Predicting Median Income from Yelp Review Language

Stephanie Chen SCHEN751@STANFORD.EDU
Michael Zhu MZHU25@STANFORD.EDU

Abstract

We investigate the relationship between the language consumers use when reviewing businesses to the socioeconomic level of those businesses' areas by predicting a zip code's median household income bracket from the language used in reviews of those businesses. We use a stochastic gradient descent model with modified Huber loss and L1 regularization and find roughly 39% accuracy in predicting income and 64% accuracy on a related task predicting the cities in which the businesses are located. In looking at the most predictive terms for each task, we find that customers of businesses in high-income areas focus on quality of service and special occasions, while customers in low-income areas focus on ease of access and location.

1. Introduction

In recent years, consumers have become more and more dependent on peer reviews to judge the quality or worth of a product or business. Instead of relying only on branded advertising or "authoritative" sources like magazine top-10 lists to choose where to spend their money, two-thirds of consumers today place their trust in recommendations by peers. [1] In addition, because of their breadth — in just the four years after Yelp's 2005 launch, its user reviews in Seattle had covered 70% of the city's restaurants, while *The Seattle Times* had only covered 5% [2] — online reviews provide a uniquely comprehensive look at the consumer-oriented business landscape of an area. We take advantage of this thoroughness as a lens through which to examine a city's income patterns and distribution.

The income distribution of an area directly affects the distribution of its infrastructure, its public facilities, and its businesses — for example, with regards to food stores, low-income neighborhoods on average have more liquor stores and four times as many grocery stores as high-income neighborhoods, whereas high-income areas have more chain supermarkets, bakeries, and specialty stores. [3] We are interested in how some of these socioeconomic and cultural differences are reflected not only in the businesses themselves but also in the language with which their customers write about them. Do restaurant patrons in different areas focus on the same traits of the food or ambience? Do customers of businesses in areas of different median income express their anger or satisfaction in different language?

Most prior work on Yelp or other online reviews has focused on predicting business rating [4][5] based on review text or leveraging review text features to build user profiles [6], essentially restricting focus to the business-consumer relationship or the consumer ecosystem. To classify reviews, prior research generally has used linear regression models and quadratic regression models, using hinge or RMS error functions; most also use a bag-of-words model for the text features,

which we do as well. Yelp reviews have also been used to predict business revenue performance [7] but not income; the closest prior research to our income bracket task uses language from Twitter or other social media activity instead. [8]

We input solely the "review" and "tip" text within the Yelp data for a set of businesses, as obtained from the Yelp dataset (explained below). We use a stochastic gradient descent linear classifier using multiclass-modified Huber loss to output a predicted income range for that business's zip code. By examining this relationship between review language and income and finding predictive language features of this relationship, we hope to reveal possible trends in the consumer focuses and language patterns of customers from different socioeconomic areas.

2. Dataset and Features

Text. We use text data from the Yelp Dataset, which includes 2.7 million reviews across 86,000 businesses in ten cities across North America and Europe; we only use the American cities (Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, and Madison), data for which comprise roughly 80% of the whole set. Data for each business includes a set of textual reviews and tips, which we use as the basis for our feature set. [10] We intentionally ignore other Yelp data such as "dollar-sign" cost rating and categorization ("fast food," "fine dining," "Chinese," etc.), as our focus is on consumers and their language patterns instead of the businesses themselves.

To create our feature set, we first replace identifying features like dollar costs ("a $15 salad") and city and state names with placeholders (e.g. _COST_, _CITY_). Location names were obtained by first filtering out the names of the major metropolitan areas covered by our data (e.g. Pittsburgh, Phoenix) and their states. We then performed a similar classification task to search for language features predictive of location and removed other highly predictive locations like suburbs and adjacent areas (ex. Scottsdale). Some generally identifying features (e.g. "casino" likely indicates a business in Las Vegas) were kept because of both their relevance to the income task and the infeasibility of judging which city features were "too" identifying. We also use placeholders for potentially useful punctuation like exclamation points and question marks, and we add tags to preserve indicators of capitalization before normalizing the text to lowercase.

We use a simple unigram bag-of-words approach for our language features, as experimentation with bigram and trigram features produced no improvement in results. We vectorize the text using scikit-learn's HashingVectorizer [9], which converts a text corpus into a sparse matrix representation of counts.

Income. Each business is labeled with an address; we map the zip code in that address to a median income from the University of Michigan's 2006-2010 Median Household Income set. We discretize incomes into four buckets — less than $40,000, $40,000 to $55,000, $55,000 to $70,000, and over $70,000 — that had relatively uniform support upon testing and that we used as the targets for prediction.

3. Methods

We use scikit-learn's stochastic gradient descent (SGDClassifier [9]) model and perform a grid search over different loss (hinge, log, modified Huber, squared hinge, perceptron, see Figure 1) and penalty (L1, L2, elasticnet with mixing parameter 0.15, see Figures 2 and 3) functions. For each classifier trained in the grid-search, we performed 3-fold cross-validation to obtain an accuracy score for each, then picked the best-scoring classifier to run on the test set.

The weight vector for the classifier is learned by performing the stochastic gradient descent update for the given loss and regularization functions on every example the training set, repeated over

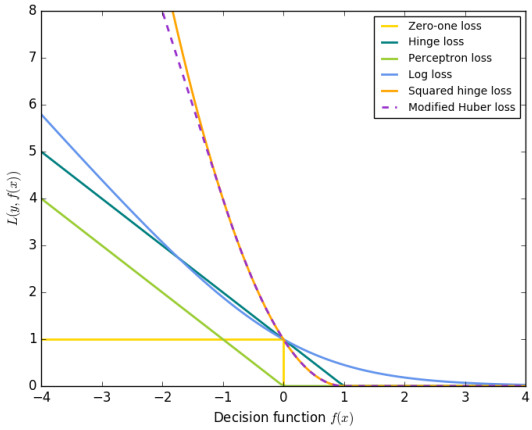

Figure 1. Comparison of loss functions [9]

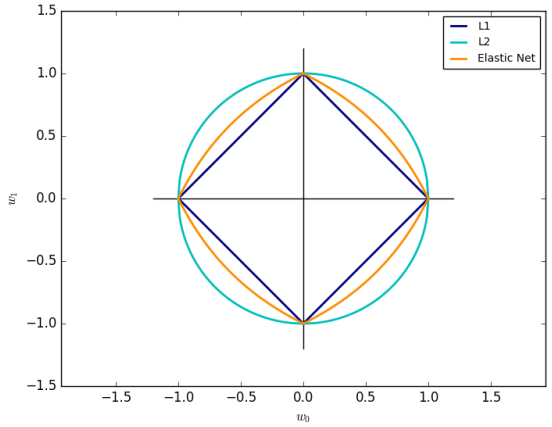

Figure 2. Contours of penalty functions [9]

some number of epochs:

$$w \leftarrow w - \eta(\alpha \frac{\partial R(w)}{\partial w} + \frac{\partial L(w^T x_i + b, y_i)}{\partial w})$$

We find that modified Huber loss with L1 loss performed the best for both classification tasks. The scikit-learn SGDClassifier uses a one-vs-all (OVA) strategy for multiclass classification, which involves training a binary classifier for each class, where samples in that class are considered positive and samples from all other classes are considered negative. The modified

$$R(w) := \sum_{i=1}^{n} |w_i|$$
$$R(w) := \frac{1}{2} \sum_{i=1}^{n} w_i^2$$
$$R(w) := \frac{\rho}{2} \sum_{i=1}^{n} w_i^2 + (1 - \rho) \sum_{i=1}^{n} |w_i|$$

Figure 3. Equations for L1, L2, and elasticnet penalty functions, respectively [9]

$$L(y, f(x)) = \begin{cases} \max(0, 1 - y\, f(x))^2 & \text{for } y\, f(x) \geq -1, \\ -4y\, f(x) & \text{otherwise.} \end{cases}$$

Figure 4. Equation for modified Huber loss function [11]

Huber loss is a generalization of the Huber loss (used for regression) for classification (see Figure 4) [11]. The Huber loss is a smooth loss function (similar to the squared hinge loss) that "brings tolerance to outliers." [11] Training a multiclass classifier using the modified Huber loss yields a probabilistic classifier, i.e. for each input the classifier calculates a probability distribution over the classes and outputs the class with the highest probability. [9]

4. Results and Discussion

Best models. We tested different methods at various stages in our classifier to find the best overall model for our data. At the data level, we experimented with different preprocessing and found that, for example, lemmatizing our text and removing stop words both produced no improvement in results. We also tested bigram and trigram feature selection and found no improvement, suggesting that single terms or keywords used by consumers were more predictive of area income than larger language patterns in phrasing and structure.

The modified Huber loss is a of the squared hinge loss, so it penalizes negative margins more harshly than the standard hinge loss. Thus the fact that the modified Huber loss performs significantly better than the hinge loss on the city prediction task suggests that perhaps there is not much variance in the city classes, so it makes sense to assign high loss to samples that have

negative margin.

We also initially tested a standard linear SVM [9] model, with hinge loss and L2 regularization. This model performed slightly better than our final SGD models on experimental subsets, with about 41% accuracy, but was computationally intractable to implement on our entire set without extensive feature reduction.

Overall results. We include here the results for both our primary income prediction task as well as the city location prediction task, which we initially performed to find location-identifying terms but which we present here as a related application of interest. For the income task, our best model as tested on the training set, using modified Huber loss with L1 regularization, resulted in 38.8% accuracy on the test set — slightly lower than test performance with a hinge loss, L2 regularization model, which resulted in 39.3% accuracy. The two accuracies are comparable, with discrepancy likely due to the variation in results from the stochastic gradient descent model, which unlike SVM is not guaranteed to converge to a global optimum.

On the city prediction task, our best model in training — again, modified Huber loss with L1 regularization — resulted in 63.8% accuracy, suggesting that there is a significantly stronger correlation between language features and city, which is consistent with our prior beliefs. It should also be noted that the support for the cities is not as evenly distributed as the support for the income buckets (e.g. there are less than 1,000 reviews for Urbana-Champaign, but Phoenix has over 30,000), which likely contributes to the higher subset accuracy.

To find which terms were actually most predictive of income brackets or city locations, we reran both tasks on experimental subsets using scikit-learn's TfidfVectorizer [9]. We do not use the tf-idf approach in our full set — while perhaps a more accurate way of weighting the relative importance of word features, TfidfVectorizer was also computationally inefficient as document state, which includes the full set of words seen, must be maintained.

Table 1. Predictive terms for median income and city classification tasks

INCOME	CITIES
UPTOWN	TOWN
VALLEY	CASINO
WINE	SMOG
AIRPORT	SCHOOL
HOSPITAL	HERE
MEMBERS	VALLEY
DOWNTOWN	WEST
BUSY	PATIO
CAMPUS	DOWNTOWN
MALL	SUN

A selection of the most predictive terms for each task are shown in Table 1, in no particular order. Even though our overall accuracy in predicting income bracket was lower than expected, the terms obtained are reasonably indicative: "movie," "wife," "wine," and "members" as well as "ask" and "receive" were output as predictive of higher-income locations, suggesting a higher-end products or services and a slant towards special occasions (e.g. dates) as well as a focus on service with "ask" and "receive." Terms output as predictive of lower-income locations included "central," "campus," "affordable," "airport," and "downtown," showing a heavy focus on location and accessibility.

Interestingly, the predictive terms for the location task were less evidently related, though our classifier performed better. Words like "desert" and "patio" are somewhat useful, potentially indicating cities like Phoenix or Las Vegas, while "school" appeared as a predictive word likely given that several of the cities have substantial university presences.

5. Conclusions and Future Work

Overall, we found that given the aggregate review and tip text of a business, the best model for

predicting the income bracket of that business's zip code district was a modified Huber loss-based stochastic gradient descent classifier with L1 regularization, which performed with roughly 39% accuracy on our test set. The same classifier performed with roughly 64% accuracy on the related task of predicting the cities in which each business was located. This might suggest that consumer language is not particularly predictive of area income; however, given that the output lists of predictive terms are, qualitatively, fairly indicative of each income bracket, it's more likely that our methodology was affected by noisy data or other issues in the feature space.

Regarding future improvements to our current framework — to begin, we'd like to expand our dataset to Yelp reviews in more cities. Our current dataset covered cities with similar median incomes, to each other and to the national median, and similar intra-city income distributions, generally with higher-income, lower-density suburbs and lower-income centers (with the exception of Las Vegas, where the Strip shifts the income distribution significantly). Underlying patterns such as frequency of Yelp usage or business density in different areas might have affected our model. With additional cities with an expanded range of median incomes (e.g. San Francisco, with a median income of over $85,000) and income distributions (e.g. New York, with drastic changes between and even within boroughs) to which to expose our model, our model might be less susceptible to some of these patterns and learn more predictive textual features.

We'd also like to improve our feature selection process — our current simple bag-of-words model takes in the entire reviews-tips aggregate for a business, and while we found that removing stop words was not useful, we could utilize scikit-learn's various selection methods to remove other low-variance features.

Finally, in terms of furthering our methodology in investigating the review-income relationship, we propose two extensions. First, a single business's type of service or product, its costs, and its customers are not necessarily fully indicative of an entire area, although observed trends in business type and patronage between high- and low-income areas are significant and often drastic, as discussed in the introduction. However, we could get a better picture of a zip code's income by combining our business predictions with a weighted average or some other aggregation method and measuring our accuracy in that task instead. Second, we're also interested in unsupervised approaches to this and other review-related relationships, for example using PCA to see if businesses could be clustered by income or by geography, based on certain textual features. This could expose even more relevant patterns present in both Yelp reviews and online reviews as a wider field of study.

References

[1] "Global Trust in Advertising: Winning Strategies for an Evolving Media Landscape", The Nielsen Company, Sep. 2015.

[2] M. Potamias, "The warm-start bias of Yelp ratings", arXiv e-print 1202.5713, Feb. 2012.

[3] M. Blanding, "The Yelp Factor: Are Consumer Reviews Good for Business?", Oct. 2011.

[4] L. V. Moore and A. V. Diez Roux, "Associations of Neighborhood Characteristics With the Location and Type of Food Stores" in *American Journal of Public Health*, Feb. 2006, Vol. 96, No. 2, pp. 325-331.

[5] M. Fan and M. Khademi, "Predicting a Business Star in Yelp from Its Reviews Text Alone", arXiv e-print 1401.0864, Jan. 2014.

[6] G. Ganu, Y. Kakodkar, A. Marian, "Improving the quality of predictions using textual information in online user reviews", in *Information Systems*, Vol. 38, No. 1, Mar. 2013, pp. 1-15.

[7] M. Luca, "Reviews, Reputation, and Revenue: The Case of Yelp.com", Harvard Business School NOM Unit Working Paper 12-016, 2011.

[8] D. Preoiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, N. Aletras, "Studying User Income through Language, Behaviour and Affect in Social Media", in *PLoS ONE*, Vol. 10, No. 9, Sep. 2015.

[9] scikit-learn. Pedregosa et al. "Scikit-learn: Machine Learning in Python". JMLR 12, pp. 2825-2830, 2011.

[10] "Yelp Dataset Challenge." [Online].

[11] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms", ICML, 2004.