ServiScan:

Making Craigslist’s “Services Offered” Section Smarter

Background & Motivation

Craigslist: A Trusted but Unstructured Marketplace

- Craigslist is a long-standing, user-driven platform for classified ads, serving millions of users across hundreds of categories.

- The **“Services Offered”** section is one of the most active parts of the platform, hosting listings for everything from tutoring to handyman services.

The Challenge

- While popular, this section suffers from:
  - Lack of structure in user-submitted text
  - Vague or duplicate listings
  - Incorrect category placements
- Users face difficulty finding relevant, trustworthy services.
- Craigslist's moderation team struggles with scale and spam.

Real-World Examples

- A dog walker lists under “computer services”
- A cleaning ad reposted across multiple cities
- Posts without pricing, availability, or contact info

The Problem

What’s Broken in “Services Offered”

- **Misclassification**
  Ads posted in incorrect subcategories reduce discoverability.

- **Messy, Unstructured Text**
  Long, unclear descriptions make browsing difficult.

- **Spam & Duplication**
  Low-effort or repeated posts clutter search results.

- **Lack of Filtering**
  No automatic way to highlight key details like price or service type.

Why It Matters

- Poor **user experience** (UX): hard to find what you need

- Declining **trust** in platform due to spam/inconsistencies

- Increased **moderation burden** on Craigslist’s internal team

Our Solution – *ServiScan*

Introducing Service Scan A lightweight, modular, AI-powered backend tool designed to make Craigslist's "Services Offered" section smarter — without changing how users post.

Key Objectives

- **Clean Listings**
  Automatically summarize long or vague descriptions.

- **Structure Free Text**
  Extract essential fields: service type, location, price, contact info.

- **Assist Moderation**
  Flag misclassified, duplicated, or suspicious content.

- **Preserve Craigslist's Simplicity**
  All enhancements happen behind the scenes — no added friction for users.

Bottom Line:
ServiScan enhances user experience and data quality while reducing internal moderation load — all without disrupting Craigslist's core design philosophy.

Data Collection & Manual Labeling

Targeted Categories: Focused on “Beauty” and “Health/Wellness” ads in Craigslist NYC.

Scraping: Used Scrapy to collect titles, price, and location of service ads.

Manual Labeling: Each ad was labeled into one of five subcategories:

- Hair-Styling
- Body-Work
- Health-Aid
- Fitness
- Others

Motivation: Body massage ads often blur lines across categories — finer distinctions reduce ambiguity.

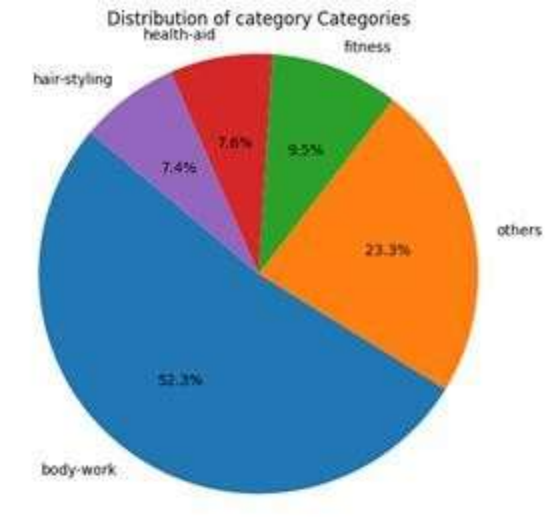
Exploratory Data Analysis (EDA)

Category Distribution:

- Dominated by “Body-Work” and “Others” categories.
- “Hair-Styling” and “Fitness” are relatively underrepresented.

Observations:

- Class imbalance evident → Considered during model evaluation.
- Initial EDA helped shape preprocessing and model expectations.



Preprocessing data

Preprocessing Pipeline

- **Text Cleaning:** Removed emojis, hyperlinks, punctuation, special characters.
- **Tokenization & Lemmatization:** Performed using NLTK.
- **Stopword Removal:** Enhanced signal-to-noise ratio.
- **Vectorization:** Used **TF-IDF** to represent text as numerical features.

Model Training & Selection

Train/Test Split: 80/20 ratio

Models Tried:

- Logistic Regression
- Random Forest
- XGBoost

Label Encoding: Subcategories converted to numerical values.

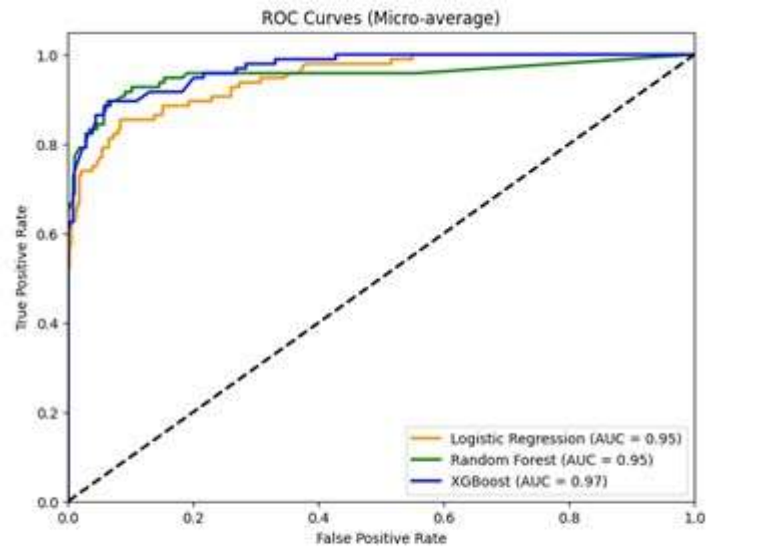
Model Evaluation & Performance

Metrics used:

- Accuracy
- F1 Score
- Precision & Recall
- ROC AUC

Best Model: XGBoost

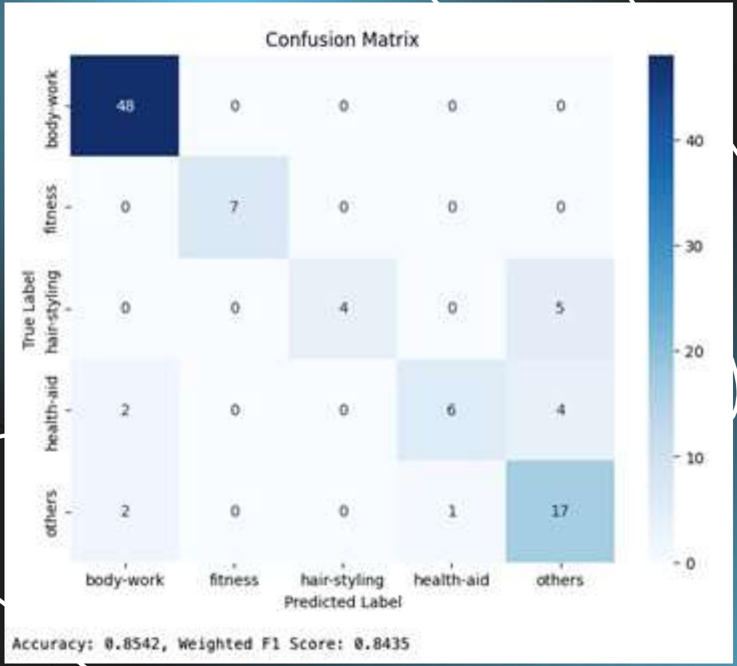
- Highest F1 score & AUC
- Handled imbalance better than others



Insights from Confusion Matrix:

- Misclassifications between Health-Aid and Body-Work
- Hair-Styling occasionally confused with Others

Overfitting Noted: Random Forest performed better on training set than test set

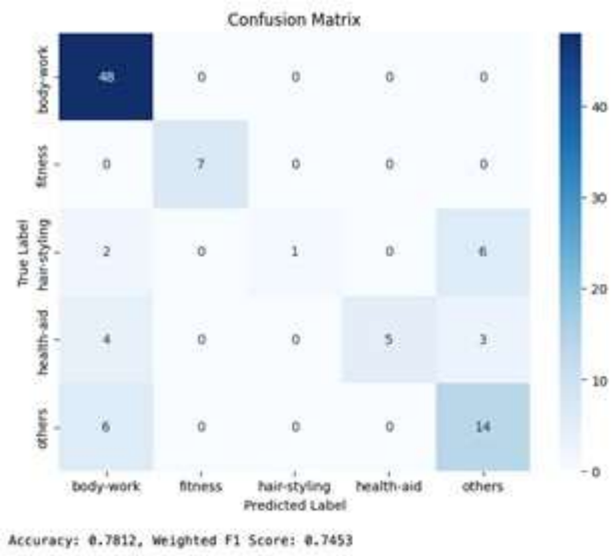


True Label \ Predicted Label	body-work	fitness	hair-styling	health-aid	others
body-work	48	0	0	0	0
fitness	0	7	0	0	0
hair-styling	0	0	4	0	5
health-aid	2	0	0	6	4
others	2	0	0	1	17

```
Accuracy: 0.8542, Weighted F1 Score: 0.8435
```

XGBoost

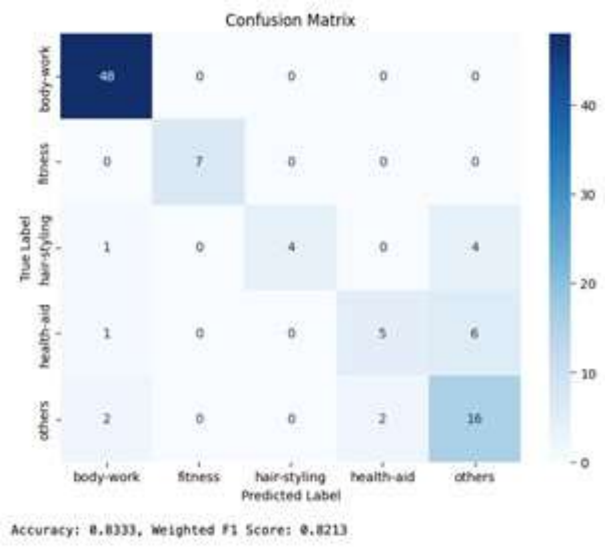
Confusion matrices on test data



	body-work	fitness	hair-styling	health-aid	others
body-work	48	0	0	0	0
fitness	0	7	0	0	0
hair-styling	2	0	1	0	6
health-aid	4	0	0	5	3
others	6	0	0	0	14

```
Accuracy: 0.7812, Weighted F1 Score: 0.7453
```

Logistic Regression

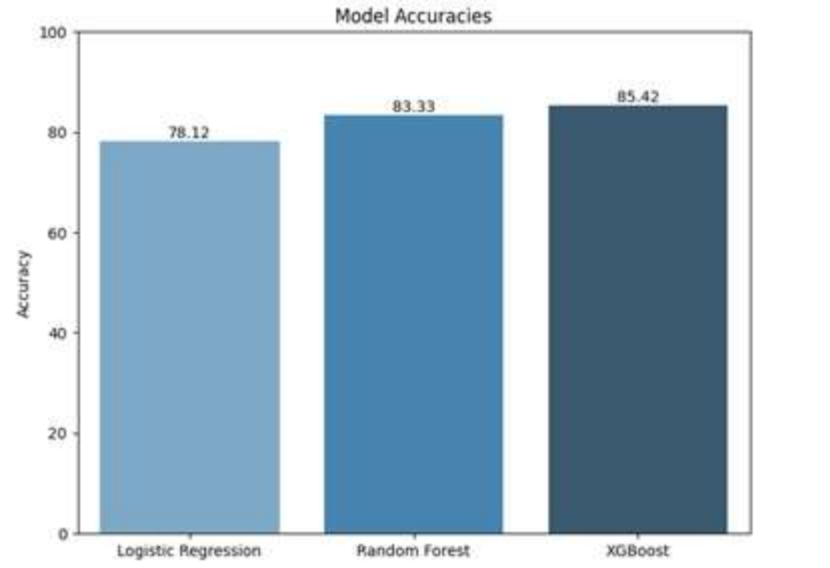


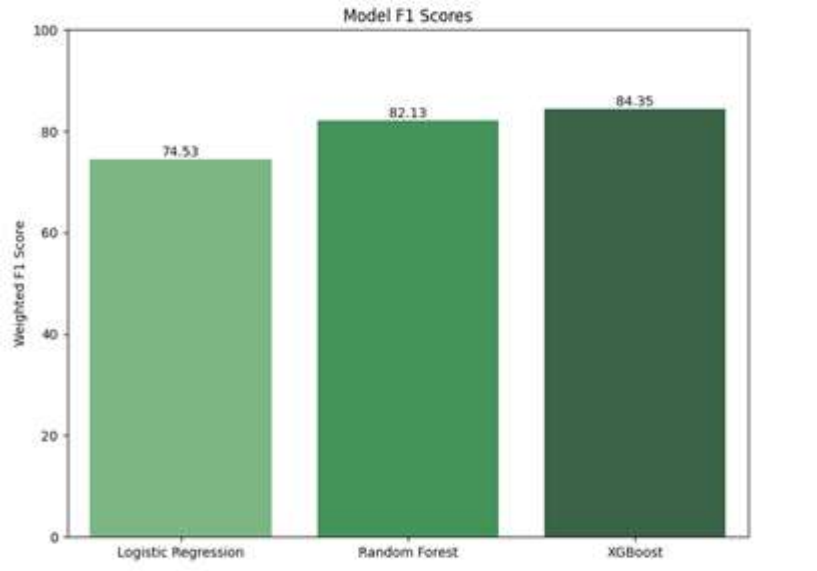
	body-work	fitness	hair-styling	health-aid	others
body-work	48	0	0	0	0
fitness	0	7	0	0	0
hair-styling	1	0	4	0	4
health-aid	1	0	0	5	6
others	2	0	0	2	16

```
Accuracy: 0.8333, Weighted F1 Score: 0.8213
```

Random Forest

Comparisons





Key Insights from Analysis

1. Craigslist's Format Causes Ambiguity
Its text-heavy, unstructured format leads to overlapping or misplaced listings — especially in service categories like "Beauty" and "Health/Wellness."

2. Manual Subcategories Improve Clarity
Finer labels (e.g., Hair-Styling, Body-Work, Health-Aid) enable better classification and reduce confusion across overlapping categories.

3. Ad Titles Alone Hold Predictive Power
Even without full post content, TF-IDF + XGBoost on titles alone yields strong classification results — proving that lightweight solutions can work well.

4. Data Imbalance Reflects Market Reality
A high proportion of Body-Work ads isn't just noise — it reflects actual user trends, offering operational insight for platform improvement.

5. XGBoost Stands Out
Among all models tested, XGBoost consistently performed best on both F1 score and AUC, showcasing its robustness for sparse, noisy classification tasks.

Strategic Takeaways for Craigslist

1. Lightweight AI = High Impact
With minimal changes to posting flow, Craigslist can deploy backend AI (like ServiScan) to categorize and clean listings automatically.

2. Prioritize Moderation at Confusion Points
The categories most frequently confused (e.g., Health-Aid vs. Others) point directly to where moderation resources can be most effective.

3. Market Patterns Should Drive Tooling
Understanding which categories dominate helps design smarter tools — not just for moderation, but also for recommendation and filtering.

4. Builds Foundation for ServiScan
This experiment validates the core feasibility of ServiScan as a backend system:
→ Accurate classification
→ Spam/misclassification detection
→ Preserved user simplicity

Thank you!