craigslist

Video Gaming Classification

Team Unstrucata

Rachel Fagan | fagan3@purdue.edu

Evan Tiffany | etiffany@purdue.edu

Vivek Rao | rao161@purdue.edu

Soyeon Baik | baik6@purdue.edu

Jai Woo Lee | lee3999@purdue.edu

Paul Chen | chen3876@purdue.edu

Background

About Craigslist

Craigslist is an online platform connecting buyers to sellers. It facilitates the sale of goods like bicycles, musical instruments, and used cell phones, as well as services like yard work, photography, and freelance work.[1] Craigslist generates revenue by charging a fee on select postings[2] and through paid posting accounts (third-party advertisers).[3] Craigslist heavily depends on users returning to its website to increase engagement. If engagement on its site goes up, Craigslist can charge paid posting accounts higher amounts.

Thus, it would be in their best interest to maximize the number of users on their website, or increase the amount of time users spend on Craigslist. An effective way of improving users' time on a page is by improving their experience. For this project, we decided to focus on user experience in the video gaming section.

Problems with the Video Gaming Category

There are several areas for improvement within Craigslist's video game category: general clutter, inconsistent filtering results, and highly specific filtering options.

(1) General clutter:

The video game category is littered with unrelated search results, from "don't play on an empty stomach" (cereal) to "3 new small appliances" (a coffee maker, slow cooker, and hand blender). This clutter negatively affects the website's user experience for the customer. If customers are unable to easily find the products

[1] Business Model of Craigslist – StudiousGuy
[2] craigslist | about | help | posting fees
[3] craigslist | about | help | paid posting accounts

that they want, they may be less likely to return to the site, either for purchasing or listing.

(2) *Inconsistent filtering results:*

Craigslist's current search filters in the video game category do not work well. For instance, if a user searches "xbox" in the Make and Model filter, only 41 listings appear in the Chicago area out of 134 listings that appear when typing "xbox" the category search bar at the top.

(3) *Highly specific filtering options:*

Additionally, when using the "Make and Model" filtering bar, the options available to customers are highly specific. For instance, if a customer was looking for an "xbox one," should the customer click on the suggested filter "microsoft xbox one," "xbox one," "xbox," or "microsoft xbox one x" in the filtering bar? The customer would have to potentially search through multiple pages because the search filters are too specific.

Business Analysis

Objectives

Our project aims to improve the user experience in the video gaming category by offering an improved filtering option by brand name. We will enable customers to search for either Nintendo, Meta, Microsoft, Sony, Arcade, or Other.

The video gaming category on Craigslist is currently not monetized with ads. By increasing customer satisfaction with this improved filter in the video gaming category, we can increase the customer retention rate, their time spent on the website, and usage

of the site overall. Craigslist's maxim is to focus on content, rather than user interface (web design), so providing a better user experience may have a significant impact. Providing a better user experience would also lead to a higher lifetime customer value in the long-term. Improving the customer experience in this category that has such a high potential customer base (video gaming is extremely popular in the United States, with around 227 million players[4]) can drive the company's revenue overall.

Data Analysis and Validation

Data Preparation

We used BeautifulSoup4 to scrape 3,011 video game listings from New York, Chicago, Washington DC, Miami, Los Angeles, and Philadelphia. From the title and description we extracted keywords to classify each game brand (e.g Sony, Microsoft) and then stored them in a dictionary of lists.

Data Processing

After we scraped the data, we began the data cleaning process. First, we removed the dollar sign in the "price" column. Second, we removed parentheses in the "place" column. Third, we made each description row into one line. Last, we lowered the case in the "title" column. Finally, we lemmatized and tokenized the data before we implemented the classification models.

[4] https://dotesports.com/general/news/more-people-in-united-states-play-video-games-than-ever-before-esa-reports

Model Building

We used six models to identify and classify the video game postings into different brand categories.

(1) Naive Bayes[5]:

This model calculates a conditional probability. Given a record to be classified represented by a vector with n features (independent variables), it assigns to this instance probabilities for each of K possible outcomes. With our dataset, Naive Bayes model has 73.97% accuracy.

(2) Logistic Regression[6]:

Logistic Regression is a simple model for classifying binary targets. However, we have multiple categories for the target variable. It makes models by multinomial logistic regression. With our dataset, the Logistic model has 84.99% accuracy.

(3) Random Forest[7]:

Random Forest is an ensemble learning method for classification. It constructs a multitude of decision trees at training time and returns the average estimates of the probability for each classification case. In our model, we set 50 as the number of estimators and maximum depth as 6. With our dataset, the Random Forest model has 63.75% accuracy. This was the best score we generated after tweaking parameters.

(4) Support Vector Classifier:

[5] Naive Bayes https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[6] Logistic Regression https://en.wikipedia.org/wiki/Logistic_regression
[7] Random Forest https://en.wikipedia.org/wiki/Random_forest

Support Vector Classifier is a supervised learning model. It maps points in space and classifies based on distance from a hyperplane/decision boundary.[8] The SVC was the best model with an accuracy of 88.45%.

(5) Neural Network:

Neural Network is trained on a hidden node architecture. Input values undergo a combination and activation function. We tried several combinations of the number of layers, and found out that three for the first layer and two for the second layer shows decent performance compared to other numbers of layers. Our model contained (2,575*3 + 3*2 + 2), i.e 7,733 parameters. With our dataset, the Neural Network model has 74.24% accuracy.

(6) Recurrent Neural Network:

We tried a Recurrent Neural Network (RNN) on our dataset after setting the maximum length of each title and review concatenation to 200. We set the dropout rate 0.1, and saw an accuracy of 25.9%. Even if we simply guessed which brand a review is associated with, we would get an accuracy of 16.67%. Therefore, the RNN is not a significant improvement.

Final Model Selection

Among the six models, we chose the SVC (Support Vector Classifier) which showed the highest accuracy rate (88.45%). Because of the size of our training set (2,257 records), neural networks performed poorly and simpler models shined here — a simple logistic regression gives us an 85% accuracy. If we had a larger dataset, more complex models may have performed better.

[8] Support Vector Machines (SVM) Algorithm Explained (monkeylearn.com)

From our confusion matrix for the SVC model, we see that the model does an excellent job of classifying postings for Arcade, Sony, Nintendo, and Microsoft. We see the model is less successful at classifying Meta postings. This is likely because our training set did not contain enough postings corresponding to Meta (see below).

	Arcade	Meta	Microsoft	Nintendo	Other	Sony
Arcade	43	0	0	1	1	1
Meta	0	4	1	3	3	3
Microsoft	0	0	147	0	4	6
Nintendo	0	0	1	173	13	8
Other	0	1	0	12	73	20
Sony	1	0	1	1	6	226

```
Sony         777
Nintendo     560
Microsoft    486
Other        274
Arcade       131
Meta          30
```

Conclusion

Delivering Value

Information Gain: To evaluate how much our model would improve the customer experience, we conducted an information gain analysis. The information gain measures the increase in the accuracy of the search rate for ad-viewers after implementing our model. For example, before our model, if ad-viewers want to search for Playstation (Sony), it shows 549 listings while there are 1,012 listings of Playstation (Sony) in total.

For all five brands, the website's search engine only shows 68% of results on average out of the total listings and leaves the rest of 32% unsearched, ultimately creating a mismatch between demand and supply, impairing customers' experience. After implementing our model, we identified a dramatic increase in the accuracy of the search rate. Out of the 647 listings in the test dataset, 593 postings or 92% of the total postings are correctly classified or searched. The search rate, or information gain rate, increased by 34% on average, from 68% to 92%. Also, by having one search engine based on our model, we can resolve the issue of inconsistent results from multiple search boxes.

Brand	Manual		Classification		Manual Rate	Classification Rate	Information Gain
	Manually Typed	Total	Classfied Number(TP)	Total(Test)			
Sony	549	1012	226	235	54%	96%	77.28%
Nintendo	558	755	173	195	74%	89%	20.04%
Microsoft	593	643	147	157	92%	94%	1.53%
Arcade	74	177	43	46	42%	93%	123.59%
Meta	24	44	4	14	55%	29%	-47.62%
Total	**1798**	**2631**	**593**	**647**	**68%**	**92%**	**34.12%**

* Others excluded (Number of observations: 380)

The improvement of search results can not only increase the retention rate but also time on site. It will have a positive impact on growth of revenue as the number of ad-viewers entering into other monetized categories increases.

Limitations

We faced a few limitations with our project. To begin, our data code for data scraping had to be split into city-specific code. With this, we were also limited in the amount of data we received from the Craigslist site. Additional data should be used to train a more robust model for this classification. Moreover, we initially hoped to add a product-type filter as well. This was infeasible through our methods without manually entering the

target variable for each of our 3,000 entries. Finally, we saw a decrease in our information gain for Meta. As this company has newly entered the videogame market, the number of ads for this brand is low. We do, however, think this is a valuable addition to the classification model.

Future Ideas

In the future, we recommend that Craigslist design and implement a filter-by-product option for customers. Customers could filter the Video Gaming Category for consoles, accessories, or video games. Additionally, Craigslist could implement our brand filtering improvement within the “posting details” user interface for advertisers across all postings. Advertisers’ descriptions of their listings would be automatically filtered into a brand category that would be confirmed by the advertiser. This would help them better list their items for purchase with more accurate descriptions. Lastly, Craigslist could implement this model by taking other variables into consideration, for example, price.