MGMT 590 ANALYZING UNSTRUCTURED DATA

Enhancement of Craigslist's Sales Potential via

Product Recommendation System

Mitchell E. Daniels, Jr.
School of Business

Team 3:

Abhishek Krovvidi
Nagarjuna Chidarala
Sathwik Kanukuntla
Shivam Mishra
Sohan Kumar Sahoo

Table of Contents

I. Background

Craigslist, an innovative American classified ads website, offers a platform for local community connections in various categories like jobs, housing, and community activities. It stands out with its free ad postings and simple, text-based design, resembling traditional paper classifieds. The website enhances user search functionality, facilitating connections across diverse needs. Despite its minimalist design, Craigslist has expanded globally, creating subsites in over 70 countries. Its unique approach of offering diverse, free services and a straightforward interface has cemented its position as a vital online community connector. The website's format resembles traditional paper classifieds, but Craigslist introduces a significant enhancement—the ability for individuals to search for and find ads that specifically interest them. This feature streamlines the process of connecting buyers with sellers, employers with job seekers, and individuals with shared community interests. In summary, Craigslist's unique combination of a non-commercial ethos, diverse categories, free postings, and global reach has positioned it as a prominent online platform for connecting people, services, and opportunities within local communities and beyond.

II. Problem Identification:

1. Difficulty Finding Complimentary Products:
    - Challenge: Craigslist lacks a sophisticated recommendation system that suggests complementary products or services based on user searches or interests. This makes it challenging for users to discover related items that could enhance or complement their original search

- Impact: Craigslist misses out on opportunities for cross-selling and upselling, as users may not be aware of additional products or services that align with their needs or preferences

2. Limited Awareness of Complimentary Products:
    - Challenge: Users may not be informed about the availability of complementary products related to their searched or desired items. Without proactive suggestions, users might miss out on valuable offerings that could meet their needs or preferences
    - Impact: This lack of awareness hampers user experience and limits the potential for users to explore and engage with a broader range of relevant content on the platform
3. Substantial Effort to Navigate the Website:
    - Challenge: The Craigslist interface is often criticized for being cumbersome, content-heavy, and overwhelming. Users may need to invest significant effort to sift through extensive listings, leading to a less-than-optimal user experience
    - Impact: The information overload and difficulty in navigation may result in user frustration, longer search times, and a potential decrease in user satisfaction. Users may not find the platform as user-friendly or efficient as other alternatives

III. Project Objective

Our team has developed an innovative automated recommendation system specifically designed for Craigslist. This system generates a curated list of product suggestions in response to user queries. The list of product recommendations is carefully curated to ensure diversity and prioritized according to factors such as geographical proximity, the freshness of the advertisement, and the

condition of the product. As a case study, we have focused on televisions (TVs) searched by a user from Indianapolis, but the framework is adaptable to various products and locations.

IV. Business Impact

Our automated recommendation system has the potential to create a win-win situation for both users and Craigslist. Users receive personalized recommendations, while the platform benefits from increased user engagement, improved brand image, and diversified revenue streams. This innovation positions Craigslist as a more competitive and dynamic marketplace in the online classifieds industry. This aims to balance user convenience with Craigslist's business goals. Strategically, this system is expected to boost user engagement and sales, offering Craigslist a competitive edge in the online marketplace. Leveraging data-driven insights, Craigslist will enhance user experience and facilitate effective marketing strategies. This comprehensive approach seeks to transform Craigslist into a more user-centric and commercially robust platform.

V. Data Analysis

A. Data Collection and Preparation:

We utilized web scraping techniques, specifically with Selenium, to extract data from two primary sources: Craigslist, focusing on classifieds, and Amazon, targeting e-commerce focusing on TV-related products. From Craigslist, we extracted data on 848 products across categories like Electronics, Appliances, and Furniture, including details like Product Name, Category, Description, Condition, Geo-Position, Posted Date, and Post ID. For Amazon, we

gathered information on 10 products, concentrating on Product Name and Description. This comprehensive data collection forms the foundation of our analysis and model development.

B. Model Analysis

1. Similarity measure:

Corpus of Amazon Product and Craigslist product descriptions: The first step involves creating a collection of Amazon products and random Craigslist products descriptions from different categories based on the searched product which is TV in our case. This corpus comprises textual representations of various products available on Amazon and Craigslist, including their names and descriptions.

Distance measure: For each Craigslist product, the system calculates the distance from the Amazon corpus using similarity measures and filters the shortest measure for each of Craigslist product. Used similarity measures are Cosine similarity, Jaccard similarity, dice similarity, overlap and vector distance.

2. Developing a classification model

I. Creation of Dataset:

We created a dataset containing two key columns:

- Product ID: Identifiers for the Craigslist products.
- Distance: The calculated shortest distance from the Amazon product.

II. Manual Classification of Craigslist Products:

Each Craigslist product in dataset was manually classified into one of two classes:

- Class 1 (1): Products related to TV.
- Class 0 (0): Products not related to TV.

III. Splitting into Training and Test Sets:

The dataset containing Craigslist product descriptions, along with their distances and manual classifications, was split into two subsets:

- Training Dataset: Used to train the Random Forest model.
- Test Dataset: Reserved for evaluating the model's performance.

IV. Random Forest Training:

Employing the training dataset, we applied the Random Forest algorithm. In this process:

- Input Features: The only parameter considered was the calculated distance
- Output/Target Variable: The manually assigned classes (related to TV or not)

V. Model Testing:

- The trained Random Forest model was then applied to the test dataset for evaluation. The model used the distance as the input feature to predict the class of each Craigslist product

VI. The above steps were repeated for different similarity measures

C. Model Validation

Evaluation of Model Performance: The 'Random Forest' model was applied on the various distance and similarity coefficients (vector distance, Dice's coefficient, Jaccard's coefficient, Cosine coefficient, and overlap coefficient). The accuracies of the random forest model for each of the above coefficients is given below:

```
Test Score for Overlap Coefficient: 0.8529411764705882
Test Score for Vector Distance: 0.8647058823529412
Test Score for Dice Coefficient: 0.788235294117647
Test Score for Jaccard coefficient: 0.8058823529411765
Test Score for Cosine Coefficient: 0.8352941176470589
Test Score for Overlap Coefficient: 0.8529411764705882
```

The model with the best accuracy is the Random Forest model for the vector distance co-efficient with a value of 86.47%. The second-best model is the Random Forest model for the overlap co-efficient with a value of 85.29%. The model with the least accuracy was the Random Forest model with the Dice co-efficient with an accuracy value of 78.82%.

D. Prioritization

I. Topic Modeling using LDA:

- Data Selection: Only Craigslist products classified as 'Related to TV' by the Random Forest model are used for topic modeling.
- TF-IDF Matrix: A Term Frequency-Inverse Document Frequency (TF-IDF) matrix is created from the Craigslist product descriptions related to TVs. This matrix represents the importance of terms within each product description.
- LDA (Latent Dirichlet Allocation): LDA is applied to the TF-IDF matrix to identify different topics and the distribution of topics for each Craigslist product related to TVs.

II. Clustering Based on Topics:

- For each Craigslist product, the LDA model assigns probabilities to different topics. The product is then assigned to the cluster corresponding to the topic with the highest probability.

III. Sorting Within Clusters:

Distance Sorting:

- Using the coordinates of the listed Craigslist products, the system calculates the distance between them. The products within each cluster are then sorted in ascending order based on this distance. This sorting algorithm ensures that product as with closer geographical proximity are prioritized within each cluster.

Condition Sorting:

- Each Craigslist product is categorized as 'good,' 'bad,' or 'excellent' based on its condition. The products within each cluster are sorted to prioritize products with an 'excellent' condition.

Ad Posted Day Sorting:

- Craigslist products have an ad posted date. The sorting algorithm uses this variable to arrange products within each cluster based on the recency of the ad posting. Recent ads are given priority in the sorted order.

VI. Conclusion

In conclusion, the implemented methodology, involving web scraping, similarity measures, and machine learning models for classification and clustering, has successfully addressed the business problem of suggesting relevant products on Craigslist based on Amazon recommended listings. By seamlessly integrating web scraping techniques to gather data from both platforms and employing advanced similarity measures, the system not only enhances the accuracy of product recommendations but also demonstrates scalability across various product categories. The classification model effectively categorizes Craigslist products, while the clustering model, leveraging topic modeling, further refines suggestions based on similarities. This systematic approach not only holds the potential to boost sales revenue by offering users personalized and contextually relevant recommendations but also contributes significantly to heightened user engagement. The success of this recommended system underscores its adaptability and effectiveness in enhancing the overall user experience across diverse product domains.

VII. Appendix – Model Flow diagram