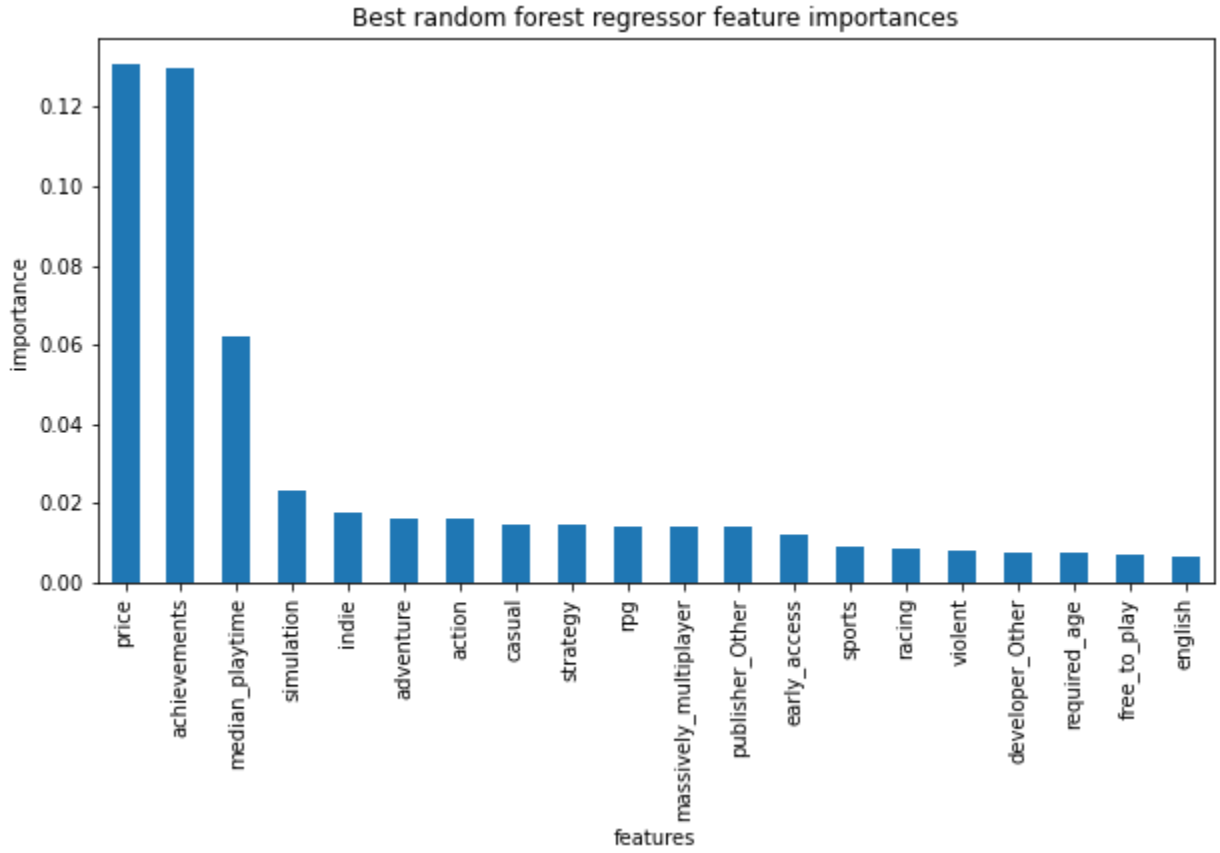
Video Game Popularity: Feature Analysis

Mackenzie Unger

A new video game designer wants to know what features to include or focus on when creating a video game to help insure their game's popularity. We decided to use percent positive ratings to determine popularity. Using a Random Forest model, we identified the top feature to be:



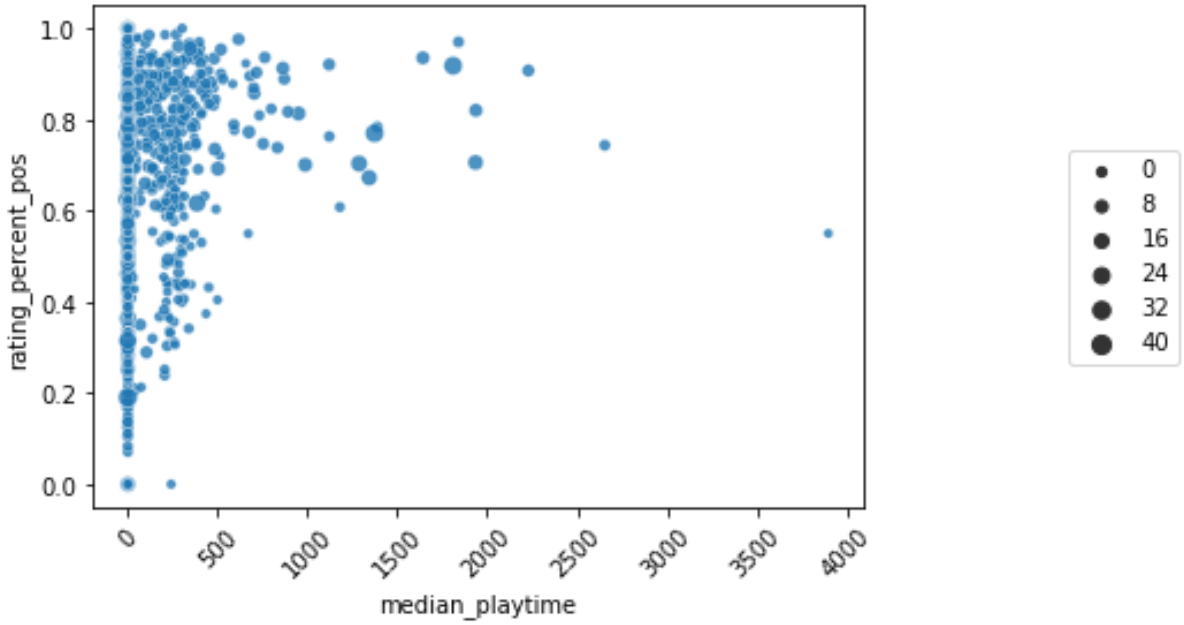
(Image features listed left to right: price, achievements, median_playtime, simulation, indie, adventure, action, casual, strategy, rpg, massively_multiplayer, publisher_Other, early_access, sports, racing, violent, developer_Other, required_age, free_to_play, english)

This information confirmed our hypotheses from Exploratory Data Analysis (EDA). Some things to note about these features:

- **Price and Free_to_Play** are both in the list of features. We may want to do some more analysis into games that are free to play, but seeing as price is ranked higher we may not need to worry about making the game free_to_play.

- **Achievements** are referring to in game milestones. These tend to be little Easter Eggs for the player as they make their way through longer games, skill unlocks, completing challenges, reaching a game milestone, ect. This is definitely tertiary focus, that can come in after the game is in creation mode, but it is good to know so the designer can keep them in mind and make note of possible achievements as the game is being developed instead of picking them after the game has been made.
- **Median Playtime** we will go more into detail on this below, but we suggest between 16 and 50 hours of playtime to complete the game
- **Simulation, Indie, Adventure, Action, Strategy, Rpg, Sports, Racing** these are genres, a developer could choose one or more of these.
- **Casual** is a play style that means that a player can enjoy the game without being competitive. These games tend to have easier settings for people who just want to enjoy the game and not as many high stress combat scenes. However lower on the list of features is **Violence**, so possibly having a difficulty setting for those who prefer casual or more violent combat can equally enjoy the game.
- **Massively Multiplayer** may be more difficult for a new game designer to make, however players enjoy getting to play with friends. Having some sort of multiplayer option whether local co-op or online, is definitely beneficial to look into.
- **Publisher and Developer**, both of these showed up as “Other”, which suggests we don’t necessarily need the backing of a larger company, this is actually reinforced by the feature **Indie**. Indie games are games made by individuals or small teams without financial or technical backing of a large game publisher. This is actually a very strong positive for our new game designer as it means now is the time for them to break into the industry, as players will be invested in trying out their game and the designer doesn’t need to try and appeal to a larger company to back their game.
- **Early Access** is a great way to start getting funds while in the final stages of creating a game. Consumers will pay to be in alpha and/or beta testing of a game that will help with funds in late stage work, as well as generate advertisement for the game, if your testers enjoy streaming on platforms such as Twitch, thus giving their viewers a sneak peak at the game and reaching more players.
- **English**, unfortunately our data set did not test for other languages, however this at least suggests having a version of the game in english will be beneficial to sales.

Going back to **Median Playtime,** in our EDA we made a few hypotheses about ideal playtime and price. From images similar to the one below, we concluded that keeping the game play between 16 and 50 hours (960 - 3000 minutes), provides the best likelihood of popularity. We also saw that longer games are able to have higher prices without losing popularity. We suggest doing further analysis to determine the price of the game after all the other features have been chosen.



(Image notes: median_playtime is in minutes, and the point size is the price of the game.)

<u>Summary of Notebooks</u>:

Wrangling:
We explored and cleaned a lot of data sets in the Data Wrangling Notebook, part of this was to decide which features and to see how clean each of the data sets were for use. In the Modeling Notebook we ended up only using the dataset in section IV, 2019 Steam Data with the tags. This dataset can be found on Kaggle (https://www.kaggle.com/nikdavis/steam-store-games?select=steam.csv). We ended up using the steam.csv and steamspy_tag_data.csv, this gave us data on developers, publishers, platforms, ratings, playtime, price, tags, and genres. Initially this data set had the tags as counts, instead of indicators. This was because users could go in and tag a game for a feature, but there wasn't a way to indicate if the tag was correct or not. A quick example we had a game that was tagged as both WW1 and WW2, and a quick Google search showed it was actually only WW2, so to fix this I picked the highest number that any of the tags for a particular game appeared and created a threshold of 75%. So if a tag was less than 75% of the maximum, we did not indicate it as a feature, and if it was greater than or equal it was included as a feature. Next we had the number

of positive and the number of negative ratings, so we created our popularity indicator by calculating the percent positive ratings.

EDA:

We still used most of the data sets in Wrangling, we explored different methods of measuring popularity. We also made many predictions and interesting discoveries mentioned at the beginning of the notebook. In the end we decided we could use Critic Score, Ratings, or Global or Regional Sales. We decided we would use Ratings in our Modeling step.

Preprocessing and Training:

We narrowed down our work to 2 data sets with an option of merging the two. We ended up only using the steam data, but this allows us to change up what is the Modeling notebook quickly if we decide we want to see if the conclusions would change. Here we chose to create an indicator for if a game was popular or not by if a game had more than a 70% positive rating.

Modeling:

We worked with several models and narrowed our focus down to two possible models, Random Forest and Gradient Boost, after hyper parameter tuning, we saw that Random Forest gave us the best model for identifying the most popular features with a ROC_AUC Score of 0.7060.