MISCLASSIFICATION OF JOB POSTINGS

craigslist

BAIM AUDitors

Chethan Manjunath
Rohan Walyat
Yamini Praneetha Ambati
Yash Kulshreshtha
Yi chen Chu

craigslist

Meet the Team

Chetan Manjunath

Rohan Walyat

Yamini Ambati

Yash Kulshreshtha

Yi Chen Chu

craigslist

AGENDA

1. Project Background
2. Business Analysis
3. Data Analysis
4. Validation
5. Conclusion

CRAIGSLIST

craigslist

Classified Advertisemenets website with diverse sections dedicated to housing, for sale, jobs, community etc.

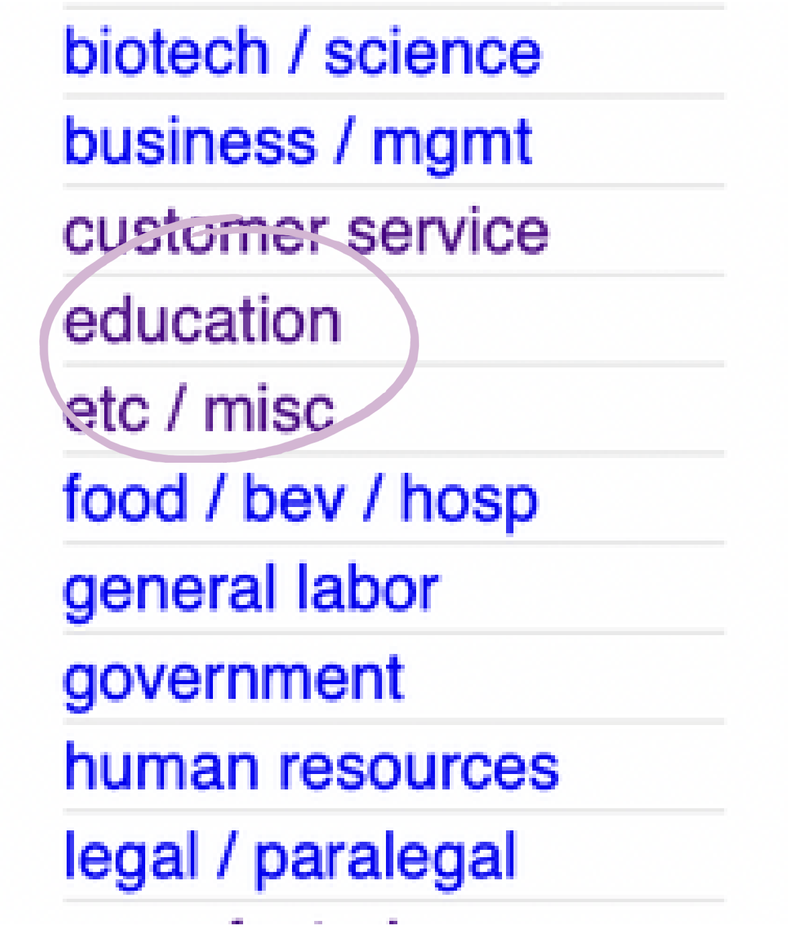
Present in 70+ countries & Millions of advertisements each day

Absence of any monitoring system to automatically label the posts

More labor intensive to search through thousands of posts before you reach the correct one

craigslist

CRAIGSLIST



Why Jobs?

- Possibility of Recession in the future
- Limited resource availability for such lower wage jobs relatively
- Absence of multiple authorized resources

craigslist

Project Objective

To improve the quality of Craigslist job section, in the education category by flagging misclassified posts

craigslist

Project Scope

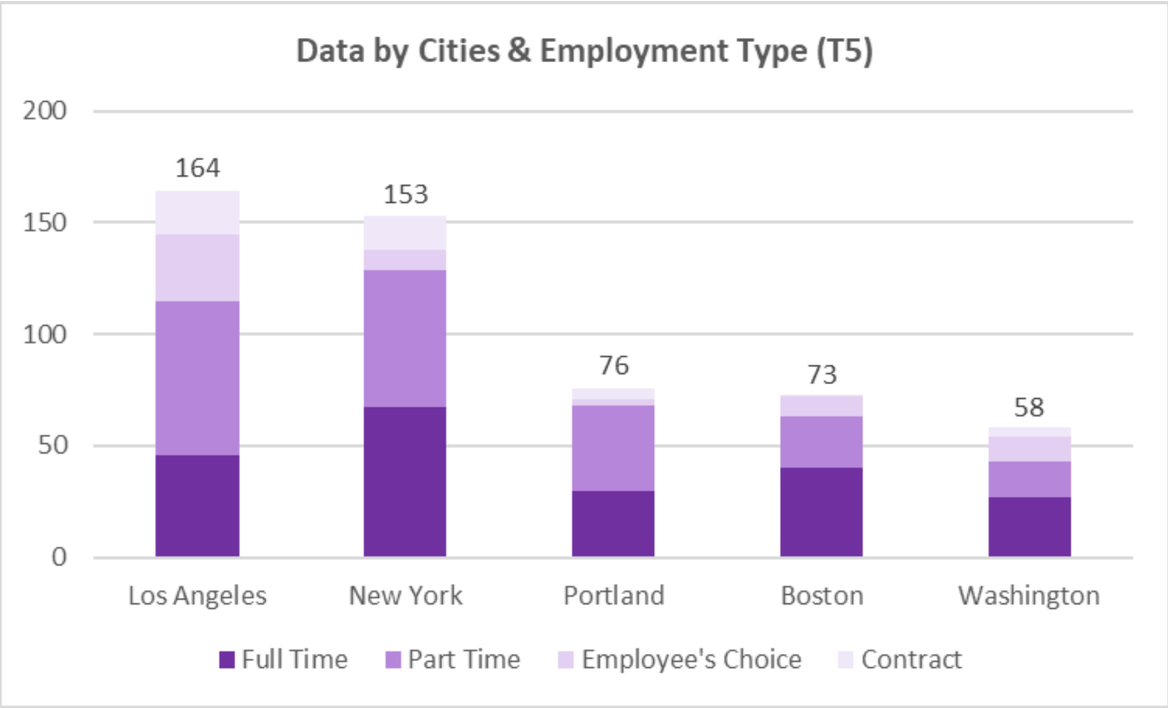
1. Collecting & cleaning the training & test dataset
2. Training model to accurately identify misclassified postings
3. Evaluating the performance of the model

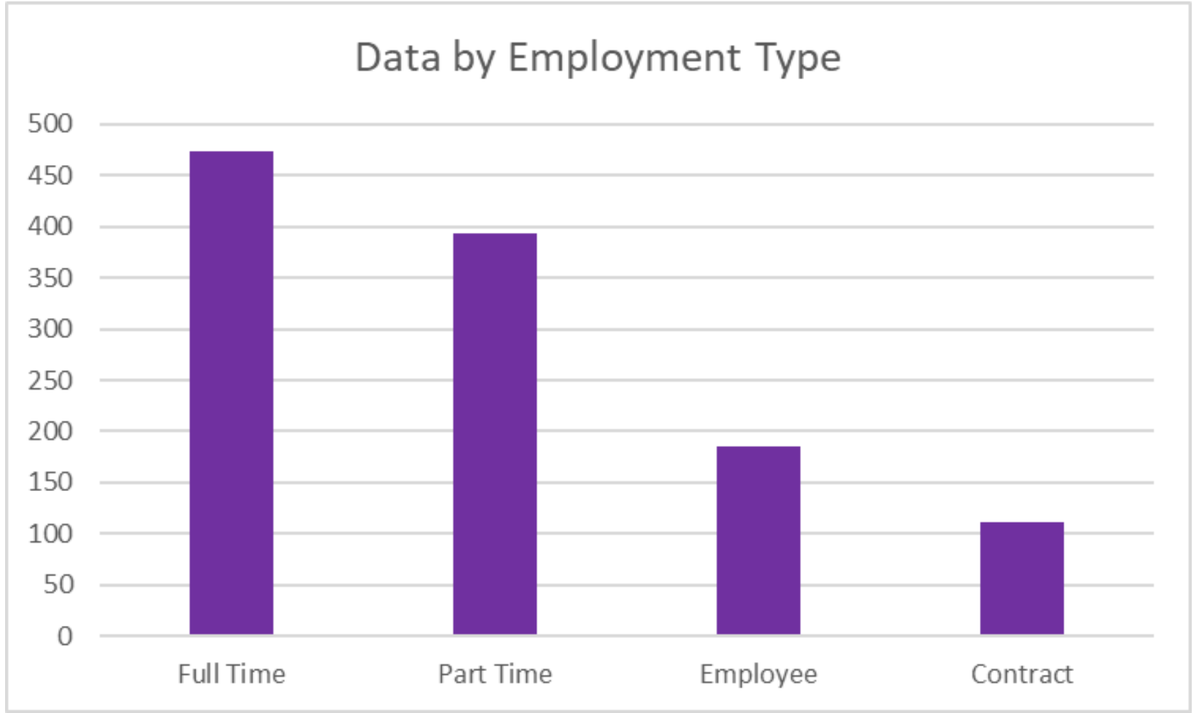
craigslist

BUSINESS ANALYSIS

craigslist

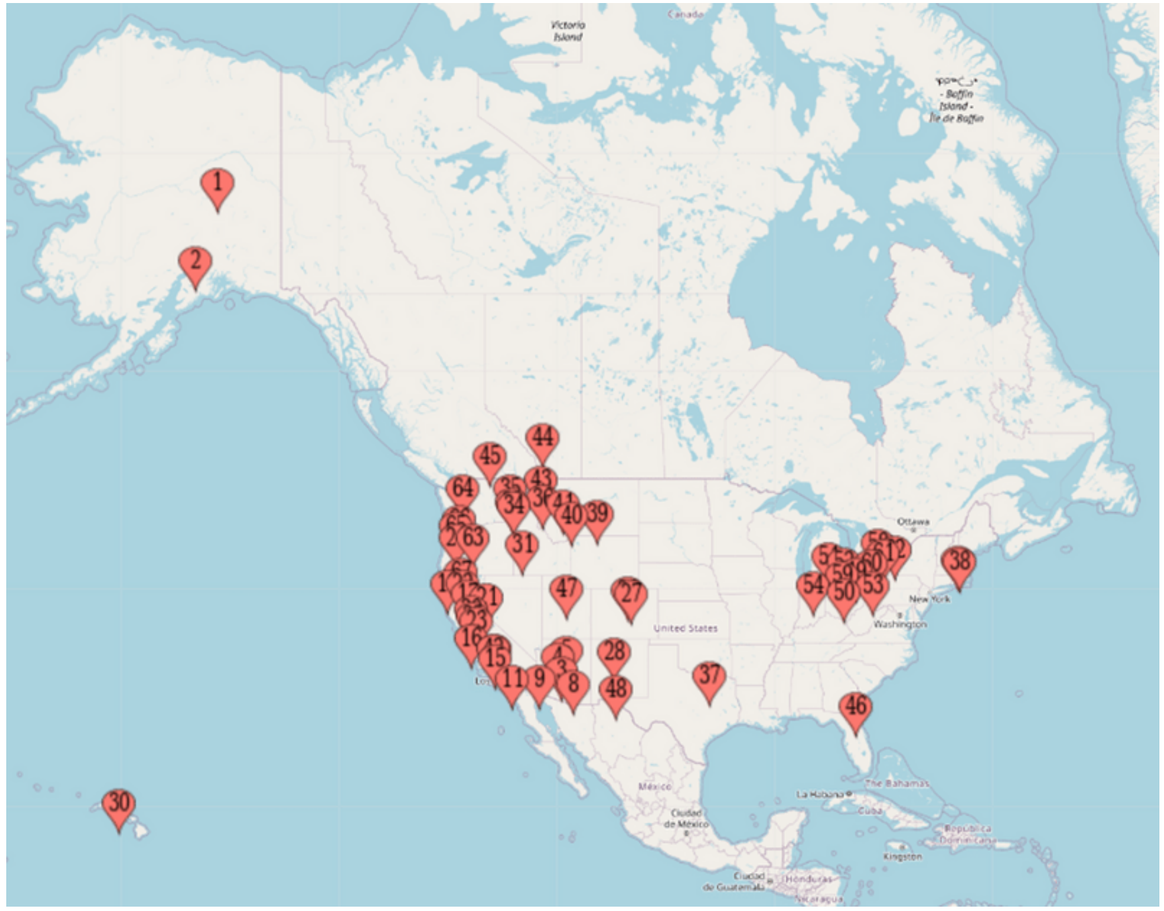
EDA





craigslist

DATA BY LOCATION



Ensured that data is representative of demographics of US

Explored data across locations and compensation

craigslist

Labeling the misclassified posts

1

Hi we are currently seeking passionate **teachers** with positive attitudes to be part of our loving team Ideal candidate will have Must have or be in process of obtaining Early Childcare **Education** ECE or Child Development Associate Degree Experience working with infant and toddlers Sheriff Card and CPR Would Certainly Be A Plus If you meet these requirements please reply with cover letter and **resume** to this ad 702 428 7471

CORRECT POST

0

I am the founder and CEO of a **medical** device company and I could use a personal assistant to help create **marketing videos** and communications and sale accounting help Must be great with people willing to learn and precise neat and able 10 15 hours per week can do some remotely 1 2 Need a young enthusiastic and calm person

MISCLASSIFIED POST

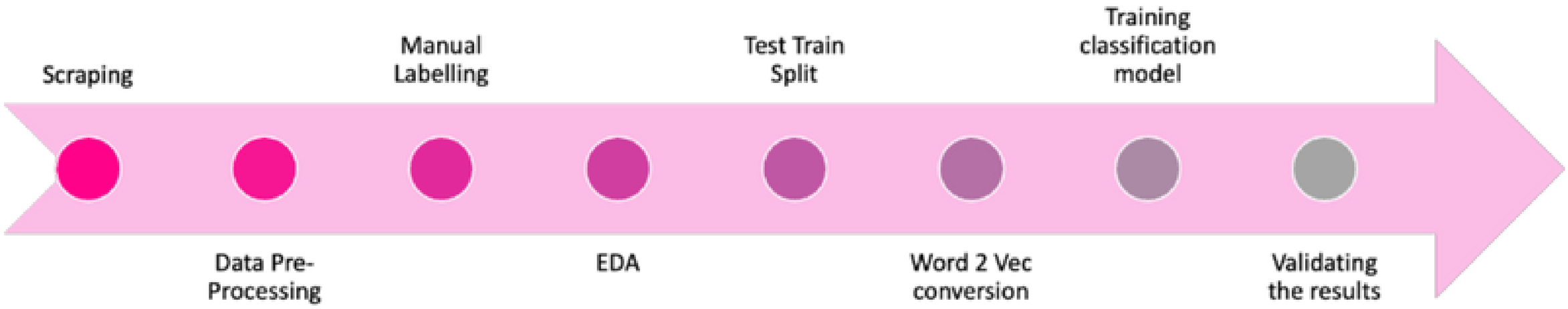
*both posts scraped from Jobs:Education

craigslist

DATA ANALYSIS

craigslist

Methodology



Scraping

craigslist

Preschool Teacher (Bucktown)

Full-time: Monday - Friday

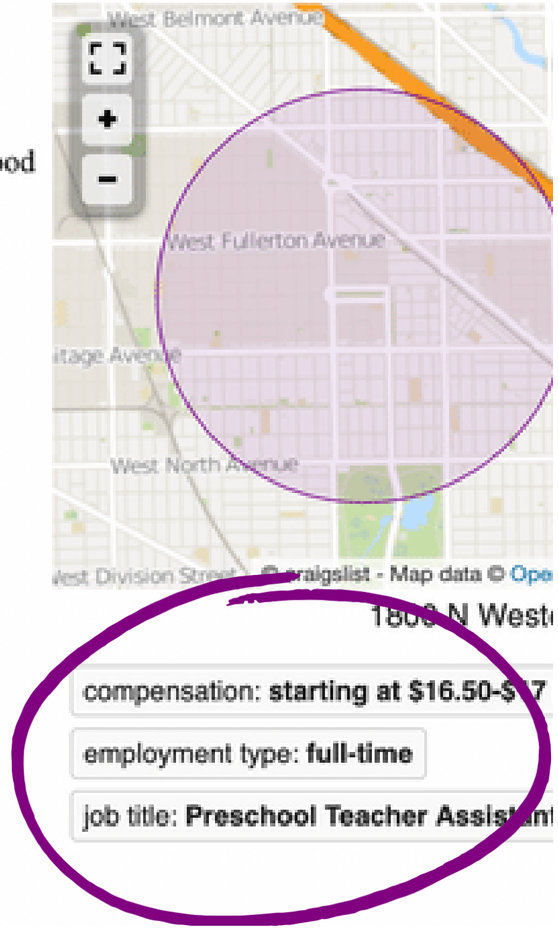
Looking to hire:

Minimum of 9 ECE credits. Preferred Candidates have a CDA or Associates degree in Early Childhood Education or related field.

Must clear background and fingerprint check, CPR certified (can be done upon hire), flexible availability, have reliable transportation, and 3 references are required.

Job duties will include but are not limited to:

Class management
Lesson planning
Attendance for center events
Willing to learn and follow the Creative Curriculum. Reggio Amelia Approach, and STEAM
Cleaning and sanitizing classrooms.
supervision of children
Establish and enforce rules of behavior for children in their classrooms
Provide basic needs for children
Provide tools and resources for children to use and explore during learning and play activities
Adapt teaching methods and materials to meet the interests and learning styles of children
Develop and maintain positive relationships with children and parents
Manage classroom activities including lessons, play, breaks and meals
Great verbal and written communication skills
Paid holidays and sick/personal/vacation days when you pass our probationary period.



HTML Parsers

BeautifulSoup
Selenium

Job Title

Description

Compensation

Employment Type

Data Preprocessing

REMOVING PUNCTUATIONS

words not in string.punctuation

TOKENIZATION

nltk.word_tokenize()

LEMMATIZATION

nltk.stem.WordNetLemmatizer()

REMOVING STOPWORDS

nltk.corpus.stopwords

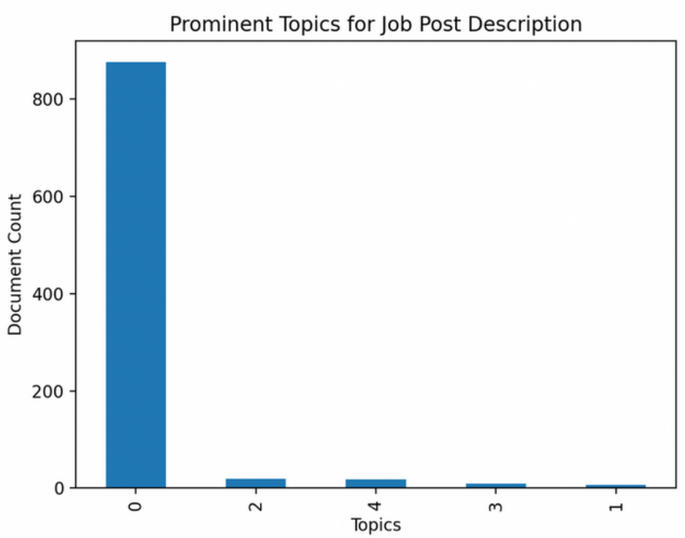
craigslist

Topic Modeling

Latent Dirichlet Allocation

Job Description

Topic 0	child student school teacher time must experience work
Topic 1	instrument lesson home song beginner teacher teach http lessonsinyourhome teacher employment lessonsinyourhome net
Topic 2	language contractor proficiency government foreign license school contractor state state license
Topic 3	camper head start camp ramapo ramapocamp com ramapocamp shasta head
Topic 4	dream youth harlem childpeace sport youth development east harlem based youth



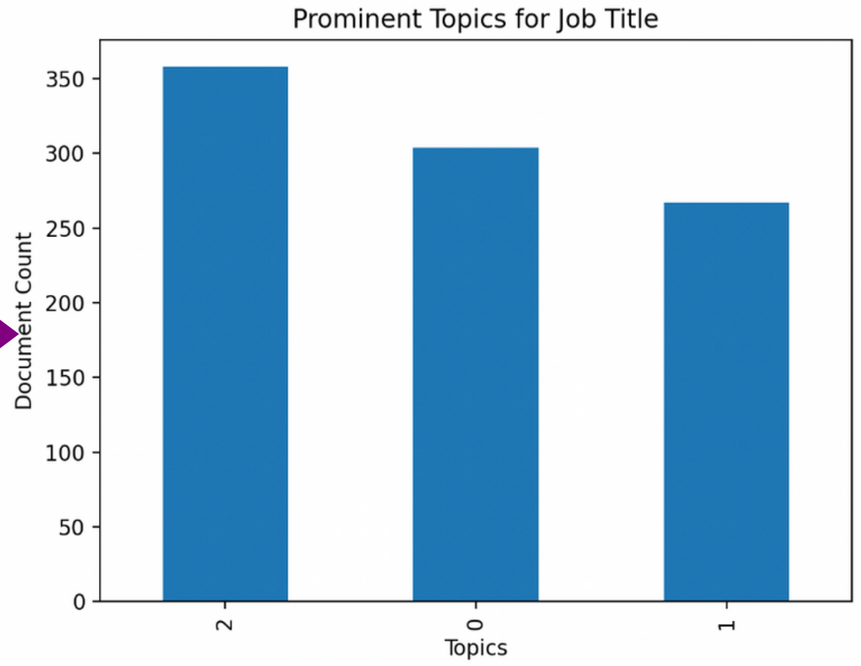
Significant posts: related to experienced school teachers

Topic Modeling

Latent Dirichlet Allocation

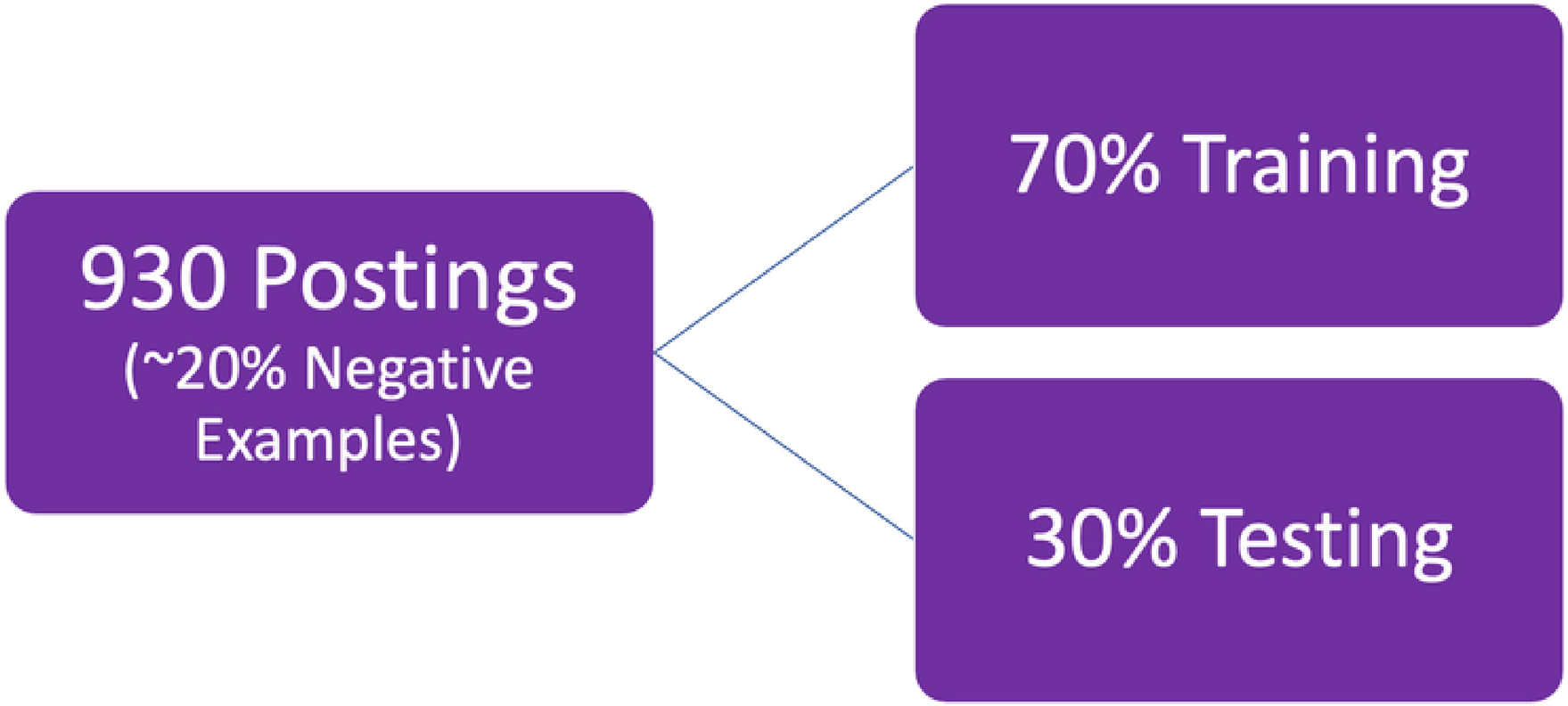
Job Titles

Topic 0	preschool teacher preschool teacher time aide part part time parttime
Topic 1	school education coach director hiring position special special education
Topic 2	teacher instructor assistant needed tutor lead teacher assistant daycare



Job Titles: Mostly standard for jobs that truly belong to the education category

Splitting the dataset



TF-IDF

sklearn.feature_extraction.text.TfidfVectorizer

Text Vectorization parameters

Title

ngram range: (1,2)

minimum document frequency: 3

Body

ngram range: (1,2)

minimum document frequency: 6

craigslist

Training the model

- Logistic Regression
- Support Vector Machine
- Gaussian Naïve-Bayes
- Multinomial Naïve-Bayes
- Random Forest
- XG Boost Classifier
- Multilayer Perceptron

craigslist

VALIDATION

Criteria - Accuracy Scores

craigslist

Model	Accuracy
Logistic Regression	86.3%
Support Vector Machine	84.9%
Multinomial Naïve Bayes	81%
Gaussian Naïve Bayes	85.6%
XG Boost Classifier	**87%**
Random Forest	83.15%
MLP Neural Network	86.7%

All models showed similar accuracy scores

XG Boost classifier had the highest score

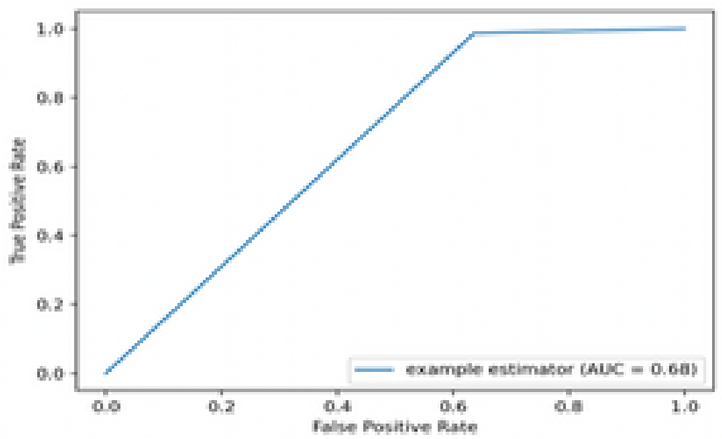
Hyper Parameter Tuning

Model	Grid Search Parameters
Logistic Regression	param ={'penalty': ['l1','**l2**'],'solver': ['newton-cg','lbfgs','**liblinear**']}
XG Boost Classifier	param = {'learning_rate': [0.001, **0.1,** 1, 10]}
Multi Layer Perceptron	param = { 'hidden_layer_sizes': [(**5,3**), (6,5), (10,3)], 'activation': ['tanh', '**relu**'], 'solver': ['sgd', '**adam**'], 'alpha': [0.001,0.05,**0.1**]}

craigslist

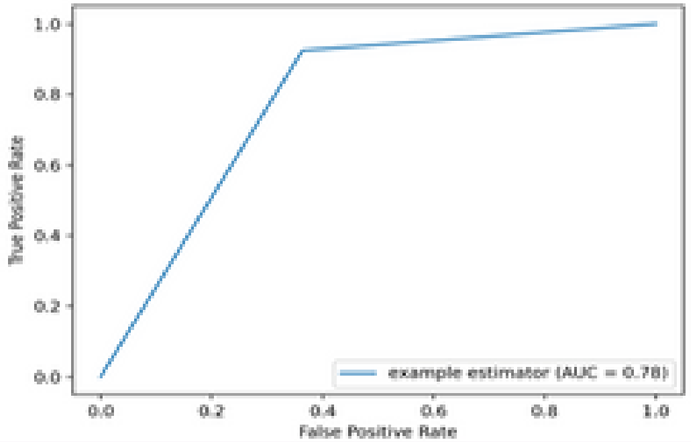
ROC Curve & Confusion Matrix

Logistic Regression



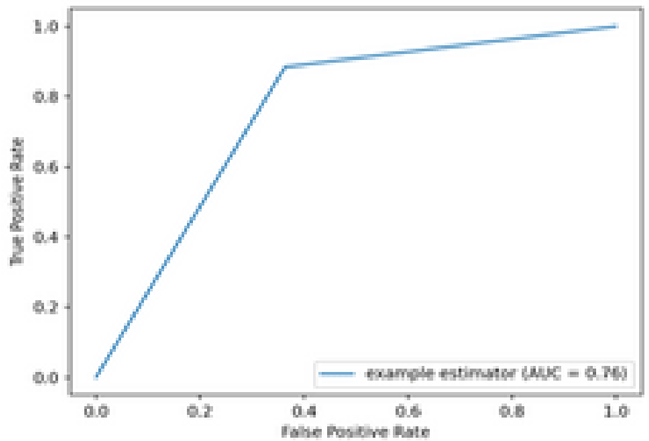
	Actual 0	Actual 1
Predicted 0	20	35
Predicted 1	3	221

XG Boost Classifier



	Actual 0	Actual 1
Predicted 0	35	20
Predicted 1	17	207

Multilayer Perceptron



	Actual 0	Actual 1
Predicted 0	35	20
Predicted 1	26	198

craigslist

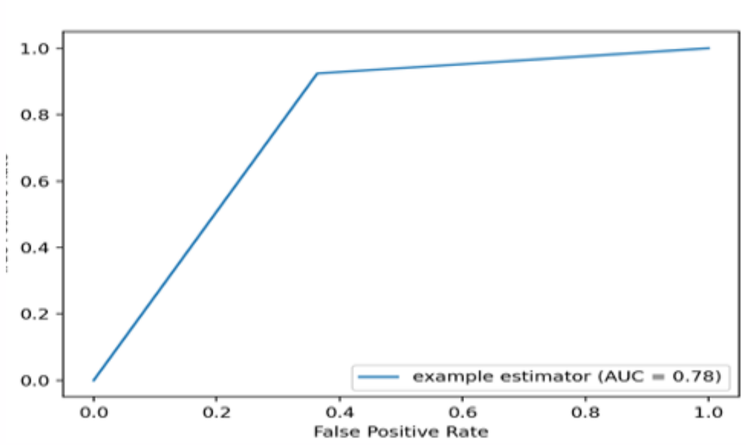
CONCLUSION

Conclusions

XG Boost Classifier gives the best results

Obtained **88%** accuracy

Area under the Roc curve: **0.78**



		Actual	
		0	1
Predicted	0	35	20
	1	17	207

Result

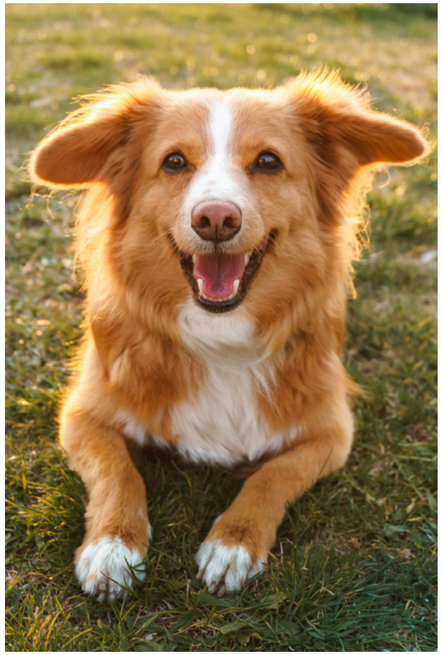
craigslist

favorite hide flag share

Training Kennel Technician at Guide Dogs for the Blind (san rafael)

Give your best for dogs who give their all.

Now Flagged as Misclassified



craigslist

Benefits for Craigslist

Building a better job portal for the education industry

Enhancing Craigslist's reputation

Increasing customer trust and satisfaction

Improving customer experience

Future work

Collection of more data

Utilizing spam repositories from local websites to capture more information

Building densely connected layers to capture rich information

craigslist

Thank you!

...and any questions?