404 Error Name Not Found- Group 2

Automate Identification of 'Best of Craigslist' posts

Final project Report

Arun Ramakrishnan
Roli Gupta
Samir Husain
Shreeansh Priyadarshi
Pranay Khandelwal

MGMT590

MGMT 590 - AUD

Final project report

Contents

Introduction

Craigslist makes money through only a few revenue streams. Most of the postings are free and there are only few avenues where they charge to post a job listing or apartment listing. This is true for only a few major cities. However, Craigslist's aim is to provide the simplest functionality in the service of society.

- Situation and Complication- In order to help boost number of users and visit time, we propose to use the existing 'Featured Listings' section. In the current situation any listing gets into featured section based on the user likes. We propose to use this list of posts as a proxy for viral posts and predict the virality of a new post.
- Solution- We plan to use machine learnings algorithm to classify posts as viral/not viral. This will help to increase customer time on site and bring in new customers. We also plan to create a separate section for Obscene posts which might hurt the user satisfaction for a few users.
- Impact- This functionality would make it easier for website users to identify interesting and amusing posts. This move will potentially bring new users as people share posts with their friends.

Project objective

We first define success metrics for the Craigslist website. Since, they do not show display ads, the only source of revenue for Craigslist are the charge they levy on Job and real estate listing posting in some area. Hence, we define the success metrics in terms of Pageviews. For craigslist to increase its reach, they would want to increase the number of people who visit their website and the time these users spend on the website. The two motives are

1. Increase traffic
2. Increase time on site

We propose adding a 'This is cool' button on craigslist page. When users click on this button, they will be taken to a page which contains post that are similar in characteristics to the posts found on the Best of Craigslist page. We hope that this section will keep users on the website for longer browsing through interesting posts, they are also more likely to share some of these websites with their friends and get new users to the website. This would help us achieve both of our defined motives.

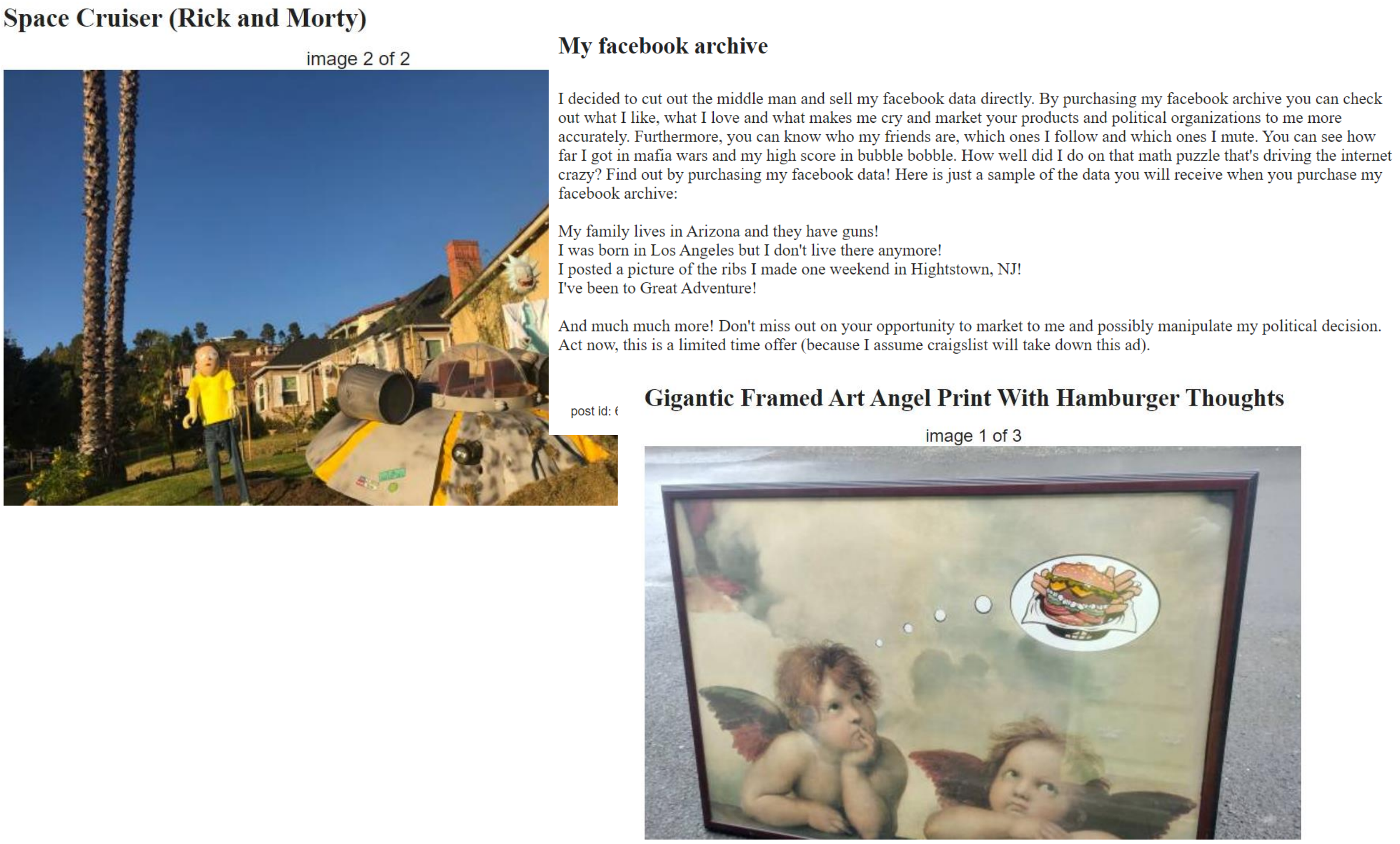
Craigslist is a website that while being the go-to classifieds site for most Americans, has also become home to a variety of funny, joke posts. We wish to work on this previously undocumented part of the website and explore ways to generate business value from the posts that people might find funny or interesting.

In order to explore more into how our idea could impact business, we set out to understand the science behind virality. Psychologists are still trying to figure out that secret ingredient that makes a person click on the 'share' button. However, there are certain characteristics of content that are found with almost all viral content. They are novelty, curiosity and emotion. Emotion is the biggest motivator among the three.

According to a research by Fractl[1], the 10 most common themes in viral content are

1. Amusement
2. Interest
3. Surprise
4. Happiness
5. Delight
6. Pleasure
7. Joy
8. Hope
9. Affection
10. Excitement

A quick glance at the 'Best of Craigslist' section will make it clear that most of the posts in this section are there due to their entertainment factor and contain one or more of the different emotions.



Space Cruiser (Rick and Morty)

image 2 of 2

My facebook archive

I decided to cut out the middle man and sell my facebook data directly. By purchasing my facebook archive you can check out what I like, what I love and what makes me cry and market your products and political organizations to me more accurately. Furthermore, you can know who my friends are, which ones I follow and which ones I mute. You can see how far I got in mafia wars and my high score in bubble bobble. How well did I do on that math puzzle that's driving the internet crazy? Find out by purchasing my facebook data! Here is just a sample of the data you will receive when you purchase my facebook archive:

My family lives in Arizona and they have guns!
I was born in Los Angeles but I don't live there anymore!
I posted a picture of the ribs I made one weekend in Hightstown, NJ!
I've been to Great Adventure!

And much much more! Don't miss out on your opportunity to market to me and possibly manipulate my political decision. Act now, this is a limited time offer (because I assume craigslist will take down this ad).

post id: (

Gigantic Framed Art Angel Print With Hamburger Thoughts

image 1 of 3

This is because, the “Best of Craigslist” section is built upon user votes and this crowdsourced nature means that posts that are viral find their place on this section very easily.

Another potential loss of business opportunity is the missing NSFW filter. This is particularly important due to diverse set of audience on the website. Using stop word lists developed by google, we aim to create a separate section with posts that contain profanity. These posts can only be accessed once a user confirms their age. Currently, craigslist advises that users under the age of 18 does not simply access certain section and this leads to them missing out pageviews from users aged under 18.

Data Analysis

Our analysis process can be broken down into the following 7 stages.

1. **Fetching unstructured data from the website** - We scraped posts from the 'Best of Craigslist' page. This makes up our 1's (featured listings) in the dataset. We scraped posts from craigslist in different categories. These are our 0's (non-featured listings) in the dataset
2. **Preprocessing** - We processed the data to remove any junk information from the dataset such as null values and insignificant variables
3. **Sampling** - We found that our target variable (featured/non-featured) is unbalanced with only ~2% of featured posts. So, we performed random down-sampling to balance the dataset
4. **Creating term document matrix** - We brought the text data into a format which our machine learning models can understand. We tokenized, lemmatized, removed stop-words and transformed the text into tf-idf matrix. This increased the size of our dataset exponentially. To further make our models simpler for faster processing, we processed the data to curb near-zero variables
5. **Final dataset** – We split the dataset into 80:20 ratio for training and testing our models respectively
6. **Model creation** – We tried various classifiers to test which would be flexible enough to fit on the training dataset and at the same time can be generalized for different datasets. The models are listed below –
    a. Logistic Regression (LR)

b. Support Vector Machines (SVM)

c. Random Forest Classifier (RF)

d. K-Nearest Neighbor (KNN)

e. Gaussian Naive Bayes (GNB)

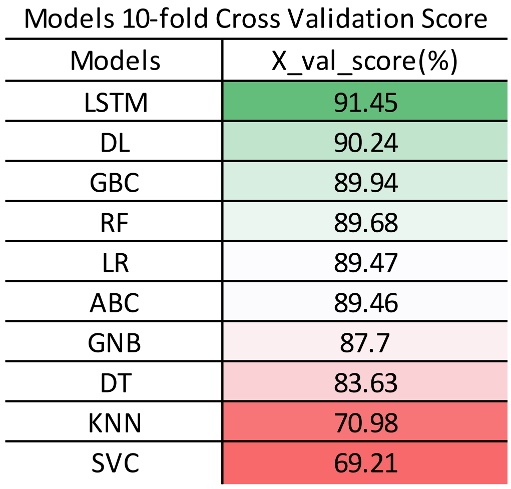
f. Decision Tree Classifier (DT)

g. Gradient Boosting Classifier (GBC)

h. Adaboost Classifier (ADC)

i. MLP Classifier (DL)

j. LSTM



Models 10-fold Cross Validation Score

Models	X_val_score(%)
LSTM	91.45
DL	90.24
GBC	89.94
RF	89.68
LR	89.47
ABC	89.46
GNB	87.7
DT	83.63
KNN	70.98
SVC	69.21

7. **Fine Tuning** – Our initial 10-fold cross-validation results showed that LSTM, deep learning, gradient boosting and random forest classifiers perform best with a maximum accuracy of ~91%. We further tried to improve these models by optimizing the model parameters using grid search. The model parameters and results post optimization are given below –

   a. LSTM Best Score: 92.6 and Best Parameters: {‘embedding_dim': 100, 'activation': ‘softmax, optimizer: ‘adam’, 'loss':’ categorical_crossentropy’, 'dropout':0.2, 'recurrent_dropout':0.2}

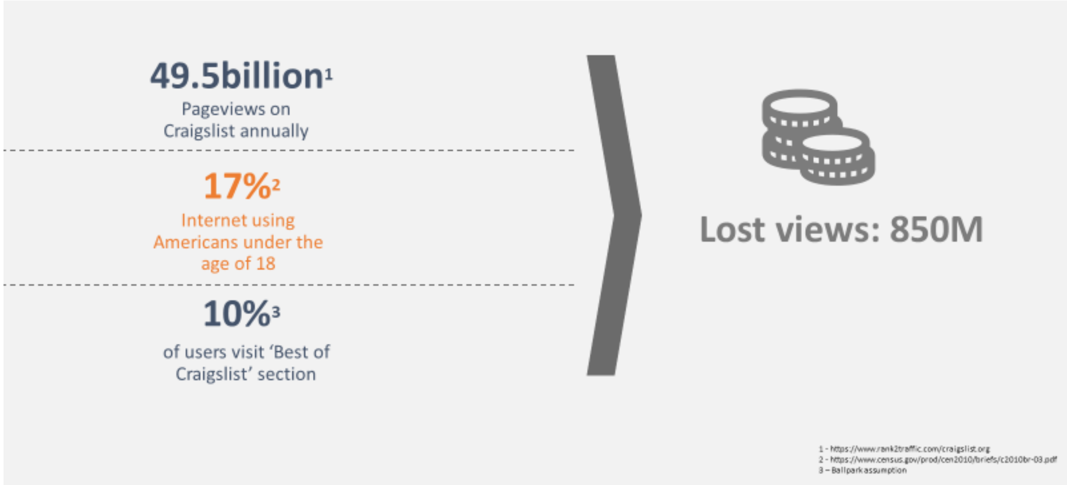
   b. DL Best Score: 90.3 and Best Parameters: {‘hidden_layer_sizes ': (2,2,4), 'activation': ‘relu’, solver': ‘adam’, 'alpha': 0.05, 'learning_rate': ‘adaptive’}

   c. GBC Best Score: 90.1 and Best Parameters: {'learning_rate': 0.05, 'max_depth': 8, 'max_features': 0.3, 'min_samples_split': 3, 'random_state': 1234}

   d. RF Best Score: 89.9 and Best Parameters: {'criterion': 'gini', 'max_features': 'log2', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 30}

Conclusion

Currently craigslist allows only users above the age of 18 to access their pages because a few of their posts might be obscene and disturbing for minors. We calculate[1] that about 850 million pageviews to Craigslist's 'Best Post' section is from under 18 users. If we create a new category or filter switch for 18+ posts and allow all users to visit regular posts, we can reclaim the lost 850 million pageviews. This will be done by using the "stop words" algorithm to identify and group mature content under a separate category. With user authorization and/or validation of his/her age, they will be able to access the mature content category.



49.5billion[1]
Pageviews on Craigslist annually

17%[2]
Internet using Americans under the age of 18

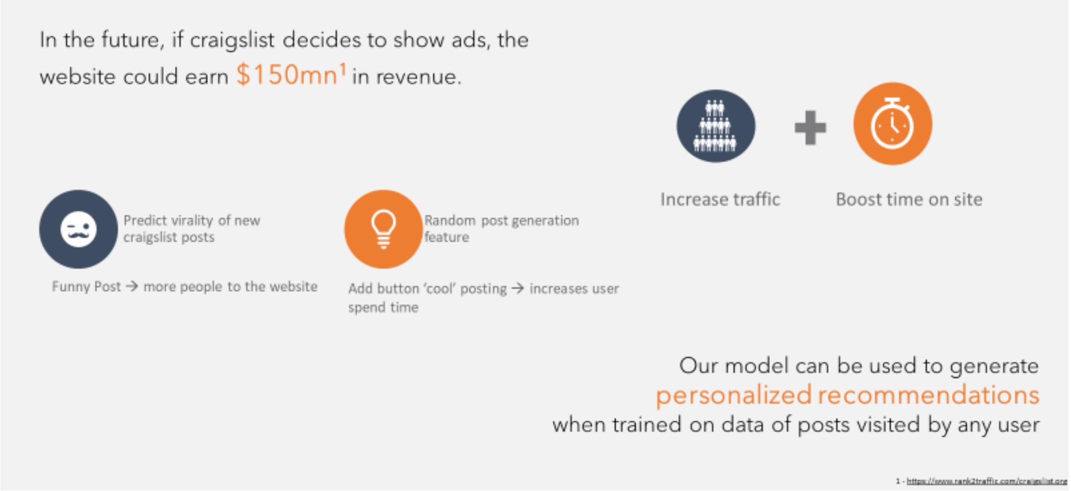
10%[3]
of users visit 'Best of Craigslist' section

Lost views: 850M

1 - https://www.rank2traffic.com/craigslist.org
2 - https://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf
3 – Ballpark assumption

Further, since our model takes microseconds to understand if a post has mature content or not, it can be applied on a real-time basis to new posts being posted by users. These will be categorized as soon as they are posted and will add to our mature content inventory over time. It is our duty as part of a society to not only protect the children from the mature content on websites but also allow them to use our website to the fullest. Hence, categorizing and separating mature content posts will not only expand our user base to people below the age of 18 years, it will also enhance customer experience and help in getting re-visits from them.

Furthermore, to increase the time spend by the existing customers on Craigslist website and usher new customers to the website, we will utilize our machine learning, LSTM model to identify popular and quirky posts that can be tagged as "amusing posts". With the model trained on the existing data from "Best of Craigslist", we can classify real-time posts that can be part of this category. This would lead to new customers to be website to check out the amusing posts. Further by adding a button called "Waste my time" or "I'm feeling lucky" the existing customers can access this category and spend some additional time going over these amusing posts.



In the future, if craigslist decides to show ads, the website could earn $150mn[1] in revenue.

Predict virality of new craigslist posts

Funny Post → more people to the website

Random post generation feature

Add button 'cool' posting → increases user spend time

Increase traffic + Boost time on site

Our model can be used to generate personalized recommendations when trained on data of posts visited by any user

1 - https://www.rank2traffic.com/craigslist.org

Risks and Challenges

Implementing a project of this magnitude comes with its own challenges. Here are a few risks and challenges that we foresee and ways to mitigate them.

- Deviation from conventional style
  - Market Survey to understand the attitude of customers towards proposed changes.
  - Devising strategy to manage public sentiment to changed experience

- NLP challenges
    - Sarcastic comments and posts can be identified in future through extensive research and development
- Operational Cost increase
    - Partnering with BA courses across countries to utilize young talents
- Object detection
    - Posts with misclassified objects can be reclassified through advanced object detection

Timeline

This can be achieved by integration of efficient process, cutting edge technology and inspirational people. Efficient processes to streamline filtration of mature content and identification of new “amusing” posing from regular posts using cutting edge cloud storage technology coupled with highly motivated web designers and data scientists, our new website will be ready to roll-out to customers within a year.

Refrences

1. http://www.frac.tl/the-role-of-emotions-in-viral-content/