

NLP Comparison of Models to Create a Survey Dialog Bot

Background

The goal of this exercise is to determine which classification technique can generate an accurate survey dialog for a restaurant business to better understand their customers. The most efficient model will be used to power an application that captures the sentiment of its customers and ranks them based on the underlying sentiment dataset. Appropriate responses will be generated by the application based on the response rankings to complete the survey dialog. The classification techniques to be tested are: Random Forest Classification, Logistic Regression and Naïve Bayes Classification.

Data Source

The dataset consists of two files: a flat csv file of a subset of Yelp reviews, as taken from Fan, 2019, and a json document of constructed dialog responses, based on the classification result rankings. The data dictionaries for each file can be seen in Appendix 1 and 2, respectively.

Data Transformation and Cleaning

Stars and Text Variables

The dataset was ultimately stripped of all other columns, except Stars and Text, as those were needed to run the binary analysis in the various classification techniques. Only the extreme star values of 1s and 5s were kept.

2-Star Observations

It was noted early in the data analysis that the dataset was dominated by 5-star ratings. 2-star ratings were reclassified as star-1 ratings and combined in attempts to balance out the dataset.

Funny Variable

The 5-star were reduced by filtering out all observations that given a 'Funny' score greater than 0 to help create further balance.

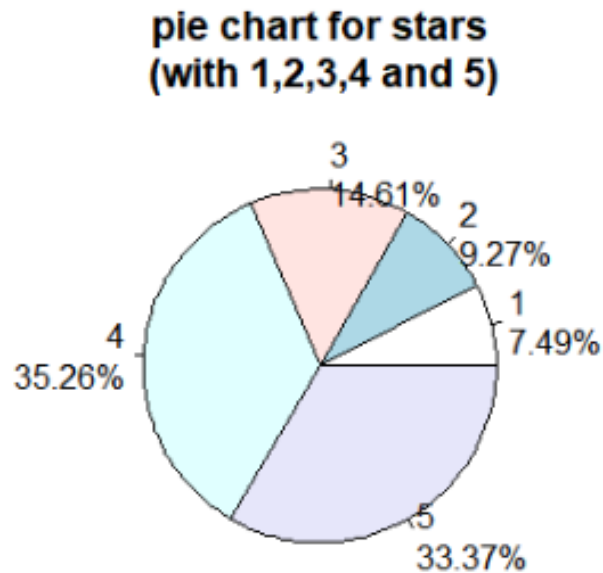
NLP Text Transformation and Cleaning

Clean Text

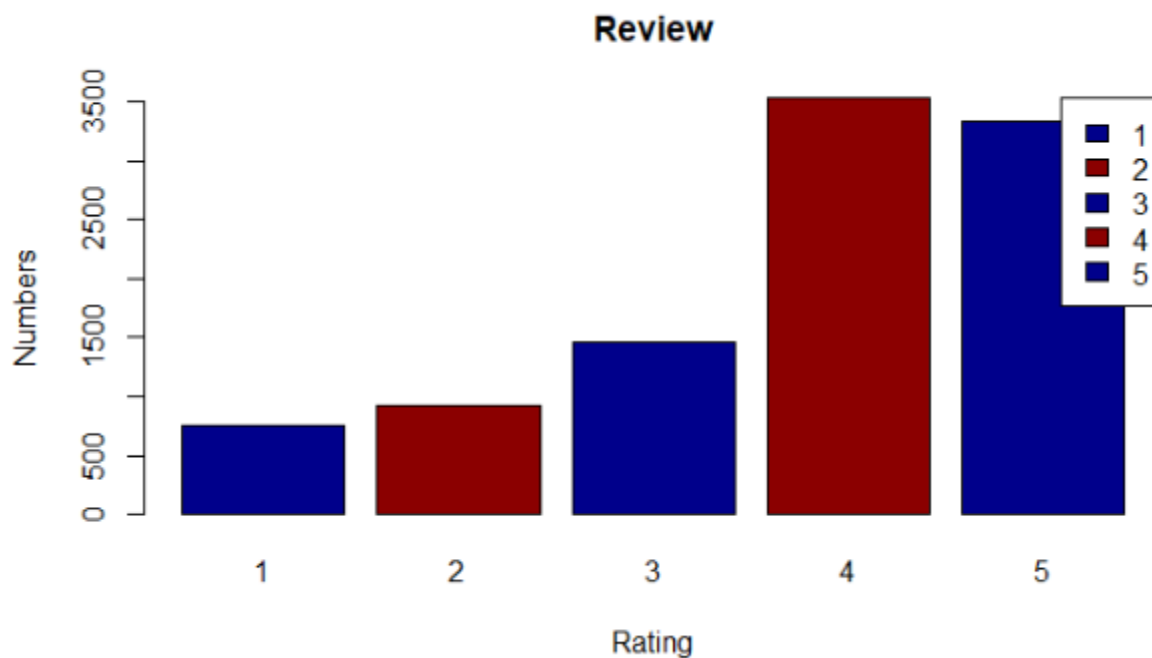
During the model preparation, the Text variable data was transformed. All unnecessary was stripped. The remaining text was tokenized into individual words and the unnecessary stem endings were removed from the stop-words. This was done in bulk for the Text variable, as well as for the individual user input text.

Data Analysis

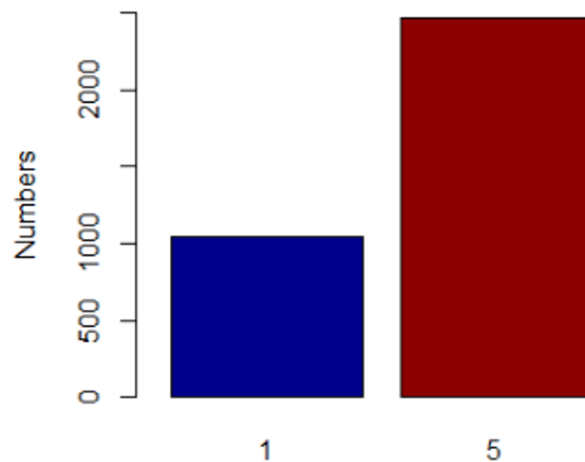
The original data was analysed to determine what type of dimension reduction would be needed to isolate only 1-star and 5-star ratings. The source code can be seen in Appendix 3.



As seen by the pie chart above, the 5-star data for the 'stars' variable held a larger number of observations than the 1-star rating.



This is bar graph, above, of the original data without any reductions and transformations which clearly depicts that we have very little 1's and 2's.



As a result, the 1-star and 2-star ratings were combined, and the 'funny' observations were removed from the 5-star data. This increased the ratio from 3:1 to 2:1, as seen above.

```
> summary(js)
business_id      date      review_id      stars      text
Length:10000    Length:10000    Length:10000    Min.   :1.000    Length:10000
Class :character Class :character Class :character 1st Qu.:3.000    Class :character
Mode  :character Mode  :character Mode  :character Median :4.000    Mode  :character
                                   Mean  :3.777
                                   3rd Qu.:5.000
                                   Max.  :5.000

type      user_id      cool      useful      funny
Length:10000 Length:10000    Min.   : 0.0000    Min.   : 0.000    Min.   : 0.0000
Class :character Class :character 1st Qu.: 0.0000    1st Qu.: 0.000    1st Qu.: 0.0000
Mode  :character Mode  :character Median : 0.0000    Median : 1.000    Median : 0.0000
                                   Mean  : 0.8768    Mean  : 1.409    Mean  : 0.7013
                                   3rd Qu.: 1.0000    3rd Qu.: 2.000    3rd Qu.: 1.0000
                                   Max.  :77.0000    Max.  :76.000    Max.  :57.0000

> |
```

The data variables, 'cool', 'useful' and 'funny' are non-normal data, especially since they are based on user sentiments. There are also many outliers, as seen with the IQR * 1.5 being lower than the Max values.

Model Development

Three datasets were randomly generated: a training dataset, containing 90%, a validation dataset containing 5% of the data, and a test dataset containing 5% of the data. Random Forest classification, Naïve-Bayes classification, and Logistic Regression models were generated to classify the Yelp text appropriately in a validation dataset. The datasets were re-generated for each model. A confusion matrix and various metrics were then calculated for each model. 1 represents star ratings of 1, while 5 represents star ratings of 5. The original data can be seen in Appendix 4.

Random Forest Classification Results (RNF)

RNF	Predicted					%	Type 1 Errors	17
Actual		1	5		Accuracy	87	Type 2 Errors	11
	1	67	17	84	Sensitivity	91		
	5	11	112	123	Specificity	80		
		78	129	207	Precision	87		

Naïve Bayes Classification Results (NBC)

The multinomial classification library in Python was used for this analysis.

NBC	Predicted					%	Type 1 Errors	30
Actual		1	5		Accuracy	83	Type 2 Errors	5
	1	54	30	84	Sensitivity	96		
	5	5	118	123	Specificity	64		
		59	148	207	Precision	80		

Logistic Regression Classification Results (LGT)

LGT	Predicted					%	Type 1 Errors	17
Actual		1	5		Accuracy	87	Type 2 Errors	9
	1	67	17	84	Sensitivity	92		
	5	9	114	123	Specificity	80		
		76	131	207	Precision	87		

Only the Random Forest Classification Logistic Regression models produced an ideal accuracy over 85%. Both models had very similar results. However, it was noted that while regenerating random datasets, the Random Forest model would fluctuate between accuracies of 83% and 87%, while the Logistic Regression model remained consistent around 87%. Hence, the Logistic Regression model was selected to generate the NLP survey classifications. The model with the test dataset can be seen below.

Logistic Regression Classification Results (LGT) with Test Dataset

LGT	Predicted			
Actual		1	5	
	1	74	10	84
	5	10	114	124
		84	124	208

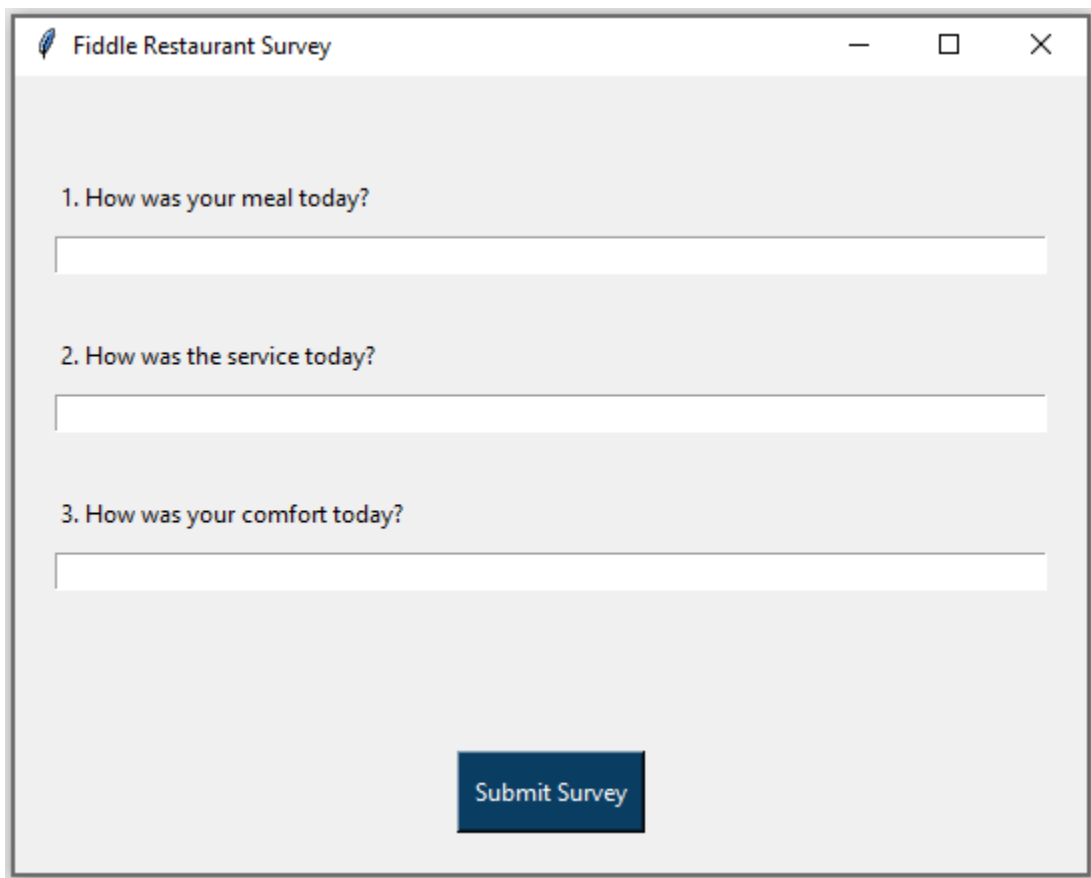
	%
Accuracy	90
Sensitivity	92
Specificity	88
Precision	92

Type 1 Errors	10
Type 2 Errors	10

Observations on NLP Survey Dialog

The survey dialog consists of three questions. Each user response is run through the classification model and the resulting probability taken from extremes values is ranked from 1 (worst) to 5 (best) to create a stand-alone classification system.

The rating is then passed to the bot dialog to respond with an appropriate message. The average score of the survey is also passed to the bot dialog to respond with an appropriate summary message. Various scenarios, grouped by trials, were simulated to see how well the application dialog responded and the results are as follows. Appendix 5 shows some of the trial responses tested and screenshots.



Fiddle Restaurant Survey

1. How was your meal today?

2. How was the service today?

3. How was your comfort today?

Submit Survey

Trial 1: All positive responses submitted

This resulted in correct dialog responses for both individual survey questions and the summary.

Trial 2: All negative responses submitted

This resulted in correct dialog responses for both individual survey questions and the summary.

Trial 3: 2 negative responses and 1 positive response submitted

This resulted in the appropriate dialog responses. The average survey score was in the medium range and the summary response reflected this.

Trial 4: 2 positive responses and 1 negative response submitted

This resulted in the appropriate dialog responses. The average survey score was in the higher range and the summary response reflected this.

Trial 5: All neutral responses submitted

This resulted in either classifying the results as a mix of high positive or high negative ranks. Sometimes, the average could result in a medium score, but not because the text was being classified as all neutral.

Trial 6: All unrelated responses submitted

Responses that have nothing to do with sentimental reviews were used, such as objective headlines. This resulted in all neutral rankings.

Trial 7: Interesting results from responses submitted

- Adding the word 'atmosphere' to a negative comment resulted in a 4 or 5-star rating.
- Some neutral comments, such as 'It was ok' received a 1-star rating.
- Other neutral comments, such as 'I was full' or 'Same as usual' received a 4 or 5-star ratings.
- Gibberish text received a 4 or 5-star ratings.
- Use of double negatives, such as 'I didn't hate this at all' received a 1 to 2-star rating.

Discussion on NLP Survey Dialog

Model Behaviour

The model behaved appropriately to the survey answers, given the circumstances under which it was created. It was trained with data that had a bias towards a very positive rating (5-star) and it was only given the two extreme ratings to work with (1-star and 5-star). The accuracy of identifying those types of responses was high and reflected its accuracy of 90%, seen in the confusion matrix. This test does prove that the data used is very important.

Biased Dataset

The Yelp dataset used was heavily dominated by 5-star ratings. Even after some dimensionality reduction, these ratings were still in a larger proportion than 1-star ratings. This can lead to higher chances of a word being categorized as positive. It is possible that words such as 'atmosphere' were found mostly in 5-star ratings and as a result, aided in classifying the text as positive.

Words taken out of context

The misclassification of double negatives suggests that insufficient contextual classification was present. It is possible that this is more of an advanced topic and should be considered at a future date. This also includes gibberish words and possible profanity.

Lack of a middle tier in the analysis

This analysis was based solely on a binary classification of 1-star and 5-star. As a result, the model had to decide what to categorize as intermediate responses. Only responses with around 50% probability of being a 1 or a 5 ended up in this category. These were mostly the unrelated text messages noted above. Furthermore, neutral responses related to sentimental reviews were classified as one of the extreme ratings. This is possibly a result of similar words being found in those categories, due to a lack of a category on its own.

Suggestions for Improvement

- Extend the Dataset
 - Try the original Yelp dataset and reduce accordingly as necessary.
 - Include dialogs to handle unrelated customer responses
- More Involved Data Cleanup
 - Perform deeper data analysis to reduce the data further for a more balanced dataset.
- Include classification of 3-star ratings
 - Switch to multi-classification techniques to accommodate non-binary data.

Conclusion

Natural Processing Language is a very powerful machine learning tool. However, it will only perform as extensively as the data and scope it is given. This exercise is only the beginning to building a more efficient survey dialog bot.

References

Fan, Z. (2019). *NLP for Yelp Reviews*. Kaggle. Retrieved August 8, 2020 from

<https://www.kaggle.com/zhenyufan/nlp-for-yelp-reviews?select=yelp.csv>

APPENDIX 1: Data Dictionary for Yelp Reviews (Fan, 2019)

Name	Description
Business_id	Unique Identifier for business
Date	Reviews collected between 17 April 05 and 4 Jan 2013
Review_id	Unique identifier for reviews
Stars	Star rating from 1 (worst) to 5 (best)
Text	User written sources of text
Type	Type of content for the text field
User_id	Unique identifier for users
Cool	Grouping of reviews that users flagged as cool
Useful	Grouping of reviews that users flagged as useful
Funny	Grouping of reviews that users flagged as funny

APPENDIX 2: Data Dictionary for Restaurant Business Bot Response Dialog

Name	Message
Star1	:(We seem to be severely failing your expectations in this area.
Star2	:(We seem to be failing your expectations in this area.
Star3	: We seem to be barely meeting your expectations in this area.
Star4	:) We seem to be meeting your expectations in this area.
Star5	:) We seem to be exceeding your expectations in this area.
Overall1	Overall, you had more of a negative experience. We will make a serious effort to address your concerns.
Overall2	Overall, you had more of a neutral experience. We will strive to improve to serve you better.
Overall3	Overall, you had more of a positive experience. We will continue to improve for you.

APPENDIX 3: Data Analysis

Creating pie chart of original data, using R

```
# we are putting the contents of star column in a table so that we can make pie chart from it
js2<-table(js$stars)
labels<- paste(names(js2),"\n",js2*100/10000,"%",sep="")
pie(js2,labels = labels,main="pie chart for stars\n (with 1,2,3,4 and 5)")
```

Creating bar plots after each data transformation, using R

```
#bar cahrt for the ratings
```

```
barplot(js2, xlab='Rating',ylab='Numbers',main="Review",
        col=c("darkblue","darkred"),
        ,legend=rownames(js2), args.legend = list(x = "topright"))
#barchart for all the numbers
js1<-table(js$stars[1:1:5])
labels<- paste(names(js1),"\n",js1,sep="")
tv<-pie(js1,labels = labels,main="pie chart for stars\n (with 1,2,3,4 and 5)")
```

```
barplot(js1, xlab='Rating',ylab='Numbers',main="Reviews",
        col=c("darkblue","darkred"),
        ,legend=rownames(1))
```

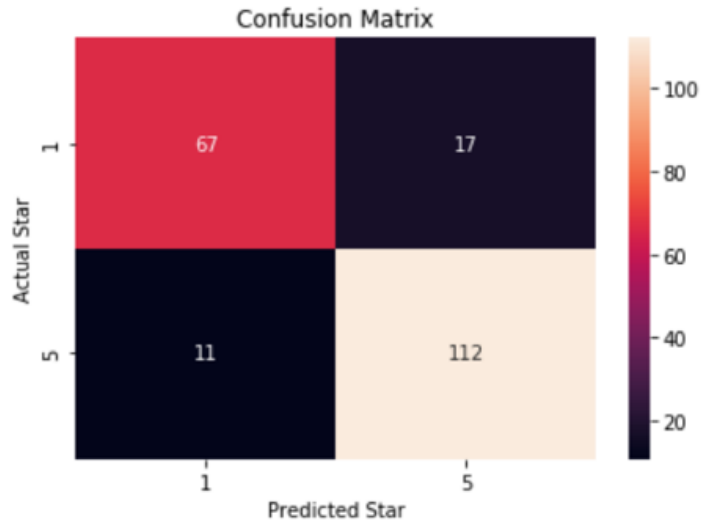
```
summary(rjs)
rjs1<-table(rjs$stars)
labels<- paste(names(rjs1),"\n",rjs1,sep="")
pie(rjs1,labels = labels,main="pie chart for stars\n (with 1,2,3,4 and 5)")
#BAr chart when we make change in the data and convert all 2's into 1's
barplot(rjs1, xlab='Rating',ylab='Numbers',main="Reviews",
        col=c("darkblue","darkred"),
        ,legend=rownames(1))
```

```
fjs1<-table(fjs$stars)
labels<- paste(names(fjs1),"\n",fjs1,sep="")
pie(fjs1,labels = labels,main="pie chart for stars\n (with 1,2,3,4 and 5)")
#Now here again we make reductions in the data and remove the funny>0 for star rating 5
barplot(fjs1, xlab='Rating',ylab='Numbers',main="Reviews",
        col=c("darkblue","darkred"),
        ,legend=rownames(1))
```

APPENDIX 4: Model Results

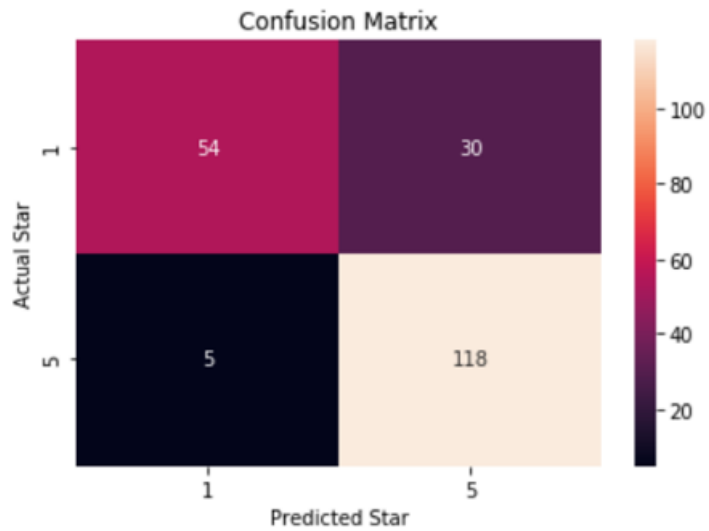
Generating Random Forest Validation Model Results

Precision: 0.868 / Recall: 0.911 / F1-Score: 0.889 / Accuracy: 0.865



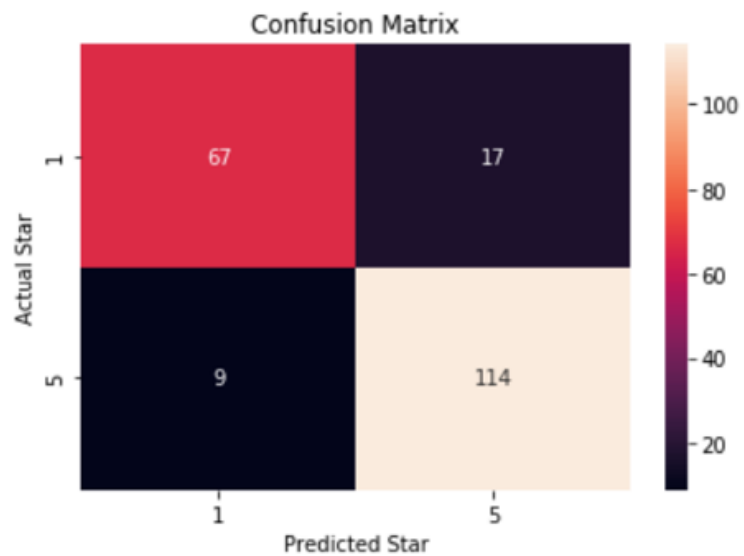
Generating Naïve Bayes Validation Model Results

Precision: 0.797 / Recall: 0.959 / F1-Score: 0.871 / Accuracy: 0.831



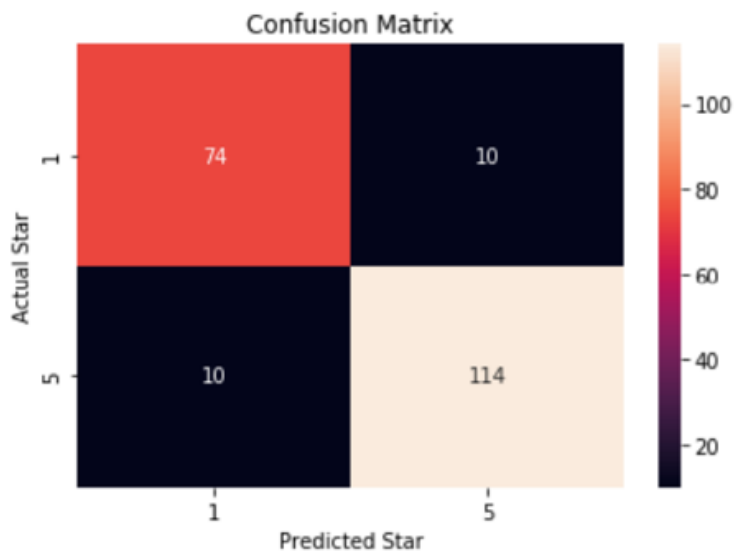
Generating Logistic Regression Validation Model Results

Precision: 0.87 / Recall: 0.927 / F1-Score: 0.898 / Accuracy: 0.874



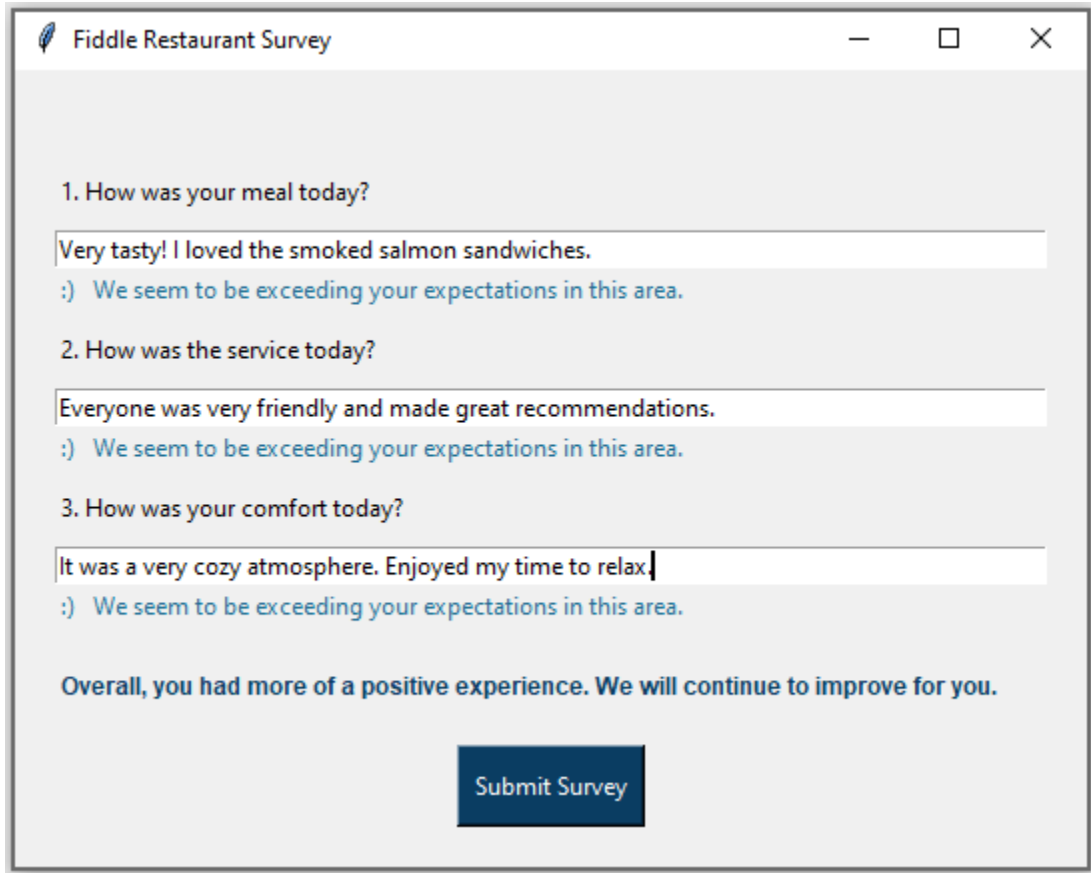
Generating Logistic Regression Test Model Results

Precision: 0.919 / Recall: 0.919 / F1-Score: 0.919 / Accuracy: 0.904



APPENDIX 5: Test Trials

Trial 1 Example: All positive responses submitted



Fiddle Restaurant Survey

1. How was your meal today?

Very tasty! I loved the smoked salmon sandwiches.

:) We seem to be exceeding your expectations in this area.

2. How was the service today?

Everyone was very friendly and made great recommendations.

:) We seem to be exceeding your expectations in this area.

3. How was your comfort today?

It was a very cozy atmosphere. Enjoyed my time to relax.

:) We seem to be exceeding your expectations in this area.

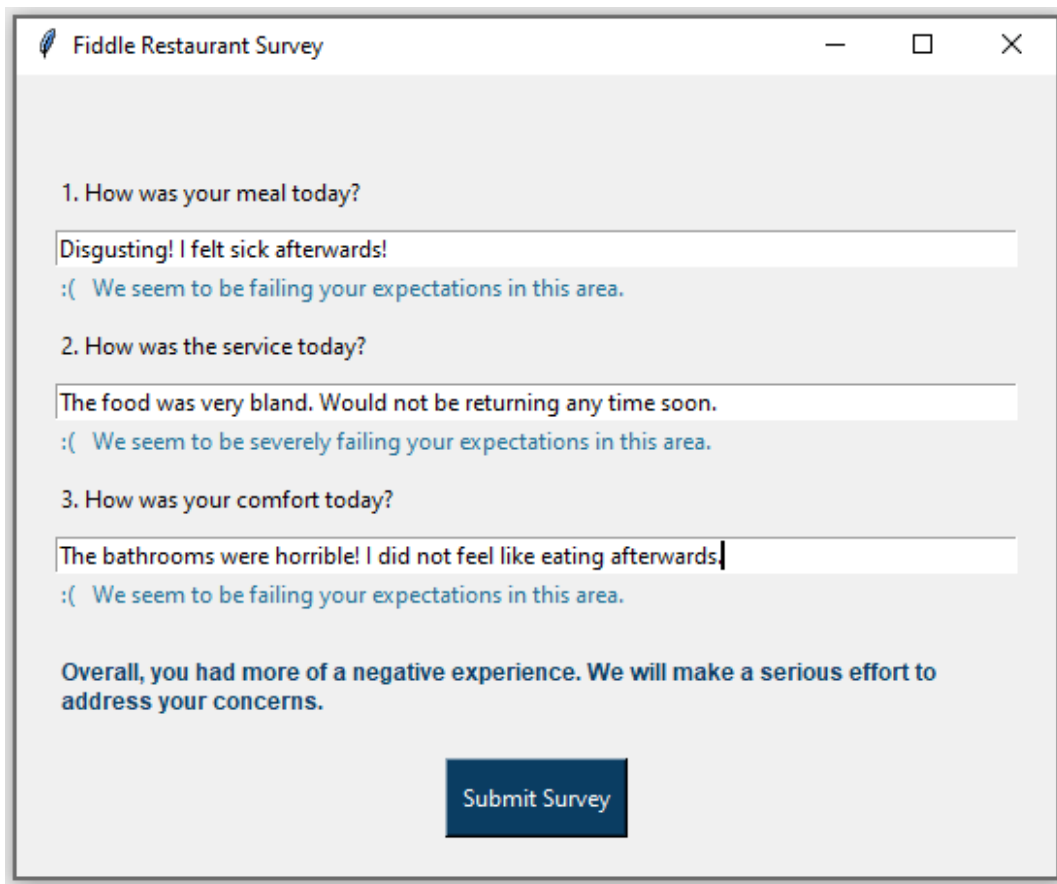
Overall, you had more of a positive experience. We will continue to improve for you.

Submit Survey

Responses used:

- Very tasty! I loved the smoked salmon sandwiches.
- Everyone was very friendly and made great recommendations.
- It was a very cozy atmosphere. Enjoyed my time to relax.

Trial 2 Example: All negative responses submitted



The screenshot shows a web browser window titled "Fiddle Restaurant Survey". It contains three questions with negative responses:

1. How was your meal today?
Disgusting! I felt sick afterwards!
:(We seem to be failing your expectations in this area.

2. How was the service today?
The food was very bland. Would not be returning any time soon.
:(We seem to be severely failing your expectations in this area.

3. How was your comfort today?
The bathrooms were horrible! I did not feel like eating afterwards.
:(We seem to be failing your expectations in this area.

Overall, you had more of a negative experience. We will make a serious effort to address your concerns.

Submit Survey

Responses used:

- Disgusting! I felt sick afterwards!
- The food was very bland. Would not be returning any time soon.
- The bathrooms were horrible!! I did not feel like eating afterwards.

Trial 3 Example: 2 negative responses and 1 positive response submitted

Fiddle Restaurant Survey

1. How was your meal today?

The produce quite bad quality today.

:(We seem to be failing your expectations in this area.

2. How was the service today?

Mad that the waiters took a very long time to respond...

:(We seem to be failing your expectations in this area.

3. How was your comfort today?

I was at least able to relax in a private setting.

:) We seem to be meeting your expectations in this area.

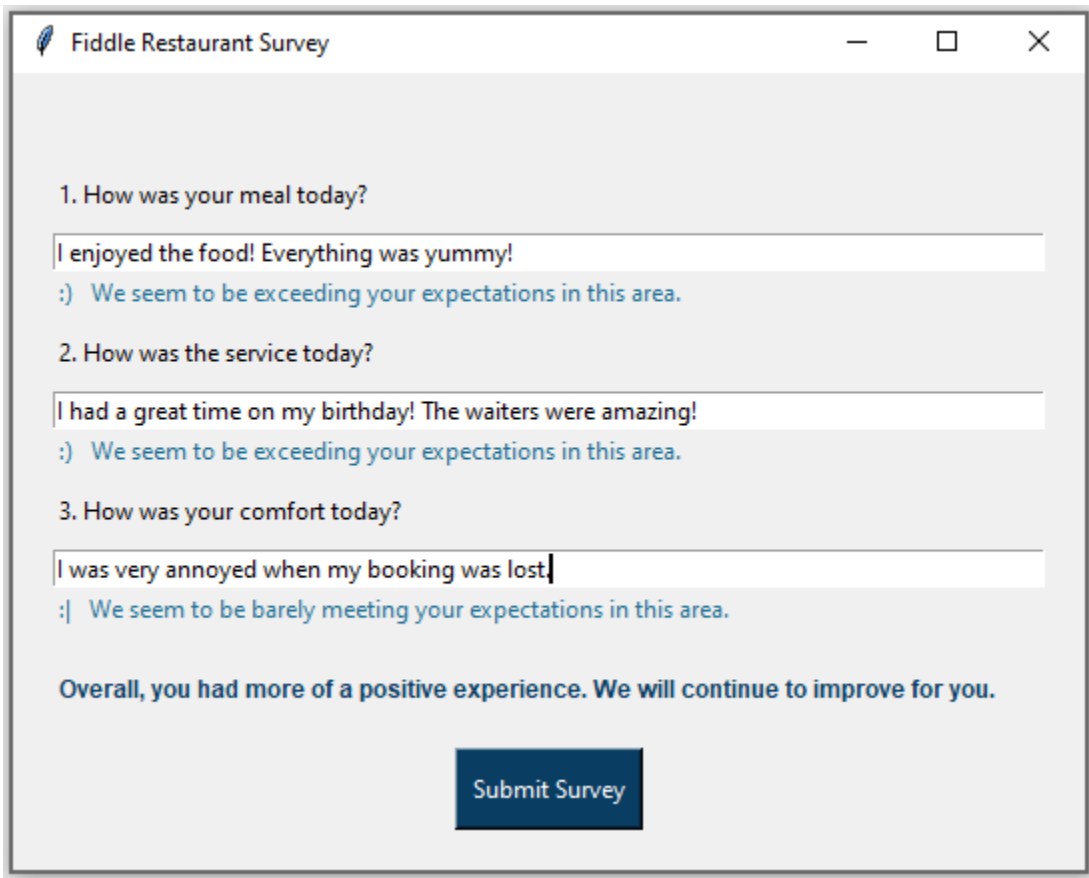
Overall, you had more of a neutral experience. We will strive to improve to serve you better.

Submit Survey

Responses used:

- The produce quite bad quality today.
- Mad that the waiters took a very long time to respond...
- I was at least able to relax in a private setting.

Trial 4 Example: 2 positive responses and 1 negative response submitted



The screenshot shows a web browser window titled "Fiddle Restaurant Survey". It contains three questions with text input fields and feedback messages:

1. How was your meal today?
I enjoyed the food! Everything was yummy!
:) We seem to be exceeding your expectations in this area.
2. How was the service today?
I had a great time on my birthday! The waiters were amazing!
:) We seem to be exceeding your expectations in this area.
3. How was your comfort today?
I was very annoyed when my booking was lost
:| We seem to be barely meeting your expectations in this area.

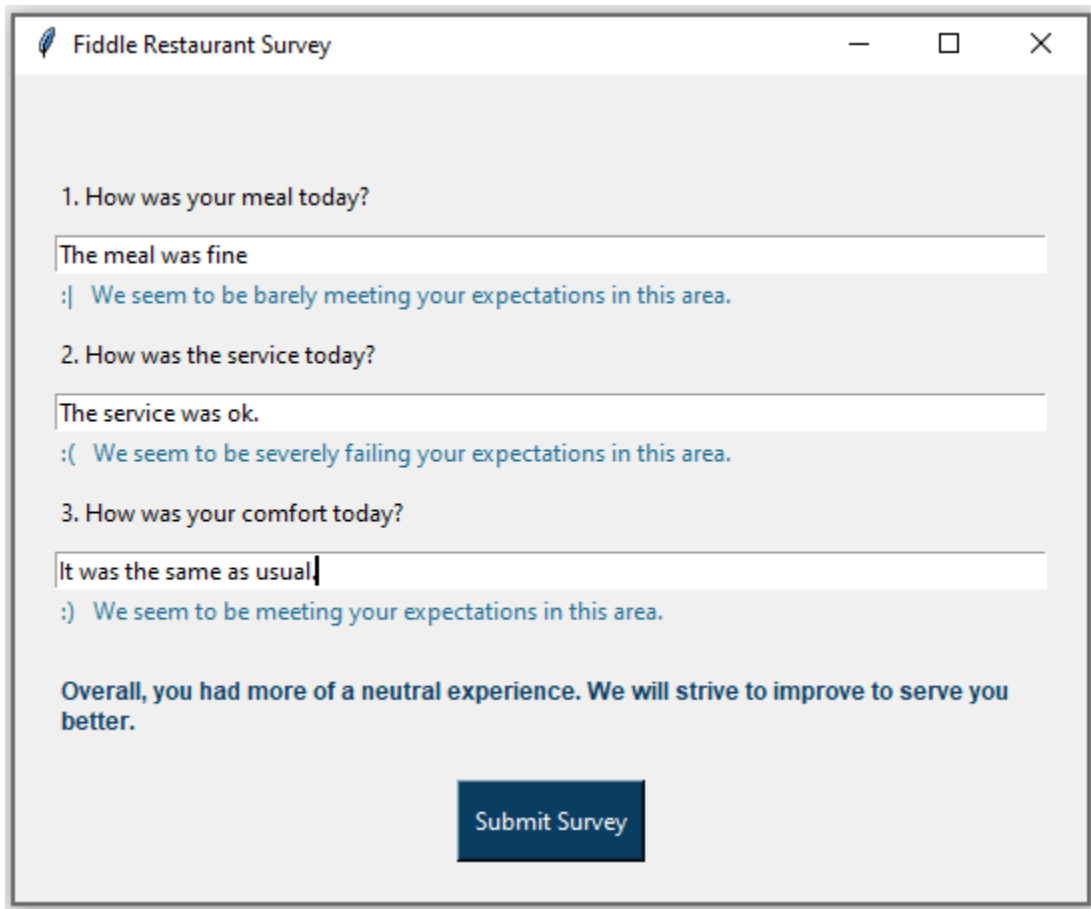
Overall, you had more of a positive experience. We will continue to improve for you.

Submit Survey

Responses used:

- I enjoyed the food! Everything was yummy!
- I had a great time on my birthday! The waiters were amazing!
- I was very annoyed when my booking was lost.

Trial 5 Example: All neutral responses submitted



Fiddle Restaurant Survey

1. How was your meal today?

The meal was fine

:| We seem to be barely meeting your expectations in this area.

2. How was the service today?

The service was ok.

:(We seem to be severely failing your expectations in this area.

3. How was your comfort today?

It was the same as usual

:) We seem to be meeting your expectations in this area.

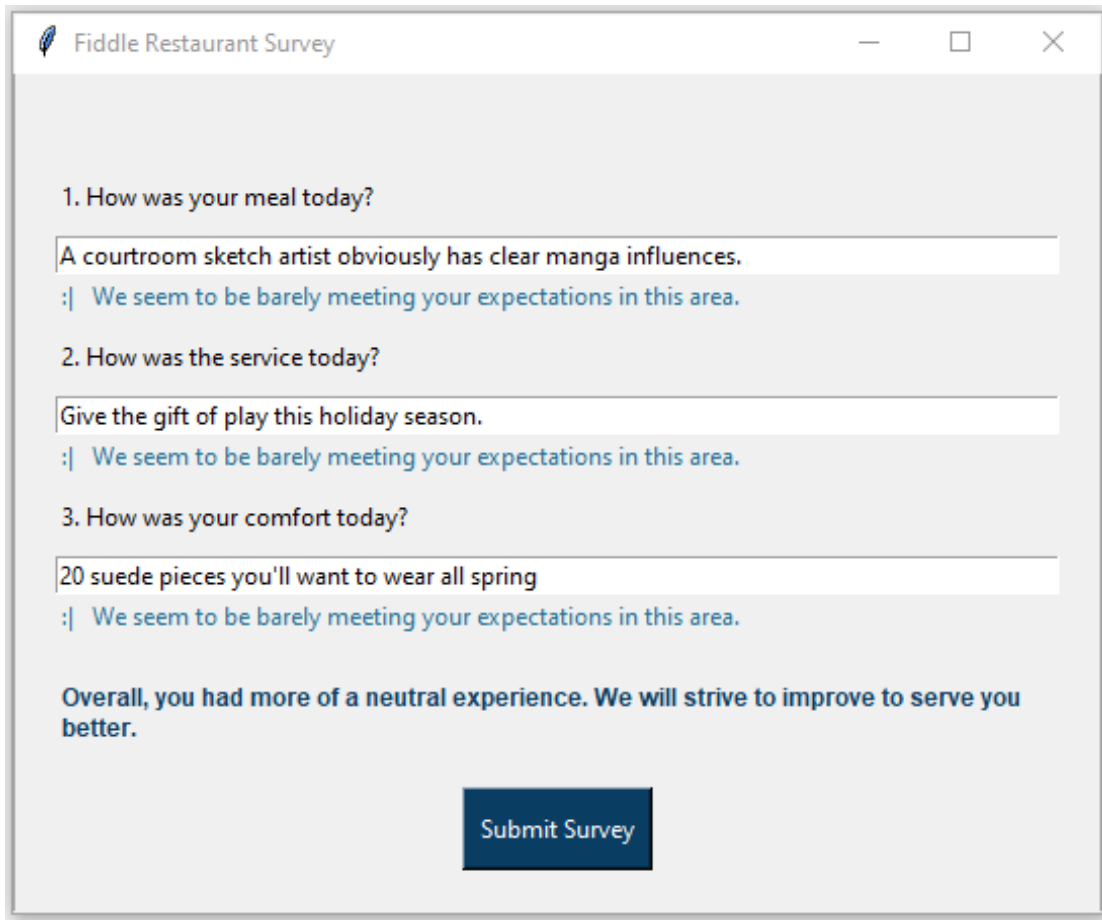
Overall, you had more of a neutral experience. We will strive to improve to serve you better.

Submit Survey

Responses used:

- The meal was fine
- The service was ok.
- It was the same as usual.

Trial 6 Example: All unrelated responses submitted



The screenshot shows a web browser window titled "Fiddle Restaurant Survey". It contains three questions, each with a text input field and a feedback message. The responses are unrelated to the questions.

1. How was your meal today?
A courtroom sketch artist obviously has clear manga influences.
:| We seem to be barely meeting your expectations in this area.

2. How was the service today?
Give the gift of play this holiday season.
:| We seem to be barely meeting your expectations in this area.

3. How was your comfort today?
20 suede pieces you'll want to wear all spring
:| We seem to be barely meeting your expectations in this area.

Overall, you had more of a neutral experience. We will strive to improve to serve you better.

Submit Survey

Responses used:

- A courtroom sketch artist obviously has clear manga influences.
- Give the gift of play this holiday season.
- 20 suede pieces you'll want to wear all spring

Trial 7 Examples: Interesting results from responses submitted

The image shows two side-by-side screenshots of a web form titled "Fiddle Restaurant Survey". Each form contains three questions and a feedback message.

Left Screenshot:

- Question 1: "1. How was your meal today?"
Input: "The setting was uncomfortable, so I did not eat."
Feedback: ":| We seem to be barely meeting your expectations in this area."
- Question 2: "2. How was the service today?"
Input: "Horrible atmosphere includes the service!"
Feedback: ":(" We seem to be failing your expectations in this area."
- Question 3: "3. How was your comfort today?"
Input: "It was a very gloomy atmosphere. A bit depressing"
Feedback: ":) We seem to be meeting your expectations in this area."

Overall, you had more of a neutral experience. We will strive to improve to serve you better.

Submit Survey

Right Screenshot:

- Question 1: "1. How was your meal today?"
Input: "I was full"
Feedback: ":) We seem to be meeting your expectations in this area."
- Question 2: "2. How was the service today?"
Input: "gfcj ktfgekuygu fdxj"
Feedback: ":) We seem to be meeting your expectations in this area."
- Question 3: "3. How was your comfort today?"
Input: "I did not have a horrible time"
Feedback: ":(" We seem to be severely failing your expectations in this area."

Overall, you had more of a neutral experience. We will strive to improve to serve you better.

Submit Survey

Responses used:

- The setting was uncomfortable, so I did not eat.
- Horrible atmosphere includes the service!
- It was a very gloomy atmosphere. A bit depressing.
- I was full
- gfcj ktfgekuygu fdxj
- I did not have a horrible time
- Was not happy about that.