SENG 474 Project

Using Regression Models to Predict Steam Game Sales

Andrew Yung

Jake Rothwell

Zheng Li

Adebayo Ogunmuyiwa

***Abstract* -The problem domain is to identify if a game is going on sale on the Steam game platform based on the available datasets. Our group will find a algorithm that computes the given datasets relating to each game in order to predict whether a particular game would be on sale or not. The output will be a continuous numerical representing the likelihood of game going on sale. If time permits, the computation results will also include the expected discount for that particular game.**

1. Problem description

The goal of our project is to draw from a dataset of Steam game sale history to predict if a certain game will go on sale. Given an input game title or appid (Steam's unique identifier), our program would output a probability percentage value of a game going on sale. Something important to note is that most Steam games will eventually go on sale in their lifespan, so we have discussed ways to narrow our problem description down to a more specific one. Another input parameter would be a certain month - the output being the probability of the input title going on sale within the given month.This would significantly narrow our problem down and make is more intuitive for users - users will want to know specifically when the game they are interested in will go on sale.

2. Related work

Previous work includes developing a music genre classifier for a final project in CSC 475 - Music Retrieval Techniques. The classifier is described as follows - given an input song, output the genre of that song. A dataset of roughly 50,000 tracks was used to train the classifier. This dataset is balanced over 13 common musical genres. Google's Tensorflow library was used for the softmax regression function, which trains on the dataset to output probability values for each of the 13 genres. For predicting the genre of a new input track, the LibRosa Python library was used to extract the features of the input track, then input the track's features into the softmax function to return an output genre.

As an additional functionality, Soundcloud's web API was used to extract the features of tracks from Soundcloud for input into the genre classifier. In retrospect, I believe this previous project will be similar to our current one in its development process and result. We believe that possessing the knowledge and experience from developing a classifier during a previous group project will ultimately be of benefit to us. Some notable differences include the type of data used - music versus raw sale history, and the type of regression - at this point in our project, we intend to use linear regression (as opposed to softmax regression).

3. Data description

The data currently exists as two separate datasets, Steamsales.Rhekua [4] containing the pricing history and SteamSpy [3] containing various information on each game. At the time of this proposal, there are around 20,000 products available on steam (not all of which are games, but can be filtered by their genre).

The attributes provided or that can be determined from the datasets are:

- Developer
- Publisher
- Score ranking of the game based on user reviews
- Ratio of positive to negative reviews

- Number of owners
- Player playtime and activity (total and past 2 weeks)
- Days since last sale
- Average number of days between each sale (if there has been more than 1 sale)
- Variance of the days between sales (if there has been more than 1 sale)

SteamSpy data can be accessed their own api [5], responding in the json format. In SteamSpys json response, there is a steam appid which can be used to request the sale history data from steamsales.rhekua.com through an http request..This data is easily pulled from the html response as each start and end date of the sale is under the same html tag class. Our script to retrieve this data is on average 3 games per second.

4. Proposition

A. Data

Data can be found on Steamsales.Rhekua [4] and SteamSpy [3] where it shows previous sales start and end dates in a sorted order. In addition, SteamSpy provides a publicly accessible csv dataset of all games that are currently on sale.

One problem that arose during the discussion of our project is how the model will be evaluated - more specifically, how the accuracy of predictions will be determined. We discovered that we must split the data into two distinct sets - previous sale data, and current game sales which would be used for checking against predictions. With the previous sale data, the labels are represented as the game going on sale or not.

B. Algorithm

Our aim is to provide a numerical prediction model using various forms of linear regression.
We plan on using Python along with the sci-kitlearn library for our model. Linear Regression seems to be the best fit for our problem at this point, however we will test all the various regression models that sci-kitlearn has to offer to evaluate the performance and accuracy of our model.Alternatively from related works, SMOreg provides an implementation for vector machine regression which categories data into several discrete classes. “The method attempts to find a boundary that correctly divides the training data into classes, and then uses this boundary to classify unseen data” [1].

C. Evaluation

Evaluation would be done using regression algorithms and the dataset would be divided into testing , training and validation datasets. Note that the output prediction will be a continuous, numerical value between 0 and 1, representing that probability bias that the game will go on sale - a “yes” or “no” output. Due to the output being a continuous probability value, we have concluded that the best model to use would be Linear Regression [2].

5. Timeline

Task	Date
Acquire initial dataset and supplementary dataset	February 14
Integrate data set structure	February 24
Validate the total dataset for completeness	February 27
Finalize and submit mid-term report	March 6
Train Algorithm and evaluate prediction performance	March 10
Explore potential interest and investigate potential regression algorithms	March 14
Tune features and parameters for tested algorithms	March 20
perform tweaks on final algorithm	March 26
Finalize and submit report	April 6

6. Task distribution

Member	Distribution
Andrew Yung	Model Development Dataset Preparation
Jake Rothwell	Model Development Documentation
Zheng Li	Model Development API Development
Adebayo Ogunmuyiwa	Model Development Predictions/Evaluation

References

[1] S. Ehrenfeld, “PREDICTING VIDEO GAME SALES USING AN ANALYSIS OF INTERNET MESSAGE BOARD DISCUSSIONS”, 2011. [Online]. Available: http://sdsu-dspace.calstate.edu/bitstream/handle/10211.10/1073/Ehrenfeld_Steven.pdf. [Accessed: 03- Feb- 2018].

[2] "sklearn.linear_model.LinearRegression — scikit-learn 0.19.1 documentation", Scikit-learn.org. [Online]. Available:http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html. [Accessed: 03- Feb- 2018].

[3] “Game sales”, SteamSpy - All the data about steam games. [Online]. Available: https://steamspy.com. [Accessed: 03- Feb- 2018].

[4] “Steam Sales”, steamsales.rhekua.com, 2010. [Online]. Available: http://steamsales.rhekua.com. [Accessed: 03- Feb- 2018].

[5] S. Galyonkin, “SteamSpy API”, steamspy.com/api.php. [Online]. Available: https://steamspy.com/api.php. [Accessed: 03- Feb- 2018].