Team10: Predicting Features of Blockbuster Games

Shuaib Balogun, Madhumitha Ganesan, Satria Priambada, Manasa Vegesna, Ting Yu, Wenxin Zhang

Project Overview

Game companies are very careful when making new games as their success can impact the survival of the business owing to the often high budget in production. Besides, it is difficult to estimate a video game's success in a dynamic market. Our project aims to help game companies predict the success of new games to achieve high investment returns. In this project, we used historical data from main-stream game databases, VGChartz and Steam API, to build machine learning models that predict a new game's success. We also analyzed user sentiment using reviews data to learn about popular game features. Our results are presented in an interactive web visualization interface.

Predicting Regional Sales

Sales number is a good indicator of how successful a game is. We made sales prediction on four different regions (North America, Europe, Japan, and Other) based on game features such as year since release, average rating, and top publisher. The prediction used 2020 data from VGChartz. Linear regression models such as GLM, and Poisson Model were tested. The Poisson Model had the lowest RMSE (500k from 4 mil scale) and was chosen for prediction.

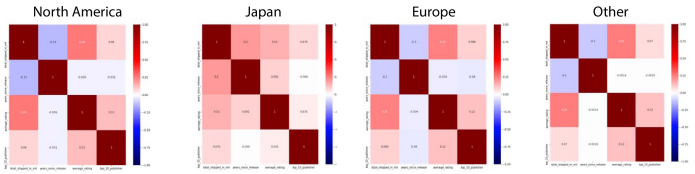

Fig. 1 Feature correlation maps for four market (features from top to bottom: Sales, year since release, average rating, and top publisher). Sales in Japan has higher correlation with years since release which can be inferred as fan loyalty to certain game title can help predict higher sales. In other three markets, year since release is negative correlated with highest sales. This may indicate a preference to new title.

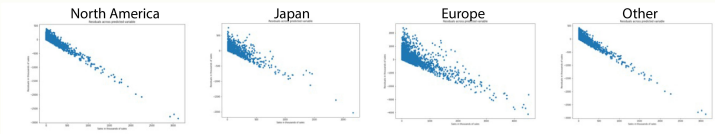

Fig. 2 Performance of Poisson linear regression model. Its better performance may imply the nature of game sales is closer to a poisson distribution with a lot of purchase near a game release and slowly decreasing afterwards.

Predicting Regional Playtime

Playtime is another significant indicator of a game's success. Players with long playtime tend to engage in more in-game purchases and purchase sequels of the game.

Model Training

The playtime prediction was performed on the same four regions as in sales prediction. A dataset consisting of 200,000 records of players' playtime in different games is used for building a regression model. Four game features are identified to be significant to playtime, including genre, rating, platform, and price, and thus used as features for regressors. Several models, including decision tree, random forest, and XGBoost were tested. The random forest model had the highest prediction accuracy and was chosen for prediction.

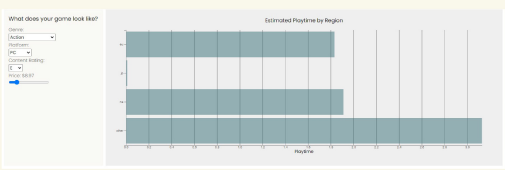

Fig 3. Web visualization for playtime. The game playtime may have a positive linear relationship with its price. Players in Europe and Japan tend to spend less time on the action games with lower prices and more time on those with higher prices. This is likely because expensive action games can have better designs, details, and graphics, which are attractive to players in those regions.

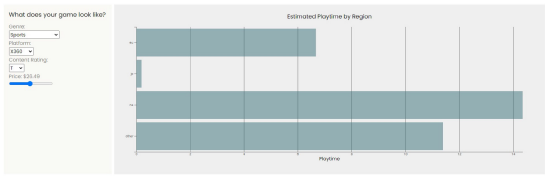

Fig. 4 Players in different regions may prefer different platforms. For sports games on the PC platform, players in Europe spend the longest time. Meanwhile, for sports games on the X360 platform, players in North America spend significantly the longest time. This can be explained by the fact that X360 is a product of Microsoft, which is based in the US, which would lead to its high popularity in North America.

Sentiment Analysis on User Reviews

Data Collection

We used a user review dataset from Kaggle of 7GB size with 21 million user reviews of 300 games. Game genre data are downloaded using Steam API to merge user reviews with genre.

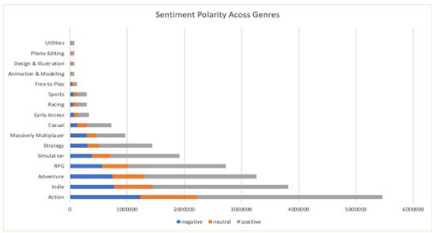

Fig. 5 Sentiment Polarity per Game Genre in 202. There is generally a higher percentage of positive reviews. "Action", "Indie", "Adventure" seem to be top performing genres on Steam.

Approach

We processed English reviews and performed sentiment analysis on review text data to generate "positive", "negative" and "neutral" sentiment labels. We then visualized most frequent user sentiments using word clouds, which effectively highlight feedback on specific game features that will aid in future game development.

Fig. 6 A word cloud with positive sentiments for a genre. Game features like "Wallpaper", "Background", "Software" are reviewed positively in this genre. Positive feedback promotes motivation for future development.

Interactive Web Visualizaton

The front-end visualization is built with d3. As users input parameters, they trigger a GET request to fetch model prediction results real time from backend python APIs built with Flask.

Fig. 7 Web visualization pipeline with a Python backend

Acknowledgement
This is a project for Georgia Tech Spring 2021 CSE6242 Data and Visual Analytics by Prof. Duen Horng (Polo) Chau, Dr. Mahdi Roozbahani, and all the awesome TAs