Craigslist

AD CLASSIFICATION

CLASSIFICATION OF ADS IN 'GENERAL' SUBCATEGORY IN SALES

Pattern Pros:
Chilla, Madhulika
Baluni, Sanyukta
Gulati, Tavish
Patil, Soham
Sheth, Parth

Contents

Executive Summary

Numerous complaints from users suggest that the site is experiencing a widespread problem that stops users from placing their adverts. Many of them had issues with Craigslist advertising being quickly flagged. This can be annoying, especially if you're new to the site. Our aim in this project to build a model that identifies misclassified posts pertaining to 3 categories: Beauty and health, Cars and Tracks and others and puts these posts into their respective subcategories. This helps the users to boost their sales and prevents substantial loss of revenue.

Initially, our dataset was quite imbalanced which means there were very few documents relating to one subcategory and resulted in poor performance of machine learning models. One way that we found to solve this problem was to use SMOTE technique for balancing our training dataset prior to fitting a model. SMOTE usually over samples the minority class by duplicating the data and handles unequal distribution of a class. We trained our model using various classification algorithms like logistic, boosting SVM, logistic and applied NLP techniques like Tokenization, Lemmatization, Tf-Idf and we manually tagged around 1846 ads in 'General' sub-category. We compared the accuracy of these models and used the model that has got the highest accuracy. We were able to attain an accuracy of 95% using linear SVC on test sets.

Background

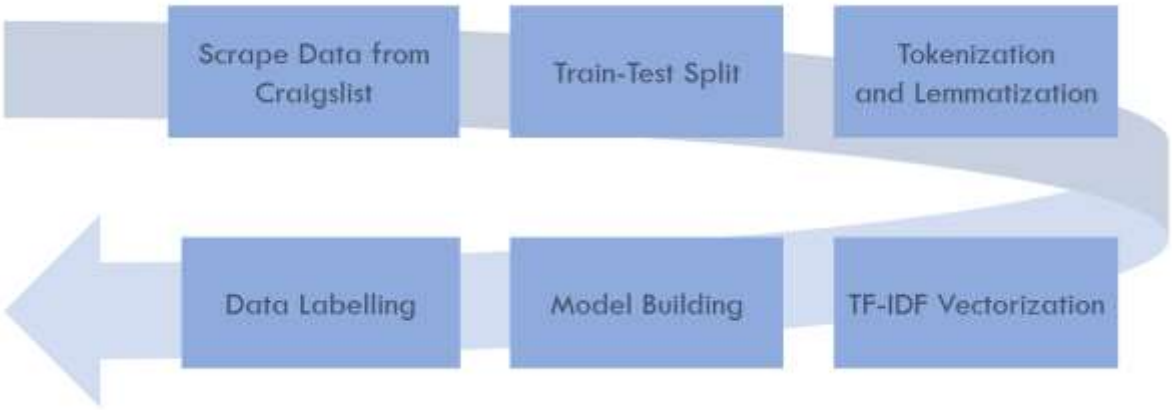
Craigslist is an American classified advertisements website that attracts large volumes of traffic from 70 different countries in 14 different languages. The platform receives more than 250 million users per month. The website features city and region-specific sites where users can post and view a wide range of ads, including jobs, apartments, garage sales, used cars, personal ads, and a whole lot more. Craigslist only allows the users to post limited number of advertisements, its pricing is very competitive. If users want to post something on Craigslist, they have to first create a Craigslist account and then visit the homepage and select a city from the dropdown and finally click "go". The site has a range of broad categories, and each category has several subcategories, making it easy to navigate to the appropriate section quickly.

Business Analysis

Miscategorized posts are one of the most common mistakes of many users when promoting their business on Craigslist. Unfortunately, this seemingly innocent mistake can get your post flagged, misclassified or removed since it is against the website's Terms of Use. For example, if you are offering painting services in your city, then it's best to choose the "Services Offered" category instead of posting it in the "For Sale" section. To reach the correct audience and improve the effectiveness of the listings, it is important to choose the correct category for the products or services. Another problem is uncategorized posts or posts being put in vague categories. Hence, audience members are missing out on relevant posts and hindering the business visibility. It also leads to more use of manpower and higher cost, as human moderators have to spend time on each of the misclassified posts and put them in correct categories. In sales category, 'Barter' and 'General' are instances of vague, catch-all subcategories.

Approximately, about 2% of posts in 'General' are for cars/trucks and 5% ads are for 'Beauty/Health'. We shall focus on classifying these posts on cars/trucks and beauty/health (Multi-class classification).

Methodology



WEB DATA SCRAPING:

- Total of 420 localities in USA, for which links were scrapped
- For each of the 420 localities, we extracted the links for subcategories

DATA PRE-PROCESSING AND TRAIN-TEST SPLIT

- Cleaned data
- Removed stop words
- Tokenized
- TF-IDF (After train-test split)
- Test data = 25% of data and Training data = 75% of data

MODELLING AND COMPARISON

- Picked the model which had the best tradeoff between precision and recall

MANUAL TAGGING

- Tagged prediction data (1846 rows in 'General' subcategory into actual categories
- Ran trained model on prediction dataset

Dataset Extraction

The library used for scraping data from craigslist was BeautifulSoup. Key points to note from our scraping include:

- Use of fake user agents – To mimic a human user's behavior in interacting with the website, to reduce likelihood of getting blocked from the website.
- Using time.sleep – To avoid overloading the website (and inadvertently causing a DDOS attack) and to avoid being blocked from the website
- Randomizing the order in which links are read – To ensure data is not too unbalanced with majority of data belonging to a few sub-categories
- Collected info on posts from first 3 pages in the subcategory

Dataset Cleaning

For the data cleaning process, we first removed all the duplicates that we had in our data. Next, we created a new column that has the information on Title+Body so that we are feeding more relevant information into our model. Next, we perform preprocessing of the data such as removing the stop words, converting text to lower case and we have also used regex to get the alphabetical data. Finally, we lemmatized the words to help map multiple words to a common root word.

Data Analysis

Models used:

- Random Forest: Random Forest classifier uses bagging and random subspace method in building each tree to create an uncorrelated forest of trees. The final decision or prediction is made based on most votes from each of the decision tree nodes.
- Logistic Regression: Logistic regression is a statistical model that uses a logistic function for predicting binary classes. Here, a logistic function or logistic curve is a sigmoid curve, and the logistic model has dependent and independent variables where the value of dependent variable is either 0 or 1.
- Logistic Regression (Multi_class=" multinomial")

- Linear SVC: The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is.
- Linear SVC (Multi_class=”OVR”)

The top 3 models based on accuracy that we got are as follows:

- Multinomial Logistic Regression
- Linear SVC
- Linear SVC with OVR

Multinomial Logistic regression is an extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. Like binary logistic regression, multinomial logistic regression uses maximum likelihood estimation to evaluate the probability of categorical membership.

SVC One vs Rest is a heuristic method for using binary classification algorithms for multiclass classification. It involves splitting the multi-class dataset into multiple binary classification problems. A binary classifier is then trained on each binary classification and predictions are made using the model that is the most confident.

Now, since the dataset that we have is highly imbalanced we have applied SMOTE to oversample the minority classes. We will then compare the results that we got before and after SMOTE.

Imbalanced data is data in which observed frequencies are very different across the different possible values of a categorical variable. Basically, there are many observations of some type and very few of another type.

SMOTE (SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE): SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This aids in resolving the overfitting issue generated by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

SMOTE is an algorithm that performs data augmentation by creating synthetic data points based on the original data points. SMOTE can be seen as an advanced version of oversampling, or as a specific algorithm for data augmentation. With SMOTE, you avoid producing duplicate data points and instead produce synthetic data points that are marginally different from the original data points.

The SMOTE algorithm works as follows:

1. Draw a random sample from the minority class.
2. For the observations in this sample, we identify the k nearest neighbors.
3. Then using one of those neighbors identify the vector between the current data point and the selected neighbor.
4. Multiply the vector by a random number between 0 and 1.
5. To obtain the synthetic data point, add this to the current data point.

This operation basically involves slightly moving the data point in the direction of its neighbor. This way, we make sure that our synthetic data point is not an exact copy of an existing data point and making sure that it is also not too different from the known observations in your minority class.

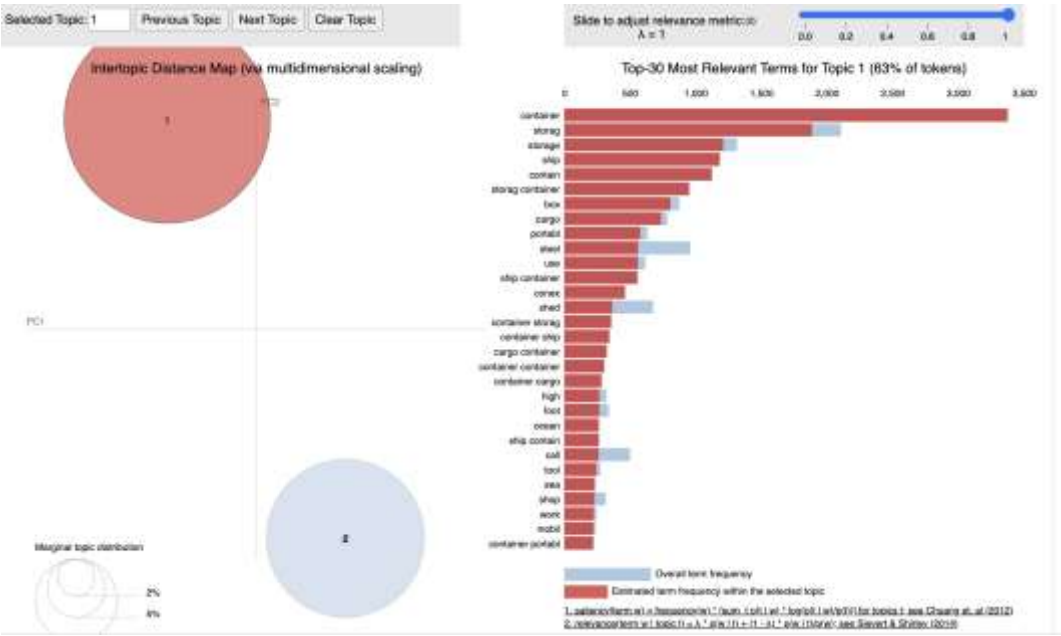
Topic Modeling - LDA : Once we performed our prediction on the prediction dataset, we were able to classify car+bike and beauty+hlth advertisements from the general category. The remaining advertisements were classified in the others category, and we decide to perform topic modeling on the description of these advertisements to check the prominent topics in these advertisements.

This was done using LDA and we selected number of topics as 2. Initially we had selected more number of topics but the words in the topics were being repeated so we decided to tune down the number of topics to 2.

```
Topic 0:
metal build carport building door
Topic 1:
container storag storage ship contain
```

Above are the top words for both topics and we saw that advertisements were mostly about storage and shipping containers or metal carports. This helps the moderators at craigslist when they have to manually tag these advertisements. They can also make a new subcategory for these advertisements to increase visibility.

Below is the inter-topic distance map. The inter-topic distance map is a visualization of the topics in a two-dimensional space. The area of these topic circles is proportional to the amount of words that belong to each topic across the dictionary.



Validation

For the top 5 models we ran, the accuracy on test dataset is as follows:

Index	Model	Accuracy
1	Linear SVC (Multi_class = 'ovr')	95.12%
2	Linear SVC	95.10%
3	Logistic Regression (Multi_class = 'multinomial')	93.07%
4	Random Forest	92.50%
5	Gradient Boosting	91.76%

For our best model, Linear SVC with multi-class classification method 'One vs rest', we used SMOTE. On the test dataset, we got the following results for the classification summary:

BEFORE SMOT

The accuracy of LinearSVC is 0.9512295081967214

	precision	recall	f1-score	support
Others	0.95	0.99	0.97	1998
beauty+hlth	0.93	0.50	0.65	157
cars+trucks	0.98	0.90	0.94	285
accuracy			0.95	2440
macro avg	0.95	0.80	0.85	2440
weighted avg	0.95	0.95	0.95	2440

AFTER SMOT

The accuracy of LinearSVC is 0.9540983606557377

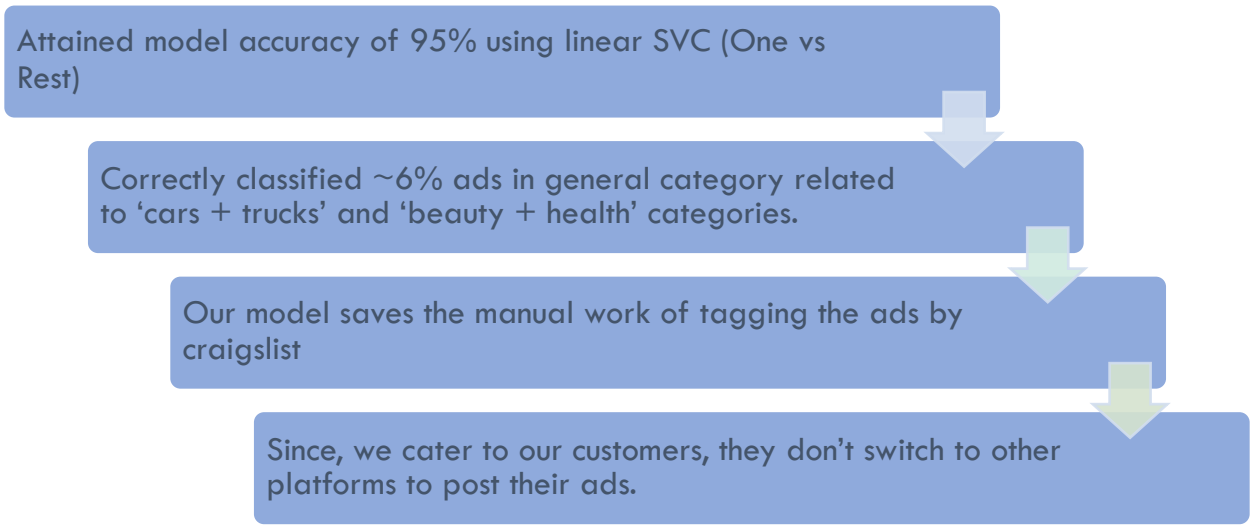
	precision	recall	f1-score	support
Others	0.96	0.98	0.97	1998
beauty+hlth	0.82	0.61	0.70	157
cars+trucks	0.96	0.93	0.95	285
accuracy			0.95	2440
macro avg	0.91	0.84	0.87	2440
weighted avg	0.95	0.95	0.95	2440

Since SMOTE was giving us a better result on account of minorities in data, we went ahead with SMOTE. We got the following result on prediction dataset:

accuracy 0.9516393442622951

	precision	recall	f1-score	support
Others	0.96	0.99	0.97	1998
beauty+hlth	0.81	0.59	0.68	157
cars+trucks	0.97	0.92	0.94	285
accuracy			0.95	2440
macro avg	0.91	0.83	0.86	2440
weighted avg	0.95	0.95	0.95	2440

Conclusion



Recommendations

- Expand the model to classify more categories from general
- Conduct grid search to find the best parameters suited to our model. Can also ensemble best models.
- Incorporate image processing and pricing for better predictions.
- Improve the model to predict other subcategories by implementing external data in our model.