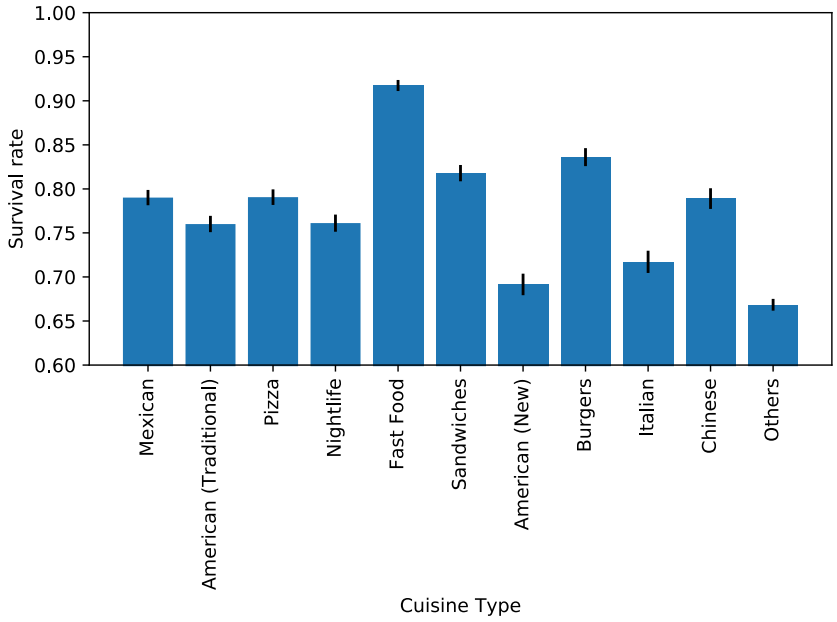
Web Appendixes: Can Consumer-Posted Photos Serve as a Leading Indicator of Restaurant Survival? Evidence from Yelp

Mengxia Zhang and Lan Luo
February 2022

Appendix A. Survival Identification and Survival-Related Summary Statistics

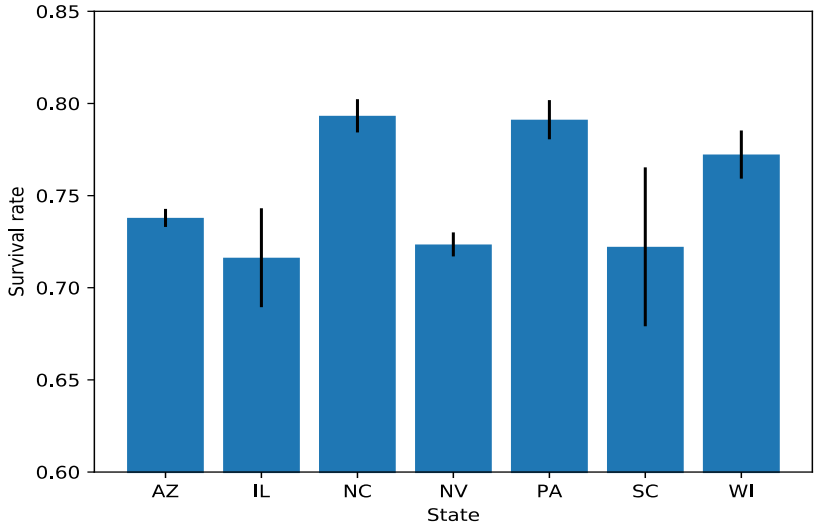
Figure A1. Example of Closure Time Identification

Figure A2. Survival Rate by Top 10 Cuisine Types



The error bars represent ± 1 times the standard error of each point estimate.

Figure A3. Survival Rate by State

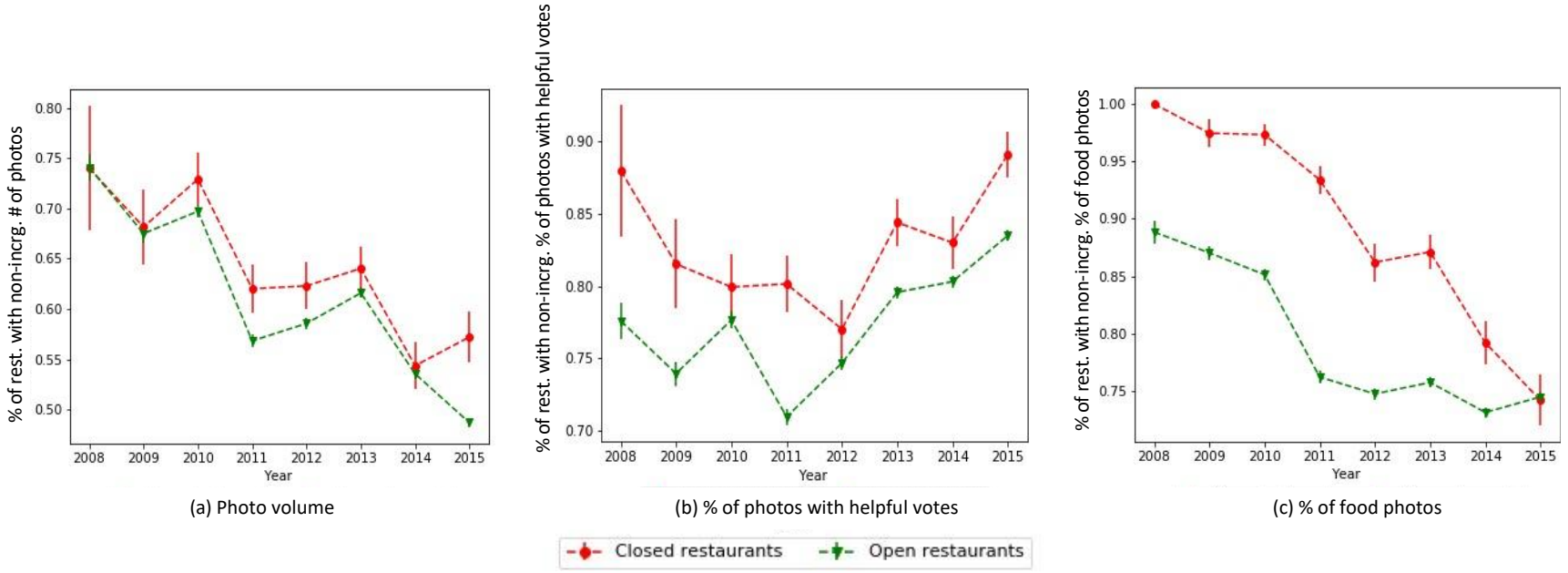


States listed alphabetically. The error bars represent ± 1 times the standard error of each point estimate.

To investigate model-free evidence supporting our conjecture that consumer-posted photos may serve as a leading indicator of restaurant survival, we compare the proportions of closed and open restaurants that have: 1) non-increasing numbers of photos; 2) non-increasing proportion of photos with helpful votes in each year; 3) non-increasing proportion of food photos in each year. We also include the proportion of food photos in this comparison because our paper suggests that food photos are more predictive of survival that other photo content. Using numbers of photos as an example, for the year 2014, we first find restaurants that were closed and those that were still open in 2014. We then check whether each restaurant received fewer photos during 2013 than in 2012. Lastly, we calculate the proportion of restaurants with a non-increasing number of photos among closed and open restaurants, respectively. To make our comparisons more reliable, we conduct such comparisons only in years with at least 20 open and at least 20 closed restaurants. This comparison is similar in spirit to a difference-in-difference approach: last year vs. the year before last difference and closed vs. open difference. Such a difference-in-difference approach is superior to plotting the total numbers of photos for open vs. closed restaurants directly because it takes the baseline differences between the two groups of restaurants into account.

Figure A4 shows the comparison results. We observe that the red dots representing proportions among closed restaurants are generally above the green triangles representing proportions among open restaurants. Such patterns indicate that, by and large, restaurants near closure are more likely to have non-increasing photo volume, proportion of photos with helpful votes, and proportion of food photos than are open restaurants. These comparisons provide model-free evidence that photo volume, proportion of photos with helpful votes, and proportion of food photos may be predictive of restaurant survival after the macro year-trend is controlled.

Figure A4. Photo Volume, % of Photos with Helpful Votes, and % of Food Photos Trends Comparisons Across Years

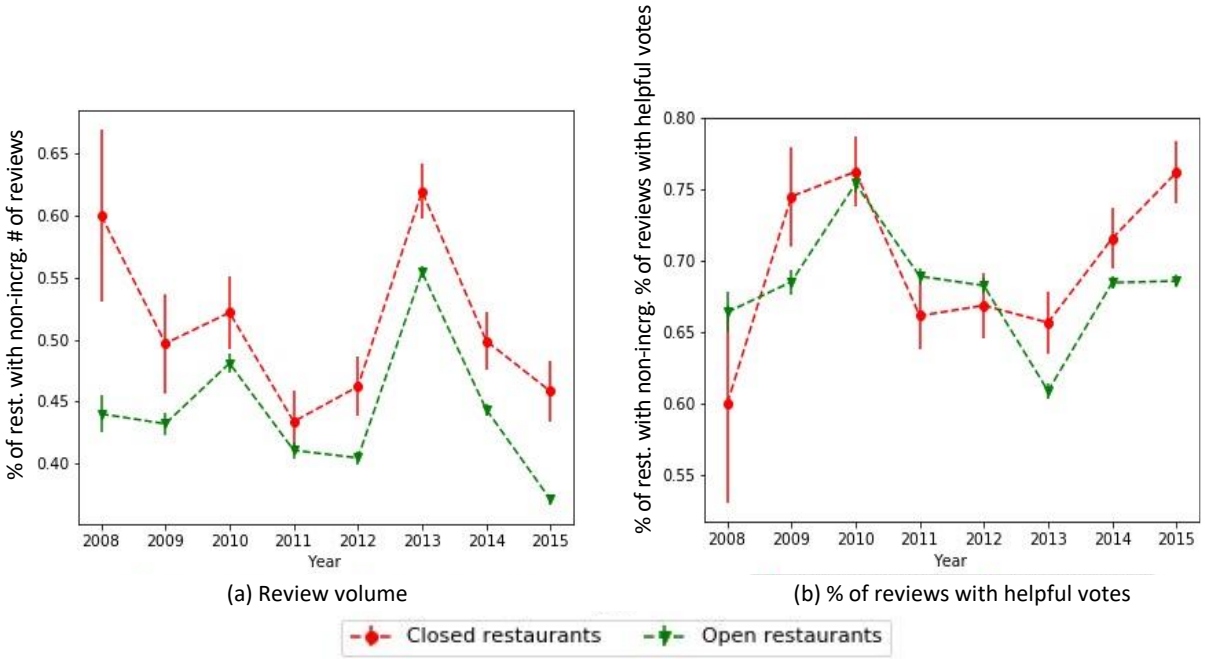


(a) Photo volume (b) % of photos with helpful votes (c) % of food photos

The error bars represent ± 1 times the standard error of each point estimate.

As a benchmark, we compare the proportions of closed and open restaurants that have: 1) non-increasing numbers of reviews; 2) non-increasing proportion of reviews with helpful votes in each year. Figure A5 shows that restaurants near closure are more likely to have non-increasing review volume than open restaurants. However, such a trend is not as apparent when it comes to the proportion of reviews with helpful votes for restaurants, indicating the proportion of reviews with helpful votes may not be as predictive of restaurant survival as its counterpart in photos.

Figure A5. Review Volume and % of Reviews with Helpful Votes Trend Comparison Across Years



(a) Review volume

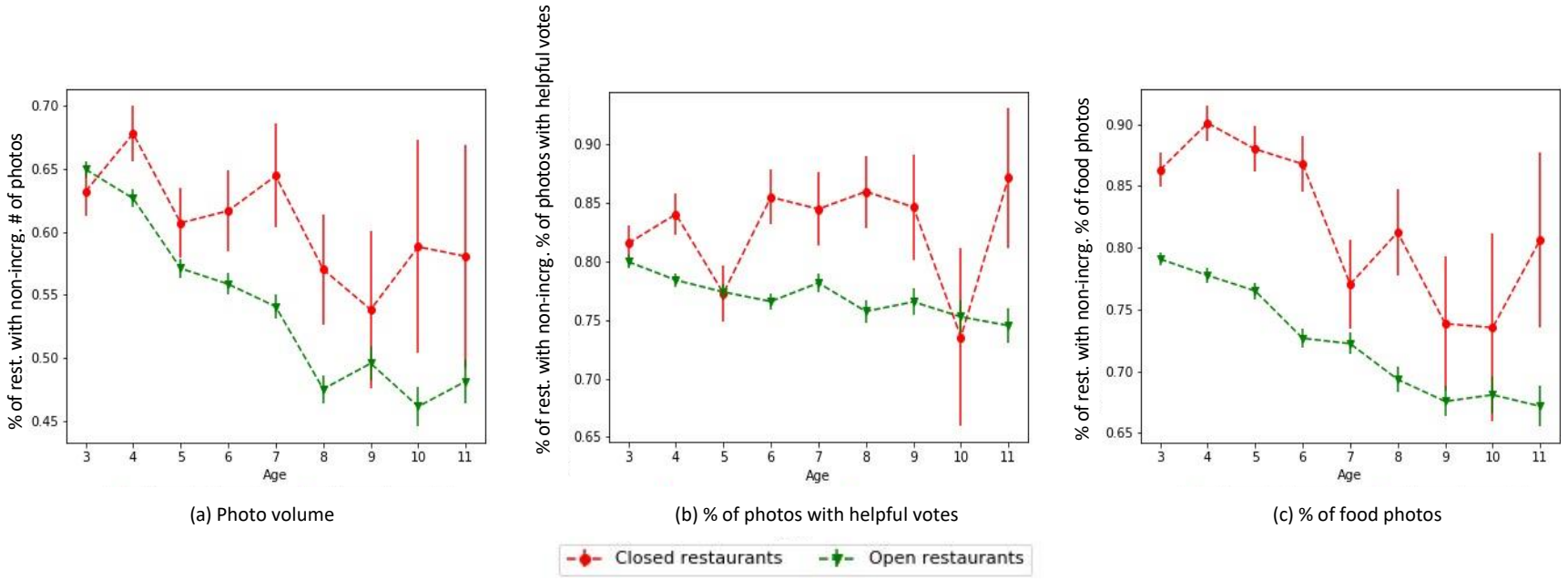
(b) % of reviews with helpful votes

The error bars represent ± 1 times the standard error of each point estimate.

We also compare the proportions of closed and open restaurants that have: 1) non-increasing numbers of photos; 2) non-increasing proportion of photos with helpful votes; 3) non-increasing proportion of food photos at each restaurant age. For example, for the age=3, we first find restaurants that were closed and those that were still open at age=3. We then check whether each restaurant received fewer photos at age=2 than at age=1. Lastly, we calculate the proportion of restaurants with a non-increasing number of photos among closed and open restaurants, respectively. To make reliable comparisons, we conduct comparisons only in ages with at least 20 open and at least 20 closed restaurants. To compare volume and helpful votes in the last two years, we also have to choose restaurants that lived more than two years and listed on Yelp for more than two years. Based on all considerations above, we plot comparisons for ages between 3 and 11 in Figures A6 an A7.

Similar to the comparison across calendar years, Figure A6 shows that the red dots representing proportions among closed restaurants are generally above the green triangles representing proportions among open restaurants. Such patterns again indicate that, by and large, restaurants near closure are more likely to have non-increasing photo volume, proportion of photos with helpful votes, and proportion of food photos than are open restaurants. These comparisons also provide model-free evidence that photos may be predictive for restaurant survival after controlling for restaurant age.

Figure A6. Photo Volume, % of Photos with Helpful Votes, and % of Food Photos Trends Comparisons Across Ages

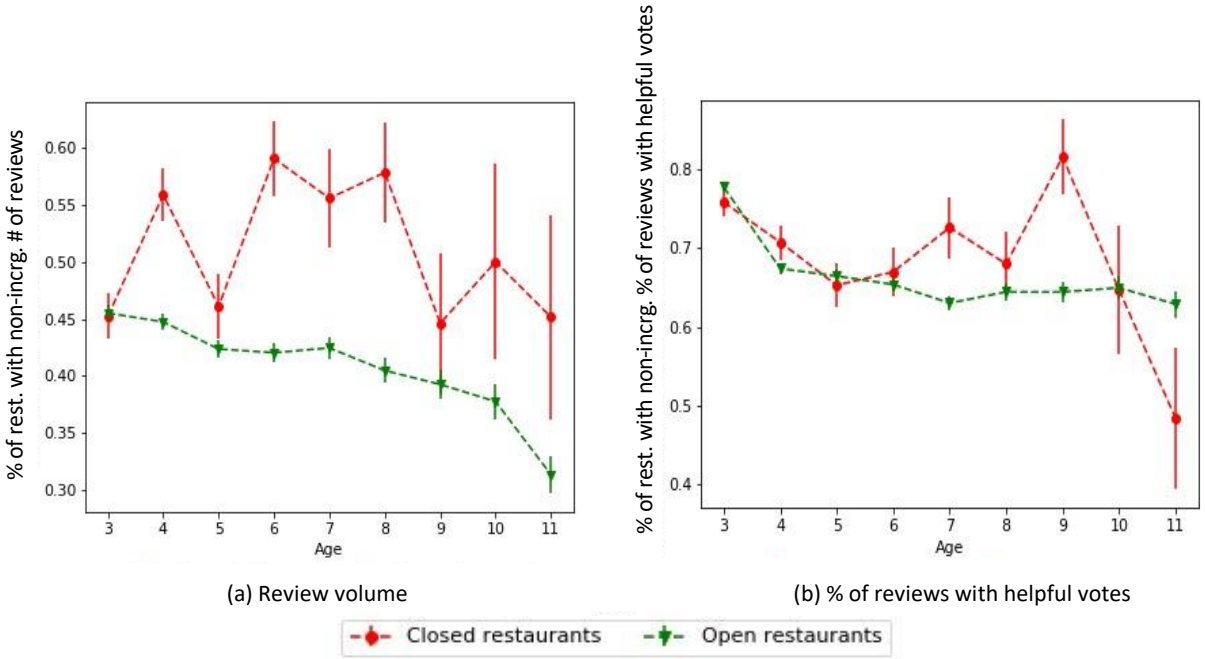


(a) Photo volume (b) % of photos with helpful votes (c) % of food photos

The error bars represent ± 1 times the standard error of each point estimate.

As a benchmark, we also compare the proportions of closed and open restaurants that have: 1) non-increasing numbers of reviews; 2) non-increasing proportion of reviews with helpful votes at each restaurant age. Figure A7 shows that restaurants near closure are more likely to have non-increasing review volume than open restaurants. Nevertheless, such a trend is not apparent for the proportion of reviews with helpful votes, indicating again the proportion of reviews with helpful votes may not be as predictive of restaurant survival as its counterpart in photos.

Figure A7. Review Volume and % of Reviews with Helpful Votes Trend Comparison Across Ages



The error bars represent ± 1 times the standard error of each point estimate.

Appendix B. Photo Analysis

Clarifai Labels

Clarifai, a photo-based API specializing in computer vision, was founded by Matthew Zeiler, author of "Visualizing and Understanding Convolutional Networks" (Zeiler and Fergus 2014). Clarifai has been a market leader in photo content detection since it won the top five places in photo classification at the ImageNet 2013 competition.

We use the Clarifai "Food" model to identify objects in food and drink photos as per the classification on Yelp. The "Food" model is designed to process photos of food and drink.[1] Upon comparing APIs provided by Google, Microsoft, Amazon, and Clarifai, we find that Clarifai can provide the most refined labels for food. For example, only Clarifai could label content as refined as "strawberry" in a photo when we compare these APIs. Figure A8 provides four examples of food and drink photos with labels and probabilities provided by the Clarifai "Food" model. The Clarifai API does a good job in recognizing a large variety of food ingredients in a photo.

Figure A8. Examples of Clarifai Labels for Food and Drink Photos

pizza	1
crust	0.999
pepperoni	0.999
dough	0.999
sauce	0.998
mozzarella	0.998
tomato	0.994
salami	0.99
basil	0.77
meat	0.715

sushi	1
salmon	0.998
nori	0.99
seafood	0.983
rice	0.952
sashimi	0.941
fish	0.94
caviar	0.911
shrimp	0.903
nigiri	0.85

corn	1
corn salad	0.987
vegetable	0.978
salad	0.977
onion	0.895
tomato	0.872
lettuce	0.615

strawberry	1
fruit	0.998
berry	0.998
mint	0.997
refreshment	0.991
ice	0.963
glass	0.943
juice	0.941

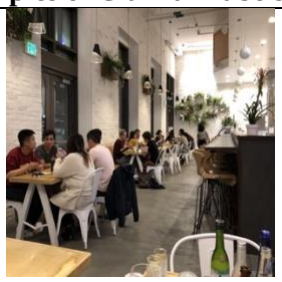
We use the Clarifai "General" model for interior, outside, menu, and other photos as per the classification of Yelp. The 'General' model recognizes over 11,000 different labels.[2] Figure A9 shows four examples of restaurant photos with labels and probabilities provided by the Clarifai "General" model.

[1] https://clarifai.com/models/food-image-recognition-model-bd367be194cf45149e75f01d59f77ba7

[2] https://clarifai.com/models/general-image-recognition-model-aaa03c23b3724a16a56b629203edc62c

Figure A9. Examples of Clarifai Labels for Interior and Outside Restaurant Photos

people	0.993
restaurant	0.991
adult	0.977
man	0.966
dining	0.957
indoors	0.941
drink	0.933
woman	0.93
table	0.93
wine	0.882
group	0.879
candle	0.866
service	0.861



restaurant	0.988
dining	0.983
table	0.974
people	0.954
interior design	0.949
chair	0.937
indoors	0.914
bar	0.9
man	0.875
group	0.875
seat	0.852
woman	0.846
cafeteria	0.842

street	0.98
city	0.968
no person	0.957
urban	0.952
building	0.932
commerce	0.9
restaurant	0.9
light	0.896
window	0.876
town	0.861

outdoors	0.979
no person	0.972
sky	0.949
light	0.942
dusk	0.931
evening	0.922
nightlife	0.916
city	0.915
street	0.886

We include only labels with more than 50 percent confidence according to the Clarifai API, resulting in a total of 5,080 unique labels. Table A1 shows the top 100 Clarifai labels.

Table A1. Top 100 Clarifai Labels

sauce	shrimp	pizza	lettuce	salsa
chicken	table	avocado	seat	street
cheese	people	drink	lamb	milk
pork	onion	cheddar	ice	duck
meat	pepper	wine	outdoors	group
bacon	chocolate	text	pasta	corn
beef	cake	city	mushroom	house
vegetable	egg	tuna	tacos	toast
garlic	butter	plate	dinner	apple
rice	salmon	coffee	ham	mozzarella
salad	seafood	refreshment	chips	ginger
sweet	sandwich	lemon	dairy product	shop
cream	bar	chair	bun	pastry
bread	chili	turkey	lunch	sign
potato	sausage	beans	curry	cuisine
food	room	tea	dish	window
indoors	soup	spinach	inside	basil
steak	beer	lobster	hotel	noodle
tomato	meal	sushi	honey	light
fish	crab	pie	ramen	lime

Topic Modeling of Clarifai Labels

We randomly sampled 10,000 photos to calibrate the LDA model for topic modeling, with 80% for training and 20% for testing. We implemented the LDA model using a Python library "GENSIM" (Rehurek and Sojka 2010), a widely used Python library for topic modeling. We used the online variational inference algorithm for the LDA training (Hoffman, Bach, and Blei 2010). We used all the 1,281 labels that Clarifai identified in more than 0.05% of photos. Removing rare labels reduces the risk that the results are influenced by outlier labels (Netzer et al. 2019; Tirunillai and Tellis 2014). We ran the model with 2 to 40 topics on the training dataset. We found that the fitted LDA with 10 topics yielded the highest topic coherence score (Röder et al. 2015) on the testing dataset. Table A2 presents the 10 topics and the most representative words for each topic based on the relevance score calculated using $\lambda = 0.5$ (Sievert and Shirley 2014). Intuitively, the representative words for a topic tend to appear together (Tirunillai and Tellis 2014).

Table A2. The 10 LDA Topics and Representative Words with Highest Relevance

(Topics are listed in an arbitrary order)

ID	Topic name	Representative words with highest relevance
1	Symbol and decoration	Illustration, symbol, sign, design, retro, art, image, text, vintage, decoration
2	Seafood	Fish, seafood, salmon, tuna, sushi, shrimp, crab, sashimi, lobster, nigiri
3	Text and information	Page, text, number, document, paper, time, information, form, order, writing
4	Meat	Pork, beef, chicken, steak, lamb, meat, duck, rib, broth, tenderloin
5	Inside	Chair, indoors, room, seat, table, bar, inside, dining, interior design, counter
6	People	Portrait, people, music, recreation, festival, wear, performance, group, celebration, facial expression
7	dessert	Chocolate, cake, cream, ice, tea, sweet, coffee, strawberry, ice cream, dessert
8	Outside	Outdoors, city, street, road, vehicle, light, urban, daylight, entrance, sky
9	food and drink	Meal, refreshment, plate, dinner, lunch, dish, cuisine, fruit, glass, drink
10	Sandwich and pizza	Cheese, bacon, sandwich, bread, cheddar, pizza, sausage, chicken, tomato, ham

In the extrapolation step, we applied the calibrated LDA parameters to extract topic distribution for each photo in our entire data set of 755,758 photos. Following the standard approach (Netzer et al. 2019), a 10-dimensional topic distribution vector was generated for each photo. Each dimension represented the empirical percentage of Clarifai labels assigned to a topic, with all percentages summing up to 1. For example, the topic distribution for one photo could be [5%, 5%, 15%, 15 %, 10%, 10%, 0%, 0 %, 20%, 20%]. We then used the average topic distribution for photos for each restaurant-year as inputs for our prediction model.

Photographic Attributes

Table A3. Definitions of Photographic Attributes

Types	Attributes	Description
Color	Brightness	Brightness refers to the overall lightness or darkness of the photo.
	Saturation	Saturation indicates color purity. Pure colors in a photo tend to be more appealing than dull or impure ones.
	Contrast	Contrast is the difference in brightness between regions.
	Clarity	The proportion of pixels with enough brightness.
	Warm hue	The proportion of warm color in a photo.
	Colorfulness	It distinguishes multi-colored photos from monochromatic, sepia, or simply low contrast photos.
Composition	Diagonal dominance	Measures how close the main region in the photo is positioned to diagonal lines.
	Rule of thirds	The "golden ratio" (about 0.618). The photo is divided into nine equal segments by two vertical and two horizontal lines. The rule of thirds says that the most important elements are positioned near the intersections.
	Physical visual balance	It measures how symmetrical a photo is around its central vertical line (horizontal physical balance) or central horizontal line (vertical physical balance).
	Color visual balance	It measures how symmetrical the colors of a photo are around its central vertical line (horizontal color balance) or central horizontal line (vertical color balance).
Figure-ground relationship	Size difference	It measures how much bigger the figure is than the ground.
	Color difference	It measures how different the color of the figure is from that of the ground.
	Texture difference	It measures how different the texture of the figure is from that of the ground.
	Depth of field	The range of distance from a camera that is acceptably sharp in a photo.

Color For the first five color features, we first convert photos from RGB space to HSV space (each pixel is a 3-dimensional vector representing hue, saturation, and value). In the following, we explain how we extract each attribute for color.

1. **Brightness** is the average of the value dimension of HSV across pixels (Datta et al. 2006). We normalized brightness to be between 0 and 1, with a higher score meaning brighter.
2. **Saturation** is the average of saturation cross pixels (Datta et al. 2006). We normalized saturation to be between 0 and 1, with a higher score meaning more saturated.
3. **Contrast** of brightness was calculated as the standard deviation of the value dimension of HSV cross pixels. We normalize the score to be between 0 and 1, with a higher score meaning higher contrast.
4. **Clarity** We first normalize the value dimension of HSV to be between 0 and 1, and we define a pixel to be of enough clarity if its value is bigger than 0.7. Then clarity is defined as the proportion of pixels with enough clarity.
5. **Warm hue** Following Wang et al. (2006), the warm hue level for the photo is the proportion of warm hue (i.e., red, orange, yellow) pixels in a photo.
6. **Colorfulness** We followed Hasler and Suesstrunk (2003) to measure colorfulness for each photo based on the RGB vector of each pixel. We normalized this metric to be between 0 and 1, with a higher score meaning more colorful.

Composition To compute the four attributes of composition, we first assigned each pixel a saliency score based on computer vision methods (Montabone and Soto 2010). Then we separated a photo into 10 segments based on the superpixel algorithm (Mori et al. 2004; Ren and Malik 2003). The segment with the highest average saliency is identified as the salient region (Zhang et al. 2018). Identifying the salient region is necessary because diagonal dominance and rule of thirds are defined with respect to the main element of a photo. In the following, we explain how we extract each attribute for composition.

1. **Diagonal dominance** We calculate the distance between the center of the salient region to each of the two diagonals of a photo. We then define diagonal dominance as the negative of the minimum of two distances (Wang et al. 2013). We normalize this score to be between 0 and 1, with a higher score meaning stronger diagonal dominance.
2. **Rule of thirds** We calculate the distance from the center of the salient region to each of the four intersections of the two horizontal lines and the two vertical lines that evenly divide the photo into nine parts (Datta et al. 2006). Then the rule of thirds is defined as the negative of the minimum of the four distances. We normalized this score to be between 0 and 1, with a higher score meaning a stronger rule of thirds.
3. **Physical visual balance** We calculate two scores for physical visual balance: vertical and horizontal physical visual balances. We first calculate the weighted center of the photo by weighting the centers of segments by their respective saliency. Then the vertical physical visual balance is defined as the negative of the vertical distance between the weighted center and the horizontal line that divides the photo into two equal parts. The horizontal physical visual balance is defined as the negative of horizontal distance between the weighted center and the vertical line that divides the photo into two equal parts (Wang et al. 2013). We normalize the two scores to be between 0 and 1, with a higher score meaning a higher balance.
4. **Color visual balance** We calculated two scores for color visual balance: vertical and horizontal color visual balances. Following Wang et al. (2013), to measure vertical color balance, we first separate the photo into two equal parts by a horizontal line. A pair of pixels is a pixel on the top part and its symmetric counterpart on the bottom part. Then the vertical color balance is defined as the negative of the average of Euclidean distance cross pixel pairs. To measure horizontal color balance, we first separate the photo into two equal parts by a vertical line. A pair of pixels is a pixel on the left part and its symmetric counterpart on the right part. Then the horizontal color balance is defined as the negative of the average of Euclidean distance cross pixel pairs. We normalize the two scores to be between 0 and 1, with a higher score meaning higher color balance.

Figure-ground relationship Figure refers to the foreground, and ground refers to the background, of a photo. For the first three figure-ground relationship features, we first use the Grabcut algorithm (Rother et al. 2004) to identify the figure and background of each photo. In the following, we explain how we extract each attribute for the figure-ground relationship.

1. **Size difference** We take the difference between the number of pixels of the figure and that of the background, normalized by the total number of pixels of the photo (Wang et al. 2013).
2. **Color difference** We first calculate the average RGB vectors for figure and ground. Then the color difference is the Euclidean distance between the two RGB vectors (Wang et al. 2013). We normalize the score to be between 0 and 1, with a higher score meaning a bigger figure-ground color difference.
3. **Texture difference** We use the Canny edge detection algorithm (Canny 1987) to detect edges in the figure and background. Then we compute the density of edges in the figure and background. The texture difference is the absolute value of the difference between figure edge density and background edge density. This score is normalized to be between 0 and 1, with a higher score meaning a higher difference.
4. **Depth of field** Professional photos usually use a low depth of field to enhance the most important element in the photo. A photo of a low depth of field is usually sharp in the center while out of focus in the surrounding area. We divide the photo into 16 equal regions. Following Datta et al. (2006), we compute the Daubechies wavelet coefficients (Daubechies 1992) in the high-frequency for each HSV

dimension of a photo. Then we calculate the depth of field by dividing the sum of wavelet of the center four regions by the sum of wavelet of the whole photo, for each HSV dimension. Thus, there are three scores of the depth of field: depth of field (hue), depth of field (saturation), and depth of field (value). A higher score means a lower depth of field.

Table A4. Summary Statistics of Photographic Attributes

	Count	Mean	Standard deviation	Minimum	Maximum
Color					
Brightness	755,758	0.53	0.13	0.00	1.00
Saturation	755,758	0.43	0.15	0.00	1.00
Contrast	755,758	0.47	0.11	0.00	0.99
Clarity	755,758	0.31	0.20	0.00	1.00
Warm hue	755,758	0.85	0.17	0.00	1.00
Colorfulness	755,758	0.25	0.10	0.00	1.00
Composition					
Diagonal dominance	755,758	0.52	0.17	0.00	1.00
Rule of thirds	755,758	0.56	0.18	0.00	1.00
Vertical physical balance	755,758	0.94	0.06	0.00	1.00
Horizontal physical balance	755,758	0.90	0.08	0.00	1.00
Vertical color balance	755,758	0.43	0.08	0.00	1.00
Horizontal color balance	755,758	0.44	0.08	0.02	1.00
Figure-ground relationship					
Size difference	755,758	0.45	0.19	0.00	1.00
Color difference	755,758	0.21	0.14	0.00	1.00
Texture difference	755,758	0.08	0.07	0.00	1.00
Depth of field (hue)	755,758	0.25	0.13	0.00	1.00
Depth of field (saturation)	755,758	0.30	0.09	0.00	1.00
Depth of field (value)	755,758	0.32	0.09	0.00	0.99

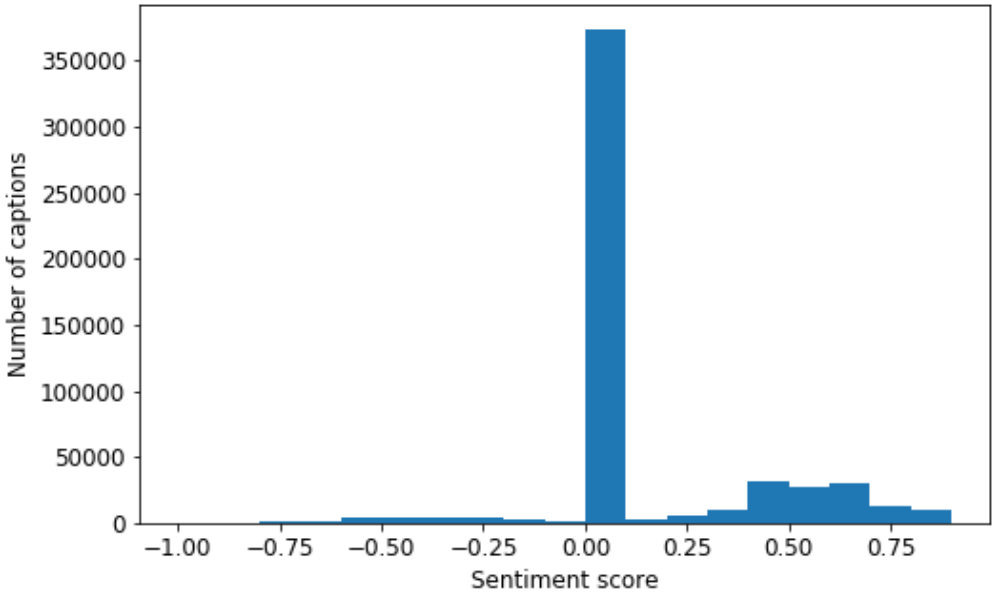
Photo Caption Sentiments

We use VADER (Hutto and Gilbert 2014) sentiment analysis to analyze photo captions. We learn that photo captions can be categorized into three categories: neutral dish names captions, positive captions, and negative captions. Table A5 shows examples of captions and respective sentiment for each category. Figure A10 demonstrates the distribution of caption sentiment.

Table A5. Examples of Caption Classification and Sentiment Scores

ID	Caption	Sentiment
Dish name captions	Sam & Emma Sandwich	0
	Red velvet pancakes	0
	Our sushi boat	0
	Vegetables	0
	Mandarin Kung Pao Chicken	0
Positive captions	Inspiration + food - love it!	0.8356
	I love being transported to another country by their food!	0.6696
	Kiffka pita so good! :)	0.6009
Negative captions	This place sucks. Flat beer bad service had to leave	-0.7351
	Can i get a bag of salt to add to my sodium overload?	-0.3612

Figure A10. Distribution of Extracted Caption Sentiment



Appendix C. Review Analysis

Restaurant Quality Dimensions

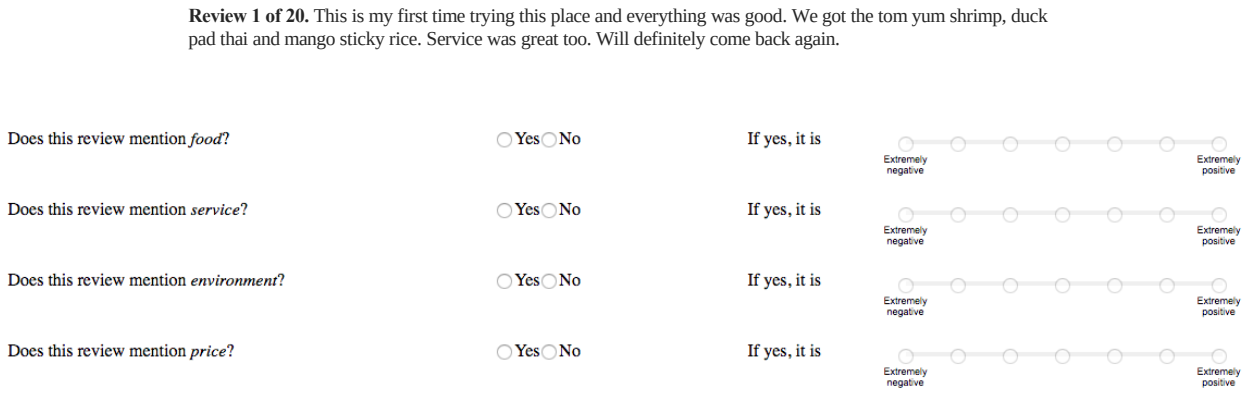
We define the four quality dimensions in Table A6, based on prior literature on restaurant quality dimensions (Bujisic et al. 2014; Hyun 2010; Ryu et al. 2012). We first label the 10,000 reviews through a Mturk survey. After an introduction of the definition for each restaurant quality dimension, each consumer was presented with 20 screens of reviews with one review per screen. The 20 reviews were randomly selected from the 10,000 reviews. See Figure A11 for a screenshot of the survey interface. On average, it took 20 minutes for each consumer to finish the survey. We removed respondents who finished the survey with less than 5 minutes, given that these participants might skim through each review too fast and did not provide reliable responses. As a result, each review is on average read by eight consumers.

We instructed each consumer to answer whether a review mentioned food, service, environment, or price of the restaurant; if so, whether the specific content was positive or negative on a 7-point Likert scale, with 1 being "extremely negative" and 7 being "extremely positive." The average of sentiment votes was used to label sentiment for each quality dimension mentioned in each review.

Table A6. Definitions of Restaurant Quality Dimensions Extracted from Reviews

Dimension	Definition
Food	food, drink, menu variety, nutrition, healthiness, plating, food presentation, serving size, freshness, etc.
Service	employee behavior, attitude, responsiveness, reliability, tangibility, empathy, assurance, process speed, wait time, etc.
Environment	interior/exterior design, décor, cleanliness, ambience, aesthetics, lighting, layout, table setting, employee appearance, location, etc.
Price	good/bad value for money, price of items, etc.

Figure A11. Restaurant Review Survey

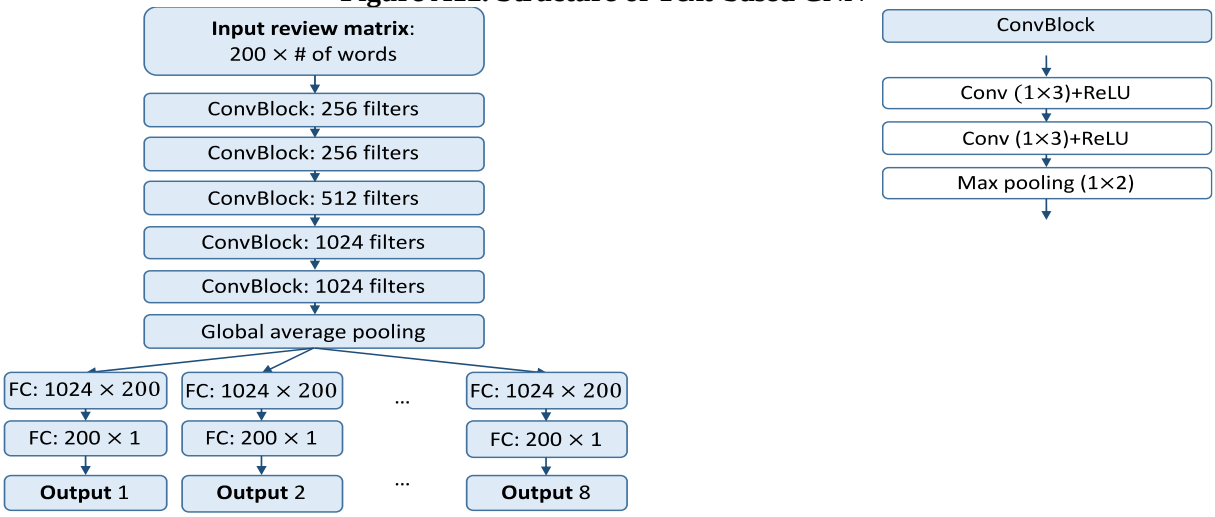


We use a deep learning model to extract the following four quality dimensions from reviews: food, service, environment, and price. Each dimension has two labels: whether the dimension was mentioned and the sentiment of the dimension. To increase the accuracy of the text-based deep learning model, we follow the standard procedure (Timoshenko and Hauser 2019) and use pre-trained word embeddings as inputs for our text-based CNN. Specifically, we use GloVe (Pennington et al. 2014) word embeddings. GloVe provides an effective measure for the linguistic or semantic similarity of the corresponding words. Under GLoVe, each word is represented by a 200-dimensional vector. Each review is then represented by a 200 × number of words matrix, which serves as the inputs for the text-based CNN.

Figure A12 depicts the structure of the text-based CNN. The left panel of the figure demonstrates the entire structure of the text-based CNN. The right panel of this figure depicts the detailed structure of

the ConvBlock (the building block of the text-based CNN). The model is constructed in a way that is similar to the photo-based CNN — VGG model (Simonyan and Zisserman 2014), just replacing 2-dimensional convolution with 1-dimensional convolution. The idea is borrowed from Kim (2014), which shows that text-based CNN structures adapted from established image-based CNN structures perform well on several benchmark datasets. We also add more convolutions and pooling operations because our text is relatively long. The input goes through five convolutional blocks (ConvBlock), one global average pooling, and two full connections (FC). The guiding principle of convolution is to identify those features that can best predict the output variable. The primary function of pooling is dimension reduction and some degree of shift invariance (LeCun et al. 1998). Average pooling calculates the average on the previous layer. The full connections can be regarded as a classification task that links the last layer before full connection with the output variable.

Figure A12. Structure of Text-based CNN



Convolution: feature extraction. A 1×3 weight matrix, "filter," multiplies with each segment of the previous layer, and the dot product becomes an element on the next layer. Different filters capture different information from the previous layer.
ReLu: non-linear transformation, which turns all negative values on the previous layer to zero.
Max pooling: subsampling method. A 1×2 window glides over the previous layer, and only the max value of each window is kept as an element on the next layer.

We employ a multitasking structure. As suggested by the deep learning literature (see Ruder (2017) for a review), multitasking can increase the accuracy of a deep learning model intuitively because the information cross tasks may be complementary. A multitasking model structure is also more efficient than a single-task structure. In our context, for each review, there are eight tasks to predict: whether a review mentions food, service, environment, and price; whether the sentiment for food, service, environment, and price is positive or negative. Under a single task text-based CNN, we would train eight separate deep learning models. With a multitasking structure, we only need to train one deep learning model. To implement the multitask CNN, we link the layer after global average pooling with a separate full connection for each task, generating a score for each task.

For tasks that aim to identify whether the review mentions one or more dimensions of restaurant quality, the label s_r^k=1 if review r mentions the quality dimension for task k, s_r^k=0 if review r does not mention the quality dimension for task k, with $k = 1, 2, 3, 4$; $r = 1, 2, 3, \ldots, 8000$. For the other four tasks $k = 5, 6, 7, 8$ that label sentiment of each quality dimension, only reviews that mention a quality dimension are kept for the respective sentiment task. For sentiment of each quality dimension, we follow the conventional procedure as in Liu et al. (2019), Zhang et al. (2018), and Zhang et al. (2015) and

convert the 7-point Likert scale sentiment to binary levels to mitigate potential noises in the data. In particular, a sentiment label is defined as positive ($s_r^k = 1$) if it receives a rating above 4; negative ($s_r^k = 0$) if a rating is less than 4; and is disregarded if its rating equals 4.

The loss function is defined in Equation (A1), which is the sum of loss for the eight tasks altogether, with $\widehat{s_r^k}$ being the predicted probability that $s_r^k = 1$. The specification is called "cross-entropy" in computer science literature, which is equivalent to negative log-likelihood.

(A1)

$$Loss = -\sum_{r=1}^{8000}\sum_{k=1}^{8}\left[s_r^k \ln\left(\widehat{s_r^k}\right) + (1 - s_r^k)\ln\left(1 - \widehat{s_r^k}\right)\right]$$

We randomly split the 10,000 reviews into 80% for calibration and 20% for out-of-sample testing. In the calibration process, the text of each review is treated as model inputs, and the outputs are the eight quality dimension scores labeled by the Mturk survey. Parameters are optimized using a stochastic gradient descent method by minimizing the above loss function. We validate the calibrated text-based CNN on the holdout dataset. Our text-based CNN yields good AUC scores for all the eight tasks on the holdout dataset, as shown in Table A7.

We then extrapolate the calibrated CNN to extract eight quality dimension scores for all reviews in our entire dataset of 1,121,069 reviews. Each review went through layers of the text-based CNN with review text and calibrated parameters as inputs and eight numerical scores between 0 and 1 representing mentioning and sentiments of the four restaurant quality dimensions as outputs. The distribution of the extracted quality scores is shown in Table A8. The calibration and extrapolation of the text-based CNN were implemented using Tensorflow, a deep learning library within Python.

Table A7. Out-of-Sample Testing Performance of the Text-based CNN

	AUC	Valid samples for testing
Mention		
Food	0.9187	1990
Service	0.9163	1986
Environment	0.8814	1980
Price	0.9449	1993
Sentiment		
Food	0.9515	1721
Service	0.9627	1325
Environment	0.8841	781
Price	0.9139	552

Table A8. Summary Statistics of Restaurant Quality Dimension Scores

	Count	Mean	Standard deviation	Minimum	Maximum
Mention					
Food	1,121,069	0.9714	0.1667	0	1
Service	1,121,069	0.9998	0.0141	0	1
Environment	1,121,069	0.4043	0.4908	0	1
Price	1,121,069	0.2570	0.4370	0	1
Sentiment					
Food	1,089,039	0.7726	0.3903	0	1
Service	1,120,880	0.7249	0.4025	0	1
Environment	453,203	0.8311	0.3384	0	1
Price	288,070	0.6017	0.4512	0	1

Topic Modeling of Reviews

For the topic modeling of reviews, we also calibrate a Latent Dirichlet allocation (LDA) (Blei et al. 2003) model to summarize the reviews' content. We use the same procedure of topic modeling on photos for reviews. We randomly sample 10,000 reviews to calibrate the LDA model, with 80% for training and 20% for testing. We implement the LDA model using a Python library "GENSIM" (Rehurek and Sojka 2010), a widely used Python library for topic modeling.

We use the online variational inference algorithm for the LDA training (Hoffman et al. 2010). We use all the 5,330 words[3] that appeared in more than 0.05% of reviews. Removing rare words reduces the risk that the results are influenced by outlier words (Netzer et al. 2019; Tirunillai and Tellis 2014). We run the model with 2 to 40 topics on the training dataset and find that the fitted LDA yields the highest topic coherence score (Röder et al. 2015) on the testing dataset when the number of topics is equal to 20. Table A9 presents the 20 topics and the most representative words for each topic based on the relevance score calculated using $\lambda = 0.5$ (Sievert and Shirley 2014). Intuitively, the representative words for a topic tend to appear together (Tirunillai and Tellis 2014).

Table A9. The 20 LDA Topics and Representative Words with Highest Relevance
(Topics are listed in an arbitrary order)

ID	Topic name	Representative words with highest relevance
1	Indian Food	Indian, naan, masala, tikka, paneer, mais, frites, tater, tot, meatloaf
2	Mixed/negative reviews on food	Sauce, chicken, flavor, salad, cheese, shrimp, steak, taste, didnt, wasnt
3	Japanese food	Sushi, roll, fish, Japanese, tuna, tofu, tempura, sashimi, spicy, chef
4	Breakfast and Brunch	Dog, toast, brunch, hot, juice, omelette, fresh, scramble, local, french
5	Satisfaction	Great, food, service, excellent, friendly, good, recommend, staff, highly, love
6	Love for a place	Best, place, ever, food, love, try, eat, phoenix, town, always
7	Happy hour and price	Happy, hour, slider, discount, alcohol, Monday, free, drink, fondue, price
8	Italian food	Bread, tomato, garlic, mozzarella, feta, meatball, Italian, crust, marinara, pasta
9	Buffet	Buffet, Vega, pasta, wine, squid, seafood, station, tray, din, selection
10	Breakfast	Coffee, breakfast, egg, pancake, waffle, biscuit, bacon, omelet, morning, benedict
11	Dissatisfaction	Floor, response, horrible, needless, boring, crew, host, havent, bore, service
12	Mexican food	Chip, Mexican, salsa, enchilada, tapa, band, tortilla, fajitas, queso, guacamole
13	Sandwich	Sandwich, beef, pork, turkey, bbq, chicken, sub, roast, meat, lettuce
14	Dessert	Dessert, pie, cake, chocolate, appetizer, creme, cheesecake, peanut, course, cream
15	General service	Order, ask, get, take, minute, say, wait, come, didnt, table
16	Fast food	Burger, shake, fry, pizza, bun, hamburger, topping, Grimaldis, smash, gourmet
17	Bar	Beer, bar, game, watch, night, bartender, selection, sport, tap, place
18	Asian food	Thai, noodle, pho, curry, Chinese, ramen, rice, bowl, pad, Vietnamese
19	Specific service	Kid, time, wait, long, service, issue, table, hostess, waiter, refill
20	Special event	Birthday, venue, football, reward, gift, private, cater, celebrate, sing, thanksgiving

In the extrapolation step, we apply the calibrated LDA parameters to extract topic distribution for each review in our entire data set of reviews. A 20-dimensional topic distribution vector is generated for each review. Each dimension represents the empirical percentage of words assigned to a topic, with all percentages summing up to 1. For example, the topic distribution for one word could be [2.5%, 2.5%,

[3] For word preprocessing, we implemented the following procedures following Tirunillai and Tellis (2014): 1) remove punctuations (i.e., keep only words, numbers, space); 2) change capitalized characters to lower case; 3) tokenize words (i.e., a sentence to list of words); 4) implement part of speech tagging and keep only nouns, verbs, adjectives, adverbs; 5) lemmatize words (e.g., "booths" to "booth", help us to normalize words); and 6) remove stop words.

7.5%, 7.5 %, 5%, 5%, 0%, 0 %, 10%, 10%, 2.5%, 2.5%, 7.5%, 7.5 %, 5%, 5%, 0%, 0 %, 10%, 10%]. We then use the average topic distribution for reviews for each restaurant-year as inputs for our predictive model.

Variety of Objects in Reviews

To measure the variety of objects in a review, we count the number of unique nouns in a review. All reviews in our dataset contain 30,927 unique nouns. All nouns are lemmatized to ensure they have the same format. E.g., "apples" are lemmatized as "apple", so "apples" and "apple" are treated as the same noun. We also remove stop words, words with fewer than two characters, and words that appear in less than 0.001% of reviews.

Top 100 Nouns

As a robustness check for using topic modeling, we use an alternative way to summarize specific content in reviews. While all reviews in our dataset contain 30,927 unique nouns, the vast majority of nouns appear in only a few reviews. Hence, we use the top 100 nouns (encompassing all nouns with more than 5% frequency in our review dataset) to capture the specific content in reviews.

Table A10. Top 100 Nouns in Reviews

food	friend	plate	option	manager
place	star	soup	check	water
time	dinner	beef	husband	room
service	experience	portion	family	fan
restaurant	fry	quality	work	crab
order	flavor	potato	town	fun
chicken	day	egg	choice	bbq
menu	taste	visit	chip	group
table	meat	waitress	coffee	party
pizza	hour	buffet	cream	end
drink	beer	shrimp	house	piece
sauce	sushi	wine	line	chocolate
price	review	feel	couple	tomato
salad	roll	pork	guy	tea
burger	location	customer	thai	style
bar	bread	home	strip	size
lunch	breakfast	selection	course	decor
night	rice	spot	wife	pasta
staff	area	appetizer	week	name
people	taco	item	owner	glass

Appendix D. Restaurant Survival Predictive Model: Gradient Boosted Trees

Technical Details

We follow Chen and Guestrin (2016) to carry out our restaurant survival predictive model using XGBoost algorithm. We provide additional technical details of our predictive model in this appendix.

The regularization term $\Omega(\theta)$ in Equation (1) of the paper aims to prevent overfitting. We carefully tune the hyper-parameters (e.g., γ and λ in $\Omega(\theta)$) before calibrating XGBoost, as described below. Following the convention of hyper-parameter tuning of XGBoost, we tune the hyper-parameters sequentially as follows.[4] Table A11 lists the hyper-parameters that we tuned as well as the values that we tried.

Table A11 Hyper-parameters for Tuning

Hyper-parameters	Definitions	Values tried
max_depth	The maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit.	[3,4,5,6]
min_child_weight	The minimum number of nodes in a leaf. Increase this value may reduce overfitting.	[1,2]
colsample_bytree	The fraction of features to use when building each tree. A lower value might reduce overfitting.	[0.6,0.8,1]
subsample	The fraction of observations to subsample at each step for building a tree. A lower value might reduce overfitting.	[0.6,0.8,1]
γ	Penalize more leaves. A higher value might reduce overfit.	[0,4,8]
λ	Regularization term on weights. Increasing this value will make a model more conservative.	[0,1,2]
η	Learning rate. Step size shrinkage used in updating to prevent overfitting	[0.05,0.1,0.3,0.5]

Ninety percent of restaurants are randomly chosen to tune the hyper-parameters. We further randomly split the restaurants with 80% for training and 20% for validation. We first find the best combination of max_depth and min_child_weight, among the 4*2 =8 combinations, as shown in the last column of Table A11, using a grid search. Given the best combination of max_depth and min_child_weight, we then pick the best values among the nine combinations of colsample_bytree and subsample. Then, given the best combination of max_depth, min_child_weight, colsample_bytree, and subsample, we choose the best values among the nine combinations of γ and λ. Finally, given the best combination of the six tuned hyper-parameters, we choose best η among the four values. Then we end up with the following values for the hyper-parameters in our XGBoost model: max_depth =4, min_child_weight=2, colsample_bytree=1, subsample=1, $\gamma = 4$, $\lambda = 2$, and $\eta = 0.1$. Additionally, we learn empirically that the model converges within 100 trees. Therefore, we choose the number of trees to be 100 in our XGBoost model.

We provide details below for the model evaluation metrics used in Tables 5-6 in Section 3.2 of the paper. ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier (Hanley and McNeil 1982). False positive rate is on the horizontal axis. False positive rate =1-specificity= $\frac{\#\ of\ false\ positive\ samples}{\#\ of\ \ real\ negative\ samples}$. True positive rate is on the vertical axis. True positive rate = recall=sensitivity=$\frac{\#\ of\ true\ positive\ samples}{\#\ of\ \ real\ positive\ samples}$. Given a false positive rate on the horizontal axis, the higher the true positive rate, the better. KL divergence =$-\frac{1}{N}\sum_{it}[y_{it}\ln\widehat{y_{it}} + (1-y_{it})\ln(1-\widehat{y_{it}})]$, which is

[4] https://blog.cambridgespark.com/hyperparameter-tuning-in-xgboost-4ff9100a3b2f

equivalent to negative log-likelihood. Pseudo $R^2 = 1 - \frac{LogLikelihood_{proposed}}{LogLikelihood_{null}}$. We reweight the data for sensitivity, specificity, and balanced accuracy, based on the discussion about the parameter "scale_pos_weight" on https://xgboost.readthedocs.io/en/latest/parameter.html and "handle imbalanced dataset" on https://xgboost.readthedocs.io/en/latest/tutorials/param_tuning.html.

Comparisons with Other Predictive Algorithms

We compare the XGBoost algorithm used in the paper with two alternative predictive algorithms: random forests and SVM (see Table A12). We also tried Lasso, hazard model, and OLS in our comparisons, which are omitted from Table A12 because their predictive performance was considerably worse than the three models presented. For random forests, we train an ensemble of 100 trees, the same as the number of trees in our XGBoost model. For SVM, we use L1 regularization and square of the hinge loss. We implement both random forests and SVM algorithms in Python. We then calculate means and standard deviations of prediction performance measured by AUC across years 2010-2015 and cross-validation iterations. For each photo or review related variable, we use the same main model specification as in the paper, namely $OnePeriod_{t-1}$+Cum_{t-1}. Table A12 shows that XGBoost dominates random forests and SVM in our context. It is not surprising because XGBoost is very flexible in handling potentially high-order interactions among predictors (Friedman 2001) and can process sparse data efficiently (Chen and Guestrin 2016).

Table A12 Out-of-Sample Time Periods and Restaurants Prediction Performance (AUC) of Different Predictive Algorithms

	XGBoost	Random forests	SVM
Baseline	0.7020 [a]	0.6596	0.6904
	(0.0047)	(0.0048)	(0.0054)
Baseline + review	0.7152 [a]	0.6607	0.7007
	(0.0048)	(0.0047)	(0.0053)
Baseline + photo	0.7612 [a]	0.7005	0.7329
	(0.0066)	(0.0061)	(0.0068)
Baseline + review + photo	0.7660 [a]	0.7081	0.7371
	(0.0065)	(0.0065)	(0.0072)
Total obs.		89,384	

Baseline model includes restaurant characteristics, competitive landscape, and macro conditions. Results are averaged over years and cross-validation iterations. Standard errors are provided in parentheses

[a] Best in the row or not significantly different from best in the row at the 0.05 level with a 2-sided test.

Yearly Results for the Main Prediction (Corresponding to Table 5)

Recall that we predict survival for out-of-sample restaurants in out-of-sample time periods in the main prediction of the paper. Table 5 shows the results aggregated by years and cross-validation iterations. Here Table A13 breaks down the prediction performance (AUC) by year. To calculate more accurate standard errors for each year, we employ 20-fold cross validation for each year here. We observe the pattern is consistent over the years in general.

Table A13 Yearly Out-of-Sample Time Periods and Restaurants Prediction Performance (AUC)

	2010	**2011**	**2012**	**2013**	**2014**	**2015**
Baseline (i.e., no UGC)	0.7217 (0.0143)	0.7158 (0.0118)	0.7103 (0.0076)	0.6945 (0.0093)	0.6946 (0.0105)	0.6660 (0.0065)
Baseline + review	0.7306 (0.0135)	0.7255 (0.0102)	**0.7305** (0.0062)	0.7119 (0.0092)	0.7197 [a] (0.0104)	0.6702 [a] (0.0072)
Baseline + photo	**0.7837** [a] (0.0105)	**0.8019** [a] (0.0092)	**0.7882** [a] (0.0056)	**0.7609** [a] (0.0075)	**0.7306** [a] (0.0082)	0.6847 [a] (0.0105)
Baseline + review + photo	**0.7898** [a] (0.0103)	**0.8063** [a] (0.0081)	**0.7943** [a] (0.0073)	**0.7742** [a] (0.0074)	**0.7414** [a] (0.0095)	**0.6909** [a] (0.0098)
Total obs.	26930	37745	49607	62299	75627	89384

Baseline model includes restaurant characteristics, competitive landscape, and macro conditions.
Results are averaged over cross-validation iterations. Standard errors are provided in parentheses.
Bold numbers indicate significant improvement over the baseline model at the 0.05 level with a 2-sided test.
[a] Best in the column or not significantly different from best in the column at the 0.05 level with a 2-sided test.

Robustness Check #1: Alternative Specifications for One-period and Cumulative Variables of Photos and Reviews

Tables A14 to A16 provide the predictive results based on three alternative variable specifications. It is worth noting that, while our main variable specification includes all restaurant observations (with the total number of observations being 89,384), these alternative variable specifications have fewer observations due to the inclusion of multi-year lags (i.e., $OnePeriod_{t-2}$, Cum_{t-2}, Cum_{t-3}, $Change_{t-1}$) in such specifications. Overall, our main variable specification has better or similar predictive performance compared to these alternative specifications. The incremental predictive power of photos is robust across these alternative modeling specifications.

Table A14 Out-of-Sample Time Periods and Restaurants Prediction Performance of Alternative Specification #1: $OnePeriod_{t-1}$+Cum_{t-2}

	Out of sample					In sample
	AUC	**KL divergence**	**Sensitivity**	**Specificity**	**Balanced accuracy**	**Pseudo R^2**
Baseline	0.6984	0.2001	0.7208	0.5402	0.6305	0.1344
	(0.0043)	(0.0041)	(0.0053)	(0.0109)	(0.0047)	(0.0025)
Baseline + review	**0.7135**	0.1981	**0.7993** [a]	0.4595	0.6294	**0.2287**
	(0.0046)	(0.0042)	(0.0048)	(0.0106)	(0.0045)	(0.0073)
Baseline + photo	**0.7606** [a]	0.1903 [a]	0.7308	**0.6210** [a]	**0.6759** [a]	**0.2495**
	(0.0065)	(0.0033)	(0.0085)	(0.0125)	(0.0061)	(0.0033)
Baseline + review + photo	**0.7662** [a]	**0.1880** [a]	**0.7629**	**0.6105** [a]	**0.6867** [a]	**0.2894** [a]
	(0.0063)	(0.0034)	(0.0082)	(0.0135)	(0.0060)	(0.0059)
Total obs.	71,665					

Baseline model includes restaurant characteristics, competitive landscape, and macro conditions. For sensitivity, specificity, and balanced accuracy, the training data are reweighted so that the total weights of surviving and closed observations are equal.
Results are averaged over years and cross-validation iterations. Standard errors are provided in parentheses.
Bold numbers indicate significant improvement over the baseline model at the 0.05 level with a 2-sided test.
[a] Best in the column or not significantly different from best in the column at the 0.05 level with a 2-sided test.

Table A15 Out-of-Sample Time Periods and Restaurants Prediction Performance of Alternative Specification #2: $OnePeriod_{t-1}$+$OnePeriod_{t-2}$+Cum_{t-3}

	Out of sample					In sample
	AUC	**KL divergence**	**Sensitivity**	**Specificity**	**Balanced accuracy**	**Pseudo R^2**
Baseline	0.6736	0.1903	0.7721	0.4314	0.6017	0.1366
	(0.0054)	(0.0045)	(0.0087)	(0.0137)	(0.0052)	(0.0037)
Baseline + review	**0.6933**	0.1888	**0.8650** [a]	0.3207	0.5928	**0.3109**
	(0.0044)	(0.0047)	(0.0068)	(0.0139)	(0.0048)	(0.0126)
Baseline + photo	**0.7465** [a]	0.1819 [b]	0.7676	**0.5418** [a]	**0.6547** [a]	**0.2896**
	(0.0071)	(0.0036)	(0.0109)	(0.0186)	(0.0067)	(0.0053)
Baseline + review + photo	**0.7484** [a]	**0.1806** [b]	**0.8204**	**0.4679**	**0.6442** [a]	**0.3592** [a]
	(0.0067)	(0.0039)	(0.0104)	(0.0204)	(0.0070)	(0.0105)
Total obs.	55,669					

Baseline model includes restaurant characteristics, competitive landscape, and macro conditions. For sensitivity, specificity, and balanced accuracy, the training data are reweighted so that the total weights of surviving and closed observations are equal.
Results are averaged over years and cross-validation iterations. Standard errors are provided in parentheses.
Bold numbers indicate significant improvement over the baseline model at the 0.05 level with a 2-sided test.
[a] Best in the column or not significantly different from best in the column at the 0.05 level with a 2-sided test.
[b] Best in the column or not significantly different from best in the column at the 0.10 level with a 2-sided test.

Table A16 Out-of-Sample Time Periods and Restaurants Prediction Performance of Alternative Specification #3: $OnePeriod_{t-1}+Cum_{t-1}+Change_{t-1}$

	Out of sample					In sample
	AUC	**KL divergence**	**Sensitivity**	**Specificity**	**Balanced accuracy**	**Pseudo R^2**
Baseline	0.6972	0.2005	0.7433	0.5131	0.6282	0.1438
	(0.0044)	(0.0041)	(0.0056)	(0.0091)	(0.0044)	(0.0029)
Baseline + review	**0.7152**	0.1987	**0.8171** [a]	0.4454	0.6312	**0.2518**
	(0.0044)	(0.0043)	(0.0050)	(0.0110)	(0.0043)	(0.0084)
Baseline + photo	**0.7683** [a]	**0.1891** [a]	0.7395	**0.6331** [a]	**0.6863** [a]	**0.2669**
	(0.0066)	(0.0031)	(0.0089)	(0.0112)	(0.0056)	(0.0036)
Baseline + review + photo	**0.7788** [a]	**0.1863** [a]	**0.7737**	**0.5991**	**0.6864** [a]	**0.3149** [a]
	(0.0065)	(0.0034)	(0.0085)	(0.0124)	(0.0057)	(0.0063)
Total obs.	71,665					

Baseline model includes restaurant characteristics, competitive landscape, and macro conditions. For sensitivity, specificity, and balanced accuracy, the training data are reweighted so that the total weights of surviving and closed observations are equal.
Results are averaged over years and cross-validation iterations. Standard errors are provided in parentheses.
Bold numbers indicate significant improvement over the baseline model at the 0.05 level with a 2-sided test.
[a] Best in the column or not significantly different from best in the column at the 0.05 level with a 2-sided test.

Robustness Check #2: Robustness Check for Photo and Review Content Measures

As a robustness check for photo and review content measures, we use content labels to capture photo and review content to carry out the comparisons in Table A17. Specifically, for photos, we replace "prop. of photos on each of the 10 LDA topics" in Table 4A in the paper with "prop. of photos containing each of the top 100 Clarifai objects" listed in Table A1. For reviews, we replace "prop. of reviews on each of the 20 LDA topics" in Table 4A with "prop. of reviews containing each of the top 100 nouns" listed in Table A10. Table A17 below provides a robustness check for Table 5. The results are qualitatively similar as the ones presented in the paper.

Table A17. Out-of-Sample Time Periods and Restaurants Performance of Prediction Using Content Labels on Photo and Review Content

	Out of sample					In sample
	AUC	**KL divergence**	**Sensitivity**	**Specificity**	**Balanced accuracy**	**Pseudo R^2**
Baseline (i.e., no UGC)	0.7020	0.1973	0.6484	0.6284	0.6384	0.1373
	(0.0047)	(0.0035)	(0.0056)	(0.0092)	(0.0046)	(0.0026)
Baseline + review	**0.7190**	0.1948	**0.7342** [a]	0.5487	0.6415	**0.2198**
	(0.0046)	(0.0035)	(0.0045)	(0.0092)	(0.0040)	(0.0060)
Baseline + photo	**0.7623** [a]	**0.1878** [a]	**0.6759**	**0.6936** [a]	**0.6848** [a]	**0.2358**
	(0.0064)	(0.0028)	(0.0086)	(0.0094)	(0.0052)	(0.0027)
Baseline + review + photo	**0.7708** [a]	**0.1850** [a]	**0.7103**	**0.6745** [a]	**0.6924** [a]	**0.2752** [a]
	(0.0061)	(0.0028)	(0.0074)	(0.0093)	(0.0050)	(0.0042)
Total obs.	89384					

Baseline model includes restaurant characteristics, competitive landscape, and macro conditions. For sensitivity, specificity, and balanced accuracy, the training data are reweighted so that the total weights of surviving and closed observations are equal.
Results are averaged over years and cross-validation iterations. Standard errors are provided in parentheses.
Bold numbers indicate significant improvement over the baseline model at the 0.05 level with a 2-sided test.
[a] Best in the column or not significantly different from best in the column at the 0.05 level with a 2-sided test.

Robustness Check #3: Prediction Survival Using Only Restaurants with Accurate Age Information

Table A18 is based on one-year-ahead predictions using the 10,368 restaurants with accurate age information (i.e., the birthdate is identified via each restaurant's Yelp/Facebook page, own website, or Google search engine, rather than approximated by the first review/photo date). The results are qualitatively the same as those in Table 5. Table A19 breaks down the prediction performance (AUC) by year. To calculate more accurate standard errors for each year, we employ 20-fold cross validation for each year. Again, we observe that the pattern is consistent over the years in general.

Table A18 Out-of-Sample Time Periods and Restaurants Performance of Prediction Using Restaurants with Accurate Age Information

	Out of sample					In sample
	AUC	**KL divergence**	**Sensitivity**	**Specificity**	**Balanced accuracy**	**Pseudo R^2**
Baseline (i.e., no UGC)	0.6450	0.1331	0.6568	0.4868	0.5718	0.1140
	(0.0089)	(0.0037)	(0.0128)	(0.0219)	(0.0073)	(0.0009)
Baseline + review	0.6669	0.1330	**0.7911** [a]	0.3801	0.5856	**0.2743**
	(0.0078)	(0.0037)	(0.0128)	(0.0243)	(0.0074)	(0.0066)
Baseline + photo	**0.7379** [a]	0.1266	**0.7345**	0.5180	**0.6263** [a]	**0.2651**
	(0.0083)	(0.0034)	(0.0122)	(0.0213)	(0.0087)	(0.0040)
Baseline + review + photo	**0.7404** [a]	0.1263	**0.7785** [a]	0.5023	**0.6404** [a]	**0.3338** [a]
	(0.0082)	(0.0033)	(0.0148)	(0.0265)	(0.0096)	(0.0042)
Total obs.	56391					

Baseline model includes restaurant characteristics, competitive landscape, and macro conditions. For sensitivity, specificity, and balanced accuracy, the training data are reweighted so that the total weights of surviving and closed observations are equal.
Results are averaged over years and cross-validation iterations. Standard errors are provided in parentheses.
Bold numbers indicate significant improvement over the baseline model at the 0.05 level with a 2-sided test.
[a] Best in the column or not significantly different from best in the column at the 0.05 level with a 2-sided test.

Table A19 Yearly Out-of-Sample Time Periods and Restaurants Performance (AUC) of Prediction Using Restaurants with Accurate Age Information

	2010	**2011**	**2012**	**2013**	**2014**	**2015**
Baseline (i.e., no UGC)	0.5802 (0.0329)	0.6359 (0.0197)	0.6851 (0.0179)	0.6696 (0.0135)	0.6478 (0.0233)	0.6371 (0.0173)
Baseline + review	0.6168 (0.0220)	0.6604 (0.0183)	0.6703 (0.0226)	0.6801 (0.0158)	0.6611 [a] (0.0183)	0.6627 [a] (0.0173)
Baseline + photo	**0.7355** [a] (0.0234)	**0.7820** [a] (0.0136)	**0.7862** [a] (0.0153)	**0.7408** [a] (0.0185)	0.7004 [a] (0.0173)	**0.6862** [a] (0.0138)
Baseline + review + photo	**0.7492** [a] (0.0199)	**0.7808** [a] (0.01580)	**0.7764** [a] (0.0169)	**0.7484** [a] (0.0160)	**0.7077** [a] (0.0180)	**0.6897** [a] (0.0168)
Total obs.	16045	22846	30413	38651	47347	56391

Baseline model includes restaurant characteristics, competitive landscape, and macro conditions.
Results are averaged over cross-validation iterations. Standard errors are provided in parentheses.
Bold numbers indicate significant improvement over the baseline model at the 0.05 level with a 2-sided test.
[a] Best in the column or not significantly different from best in the column at the 0.05 level with a 2-sided test.

Robustness Check #4: Prediction without Period 0

Table A20 is based on one-year-ahead predictions without period 0. As such, the 1,723 restaurants with only one period observation in our data are dropped from the analysis. The results are qualitatively the same as those in Table 5. Table A21 breaks down the prediction performance (AUC) by year. To calculate more accurate standard errors for each year, we employ 20-fold cross validation for each year. Again, we observe that the pattern is consistent over the years in general.

Table A20 Out-of-Sample Time Periods and Restaurants Performance of Prediction Without Period 0

	Out of sample					In sample
	AUC	**KL divergence**	**Sensitivity**	**Specificity**	**Balanced accuracy**	**Pseudo R^2**
Baseline (i.e., no UGC)	0.6973	0.2001	0.7247	0.5355	0.6301	0.1344
	(0.0051)	(0.0042)	(0.0049)	(0.0091)	(0.0046)	(0.0026)
Baseline + review	**0.7159**	0.1977	**0.8036** [a]	0.4636	0.6336	**0.2397**
	(0.0045)	(0.0042)	(0.0053)	(0.0106)	(0.0047)	(0.0086)
Baseline + photo	**0.7667** [a]	**0.1888** [a]	0.7310	**0.6377** [a]	**0.6843** [a]	**0.2608**
	(0.0064)	(0.0032)	(0.0089)	(0.0101)	(0.0052)	(0.0033)
Baseline + review + photo	**0.7746** [a]	**0.1864** [a]	**0.7628**	**0.6088**	**0.6858** [a]	**0.3077** [a]
	(0.0062)	(0.0033)	(0.0087)	(0.0102)	(0.0056)	(0.0063)
Total obs.	71665					

Baseline model includes restaurant characteristics, competitive landscape, and macro conditions. For sensitivity, specificity, and balanced accuracy, the training data are reweighted so that the total weights of surviving and closed observations are equal.
Results are averaged over years and cross-validation iterations. Standard errors are provided in parentheses.
Bold numbers indicate significant improvement over the baseline model at the 0.05 level with a 2-sided test.
[a] Best in the column or not significantly different from best in the column at the 0.05 level with a 2-sided test.

Table A21 Yearly Out-of-Sample Time Periods and Restaurants Prediction Performance (AUC) of Prediction Without Period 0

	2010	**2011**	**2012**	**2013**	**2014**	**2015**
Baseline (i.e., no UGC)	0.6926	0.7123	0.7078	0.6930	0.6905	0.6767
	(0.0183)	(0.0126)	(0.0104)	(0.0104)	(0.0120)	(0.0071)
Baseline + review	0.7097	0.7273	**0.7381**	0.7198	**0.7246** [a]	0.6901 [a]
	(0.0145)	(0.0116)	(0.0074)	(0.0118)	(0.0124)	(0.0063)
Baseline + photo	**0.7861** [a]	**0.8094** [a]	**0.7993** [a]	**0.7718** [a]	**0.7375** [a]	0.6972 [a]
	(0.0123)	(0.0082)	(0.0067)	(0.0093)	(0.0102)	(0.0099)
Baseline + review + photo	**0.7875** [a]	**0.8159** [a]	**0.8083** [a]	**0.7866** [a]	**0.7508** [a]	**0.6992** [a]
	(0.0121)	(0.0081)	(0.0080)	(0.0093)	(0.0111)	(0.0092)
Total obs.	16882	25724	35867	47042	58991	71665

Baseline model includes restaurant characteristics, competitive landscape, and macro conditions.
Results are averaged over cross-validation iterations. Standard errors are provided in parentheses.
Bold numbers indicate significant improvement over the baseline model at the 0.05 level with a 2-sided test.
[a] Best in the column or not significantly different from best in the column at the 0.05 level with a 2-sided test.

Robustness Check #5: Predicting Out-of-Sample Time Periods (No Cross-Validation between Restaurants)

In the main paper we predict survival in out-of-sample time periods for out-of-sample restaurants. As a robustness check, here we carry out an alternative forecasting scenario by using data up to period *t* -1 to predict the survival of all open restaurants in period *t*, which is likely to be one application of our algorithm in practice. Table A22 – A24 present the results. The results are qualitatively the same as those in the paper.

Table A22 Out-of-Sample Time Periods Prediction Performance

	Out of sample					In sample
	AUC	**Pseudo R^2**	**Sensitivity**	**Specificity**	**Balanced accuracy**	**Pseudo R^2**
Baseline (i.e., no UGC)	0.7010	0.1973	0.6455	0.6272	0.6364	0.1370
	(0.0080)	(0.0077)	(0.0159)	(0.0083)	(0.0047)	(0.0088)
Baseline + review	0.7165	0.1951	**0.7244** [a]	0.5648	0.6446	**0.2031**
	(0.0084)	(0.0076)	(0.0141)	(0.0166)	(0.0054)	(0.0185)
Baseline + photo	**0.7595** [a]	0.1880	0.6712 [a]	**0.7027** [a]	**0.6870** [a]	**0.2286**
	(0.0161)	(0.0043)	(0.0262)	(0.0119)	(0.0103)	(0.0089)
Baseline + review + photo	**0.7672** [a]	0.1854	**0.7017** [a]	**0.6768** [a]	**0.6892** [a]	**0.2598** [a]
	(0.0164)	(0.0044)	(0.0238)	(0.0145)	(0.0103)	(0.0133)
Total obs.				89384		

Baseline model includes restaurant characteristics, competitive landscape, and macro conditions. For sensitivity, specificity, and balanced accuracy, the training data are reweighted so that the total weights of open and closed observations are equal.
Results are averaged over years and cross-validation iterations. Standard errors are provided in parentheses.
Bold numbers indicate significant improvement over the baseline model at the 0.05 level with a 2-sided test.
[a] Best in the column or not significantly different from best in the column at the 0.05 level with a 2-sided test.

Table A23 Yearly Out-of-Sample Time Periods Prediction Performance (AUC)

	2010	**2011**	**2012**	**2013**	**2014**	**2015**
Baseline (i.e., no UGC)	0.7286	0.7161	0.7104	0.6901	0.6914	0.6696
Baseline + review	0.7378	0.7255	0.7299	0.7121	0.7197	0.6739
Baseline + photo	0.7875	0.8029	0.7854	0.7622	0.7325	0.6867
Baseline + review + photo	0.7953	0.8091	0.7908	0.7729	0.7463	0.6890
Total obs.	26930	37745	49607	62299	75627	89384

Baseline model includes restaurant characteristics, competitive landscape, and macro conditions.
Because there is no cross validation among restaurants, standard errors for each year are not available.

Table A24 Out-of-Sample Time Periods Prediction Performance – Different Aspects of Photos

	Out of sample					In sample
	AUC	**KL divergence**	**Sensitivity**	**Specificity**	**Balanced Accuracy**	**Pseudo R^2**
Baseline	0.7010	0.1973	0.6455	0.6272	0.6364	0.1370
	(0.0080)	(0.0077)	(0.0159)	(0.0083)	(0.0047)	(0.0088)
Baseline + photographic attributes	0.6996	0.1973	**0.6908** [a]	0.5772	0.6340	**0.1702**
	(0.0070)	(0.0078)	(0.0174)	(0.0145)	(0.0028)	(0.0118)
Baseline + photo caption	0.7018	0.1970	0.6668 [a]	0.6119	0.6393	0.1453
	(0.0082)	(0.0076)	(0.0140)	(0.0077)	(0.0053)	(0.0091)
Baseline + photo volume	0.7034	0.1968	0.6588 [a]	0.6170	0.6379	0.1381
	(0.0074)	(0.0078)	(0.0147)	(0.0139)	(0.0030)	(0.0077)
Baseline + helpful votes	0.7171 [a]	0.1960	0.6542 [a]	0.6427	**0.6485**	**0.1646**
	(0.0068)	(0.0073)	(0.0304)	(0.0243)	(0.0038)	(0.0107)
Baseline + photo content	**0.7525** [a]	0.1876	0.6672 [a]	**0.6957** [a]	**0.6815** [a]	**0.2013** [a]
	(0.0203)	(0.0044)	(0.0184)	(0.0126)	(0.0153)	(0.0070)
Total obs.				89384		

Baseline model includes restaurant characteristics, competitive landscape, and macro conditions. For sensitivity, specificity, and balanced accuracy, the training data are reweighted so that the total weights of surviving and closed observations are equal.
Results are averaged over years and cross-validation iterations. Standard errors are provided in parentheses.
Bold numbers indicate significant improvement over the baseline model at the 0.05 level with a 2-sided test.
[a] Best in the column or not significantly different from best in the column at the 0.05 level with a 2-sided test.

Robustness Check #6: Prediction without Age 0 Restaurant-Year Observations

Although all restaurants don't have UGC in period zero, such zero UGC can have different meanings for new and existing restaurants. For example, restaurant A was founded in 2010 (restaurant A's period 1), and thus it was not eligible to receive any UGC in 2009 (age = 0 in restaurant A's period 0). Differently, restaurant B was founded in 2008, but it received its first review in 2010 (restaurant B's period 1) because it did not receive any UGC in 2009 (age = 2 in restaurant B's period 0). In such cases, while both restaurants have zero UGC in period 0, the reasons are different. Hence, we conduct another robustness check by dropping observations with age = 0. In such cases, restaurant A's age = 0 in period 0 observation is dropped because it cannot receive UGC in period 0. But restaurant B's age = 2 in period 0 observation remains because zero UGC is meaningful to be accounted for in such cases. Table A25 presents the results which are qualitatively the same as those in Table 5.

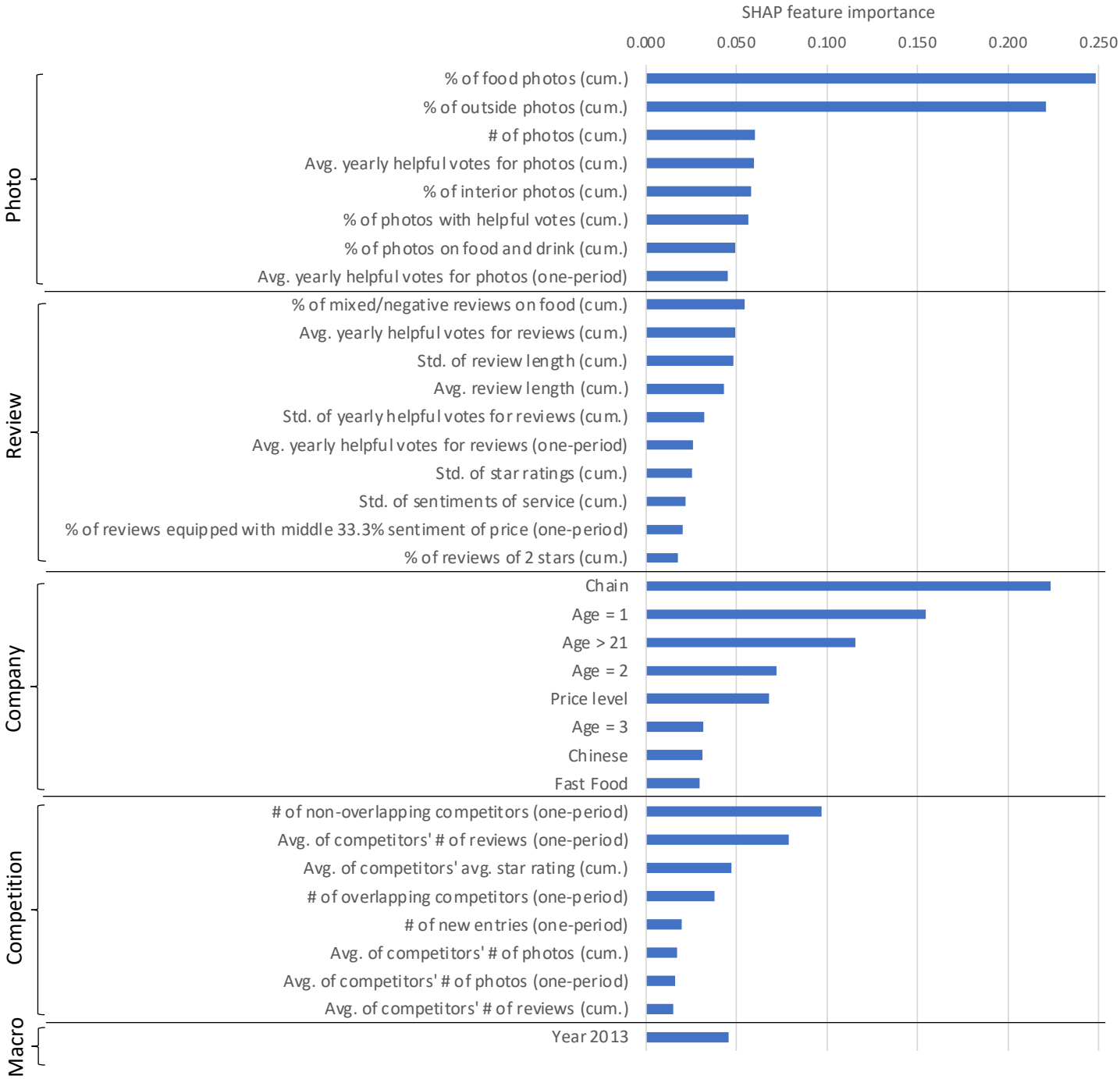
Figure A13 presents the top 35 predictors of restaurant survival in one-year-ahead prediction without age 0. The top predictors are similar to those Figure 4 of the paper in general. In particular, Figure A13 shows that the eight photo-related variables among the top 35 predictors are the same as those in Figure 4 of the paper. Again, the most predictive photos variables, % of food photos and % of outside photos, are related to photo content. The variable, average yearly helpful votes for photos, is also very predictive of restaurant survival. Moreover, after dropping observations with age=0, the SHAP feature importance values increase for these top photo variables overall. This is intuitive. For example, the SHAP feature importance of % food photos increases from 0.213 (Figure 4 of the paper) to 0.248 (Figure A13).

Table A25 Out-of-Sample Time Periods and Restaurants Performance of Prediction Without Age 0

	Out of sample					In sample
	AUC	**KL divergence**	**Sensitivity**	**Specificity**	**Balanced accuracy**	**Pseudo R^2**
Baseline (i.e., no UGC)	0.7070	0.1908	0.6828	0.5977	0.6403	0.1513
	(0.0054)	(0.0036)	(0.0062)	(0.0100)	(0.0050)	(0.0031)
Baseline + review	**0.7244**	0.1884	**0.7658** [a]	0.5295	0.6476	**0.2456**
	(0.0050)	(0.0035)	(0.0057)	(0.0101)	(0.0047)	(0.0078)
Baseline + photo	**0.7712** [a]	**0.1809** [a]	**0.7136**	**0.6678** [a]	**0.6907** [a]	**0.2692**
	(0.0068)	(0.0027)	(0.0099)	(0.0098)	(0.0053)	(0.0038)
Baseline + review + photo	**0.7801** [a]	**0.1778** [a]	**0.7418**	**0.6552** [a]	**0.6985** [a]	**0.3129** [a]
	(0.0067)	(0.0027)	(0.0088)	(0.0105)	(0.0063)	(0.0060)
Total obs.	80390					

Baseline model includes restaurant characteristics, competitive landscape, and macro conditions. For sensitivity, specificity, and balanced accuracy, the training data are reweighted so that the total weights of surviving and closed observations are equal.
Results are averaged over years and cross-validation iterations. Standard errors are provided in parentheses.
Bold numbers indicate significant improvement over the baseline model at the 0.05 level with a 2-sided test.
[a] Best in the column or not significantly different from best in the column at the 0.05 level with a 2-sided test.

Figure A13. Top 35 Predictors of Restaurant Survival in One-Year-Ahead Prediction without Age 0



Importance weights are based on predicting survival in 2015.
Within each type of factors (e.g., photo), variables are ordered by their predictive power.

Robustness Check #7: Prediction with $dummy_{pre2004\ based\ on\ UGC}$

The first photo/review on Yelp is truncated in year 2004. As such, when birthdate = 2004 is inferred from the year of the first photo/review on Yelp, such age information might be less accurate. Hence, we add a dummy (1 when birthyear=2004 based on UGC, 0 otherwise) as an additional variable in the model as a robustness check. We observe that only eight restaurants started in 2004 in the entire sample. Among these, four restaurants have their birthdates inferred from Yelp reviews or photos. And these four restaurants only have 36 restaurant-year observations in total. Table A26 below presents the results of the robustness check. The results here are qualitatively consistent with to those in Table 5 of the paper.

Table A26 Out-of-Sample Time Periods and Restaurants Performance of Prediction With $dummy_{pre2004\ based\ on\ UGC}$

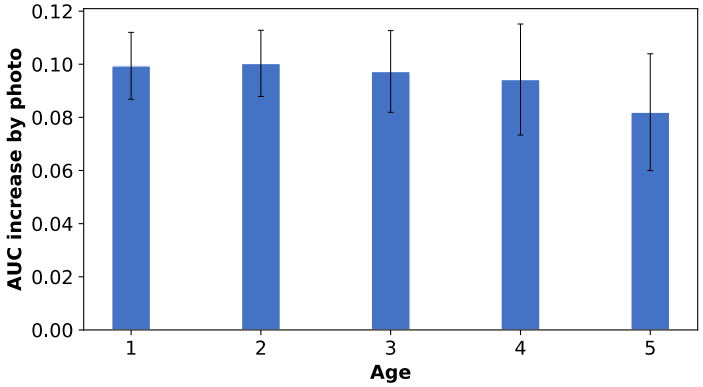
	Out of sample					In sample
	AUC	**KL divergence**	**Sensitivity**	**Specificity**	**Balanced accuracy**	**Pseudo R^2**
Baseline (i.e., no UGC)	0.6991	0.1975	0.6496	0.6321	0.6408	0.1376
	(0.0048)	(0.0037)	(0.0061)	(0.0097)	(0.0045)	(0.0026)
Baseline + review	**0.7144**	0.1956	**0.7274** [a]	0.5657	0.6465	0.2103
	(0.0046)	(0.0036)	(0.0056)	(0.0099)	(0.0046)	(0.0063)
Baseline + photo	**0.7583** [a]	**0.1881** [a]	**0.6740**	**0.6923** [a]	**0.6831** [a]	0.2334
	(0.0065)	(0.0029)	(0.0090)	(0.0090)	(0.0051)	(0.0029)
Baseline + review + photo	**0.7657** [a]	**0.1856** [a]	**0.7032**	**0.6787** [a]	**0.6910** [a]	**0.2671** [a]
	(0.0063)	(0.0029)	(0.0081)	(0.0089)	(0.0053)	(0.0047)
Total obs.						

Baseline model includes restaurant characteristics, competitive landscape, and macro conditions. For sensitivity, specificity, and balanced accuracy, the training data are reweighted so that the total weights of surviving and closed observations are equal.
Results are averaged over years and cross-validation iterations. Standard errors are provided in parentheses.
Bold numbers indicate significant improvement over the baseline model at the 0.05 level with a 2-sided test.
[a] Best in the column or not significantly different from best in the column at the 0.05 level with a 2-sided test.

Photos' Predictive Power Broken Down by Each Age for Age <=5

Because the survival of younger restaurants is more volatile[5], we take a closer look at photos' predictive power by each age for age<=5 in Figure A13. We calculate the AUC increase from photos by examining the difference between $AUC_{Baseline+\,review\,+\,photo}$ and $AUC_{Baseline+\,review}$ for each age group, respectively.[6] The figure shows that photos increase prediction accuracy for restaurants at each age younger than or equal five years old.

Figure A14 Photo's Prediction Power by Each Age (Age<=5)



Robustness Checks for Multiple-Year Survival Prediction

We report results from three alternative specifications for multiple-year (Δt) survival prediction as robustness checks below. We let Δt=1,2,3 years. Compared with the main specification in the paper where we predict survival during the future Δt years (e.g., forecasting whether a restaurant would survive until the end of 2013), these alternative specifications break down the prediction for each year (e.g., predicting whether a restaurant would survive in 2011, 2012, or 2013). To make these specifications comparable with the main specification in the paper, we multiply the predicted survival probability for each future year to derive the predicted survival probability during the future Δt years (e.g., $\widehat{y_{i2010+3}} = \hat{y}_{i2011} * \hat{y}_{i2012} * \hat{y}_{i2013}$).

Same for all the approaches below, for each Δt, we train separate models (baseline; baseline + review; baseline + photo; baseline + review + photo) with data till t and 10-fold cross-validation. Then for each Δt, we report average performance across t and cross-validation iterations.

First specification We simply use $\boldsymbol{X}_{it}$ to predict $y_{it+\Delta t}$. For example, we use all information up till year 2010 to predict survival probability in year 2013, with no additional assumptions made.

Second specification We predict $y_{it+\Delta t}$, assuming $x_{i,t+\Delta t-1} = \cdots = x_{i,t+1} = x_{it}$ for all one-period variables, with cumulative variables in future periods derived from the cumulative variables in the current period and the respective one-period variables in future periods. For example, we assume # of photos$_{it}$ (one-period) = # of photos$_{it-1}$ (one-period). And # of photos$_{it}$ (cum.) = # of photos$_{it-1}$ (cum.) + # of photos$_{it}$ (one-period).

Third specification This specification is based on the distributed lag model (Mela et al. 1997). In this model, the dependent variable $y_{it+\Delta t}$ is a function of $\boldsymbol{X}_{it+\Delta t-1}$ and lagged dependent variable $y_{it+\Delta t-1}$. We do not observe $\boldsymbol{X}_{it+\Delta t-1}$ and $y_{it+\Delta t-1}$ in current year t, when $\Delta t > 1$. Therefore, we use

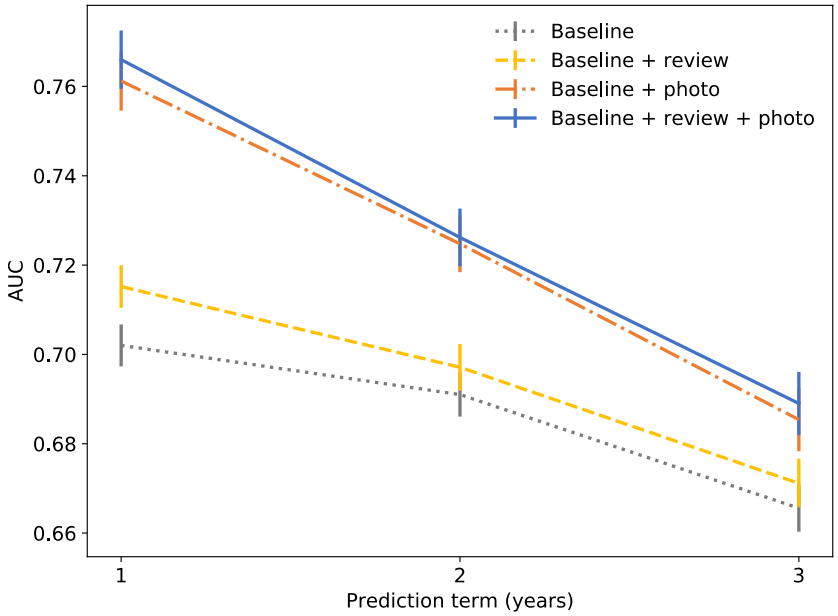
[5]The failure rates for the three groups (1<=age<=3; 3<age<=21; age>21) in Table 7 are 10%, 4%, and 2%, respectively. Please note that the failure rates are calculated at the restaurant-year level (instead of restaurant level) since restaurant age changes over the years.

[6] In Table 7, we train a separate model for each age bracket (young, mid-aged, established) to fully calibrate the model for each type of restaurants. Because each age in Figure A13 has a relatively small number of observations, we do not train a separate model for each age. Instead, we plot the AUC difference for each age separately, using the model trained on all ages.

$\boldsymbol{X}_{it}$ to predict y_{it+1}, use $\widehat{y_{it+1}}$ and $\widehat{\boldsymbol{X}_{it+1}}$ to predict y_{it+2}, use $\widehat{y_{it+2}}$ and $\widehat{\boldsymbol{X}_{it+2}}$ to predict y_{it+3}, with $\widehat{\boldsymbol{X}_{it+1}}$ and $\widehat{\boldsymbol{X}_{it+2}}$ derived in the same fashion as in the second specification above.

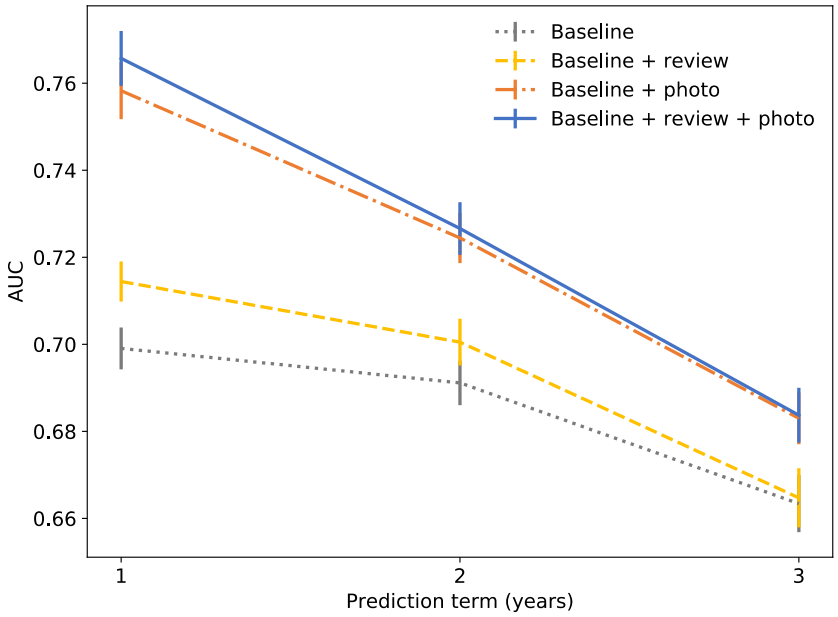
Figures A14-A16 show that, under all these three alternative specifications, photos can predict three-year survival while reviews can only predict one-year survival. It is consistent with the pattern of main prediction in the paper.

Figure A15 Multiple-Year Survival Prediction
(First Specification)



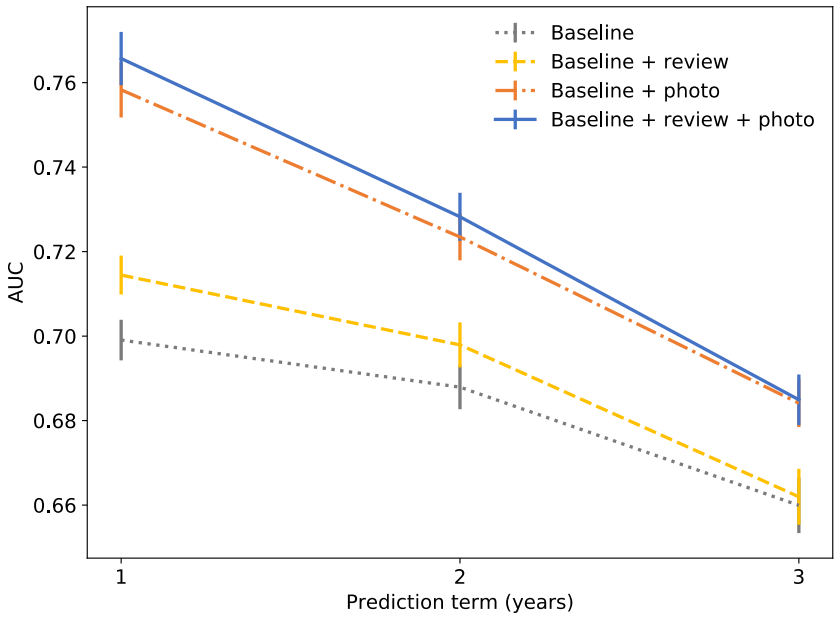
Baseline includes restaurant characteristics, competition landscape, and macro factors.
Results are averaged over years and cross-validation iterations.
The error bars represent ± 1 times the standard error of each point estimate.

Figure A16 Multiple-Year Survival Prediction Assuming $x_{i,t+1} = x_{it}$ for All One-Period Variables
(Second Specification)



Baseline includes restaurant characteristics, competition landscape, and macro factors.
Results are averaged over years and cross-validation iterations.
The error bars represent ± 1 times the standard error of each point estimate.

Figure A17 Multiple-Year Survival Prediction Using Distributed Lag Model
(Third Specification)

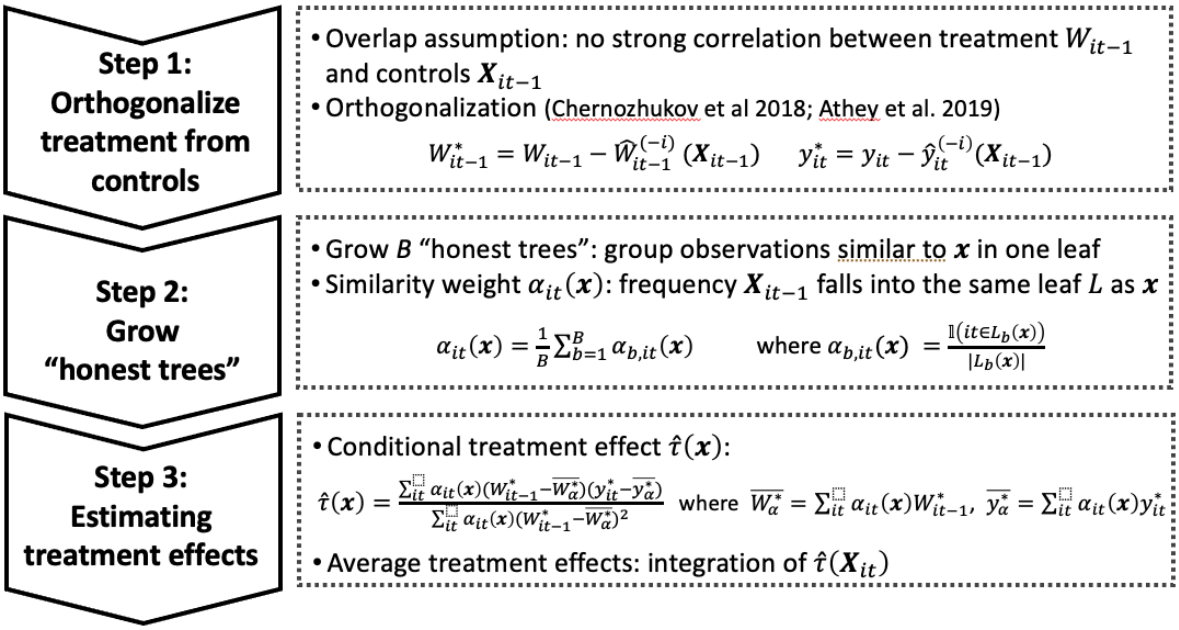


Baseline includes restaurant characteristics, competition landscape, and macro factors.
Results are averaged over years and cross-validation iterations.
The error bars represent ± 1 times the standard error of each point estimate.

Appendix E. Cluster-Robust Causal Forests

Figure A17 summarizes the estimation procedures of our causal forests. We provide more details for the estimation procedures below.

Figure A18 Causal Forest Estimation Procedures



$\boldsymbol{x}$ denotes a set of values of $\boldsymbol{X}_{it-1}$.

Consolidating Highly Correlated Variables

Before carrying out the causal forests estimation, we consolidate variables to reduce the correlations between treatments and controls as required by the overlap assumption (Wager and Athey 2018). Table A25 presents the complete list of consolidated variables. When both cumulative and one-period variables of the same measure (e.g., cum. and one-period avg. yearly helpful votes for photos) are among the top 35 most informative variables in Figure 4 of the main paper, we use cumulative rather than one-period lag variables because the former often times are more predictive of restaurant survival than the latter. When two variables carry similar meanings (e.g., prop. of photos with helpful votes, avg. yearly helpful votes for photos), we keep the one with higher SHAP feature importance from Figure 4 in the main paper. When the average and the standard deviation of the same measure are highly correlated (e.g., avg. and std. of yearly helpful votes for reviews), we keep the average for ease of interpretation.

Table A27 The List of Consolidated Variables

Variables before consolidation	Treatment variable used in causal forests
# of photos (cum.) # of reviews (cum.)	# of photos and reviews (cum.)
Prop. of photos with helpful votes (cum.) Avg. yearly helpful votes for photos (cum.) Avg. yearly helpful votes for photos (one-period)	Prop. of photos with helpful votes (cum.)
Avg. yearly helpful votes for reviews (cum.) Std. of yearly helpful votes for reviews (cum.) Avg. yearly helpful votes for reviews (one-period)	Avg. yearly helpful votes for reviews (cum.)
Avg. review length (cum.) Std. of review length (cum.)	Avg. review length (cum.)
# of overlapping competitors (one-period) # of non-overlapping competitors (one-period) # of new entries (one-period) # of new exits (one-period)	# of competitors (one-period)
Avg. of competitors' # of photos (cum.) Avg. of competitors' # of reviews (cum.) Avg. of competitors' # of reviews (one-period)	Avg. of competitors' # of photos and reviews (cum.)

Correlations Between Treatment and Control Variables Before and After Orthogonalization

We carry out an orthogonalization procedure (Step 1 in Figure A17) to further reduce correlations between treatment variables and controls. We report correlations between treatment and control variables before and after orthogonalization. The idea of examining such correlations is similar to the matching quality check under traditional propensity score matching methods (Austin 2009). Table A26 displays the maximum correlation coefficient between each treatment variable and their controls before and after orthogonalization. For example, when the cumulative number of photos and reviews is the treatment variable, we check its correlation with each control variable (i.e., all other treatment variables in the second column of Table A26, restaurant quality dimensions, zip codes, and years). Before orthogonalization, the maximum correlation between cumulative UGC volume and controls was 0.40. After orthogonalization, the maximum correlation is reduced to 0.08. It is evident that the correlations between the treatment and control variables are greatly reduced after orthogonalization.

As a robustness check, we also double-check the quality of the propensity scores $\widehat{W}_{it-1}^{(-i)}(\boldsymbol{X}_{it-1})$ (the second term of the first part of Equation 3 in the paper) generated in the orthogonalization step. We choose to perform this robustness check on binary treatment variables, given that such a conventional balance check method is only applicable to binary variables. The intuition is, if the quality of the propensity scores is good, we would see that the control variables adjusted by the propensity scores are balanced across the treated and untreated conditions. For example, suppose that the treatment variable is chain status and that one control variable is age. To check the extent to which age is balanced in the observations of chain and independent restaurants, we compute the normalized balance diagnostic d (Austin 2011) for age as in Equation (A2) before and after the adjustment (d_{raw}, $d_{adjusted}$). $\overline{age}_D$ is the average of age and ${s_{age}}_D^2$ is the variance of age for observations of $D \in (chain, independent)$, respectively. Following Austin (2011) , we define weight for adjustment as $\varphi_{it-1} = \left[\frac{chain_i}{\widehat{chain}_{it-1}^{(-i)}(\boldsymbol{X}_{it-1})} + \right.$

$\frac{1-chain_i}{1-\widehat{chain}_{it-1}^{(-i)}(\boldsymbol{X}_{it-1})}\Big]$[7]. Then following Austin and Stuart (2015), $\overline{age}_{D,adjusted} = \frac{\sum_{i \in D} \varphi_{it} age_{it}}{\sum_{i \in D} \varphi_{it}}$ and $s_{age}{}^2_{D,adjusted} = \frac{\sum_{i \in D} \varphi_{it}}{(\sum_{i \in D} \varphi_{it})^2 - \sum_{i \in D} \varphi_{it}^2} \sum_{i \in D} \varphi_{it} \left(age_{it} - \overline{age}_{D,adjusted}\right)^2$, with $D \in (chain, independent)$. The normalized balance diagnostic d is calculated for other control variables in the same fashion.

(A2) $$\boldsymbol{d_{raw}} = \frac{\overline{\boldsymbol{age}}_{\boldsymbol{chain}} - \overline{\boldsymbol{age}}_{\boldsymbol{independent}}}{\sqrt{\frac{s_{\boldsymbol{age}}^2{}_{\boldsymbol{chain}} + s_{\boldsymbol{age}}^2{}_{\boldsymbol{independent}}}{2}}}, \quad \boldsymbol{d_{adjusted}} = \frac{\overline{\boldsymbol{age}}_{\boldsymbol{chain,adjusted}} - \overline{\boldsymbol{age}}_{\boldsymbol{independent,adjusted}}}{\sqrt{\frac{s_{\boldsymbol{age}}^2{}_{\boldsymbol{chain,adjusted}} + s_{\boldsymbol{age}}^2{}_{\boldsymbol{independent,adjusted}}}{2}}}$$

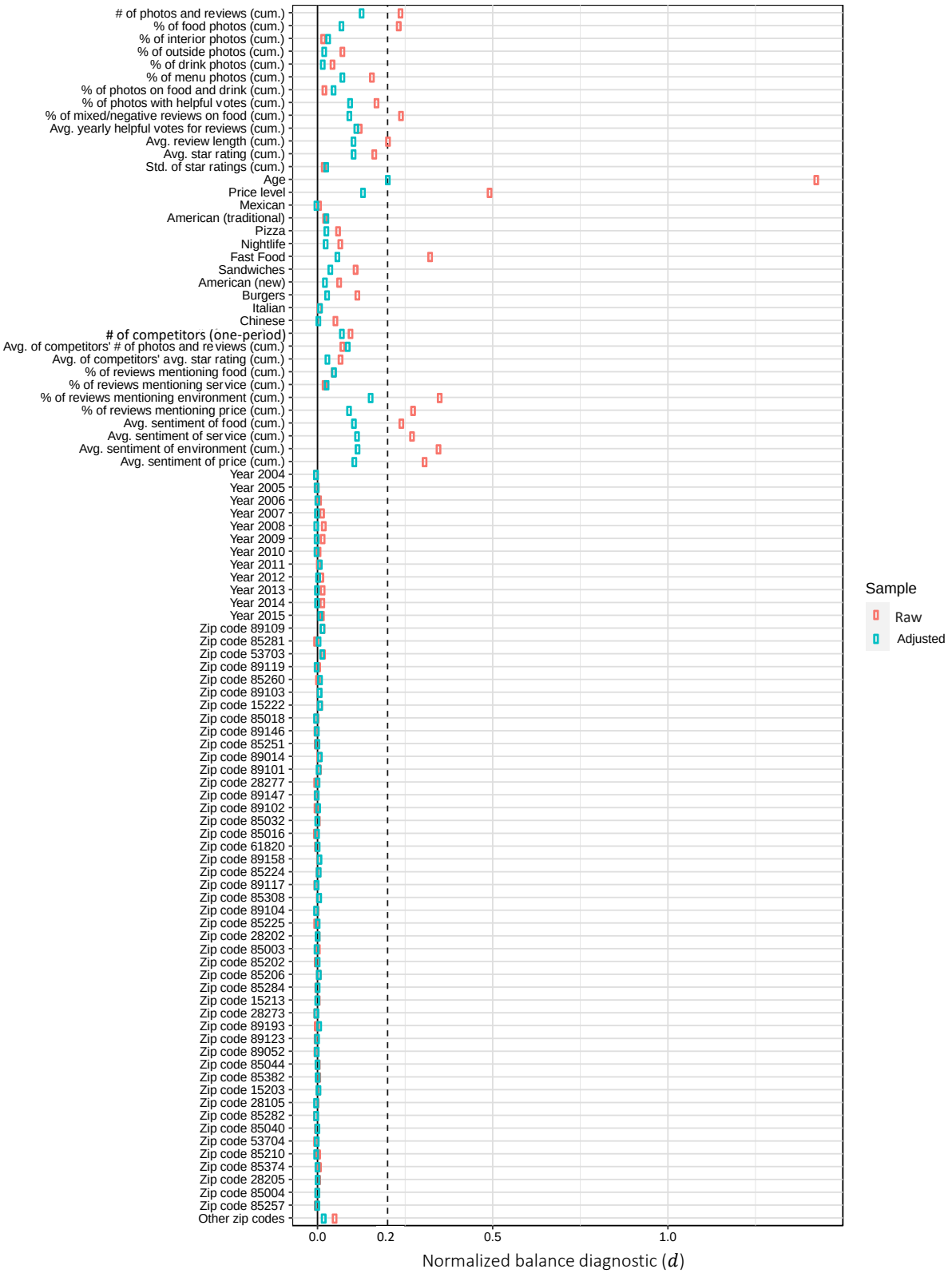
Figure A18 visualizes d_{raw} and $d_{adjusted}$ when chain status is the treatment variable, where the x-axis is the normalized balance diagnostic d and y-axis lists names of all control variables. As per Austin (2009), $d \leq 0.2$ indicates adequate balance in control variables between treated (chain restaurants) and untreated (independent restaurants) groups. For example, in Figure 18, d_{raw} for age is bigger than 0.2, while the $d_{adjusted}$ for age is smaller than 0.2, indicating that the control variable age is balanced across the chain and independent restaurants after being adjusted by the propensity scores generated by the causal forests. Figure 18 shows $d_{adjusted} \leq 0.2$ for all control variables.

[7] This weight is conventionally used by inverse probability of treatment weighting methods to estimate the average treatment effect (ATE), where $\widehat{chain}_{it-1}^{(-i)}(\boldsymbol{X}_{it-1})$ is the probability of being treated (chain restaurants in this example) and $1 - \widehat{chain}_{it-1}^{(-i)}(\boldsymbol{X}_{it-1})$ is the probability of being untreated (independent restaurants in this example) (Austin 2011). Namely, each sample's weight is equal to the inverse of the probability of receiving the condition the subject received.

Table A28 Correlations between Treatment and Control Variables Before and After Orthogonalization

Type	Treatment variable	Maximum correlation coefficient between treatment and controls before orthogonalization	Maximum correlation coefficient between treatment and controls after orthogonalization
Photo and review	# of photos and reviews (cum.)	0.40	0.08
Photo	% of food photos (cum.)	0.36	0.07
	% of outside photos (cum.)	0.23	0.06
	% of interior photos (cum.)	0.19	0.06
	% of dink photos (cum.)	0.16	0.03
	% of menu photos (cum.)	0.09	0.04
	% of photos on food and drink (cum.)	0.36	0.05
	% of photos with helpful votes (cum.)	0.31	0.04
Review	% of mixed/negative reviews on food (cum.)	0.23	0.04
	Avg. yearly helpful votes for reviews (cum.)	0.20	0.06
	Avg. review length (cum.)	0.23	0.08
	Avg. star rating (cum.)	0.32	0.05
	Std. of star ratings (cum.)	0.32	0.04
Company	Chain	0.56	0.05
	Age <= 1	0.26	0.04
	Age 2-3	0.21	0.03
	Age 4-7	0.21	0.03
	Age 8-21	0.11	0.03
	Age 22-42	0.24	0.03
	Age > 42	0.50	0.05
	Price level	0.28	0.09
	Mexican	0.14	0.03
	American (traditional)	0.25	0.05
	Pizza	0.33	0.05
	Nightlife	0.27	0.06
	Fast Food	0.47	0.04
	Sandwiches	0.17	0.04
	American (new)	0.20	0.05
	Burgers	0.29	0.03
	Italian	0.33	0.03
	Chinese	0.14	0.05
Competition	# of competitors (one-period)	0.58	0.20
	Avg. of competitors' # of photos and reviews (cum.)	0.55	0.19
	Avg. of competitors' avg. star rating (cum.)	0.23	0.06

Figure A19 An Example of Control Variable Balance Check before and after Orthogonalization (Treatment Variable: Chain)



y-axis lists names of all control variables when chain status is the treatment variable.

Building Honest Trees

In Step 2 of Figure A17, we follow Athey et al. (2019) by building honest trees to weight the similarity between a set of values of controls $\boldsymbol{x}$ and an arbitrary $\boldsymbol{X}_{it-1}$. For a consistent estimation of the treatment effects, honest trees use separate subsamples of observations for placing the splits and calculating similarity weights. Let us grow B (B=2000) trees, with each tree (indexed by b) built with a random subsample of observations $\mathcal{S}_b$ and a random subsample of control variables. Each tree aims to group observations $\boldsymbol{X}_{it-1}$ similar to $\boldsymbol{x}$ in one leaf. Denote the leaf containing $\boldsymbol{x}$ as $L_b(\boldsymbol{x})$. The similarity between observation $\boldsymbol{X}_{it-1}$ and $\boldsymbol{x}$ measured by tree b is thus defined as $\alpha_{b,it}(\boldsymbol{x}) = \frac{\mathbb{I}(\boldsymbol{X}_{it-1} \in L_b(\boldsymbol{x}))}{|L_b(\boldsymbol{x})|}$. Then, the overall similarity between $\boldsymbol{X}_{it-1}$ and $\boldsymbol{x}$, denoted by $\alpha_{it}(\boldsymbol{x})$, captures the frequency the $\boldsymbol{X}_{it-1}$ observation falls into the same leaf as $\boldsymbol{x}$: $\alpha_{it}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} \alpha_{b,it}(\boldsymbol{x})$.

As discussed in the paper, our cluster-robust causal forests account for within-restaurant variation by allowing clustered errors. Technically, to allow errors to be correlated within clusters, we first draw a subsample of clusters $I_b \subseteq \{1,2,\dots,I\}$ when building each tree; then $\mathcal{S}_b$ is formed by all observations from the clusters I_b.[8]

Robustness Check for Causal Forests Estimation

Related to robustness check #6 for the restaurant survival predictive model in Appendix D, we re-estimate the causal forests by dropping age 0. Table A28 below presents the causal forests estimates for treatment variables. The results are qualitatively consistent with those in Table 8 of the paper. In particular, for photo-related treatment variables, the significance and directions of effects for variables with a significant effect are the same as those in Table 8 of the paper. Additionally, magnitudes of effects for photo variables with a significant effect are similar to those in Table 8 of the paper.

[8] https://grf-labs.github.io/grf/REFERENCE.html#cluster-robust-estimation

Table A29. Photo-Related Results of Cluster-Robust Causal Forests without Age=0 Restaurant-Year Observations

	Treatment variable	Parameter Estimate	SE
Photo and review	# of photos and reviews (cum.)	-0.0284	0.0147
Photo	% of food photos (cum.)	0.0480 ***	0.0056
	% of outside photos (cum.)	0.0305 ***	0.0092
	% of interior photos (cum.)	0.0348 **	0.0121
	% of dink photos (cum.)	-0.0503	0.1257
	% of menu photos (cum.)	0.0095	0.0566
	% of photos on food and drink (cum.)	-0.0098	0.0130
	% of photos with helpful votes (cum.)	0.0540 ***	0.0037
Review	% of mixed/negative reviews on food (cum.)	-0.0747 ***	0.0130
	Avg. yearly helpful votes for reviews (cum.)	0.0344 ***	0.0042
	Avg. review length (cum.)	-0.0036 ***	0.0005
	Avg. star rating (cum.)	0.0051 **	0.0016
	Std. of star ratings (cum.)	-0.0238 ***	0.0046
Company	Chain	0.0226 ***	0.0024
	Age <= 1	-0.0758 ***	0.0066
	Age 2-3	-0.0266 ***	0.0032
	Age 4-7	0.0011	0.0027
	Age 8-21	0.0236 ***	0.0017
	Age 22-42	0.0201 ***	0.0020
	Age > 42	0.0209 ***	0.0026
	Price level	-0.0067 ***	0.0017
	Mexican	0.0152 ***	0.0028
	American (traditional)	0.0068 **	0.0025
	Pizza	0.0023	0.0030
	Nightlife	0.0142 ***	0.0028
	Fast Food	0.0094 ***	0.0029
	Sandwiches	0.0049	0.0028
	American (new)	-0.0017	0.0034
	Burgers	-0.0007	0.0031
	Italian	-0.0001	0.0032
	Chinese	0.0329 ***	0.0031
Competition	# of competitors (one-period)	-0.0141 ***	0.0036
	Avg. of competitors' # of photos and reviews (cum.)	-0.0658 ***	0.0112
	Avg. of competitors' avg. star rating (cum.)	-0.0166 *	0.0079
Additional controls	Restaurant quality dimensions including: 1) Prop. of reviews mentioning each restaurant quality dimension (food, service, environment, price) (cum.) 2) Avg. sentiment of each dimension (cum.) Zip codes Year dummies		

*** p<0.001, ** p<0.01, * p<0.05. A positive sign means positive association with restaurant survival. Each row is a separate cluster-robust causal forests estimation. We rescaled "avg. review length (cum.)," "# of competitors (one-period)," and "avg. of competitors' # of photos and reviews (cum.)" by 1/100. Avg. sentiment of each quality dimension is not included as controls when "avg. star rating" is the treatment variable due to high correlation.

References for Appendices

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *Annals of Statistics* 47(2):1148–78.

Austin, Peter C. 2009. "Balance Diagnostics for Comparing the Distribution of Baseline Covariates between Treatment Groups in Propensity-Score Matched Samples." *Statistics in Medicine* 28(25):3083--3107.

Austin, Peter C. 2011. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behavioral Research* 46(3):399–424.

Austin, Peter C., and Elizabeth A. Stuart. 2015. "Moving towards Best Practice When Using Inverse Probability of Treatment Weighting (IPTW) Using the Propensity Score to Estimate Causal Treatment Effects in Observational Studies." *Statistics in Medicine* 34(28):3661–79.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3(Jan):993–1022.

Bujisic, Milos, Joe Hutchinson, and H. G. Parsa. 2014. "The Effects of Restaurant Quality Attributes on Customer Behavioral Intentions." *International Journal of Contemporary Hospitality Management* 26(8):1270–91.

Canny, John. 1987. "A Computational Approach to Edge Detection." *Readings in Computer Vision* (184–203).

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* 785–94.

Datta, Ritendra, Dhiraj Joshi, Jia Li, and James Z. Wang. 2006. "Studying Aesthetics in Photographic Images Using a Computational Approach." *European Conference on Computer Vision* 288–301.

Daubechies, Ingrid. 1992. *Ten Lectures on Wavelets*. Siam.

Friedman, Jerome. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 1189–1232.

Hanley, James A., and Barbara J. McNeil. 1982. "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve." *Radiology* 143(1):29–36.

Hasler, David, and Sabine E. Suesstrunk. 2003. "Measuring Colorfulness in Natural Images." *Human Vsion and Electronic Imaging VIII* 5007(87–96).

Hoffman, Matthew, Francis Bach, and David Blei. 2010. "Online Learning for Latent Dirichlet Allocation." *Advances in Neural Information Processing Systems* 23:856–64.

Hutto, Clayton J., and Eric Gilbert. 2014. "Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Eighth International AAAI Conference on Weblogs and Social Media*.

Hyun, Sunghyup Sean. 2010. "Predictors of Relationship Quality and Loyalty in the Chain Restaurant Industry." *Cornell Hospitality Quarterly* 51(2):251–67.

Kim, Yoon. 2014. "Convolutional Neural Networks for Sentence Classification." *ArXiv Preprint ArXiv:1408.5882*.

LeCun, Yann, Leon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. "Gradient-Based Learning Applied to Document Recognition." *Proceedings of the IEEE* 2278–2324.

Liu, Xiao, Dokyun Lee, and Kannan Srinivasan. 2019. "Large Scale Cross Category Analysis of Consumer Review Content on Sales Conversion Leveraging Deep Learning." *Journal of Marketing Research*.

Mela, Carl F., Sunil Gupta, and Donald R. Lehmann. 1997. "The Long-Term Impact of Promotion and Advertising on Consumer Brand Choice." *Journal of Marketing Research* XXXIV:248–61.

Montabone, Sebastian, and Alvaro Soto. 2010. "Human Detection Using a Mobile Platform and Novel Features Derived from a Visual Saliency Mechanism." *Image and Vision Computing* 28(3):391–402.

Mori, Greg, Xiaofeng Ren, Alexei A. Efros, and Jitendra Malik. 2004. "Recovering Human Body Configurations: Combining Segmentation and Recognition." *Proceedings of the 2004 IEEE*

Computer Society Conference on Computer Vision and Pattern Recognition 2:II--II.
Netzer, Oded, Alain Lemaire, and Michal Herzenstein. 2019. "When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications." *Journal of Marketing Research* 56(6):960–80.
Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "Glove: Global Vectors for Word Representation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* 1532–43.
Rehurek, Radim, and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora." *Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*.
Ren, Xiaofeng, and Jitendra Malik. 2003. "Learning a Classification Model for Segmentation." *Proceedings of 9th International Conference of Computer Vision* 1:10–17.
Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. "Exploring the Space of Topic Coherence Measures." *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* 399--408.
Rother, Carsten, Vladimir Kolmogorov, and Andrew Blake. 2004. "Grabcut: Interactive Foreground Extraction Using Iterated Graph Cuts." *ACM Transactions on Graphics (TOG)* 23(3):309–14.
Ruder, Sebastian. 2017. "An Overview of Multi-Task Learning in Deep Neural Networks." *ArXiv Preprint ArXiv:1706.05098*.
Ryu, Kisang, Hye-Rin Lee, and Woo Gon Kim. 2012. "The Influence of the Quality of the Physical Environment, Food, and Service on Restaurant Image, Customer Perceived Value, Customer Satisfaction, and Behavioral Intentions." *International Journal of Contemporary Hospitality Management* 24(2):200–223.
Sievert, Carson, and Kenneth Shirley. 2014. "LDAvis: A Method for Visualizing and Interpreting Topics." *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* 63–70.
Simonyan, Karen, and Andrew Zisserman. 2014. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *ArXiv Preprint ArXiv:1409.1556*.
Timoshenko, Artem, and John R. Hauser. 2019. "Identifying Customer Needs From User-Generated Content." *Marketing Science* 38(1):1–20.
Tirunillai, Seshadri, and Gerard Tellis. 2014. "Extracting Dimensions of Consumer Satisfaction with Quality From Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation." *Journal of Marketing Research* 51:463–79.
Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113(523):1128–1242.
Wang, Wei-ning, Ying-lin Yu, and Sheng-ming Jiang. 2006. "Image Retrieval by Emotional Semantics: A Study of Emotional Space and Feature Extraction." *IEEE International Conference on Systems, Man and Cybernetics* 4:3534–39.
Wang, Xiaohui, Jia Jia, Jiaming Yin, and Lianhong Cai. 2013. "Interpretable Aesthetic Features for Affective Image Classification." *IEEE International Conference on Image Processing* 3230–34.
Zeiler, Matthew D., and Rob Fergus. 2014. "Visualizing and Understanding Convolutional Networks." *European Conference on Computer Vision* 818–33.
Zhang, Shunyuan, Dokyun Lee, Param Singh, and Kannan Srinivasan. 2018. "How Much Is an Image Worth? Airbnb Property Demand Analytics Leveraging A Scalable Image Classification Algorithm." *Working Paper*.
Zhang, Xiang, Junbo Zhao, and Yann LeCun. 2015. "Character-Level Convolutional Networks for Text Classification." *Advances in Neural Information Processing Systems* 649–57.