

**MGMT 59000 Analysing Unstructured Data**

Optimizing the user experience for CL’s sporting goods page

End Project Report

Team Hello World! – Aaron Chen, Vinod Iyer, Sriya Musunuru

**> Understanding the business context**

If you have ever used or considered using Craigslist to sell-off your ancient, bulky washing machine when moving into a new apartment or scanned through lists and lists of webpages for a second-hand road bike to help you commute within your college campus, chances are that you would have been shocked to find out that the website’s architecture is as chunky, old-fashioned and dismembered as some of the product listings therein. Why then, a company that was forged during the internet era of uber cool API updates and next-gen website optimization techniques wilfully opt to function as a living relic of our not-so-distant past? Craig Newmark, the founder of Craigslist, answered this very question in 2009:

*“If most people are good and their needs are simple, all you have to do to serve them well is build a minimal infrastructure allowing them to get together and work things out for themselves. Any additional features are almost certainly superfluous and could even be damaging.”*

To further understand where Craig was getting at, we have to zero down on the website’s user demographics. Thankfully though, Craig has got that covered as well. In his words:

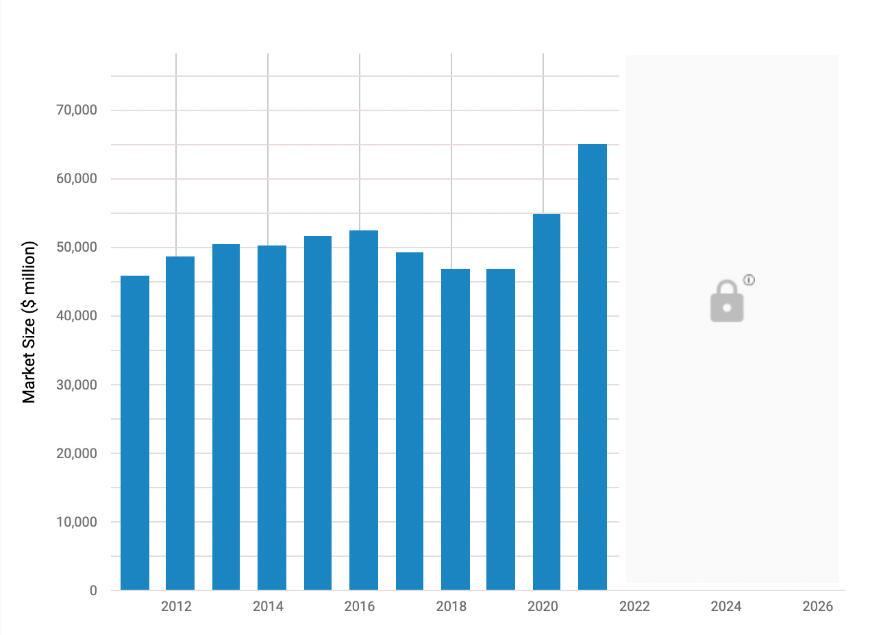
*“Craigslist is about helping people put food on the table, it’s about helping them find a table to put the food on, and lastly, it’s about helping people find a roof under which to put the table with the food.”*

What he was alluding to is that Craigslist was found as an organization that intends to do well by doing good. Any number of snazzy updates would not make an iota of difference if it precludes some sections of the society from taking advantage of it.

**> Getting to the problem – USA sports goods industry**

It can be concluded that Craig considers user inclusivity (even at the cost of technological advancement) not just a crucial metric but also a prime reason for the company’s success. We used these observations as important inputs to evaluate the website for modifications.

The sporting goods store market size in the US in 2021 is $65.1 Billion. From 2012, this industry has seen a constant rise from $47 Billion but saw a sudden dip from 2017. This trend can be seen in the below graph:



It can also be seen that in 2021 this industry saw a rise and hit an all-time high of $65.1 Billion in market size and a revenue of $14.5 Billion, which goes to say that sporting goods are performing very well. It saw an 18.6% annual increase in the USA and the average annual growth rate for the last five years was 4.4%. Hence, we can conclude that industry is going to do very well in the coming years and Craigslist should make a strong presence in this industry as well, in order to achieve everything that the founders envisioned.

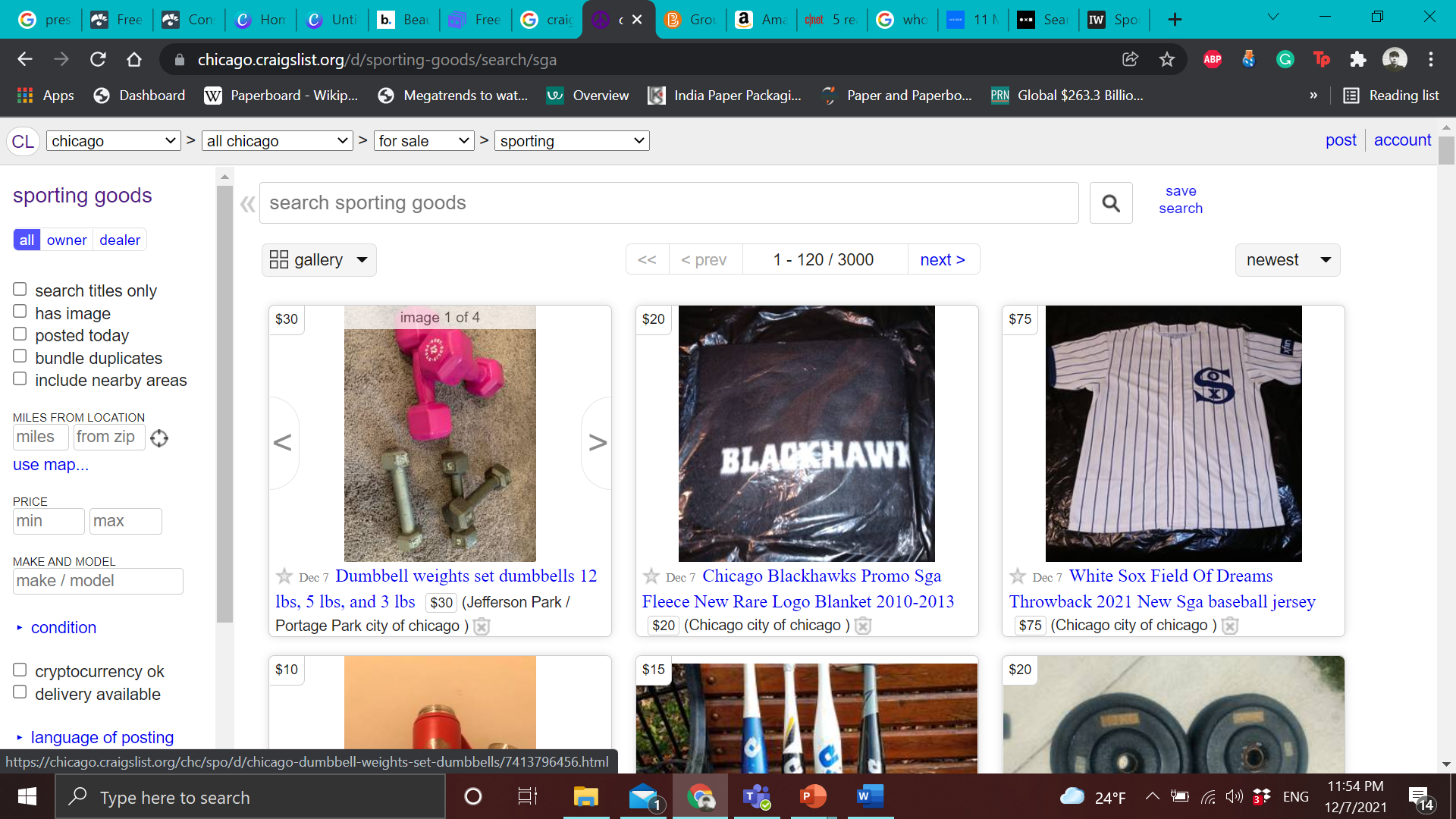
**> Formulating our proposed solution**

As of today, Craigslist has <5000 active sports goods listings across all cities which is very low compared to their potential and their current customer base in other sections. Though their number of listings are low, after manually going through 2400 listings, the numbers of sports categories is above 25. Having over 25 categories with no filters would ruin a user experience.

The above studies show that there is a wide gap between the popularity of sporting goods in USA compared to the relative popularity of the sporting goods page on Craigslist. We, as a group, believe that there is no specific reason for this section of the website to lag behind when compared to other sections on Craigslist, for instance, housing and jobs. When you compare it to the quantum of sporting goods traded on an everyday level on sites like the Facebook Marketplace, our intuition becomes even stronger.

Hence, we would like to use text classification algorithms to classify these listings into various categories which would help improve a user’s experience, which in turn could increase their customer base.

We attribute this relative underperformance to the current user experience of the sporting goods page (<https://craigslist.org/d/sporting-goods/search/sga>), which looks like below:



As you can see, there are no filters for the page. Hence, our proposed improvement solution for Craigslist is:

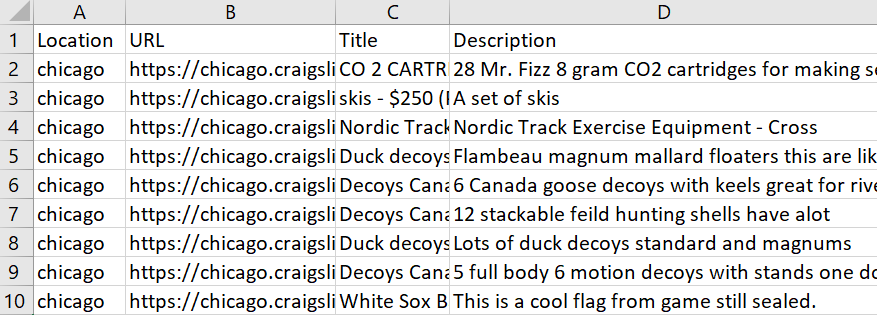
**Craigslist Classification Model**

**Optimizing the user experience on the sports goods section of Craigslist by bucketing the listings into relevant categories in order to better serve the end user.**

**> Data requirements and collection scheme**

For the purposes of classifying the sports goods listing page on Craigslist, we followed a stepwise procedure as follows:

>> Firstly, we scraped the webpages for sporting goods (Listing title and Listing description) traded in the 20 major cities of USA using Selenium and Beautiful Soup packages. We have attached the python codes for scraping along-with this report and the screenshot for the output CSV file is as below:



>> We were able to retrieve about 2400 listings from the 20 cities. The list of the cities is as follows: Chicago City, New York City, San Francisco Bay Area, Washington DC, Philadelphia, Los Angeles, Boston, Detroit, Houston, Phoenix City, Atlanta, Austin, Denver, Dallas, Las Vegas, Miami City, Minneapolis, Portland, Raleigh and San Diego.

> **Preprocessing and LDA Topic Modelling**

>> Preprocessing

We followed the following steps to preprocess our data and make it ready to feed it into models.

# Tokenization: Tokenization is the process of breaking text documents apart into atomic terms. In text analytics, tokens are broken down into words. A sentence of 10 words, then, would contain 10 tokens. We used the NLTK package for the same.

# Stop words Removal: Stop words are the set of commonly used words in any language. Stop words are commonly eliminated from many text processing applications because these words can be completely misleading and does not provide a lot of information.

# Stemming: Stemming is the process of reducing a word to its root word to make sure not a lot of same words are being repeated.

# Lemmatization: Lemmatization is the algorithmic process of finding the lemma of a word completely depending on its meaning.

# Vectorization: This is the process of converting text into a numerical representation and provide weights of every word based on frequency. We used TF-IDF to vectorize our document with a minimum document frequency of 15 and n-gram between 1-4.

>> LDA Topic Modelling

Topic modelling is a type of statistical method for figuring out broad topics within a document.  It builds a topic per document model and words per topic model, modelled as Dirichlet distributions.

The topics obtained are as below. We used a part of this result along-with our logic in order to classify sports categories. The final categories we trained the model with are: Hiking, Hunting, Net-based Sports, Water Sports, Winter Sports, Golf, Health and Fitness, Roller Sports, Baseball and Miscellaneous.



> **Descriptive Analysis and Modelling the data**

>> Descriptive Analysis

# Analysing popularity of categories by city: We have attached the popularity chart for all the 20 cities below, you can double click on the graph to check out the categories.

# In order to check if the data we scraped is correct or not, we checked if colder cities have winter sports as a popular category. Similarly, we checked for bigger cities having health and fitness as a popular choice.



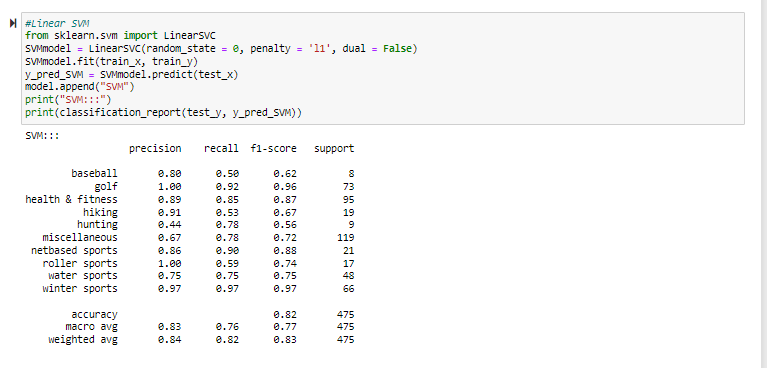
>> Modelling

# Models used: Support Vector Classifier (SVC), Multinomial Naïve Bayes, Random Forest, Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN)

# Packages used: scikit-learn, NLTK, seaborn, matplotlib, imblearn, numpy, pandas

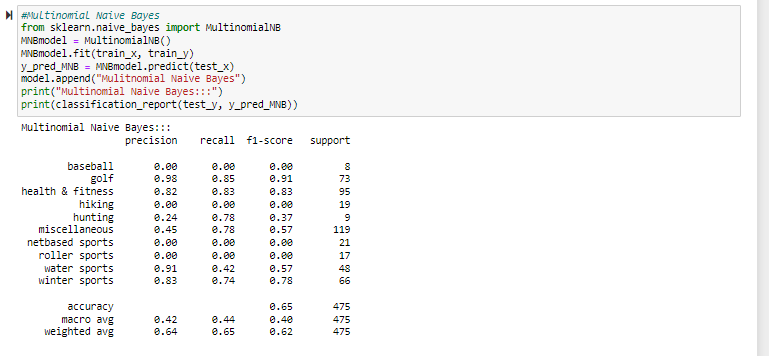
# Support Vector Classifier (SVC): We used LinearSVC(), a model based on support vector machines, as it is ideal for multiclass text classification. The model provided by scikit-learn uses the one-vs-rest strategy where the multiclass classification problem is broken down into several binary classification problems in which each class has a binary class. The class that has the highest accuracy wins and given as the prediction. This model gave the best accuracy out of all the models used on this particular dataset with random\_state as 0 and penalty set to 11.

Results:



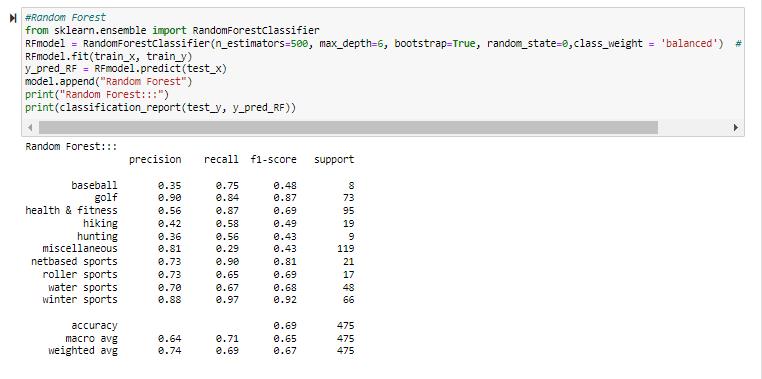
**#** Multinomial Naïve Bayes Results: The Multinomial Naïve Bayes algorithm calculates the tag’s likelihood for a given sample. It is also a suitable class for classification because MNB is based on the naïve bayes probabilistic classifier and follows a multinomial distribution.

Results:



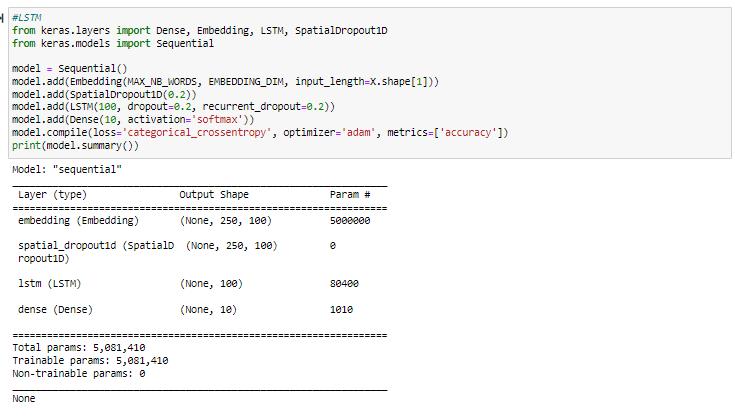
# Random Forest: It is an ensemble classifier where multiple decision tree classifiers are used to estimate the predictor. We chose this model expecting the model to perform better as the random forest classifier uses averages to improve prediction and control overfitting. Here, the max\_depth of the sub-trees is set to 6 and 5000 n\_estimators are used with 0 as random\_state.

Results:



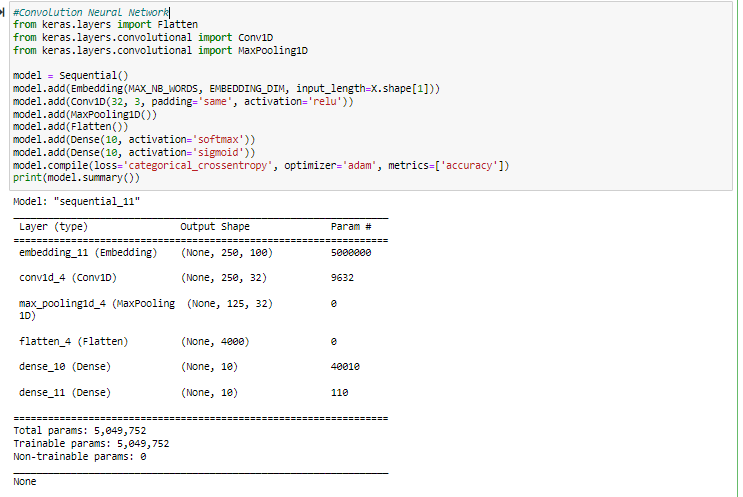
# Long Short-Term Memory (LSTM): For complex text analysis problems like machine translation and speech recognition, LSTM is an ideal model. It is a type of recurrent neural network and can remember the order information. In defining the Sequential model, the Embedding, SpatialDroupout, LSTM and finally, a dense layer with the softmax activation layer is applied. Since it is a multiclass classification problem, the ‘categorical\_crossentropy’ loss function is used. A dropout rate of 0.2 was maintained.

Results:



# Convolutional Neural Network (CNN): Though CNN works best for image classification problems, we tried the model to check if it would run in a similar fashion for text classification as well, as the model persists information from the last run. Here, embeddings layer, 1D Convolution, Max Pooling, and Flatten layers are used to form the neural network. A couple of dense layers are added the end using the `softmax` and `sigmoid` activation functions are used.

We learnt that CNN is not an ideal choice given the accuracy and performance of the other parameters. TextCNN() would’ve been a better choice.



Model Comparison:

**Chart, bar chart

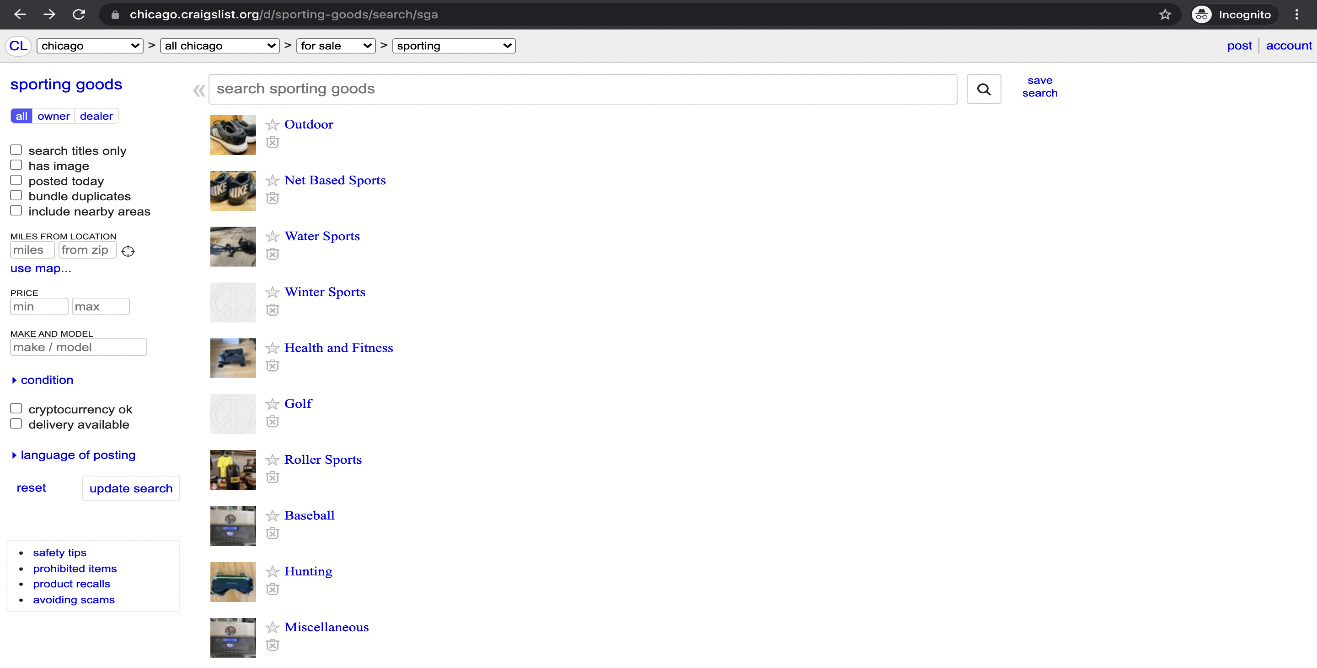
Description automatically generated**

Picking `accuracy` as the metric for selecting the best model had given the above results. Linear SVM comes out to be the best model with an accuracy of 84.13% followed by Multinomial Naïve Bayes and Radom Forest.

Using these models and by fine-tuning the parameters, the sporting goods categories can be accurately classified and the each of the listings can be grouped under their respective category on Craigslist.

> **Results and Way Forward**

>> Results: The proposed output page after our classification would look something like this:



>> Way forward:

# Implement an A/B Testing phase to test for significance

# Given its successful, find the ideal combination of categories & subcategories for the website

# Incorporate classification based on Image recognition to improve the classification accuracy

---------------------------------------------- *end* -------------------------------------------------