ANALYZING UNSTRUCTURED DATA FALL 2023

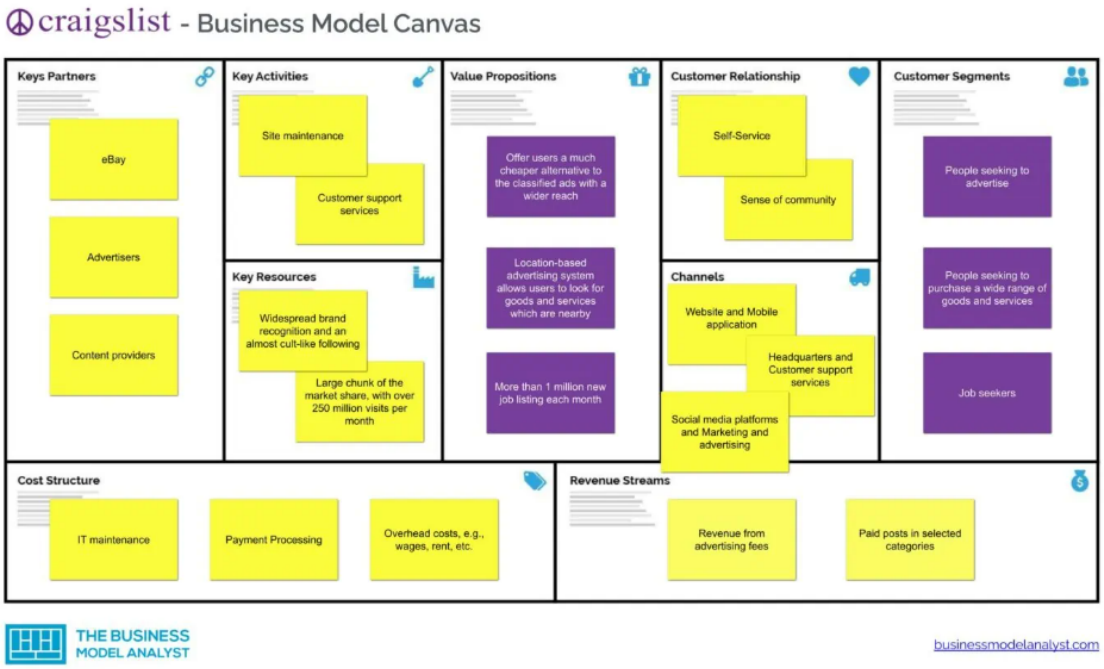
MEMBERS
Akanksha Singh
Keertana Madan
Aishwarya Ajaykumar
Harshraj Jadeja
Devarshi Sharma

BACKGROUND

Craigslist: A Pioneering Marketplace and Its Quest for Enhanced Categorization

Client Overview: Craigslist, a household name in classified advertisements, began as a simple email distribution list in 1995 and has since burgeoned into an online behemoth. Headquartered in San Francisco, California, it has etched its presence in over 700 cities across 70 countries, providing a versatile platform for job listings, housing, personal ads, and various services. Its transition from a non-profit to a for-profit entity in 1999 marked a significant turn in its operational ethos, now serving over 500 million monthly visitors worldwide.

Business Model and Operations: Craigslist stands out for its democratic approach, allowing virtually anyone to post ads in various categories, mostly free of charge. This has cultivated a vibrant, diverse marketplace that's underpinned by a revenue model focused on specific category fees, striking a balance between accessibility and profitability. The platform's business model canvas reveals a robust ecosystem, with revenue primarily generated through these listing fees, which can range from $3 to $75. Despite these charges, Craigslist's commitment to affordability has never wavered, ensuring it remains an accessible platform for all.



craigslist - Business Model Canvas

Keys Partners

- eBay
- Advertisers
- Content providers

Key Activities

- Site maintenance
- Customer support services

Key Resources

- Widespread brand recognition and an almost cult-like following
- Large chunk of the market share, with over 250 million visits per month

Value Propositions

- Offer users a much cheaper alternative to the classified ads with a wider reach
- Location-based advertising system allows users to look for goods and services which are nearby
- More than 1 million new job listing each month

Customer Relationship

- Self-Service
- Sense of community

Channels

- Website and Mobile application
- Headquarters and Customer support services
- Social media platforms and Marketing and advertising

Customer Segments

- People seeking to advertise
- People seeking to purchase a wide range of goods and services
- Job seekers

Cost Structure

- IT maintenance
- Payment Processing
- Overhead costs, e.g., wages, rent, etc.

Revenue Streams

- Revenue from advertising fees
- Paid posts in selected categories

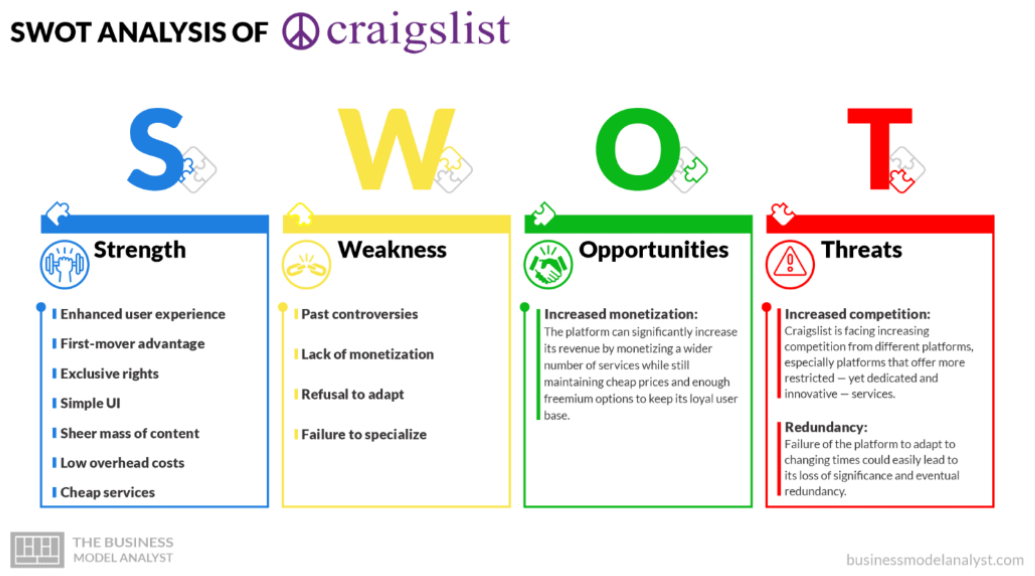
THE BUSINESS MODEL ANALYST

businessmodelanalyst.com

Subsection of Interest: Computer and Computer Parts Listings: The computer and computer parts section of Craigslist represents a bustling digital marketplace where individuals and companies converge to buy and sell. Yet, it is here that a significant challenge has arisen: the misclassification and inefficient categorization of listings, which hampers the user experience and dilutes the platform's usability.

Problem Definition: In a subsection awash with diverse offerings, users frequently encounter a disjointed search experience, finding unrelated items such as iPads when searching for laptops. This indicates a critical weakness in the platform's categorization capabilities—a flaw that undermines Craigslist's strength in simplicity and sheer volume of content.

The Proposed Intelligent Categorization and Tagging System: In response to these issues, we propose an intelligent categorization and tagging system tailored for the computer and computer parts listings. By leveraging advanced NLP and image recognition algorithms, our solution promises to auto-suggest the most fitting categories and generate precise tags, thereby refining the listing process and enhancing search efficiency.



Strength

- Enhanced user experience
- First-mover advantage
- Exclusive rights
- Simple UI
- Sheer mass of content
- Low overhead costs
- Cheap services

Weakness

- Past controversies
- Lack of monetization
- Refusal to adapt
- Failure to specialize

Opportunities

Increased monetization: The platform can significantly increase its revenue by monetizing a wider number of services while still maintaining cheap prices and enough freemium options to keep its loyal user base.

Threats

Increased competition: Craigslist is facing increasing competition from different platforms, especially platforms that offer more restricted – yet dedicated and innovative – services.

Redundancy: Failure of the platform to adapt to changing times could easily lead to its loss of significance and eventual redundancy.

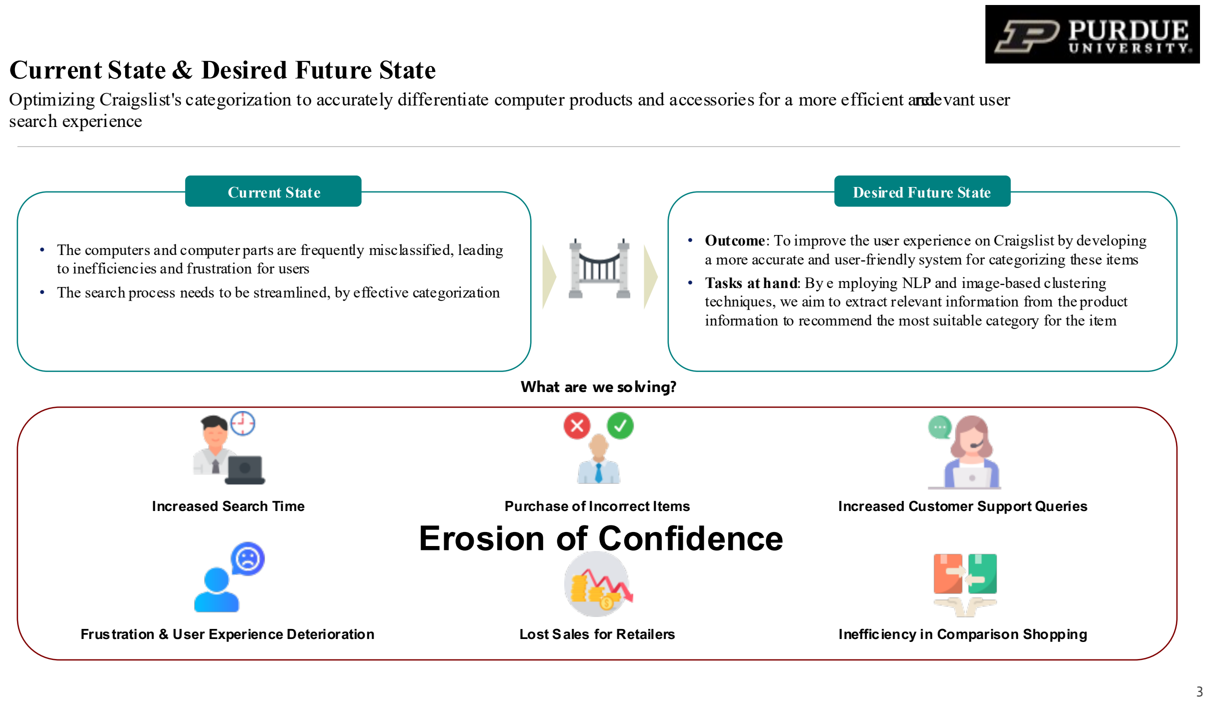
Addressing Weaknesses and Leveraging Strengths: Our solution directly engages with Craigslist's weakness in monetization and reluctance to adapt by potentially increasing user engagement—paving the way for new monetization avenues through premium features and ad placements. By minimizing the need for manual oversight of misclassified ads, we also anticipate a reduction in operational costs, transforming a weakness into a strategic advantage.

Enhancing Precision and Relevance in Searches: By ensuring that computer accessories and specific types like laptops are correctly categorized, the system aims to sharpen the precision of searches, bolstering the visibility of accurately listed items. This improvement is not just about enhancing user experience; it's about keeping Craigslist competitive in a market that increasingly values specialization and precision.

Strategic Adaptation for Future Growth: The adoption of this system represents a strategic adaptation for Craigslist. It aligns with the platform's foundational strengths—enhanced user experience, low overhead costs, and a simple UI—while preparing it for future growth opportunities. This move is a proactive step towards specialization within the broad marketplace, ensuring Craigslist's relevance and competitive edge are maintained.

BUSINESS ANALYSIS

Intelligent Categorization and Tagging System: Transforming Craigslist's Computer Listings



Current State & Desired Future State

Optimizing Craigslist's categorization to accurately differentiate computer products and accessories for a more efficient and relevant user search experience

Current State

- The computers and computer parts are frequently misclassified, leading to inefficiencies and frustration for users
- The search process needs to be streamlined, by effective categorization

Desired Future State

- **Outcome**: To improve the user experience on Craigslist by developing a more accurate and user-friendly system for categorizing these items
- **Tasks at hand**: By e mploying NLP and image-based clustering techniques, we aim to extract relevant information from the product information to recommend the most suitable category for the item

What are we solving?

Erosion of Confidence

Increased Search Time

Purchase of Incorrect Items

Increased Customer Support Queries

Frustration & User Experience Deterioration

Lost Sales for Retailers

Inefficiency in Comparison Shopping

3

Introduction and Strategic Vision

Craigslist has long been a foundational pillar in the digital classifieds market. Its inception heralded a new era in online peer-to-peer transactions, largely due to its expansive reach and the simplicity of its user interface. However, as the digital marketplace evolves, it becomes increasingly clear that a smarter, more user-focused approach is necessary. The introduction of an intelligent categorization and tagging system represents a pivotal change for Craigslist, utilizing advanced machine learning and data analytics techniques to significantly enhance user experience and open up new avenues for monetization.

Objective: Refining Categorization

Central to our initiative is the resolution of the pervasive issue of misclassified listings, especially within the computer and computer parts section, a highly trafficked category on Craigslist. By deploying sophisticated algorithms, our aim transcends mere correct placement of listings; we seek to ensure these listings are easily discoverable by potential buyers, thereby enhancing the overall marketplace efficiency.

Optimizing the User Journey

A user's journey, spanning from the moment of listing an item to finding the desired product, should be a streamlined and intuitive experience. Presently, this journey is often hindered by issues like the misplacement of computer accessories or the appearance of unrelated products in specific search queries. Our project directly addresses these challenges, seeking to provide a seamless and effective process for both buyers and sellers.

Technical Implementation: Advanced Categorization and Tagging

We propose the development of a cutting-edge system employing natural language processing (NLP) to analyze textual content and machine learning techniques for image interpretation. This dual-pronged approach is designed to intelligently suggest the most fitting categories for listings while generating relevant tags to improve searchability and visibility.

Data Analysis: The Foundation of Intelligent Categorization

At the core of our approach lies a comprehensive analysis of listing descriptions and associated images. This data forms the backbone of our categorization algorithms, ensuring the system is not only theoretically robust but also practically attuned to the real content of the listings.

User-Centric Refinement

A pivotal aspect of our project is the ongoing analysis of user search queries and feedback. Understanding user interactions with the system allows us to continuously refine our algorithms, aligning the outputs of the system with user expectations and behaviors, thereby enhancing the overall user experience.

Addressing Classified Ad Challenges

This system directly confronts inherent challenges in classified advertising, such as the limited reach of improperly categorized ads and the heightened costs associated with broad, non-targeted advertising efforts. By improving the precision of our categorization, we aim to ensure each ad reaches its intended audience more effectively, reducing inefficiencies and enhancing the likelihood of successful transactions.

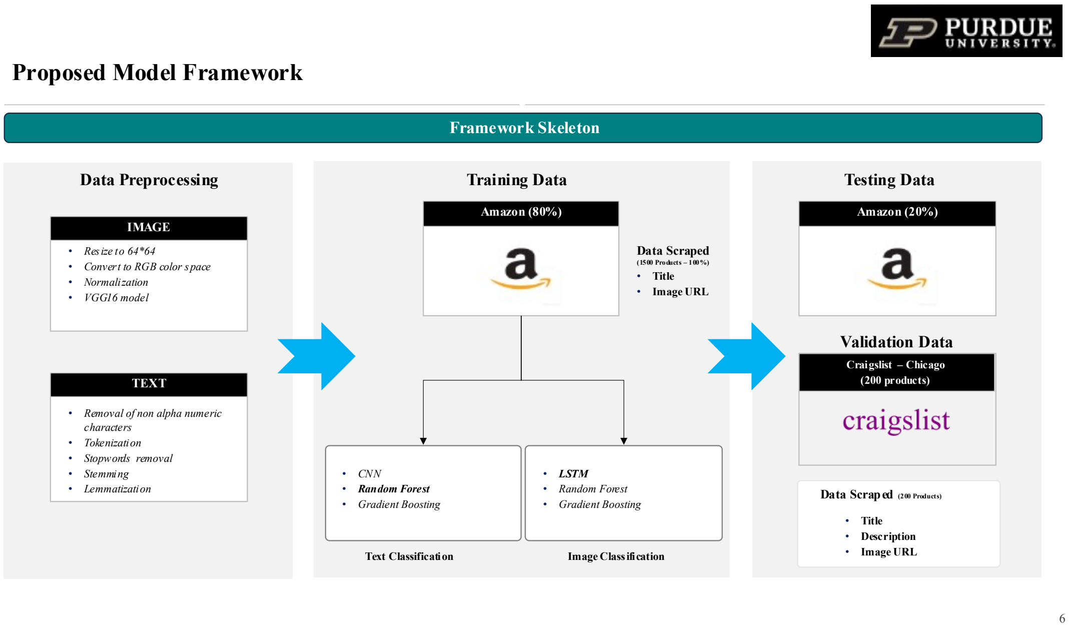
Real-World Impact

The implications of this project go beyond the digital boundaries of Craigslist. By establishing a new standard in classified ad categorization, we are poised to influence the broader online advertising ecosystem, advocating for a shift towards more intelligent, data-driven approaches.

Conclusion: A New Era for Online Classifieds

In conclusion, this project represents more than a mere technological upgrade. It signifies a strategic shift towards a future where classified listings are intelligently categorized, easily discoverable, and precisely targeted. This initiative is set to not only elevate user satisfaction but also solidify Craigslist's position as a leader and innovator in the online classifieds space.

DATA ANALYSIS



Proposed Model Framework

Framework Skeleton

Data Preprocessing

IMAGE

- *Resize to 64*64*
- *Convert to RGB color space*
- *Normalization*
- *VGG16 model*

TEXT

- *Removal of non alpha numeric characters*
- *Tokenization*
- *Stopwords removal*
- *Stemming*
- *Lemmatization*

Training Data

Amazon (80%)

Data Scraped (1500 Products – 100%)

- **Title**
- **Image URL**

- *CNN*
- ***Random Forest***
- *Gradient Boosting*

Text Classification

- ***LSTM***
- *Random Forest*
- *Gradient Boosting*

Image Classification

Testing Data

Amazon (20%)

Validation Data

Craigslist – Chicago (200 products)

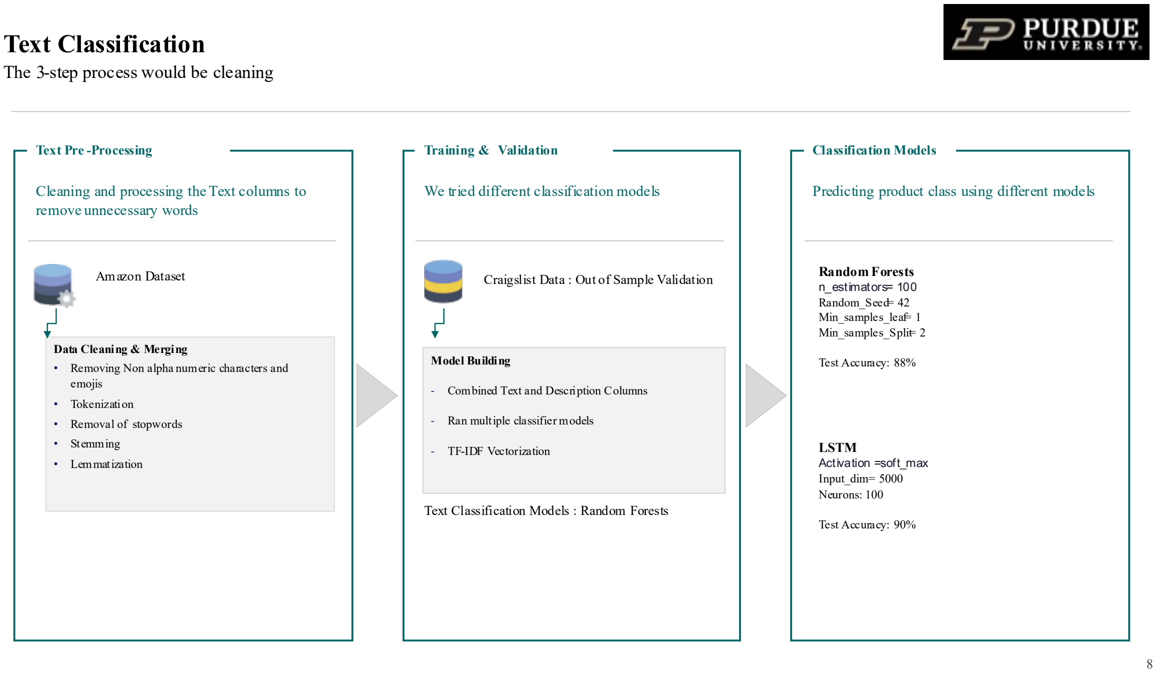
craigslist

Data Scraped (200 Products)

- **Title**
- **Description**
- **Image URL**

6

Text Classification preparation-



Text Classification

The 3-step process would be cleaning

Text Pre-Processing

Cleaning and processing the Text columns to remove unnecessary words

Amazon Dataset

Data Cleaning & Merging

- Removing Non alpha numeric characters and emojis
- Tokenization
- Removal of stopwords
- Stemming
- Lemmatization

Training & Validation

We tried different classification models

Craigslist Data : Out of Sample Validation

Model Building

- Combined Text and Description Columns
- Ran multiple classifier models
- TF-IDF Vectorization

Text Classification Models : Random Forests

Classification Models

Predicting product class using different models

Random Forests
n_estimators= 100
Random_Seed= 42
Min_samples_leaf= 1
Min_samples_Split= 2

Test Accuracy: 88%

LSTM
Activation =soft_max
Input_dim= 5000
Neurons: 100

Test Accuracy: 90%

8

Leveraging the comprehensive Amazon Dataset, we embark on a rigorous cleaning process to refine the textual columns, ensuring the elimination of superfluous elements that could potentially distort our results. This process encompasses several sophisticated steps:

- **Data Cleaning:** We start by purging all non-alphanumeric characters, including punctuation and symbols. This also involves the removal of emojis and any graphical characters that do not contribute to textual meaning, thus homogenizing the data into a purely textual format.

- **Tokenization**: Following sanitization, we segment the text into its atomic elements, known as tokens. This granular breakdown involves splitting the text into individual words or terms, which serves as the foundational step for more advanced text analysis techniques.

- **Stopwords Elimination**: Our process then identifies and discards stopwords. These are typically common words such as 'the', 'is', and 'at', which are pervasive in language but generally carry negligible analytical weight for our purposes. Removing these allows us to concentrate on more impactful terms.

- **Stemming**: To further distill the text, we apply stemming algorithms which truncate words to their root forms. This reductionist approach often involves cutting off derivational affixes, enabling the grouping of different forms of the same word, thereby reducing complexity and variability within the text data.

- **Lemmatization**: As a more advanced alternative to stemming, lemmatization considers the morphological analysis of the words. By understanding the context and discerning the specific part of speech, lemmatization ensures that words are reduced to their canonical or 'dictionary' form (lemma). This not only aids in achieving a more accurate representation of the language but also helps in maintaining the semantic integrity of the text.

- **TF-IDF Vectorization**: With the text cleaned and normalized, we proceed to transform it into a numerical format interpretable by machine learning classifiers. This transformation is achieved through TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. TF-IDF is a statistical measure that evaluates the importance of a word in a document relative to a collection of documents (corpus). It enhances the value of less frequent, more topical terms, while diminishing the impact of recurrent but less significant words.

Image Classification preparation-

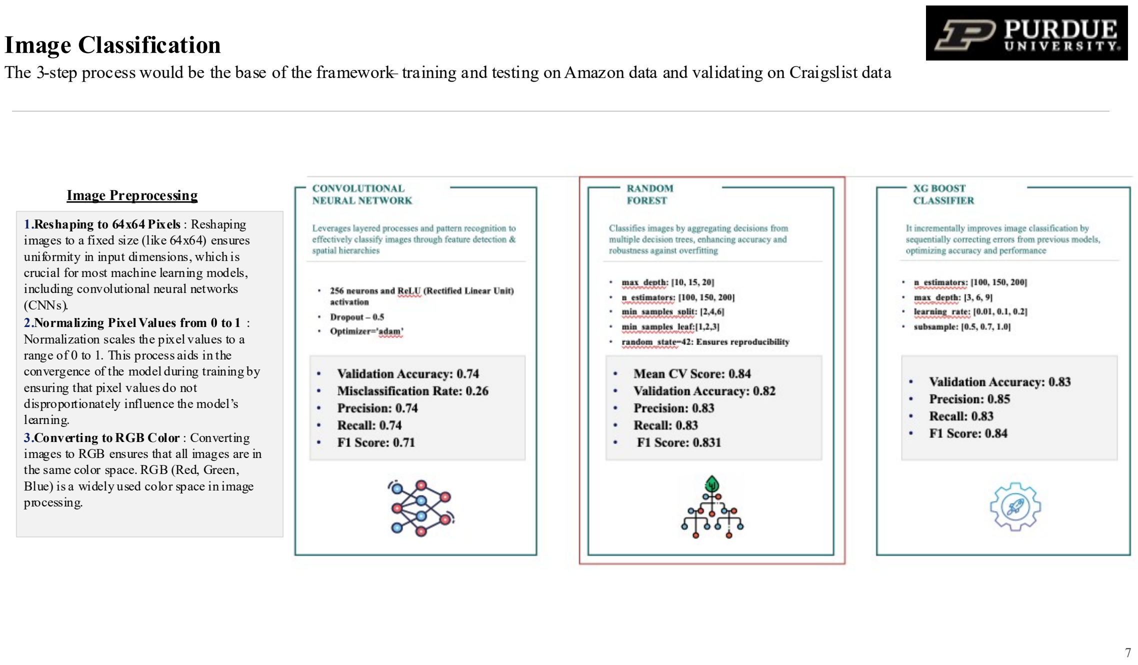


Image Classification

The 3-step process would be the base of the framework- training and testing on Amazon data and validating on Craigslist data

Image Preprocessing

1.**Reshaping to 64x64 Pixels** : Reshaping images to a fixed size (like 64x64) ensures uniformity in input dimensions, which is crucial for most machine learning models, including convolutional neural networks (CNNs).
2.**Normalizing Pixel Values from 0 to 1** : Normalization scales the pixel values to a range of 0 to 1. This process aids in the convergence of the model during training by ensuring that pixel values do not disproportionately influence the model's learning.
3.**Converting to RGB Color** : Converting images to RGB ensures that all images are in the same color space. RGB (Red, Green, Blue) is a widely used color space in image processing.

CONVOLUTIONAL NEURAL NETWORK

Leverages layered processes and pattern recognition to effectively classify images through feature detection & spatial hierarchies

- **256 neurons and ReLU (Rectified Linear Unit) activation**
- **Dropout – 0.5**
- **Optimizer='adam'**

- **Validation Accuracy: 0.74**
- **Misclassification Rate: 0.26**
- **Precision: 0.74**
- **Recall: 0.74**
- **F1 Score: 0.71**

RANDOM FOREST

Classifies images by aggregating decisions from multiple decision trees, enhancing accuracy and robustness against overfitting

- **max_depth: [10, 15, 20]**
- **n_estimators: [100, 150, 200]**
- **min_samples_split: [2,4,6]**
- **min_samples_leaf:[1,2,3]**
- **random_state=42: Ensures reproducibility**

- **Mean CV Score: 0.84**
- **Validation Accuracy: 0.82**
- **Precision: 0.83**
- **Recall: 0.83**
- **F1 Score: 0.831**

XG BOOST CLASSIFIER

It incrementally improves image classification by sequentially correcting errors from previous models, optimizing accuracy and performance

- **n_estimators: [100, 150, 200]**
- **max_depth: [3, 6, 9]**
- **learning_rate: [0.01, 0.1, 0.2]**
- **subsample: [0.5, 0.7, 1.0]**

- **Validation Accuracy: 0.83**
- **Precision: 0.85**
- **Recall: 0.83**
- **F1 Score: 0.84**

7

□ **Reshaping to 64x64 Pixels:** Purpose: Reshaping images to a fixed size (like 64x64) ensures uniformity in input dimensions, which is crucial for most machine learning models, including convolutional neural networks (CNNs).

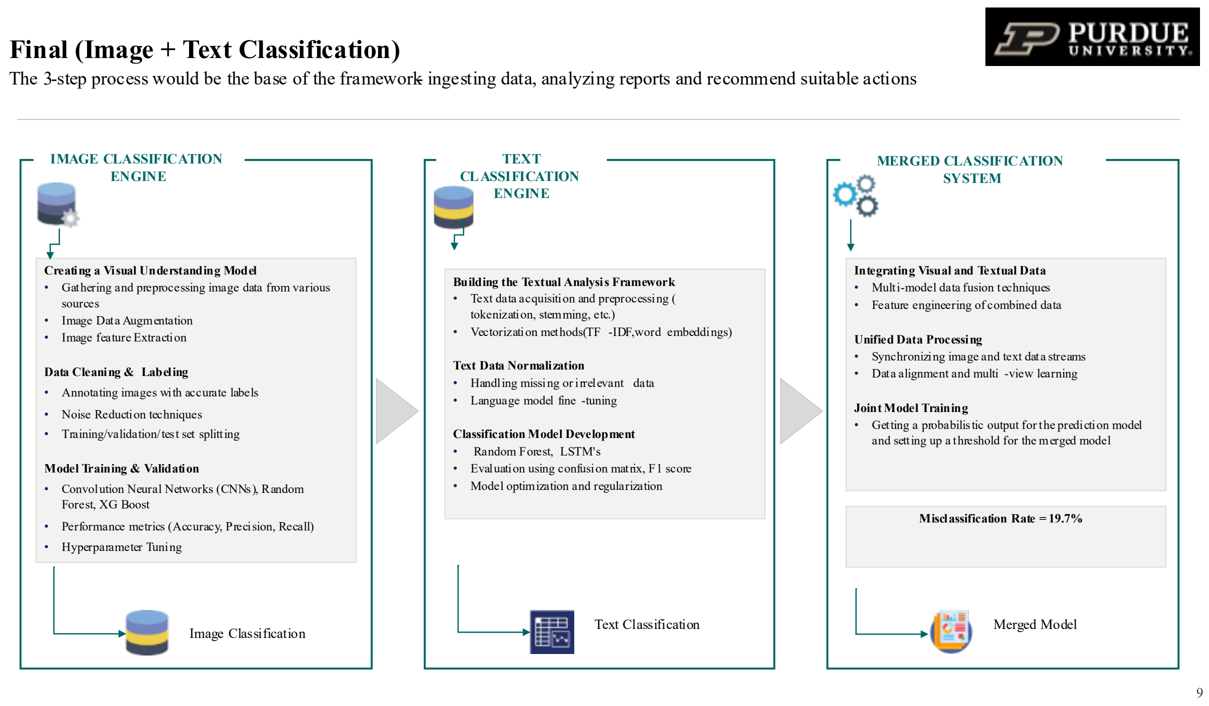
□ **Purpose:** Normalization scales the pixel values to a range of 0 to 1. This process aids in the convergence of the model during training by ensuring that pixel values do not disproportionately influence the model's learning.

□ **Uniform Color Space:** Converting images to RGB ensures that all images are in the same color space. RGB (Red, Green, Blue) is a widely used color space in image processing.

□ **VGG16 for CNN:** An architecture that is composed of 16 layers that have weights. This includes 13 convolutional layers and 3 fully connected layers at the end. It uses small 3×3 convolutional filters throughout the network, which was a distinctive feature at the time of its introduction. The network also includes several max pooling layers interspersed between the convolutional layers.

□ **ReLU Activation Function for CNN:**The Rectified Linear Unit (ReLU) activation function is used throughout the network. This choice helps in alleviating the vanishing gradient problem, allowing for faster training

Image classification has been used to classify craigslist products across computer and computer accessory categories. The model has been trained on images obtained for similar categories from Amazon and has been tested on Craigslist data.
The models used for classification are Random Forest, XGBoost and Convolutional Neural Network with Random Forest performing the best. The validation metrics used is Accuracy.



Final (Image + Text Classification)

The 3-step process would be the base of the framework ingesting data, analyzing reports and recommend suitable actions

IMAGE CLASSIFICATION ENGINE

Creating a Visual Understanding Model
- Gathering and preprocessing image data from various sources
- Image Data Augmentation
- Image feature Extraction

Data Cleaning & Labeling
- Annotating images with accurate labels
- Noise Reduction techniques
- Training/validation/test set splitting

Model Training & Validation
- Convolution Neural Networks (CNNs), Random Forest, XG Boost
- Performance metrics (Accuracy, Precision, Recall)
- Hyperparameter Tuning

Image Classification

TEXT CLASSIFICATION ENGINE

Building the Textual Analysis Framework
- Text data acquisition and preprocessing (tokenization, stemming, etc.)
- Vectorization methods(TF -IDF,word embeddings)

Text Data Normalization
- Handling missing or irrelevant data
- Language model fine -tuning

Classification Model Development
- Random Forest, LSTM's
- Evaluation using confusion matrix, F1 score
- Model optimization and regularization

Text Classification

MERGED CLASSIFICATION SYSTEM

Integrating Visual and Textual Data
- Multi-model data fusion techniques
- Feature engineering of combined data

Unified Data Processing
- Synchronizing image and text data streams
- Data alignment and multi -view learning

Joint Model Training
- Getting a probabilistic output for the prediction model and setting up a threshold for the merged model

Misclassification Rate = 19.7%

Merged Model

9

Training and Validation

After completing the pre-processing of our data, we transitioned into the critical phase of training and validation. We experimented with a variety of classification algorithms to ascertain the most compatible model for our dataset. In our pursuit of robustness, we employed a holdout validation method using Craigslist data. This approach is particularly effective as it subjects our models to a rigorous test — evaluating their performance on completely unseen data, thus providing us with a realistic gauge of their predictive capabilities.

MODEL VALIDATION

Image Classification Model Performance Summary

- The **Convolutional Neural Network (CNN)** features 256 neurons and ReLU activation, with a 0.5 dropout and 'adam' optimizer, achieving a 0.74 validation accuracy, precision, recall, and a 0.71 F1 score.
- **Random Forest** uses various max_depth [18, 15, 20], n_estimators [100, 150, 200], and min_samples_split [2,4,6], achieving better results than CNN with a mean CV score of 0.84 and F1 score of 0.831.
- **XG Boost** stands out with the best metrics, using n_estimators [100, 150, 200] and max_depth [3, 6, 9], reaching a validation accuracy of 0.83, precision of 0.85, and the highest F1 score of 0.84.

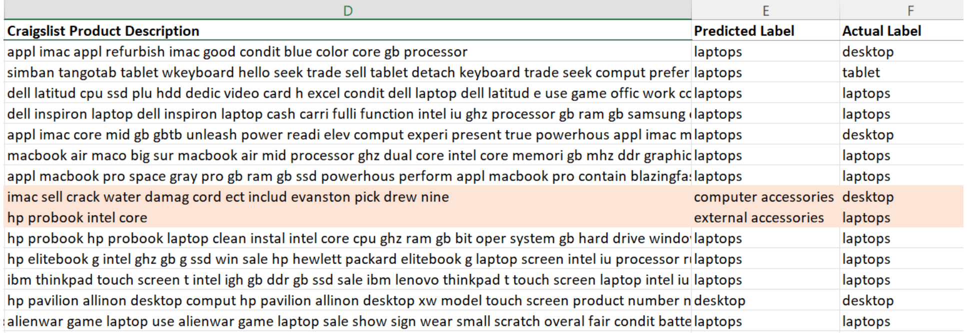
Text Classification Model Performance Summary

- The **Random Forest** model, using 100 estimators and a fixed random seed of 42, achieved an 81.01% holdout accuracy and 89% test accuracy, proving its effectiveness in classifying the Craigslist dataset.
- **Logistic Regression** served as a solid baseline with a 75.95% holdout accuracy. Multinomial Naive Bayes showed proficiency in text data classification with a 77.22% holdout accuracy.
- **Support Vector Machine (SVM)** underperformed significantly, with just 29.11% holdout accuracy, indicating a poor fit for this data.
- An **LSTM** Network, despite a high 90% test accuracy, was less effective in out-of-sample prediction, likely due to limited data size and the nature of the data. Ultimately, Random Forest stood out for its robust validation performance and was chosen for our text classification needs.

Final Classification

- Random Forest was the selected model for both text and image classification as it had the best performance on holdout (Craigslist) data.
- If the image and text classification gave different classes, the probabilities of image and text were compared for the final classification and the one with higher probability was assigned as the final class
- We were able to achieve a misclassification rate of 19.7% on Craiglist test data, significantly lower than the websites ~ (30- 40%)

An example of the Predictions using the Combined model is as follows



D	E	F
Craigslist Product Description	**Predicted Label**	**Actual Label**
appl imac appl refurbish imac good condit blue color core gb processor	laptops	desktop
simban tangotab tablet wkeyboard hello seek trade sell tablet detach keyboard trade seek comput prefer	laptops	tablet
dell latitud cpu ssd plu hdd dedic video card h excel condit dell laptop dell latitud e use game offic work cc	laptops	laptops
dell inspiron laptop dell inspiron laptop cash carri fulli function intel iu ghz processor gb ram gb samsung	laptops	laptops
appl imac core mid gb gbtb unleash power readi elev comput experi present true powerhous appl imac m	laptops	desktop
macbook air maco big sur macbook air mid processor ghz dual core intel core memori gb mhz ddr graphic	laptops	laptops
appl macbook pro space gray pro gb ram gb ssd powerhous perform appl macbook pro contain blazingfa	laptops	laptops
imac sell crack water damag cord ect includ evanston pick drew nine	computer accessories	desktop
hp probook intel core	external accessories	laptops
hp probook hp probook laptop clean instal intel core cpu ghz ram gb bit oper system gb hard drive windo	laptops	laptops
hp elitebook g intel ghz gb g ssd win sale hp hewlett packard elitebook g laptop screen intel iu processor r	laptops	laptops
ibm thinkpad touch screen t intel igh gb ddr gb ssd sale ibm lenovo thinkpad t touch screen laptop intel iu	laptops	laptops
hp pavilion allinon desktop comput hp pavilion allinon desktop xw model touch screen product number n	desktop	desktop
alienwar game laptop use alienwar game laptop sale show sign wear small scratch overal fair condit batte	laptops	laptops

The example presented highlights the Combined model's capability in handling a wide range of classifications, adeptly identifying many instances correctly while also encountering some challenges with misclassifications and ambiguities. This performance reflects the model's substantial proficiency in interpreting complex text, yet subtly suggests areas for nuanced improvements, particularly in addressing the subtleties of highly ambiguous or less clear content.

CONCLUSION

Our comprehensive analysis, underpinned by an innovative AI-based categorization and tagging system, is poised to revolutionize Craigslist's approach to classified advertisements, particularly in the computer and computer parts section. The intelligent system, leveraging advanced NLP and machine learning algorithms, promises a significant improvement in the accuracy of listings, directly addressing the prevalent issue of misclassification.

This enhancement in categorization not only resolves the immediate problem of finding relevant listings but also streamlines the user journey from listing to purchase. It significantly improves user experience, making the platform more intuitive and user-friendly. Additionally, the image classification model, trained and validated using robust datasets, ensures that computer-related products are categorized with high precision.

For Craigslist, this translates into increased user engagement and potential for monetization. The precision in categorization and tagging opens new avenues for targeted advertising and premium listing features, thereby increasing revenue opportunities. Furthermore, by reducing the manual effort required in managing listings and ensuring higher accuracy, our solution effectively decreases operational costs.

The real-world impact of our analysis and the implemented system extends beyond Craigslist. By setting a new standard in classified ad categorization, we influence the broader online advertising ecosystem, promoting a shift towards intelligent, data-driven approaches. This not only enhances the efficiency of online marketplaces but also elevates the overall quality of digital advertising.

In conclusion, our project offers substantial value to Craigslist by resolving key operational challenges, enhancing user experience, and opening new paths for revenue generation. This strategic adaptation aligns with Craigslist's foundational strengths and positions it for sustained growth and competitiveness in the evolving digital marketplace.

References:
https://businessmodelanalyst.com/craigslist-business model/#Craigslist_Value_Propositions