CRAIGSLIST JOB CLASSIFICATION & SCAM DETECTION

MGMT 590 – AUD FINAL PROJECT

TEXNOMICS

ADITHYA KOTHARI

HROHAAN MALHOTRA

NAVEEN KUMAR

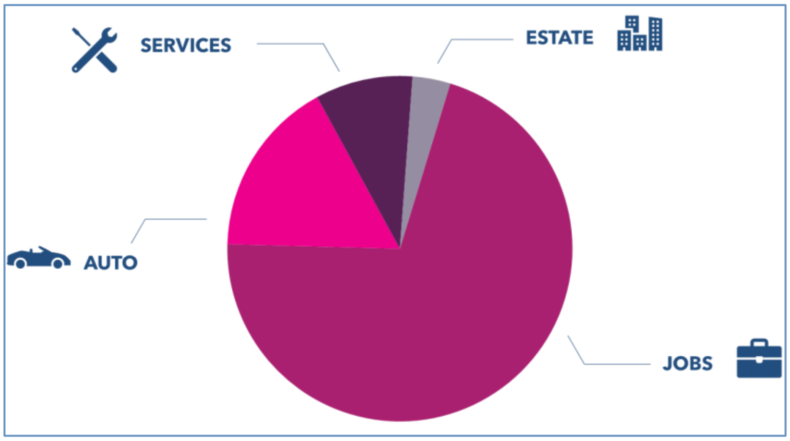
SHUBHAM GOYANKA

ABOUT

Craigslist (stylized as craigslist) is an American classified advertisements website with sections devoted to jobs, housing, for sale, items wanted, services, community service, gigs, résumés, and discussion forums.

Craigslist has more than 50 million unique monthly users with more than 20 billion page views per month making it one of the most visited websites in the USA. The company recorded a revenue of USD 694 million in 2016, with a net profit of USD 500 million.

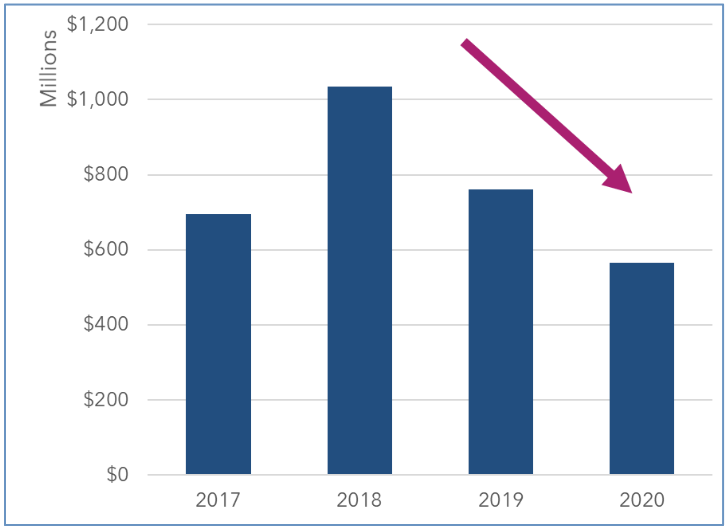
REVENUE



We see that almost 70% of their revenue comes from their job postings. This is reflective of their pricing strategy which is geared towards charging customers for posting jobs. Their top 3 sources of revenue are

- $75 for job posts in the bay area
- $25 for job posts elsewhere
- $10 for apartment posts

YOY REVENUE



It is quite evident from the graph that the YoY revenue is dropping after reaching its peak of over $1 billion in 2018. It is also clear that the job postings are not performing well since the job posts are nearing all-time lows.

PAIN POINTS

After analyzing the websites of competitors of craigslist such as indeed, Upwork, etc. We saw a clear difference in the quality of jobs present on both sites. Their competition has a better mechanism for categorizing jobs, filtering fake jobs, and matching them to relevant users. The first step would be to increase the quality of jobs posted by

- Correctly classifying the jobs
- Identify and initiate action against scam jobs

PROBLEM STATEMENT

Build a model to correctly categorize jobs and then go on to detect if the jobs are fake now. Generate additional insights that may help craigslist pull their revenues back up.

We are confident that with the use of this model, craigslist will be able to drastically to improve both the quality of jobs posted and the overall user experience.

OBJECTIVE

The goals we are trying to achieve are

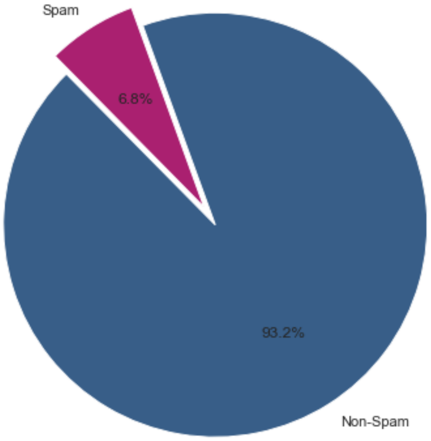
- Gather high-quality data to serve as a base for our analysis and modeling
- Generate insights from gathered data to understand the job posts and their various features
- Build excellent machine learning models to correctly classify jobs and identify the fake ones
- Suggest strong recommendations based on the analysis and insights generated from this entire process that will help Craigslist increase their YoY revenue.

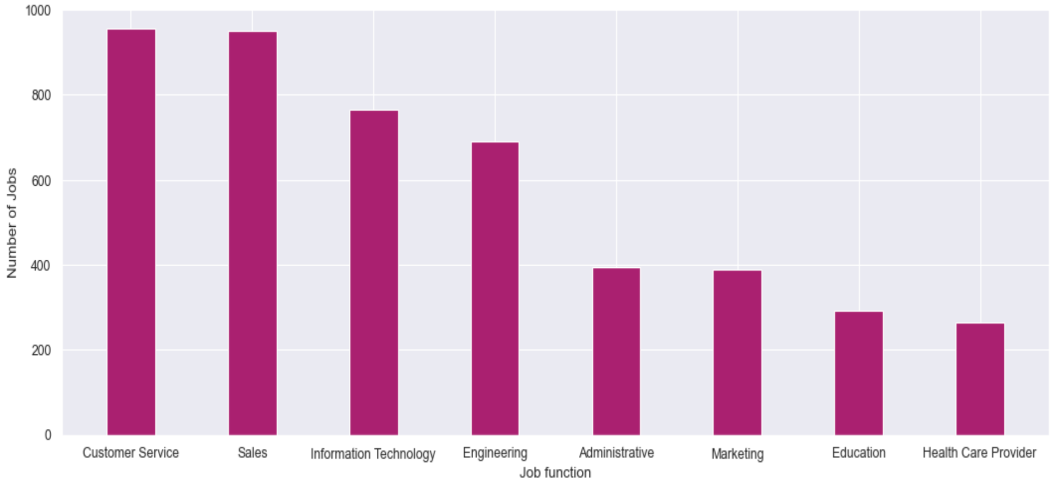
DATA REQUIREMENT AND COLLECTION

To build a machine learning classifier to correctly categorize jobs into correct functions and filter the fake/scam jobs, we need the data having job descriptions and corresponding fraudulent flags. The dataset may also include the following features - job title, location, department, industry, function, requirements, benefits, employment type, company logo, if any, required experience, required education.

To obtain the relevant data, we are gathering data from two sources.

- Web scraping to get listings from the Craigslist jobs section. We manually label the data depending on whether the job listing is authentic or fake/scam.
- We enrich the dataset by getting third-party data. Specifically, we use EMSCAD - Employment Scam Aegean Dataset (http://emscad.samos.aegean.gr/), which is a labeled data set for job scams.





DATA PREPARATION

We followed a six-step process to reduce the dimensionality of our input text data and convert it into a numerical representation. The steps are listed below:

- Stemming
- Lemmatization
- Remove stop words and convert to lowercase
- Convert text to TF-IDF representation (unigrams and bigrams, with minimum document frequency as 2)
- Convert text to Word Embedding representation (Glove Embedding)
- Handle class imbalance using K-means SMOTE

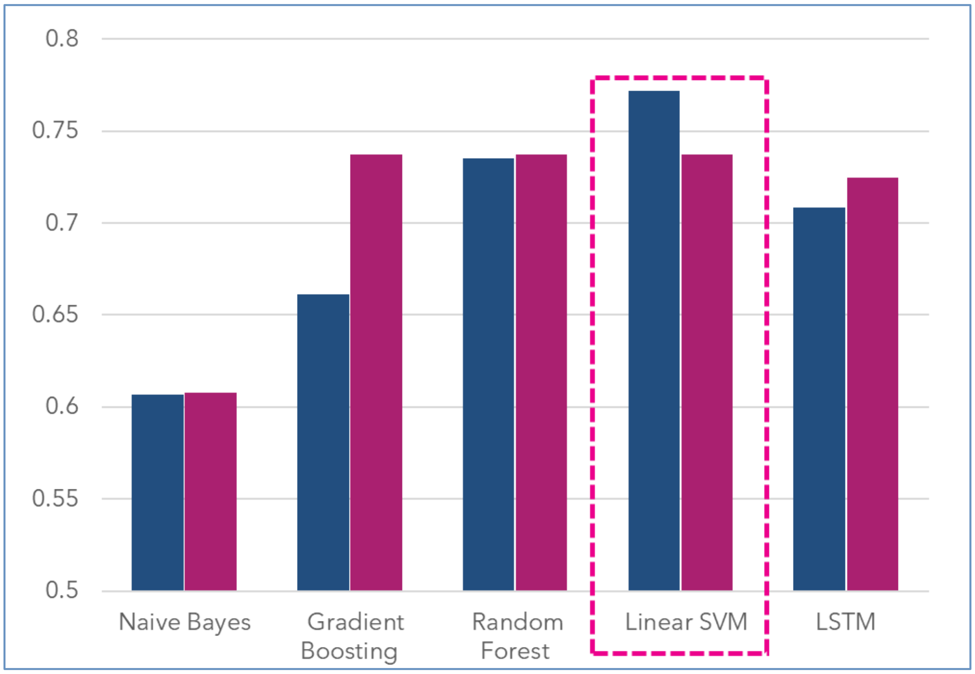
DATA MODELLING – JOB CATEGORY CLASSIFICATION

Job function category classification is a multi-class classification problem where we are trying to correctly classify the job advertisements listed on craigslist into the correct categories. The input/predictors we have used here are the job title and description. The output target variable has total of 8 classes - Customer Service, Sales, Education, Engineering, Marketing, Administrative, Healthcare, Information technology.

We tried multiple models :

- Naïve Bayes (Default parameters)
- Random Forest (n_estimators=500, max_depth=6, bootstrap=True, class_weight = 'balanced')
- Support Vector Machines (penalty = l1)
- Boosting (n_estimators=100, learning_rate=1.0, max_depth=1)
- LSTM (Glove embedding, 100 layers,, dropout=0.3, recurrent_dropout=0.3 with two dense layers of 1024, one dense layer of 8 neurons and softmax layer at the end, categorical crossentropy loss with adam optimizer)

The linear SVM performed the best with an F1 score of 0.77 while the Naïve Bayes performed the worst with an F1 score of 0.62



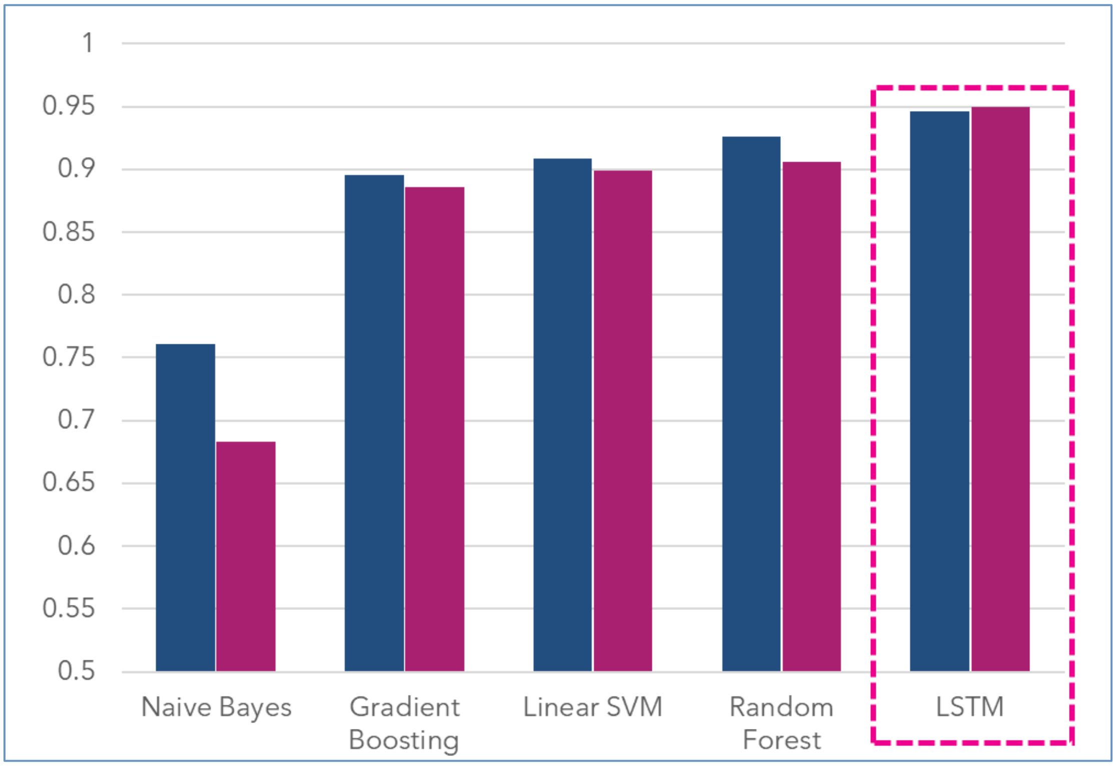
DATA MODELLING –FRAUDULENT JOB CLASSIFICATION

Fraudulent job classification is a binary class classification problem where we are trying to correctly classify if the job advertisements listed on craigslist is a scam or not. The input/predictors we have used here are the job title and description. The output target variable is fraudulent with 0 and 1 classes.

We tried multiple models :

- Naïve Bayes (Default parameters)
- Random Forest (n_estimators=500, max_depth=6, bootstrap=True, class_weight = 'balanced')
- Support Vector Machines (penalty = l1)
- Boosting (n_estimators=100, learning_rate=1.0, max_depth=1)
- LSTM – (Glove embedding, 100 layers, dropout=0.3, recurrent_dropout=0.3 with two dense layers of 1024, one dense layer of 1 neuron and sigmoid layer at the end, binary crossentropy loss with adam optimizer)

The LSTM performed the best with an F1 score of 0.95 while the Naïve Bayes performed the worst with an F1 score of 0.76.



INSIGHTS

SCAM Jobs

Scam jobs worsen the user experience which affects all stakeholders. For the person who applies for the job, the information in the specific section is misleading. For the person who posts the jobs, the scam jobs in a way takes their listing's position which hampers the brand image of Craigslist.

Job Categorization

Incorrect categorization of a job would lead to potential applicants not being able to reach the listing. This impacts Craigslist massively as the organization posting the job would not get these applicants. This could lead them to doubt the efficacy of Craigslist in acting as a platform for job seekers and employers. Job listings have been Craigslist's highest revenue-generating avenue for the past years and it is important to provide a seamless experience to all stakeholders for the growth of the organization.

One Size Does Not Fit All

Model selection should be based on the task, the available data and testing. In our analysis we found that the Linear SVM performs better for job categorization which is a multi-classification problem. On the other hand, the LSTM model performs better for job scam detection, which was a binary classification problem.

Resampling

Over-sampling techniques like SMOTE improve model performance by a significant margin. In our case to cater the problem statement fraudulent job classification, we initially found only 7% of the data to be fraudulent. So using the SMOTE technique, we were able to perform oversampling, leading to a balanced dataset with 50% fraudulent and 50% non-fraudulent jobs.

RECOMMENDATIONS

Eliminate spam jobs

Use our model to identify spammers and initiate strict action against such accounts since this directly impacts the brand image.

Better categorize jobs

The opportunity cost of misclassified jobs is high. Categorizing the jobs correctly helps in customer retention.

Shift focus on NY. Charge $75 for jobs posts

Features of jobs posted in New York are very similar to jobs in San Francisco. We would recommend shifting the focus to New York and charging $75 for job posts. Assuming the jobs posted to remain constant, this increases revenue from jobs by 13.65%.

Add more checks

We see in our analysis that job posts with a company logo are less likely to be a spam. So, we propose craiglist to have more similar checks to ensure the authenticity of the listings.

FUTURE SCOPE

Pricing strategy:

We can collect more data enriching our dataset with more predictors. These predictors can help us devise a pricing strategy for jobs across cities in the US. The features we can use here would be population demographics, types of jobs listed in areas and types of businesses posting the jobs. This would provide us with the optimum pricing strategy for each city, opening new revenue streams.

Customer experience

We will understand more variables that have a big impact on the validity of the posting. These models can be extended to other revenue sources as well where content filtering could help improve the customer experience.