Noah Estrada-Rand, Charlie Filce, Brayden Hoskins

Dr. Doosti

MGSC 310

13 December 2019

Steam Store Data Analysis Write Up

**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.

**Dataset Summary**

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.

**Initial Findings**

Initially exploring the dataset, we found that the price and average\_playtime were the most highly correlated variables with a game being successful. 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.

Moreover, we initially believed that more achievements in a game would lead to higher average playtime. This was not the case, however, as we found some of the games with the lowest achievements have the highest average playtime. In addition, it was discovered that price did not necessarily translate to stable changes in average playtime. This was characterized by the equal dispersion of high average playtimes amongst the different price ranges. Furthermore, it is interesting to notice that successful games were equally distributed among the varying price ranges. Both of these points were important in beginning our model building as the two key variables we believed to be driving a game’s success did not necessarily do so.

We next looked to explore a relationship between different genres of games and the prices of those games. While many of the genres had similar price ranges, it was interesting to notice that casual games clearly had the lowest mean price and range of prices. Moreover, “other” and “indie” games appeared to have the highest overall prices, perhaps indicative of the inability of smaller studios to produce the game at a reduced cost. We would also take this into account as we further narrowed our search for optimal predictor variables.

**Models/Results**

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

*Logistic Regreesion*

In building our logistic regression model, we first attempted using all variables as predictors of our successfulGame 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.

Looking at price, the model suggested 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, it was 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 results 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 for a game to be 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.

However, when using this model with a .04 cutoff to determine if a game would be successful or not, we found that the model had good accuracy, but poor ability to predict successful games accurately, thus we looked to find a stronger more accurate model.

*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.

Results yielded that 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 indicated 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.

When testing this models predictive power, we saw a significant increase in its ability to predict successful games while maintaining its overall accuracy and ability to predict unsuccessful games. Thus this model was seen as a large improvement to the previous logistic model.

*Pruned Tree Model*

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 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. With this in mind, we set about using the tree model to make predictions on our test data. Results found that while accuracy of the predictions, and the ability to predict unsuccessful games were maintained, the ability to predict successful games suffered. This set the pruned tree model akin to the logistic model in terms of performance leaving our random forest model as the sole victor in terms of predictive performance.

**Conclusions**

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.