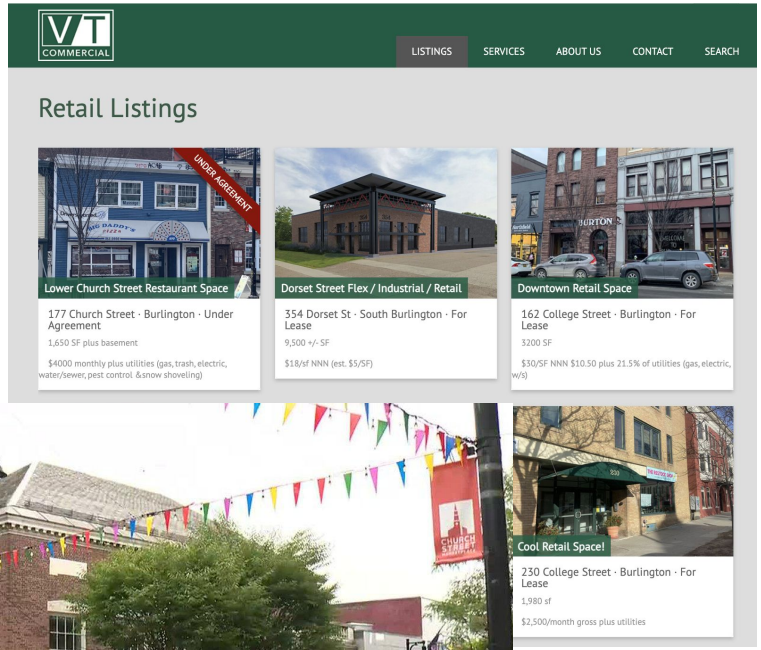
A Prediction Model for Burlington Restaurant Ratings Based on Review Features

Caleb Oliveira and Emma O’Brien

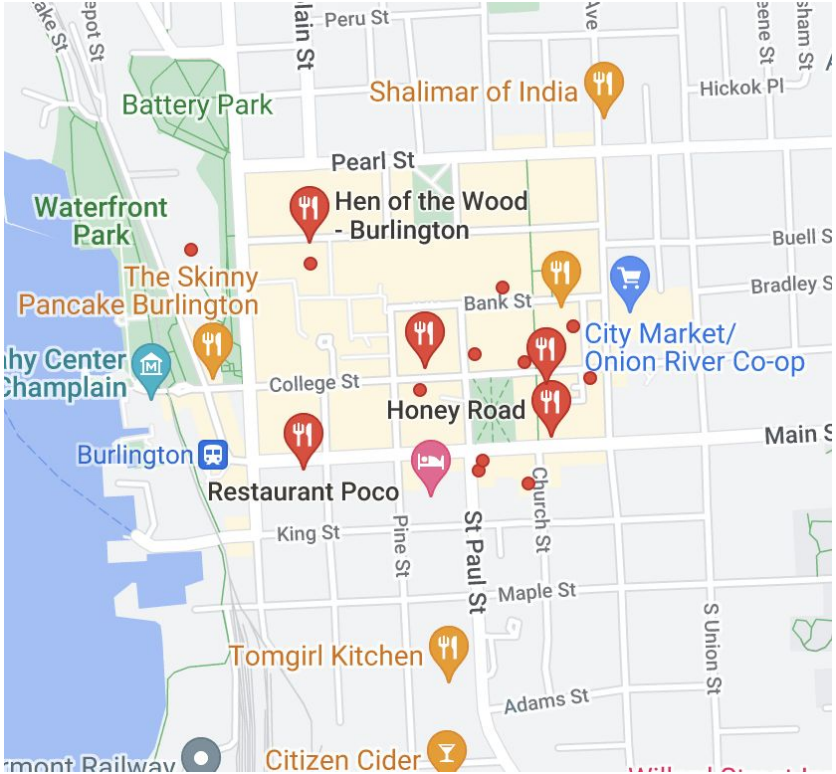
Motivation

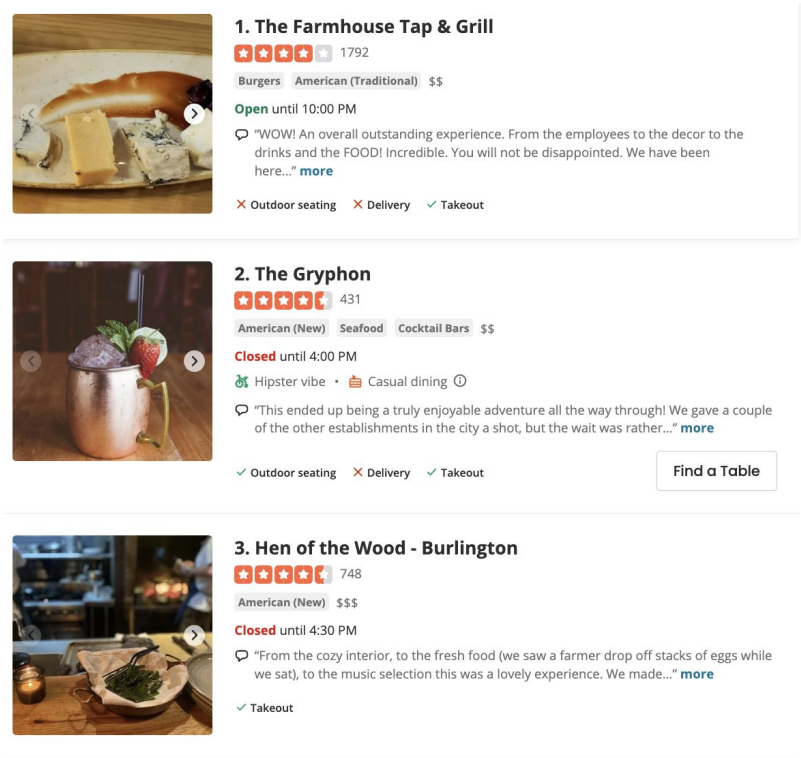
- Next time you walk down Church St take a look at the changes in restaurants
- What makes a restaurant successful in Burlington?
- By analyzing user reviews and restaurant information on Yelp, can we discover any insights into customer preferences?



What we planned:

- Collect online review data for local restaurants
- Identify trends
- Draw conclusions on what makes a restaurant successful in Burlington





How we planned to do it:

- Curate a list of local restaurants and webscrape review data from well-known sources
- Filter data and identify good features and general trends

Related Work 1

Sentiment Analysis of Customer Reviews of Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence: Systematic Review

by Anirban Adak [1], Biswajeet Pradhan [1,2,3,*] and Nagesh Shukla [1]

[1] Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), School of Civil and Environmental Engineering, Faculty of Engineering & IT, University of Technology Sydney, Sydney, NSW 2007, Australia

Table 3. Common complaint types in FDS.

Complaint Types	References
Service, missing item, problem with order, missing order, rude service	[4,15,19,32,33,34]
Food, food quality, food taste	[4,15,19,32,33,34]
Place, location	[19,27,35]
Experience, environment, ambiance, dining atmosphere	[4,15,27,35,36]
Value for money, restaurant value, cost	[4,15,27,35,36]
Time, slow service, slow delivery	[19,33]

Adak, Anirban, Biswajeet Pradhan, and Nagesh Shukla. 2022. "Sentiment Analysis of Customer Reviews of Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence: Systematic Review" Foods 11, no. 10: 1500. https://doi.org/10.3390/foods11101500

- Study looked at customer reviews from food service delivery companies to see if they could find ways to increase customer satisfaction
- Used sentiment analysis, machine learning models, deep learning models, and explained artificial intelligence methods to predict customer sentiment
- Got good results with Deep Learning, but this lacked explainability

Related Work 2

- Study looked at a dataset of 4000 restaurant reviews in Karachi (a large city in Pakistan)
- Used sentiment analysis to determine if the review was negative or positive, then using text classification techniques classified the reviews by food taste, ambiance, service, and value
- Found the most success with a random forest algorithm (95% accuracy)

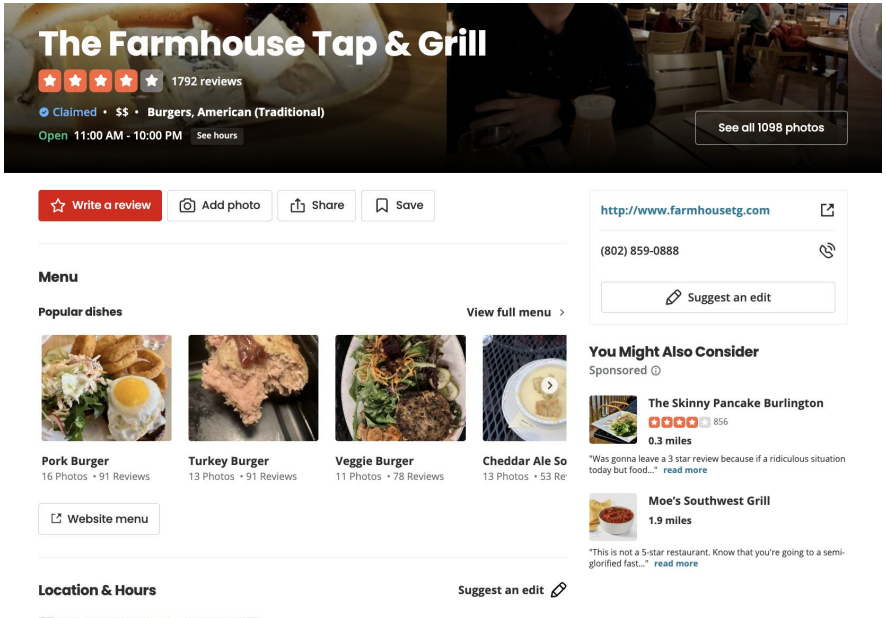
Sentiment Analysis and Classification of Restaurant Reviews using Machine Learning

Publisher: IEEE | Cite This | PDF

	The ratio of training versus testing dataset - 75:25											
Algorithm	**Naïve Bayes**				**Logistic Regression**				**Random Forest**			
Categories	**P %**	**R %**	**F %**	**Acc.%**	**P %**	**R %**	**F %**	**Acc.%**	**P %**	**R %**	**F %**	**Acc.%**
Food Taste	82	99	89	84	93	98	96	86	95	97	96	94
Value for Money	92	51	66		92	91	93		94	91	92	
Ambiance	92	62	73		93	85	89		94	90	92	
Service	93	62	74		91	87	90		94	90	92	

K. Zahoor, N. Z. Bawany and S. Hamid, "Sentiment Analysis and Classification of Restaurant Reviews using Machine Learning," 2020 21st International Arab Conference on Information Technology (ACIT), Giza, Egypt, 2020, pp. 1-6, doi: 10.1109/ACIT50332.2020.9300098.

Data Collecting: Web Scraping



- Used Yelp as a source
- Scrapped from list of 110 local restaurants using Beautiful Soup!
- Collected name, rating, date, location, cuisine type, and text reviews

Data Collecting: Web Scraping


```
https://www.yelp.com/biz/the-farmhouse-tap-and-grill-burlington

<!DOCTYPE html>
<html lang="en-US" prefix="og: http://ogp.me/ns#" style="margin: 0;padding: 0; border: 0; font-size: 100%; font: inhe
rit; vertical-align: baseline;"><head><script>document.documentElement.className=document.documentElement.className.r
eplace(no-j/,"js");</script><meta content="text/html; charset=utf-8" http-equiv="Content-Type"/><meta content="en-US"
http-equiv="Content-Language"/><meta content="width=device-width, initial-scale=1, shrink-to-fit=no" name="viewport"/
><link content="#FF1A1A" href="https://s3-media0.fl.yelpcdn.com/assets/srv0/yelp_large_assets/b2bb2fb0ec9c/assets/im
g/logos/yelp_burst.svg" rel="mask-icon" sizes="any"/><link href="https://s3-media0.fl.yelpcdn.com/assets/srv0/yelp_la
rge_assets/dcfe403147fc/assets/img/logos/favicon.ico" rel="shortcut icon"/><script> window.ga=window.ga||function()
{(ga.q=ga.q||[]).push(arguments)};ga.l=+new Date;window.ygaPageStartTime=new Date().getTime();</script><script>
            window.yelp = window.yelp || {};
            window.yelp.cookieTypePreferencesHeader = '["ANALYTICS","FUNCTIONAL","INTERNAL","STRICTLY_NECESSARY","TAR
GETING"]';
            window.OneTrust = window.OneTrust || {};
            window.OneTrust.dataSubjectParams = window.OneTrust.dataSubjectParams || {};
            window.OneTrust.dataSubjectParams.id = '';
            window.OneTrust.dataSubjectParams.isAnonymous = true;

            function addCookieManagementScript() {
                var tag = document.createElement('script');
                tag.async = true;
                tag.src = 'https://cdn.cookielaw.org/consent/5fbcd417-fbc1-4fcc-abad-f622f1bba72d/otSDKStub.js';
                tag.type = 'text/javascript';
                tag.charset = 'UTF-8';
                tag.setAttribute('data-domain-script', '5fbcd417-fbc1-4fcc-abad-f622f1bba72d');
                tag.setAttribute('data-document-language', 'true');

                document.head.appendChild(tag);
            }

            if ('requestIdleCallback' in window) {
                requestIdleCallback(addCookieManagementScript, { timeout: 4000 });
            } else {
                setTimeout(addCookieManagementScript, 2000);
            }
        </script><script async="" src="https://www.google-analytics.com/analytics.js"></script><script async="" src
="https://www.googletagmanager.com/gtag/js?id=G-K9Z2ZEVC8C&amp;l=ga4DataLayer"></script><script>window.ga4DataLayer=w
indow.ga4DataLayer||[];function ga4Tag(){ga4DataLayer.push(arguments);};ga4Tag("js",new Date());ga4Tag("config","G-K9
Z2ZEVC8C",{send_page_view:false});</script><meta id="emotion-container"/><style data-emotion-css="0 xl0fej ymzf9m zar
4qi 13merx8 e29med qyp8bo 1nrzw89 w8rns 1j8qci xlzvdl 1sb02f4 19v1rkv 148iqgq chan6m 12i50in 6a0jil qgunke 1egxyvc 1i
dmmu3 sw4sjz 1v6h3fb 1evauet 1xvm99p 1enow5j 12il5ah 7z2yta 1llfbb7 aurft1 z8bp47 1r871ch 1x88bq4 1jr0qsc 1qjitbm 1iu
vwzi omlvsn ije6j5 92lxig 1m77lek 9vivtc ralh0w 1u89q5b wonnhn vzslx5 1pz4y59 ux5mu6 1fnccdf rom85o inq9gi 1c2abjj 1o
ibaro vently ptoa9p 1hqkluu igpa9p 83zf8j 1lr1m88 1s9yq3a 1adhs7a 1x398nr 1tbr2yw 1e6spgt tco4il 1afdtvl 1nx1n9o 3u4t
y l56cg5 15j7fnr war30n 1nwtuv2 14zroxr 1hvmmx1 1155t4a gutk1c f4x8dl 1rat78m 1v8f4td bumo7c dz6lzi foyide 1lx34my ah
goya 1kq79li 79elbk 1rjq8gp 16agjqd 174w6ul gnym5v rlqqlq 1gquror 12anxc3 cgap95 4zp531 45sgf 1h0y4i5 1k01157 rvoh79
160jk0r 1nzu07n 1c994ch 1dvzjst 1hyrt2g na3oda 1s86a4e i9xpmd v2jytq 597eco r6zr65 qcwfp3 165w7u0 p3k0ud z8jbqp xb3nz
6 dh50jt 1jxkguo 182o411 1jac6c2 cwr0q8 ddqa90 fksa8h n65gq3 1vg9j2 19sk4h4 147xtl9 11i4m5w f14jpt 1se8maq 1x9ee72 ap
lobz 18182hy 1fdy0l5 1ir4e44 13jrydf 1d810fm 1qdduzb 1ytpb90 tza3mu agyoef 1p9ibgf mpwjkl xkn8u7 1inzsql nyjpex 142pe
7r r9996t 1jtriov 29kerx 1hgawz4 1iog7m3 106vfgv xxqqxs 1wayfxy 1j6t045 4g6ai3 lziufc 1xkgsub w1k3wh ee59jl 1oiyjl1 l
sbzdp 9zvmc3 1p80smk xp8w2v 1vhakgw 1lnf5f9 2sacua">html,body,div,span,applet,object,iframe,h1,h2,h3,h4,h5,h6,p,block
quote,pre,a,abbr,acronym,address,big,cite,code,del,dfn,em,img,ins,kbd,q,s,samp,small,strike,strong,sub,sup,tt,var,b,
u,i,center,dl,dt,dd,ol,ul,li,fieldset,form,label,legend,table,caption,tbody,tfoot,thead,tr,th,td,article,aside,canva
s,details,embed,figure,figcaption,footer,header,hgroup,menu,nav,output,ruby,section,summary,time,mark,audio,video{mar
gin:0;padding:0;border:0;font-size:100%;font:inherit;vertical-align:baseline;}article,aside,details,figcaption,figur
e,footer,header,hgroup,menu,nav,section{display:block;}body{line-height:1;}ol,ul{list-style:none;}blockquote,q{quote
s:none;}blockquote:before,blockquote:after,q:before,q:after{content:'';content:none;}table{border-collapse:collapse;b
order-spacing:0;}.hidden-non-responsive-inline,.hidden-non-responsive-inline-block,.hidden-non-responsive-block,.hidd
en-non-responsive-table-cell{display:none !important;}@media only screen and (max-width:600px){.responsive .responsiv
e-small-display-inline-block{display:inline-block !important;}.responsive .responsive-small-display-block{display:blo
ck !important;}.responsive .responsive-small-display-inline{display:inline !important;}}@media only screen and (max-w
idth:780px){.responsive-biz .responsive-small-display-inline-block{display:inline-block !important;}.responsive-biz .
responsive-small-display-block{display:block !important;}.responsive-biz .responsive-small-display-inline{display:inl
ine !important;}}@media only screen and (max-width:1024px){.responsive .responsive-medium-display-inline-block{displa
y:inline-block !important;}.responsive .responsive-medium-display-block{display:block !important;}.responsive .respon
sive-medium-display-inline{display:inline !important;}}@media only screen and (max-width:1240px){.responsive-biz .res
ponsive-medium-display-inline-block{display:inline-block !important;}.responsive-biz .responsive-medium-display-block
```


→

Web Scraping Last Checkpoint: 15 hours ago (autosaved)

Logout

View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

```python
def extract_data(filename):
    """
    Function to get the information from restaurant's Yelp page and put each review in a list
    param filename: address of restaurant's Yelp site
    return: list of basic restaurant info (name, street, category, price), list of separate reviews
    """
    page = requests.get(filename)
    soup = BeautifulSoup(page.content, 'html.parser')
    review_lines = soup.find_all('script', type = 'application/json')

    #create a list to store basic restaurant info
    restaurant_info = []

    #get restaurant name and street and add to info list
    try:
        info = soup.title.string.split('-')
    except:
        print('Could not get soup :/')
        return [], []
    restaurant = info[0]
    restaurant_info.append(restaurant)
    address = info[2][1:len(info[2])-17]
    try:
        street = address[address.index(' ')+1:]
    except:
        street = 'NA'
    restaurant_info.append(street)

    string_review_line = str(review_lines[0])
    #find food category, price and add to info list
    info2= string_review_line[string_review_line.index('category_aliases'):string_review_line.index('review_count')].sp
    category = info2[1][1:len(info2[1])-9]
    restaurant_info.append(category)
    price_tags = info2[6][1:len(info2[6])-9]
    restaurant_info.append(price_tags)

    #find overall restaurant rating
    string_rating_line = str(review_lines[1])
    average_rating= string_rating_line[string_rating_line.index('stars from')-4:string_rating_line.index('stars from')]
    restaurant_info.append(average_rating)

    #find reviews
    split_start = string_review_line.split('comment')
    try:
        split_end = split_start[4].split('photos')
    except:
        print("could not read reviews :(")
        return restaurant_info, []
    #create list of all the different reviews/ratings of customers
    reviews = []
    for i in range(len(split_start)):
        reviews.append(split_start[i].split('photos')[0])
    return restaurant_info, reviews
```

Data Collecting: Web Scraping

	Name	Street	Category	Price Tags	Average Rating	Review	Date	Rating
0	SUSHI MAEDA	Cherry St, Bur	japanese,restaurants,sush	ul	3.5	The Sushi is definitely quality however, wait ...	2/18/2023	4
1	SUSHI MAEDA	Cherry St, Bur	japanese,restaurants,sush	ul	3.5	This place was delicious. Service was great. T...	8/31/2022	5
2	SUSHI MAEDA	Cherry St, Bur	japanese,restaurants,sush	ul	3.5	This place is legit. Fantastic Japanese food. ...	7/20/2022	5
3	SUSHI MAEDA	Cherry St, Bur	japanese,restaurants,sush	ul	3.5	Totemo oishii desu. @sushimaeda. Finally. The ...	6/6/2022	5
4	SUSHI MAEDA	Cherry St, Bur	japanese,restaurants,sush	ul	3.5	Food was slow to get to the table even though ...	9/27/2022	1
...	...	...	...	...	...	...	...	...
3129	THREE NEEDS TAPROOM & BREWERY	Pearl St	breweries,food,nightlife,poolhall	2	3.0	Voted best place to play pool in Burlington (2...	8/23/2016	5
3130	THREE NEEDS TAPROOM & BREWERY	Pearl St	breweries,food,nightlife,poolhall	2	3.0	update: Everytime I go here I leave pissed off...	9/10/2016	1
3131	THREE NEEDS TAPROOM & BREWERY	Pearl St	breweries,food,nightlife,poolhall	2	3.0	An amazing pizza menu with no option for dairy...	2/23/2015	1
3132	THREE NEEDS TAPROOM & BREWERY	Pearl St	breweries,food,nightlife,poolhall	2	3.0	Cool spot to grab a beer. Service is quick esp...	10/23/2017	5
3133	THREE NEEDS TAPROOM & BREWERY	Pearl St	breweries,food,nightlife,poolhall	2	3.0	This place is so chill! My husband and I stopp...	7/7/2019	5

3134 rows × 8 columns

Data Cleaning

```
nt in Burlington. They constantly have different chef
y. \xa0Everyone&amp;#39;s food was great. \xa0Seems l
```

- Lots of garbage
- Misspelled words + woes of scraping
- 110 Restaurants
- 3000+ Reviews
- Missing data

```
it&amp;#39;s right up there with them.<br&gt;<br&gt;The owner needs to show
```

```
<br&gt;<br&gt;This place is great if you&amp;#39;re looking for good sushi
```

```
# Clean Reviews
df['Review text'] = df['Review text'].str.replace(';', '\'')
df['Review text'] = df['Review text'].str.replace('<br&gt;', '')
df['Review text'] = df['Review text'].str.replace('&', '')
df['Review text'] = df['Review text'].str.replace('amp;', '')
df['Review text'] = df['Review text'].str.replace('  ', ' ')
df['Review text'] = df['Review text'].str.replace('#39\'', ' ')

# Make all lowercase
df['Review text'] = df['Review text'].apply(str.lower)

# Add custom stop words
custom_stop_words = ['food', 'burlington', 'would', '-', 'place',
stop = stop + custom_stop_words
```

```
df['Street'].unique()
```

```
array(['Cherry St, Bur', 'Battery St, Bur', 'N Champlain St, Bur',
       'Church St, Bur', 'Riverside Ave, Bur', 'N Winooski Ave, Bur',
       'Bar', 'Tru', 'Pine St, Bur', 'Colchester Ave, Bur',
       'Main St, Bur', 'Maple St, Bur', 'Bank St', 'Phot', 'Cherry St',
       'Church St', 'S Champlain St', 'Lake St', 'Center St',
       'Saint Paul St', 'Main St', 'Photo', 'N Winooski Ave', 'North Ave',
       'Riverside Ave', 'St Paul St', 'North St', 'Battery St',
       'Mechanics Ln', 'N St', 'Center St, Bur', 'Pine St',
       'South Union St', 'Pearl St', nan, 'College St', 'Henry St',
       'Shelburne Rd', 'Institute Rd', 'College St, Bur', 'Oak St',
       'S Union St', 'N Ave, Bur', 'Tracy Dr', 'Flynn Ave', 'Lawson Ln',
       'Pho', 'Colchester Ave', 'Maple St'], dtype=object)
```

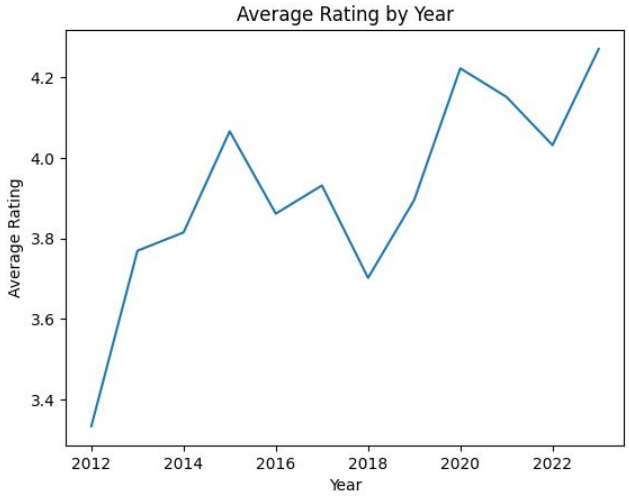
Data Cleaning

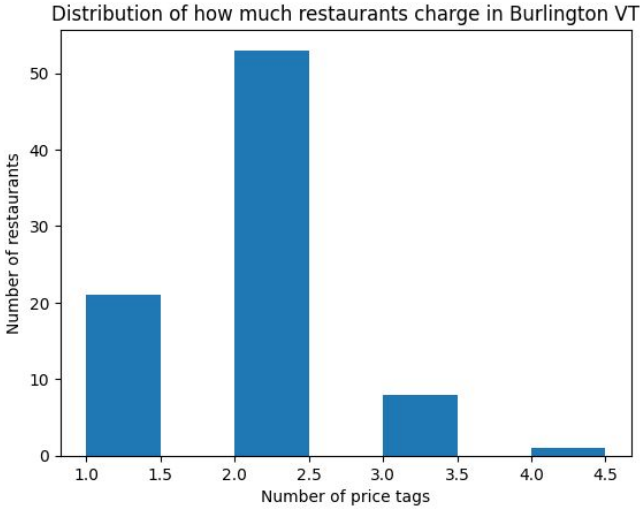
Even more cleaning!

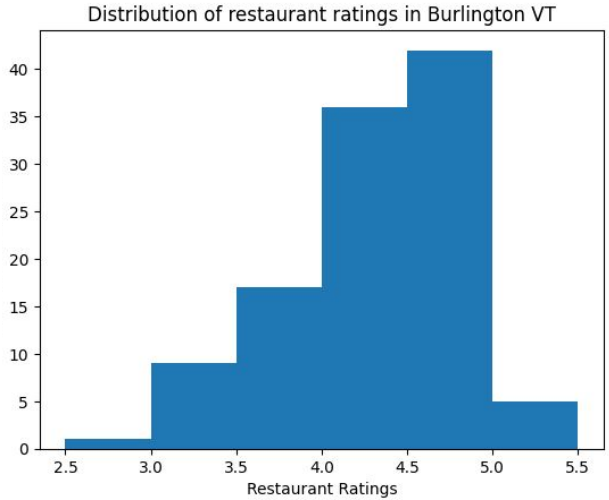
	Name	Street	Price Tags	Average Rating	Review text	Category 1	Category 2	Customer Satisfaction
0	THE FARMHOUSE TAP & GRILL	Bank_St	2.0	4.0	julie another wonderful experience farmhouse! ...	burgers	tradamerica	High
1	HEN OF THE WOOD	NaN	3.0	4.5	cozy interior, fresh (we saw farmer drop stack...	newamerican	NaN	High
2	THE GRYPHON	Main_St	2.0	4.5	best brunch. bloody mary' die for, love unique...	nightlife	newamerican	High
3	JUNIPER BAR & RESTAURANT	Cherry_St	2.0	4.0	met family crazy thanksgiving weekend. great p...	tradamerica	NaN	High
4	HONEY ROAD	Church_St	2.0	4.5	review mentioned difficult get reservation res...	NaN	NaN	High
...	...	...	...	...	...	...	...	...
105	KRU COFFEE	Church_St	1.0	4.5	stopped last day struggling decide grab coffee...	coffee	sandwiche	High
106	HALVORSON'S UPSTREET CAFE	Church_St	2.0	3.0	great food. cozy atmosphere. made reservations...	tradamerica	beer_and_wine	Low
107	PIESANOS	Main_St	2.0	3.0	1 1/2 hour delivery wait...i get it, it' busy,...	pizza	italian	Low
108	KESTREL COFFEE ROASTERS	Maple_St	NaN	4.5	sweet little roastery free short term streetsi...	cafes	coffee	High
109	NECTAR'S	Main_St	2.0	3.0	boyfriend trivia years. first time. great atmo...	burgers	chicken_wings	Low

110 rows × 8 columns

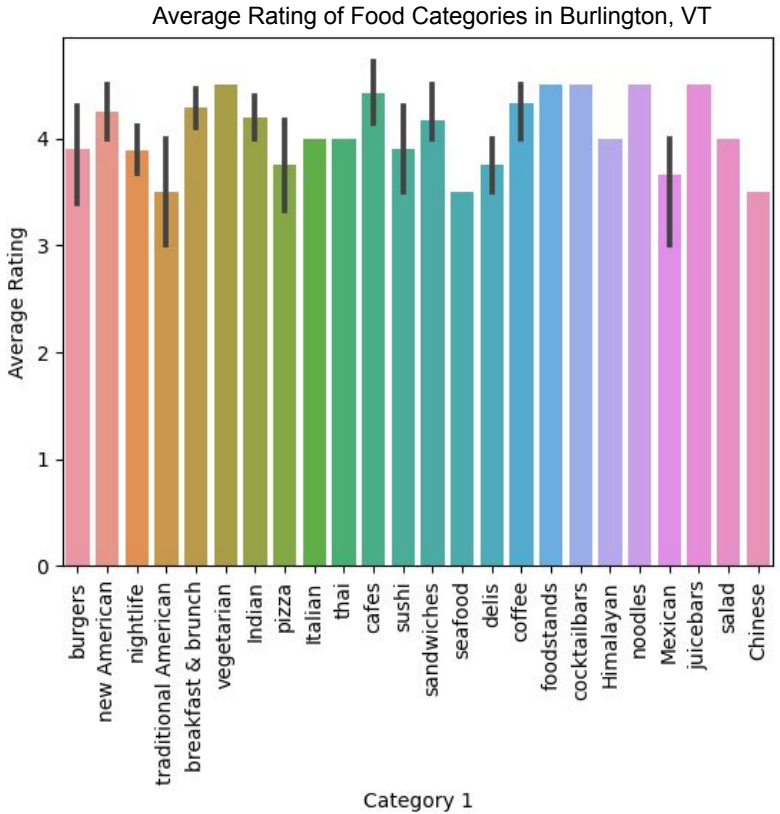
Data Exploration

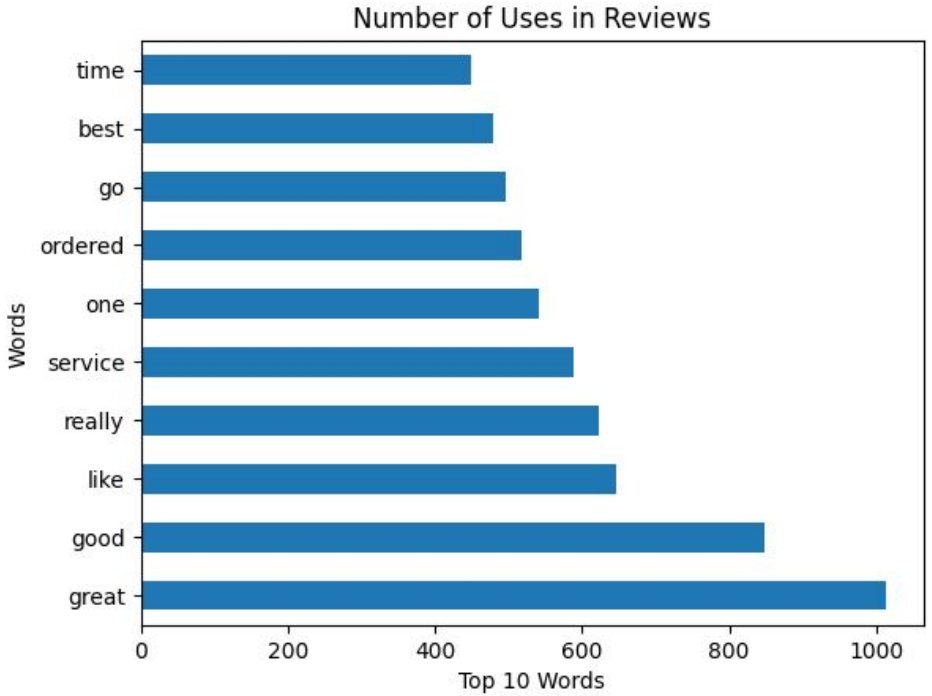






Further Exploration

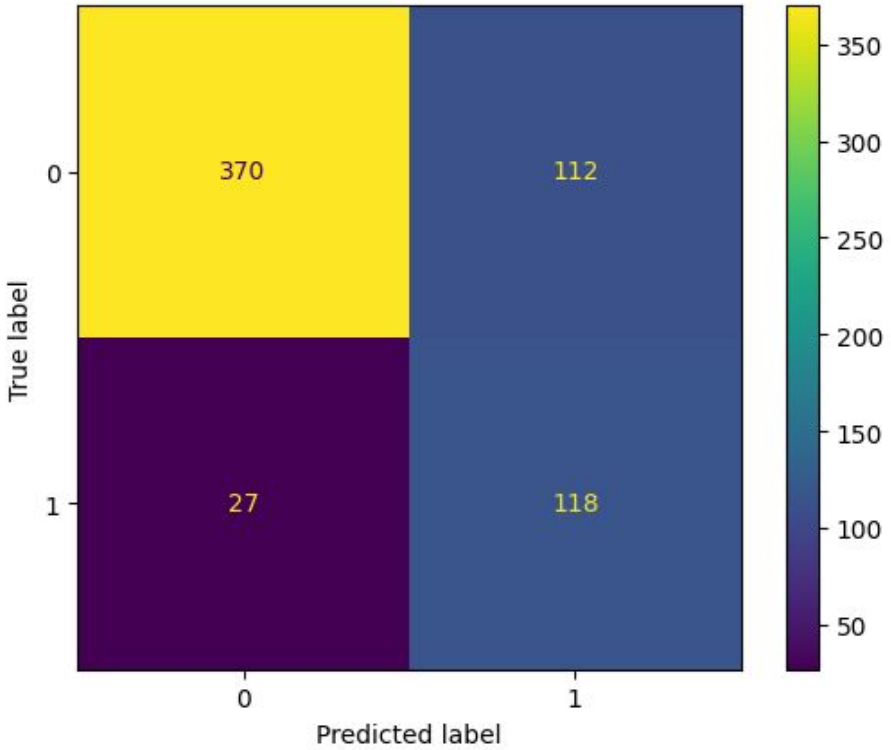




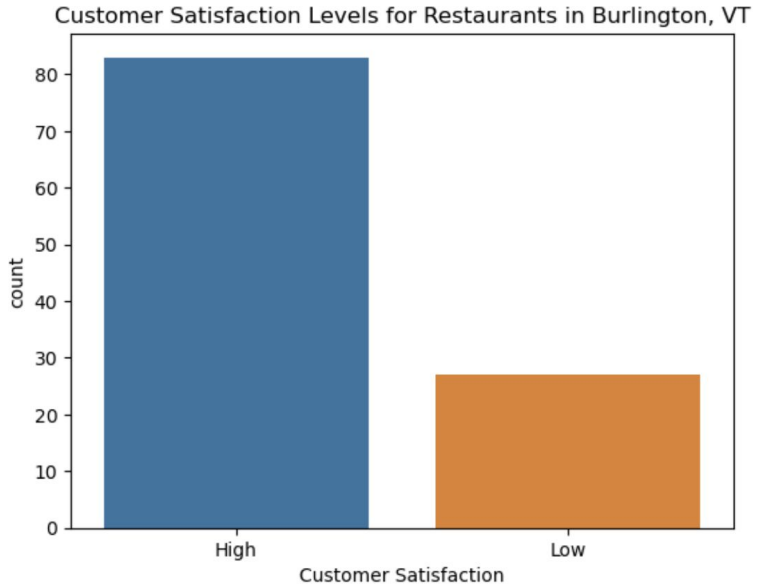
Logistic Regression Model Based on Word Frequencies

- Prediction for each individual review in dataset.
- 78% Validation Accuracy for individual reviews

	feature	coef
21	delicious	-1.944369
44	highly	-1.585731
2	amazing	-1.454004
31	excellent	-1.184753
48	love	-1.177975
..	...	...
27	even	0.536201
74	said	0.548959
98	went	0.596276
71	restaurant	0.607097
68	pretty	0.650434



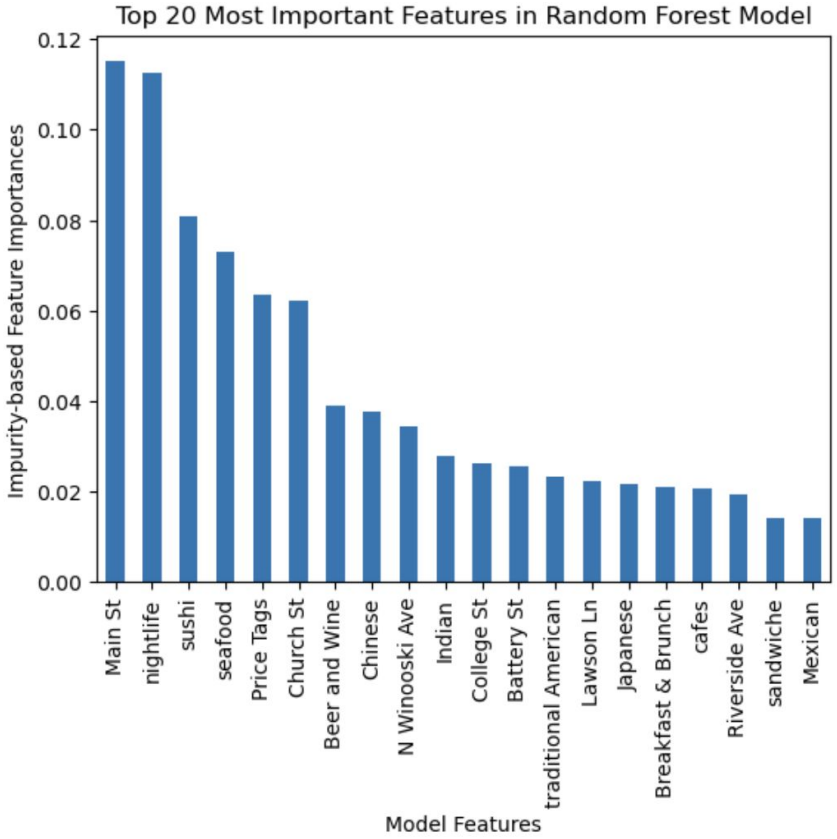
Random Forest Model

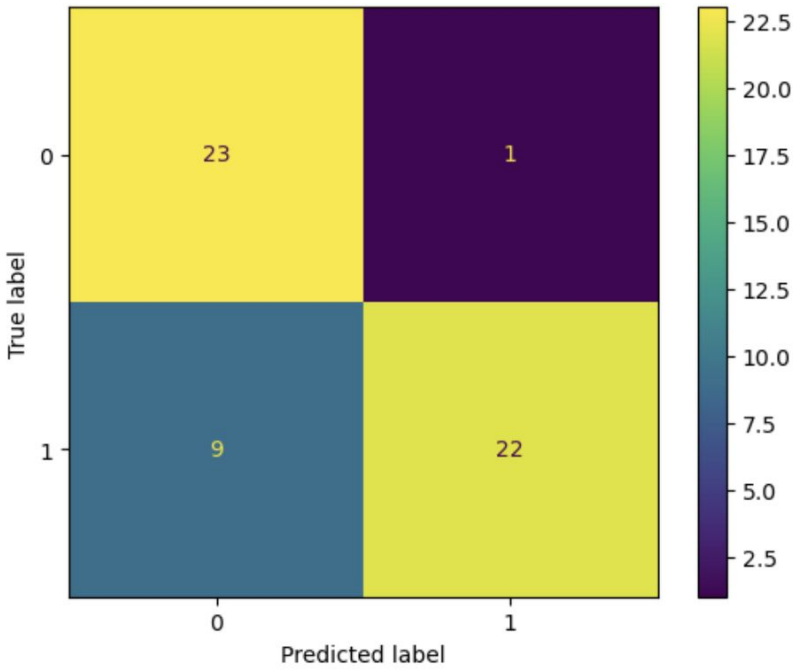


- Problem: Imbalanced data
- Solution: Random Oversampling!
- One hot encoding to transforms categorical data into numerical
- Hyperparameter tuning using RandomizedSearchCV

```
#fit model to our training data
from sklearn.ensemble import RandomForestClassifier
random_forest = RandomForestClassifier(n_estimators = 90, min_samples_split = 5,
                                       min_samples_leaf = 2, max_features = 'sqrt',
                                       max_depth = None, bootstrap=False, random_state=42)
#train random forest model
random_forest.fit(x_train, y_train.values.ravel())
#change array shape with .values.ravel()
```

Random Forest continued

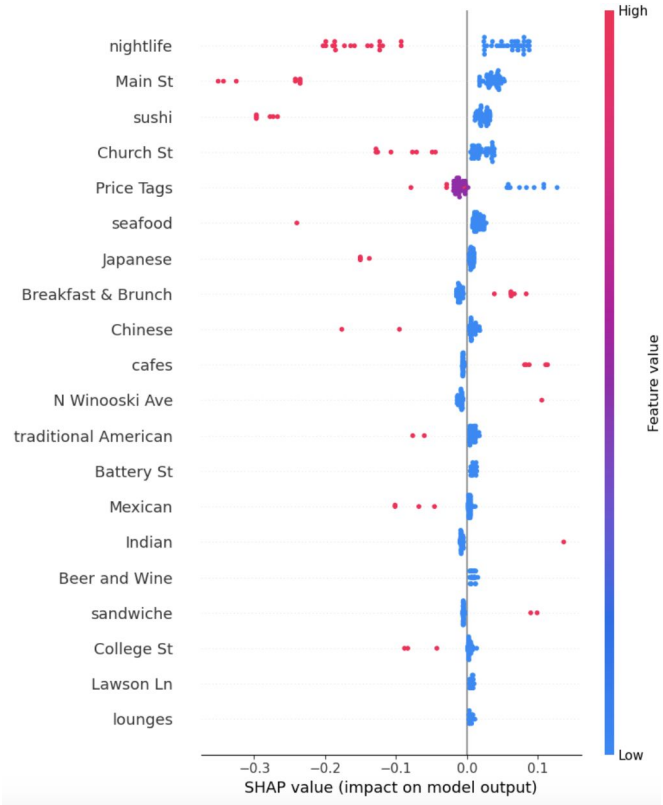




Accuracy: 0.8181818181818182
Precision: 0.9565217391304348
Recall: 0.7096774193548387

But are the features good or bad for customer satisfaction?

- SHAP values can break down a prediction and show the impact of features
- In this plot, positive SHAP values lead to the model predicting that the restaurant will have high customer satisfaction
- Red dot represents a restaurant with this feature, blue occurs when restaurants do not have this feature
- A restaurant having a lower price appears to have a positive impact on customer satisfaction
- Nightlife and Main St appear to have a negative impact



Conclusions

- Location seems play a role in customer satisfaction (Main St appears to have a negative impact)
- Lower food prices (1 price tag on Yelp) seem have a positive impact on customer satisfaction
- Categories like sushi, seafood, Chinese, Japanese, Indian, and Mexican foods impact the model
- Nightlife seems to be have an association with lower customer satisfaction levels

Conclusions

- Perhaps this can be used to promote growth and success for local businesses and enhance the overall dining experience for the community
- Hopefully this can help restaurant owners better understand their customers' preferences and improve their offerings!

Limitations and Further work

- Models tailored to Burlington
- Limited data adds importance to features that may not be that important
- Unbalanced data
- Not enough data to examine individual variables in a meaningful manner
- How accurate is Yelp?
- Would be helpful to add more variables (yearly earnings, years in business, etc)
- Adding more restaurants from the surrounding area or trying this again in a city with more restaurants