Cell Phone Pricing Prediction

MGMT 590: Analyzing Unstructured Data

December 2021

**Background**

-------------------------------------------------------------------------------------------------------------------------

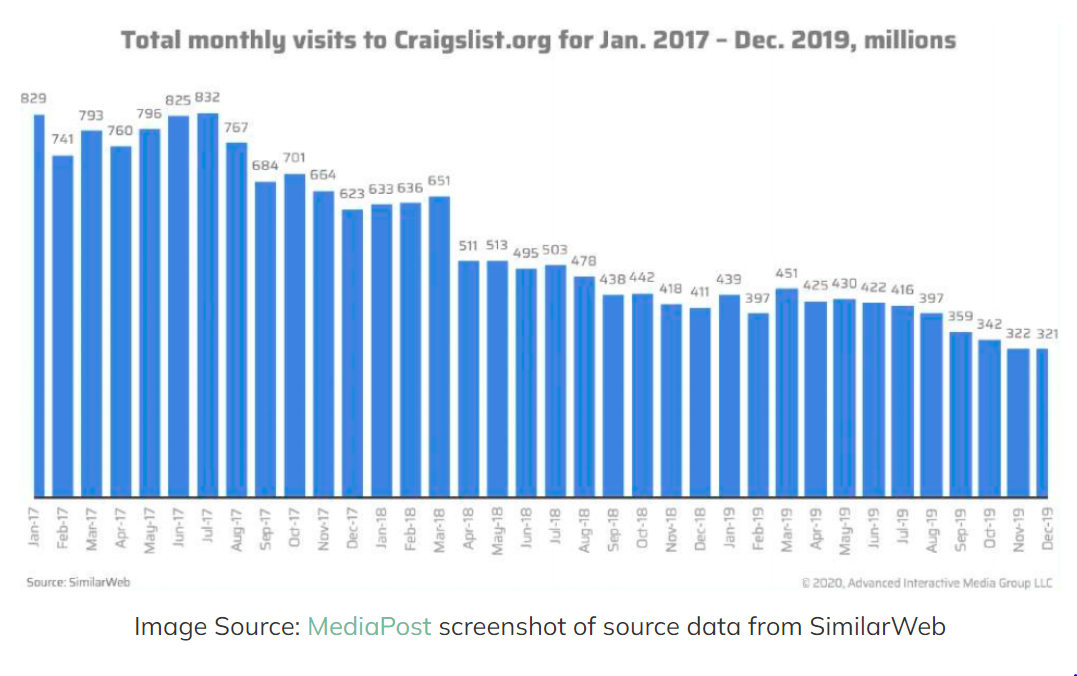
**Business Scenario and Problem**

* **Scenario Description**

Craigslist is an American classified advertisements website that allows users to hold conversations or sell their items. It provides many different sections where people can search for jobs, housing, etc. However, it also makes users spend more time extracting the relevant information they are looking for. Especially when users are browsing the website, finding something they are interested in, and considering buying the products, they would like to know whether the price listed on the certain post is above, under, or the same as the price of similar items on Craigslist.

* **Problem Identification**

Craigslist was the king of classifieds, but the shine is off its crown. The problem that Craigslist encountered was a drastic decline in monthly visits. The decline stems from several factors, such as increasing competition and the Covid-19 pandemic. However, the decline started several years ago. The decline stems from both external and internal factors. External factors involve increasing competition as more and more companies enter this field, such as Facebook and Airbnb. Covid-19 pandemic is another unpredictable external factor. As the pandemic ravaged the world, people were locked at home, and the global economy shrank. Internal factors result from an unfriendly user experience. As people can post whatever they want, the contents are usually unstructured, rendering people to spend more time finding the information or product they want. As customers spend more searching for the information they want and comparing the prices of similar items, they might have bad user experiences, in turn decreasing consumer satisfaction on Craigslist.



**(Source:AIM Group:** [**Craigslist: Traffic, revenue fell radically in 2020**](https://aimgroup.com/2021/02/11/craigslist-traffic-revenue-fell-radically-in-2020/)**)**

Customers always want to purchase an item that is a good bargain, so they spend a lot of time comparing similar products from our website or even among different websites.

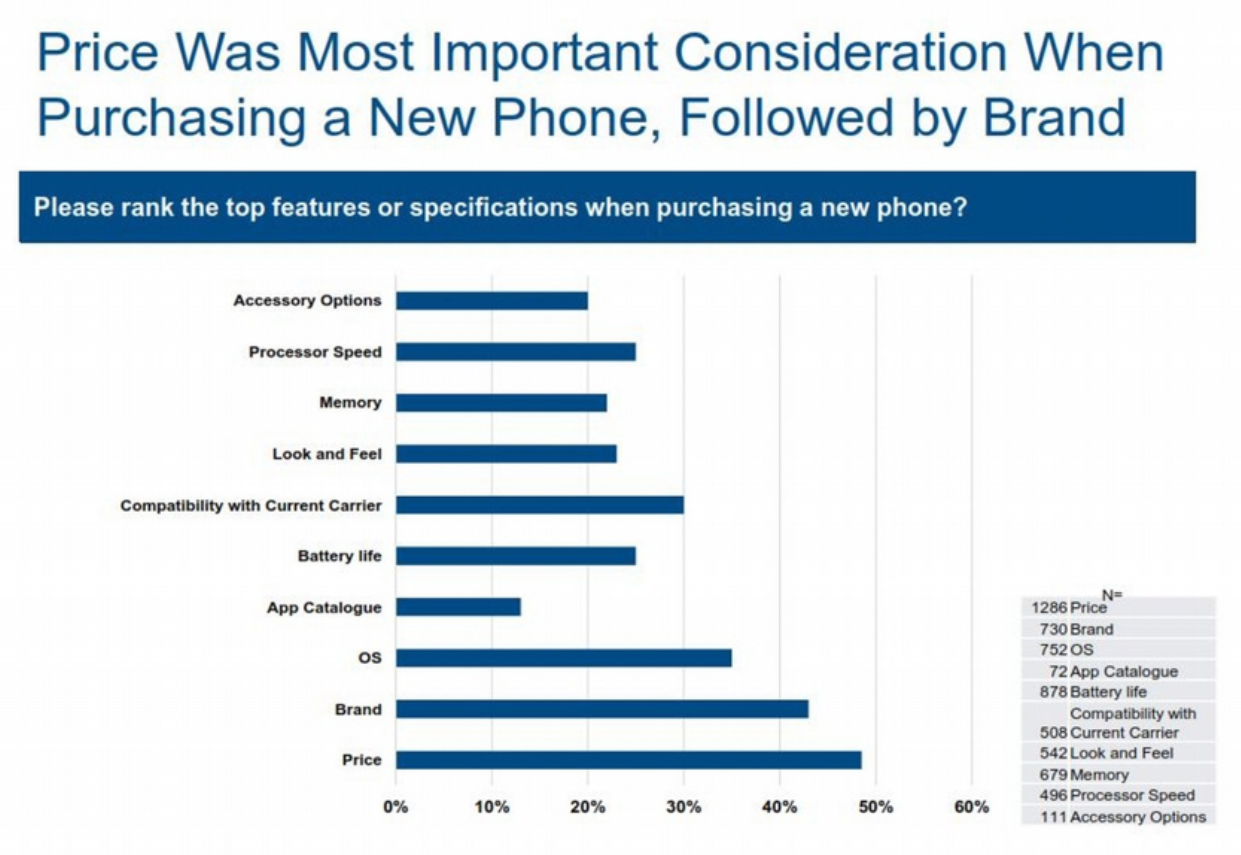
We choose the cell phone section for two reasons. First, the used phone market grows quickly, although the new smartphone market is gradually saturated. Second, the prices of used cell phones are relatively structured, stable. In contrast with the cell phone section, the price of tickets for a ball game may fluctuate drastically. The price of a ticket will become 0 after the game ends, but cell phones are not time-sensitive. Several features needed to be extracted from the description to find the similarity of items, such as conditions, names, numbers, and brands. We can predict the price of a cell phone based on such features.

**Business Analysis**

-------------------------------------------------------------------------------------------------------------------------

**Problem Improvement Proposal**

Study shows that customers care about the price the most when purchasing a new phone. Therefore, we provide a swift tool to predict prices that our customers and sellers can use as a benchmark. When sellers write a description for a new product that they are willing to sell, a corresponding price would be recommended based on the content of the title and description. Once we implement this feature, sellers would be more careful when they are pricing new products or writing descriptions, making the customers collect information more accessible and refining the user experience.



(Source: medium [hyperlink](https://medium.com/@GalaxyeSolution/behind-the-boom-consumer-trends-driving-the-used-phones-sales-upswing-464e89497bdb))

We have collected enough data to train different models and use these models to predict a price for a new product. Different models have their advantages and disadvantages, so we train different models and compare the accuracies of results. The models we built include logistic regression, random forest, decision tree, neural network, LSTM, and SVM.

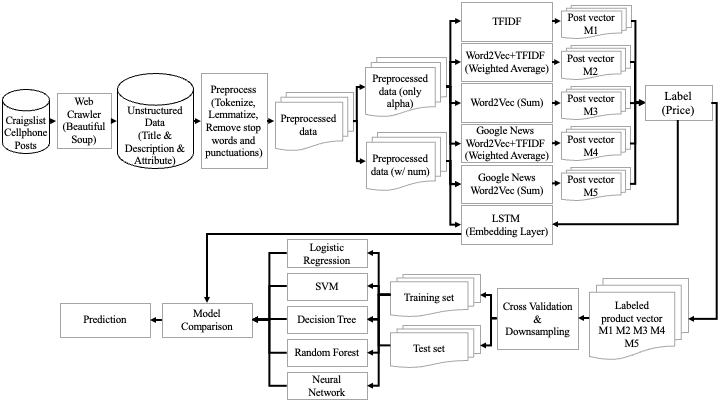
We split the data into a training set and a validation set, use the training set to train the models, and apply the validation set to see how the models perform. If a model performs well, we can use the dataset ensemble to train a model and apply this model to predict results.

By leveraging the functionality, customers could easily compare the price of a product to the prices of items that have a similar description content. When the posted price is overvalued, customers could reasonably assume that there would be a chance that they could find a better deal in another post. On the other hand, when a posted price is undervalued, customers could understand that the product is a bargain or that there might be some risk due to the low price and decide whether they would like to afford the risk. The additional information can help customers make better decisions and increase customer engagement. We will apply this feature to other subsections and create more value for Craigslist.

**Data Analysis**

-------------------------------------------------------------------------------------------------------------------------

**Process Overview**

****Our project starts with scraping data from Craigslist cell phone posts. Next, we perform a series of preprocesses and produce two types of preprocessed documents. After that, we apply six text representations and generate five types of post vectors, referred to as M1, M2, M3, M4, and M5, in the above graph. Once x, post vector, and y, label, are prepared, we split our data and train classification models.

**Data Requirement and collection scheme**

* **Collection Scheme**

1. Collect Title, Location, Price, Description, Attributes, Post ID, Post Date, Update Date, Notice, and Url of each post under the for sale/cell phones section by searching the area 200 miles from the Los Angeles area. The URL of each post is identified by post id. To better track the status of web scraping, we get the URL of each post to check and ensure that we do not collect any other data during scraping.
2. Use python web scraping package, beautiful soup to scrape all the searching results.

We employ the following methods to avoid getting blocked by the website:

* time.sleep(): We add time.sleep() to stop the script from sending requests too frequently. We set the script to sleep from the 20s to 25s between every page.
* User agents: We create a list of user agents to make the web crawler pretend like a real web browser.

1. Create an output CSV file that contains 10 columns (Title, Location, Price, Description, Attributes, Post ID, Update Date, Notice, and Url) and 1573 rows.

**Data Analysis Process**

* **Preprocess**

We remove irrelevant columns such as Post ID, Location, and so on from the raw data. ”QR Code Link to This Post” for each description and column names of each value in the attributes are removed as well. Then, we combine the content of the title, description, and attributes columns as our model input.

After tokenizing, lemmatizing, and removing stop words and punctuations, we create two types of preprocessed documents. One document contains only alphabetic words, and the other includes both numbers and alphabetic words. Cell phone model names usually have numbers; we assume that if text documents include complete model names, models will produce better performance.

As for the target variable, we remove outlier values and classify prices into 17 price ranges.

* **Text Representation**

We apply six types of vector representations. The representations are listed below.

1. **TF-IDF**: Generate a probability of each word by its importance in the post
2. **Word2Vec (self-trained) (sum)**: Sum up all the word2vec word vectors in a post and use them to represent each post. The word2vec model is trained by the Craigslist cell phone processed data.
3. **Word2Vec (self-trained)+TF-IDF:** Use TF-IDF weighted value to do weighted average on word vectors, thereby generating a weighted average vector that represents a post vector. The word2vec model is trained by the Craigslist cell phone processed data.
4. **Word2Vec (pre-trained) (sum)**: Sum up all the pre-trained word2vec word vectors in a post and use them to represent each post. The word2vec model is trained by google news.
5. **Word2Vec (pre-trained) +TF-IDF**: Utilize the vector from google news word2vec, a pre-trained model, to represent each word and combine it with TF-IDF, generating a vector that represents a post vector
6. **LSTM:** Contains embedding layer to do text representation

* **Labeling**

The post vectors are labeled by the price of the post. Each price range is assigned with a label. ($0-$50: label 0, $51-$100: label 1, $101-$150: label 2,etc. and a total of 17 labels).

* **Down Sampling**

The target labels are highly right-skewed. Almost half of the target values belong to the label 0. This would lead to a problem that when we train the model, the model would learn to predict 0 all the time. To deal with the imbalanced issue, we randomly selected some samples from our majority class and set the sample size to be the same as the number of the second largest class.

* **Cross-Validation**

We set the random seed of the split data function to be the same so each model can be comparable. Data is randomly partitioned into training and test data with a split ratio of 0.9. Models are trained by the training data and evaluated by the test data.

* **Classification Models**

Six classificationmodels are applied in this analysis: Logistic Regression, SVM, Decision Tree, Random Forest, Neural Network, and LSTM.

* **Model Comparison**

We choose the highest accuracy on the test set from each combination of preprocessed data, text representation, and classification model as the prediction model.

* **Prediction**

The highest accuracy combination of models would be input with the new data (title, description, and attributes) to predict the suitable price for the post item.

**Vlidation**

-------------------------------------------------------------------------------------------------------------------------

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Document | Text-Representation Method | **Logistic Regression** | **Decision Tree** | **Random Forest** | **SVM** | **Neural Network** | **LSTM** |
| **w/ num** | **TFIDF** | 37.62% | 37.62% | 20.79% | 44.55% | 14.85% | 16.83% |
| **Weighted Avg** | 43.56% | 36.63% | 26.73% | 43.56% | 35.64% |
| **sum** | 42.57% | 37.62% | 27.72% | 47.52% | 14.85% |
| **pretrained\_sum** | 38.61% | 33.66% | 30.69% | 31.68% | 14.85% |
| **pretrained\_weighted\_avg** | 33.66% | 33.66% | 20.79% | 38.61% | 14.85% |
| **only alpha** | **TFIDF** | 35.64% | 35.64% | 25.74% | 43.56% | 32.67% | 16.85% |
| **Weighted Avg** | 36.63% | 32.67% | 25.74% | 36.63% | 25.74% |
| **sum** | 36.63% | 34.65% | 26.73% | 29.70% | 21.78% |
| **pretrained\_sum** | 36.63% | 30.69% | 28.71% | 39.60% | 26.73% |
| **Pretrained\_Weighted Avg** | 34.65% | 31.68% | 26.73% | 36.63% | 14.85% |

The table above represents the heatmap of six model comparisons by accuracy.

SVM model has the best model performance since its overall average accuracy is higher than the other five models. The second and third well-performance models are the Logistic Regression Model and Regression Model. Their accuracies are all above 30% with each different document or text representation method. On the other hand, LSTM has the worst performance in our model comparison as its accuracy is about 16.84% with either number included document or number excluded document.

We intuitively think that TFIDF could be a good leverage of weight. Therefore, using TFIDF to calculate the weighted average of the Word2vec vector should have the highest accuracy. However, the SVM model with the number included in the document and using the sum of self-trained word2vec vectors as the text representation method has the highest performance (accuracy 47.52%).

**Conclusion**

-------------------------------------------------------------------------------------------------------------------------

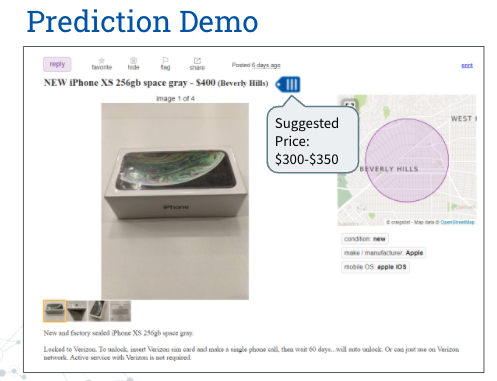
After trying on 2 kinds of processed data, 6 types of representations, and 6 types of classification models, we found out that our best model is built by the data contains numbers, preprocessed by self-trained Word2Vec aggregation, and trained using the SVM method, which results in an accuracy close to 48%.

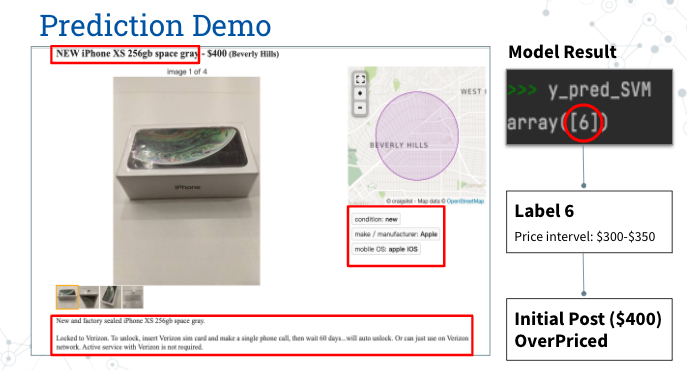
There are multiple findings that we have identified from the comparison among the outcomes from different models. The model trained with the numbers included outperformed the model that only contains words. Therefore, we infer that since we were trying to predict cellphone products, the product model number of the phone is highly related to their price, contributing to the difference in performance. Furthermore, we found out that google news pre-trained words have a low relationship to our data. We assume that Google news might have less cell phone selling information content. Thus, the Word2Vec model we pre-trained by Craigslist performed better than the one using Google news.

For further improvements, we could make more effort on the data preprocessing, since only when the representation and the label are well defined will the model provide better results. For example, removing the repeated word in a single post so that the weight of every single word would be calculated more precisely. Additionally, the decision of the price range could also make a massive difference to the outcome. Moreover, since all of our models were built in a default design, we believe that tuning would contribute to an improvement to the accuracy of the models, making the model more robust. Ensemble model would also be an option for us to try further. Lastly, we only used approximately a thousand posts on Craigslist to train the model. We believe that if more data could be collected, the performance of the models would be a lot better.

In conclusion, price is a volatile target that is relatively hard to predict. However, the accuracy of our best model is up to three times higher than the probability when randomly selecting the price range within the total 17 categories. Furthermore, our text representation using the sum of Word2Vec in a post outperformed the one using TFIDF since TFIDF only represents the possibility that each word appears within the collection of posts. On the other hand, the sum of Word2Vec uses the vector trained from the Word2Vec model of Craigslist, making a more accurate prediction of the target price.

The conceptual design of the user interface would be as follows:





The graph above is a demonstration showing how the model works. By inputting the product name “NEW iPhone XS 256GB space gray”, the product description, and the labels framed by the red frame in the graph, the model would automatically output the predicted price learned from the other posts that have similar content. Given a predicted price, customers could have a basic understanding of whether the price from the current post is relatively overvalued or undervalued. The additional information would boost the user experience since sellers would pay more attention to their pricing and the product description. The function would genuinely add value to Craigslist’s business.

To summarize the above, we could prove that our current price prediction model based on product description is definitely feasible to put into practice. We would recommend that Craiglist apply the model to their user interface to enhance user experience and improve customer satisfaction.