Predict Online Review Helpfulness

Ali R Kaya

Table of Contents

List of Figures

List of Tables

Introduction

Customer reviews are an essential source of information for many potential buyers. It is possible to reach thousands of product reviews via different mediums. However, it is not the best practice to read all customer reviews before purchasing a product. This kind of approach will come with a high cost of time. For this reason, the potential buyers need to reach the most helpful customer reviews with minimum time and effort to use their resources more efficiently.

On the other hand, businesses will benefit from providing helpful reviews to potential buyers' convenience because helpful reviews may help firms establish profitable business relationships by increasing the likelihood of purchase, providing material information about the product, or improving customer service. For this reason, a business must identify helpful reviews, extract material information from them, and present those to the convenience of potential buyers.

This project is interested in predicting helpful reviews for the restaurant businesses since each industry has its unique vocabulary and features. For example, a home service business review will be different from a review of a shopping business. What customers look for in a helpful review will differ by business. For this reason, it is essential to focus on a single industry for robust results.

About Yelp Data

Yelp academic dataset is available for free and open to the public. It's a well-known dataset in NLP (Natural Language Processing) research due to the detailed information provided to each business, user, and customer review. Moreover, the dataset consists of six files:

- **business.json** contains information about each company such as name and location, attributes, working hours, etc.
- **review.json** has information about each posted review such as user id, star rating, the customer review itself, number of useful votes, etc.
- **user.json** provides information about each Yelp user such as first name, the total number of reviews, the list of friends, the average star rating, etc.
- **checkin.json** has information about check-in for each business, such as business id and date.
- **tip.json** (the shorter version of reviews and conveys quick suggestions to the businesses) provides information such as the tip itself, the number of compliments and dates, etc.
- **photo.json** contains information about each photo uploaded to Yelp, such as photo id and photo label, etc.

It can be acquired from Yelp's official website by filling a form indicating that the it will only be used for academic and research purposes.

It is important to note that Yelp's academic dataset does not contain every business in the United States. Instead, it provides samples of firms in various categories. As we will see in the Exploratory Data Analysis section, firms in the United States are clustered around a few states. For this reason, our inferences about those businesses may not be generalized nationwide.

Customer Reviews and eWOM

Online customer reviews can be defined as comments or opinions on a specific product posted on the company or a third-party website by peers (Mudambi and Schuff, 2010). Moreover, they can be seen as an outcome of a customer's experience with a product and an input for a potential customer's buying process. For this reason, customer reviews should not be considered as a narrow two-way relationship between the customer and the brand. Still, they may result in a more significant effect on business performance by affecting potential customers.

The consumer buying process consists of the following steps: problem recognition, information search, evaluation of alternatives, purchase decision, and postpurchase behavior (Kotler and Keller, 2016). Consumers may not go through each step anytime they purchase a product. However, depending on the importance of the purchase, prior experience, etc., they will go through several steps in the process, including information search and postpurchase behavior. Accordingly, customer reviews will be one of the primary sources of collecting information for the potential buyers and an essential medium for the customers who would like to share their experiences with the product and purchasing process.

On the other hand, in the information search process, customers would like to access word-of-mouth (WOM) to mitigate uncertainty and perceived risk (Xie, etc., 2014). Traditionally, WOM is done by contacting friends, peers, people who have experience with the product, etc. However, as e-commerce is prevalent, online customer reviews can be considered a part of WOM and can be defined as eWOM (electronic word-of-mouth) (Bronner and Hoog, 2010).

Why are Helpful Reviews Important?

As discussed in the above section, customer reviews are primary sources of information for potential buyers on the internet. However, a company has to do more than just presenting customer reviews to potential buyers. As the number of customer reviews grows, the amount of resources that potential buyers have to allocate to complete the purchasing process also grows. Moreover, low-quality customer reviews can change potential buyers' minds and cause a loss of business relationships. For this reason, a business needs to select and present helpful reviews to build and sustain profitable relationships with customers (Park, 2018).

For example, Amazon orders customer reviews based on their helpfulness and asks each reader to vote if the review was helpful. It is possible to determine the helpfulness of customer reviews manually by other customers. However, this kind of approach comes with its caveats:

1. Potential buyers or consumers will not be able to spend the required amount of time reading all customer reviews and voting for them.
2. Most customer reviews will not be voted (as we will see in the Yelp dataset, 93.4% have less than five votes).
3. The amount of time spent for a review to be recognized as helpful can vary greatly and can cause the loss of potential business relationships.

As a result, predicting customer reviews' helpfulness gives business control over the customer reviews by promoting the possible ones. It is also convenient for potential buyers since it significantly

reduces the search for material information about the product. Finally, it increases the dataset's efficiency by going through each customer's reviews instead of focusing on a limited number of customer reviews.

Exploratory Data Analysis

This section will provide summary statistics, graphs, and tables about the businesses, customer reviews, and users in the Yelp academic dataset. In the first step, we will have a holistic view of the dataset.

In the second step, we will focus mainly on the restaurant industry by investigating the distribution of star ratings, review counts, the relationship between time and helpful reviews, and most common phrases in 1-star and 5-star reviews, etc.

Statistical Summaries

We want to start with a brief description of the Yelp academic dataset. It consists of 209,393 businesses, 1,968,703 Yelp users and 8,021,122 customer reviews.

Table 1: Statistical Summary of Reviews

	Stars	Helpful	Funny	Cool	Is Helpful?	Is Funny?	Is Cool?
count	8,021,122	8,021,122	8,021,122	8,021,122	8,021,122	8,021,122	8,021,122
mean	3.704	1.323	4.596	5.746	0.459	0.200	0.245
std	1.490	3.551	2.188	2.477	0.498	0.400	0.430
min	1.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	3.000	0.000	0.000	0.000	0.000	0.000	0.000
50%	4.000	0.000	0.000	0.000	0.000	0.000	0.000
75%	5.000	1.000	0.000	0.000	1.000	0.000	0.000
max	5.000	1,122.000	976.000	502.000	1.000	1.000	1.000

The businesses have an average of 36 customer reviews with a maximum of 10,129. Also, the average number of reviews posted by a user is 22, with a maximum of 14,455. Almost half of the customer reviews (46%) has at least one helpful vote, and the average number of helpful votes is 1.32 with a maximum of 1,122. Moreover, the average number of helpful votes given by a user is 40. Harold (Yelp only provides first names) is the user who gave the most significant number of helpful votes (197,130) and a Yelp community member since 2012.

On the other hand, the average star rating for all businesses is 3.6. However, Yelp marked approximately 20% of the companies as closed. Finally, the restaurant industry, which is the most significant industry, accounts for 30.5% of all businesses and is followed by the shopping industry, 16.5%.

Table 2: Statistical Summary of Businesses

	Stars	Review Counts	Is Open?	Restaurants	Shopping
count	209,393	209,393	209,393	209,393	209,393
Mean	3.538	36.938	0.807	0.305	0.165
Std	1.024	123.344	0.395	0.461	0.372
Min	1.000	3.000	0.000	0.000	0.000
25%	3.000	4.000	1.000	0.000	0.000
50%	3.500	9.000	1.000	0.000	0.000
75%	4.500	27.000	1.000	1.000	0.000
Max	5.000	10,129.000	1.000	1.000	1.000

Table 3: Statistical Summary of Yelp Users

	Review Count	Helpful	Funny	Cool	Fans	Average Stars
count	1,968,703	1,968,703	1,968,703	1,968,703	1,968,703	1,968,703
mean	22.169	39.827	17.034	21.708	1.459	3.648
std	76.742	513.354	355.057	445.719	16.675	1.173
min	0.000	0.000	0.000	0.000	0.000	1.000
25%	2.000	0.000	0.000	0.000	0.000	3.000
50%	5.000	3.000	0.000	0.000	0.000	3.880
75%	15.000	13.000	3.000	3.000	0.000	4.570
max	14,455.000	197,130.000	165,861.000	191,359.000	11,568.000	5.000

Where the Businesses are Located

The businesses spread around Canada and the United States. As mentioned in the introduction, Yelp only gives a sample of the companies clustered around a few states. In Figure 1, we provided the cluster centers with the corresponding number of businesses. Also, in Table 4, the list of top 10 states (including Canada) with the most companies are presented.

While almost ¼ (43,693 out of 168,903) of all open businesses placed in Canada, Ontario is the most populous state with 36,627 companies. Quebec and Alberta follow it with the number of companies 10,223 and 8,682, respectively.

In the United States, the companies grouped around Arizona-Nevada, Ohio-Wisconsin-Illinois, and North Carolina. However, Arizona and Nevada have 48% of all businesses in the US.

Figure 1: Where Businesses are Located

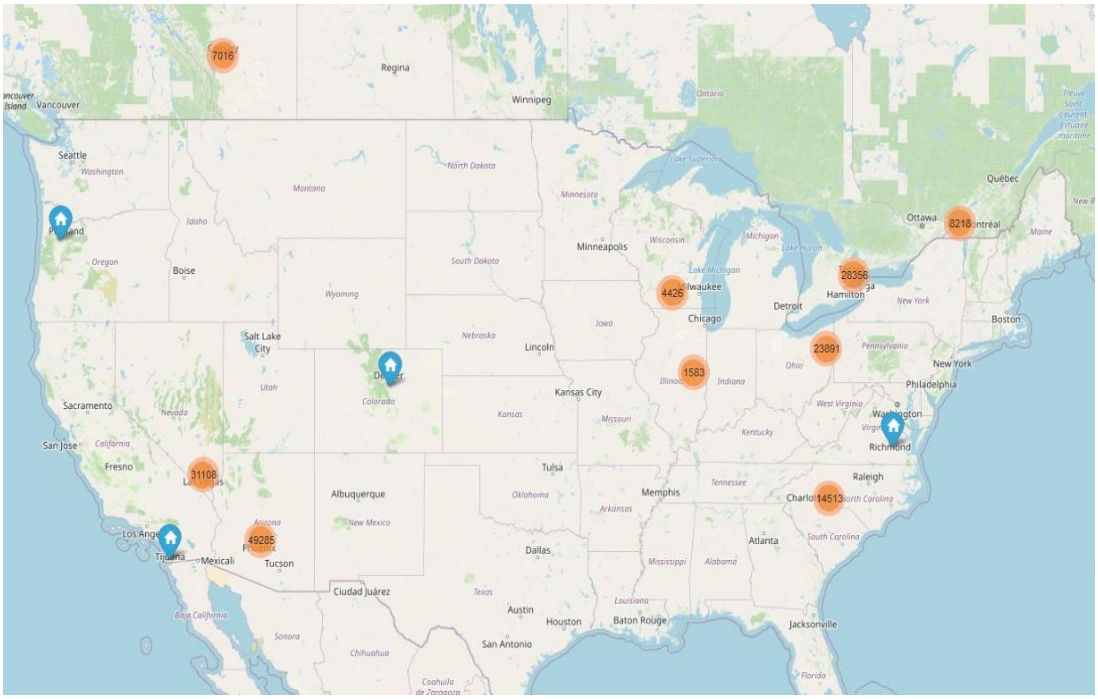


Table 4: Top 10 States with Most Businesses (Including Canada)

State	Number or Businesses
AZ	60,803
NV	39,084
ON	36,627
OH	16,392
NC	16,218
PA	12,376
QC	10,233
AB	8,682
WI	5,525
IL	2,034

Businesses in the United States

As seen in Table 4, the most populous states in terms of the number of businesses are Arizona and Nevada. Ohio and North Carolina have approximately the same number of companies. Pennsylvania is the last state in the US that has more than 10k businesses.

In this project, we mainly focus on the businesses operating in the US's restaurant industry. For this reason, it is essential to select firms in the US. Accordingly, we generated a dummy variable, **in_US**, which indicates if a business operates in the US. Additionally, we take advantage of another dummy variable, **is_open**, provided by Yelp, which shows if a company still operates. Table 5 includes necessary information about businesses in the US.

Table 5: Number of Operating Businesses in the US

Status	Number of Businesses
Open	125,210
Closed	28,633

What are the Industries?[1]

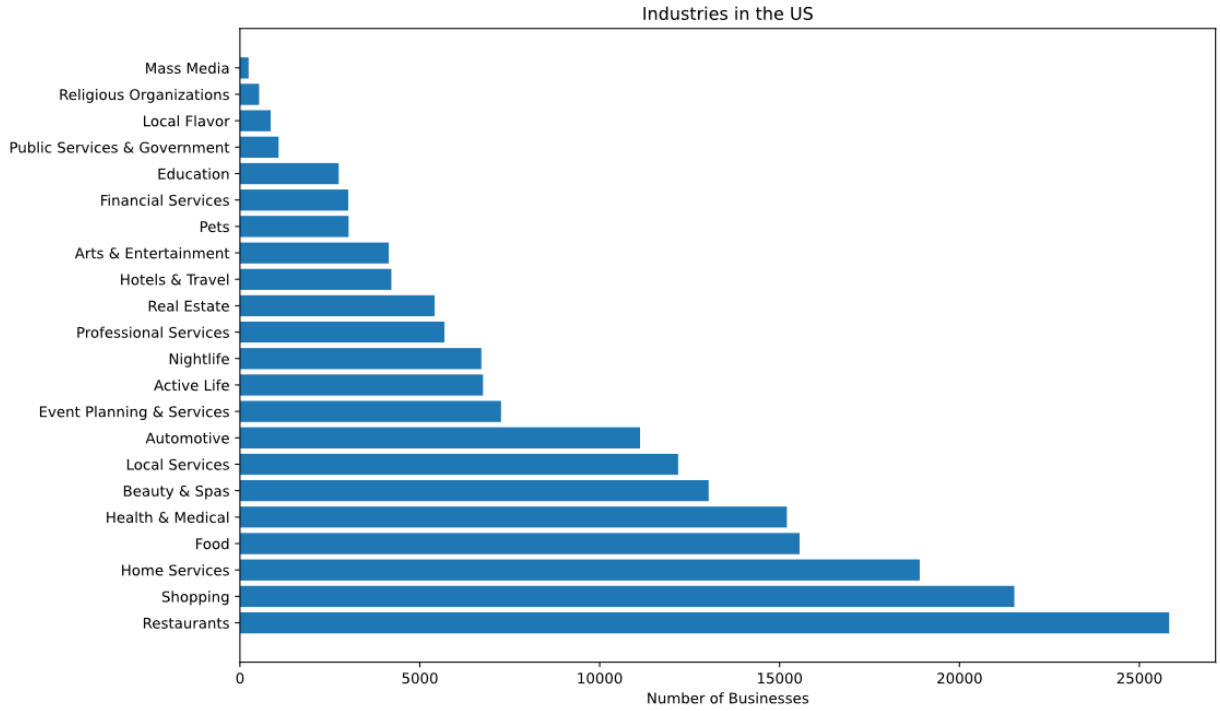
Yelp's dataset provides industry information for each business. However, 374 firms do not associate with any industry. For this reason, they are dropped from the study. The remaining 124,836 firms are divided into 22 industries. The biggest industry is the restaurant industry, which is followed by the shopping industry. They almost account for half of the businesses in the US.

Table 6: Top 10 Businesses (All Industries)

Business	Number of Branches	Overall %
Starbucks	698	0.559
Subway Restaurants	609	0.488
McDonald's	536	0.429
Walgreens	316	0.253
Taco Bell	294	0.236
Circle K	278	0.223
Burger King	273	0.219
CVS Pharmacy	272	0.218
The UPS Store	271	0.217
Wendy's	232	0.186

[1] In this and the following sections, we will only consider the businesses in the US and will refer to them as businesses.

Figure 2: Businesses by Category



Businesses in Restaurant Industry

While the restaurant businesses account for 30.5% of all companies, they are the customer reviews dataset's dominant category. Approximately 63% of all customer reviews (5,055,992 out of 8,021,122) belong to the restaurant industry.

Table 7: Top 10 Businesses (Restaurant Industry)

Business	Number of Branches
Subway Restaurants	609
McDonald's	536
Taco Bell	294
Burger King	273
Wendy's	232
Pizza Hut	232
Jack in the Box	182
Chipotle Mexican Grill	168
Jimmy John’s	157
Panda Express	145

More on Restaurant Industry

In this section, the focus will be on the restaurant industry. We will look at statistical summaries of customer reviews, users, and businesses. We will explore the most common company names and the most common phrases in 1-star and 5-star customer reviews, the businesses with the highest number of reviews, and the users that posted the most reviews, etc. Finally, we will look at the most helpful, funniest, and coolest review.

Statistical Summaries about Restaurant Industry

Businesses in the shopping industry have 3.6-star ratings and 22 reviews on average. The least number of reviews submitted for a shopping business is three, and the highest number of reviews is 3,873. Also, 1.2% and 1% of the restaurant businesses are in the shopping and hotels & travel industries.

Table 8: Statistical Summary of Restaurant Businesses

	Stars	Review Count	Restaurants	Shopping	Hotels & Travel
count	25,827	25,827	25,827	25,827	25,827
mean	3.438	130.241	1.0	0.012	0.010
std	0.862	285.422	0.0	0.110	0.099
min	1.000	3.000	1.0	0.000	0.000
25%	3.000	16.000	1.0	0.000	0.000
50%	3.500	46.000	1.0	0.000	0.000
75%	4.000	132.000	1.0	0.000	0.000
max	5.000	10129.000	1.0	1.000	1.000

Table 9: Statistical Summary of Restaurant Businesses' Reviews

	Stars	Helpful	Funny	Cool	Is Helpful?	Is Funny?	Is Cool?
count	3,487,937	3,487,937	3,487,937	3,487,937	3,487,937	3,487,937	3,487,937
mean	3.781	1.046	0.431	0.578	0.405	0.191	0.241
std	1.412	3.157	1.990	2.623	0.491	0.393	0.427
min	1.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	3.000	0.000	0.000	0.000	0.000	0.000	0.000
50%	4.000	0.000	0.000	0.000	0.000	0.000	0.000
75%	5.000	1.000	0.000	0.000	1.000	0.000	0.000
max	5.000	758.000	786.000	321.000	1.000	1.000	1.000

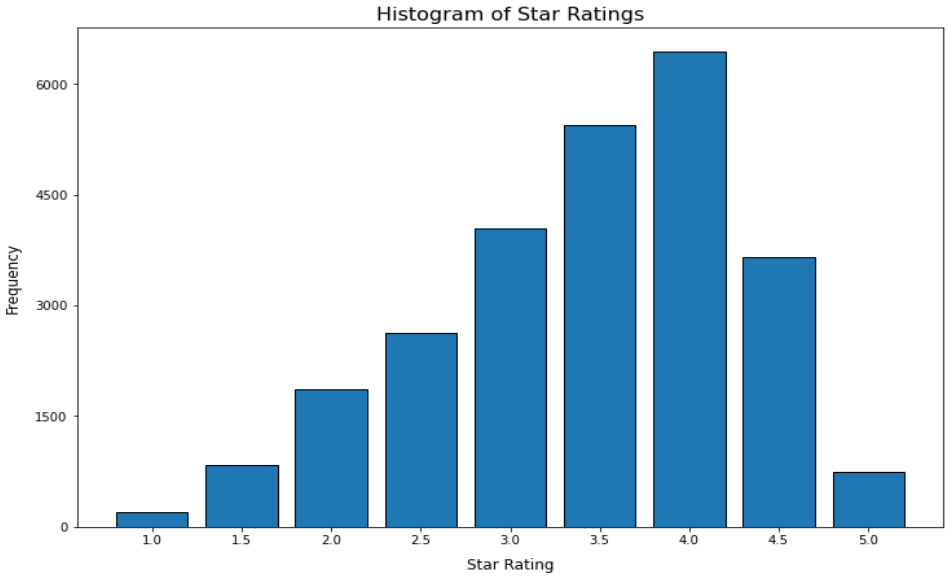
On the other hand, the total number of reviews submitted to the restaurant businesses is 3,487,937, with a mean helpful vote of 1.046. Moreover, 40.5% of all customer reviews have at least one helpful vote, 19.1% have at least one funny vote, and 24.1% have at least one cool vote. The highest number of helpful votes given to a review is 758, the highest number of funny votes and cool votes given to a review are 786 and 321.

Finally, the standard deviation of helpful, funny, and cool reviews shows us that the reviews tend to have a small number of votes, in general. In other words, we may say that a significant number of customer reviews that have at least one of the three kinds of votes (helpful, funny, or cool) have less than ten votes.

Distribution of Business Star Ratings

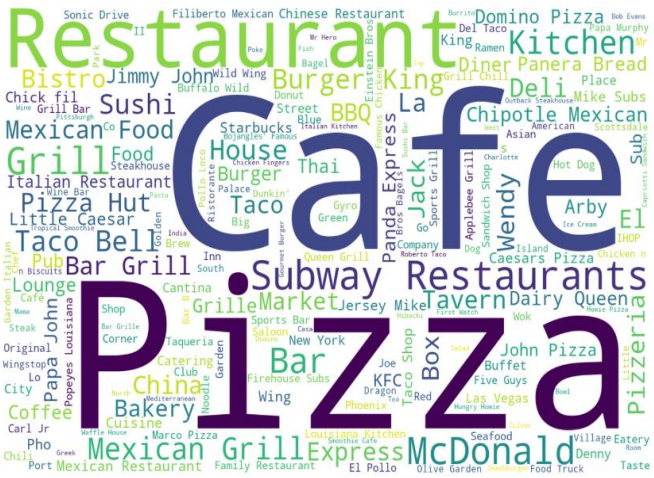
The histogram of star rating is right-skewed, which means that the number of businesses with a higher star rating than 3.0 is much greater than the number of companies with a lower star. It is possible to take star ratings of a business as an indicator of customer satisfaction. For this reason, we can say that a small number of companies can meet their customers' expectations by the goods and services. However, a more significant number of businesses have to improve their business to increase their customers' satisfaction.

Figure 3: Distribution of Star Ratings of the Restaurant Industry



Most Common Words in the Restaurant Industry

Figure 4: Most Common Restaurant Business Names



In a word cloud, a word's frequency in a corpus is directly proportional to its size. In Figure 4, we presented the most common words chosen by the restaurant business owners. 'Restaurant,' 'Cafe,' 'Pizza,' 'Bistro,' 'Grill,' and 'Kitchen' are among the restaurant businesses' most common names.

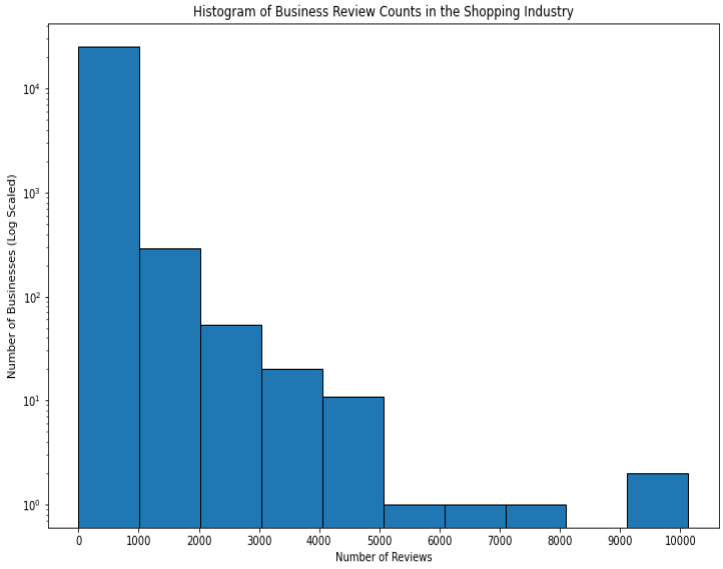
Moreover, we can see that some famous chain restaurants in the figure. Their size is a relative indication of how many branches they have. Also, we see that the name of the cuisine is also highly used in the restaurant businesses, such as ‘Mexican’, ‘China’, and ‘Italian’.

Distribution of Business Review Counts

In Figure 5, we provided the histogram of review counts. The y-axis of the figure is logarithmically scaled, which allows showing a wide range of data compactly. In the figure, the markers on the y-axis increases by multiples of 10. In other words, even though the gaps between the markers on the y-axis are equal on the figure, they grow exponentially.

In the figure, we see that only a few businesses have more than a thousand customer reviews. Accordingly, almost all companies stay below the threshold.

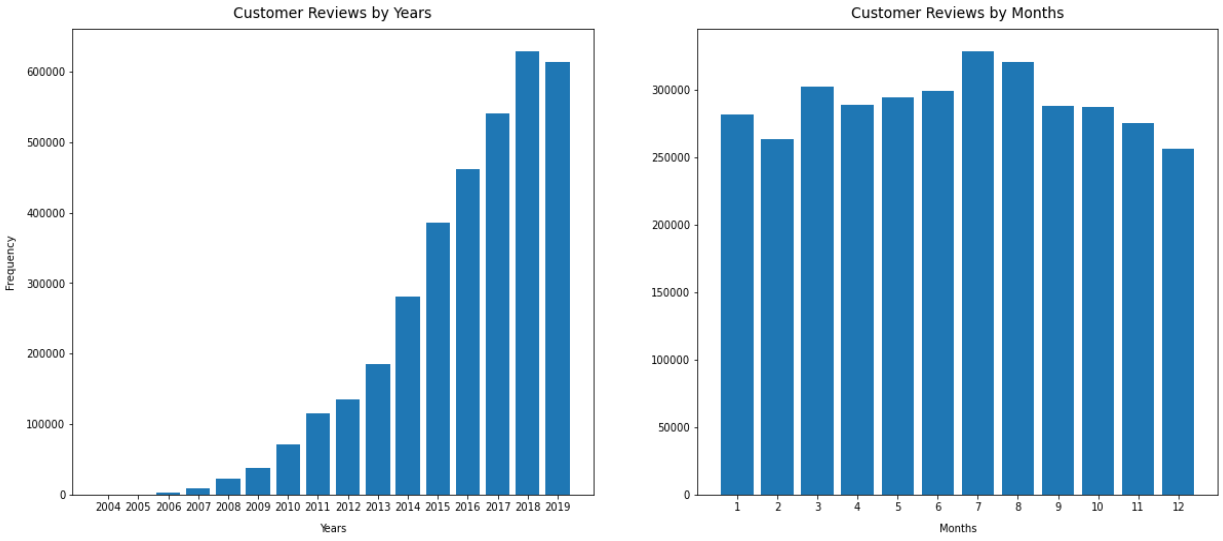
Figure 5: Distribution of Review Counts per Business



Distribution of Reviews by Time

As seen in the figure, the number of reviews submitted steadily increased until 2018. This kind of trend may mean that people started to adopt the culture that Yelp brings to their lives widely. Even though the number of reviews submitted increases yearly, it does not change monthly. People tend to review most in July and August, which is expected because of the weather conditions.

Figure 6: Distribution of Reviews by Time



Distribution of Helpful Reviews

As seen in Figure 7 and Table 10, the number of reviews decreases as we increase the number of helpful votes. It may require to have a cut-off point to determine the helpful reviews. Because a customer reviews with one helpful vote should differ from the one with a hundred helpful votes. For this reason, implementing a cut-off may help us better predict helpful reviews.

Figure 7: Distribution of Helpful Reviews

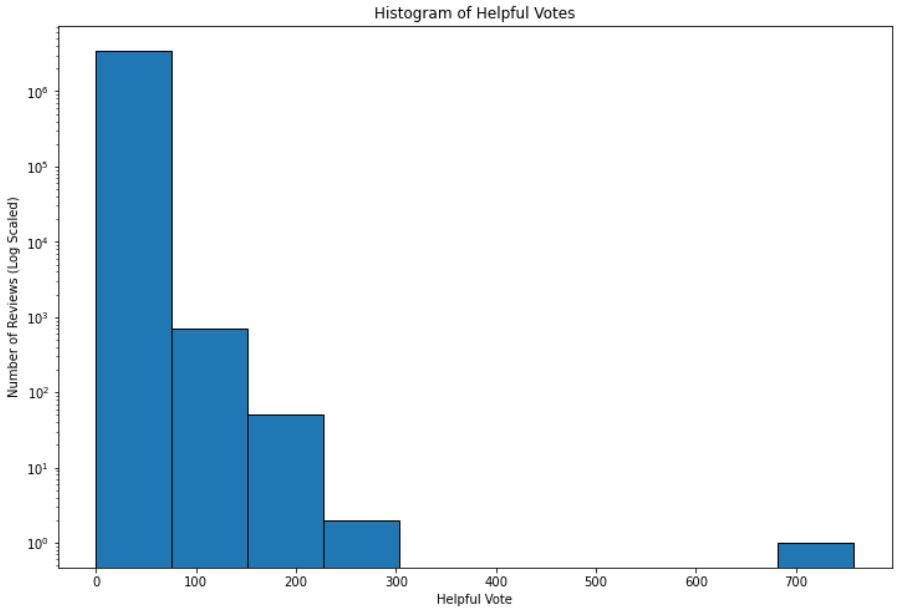


Table 10: Distribution of Helpful Reviews

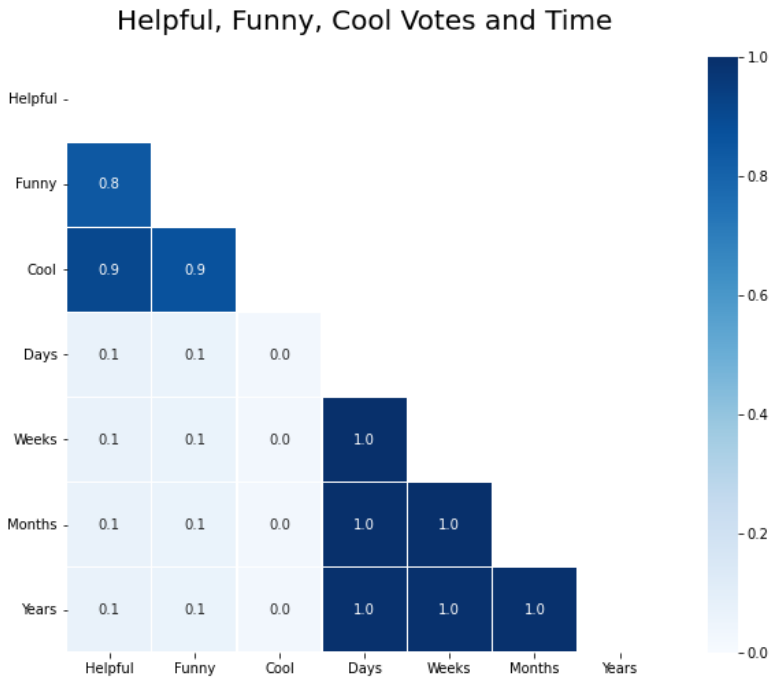
Number of Helpful Votes	Number of Reviews
1-5	1,298,408
6-10	75,439
11-20	25,915
21-30	6,125
31-40	2,429
51-100	1,603
41-50	1,173
101-200	369
201-500	13
501-1000	1

The Relationship Between Review and It's Age

In Figure 8, we see a heatmap of correlation among numerical values in the reviews data. To investigate the relationship between helpful, funny, and cool reviews and time, we generated three measures such as year, month, and day using modular arithmetic. We used an anchor (12/15/2020) to extract all reviews' age in terms of years, months, and days.

As we see in the figure, helpful, funny, and cool reviews are highly positively correlated, leading to multicollinearity. It is the situation when one feature can be predicted by using another feature with great accuracy. For this reason, we will discard funny and cool counts of the reviews from the study.

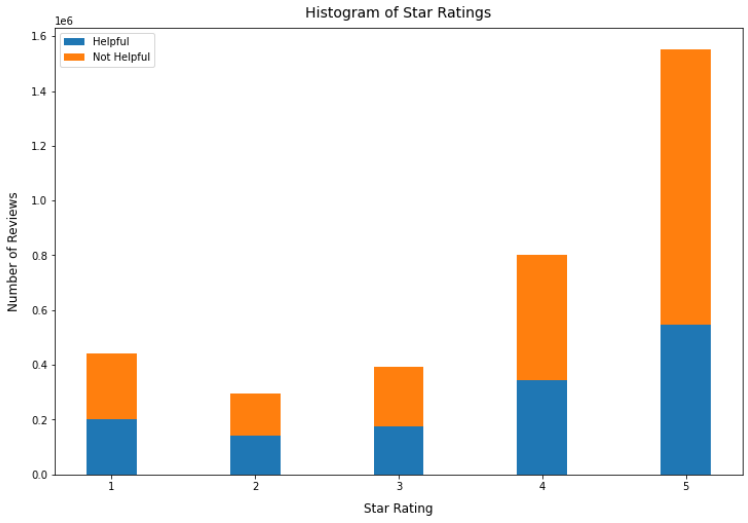
Figure 8: Correlation Among Helpful, Funny, Cool Reviews and Time



Distribution of Review Star Ratings and Helpful Reviews

We want to investigate how helpful votes are distributed with respect to star ratings since star ratings can determine how many helpful votes a review will have. The figure shows that the number of helpful reviews and unhelpful reviews are pretty close in almost all categories. Additionally, star rating distribution is right-skewed, which means that most of the reviews have 5-stars.

Figure 9: Distribution of Helpful Reviews by Star Ratings



Most Common Phrases in Reviews by Review Star Ratings

As expected the vocabulary changes by the star rating. Reviews that have 1-star rating tend to have more negative and discussion-related vocabulary such as ‘asked, ‘told’, ‘said’, ‘manager’ and ‘will never’. On the other hand, 5-star reviews have more positive vocabulary such as ‘love, ‘delicious’, ‘amazing’, ‘will definitely go back’, and ‘well’.

Figure 10: Most Common Phrases in 1-Star and 5-Star Reviews

Top 10 Users with Most Reviews

Yelp's academic dataset provides detailed information for each user. To identify users who submitted a review to a restaurant business, first, we identified the restaurant businesses. By using those ids, we discard all customer reviews that are not relevant to the restaurant industry. Finally, we grouped the data by user id and sorted it in descending order.

Table 11: Top 10 Yelp Users with Most Reviews

Name	Member Since	How Many Times Elite?	Average Star Rating	Number of Fans	Number of Reviews
Brad	2009	0	3.11	77	1259
Stefany	2011	7	3.39	785	1166
Michael	2008	7	3.90	1090	915
Karen	2006	6	3.88	479	832
Norm	2008	9	3.75	319	815
Jennifer	2010	7	3.61	98	810
Jennifer	2009	9	4.05	185	682
Deni	2010	5	3.62	154	639
Pepper	2011	0	3.35	110	626
DJ	2010	2	3.65	121	599

Top 10 Businesses with Most Reviews

To develop the businesses with the most significant number of reviews, we grouped the data by business ids and sorted it in descending order. As expected, all of the firms are located in Las Vegas, and more than half of them have at least a 4.0-star rating. Table 12 provides the necessary information about the firms.

Table 12: Top 10 Businesses with Most Reviews

Business	City	State	Average Star Rating	Number of Reviews
Bacchanal Buffet	Las Vegas	NV	4.0	10,417
Mon Ami Gabi	Las Vegas	NV	4.0	9,536
Wicked Spoon	Las Vegas	NV	3.5	7,594
Hash House A Go Go	Las Vegas	NV	4.0	6,859
Earl of Sandwich	Las Vegas	NV	4.5	5,370
Yardbird Southern Table & Bar	Las Vegas	NV	4.5	4,979
The Cosmopolitan of Las Vegas	Las Vegas	NV	4.0	4,973
The Buffet At Wynn	Las Vegas	NV	3.5	4,953
Secret Pizza	Las Vegas	NV	4.0	4,882
Luxor Hotel and Casino Las Vegas	Las Vegas	NV	2.5	4,819

The Funniest, Coolest and Most Helpful Review

The most helpful, funniest, and coolest review is posted by Doug, who has 684 fans and a member since 2019. The review is about Rio All-Suite Hotel & Casino, which locates in Las Vegas:

The rumor is that this hotel is about to be torn down. Staying here will make an ardent atheist pray that this is true. You can even see it in the eyes of the employees. As friendly as they are, you can tell that nobody wants to be here.

The Rio is like being in 1986. By that I mean it's like you were still driving your 1986 Ford Tempo 33 years later, held together with gaffer's tape and surgical mesh, riding on the rusted rims. Vegas isn't what it used to be, anyone who's come here over the last two or three decades can attest. The Rio isn't even what it was when they last updated their Expedia page.

You have more than too many options when it comes to finding a hotel in Vegas. What finally sold me on the Rio, aside from having regular decent stays here in the past, was this from Expedia.

"Dining options include a seafood buffet, a Japanese restaurant and sushi bar, a wine cellar and tasting room, a New England-style seafood restaurant, an American grill, a South American café, an Indian restaurant, and room service."

There is no seafood buffet. It has long since been discontinued. As has the South American cafe. You find all this out piecemeal by asking everyone and anyone who has worked here long enough to remember. There used to be a Japanese restaurant. The folks at the American grille told me that it didn't exist. It had in fact closed eons ago. Yet I kept passing signs across the casino floor saying Sushi at Club 172. When I would call the front desk, they were unaware of this place. After eating there, I kept asking any employee in any department where sushi could be found. Turns out this was a new sushi place and they rent the spot no differently than a nail salon in a strip mall. The property itself doesn't know or care who may be in or out of business. They just gaze into an unknown future, waiting for the wrecking ball to swing.

There is a poster in the elevator for a celebrity host on a limited run at Chippendales's. It was from last year. I wouldn't doubt that they didn't put it up until after he was long gone. The casino floor is littered with barkers outside of every shop or passageway, hustling everything from haircuts to time-share, like beggars who ask for investments rather than pocket change. The ATM fee is 9.99 to take out a twenty, onerous even by Las Vegas standards. Thankful, and rightfully, my hometown bank treated this charge as an illegal transaction and declined it.

On the first of my six days here, I called down with a litany of these complaints from the 1986 phone in the room. But I could barely hear the front desk because the phone was so old. That simply added one more complaint.

On the second night at midnight, my manager showed up drunk at my door like Oscar Madison from the Odd Couple after having a tiff with his galpal. I didn't wake to his repeated calls and banging at my door (Thanks, Xanax!) so he simply slept like a homeless person outside of my suite in the hallway for almost six hours. Like so many piles of garbage or room service trays during my stay here, he wasn't noticed or removed.

In defense of the Rio, the suite I stayed in was a full 1600 square feet that averaged less than 200 dollars a night, with a full view of the Strip. Despite the colors of beige-on-brown-on-cream - the vibrant spectrum that brings to mind Rio de Janeiro - and the highway salvage living room furniture that is currently spitting feathers from every tear and crack, it's still a pretty decent price. So long as you avoid gambling. Go to a Station casino for that.

I won't say how much I lost gambling here but I'd estimate that my Rewards/Players Club card paid me about one penny for every twenty bucks. I've gambled less at other casinos and been offered free rooms, my own private concierge and amenities too numerous to count. Here I got half-off a buffet.

The hallway reeks of cigarettes and now-legalized recreational weed. This is a positive. Penn & Teller, as well as The Comedy Cellar with Mark Cohen and Guests are also a bonus. Pet-friendly another plus.

Still, I give it one-star. Because who reads a two-star review? It's just business.

Data Cleaning and Feature Extraction

In the EDA section, we explored Yelp's academic dataset and identified our sample as the businesses that still operate in the restaurant industry and are located in the US. As a result, the sample data consists of:

- 3,487,937 customer reviews
- 1,116,634 users and
- 25,827 businesses

As a first step, we dropped all duplicated, 9,073, reviews and removed a null value from the reviews. Thus, the reviews corpus is ready for feature extraction, such as number of sentences, number of words, number of unique words, number of punctuations, number of stop words, number of uppercase and title case words, number of letters, and average word length.

We explored the data based on the extracted features and designed for the data cleaning process in the second step. Additionally, we extracted another set of features, such as number of photos, URLs, price, time and emoticons.

Finally, we performed data cleaning processes and vectorized the reviews corpus using TF-IDF matrix. Moreover, we implemented a cut-off for the minimum number of documents, such as 0.03. In other words, the matrix only consists of terms that appears more than 3% of all documents (reviews).

Basic Text Features

This section will extract and explore basic features to have a broad understanding of the corpus. By doing so, we hope to detect the anomalies in the reviews. Also, we used a logarithmic scale in the histograms to identify extreme values.

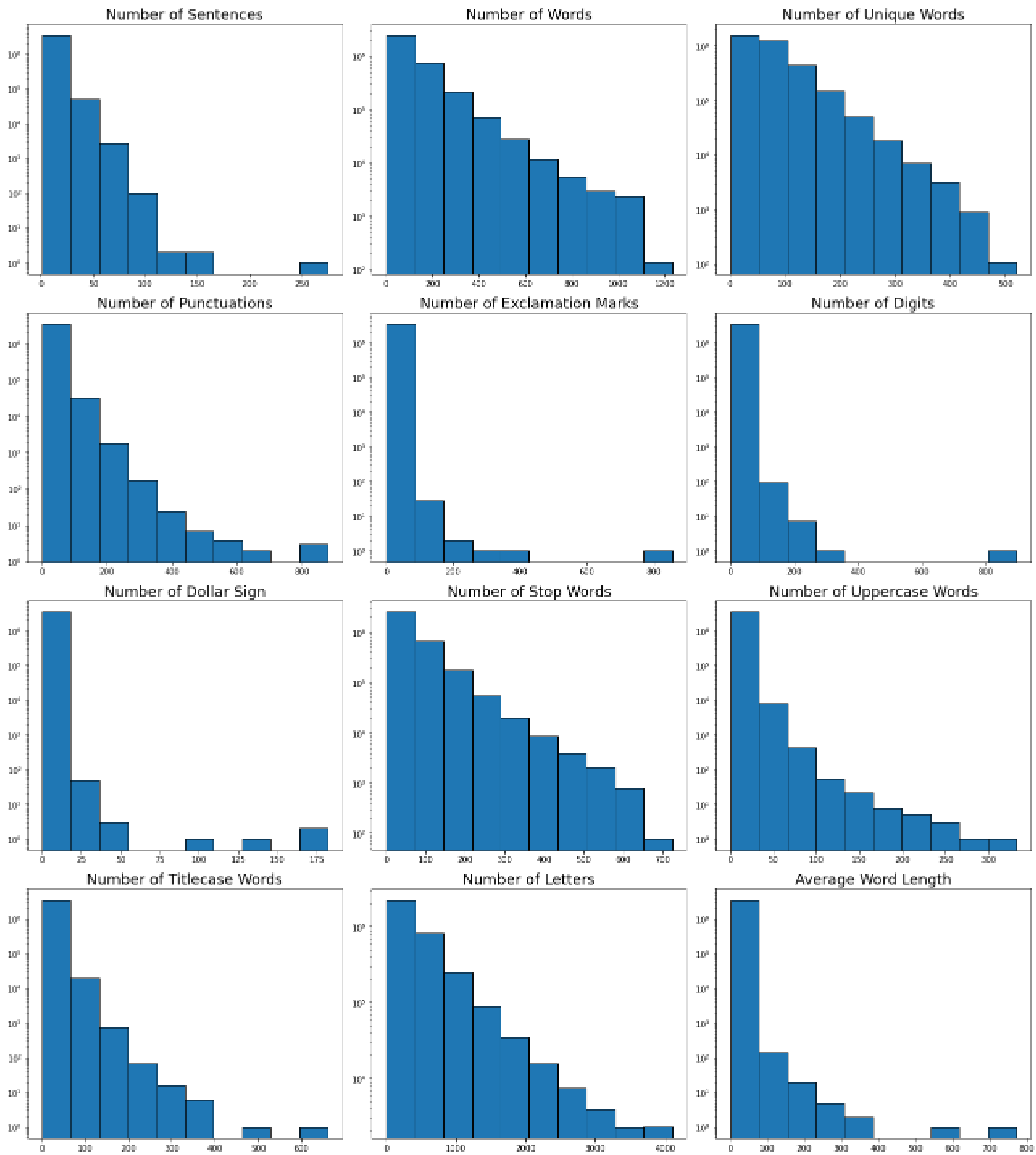
We used the following features:

- Number of Sentences
- Number of Words
- Number of Unique Words
- Number of Punctuations
- Number of Exclamation Marks
- Number of Digits
- Number of Dollar Sign
- Number of Stop Words
- Number of Uppercase Words
- Number of Titlecase Words
- Number of Letters
- Average Word Length

As seen from Figure 11, most reviews have less than 100 sentences, but it is perfectly normal to have incredibly long customer reviews. However, the longest review consists of more than 250 sentences.

In English, a sentence has, on average, 20 words, and each word has 4.7 letters. Thus, an average sentence in English has 94 letters. Accordingly, any properly constructed review may not be able to have more than 50 sentences. For this reason, we will look at the longest review.

Figure 11: Histograms of Extracted Features



Ugh. I have so much to say about this place. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings. Hot wings.

Hot wings. Hot wings.

They have other stuff that's really good. But the hot wings are amaaaaaaazing.

(I like this place)

As seen from the example above, the sentence is wordy and does not contain any material information about the place, such as price, time, and photo.

On the other hand, we focused on the number of dollar signs, exclamation marks, digits, uppercase words, and average word length. Because each feature may reveal essential information to the customers, such as the number of dollar signs and digits increases in a review, it may contain more information about the price. Also, the number of uppercase words and exclamation marks can express customer emotions and be perceived as helpful by others.

However, the average length of words is an essential feature to identify anomalies in the corpus. Since it is approximately 5 in English, we can locate any non-English review. Figure 12 shows the relationship between the average length of words in a review and its language. As the average length of words increases, it is more likely to be written in a different language.

After we investigated the corpus, we examined the relationship between the extracted features and the target variable. However, we first generated a new dataset without the outliers because outliers can

suppress the true relationship between the features and the number of helpful votes. Figure 13 presents the scatter plots.

Figure 12: Example of Non-English Reviews

	review	average_length_words
11146	フェニックス(地元の発音では、フィーニックス)のスカイハーバー国際空港の４番ターミナルにあり...	25.625
19091	ここもレストランではなく、先にカウンターで注文するファストフード店です。\nフォーなどが食べ...	48.800
20006	南部的なフライドチキンが楽しめる全国チェーン。チキンのからあげ等が大好きな日本人にとって、思...	124.000
26676	店員さんがフレンドリー！ \n味も美味しくディナーと次の日のランチも来てしまいました\nディナ...	21.000
32981	好吃好吃好吃好吃好吃好吃好吃好吃好吃今天和我男朋友一起来吃面，好吃好吃好吃好吃我写了那么长了...	92.000

As expected, the average word length clustered around four, and the number of sentences stayed below 50 for most reviews. The number of exclamation marks, punctuations, digits, uppercase words, and title case words also tends to stay lower. Table 13 provides the correlation coefficients between each feature and the target variable. Also, Figure 14 presents the correlation among the extracted features.

Figure 13: The Relationship Between Features and the Target Variable

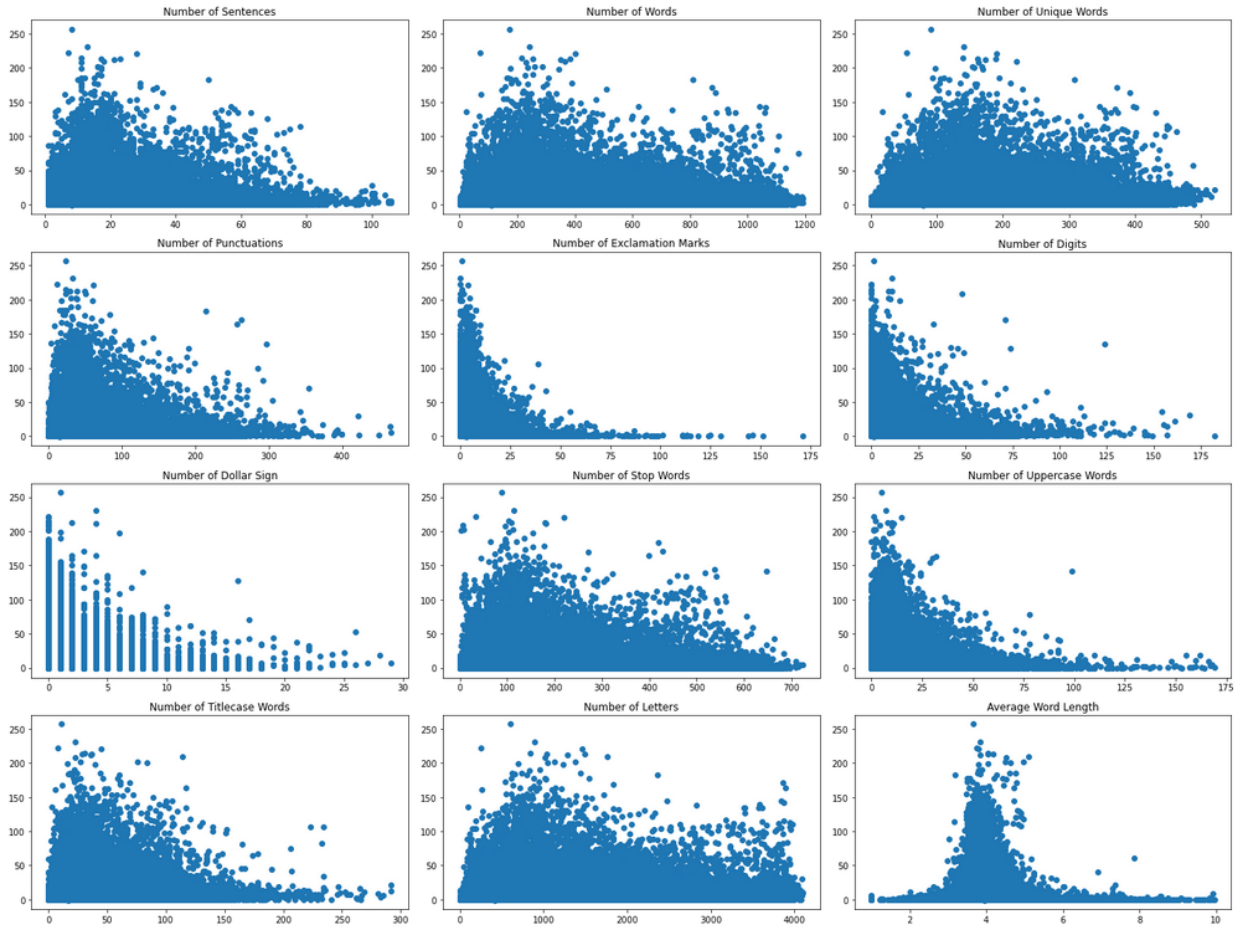


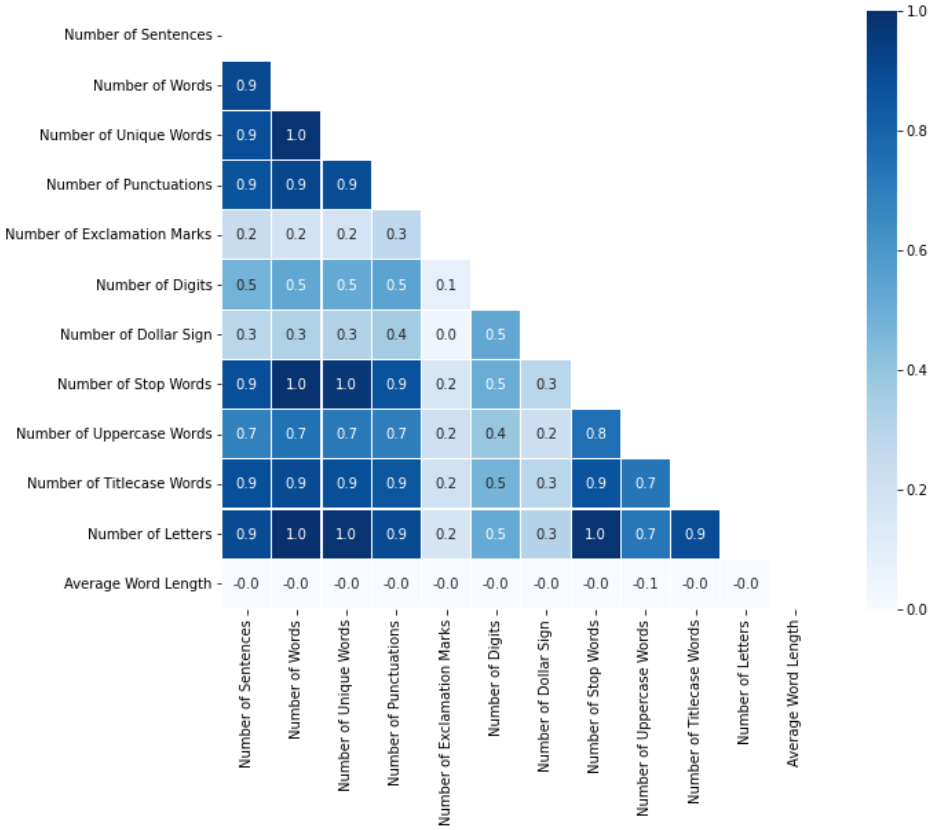
Table 13: The Relationship Between Extracted Features and Helpful Votes

Feature	Correlation with Helpful Votes
Number of Sentences	0.266259
Number of Words	0.281358
Number of Unique Words	0.287694
Number of Punctuations	0.280379
Number of Exclamation Marks	0.066640
Number of Digits	0.164860
Number of Dollar Sign	0.123075
Number of Stop Words	0.261895
Number of Uppercase Words	0.207940
Number of Titlecase Words	0.294231
Number of Letters	0.282718
Average Word Length	-0.018440

All extracted features but the average word length have a positive relationship with the target variable. However, the magnitudes do not vary among features. It is expected that the number of sentences, words, unique words, letters, and punctuations are highly correlated.

As seen in Figure 14, almost all features are highly correlated, and using all features as the determinants of the target variable will cause multicollinearity. For this reason, we only selected the number of unique words as an independent variable and discarded the remaining features. It was possible to extract only the number of unique words initially; however, this would not provide the best opportunity to get familiar with the corpus. As a result, we used the extracted features for experimental purposes and decided to generate another set of predictive variables.

Figure 14: Relationship among Extracted Features



Data Cleaning

As a first step, we used a language detection algorithm to locate English reviews. A small portion of the reviews, 0.002, was in a different language and discarded from the study. We also focused on the businesses that have more than a thousand reviews to have a homogeneous corpus.

Even though we decreased the average length of words from 800 to 30, there were still non-English characters in the corpus. Those reviews used mixed languages, so that we wanted to remove only the non-English characters.

Before we moved to the text normalization process, we have to generate the predictive variables, such as the number of photos, URLs, price, time, and emoticons. Since text normalization requires all that information to be removed from the text, it must be done in advance.

To extract those features, we used Regular Expressions, which finds the pre-defined patterns in the corpus. Since the patterns have to be hard-coded, there may be left-over information after the matching process. However, it should not cause any problem because we covered the most used cases in the patterns. Table 14 presents the relationship between each feature and the target variable.

Table 14: Predictive Features and Helpful Votes

Feature	Pearson R	Significance
PHOTO	0.07	0.00
URL	0.05	0.00
PRICE	0.13	0.00
TIME	0.07	0.00
EMOTICON	0.16	0.00

We, finally, moved to the text normalization process in which we performed the following steps:

- Replace Chinese and Japanese characters with whitespace
- Whitespace formatting
- Reduce duplicated letters (Ex. Soooooooooooooooo → So)
- Replace spaced words (Ex. A M A Z I N G → AMAZING)
- Fix contractions (Ex. I'm → I am)
- Remove hashtags (#) and mentions (@)
- Remove punctuations
- Remove digits
- Lowercase terms
- Remove stop words
- Lemmatize and
- Stemmer

After the text normalization process, we finally dropped all null values from the cleaned corpus. Text normalization may result in the removal of all words and letters from a review. For this reason, we dropped all such documents from the corpus.

In the last step, we generated a TF-IDF matrix from the cleaned corpus. However, we implemented a cut-off to remove noise from the data. We used 3% as the lower and 90% as the upper threshold. As a

result, the number of terms in the corpus decreased from 100,847 to 243, accounting for 53% of the corpus. Table 15 shows the most used words in descending order.

Table 15: Most Frequent Terms in the Corpus

Term	Frequency	Term	Frequency
food	535,503	love	195,117
good	455,635	wait	194,598
place	437,241	restaur	193,228
great	383,019	eat	187,496
time	314,294	friend	182,940
order	312,467	amaz	152,153
servic	308,846	delici	150,732
make	228,983	nice	148,213
back	218,896	tabl	140,939
vega	196,308	drink	138,937

Predict Helpful Reviews

After implementing data cleaning and feature extraction steps, we will get initial results from the analysis. We used three different training sets initially. The TF-IDF matrix, extracted features, and all features together. Also, we used Naïve Bayes and Decision Tree algorithms to obtain initial results.

On the other hand, we generated a binary target variable by transforming helpful votes. Any review with five or more helpful votes is recorded as a helpful review, and the remaining ones unhelpful. We used Naïve Bayes and Decision Tree algorithms to obtain initial results. Confusion matrices are presented in the below figures.

Figure 15: Confusion Matrices – Predict Helpful Reviews with TF-IDF Matrix

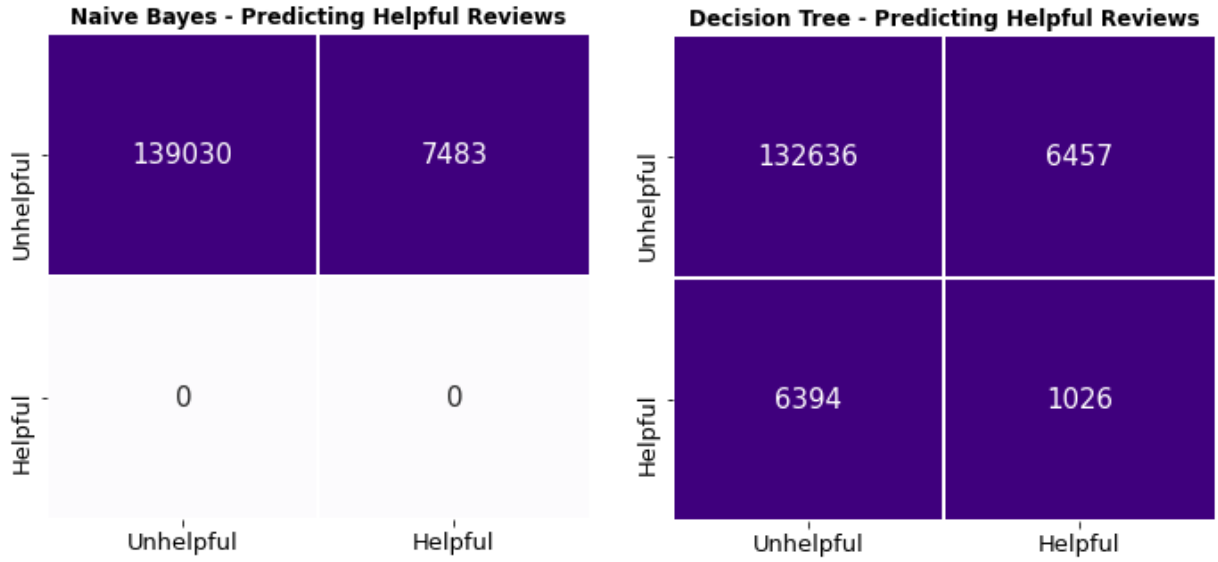
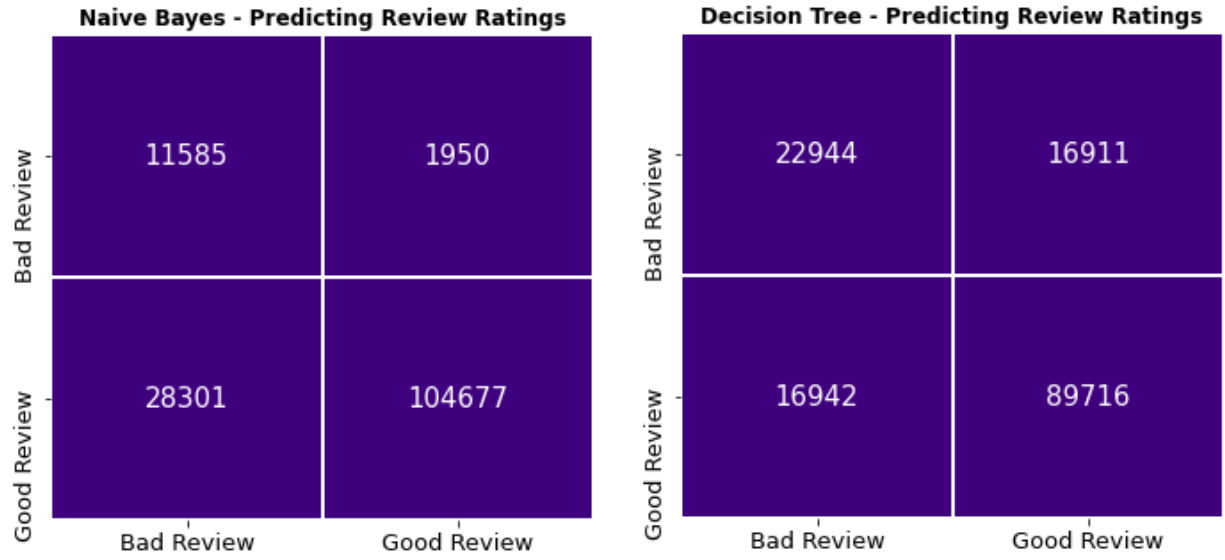


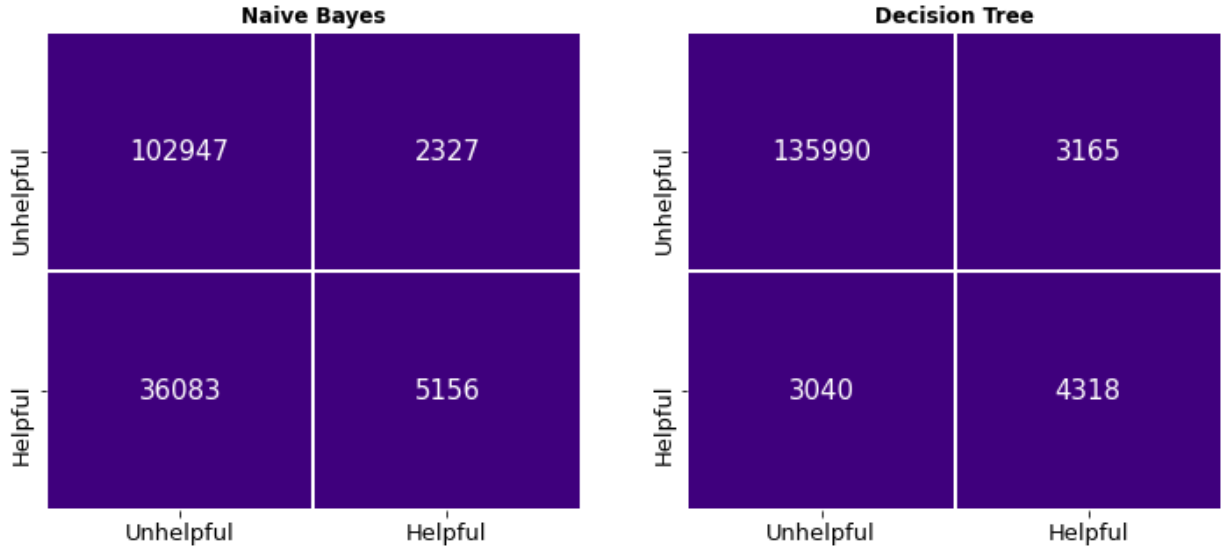
Figure 16: Confusion Matrices - Predict Star Rating with TF-IDF Matrix



Initially, we used the TF-IDF matrix to predict helpful reviews; however, the results were not promising. To check if the matrix works, we predicted star ratings of the reviews with the same matrix. We transformed the star ratings into a binary variable by assigning 1 to reviews with four or higher ratings and 0 to the remaining ones. The results are presented in Figure 15 and 16.

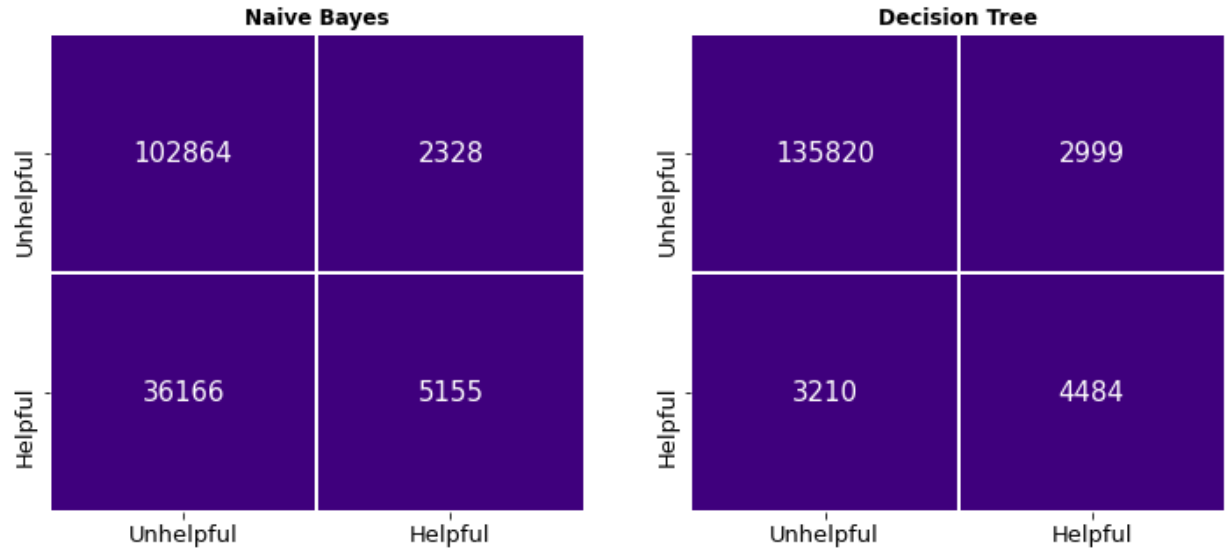
The TF-IDF matrix works when we predict the star rating; however, it does not predict helpful reviews. For this reason, we add the extracted features to the TF-IDF matrix and rerun the algorithms. The results are shown in Figure 17.

Figure 17: Confusion Matrices - Predict Helpful Reviews Using All Features



Finally, we used only the extracted features to predict helpful reviews, and the results are provided in the below figures.

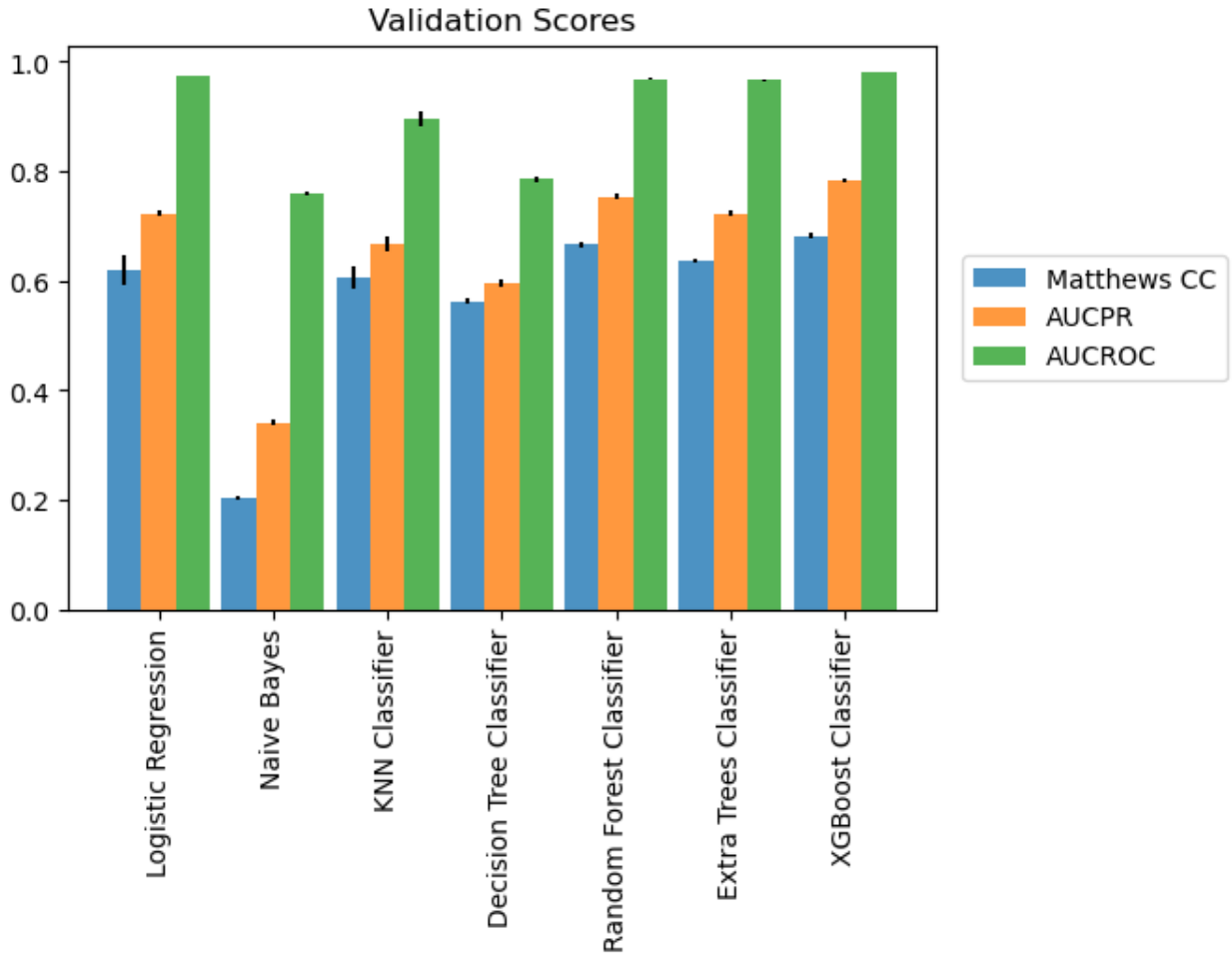
Figure 18: Confusion Matrices - Predict Helpful Reviews Using Extracted Features



As seen from the confusion matrices, the TF-IDF matrix is suitable for predicting the review rating, and the algorithms show much better performances. The TF-IDF matrix, combined with the extracted features, does not improve the model performance, either. Thus, we will use only the extracted features to predict the helpful reviews.

After deciding the set of features, we did cross-validation to see the algorithms' performances. The black bars represent the variation in the validation scores. ROC (Receiver Operating Characteristic) is the

Figure 19: Cross Validation Scores



highest score for each algorithm, and Matthews Correlation Coefficient is the lowest. However, Naïve Bayes has inconsistent scores when compared with the remaining algorithms. For this reason, it is dropped from the study. In Figures 20 and 21, we provided the ROC and PR curves.

The black dashed line on the ROC curves and the blue dashed line on the PR curves represent Dummy Classifier, which does not consider the input matrix while making predictions. For this reason, any algorithm which goes below the dashed line is considered 'garbage.'

After the cross-validation section, we trained the algorithms using the whole training data and test them on the test data. Tables 16 and 17 provide the training and test scores.

Figure 20: ROC (Receiver Operating Characteristic) Curves

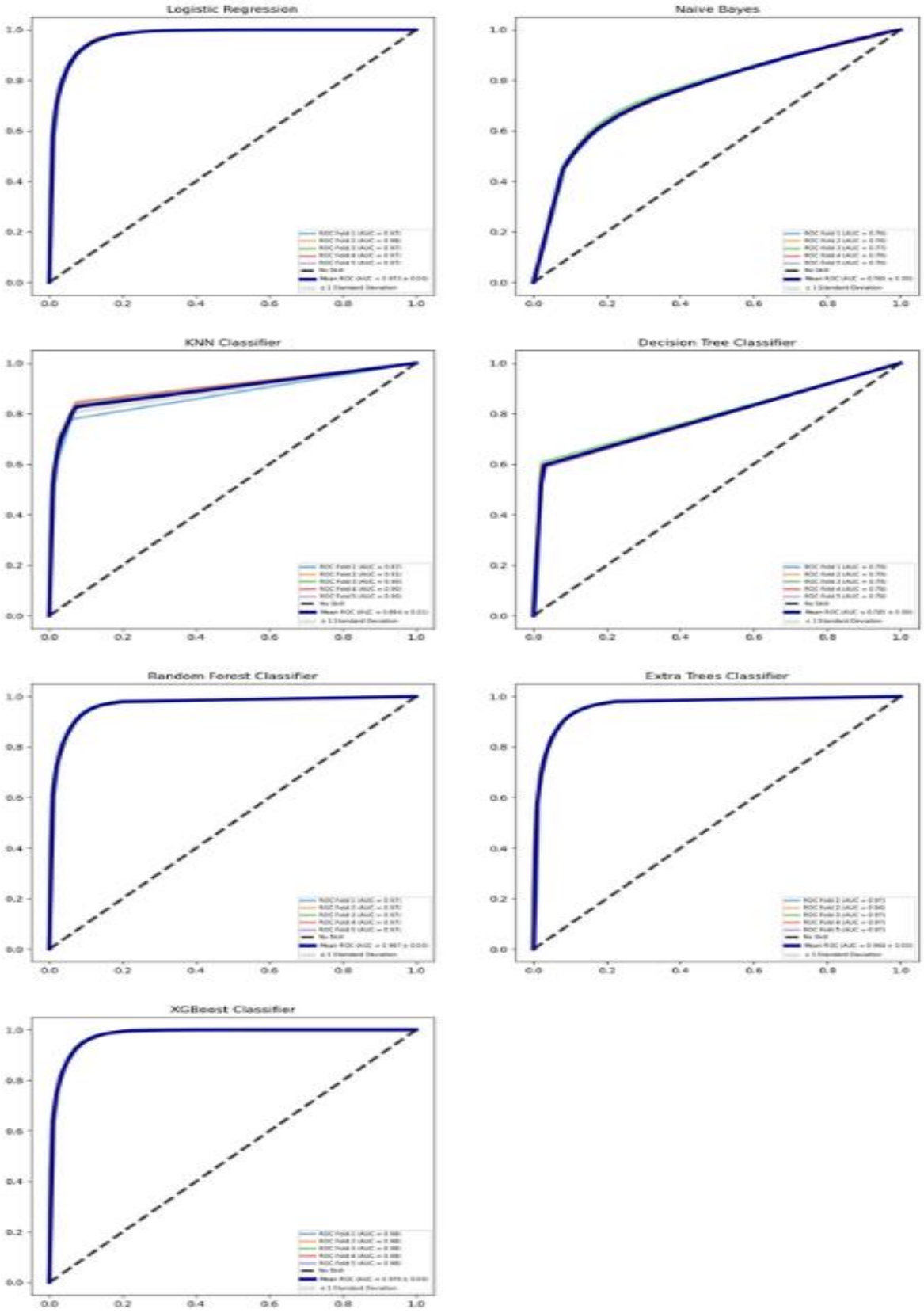


Figure 21: PR (Precision-Recall) Curves

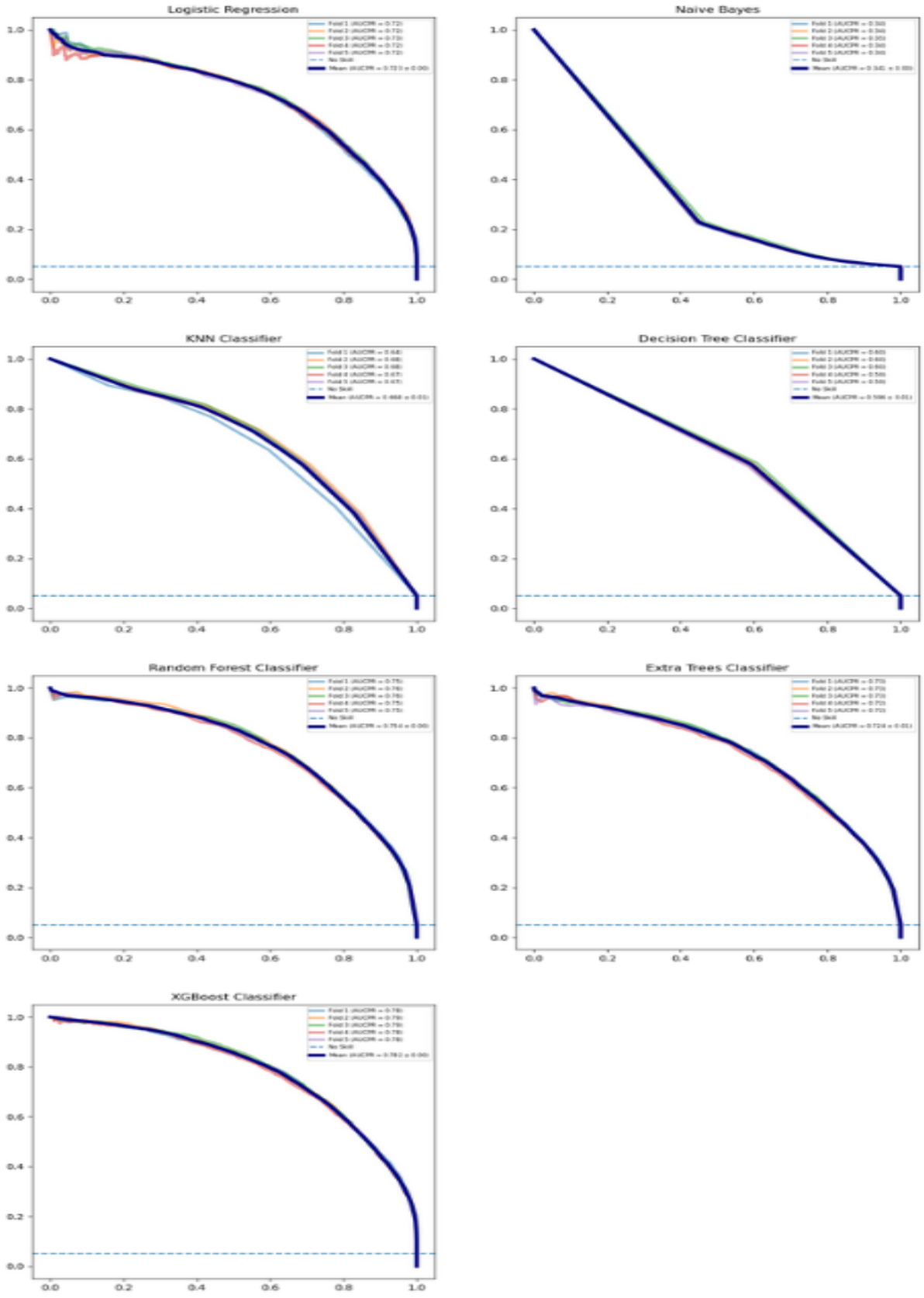


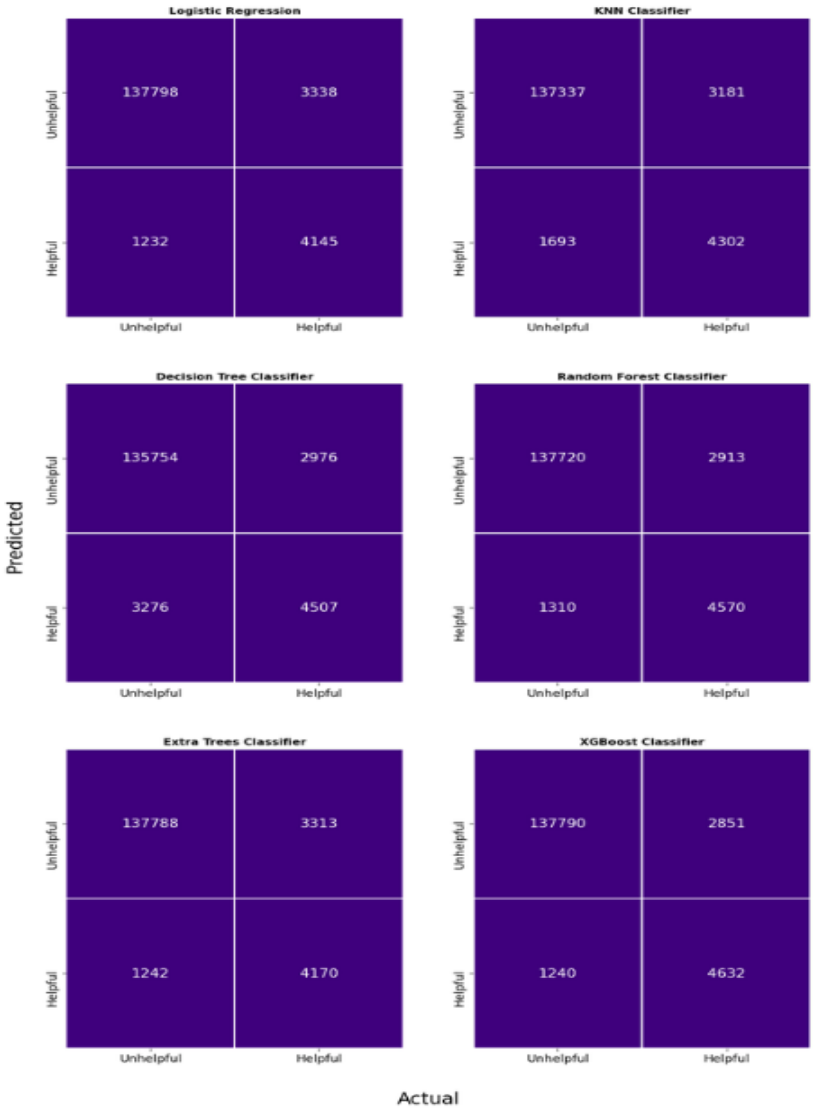
Table 16: Training Scores

	Matthews CC	ROC Score	PR Score
Logistic Regression	0.620788	0.974573	0.723073
KNN Classifier	0.709184	0.989005	0.845982
Decision Tree Classifier	0.999946	1.000000	1.000000
Random Forest Classifier	0.999750	1.000000	1.000000
Extra Trees Classifier	0.999946	1.000000	1.000000
XGBoost Classifier	0.703560	0.983709	0.813996

Table 17: Test Scores

	Matthews CC	ROC Score	PR Score
Logistic Regression	0.638187	0.975003	0.727144
KNN Classifier	0.625358	0.903029	0.677878
Decision Tree Classifier	0.568085	0.789368	0.600847
Random Forest Classifier	0.674447	0.970229	0.757467
Extra Trees Classifier	0.640015	0.964758	0.724107
XGBoost Classifier	0.684751	0.980687	0.785370

Figure 22: Confusion Matrices (Default Parameters)



Based on the cross-validation scores, we dropped Naïve Bayes from the study. The differences between the training and the set scores are not significant for ROC, but PR and Matthews CC scores differ significantly for tree-based algorithms (excluding XGBoost). Accordingly, those algorithms may not be good choices for our data. The confusion matrices are presented in Figure 22.

According to the test scores and the confusion matrices, we can say that the best algorithm for detecting helpful reviews is XGBoost, which has the lowest False Positive and False Negative and the highest True Positive scores.

Hyperparameter Optimization

In the hyperparameter optimization process, we used Stratified 5-Fold cross-validation with GridSearchCV. It tests the algorithms using a different part of the data from scrap in each iteration by going over each parameter in the hyperparameter space in each iteration. It may require a significant amount of time depending on the hyperparameter space's size, the algorithm's characteristics, and the dataset's size. Table 18 presents scores with default parameters and the optimized parameters.

We used the following parameters in the hyperparameter optimization process:

- Logistic Regression: **solver, C, class_weight**
- K-Nearest Neighbor Classifier: **n_neigbors**
- Decision Tree Classifier: **max_features, class_weight**
- Random Forest Classifier: **n_estimators, max_features, class_weight**
- Extra Trees Classifier: **n_estimators, max_features, class_weight**
- XGBoost Classifier: **eta, colsample_bytree**

Table 18: Scores with Default and Optimized Parameters

	Score (Default Parameters)	Score (Optimized Parameters)	Change
Logistic Regression	0.975	0.976	+ 0.001
KNN Classifier	0.897	0.950	+ 0.053
Decision Tree Classifier	0.785	0.785	0.000
Random Forest Classifier	0.969	0.976	+ 0.007
Extra Trees Classifier	0.965	0.972	+ 0.007
XGBoost Classifier	0.980	0.981	+ 0.001

As seen from the above table, KNN benefited the most from the hyperparameter optimization process. In the final step, we will train and test the algorithms with the optimized parameters and find the best algorithm to predict helpful reviews.

Final Evaluation

In the final evaluation step, we will assess algorithms based on two criteria:

1. Confusion matrix
2. Recall rate (over assigned 0.95 probability)

We aim to predict helpful reviews based on the extracted features but approaching predicted reviews as a pure classification problem may not benefit any parties. In other words, it will result in only two

classes of reviews, such as helpful and unhelpful. However, it does not say anything about how likely one review to be a helpful review. For this reason, we will use predicted probabilities as a second criterion and set the threshold as 0.95. By doing so, we hope to get reviews that are more likely to be helpful review.

Figure 23 and Table 19 provide confusion matrices and test scores, respectively.

Figure 23: Confusion Matrices - Optimized Parameters

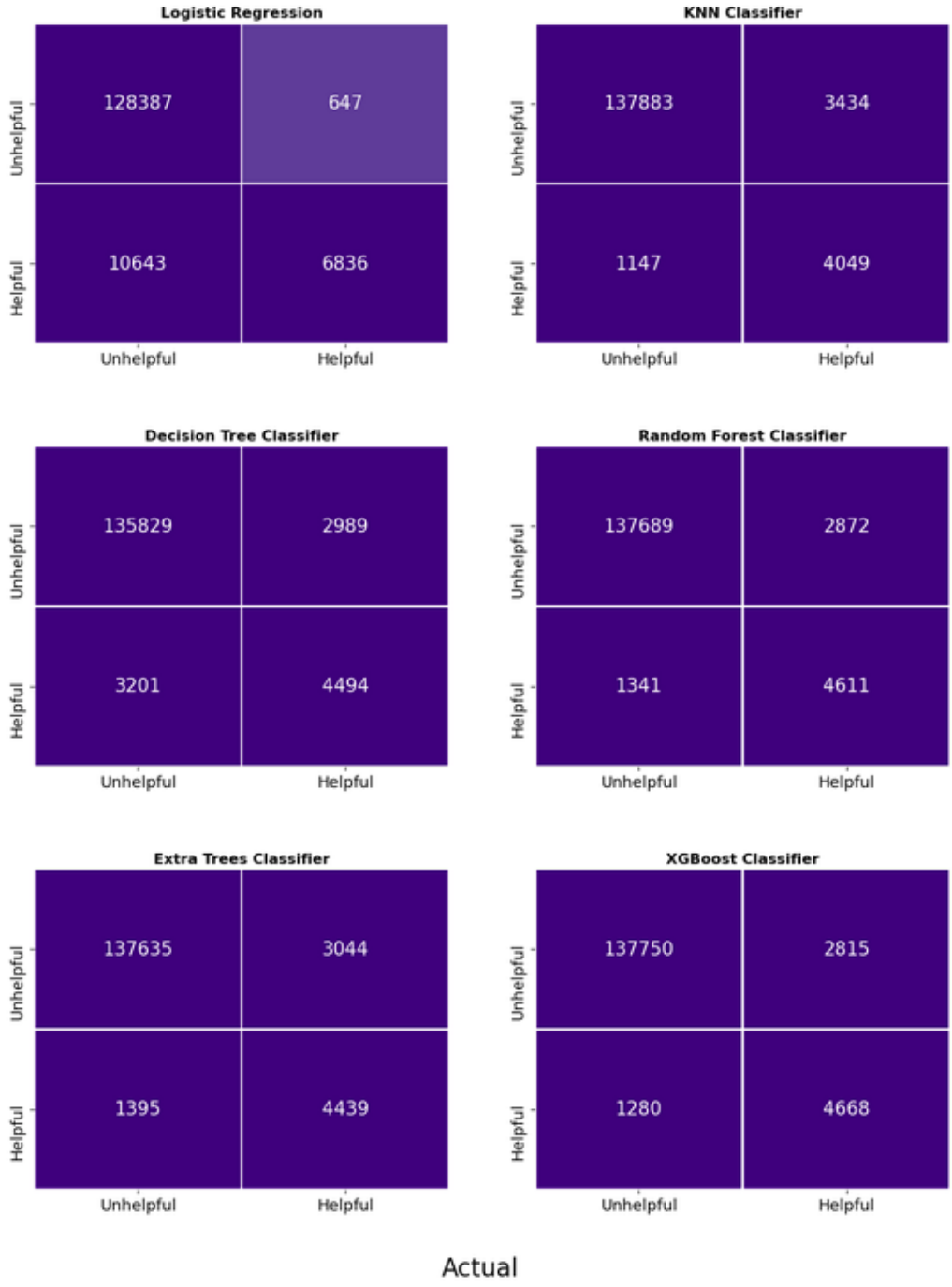


Table 19: Test Scores (Optimized Parameters)

	Matthews CC	ROC Score	PR Score
Logistic Regression	0.568460	0.976048	0.727485
KNN Classifier	0.634249	0.952136	0.720750
Decision Tree Classifier	0.569961	0.788769	0.602489
Random Forest Classifier	0.676388	0.976074	0.766688
Extra Trees Classifier	0.656591	0.972061	0.738463
XGBoost Classifier	0.685592	0.980665	0.783915

We want to build a model that recommends the reviews which are most likely to be helpful reviews rather than generating a pure classification model. For this reason, we are interested in the assigned probabilities, not the classes. Moreover, we implement a cut-off of 0.95 for the assigned probabilities. Finally, we evaluated the models based on their confusion matrices and the recall rate for the reviews with at least 0.95 assigned probability.

We located the reviews in the test set with the highest helpful votes and checked the algorithms' performance on those reviews. Among 146,513 reviews in the test set, there are only 11 reviews with the top 10 helpful votes. In the first place, we aim to decide if an algorithm can distinguish helpful reviews from unhelpful reviews. Also, we want to check how many of the predicted helpful reviews are among the ones that have the top 10 helpful votes.

In the confusion matrices, we see that:

1. Logistic Regression predicted the highest number of helpful reviews at the expense of false positives. Moreover, it has the lowest number of false negatives among the algorithms.
2. KNN predicted an average number of helpful reviews with the lowest false positive rate. However, it has the most significant number of false negatives.
3. All other algorithms stay in the spectrum where the edges are Logistic Regression and KNN algorithms.

Table 20 provides the recall rates of the algorithms for the reviews with 0.95 or above assigned probability.

Table 20: Recall Rate (Reviews 0.95 or above Assigned Probability)

	Recall Rate	Predicted Value		True Value
Logistic Regression	71.98 %	4,853	out of	6,742
KNN Classifier	95.16 %	904	out of	950
Decision Tree Classifier	58.40 %	4,497	out of	7,700
Random Forest Classifier	96.86 %	924	out of	954
Extra Trees Classifier	95.65 %	593	out of	620
XGBoost Classifier	97.91 %	656	out of	670

XGBoost and Random Forest are the top algorithms with the highest recall rates. On the other hand, Decision Tree and Logistic Regression have the lowest recall rates among the algorithms. For this reason, we can say that those algorithms have a more naive approach than the remaining algorithms.

In this project, our aim is not solely to approach the problem as a pure classification problem but to recommend among the freshly posted reviews that would have the highest helpful votes. Thus, we will investigate how much success an algorithm has in identifying helpful reviews with the highest number of helpful votes. Since those reviews are at the top, they should have assigned probabilities that are close to 1.0. Table 21 provides the necessary information.

Table 21: Recall Rate (Top 10 Helpful Reviews)

	Recall Rate	Predicted Value		True Value
Logistic Regression	100.00 %	11	out of	11
KNN Classifier	100.00 %	11	out of	11
Decision Tree Classifier	100.00 %	11	out of	11
Random Forest Classifier	72.73 %	8	out of	11
Extra Trees Classifier	81.82 %	9	out of	11
XGBoost Classifier	72.73 %	8	out of	11

We set 0.95 as the threshold for the assigned probabilities so that the business owners (the restaurants in this case) will be more likely to provide customers with the reviews that can attract their attention and present what they are looking for. They can also promote those reviews among the others so that it will be easier for the customers to reach the relevant information about the business and the product(s).

Based on the algorithms' performance in the confusion matrices and the top 5% predicted reviews, we can say that KNN is the most practical algorithm. Even though XGBoost has the best performing results, it has some flaws:

1. KNN hits a 100% recall rate for the top 10 helpful reviews, but XGBoost stays at 72.73%.
2. Even though KNN has a lower recall rate in general, it has the most significant number of correctly predicted helpful reviews.
3. KNN provides a broader pool of helpful reviews for the business owner to hand-pick if necessary.

For those reasons, we believe that KNN is the best algorithm for our purpose in this project. We will provide some examples in the next chapters.

The Most Important Features

In Figure 24, we provided the most important features to predict helpful reviews. According to the figure, the number of emoticons in a review is the most significant features to detect helpful reviews. It is followed by the average useful votes that the reviewer has and the reviews' star rating.

Learning Curves

Figure 25 provides the learning curves for the algorithms. It shows how the performance of an algorithm as it is fed with more data. We kept the test set stable and trained the algorithms using different training sets in different sizes, such as 1k, 10k, 50k, 100k, 250k, and whole training data.

Based on the figure, we can say KNN grows exponentially; however, the remaining algorithms do not significantly change the performance. Thus, we expect KNN to be more efficient as we provide a bulker training set. On the other hand, KNN, by its nature, is an extremely slow algorithm compared with XGBoost. It is possible to take advantage of parallel programming, but it will not beat XGBoost by any means. For this reason, it is possible to switch to another algorithm for time-saving purposes.

Figure 24: The Most Important Features

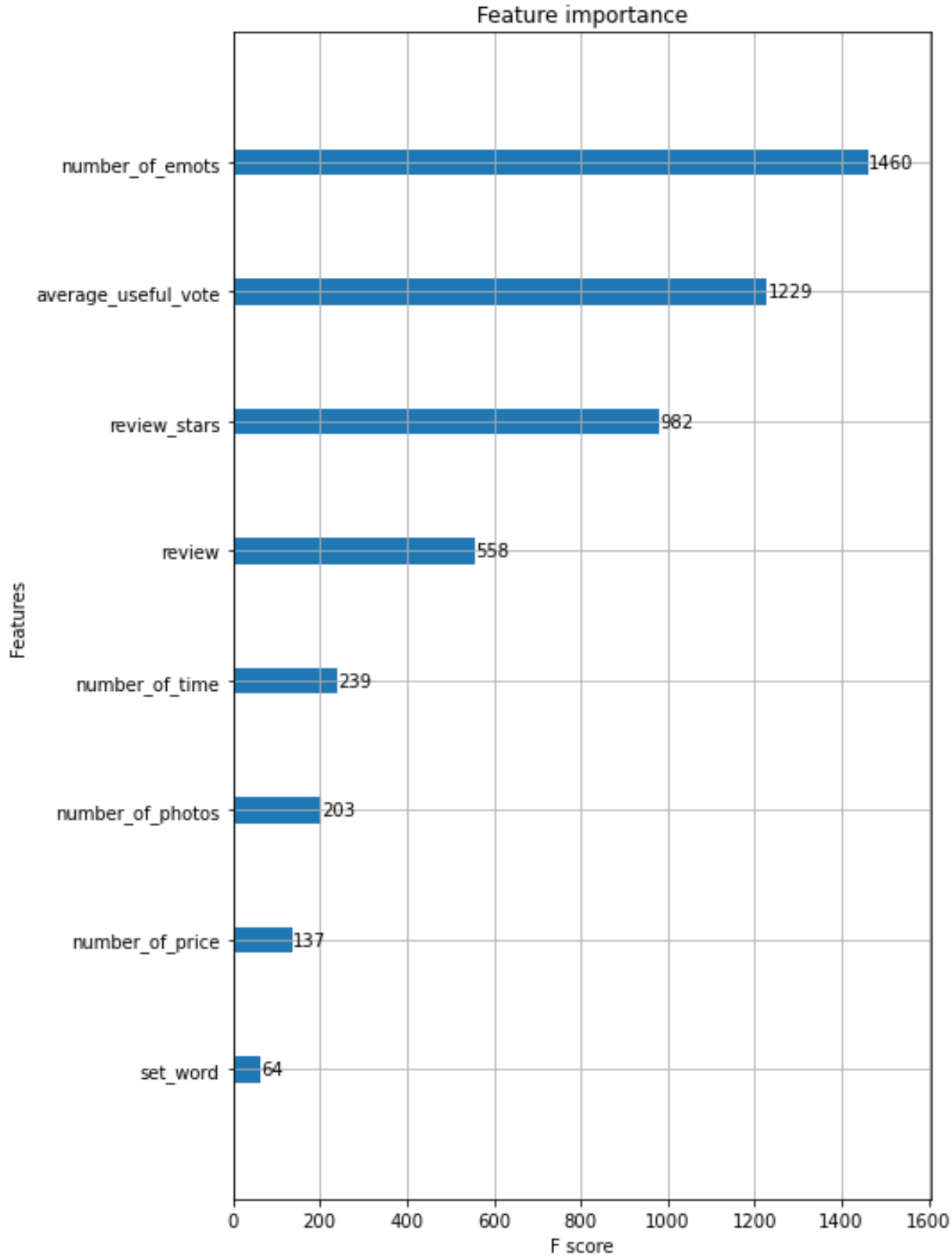
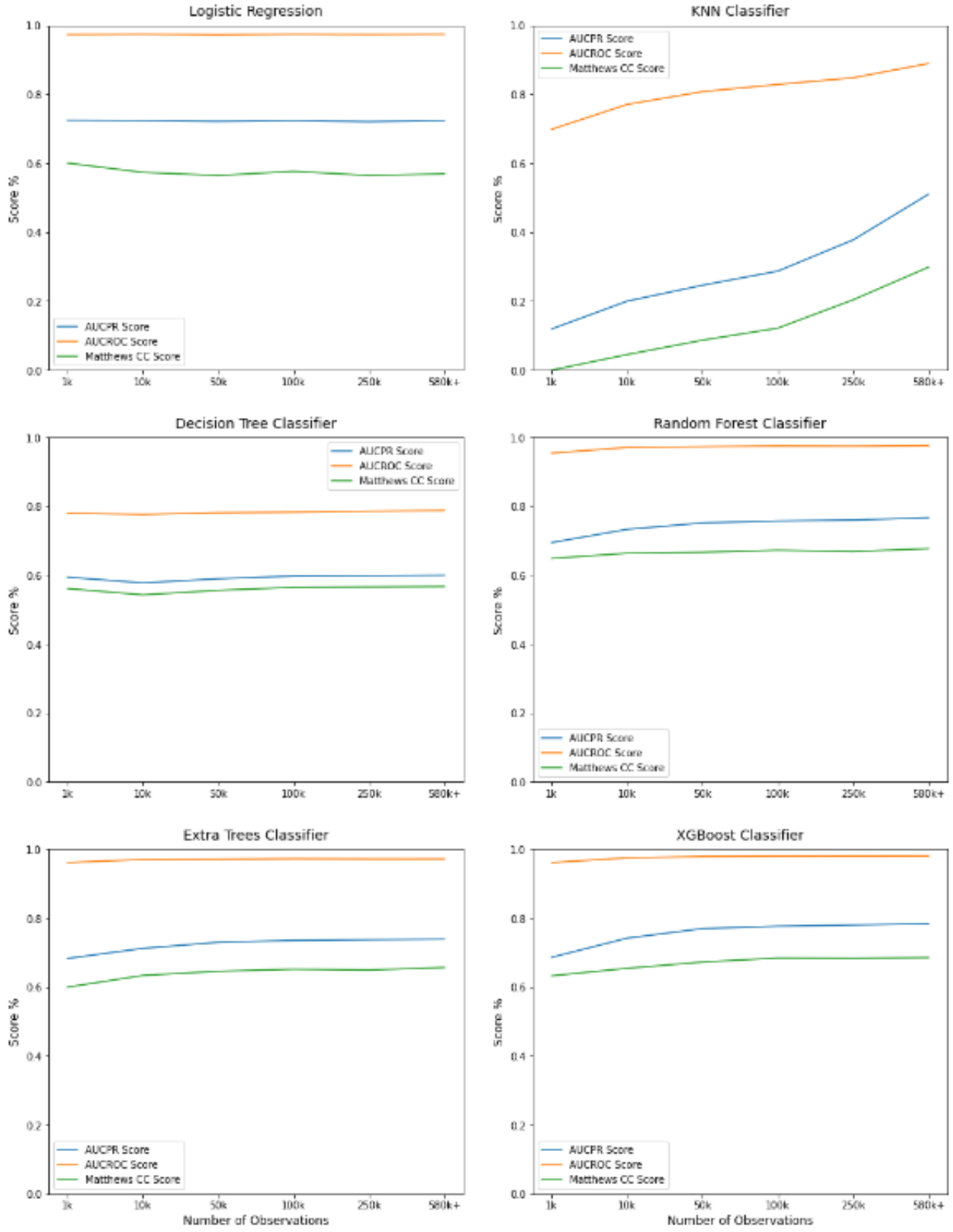


Figure 25: The Learning Curves



Conclusion

In this project, we aimed to develop a model that can predict if a freshly posted review will be a helpful review. If so, the businesses can benefit from this kind of an approach by promoting those reviews and letting the customers enjoy them.

A helpful review's essential components are the number of customers who voted for the review and the amount of time passed since it was posted. By doing so, we hoped to save the amount of time that would require a review to be recognized as a helpful review and provide those reviews for the customers' convenience in advance.

We started with the review corpus in order to identify the helpful reviews. First, we implemented text cleaning steps in order to ready the text for vectorization. Later, we used the TF-IDF method to vectorize the text and set the following cut-off points for the minimum, and the maximum number of documents for a word has to appear as 3% and 90%, respectively.

However, the TF-IDF vector was not an efficient way to identify helpful reviews as it was detecting the star rating of the reviews. For this reason, we employed the features extracted from the reviews, such as the number of photos, the number of price information, the average helpful vote that the writer has, etc. As a result, we improved the model performances.

Later, we did hyperparameter optimization using the extracted features to find the parameter values that explain the data best. Finally, in this notebook, we trained the models using the whole training set and evaluated them with the test set.

To increase the recall rate, we focused on the reviews that have at least 0.95 assigned probability. As a result, we got the highest recall rate by using XGBoost Classifier. However, the best result is acquired by the KNN classifier.

In the final step, we evaluated the model performances by training models with six subsets of the training set in which they differ in the number of reviews. However, we kept the test set stable. The aim is to get an idea of how the model performances change as the dataset grows. We found that the best improvement experienced by the KNN classifier.

As a result, we think that the best algorithm to identify helpful reviews with the given feature set and the given conditions such as 0.95 probability cut-off is the KNN algorithm. We claim KNN can detect the reviews that have the highest number of helpful votes with great accuracy. Even though its recall rate is lower than some of the other algorithms, it is best to recommend the highest number of helpful reviews.

Here, we give an example of helpful reviews that have assigned probability over 0.95 with the number of helpful votes. The business can benefit from machine learning applications to detect helpful reviews in advance and enhance their customer relationship.