Final Report

Brayden Hoskins, Noah Estrada-Rand, Charlie Filce

December 12, 2019

Business Problem

Our group was interested in finding what characteristics of video games lead to a successful game. Our results would give developers an insight into what specific features should be focused on when developing their games to maximize success. Since there are many different variables within each game, we looked to find key indicators of a successful game.

Motivation

The gaming industry is projected to reach an annual revenue of $230 billion dollars by the year 2020. Since the gaming industry is quickly expanding into one of the highest grossing industries in the world, we wanted to see what types of games are pushing this industry to the top. When looking at different successful games in our data set, at times there can be a large disconnect between the resources a game was developed with and the total number of copies sold. Because of this we wanted to analyze what specific features led to the less well known games being considered as successful as the bigger, more popular games.

Summary Statistics

The Steam Store dataset consisted of 27,075 observations, or different games sold on the Steam store, and 18 different features, or characteristics of the games. This data set was provided by scraping data from the Steam store by a third party service, Steam Spy. Our data set consisted of 7 numeric variables: Required Age, Number of Achievements, Positive Ratings, Negative Ratings, Average Playtime, Median Playtime, and Price. The dataset also consisted of 11 factored variables: Application Id, Name, Release Date, English, Developer, Publisher, Platform, Categories, Genres, Steam Spy set, and Number Of Owners. We also created 2 factored variables: Simple_Categories, and SuccessfulGame. Ultimately we kept only the variables of required_age, genres, achievements, average_playtime, price, simple_categories, and successfulGame.

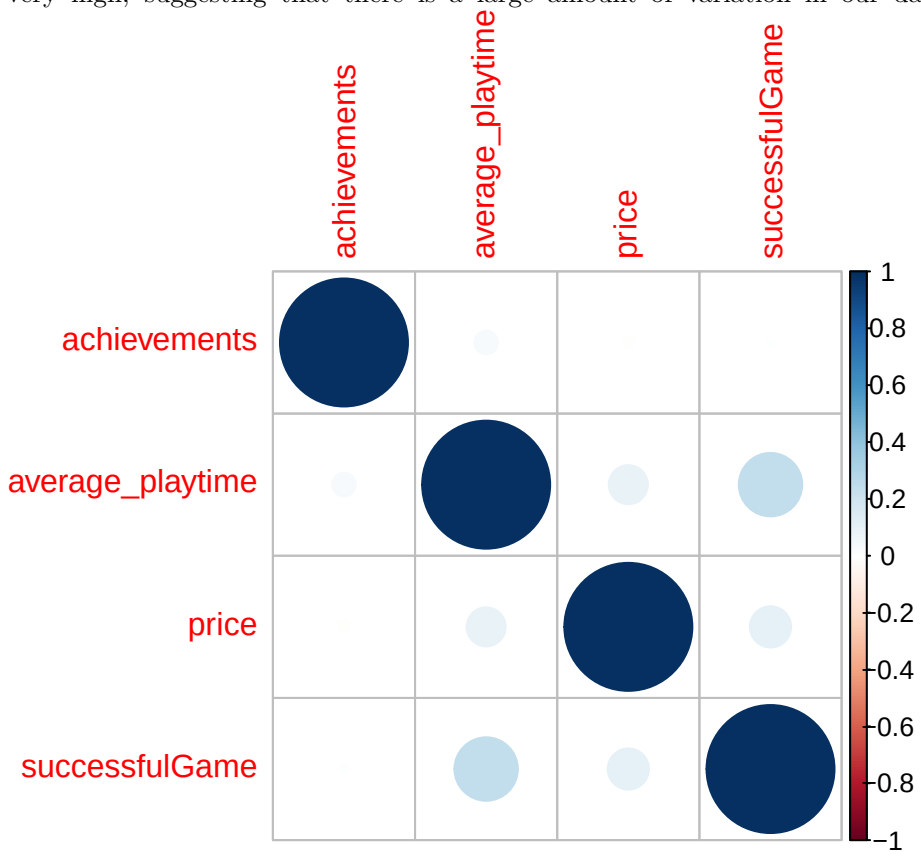
The drastic decrease in number of variables is due to the fact that many of the variables proved to be unusable in their current state. Many of them contained three semicolon delimited values, ultimately making it difficult to distinguish a discernable category or value for that particular variable. Moreover, in regards to variables such as publisher and developer, it was not possible to build meaningful evaluations off of the data as there were over 10,000 different possible classes. As such, we used the variables that were still impactful and meaningful to our analyses. From these remaining variables, the following summary statistics were calculated.

```
## Non-numerical variable(s) ignored: required_age, genres, simple_categories
```

```
## Descriptive Statistics
## steam
## N: 26876
##
##                     achievements   average_playtime      price   successfulGame
## ----------------- -------------- ------------------ ---------- ----------------
##               Mean          45.31             117.46       7.33             0.02
##            Std.Dev         353.96             827.80       7.34             0.14
##                Min           0.00               0.00       0.00             0.00
##                 Q1           0.00               0.00       2.16             0.00
##             Median           7.00               0.00       5.11             0.00
##                 Q3          23.00               0.00       9.20             0.00
##                Max        9821.00           38805.00      49.91             1.00
```

```
##                MAD      10.38         0.00      5.69         0.00
##                IQR      23.00         0.00      7.04         0.00
##                 CV       7.81         7.05      1.00         7.04
##           Skewness      13.38        22.16      1.89         6.89
##       SE.Skewness       0.01         0.01      0.01         0.01
##           Kurtosis     189.74       676.18      4.66        45.54
##            N.Valid   26876.00     26876.00  26876.00     26876.00
##          Pct.Valid     100.00       100.00    100.00       100.00
```

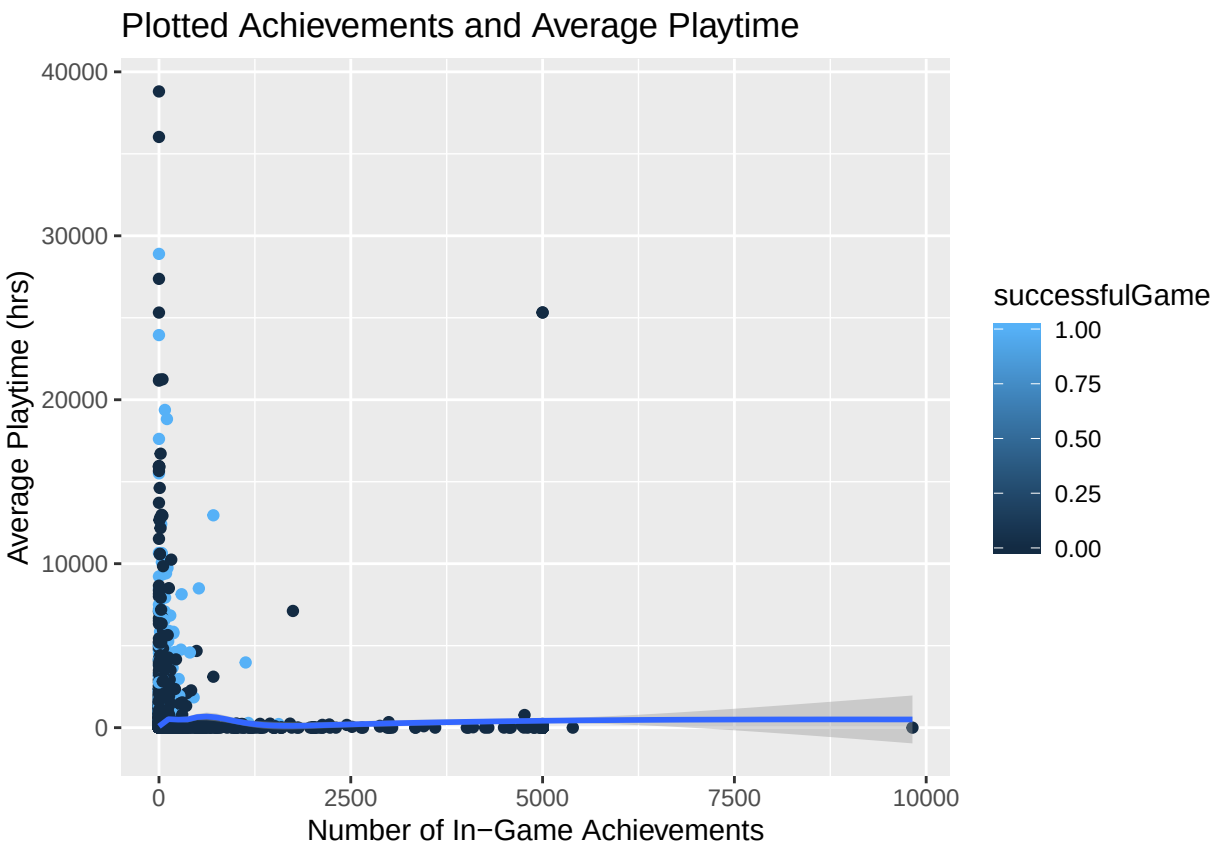
From the table above there are a few significant things to point out. Our data produced a mean of .02 for the number of successful games. This is significant because it points out that the overwhelming majority of the games in this data set we consider to be unsuccessful games. There were some extreme outliers for achievements and average playtime, so when we ran our models we removed these observations from the dataset to prevent skewing our results. We also found it very interesting that the number of achievements in games was rather low, and it would most likely drop a significant amount if we had removed the outliers when we generated this table. The standard deviations for average playtime and achievements is very high, suggesting that there is a large amount of variation in our data set.



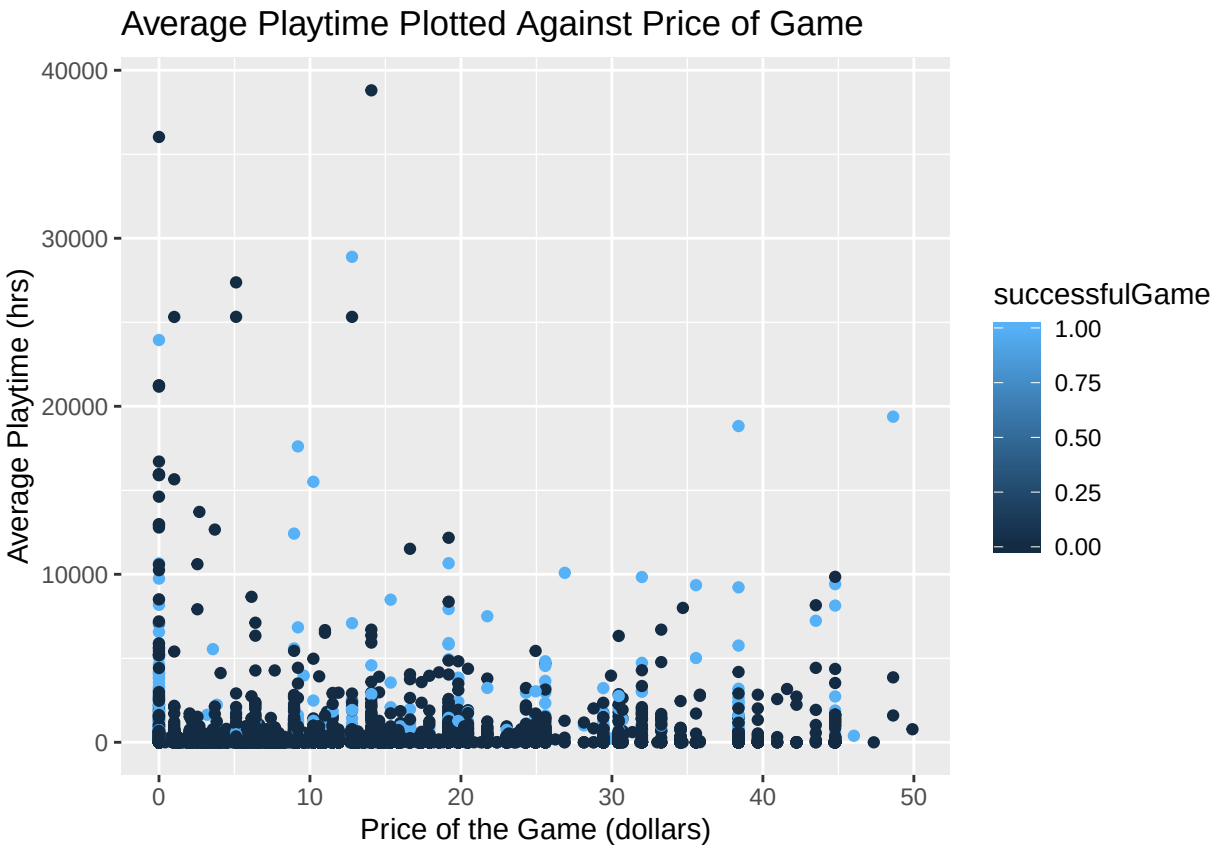
From the correlation matrix above you can see that there are 2 variables that are correlated with a successful game: price and average playtime. This made sense to us for two main reasons: the first is that if a games price is low or free it is more accessible to the public, and the second reason is that if a game has a high average playtime there must be a lot of content in the game that keeps its players for long periods of time. If the game is more accessible to the public then more players will be drawn to it, thus increasing popularity and the potential for players to buy in game cosmetics or expansions. If a game has a high average playtime, it shows that there is enough content to keep players interested in the game. To us average playtime provided a lot of information into a games retention rate of its players, if the retention rate was higher it is more likely that the game is successful.

Summary Plots

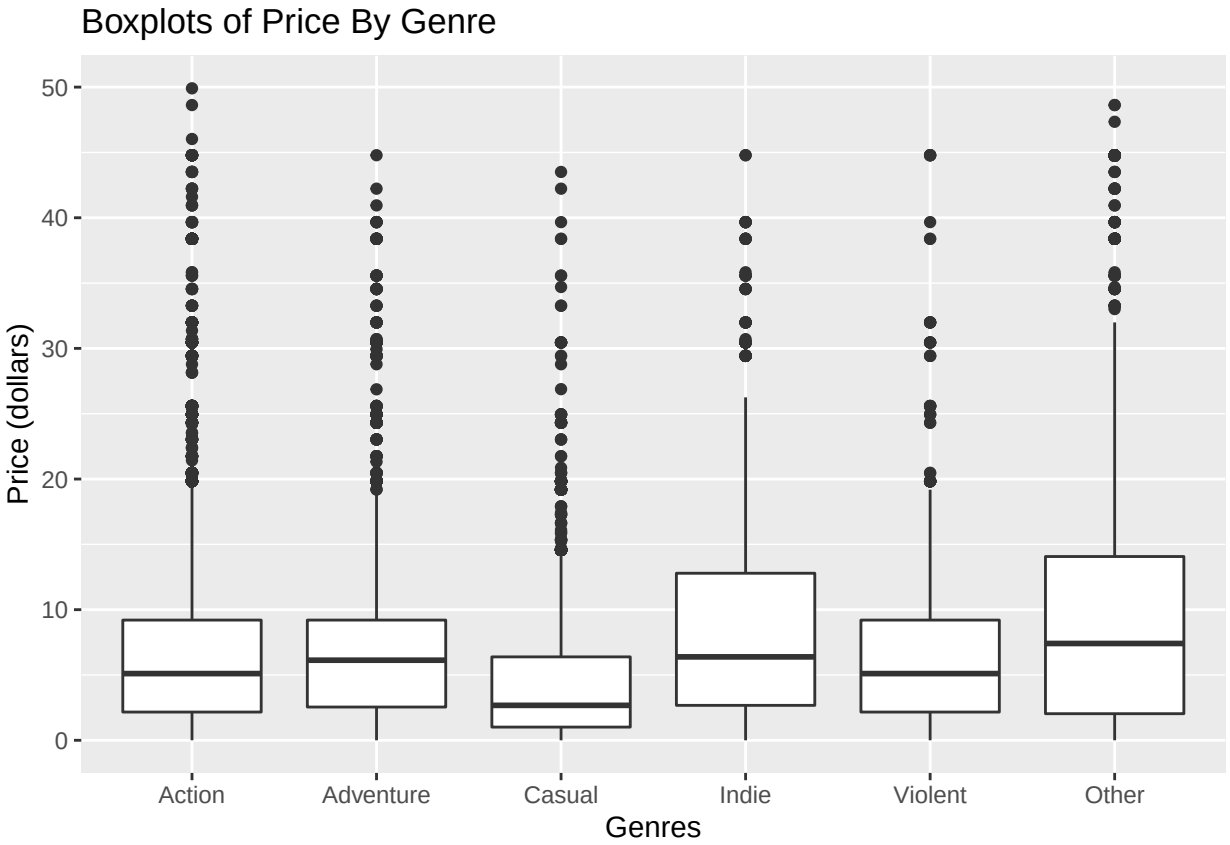
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



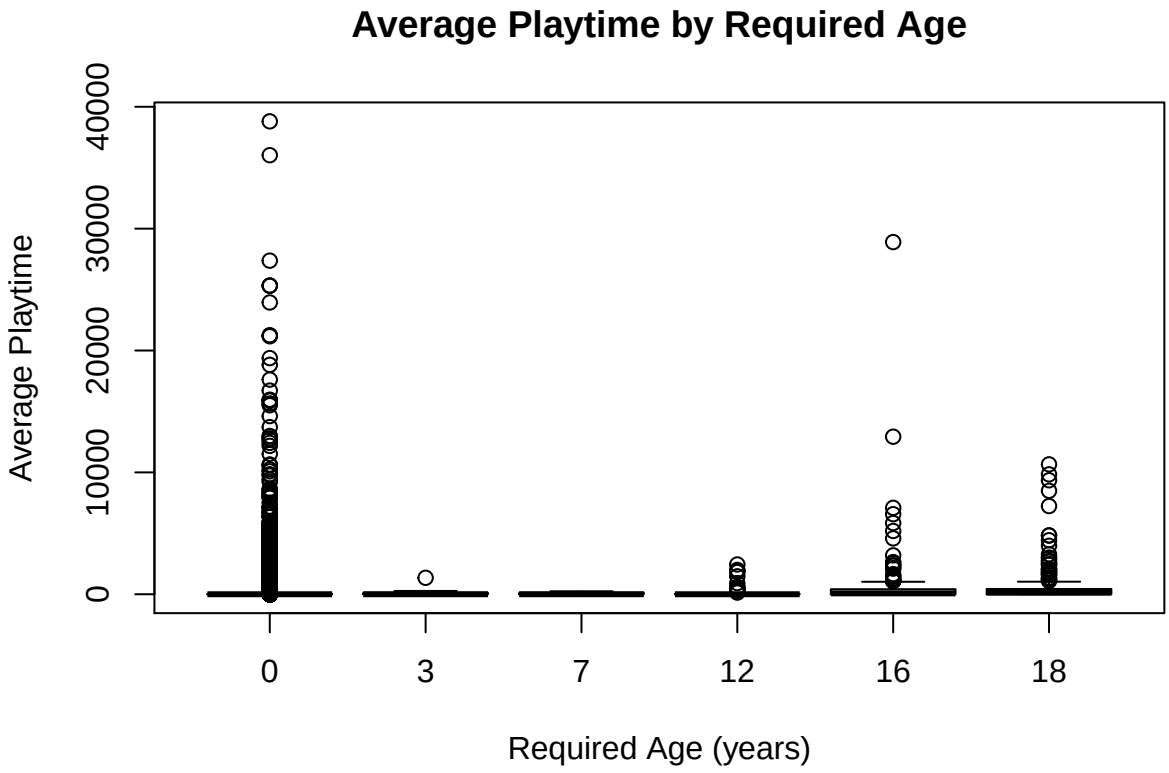
In looking to find impactful variables to base our analyses on, we initially believed that more achievements in a game would lead to higher average playtime. This was not the case, however, as the plot above clearly shows that some of the games with the lowest achievements have the highest average playtime.



From the plot above it becomes clear that price does not necessarily translate to higher average playtime. This is characterized by the equal dispersion of high average playtimes amongst the different price ranges. Furthermore, it is intersting to notice that successfulgames are also equally distributed among the varying price ranges.



The boxplot above shows the distribution of prices for video games against the different kinds of genres. While many of the genres have similar ranges, it is interesting to notice that casual games clearly have the lowest mean price and range of prices. Moreover, it is important to notice the `other'' and`indie" categories as having th widest range of prices, perhaps due to the fact that many of these games are perhaps created by independent studios, leading them to demand higher prices as they are unable to produce the games for lower costs.



As seen in the plot above, it appears that games with essentially no age requirement have a much broader range of average playtimes than any other level of required age. This intuitively makes sense because the more people that can play a game, the more the game will be played, and thus a higher average playtime overall.

Models

In looking to fit models to help solve our business problem, we decided to use logistic regression, random forest, and tree models.

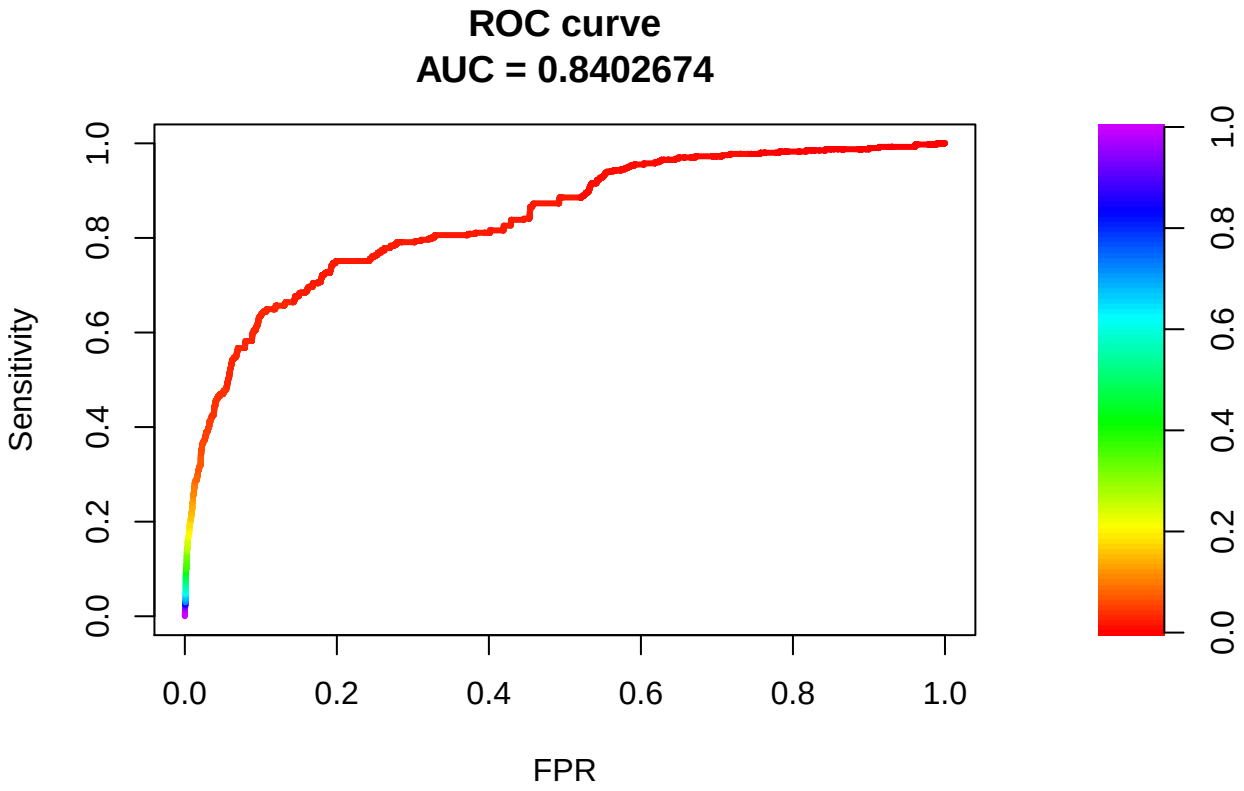
Logistic Regression

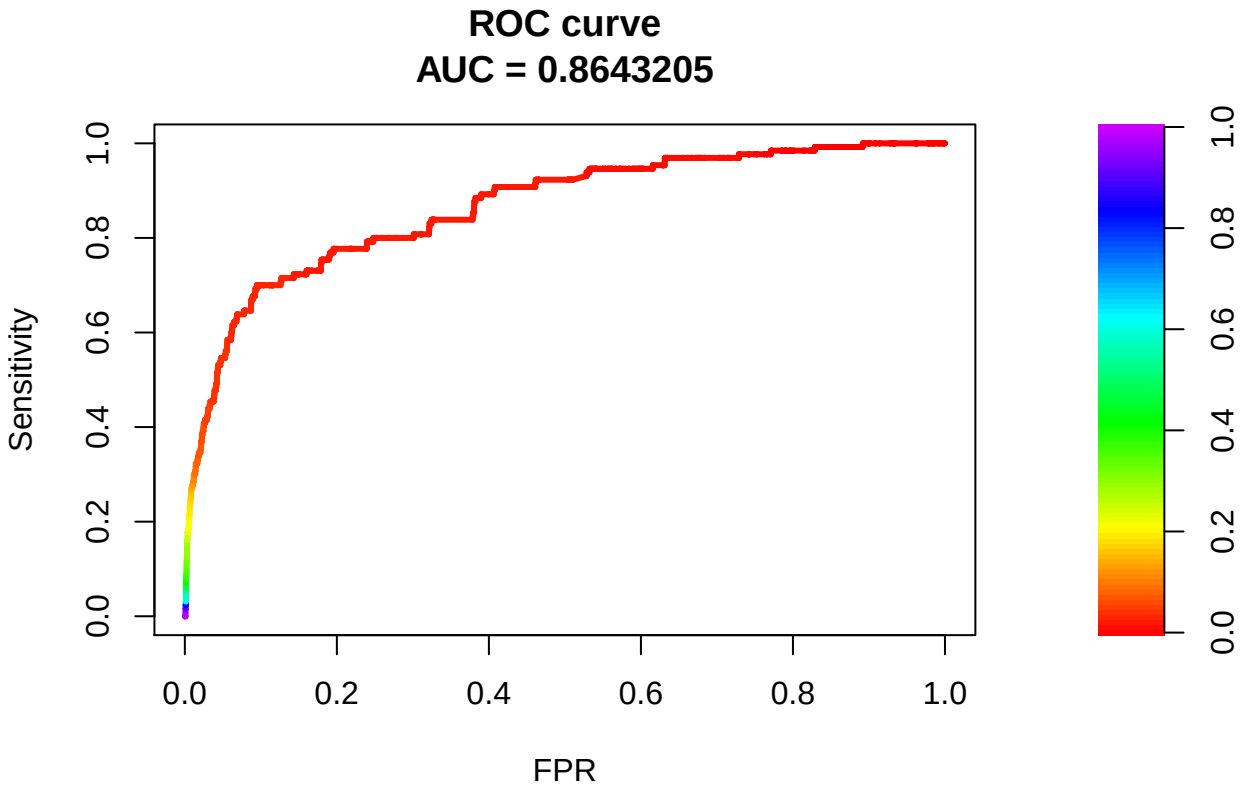
In building our logistic regression model, we first attempted using all variables as predictors of our successful-Game variable. From this initial model we were able to see which variables were statistically significant and proceeded to use these significant variables to build a new logistic model. The variables used as predictors were price, genres, required age, and average playtime.

```
##                              (Intercept)
##                            2.005074e-02
##                                   price
##                            1.035616e+00
##    relevel(genres, ref = "Other")Action
##                            7.948691e-01
## relevel(genres, ref = "Other")Adventure
##                            2.321590e-01
##    relevel(genres, ref = "Other")Casual
##                            1.396069e-01
```

```
##     relevel(genres, ref = "Other")Indie
##                           2.174737e-01
##   relevel(genres, ref = "Other")Violent
##                           8.408791e-02
##                          required_age3
##                           4.935156e-05
##                          required_age7
##                           2.293985e-05
##                         required_age12
##                           2.726689e+00
##                         required_age16
##                           5.454452e+00
##                         required_age18
##                           1.042955e+01
##                       average_playtime
##                           1.000477e+00
```

Looking at price, the model suggests that for every unit increase in price, a dollar, the likelihood of the game being successful rises by three percent. Moreover, looking at the different genres of games, it becomes clear that any game not in the "Other" category leads to a decrease in likelihood of a game being successful. Lastly, when considering the coefficients for required age, it is interesting to notice that a game being or required age 12, 16, or 18 leads to a significant increase in the likelihood of a game being successful. We interpreted the coefficients above as indicating that having a game considered in the "Other" category as well as making the age requirement higher leads to the largest increases in likelihood of a game being successful. However, it is very important to take this interpretation lightly as we must also consider what other variables may play into this effect. In the case of age requirements, perhaps it is the more mature content of the video games, and not necessarily the age requirement itself that drives the increase in likelihood of a game being successful or not. Moreover, a game being considered in the "Other" genre may increase the likelihood of a game succeeding as it breaks the traditional genre molds of most games, indicative of innovative and novel game development.


ROC curve
AUC = 0.8402674
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
FPR
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0



Looking at the ROC plots above, it is clear to see that the optimal cutoff for our logistic model resides between 0.1 and 0, indicative of a very low cutoff to maximize true positives and minimize false positives.

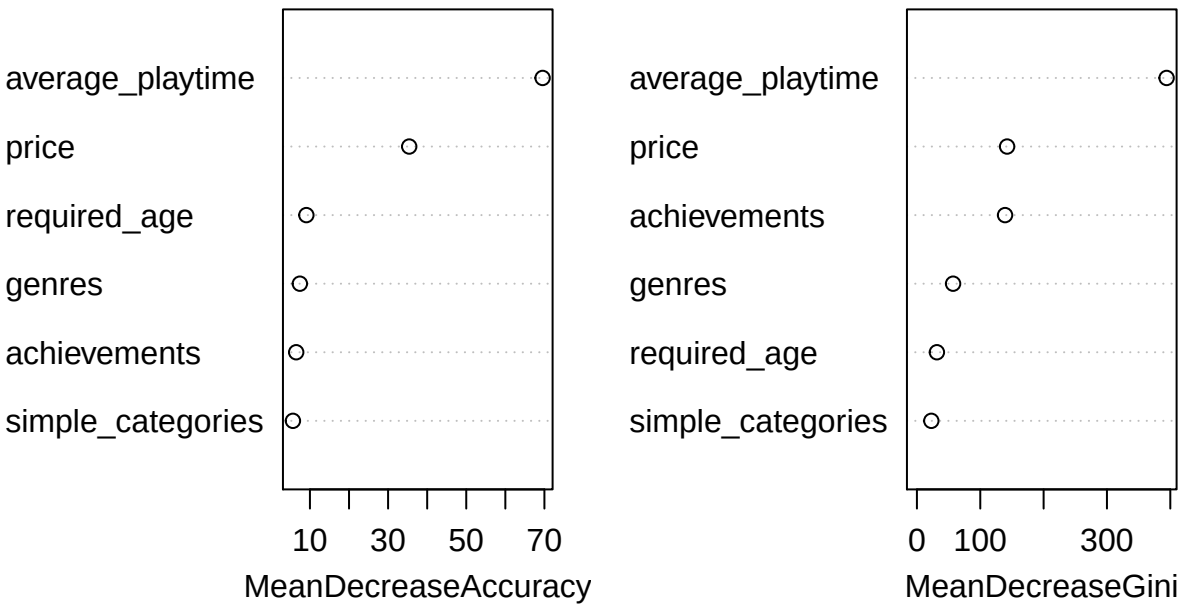
```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  20157
##
##
##                          | steam_train$successfulGame
## steam_train$pred_class |         0 |         1 | Row Total |
## -----------------------|-----------|-----------|-----------|
##                      0 |     18916 |       217 |     19133 |
## -----------------------|-----------|-----------|-----------|
##                      1 |       839 |       185 |      1024 |
## -----------------------|-----------|-----------|-----------|
##           Column Total |     19755 |       402 |     20157 |
## -----------------------|-----------|-----------|-----------|
##
##

##
##
```

```
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  6719
##
##
##                       | steam_test$successfulGame
## steam_test$pred_class |         0 |         1 | Row Total |
## ----------------------|-----------|-----------|-----------|
##                     0 |      6310 |        63 |      6373 |
## ----------------------|-----------|-----------|-----------|
##                     1 |       279 |        67 |       346 |
## ----------------------|-----------|-----------|-----------|
##          Column Total |      6589 |       130 |      6719 |
## ----------------------|-----------|-----------|-----------|
##
##
```

After considering different cutoffs between 0.1 and 0, we decided to use a cutoff of .04 as going any closer to 0 would increase sensitivity while sharply penalizing specificity. Although the accuracy, and the specificity of this model is relatively strong on both test and training datasets, it becomes clear that there is much to be desired in terms of sensitivity. With this in mind we set out to improve sensitivity while maintaining the optimal levels of specificity and accuracy.

RandomForest

Next, we began to build a random forest model to hopefully increase the sensitivity of our predictions. In order to do so, we used cross validation to find the best number of variables to consider at each split of each tree. We plotted the out of bag error against the number of variables tried and found the optimal number to be 2. With this in mind we built our random forest model which yielded the insightful results.



As seen in the importance plot the variables of average playtime and price are the two most important variables in the 500 trees made in the random forest model. This largely aligned with the "random forest explained" results that are included outside of this report due to their formatting. The figures included in that report indicate that average playtime and required age were used as roots the most while also having the lowest average depth in each tree, once again highlighting the most important variables as predictors.

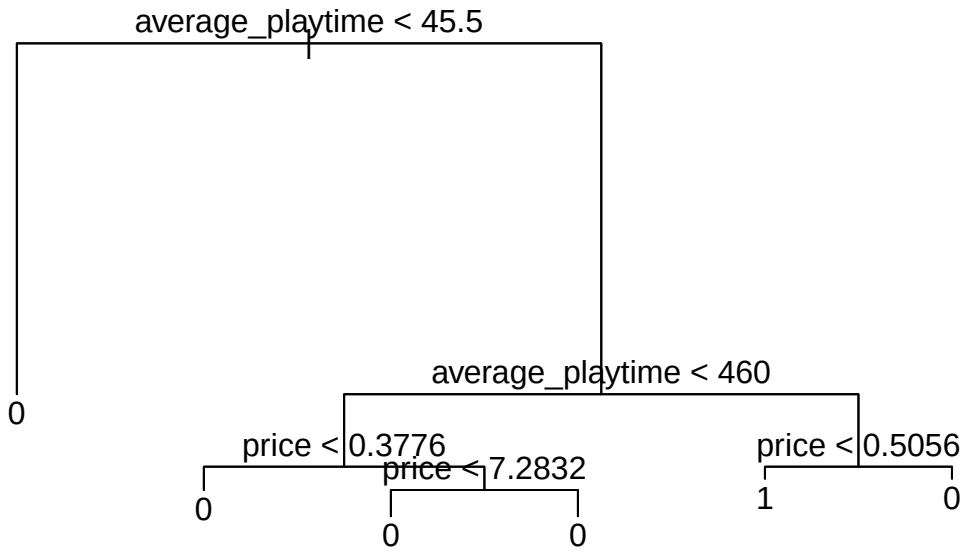
```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  20157
##
##
##                                    | preds_train_bagged$bag_preds
## preds_train_bagged$successfulGame |         0 |         1 | Row Total |
## ----------------------------------|-----------|-----------|-----------|
##                                 0 |     19631 |       124 |     19755 |
## ----------------------------------|-----------|-----------|-----------|
##                                 1 |       295 |       107 |       402 |
## ----------------------------------|-----------|-----------|-----------|
##                      Column Total |     19926 |       231 |     20157 |
## ----------------------------------|-----------|-----------|-----------|
##
```

```
##
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  6719
##
##
##                                  | preds_test_bagged$bag_preds
## preds_test_bagged$successfulGame |         0 |         1 | Row Total |
## ---------------------------------|-----------|-----------|-----------|
##                                0 |      6558 |        31 |      6589 |
## ---------------------------------|-----------|-----------|-----------|
##                                1 |        95 |        35 |       130 |
## ---------------------------------|-----------|-----------|-----------|
##                     Column Total |      6653 |        66 |      6719 |
## ---------------------------------|-----------|-----------|-----------|
##
##
```

After making predictions using the random forest model, the above confusion matrices were produced. Here it is clear that the random forest model was able to maintain the high level of accuracy and specificity while also increasing the sensitivity of the model. An increase of roughly 20-30% in sensitivity was able to make this model much more effective at predicting if a game is successful. Thus this model was able to outperform the initial logistic regression model.

Pruned Tree

While the random forest model offered us increased performance compared to the logit model, we still looked for the possibility of a simpler, yet equally effective model. Thus, we fitted a basic tree model using all possible variables as predictors. We then pruned the tree to find the number of splits to give us the lowest amount of deviance from the truth data, only to find the pruned tree to be identical to the initial tree.



The pruned tree model appeared to draw the same conclusions as the random forest model, indicating average playtime and price to be the most important variables to create splits off of.

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  20157
##
##
##                                  | basic_preds_train$preds
## basic_preds_train$successfulGame |         0 |         1 | Row Total |
## ---------------------------------|-----------|-----------|-----------|
##                                0 |     19699 |        56 |     19755 |
## ---------------------------------|-----------|-----------|-----------|
##                                1 |       337 |        65 |       402 |
## ---------------------------------|-----------|-----------|-----------|
##                     Column Total |     20036 |       121 |     20157 |
## ---------------------------------|-----------|-----------|-----------|
##
##

##
##
```

```
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  6719
##
##
##                                  | basic_preds_test$preds
## basic_preds_test$successfulGame |         0 |         1 | Row Total |
## --------------------------------|-----------|-----------|-----------|
##                               0 |      6569 |        20 |      6589 |
## --------------------------------|-----------|-----------|-----------|
##                               1 |       115 |        15 |       130 |
## --------------------------------|-----------|-----------|-----------|
##                    Column Total |      6684 |        35 |      6719 |
## --------------------------------|-----------|-----------|-----------|
##
##
```

From the pruned tree model, we once again made predictions using the model to observe its performance against the truth data. Ultimately, we found the tree model to perform similarly to the logistic model in regards to sensitivity, whilst maintaining high accuracy and specificity, just as the other two models had. Thus, while this model was able to offer a human readable model, it offered lower sensitivity than the random forest model.

Conclusion

From the results of the Logistic Model, we recommend game developers place a focus on restricting the age group of their audience to 18+, because the current successful games are often more violent or feature mature themes, and are therefore restricted to the 18+ audience. In both the Random Forest Model and the Pruned Tree Model, we saw that Playtime and Price are the most important variables in determining the success of a game. When using these models, we suggest using the pruned tree to make fast, easily read decisions and the Random Forest when in a situation where there is time to compute. Overall, we recommend that game developers create games with a free to play model and maximize profits through in-game stores and subscription based perks. According to 2/3 of our models, this approach will maximize the amount of copies sold and help the game reach the largest audience. Moving forward, in order to further refine our models, we would like to know what types of in-game purchases have the best results between battle passes, loot boxes and other game-specific purchases. Our current data doesn't provide information on this, so we would need new datasets that provide in-game purchase statistics.