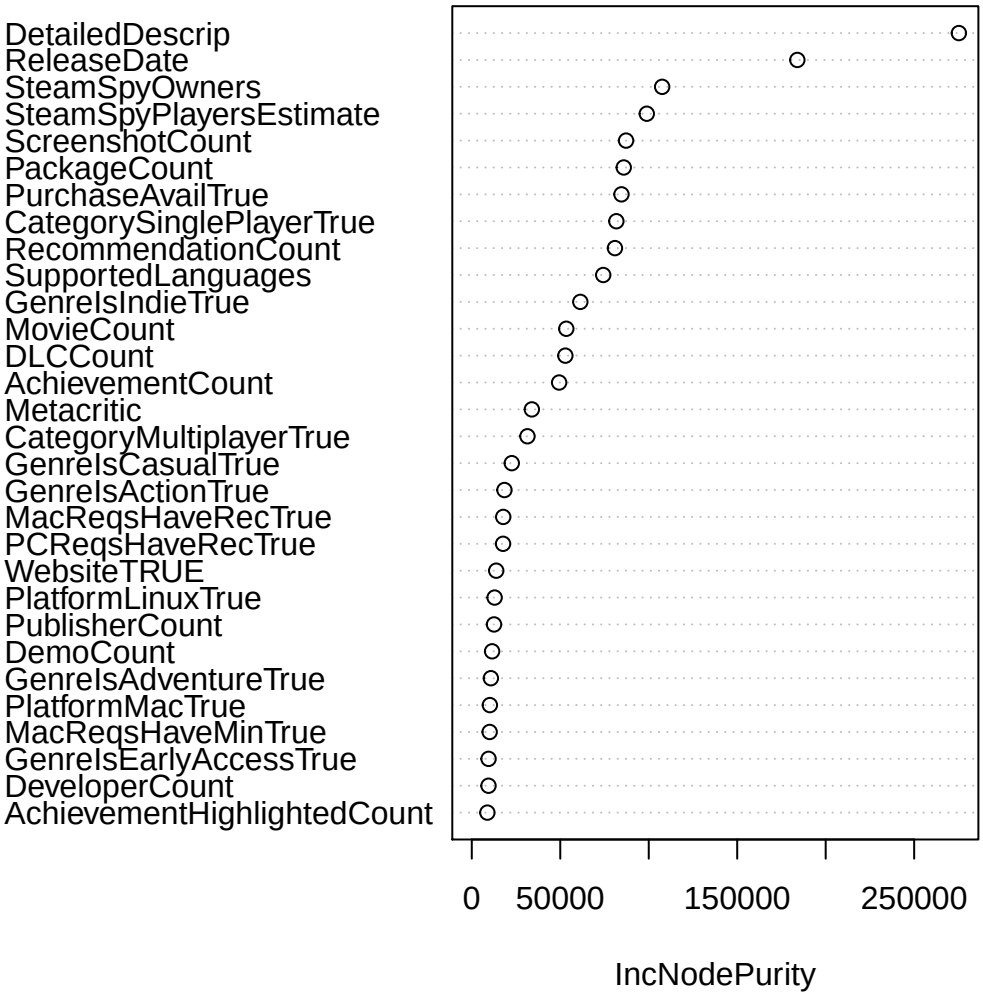
```
## GenreIsMassivelyMultiplayerTrue                 CategoryMMOTrue
##                       0.4200411                       0.4090913
##           GenreIsFreeToPlayTrue       CategoryInAppPurchaseTrue
##                       0.3672820                       0.2548077
##         CategoryMultiplayerTrue                CategoryCoopTrue
##                       0.2514441                       0.2013751
##         SteamSpyPlayersEstimate             RecommendationCount
##                       0.1498070                       0.1262579
##                      MovieCount                   GenreIsRPGTrue
##                       0.1254186                       0.1207485
##  [ reached getOption("max.print") -- omitted 44 entries ]
```

Table 8

```
##              Metacritic            PackageCount   ControllerSupportTrue
##               0.3143453               0.2787179               0.2662536
##             RequiredAge      SupportedLanguages            PriceInitial
##               0.1917889               0.1828120               0.1818398
## SteamSpyPlayersEstimate        AchievementCount     RecommendationCount
##               0.1721295               0.1696391               0.1674830
## CategoryMultiplayerTrue
##               0.1410278
##  [ reached getOption("max.print") -- omitted 44 entries ]
```

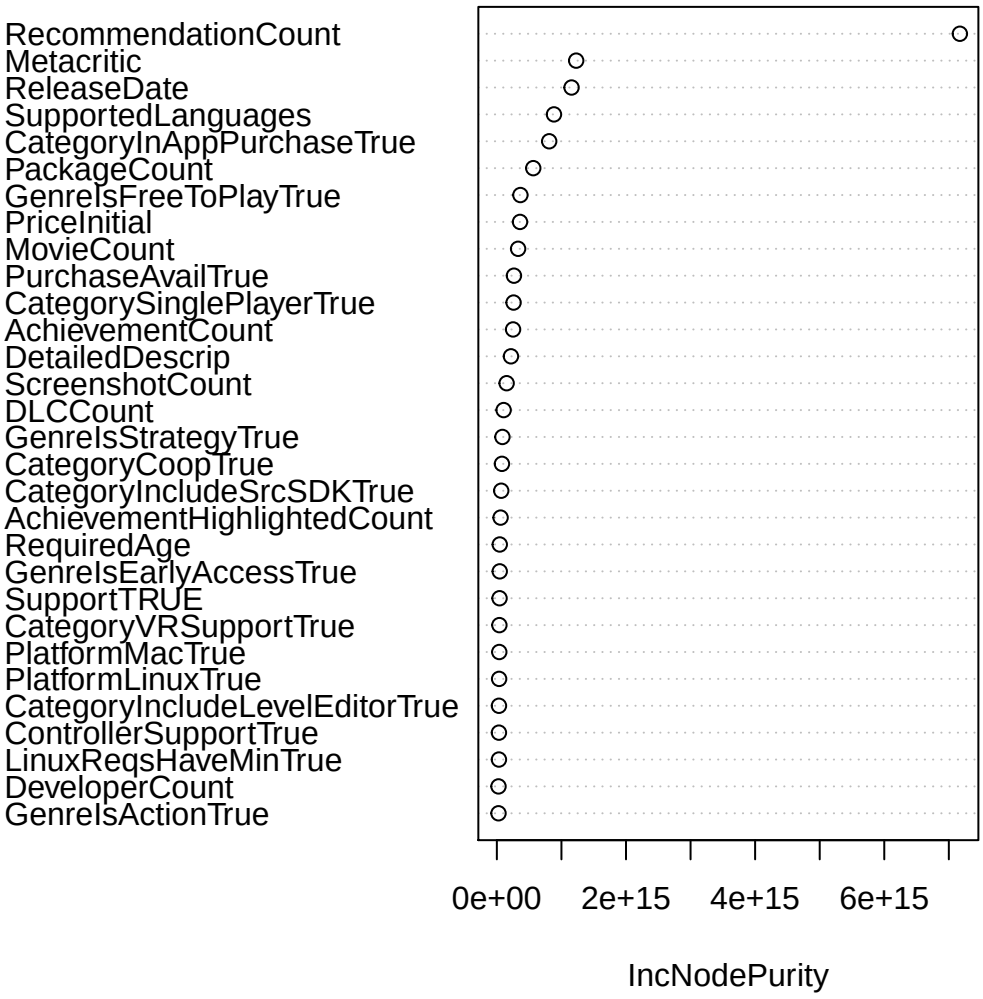
Random Forest

Figure 1



Initial Price This is the plot of the importance of each feature in terms of predicting initial price.

Figure 2



Sales This is the plot of the importance of each feature in terms of predicting sales.

Conclusion

PCA

Initial Price This first component (Table 1) 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 free 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. The second component (Table 2) 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. The third component (Table 3) describes niche indie games. We can conclude this because of the positive emphasis on the coefficients of

Single Player Games, Casual Games, and Indie Games as well as a negative emphasis on Number of owners and Number of Players. The fourth component (Table 4) 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.

Sales Similar to above, we were able to glean some information from the first 4 principal components. The first component (Table 5) shows that the most prominent group of users buy games that have a reputable studio behind them, have detailed and rich media on their homepage and are available on many platforms. The second component (Table 6) takes into account a group on the opposite spectrum, who buys games that aren't available on mainstream platforms like indie games. Component 3 (Table 7) describes massively multiplayer online games that are free to play, something that comprises a large section of the market. Finally component 4 (Table 8) shows that a large portion of the market is older gamers influenced by reviews and the pricing of the games.

Random Forest

Initial Price From Figure 1 we can see a large difference between the top 2 features and the rest of the graph. The description and release date have the most impact on the accuracy of the random forest, showing that how much effort a company puts into marketing and the when the game was published matters the most when a company is pricing its products.

Sales As for Figure 2, a similar difference can be seen between the top 3 features and the rest of the graph. However, there is an even larger difference within the top 3, with recommendations shooting past the rest of the features by a large margin. This shows that most people take into account what other people are saying about a game when they are considering what to purchase. They then look at when the game was published and what reviewers are saying about the product.

Summary From our perspective as future game developers, we can identify areas that would allow us to determine the best price and predict where we would get the highest amount of profit. From the PCA analysis, we determined that fleshing out the website and adding vivid descriptions and multimedia will allow us to attract as many players as possible. In addition, making a casual single player game will make it more likely for players to tolerate a higher price.

From the Random Forest analysis, we can see similar results. In order to price our product higher, we need to ensure that our marketing campaign generates enough interest to justify the price. In order to attract as many players as possible, the game needs to be released to reviewers and perhaps a large beta test player pool needs to be generated so that people can spread word of the game around by mouth.