Data Appendix

Project: SentiCook
Author(s): Srujana Yalamanchili, Henry Allen, Anika Tripathi
Date: 09/26/2025
Analysis file: Recipe Reviews and User Feedback Dataset.csv

1. Analysis Data File

1.1 Unit of Observation

Each row in the analysis file represents a single user review of a specific recipe at a point in time.

1.2 Scope of the Analysis File

The analysis file contains only complete-case observations. All rows with any missing values were removed, and fields that were not required for analysis were dropped to streamline the dataset.

1.3 Provenance & Processing

The analysis file was created from the raw file Recipe Reviews and User Feedback Dataset.csv. First, all rows containing missing values were removed. Next, the following fields were dropped because they were not necessary for the analysis: recipe_code, comment_id, user_id, user_name, created_at, reply_count, and best_score. During analysis, we created the variable star_bin from the original stars rating to facilitate interpretation; the categories are neg for ratings 0–2, neu for a rating of 3, and pos for ratings 4–5. We also compute derived engagement measures where noted below.

1.4 Reproducibility

The steps described above can be reproduced using the code already documented in the Github. Tables and figures referenced in this appendix come from our exploratory analysis and model output findings.

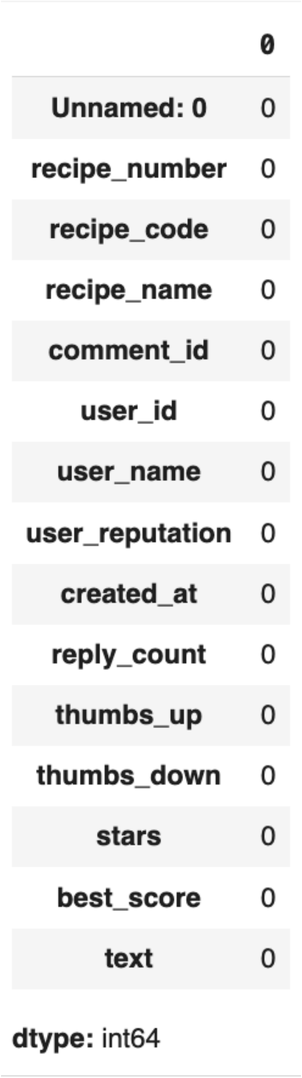
2. File-Level Diagnostics

2.1 Dimensions

The final analysis file contains 18,180 rows and 8 columns.

2.2 Missingness (Post-cleaning)

Because a complete-case filter was applied, the variables retained in the analysis file have zero missing values. Table A1 reports missingness for each variable and should confirm that all counts are equal to zero.



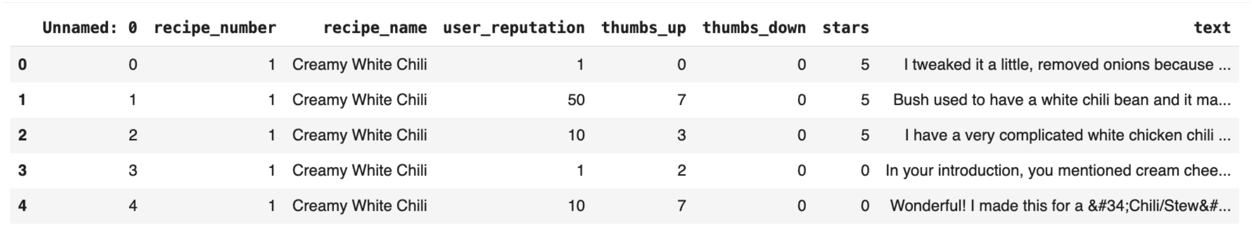
	0
Unnamed: 0	0
recipe_number	0
recipe_code	0
recipe_name	0
comment_id	0
user_id	0
user_name	0
user_reputation	0
created_at	0
reply_count	0
thumbs_up	0
thumbs_down	0
stars	0
best_score	0
text	0

dtype: int64

Table A1

2.3 Variable Inventory

Table A2 lists the final variables kept in the final analysis file.



	Unnamed: 0	recipe_number	recipe_name	user_reputation	thumbs_up	thumbs_down	stars	text
0	0	1	Creamy White Chili	1	0	0	5	I tweaked it a little, removed onions because ...
1	1	1	Creamy White Chili	50	7	0	5	Bush used to have a white chili bean and it ma...
2	2	1	Creamy White Chili	10	3	0	5	I have a very complicated white chicken chili ...
3	3	1	Creamy White Chili	1	2	0	0	In your introduction, you mentioned cream chee...
4	4	1	Creamy White Chili	10	7	0	0	Wonderful! I made this for a &#34;Chili/Stew&#...

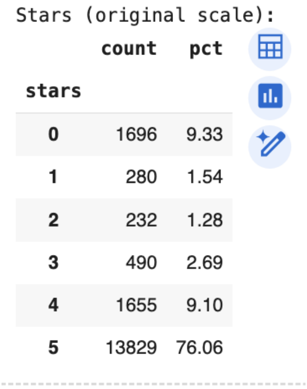
Table A2

3. Variables (Codebook Entries)

Each subsection provides a definition, a brief description of how the variable was processed, a statement on missingness in n(m) format, and descriptive statistics with references to the appropriate table and figure.

3.1 stars (quantitative; 0–5)

The variable stars records the reviewer's rating on an integer scale from 0 to 5. It is carried forward directly from the raw file and is coerced to numeric if necessary; no transformations are applied beyond the complete-case filter. The variable has n(m) = 18,180(0) in the analysis file. Descriptive statistics for stars appear in Table A3, which reports the count and percentage of each star within the dataset. The distribution of stars is shown in Figure A1 using a histogram (count plot), which clearly indicates that 5-star ratings are most common.



Stars (original scale):

stars	count	pct
0	1696	9.33
1	280	1.54
2	232	1.28
3	490	2.69
4	1655	9.10
5	13829	76.06

Table A3

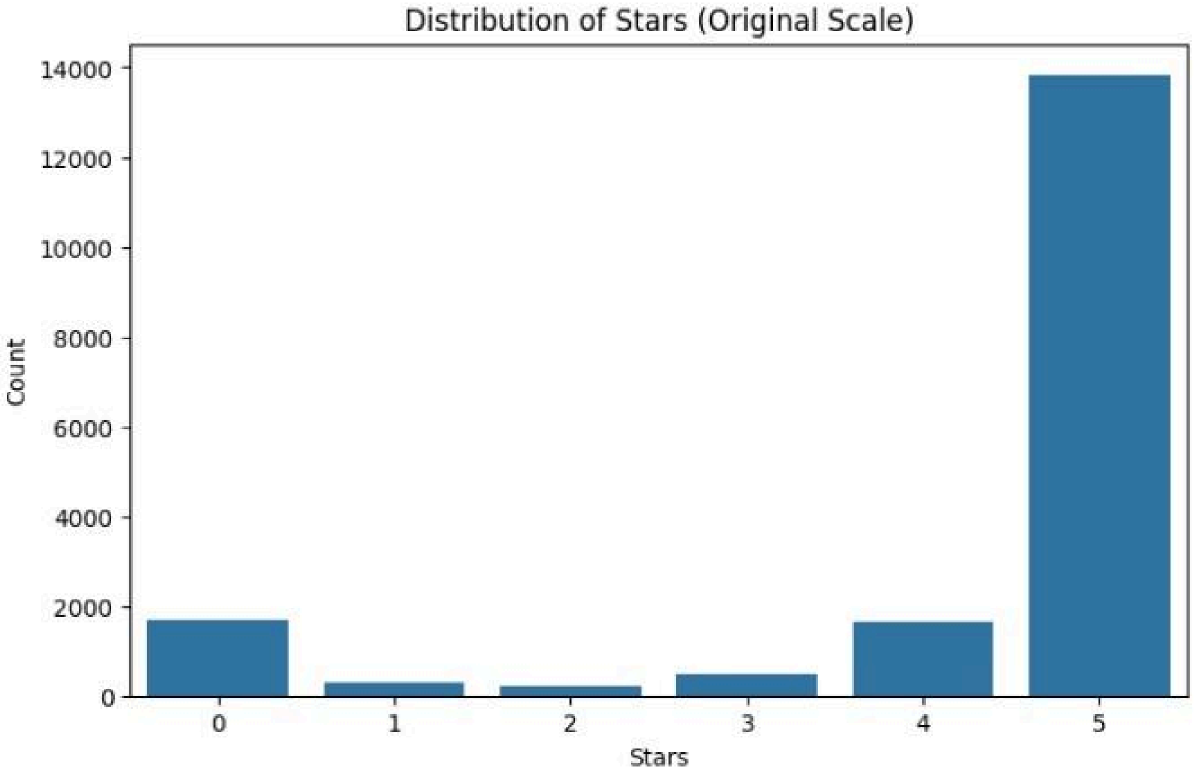


Figure A1

3.2 star_bin (categorical; derived)

The variable star_bin groups the original stars rating into three categories to simplify interpretation: neg for ratings from 0 to 2, neu for a rating of 3, and pos for ratings from 4 to 5. It is deterministically derived from stars using fixed cut points and therefore inherits the complete-case status of the parent variable. The variable has n(m) = 18,180(0) in the analysis file. The counts of each category are reported in Table A4. Figure A2 displays the corresponding bar chart and visually confirms the strong right-skew toward positive ratings.

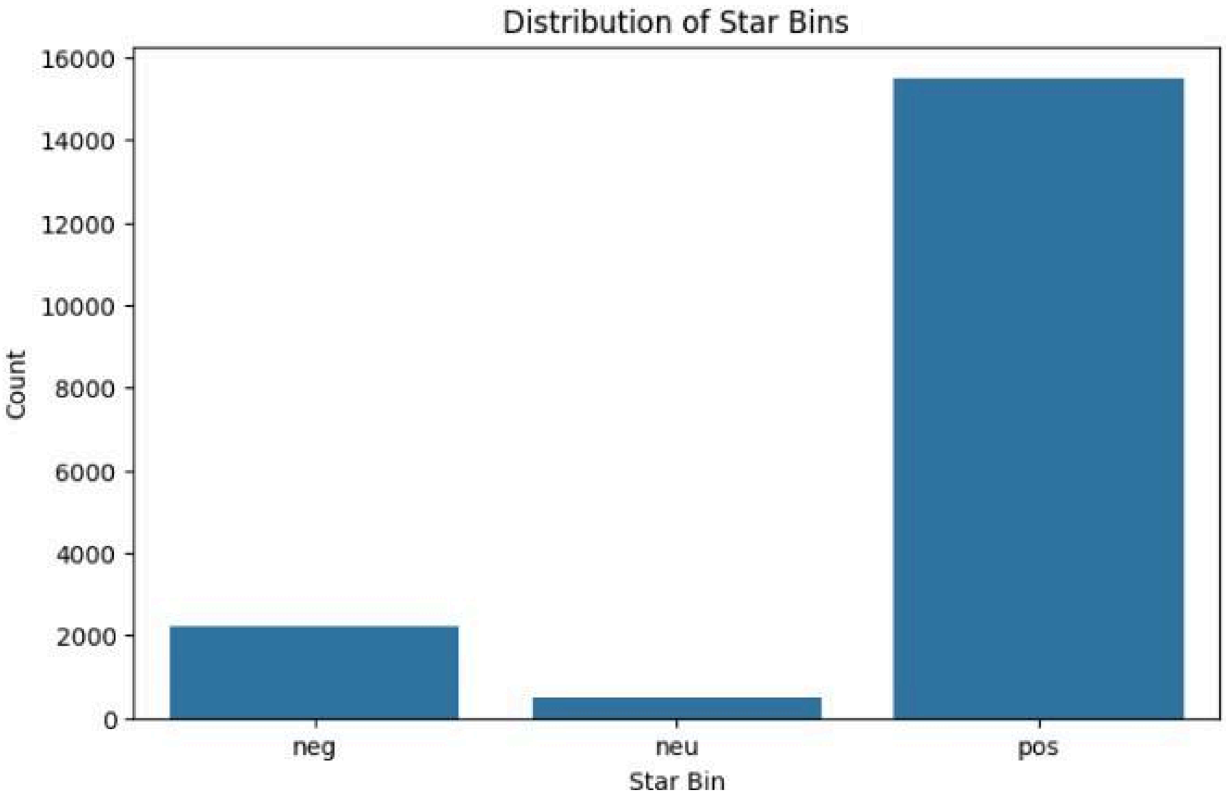
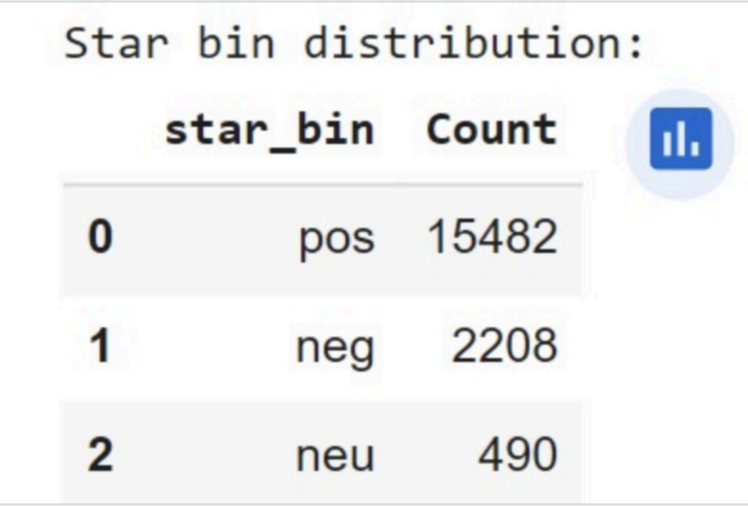


Figure A2



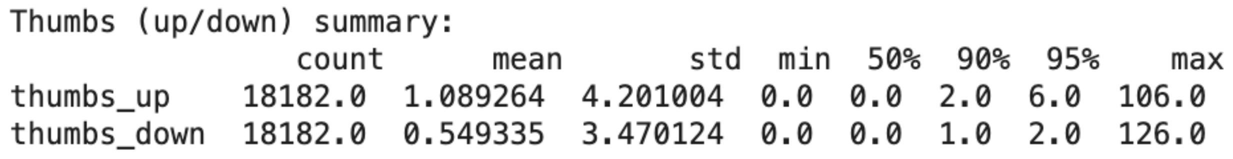
Star bin distribution:

	star_bin	Count
0	pos	15482
1	neg	2208
2	neu	490

Table A4

3.3 thumbs_up (quantitative; nonnegative integer)

The variable thumbs_up records the number of helpful votes received by each review. It is carried forward directly from the raw file and is coerced to numeric where needed to ensure valid aggregation and plotting. The variable has n(m) = 18,180(0) in the analysis file. Table A5 reports the descriptive statistics for thumbs_up (including the mean, standard deviation, median, selected upper percentiles such as p90 and p95, and the maximum). Figure A3 shows the distribution of thumbs_up with a histogram, which exhibits a heavy right tail with many observations near zero and a small number of reviews with high engagement.



```
Thumbs (up/down) summary:
               count      mean       std  min  50%  90%  95%    max
thumbs_up    18182.0  1.089264  4.201004  0.0  0.0  2.0  6.0  106.0
thumbs_down  18182.0  0.549335  3.470124  0.0  0.0  1.0  2.0  126.0
```

Table A5

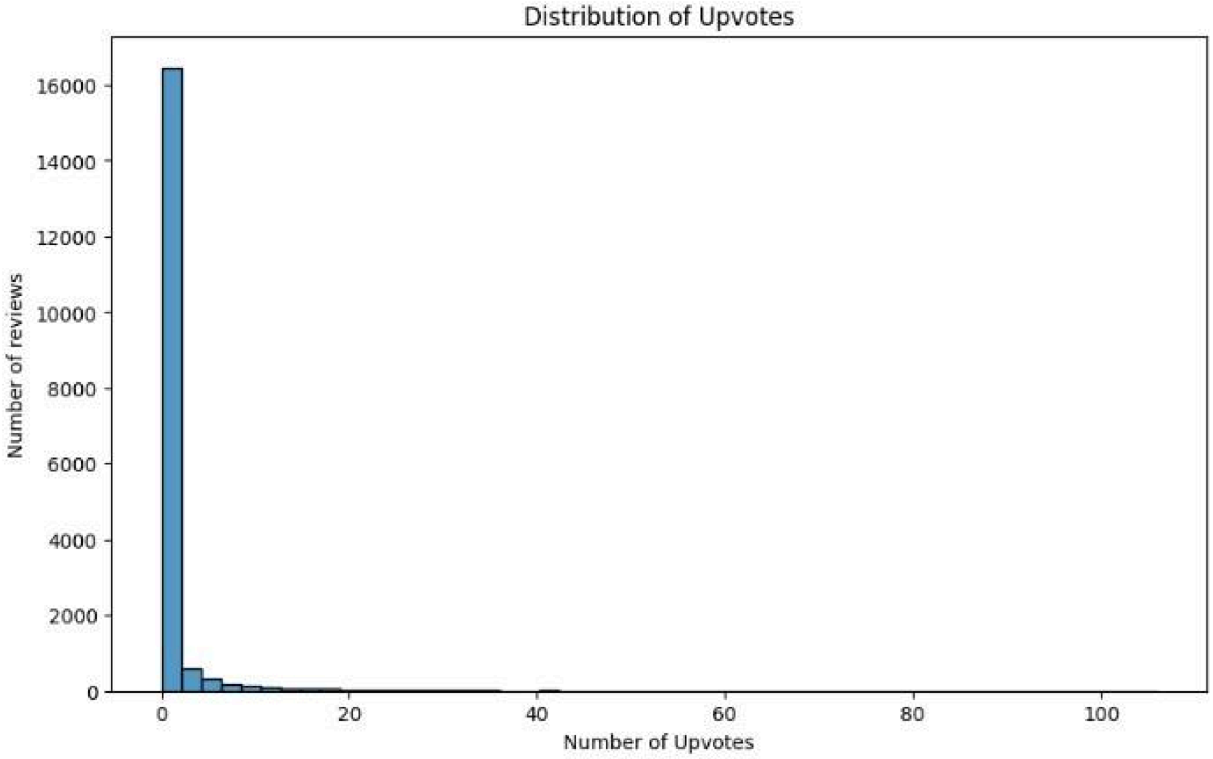


Figure A3

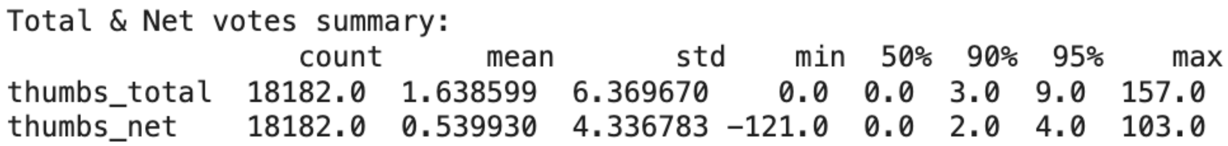
3.4 thumbs_down (quantitative; nonnegative integer)

The variable thumbs_down records the number of unhelpful votes received by each review. It is carried forward from the raw file and is coerced to numeric if necessary, with no transformation beyond the complete-case filter applied earlier in cleaning. The variable has n(m) = 18,180(0) in the analysis file. Table A5 reports the descriptive statistics for thumbs_down (including the mean, standard deviation, median, upper percentiles, and maximum). Refer to Table A5 above for the summary. As with thumbs_up, this distribution is expected to be sparse with occasional high values.

3.5 (Derived measures)

3.5.1 thumbs_total and thumbs_net (derived, quantitative)

To summarize engagement magnitude and direction, we computed thumbs_total as the sum of thumbs_up and thumbs_down and thumbs_net as their difference. Table A6 reports the descriptive statistics for these measures.

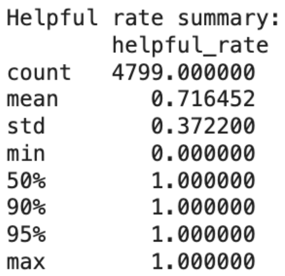


```
Total & Net votes summary:
                 count      mean       std    min  50%  90%  95%    max
thumbs_total  18182.0  1.638599  6.369670    0.0  0.0  3.0  9.0  157.0
thumbs_net    18182.0  0.539930  4.336783 -121.0  0.0  2.0  4.0  103.0
```

Table A6

3.5.2 helpful_rate (derived, proportion)

We defined helpful_rate as thumbs_up / (thumbs_up + thumbs_down) and left it undefined when no votes were present. Table A7 reports the distribution of this proportion for the 4,799 reviews with at least one vote; the mean helpfulness share is 0.716, the standard deviation is 0.372, and the median is 1.000.

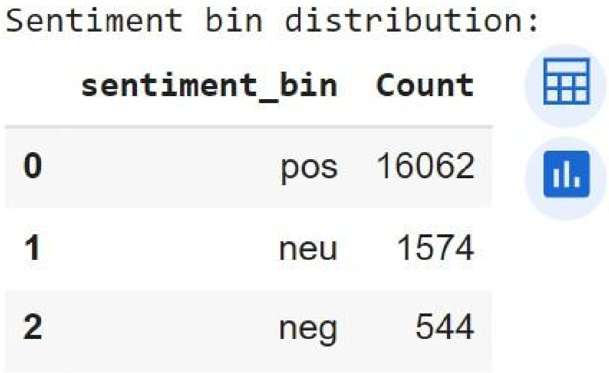


```
Helpful rate summary:
       helpful_rate
count   4799.000000
mean       0.716452
std        0.372200
min        0.000000
50%        1.000000
90%        1.000000
95%        1.000000
max        1.000000
```

Table A7

3.5.3 sentiment_bin (derived, categorical)

We derived sentiment_bin from the review text using our sentiment pipeline, classifying reviews as pos, neu, or neg. The distribution is pos = 16,062 (88.35%), neu = 1,574 (8.66%), and neg = 544 (2.99%), as shown in Table A8.



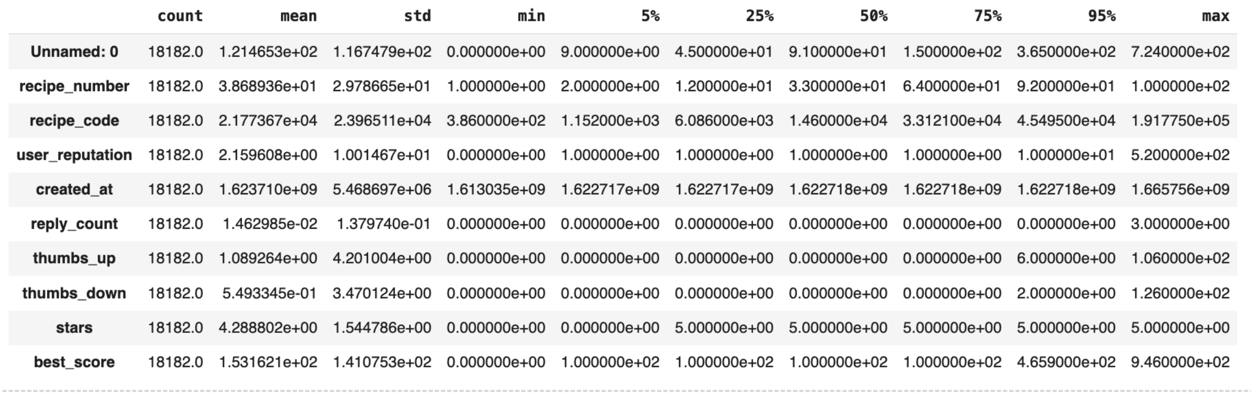
Sentiment bin distribution:

	sentiment_bin	Count
0	pos	16062
1	neu	1574
2	neg	544

Table A8

3.5.4 Additional numeric descriptive statistics (raw file, pre-cleaning)

For transparency about the source data, we include a summary table of numeric variables from the raw file before cleaning. This table includes several fields that were later dropped (for example, recipe_code, created_at, and best_score), so it should not be interpreted as part of the analysis dataset. Table A9 presents these pre-cleaning numeric summaries.

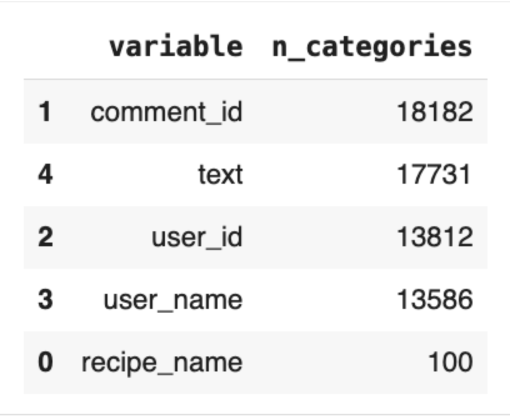


	count	mean	std	min	5%	25%	50%	75%	95%	max
Unnamed: 0	18182.0	1.214653e+02	1.167479e+02	0.000000e+00	9.000000e+00	4.500000e+01	9.100000e+01	1.500000e+02	3.650000e+02	7.240000e+02
recipe_number	18182.0	3.868936e+01	2.978665e+01	1.000000e+00	2.000000e+00	1.200000e+01	3.300000e+01	6.400000e+01	9.200000e+01	1.000000e+02
recipe_code	18182.0	2.177367e+04	2.396511e+04	3.860000e+02	1.152000e+03	6.086000e+03	1.460000e+04	3.312100e+04	4.549500e+04	1.917750e+05
user_reputation	18182.0	2.159608e+00	1.001467e+01	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+01	5.200000e+02
created_at	18182.0	1.623710e+09	5.468697e+06	1.613035e+09	1.622717e+09	1.622717e+09	1.622718e+09	1.622718e+09	1.622718e+09	1.665756e+09
reply_count	18182.0	1.462985e-02	1.379740e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	3.000000e+00
thumbs_up	18182.0	1.089264e+00	4.201004e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	6.000000e+00	1.060000e+02
thumbs_down	18182.0	5.493345e-01	3.470124e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2.000000e+00	1.260000e+02
stars	18182.0	4.288802e+00	1.544786e+00	0.000000e+00	0.000000e+00	5.000000e+00	5.000000e+00	5.000000e+00	5.000000e+00	5.000000e+00
best_score	18182.0	1.531621e+02	1.410753e+02	0.000000e+00	1.000000e+02	1.000000e+02	1.000000e+02	1.000000e+02	4.659000e+02	9.460000e+02

Table A9

3.5.5 Categorical cardinalities (raw file, pre-cleaning)

We also report the number of distinct values for selected categorical variables in the raw file before cleaning. This diagnostic highlights the breadth of identifiers and text fields that were not carried into the analysis data. Table A10 presents the distinct counts for comment_id, text, user_id, user_name, and recipe_name.



	variable	n_categories
1	comment_id	18182
4	text	17731
2	user_id	13812
3	user_name	13586
0	recipe_name	100

Table A10

4. Model Outputs

4.1 Evaluation setup and unit of observation

The evaluation uses a held-out test set of 3,636 reviews drawn from the analysis file. The unit of observation is still a single review of a particular recipe at a point in time. Class supports in the test set are neg = 442, neu = 98, and pos = 3,096. The figures referenced below are exported to the repository folder OUTPUT/model visualizations/.

4.2 Helpfulness vs. alignment between text and stars

To assess whether the crowd finds "aligned" comments more helpful, we define an alignment indicator that equals one when the text-sentiment label (sentiment_bin) matches the rating bin (star_bin) and zero otherwise. Across the full analysis file, 3,533 reviews are misaligned and 14,647 are aligned. Mean helpful votes are higher for misaligned comments (1.3515) than for aligned comments (1.0261), while both groups have a median of zero and heavy right tails (maximums of 80 and 106, respectively). The figures for these are shown below; where Figure A4 shows the mean number of votes by alignment, Figure A5 shows the distribution of helpful votes by alignment, and Table A11 shows a summary of helpful votes by alignment.

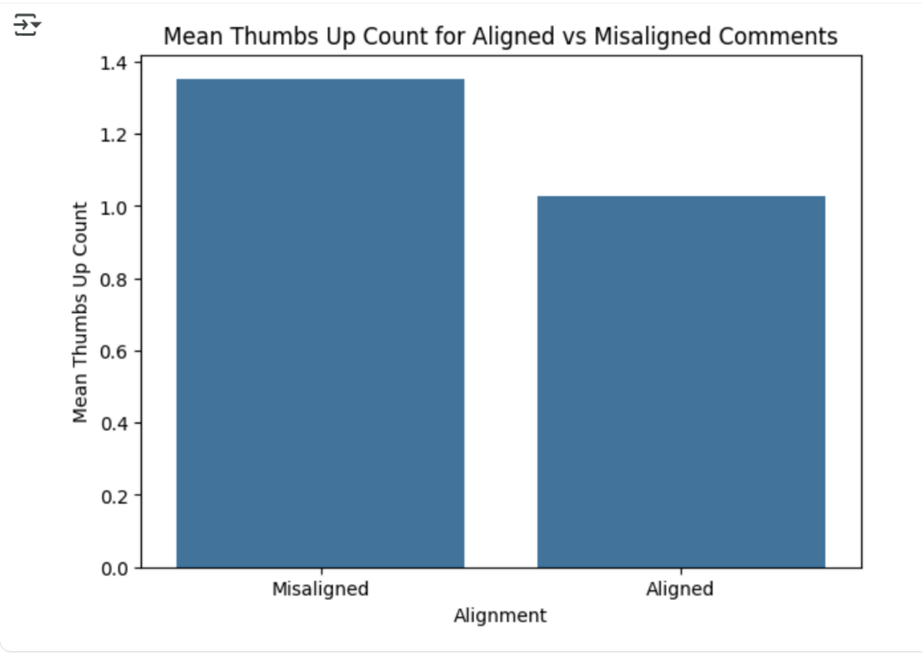


Figure A4

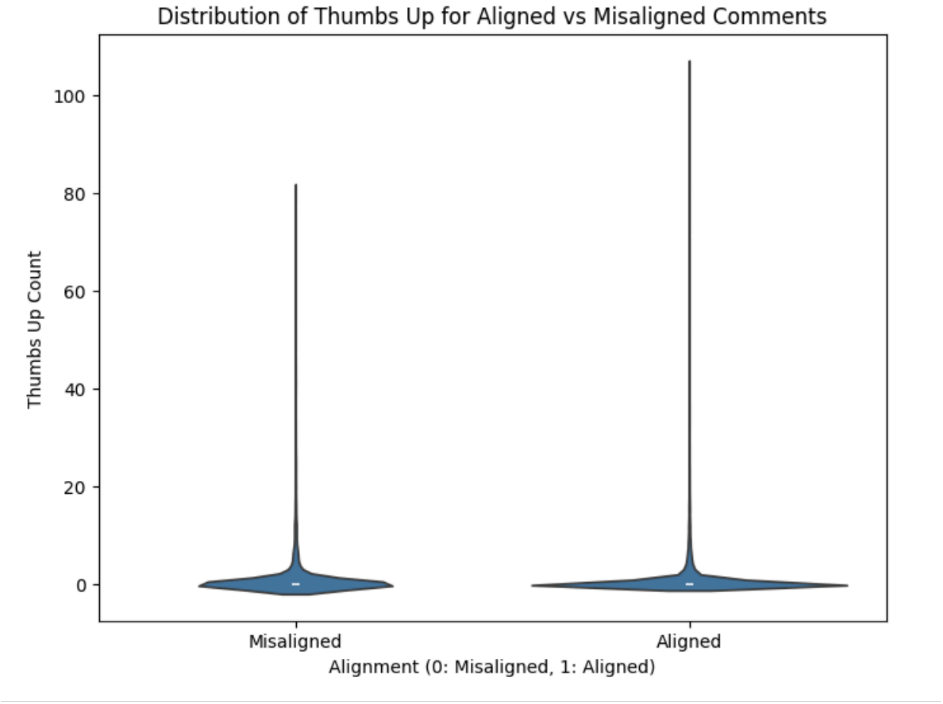
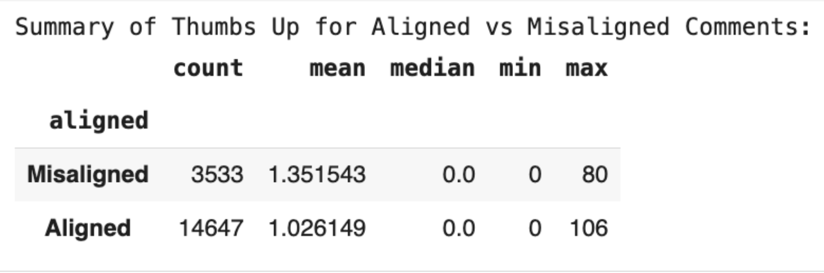


Figure A5



Summary of Thumbs Up for Aligned vs Misaligned Comments:

aligned	count	mean	median	min	max
Misaligned	3533	1.351543	0.0	0	80
Aligned	14647	1.026149	0.0	0	106

Table A11

4.3 Baseline sentiment classifier (unweighted)

We first evaluate a multi-class classifier (neg/neu/pos) trained without class weighting. From there, we created a confusion matrix, classification report, and ROC curves. Figure A6 shows the baseline confusion matrix which identifies positive reviews well and generally is able to map neutral and negative reviews. Table A12 is a baseline classification report with information such as precision, recall, f1-score, and support for all value classes. Figure A7 is a ROC curves visualization which looks at One-vs-Rest. These first model visualizations are shown below.

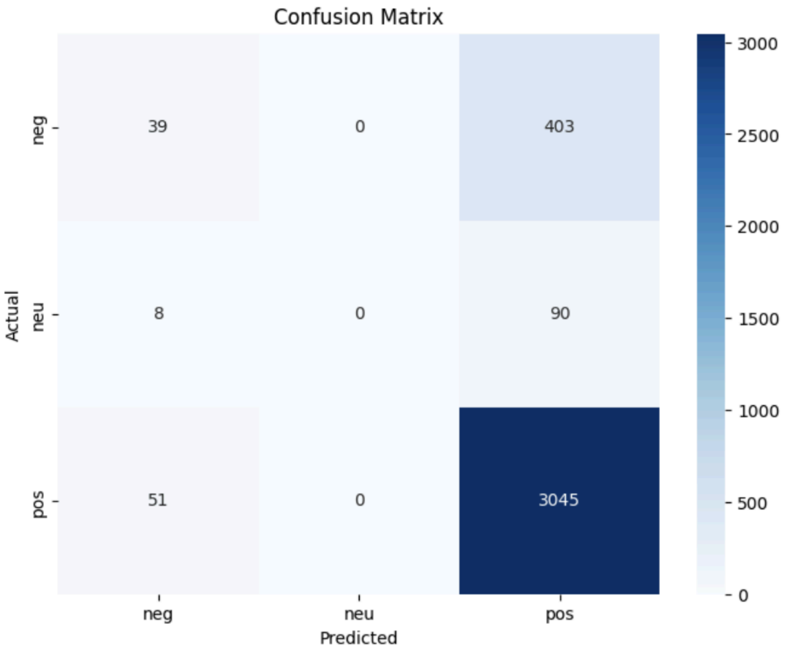
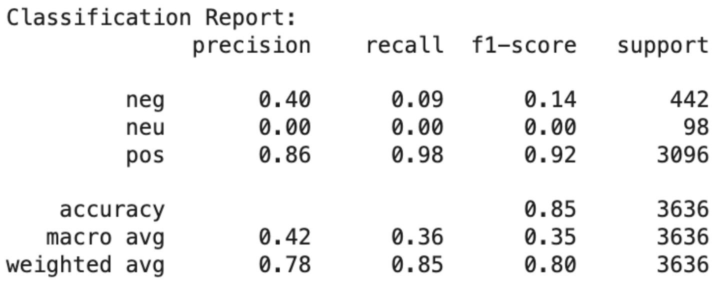


Figure A6



```
Classification Report:
              precision    recall  f1-score   support

         neg       0.40      0.09      0.14       442
         neu       0.00      0.00      0.00        98
         pos       0.86      0.98      0.92      3096

    accuracy                           0.85      3636
   macro avg       0.42      0.36      0.35      3636
weighted avg       0.78      0.85      0.80      3636
```

Table A12

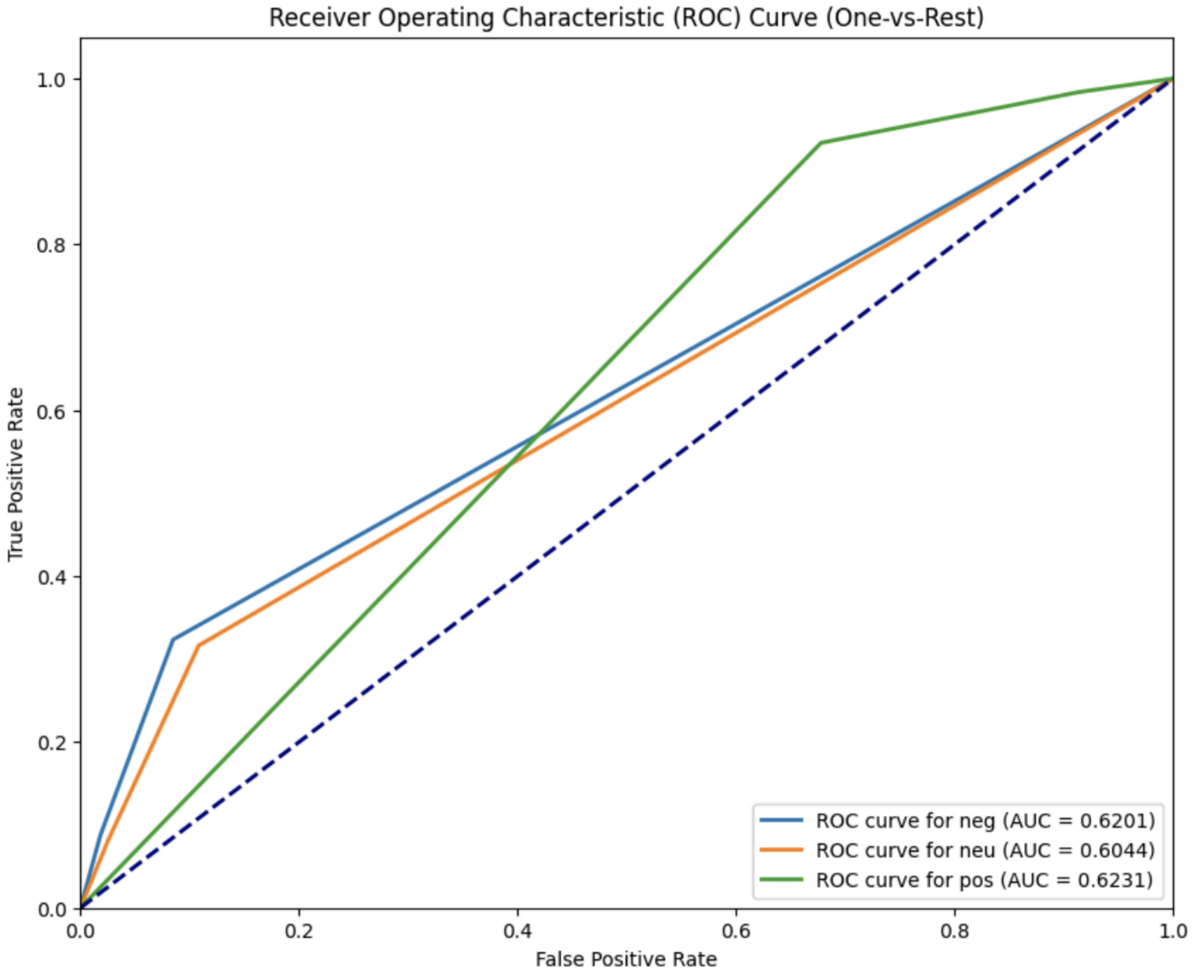


Figure A7

4.4 Class-weighted sentiment classifier

From the first model, we decided to re-train the classifier with class weights to mitigate imbalance and evaluate on the same test set. From there, we recreated the same figures and tables, now accounting for the weighted classification to make up for the prior model accustoming only to the positive class. Figure A7 is a confusion matrix for the weighted model, Table A13 is a classification report for the weighted model, and lastly Figure A9 is a ROC curve (One vs. Rest) for the weighted model. Those figures are shown below.

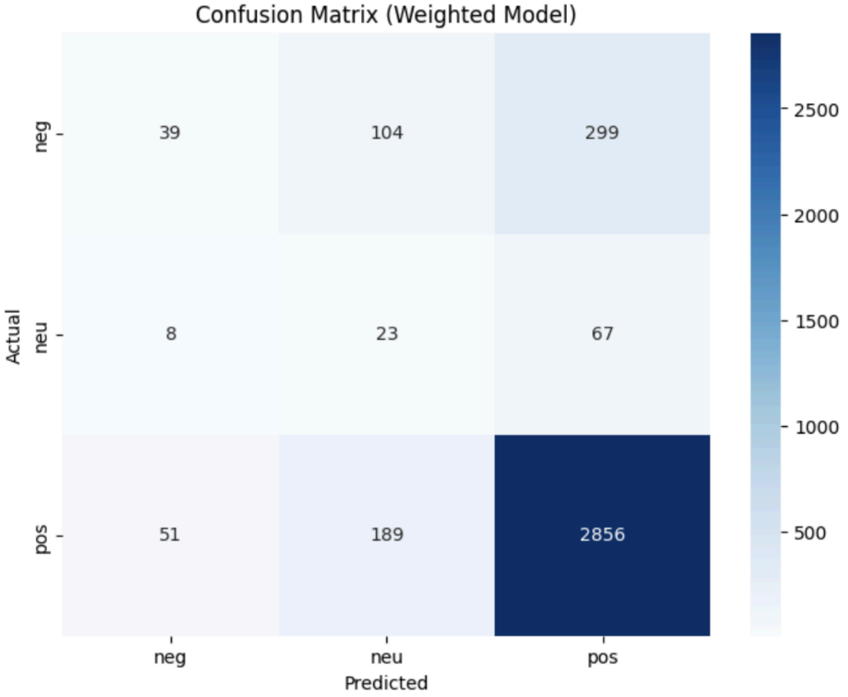
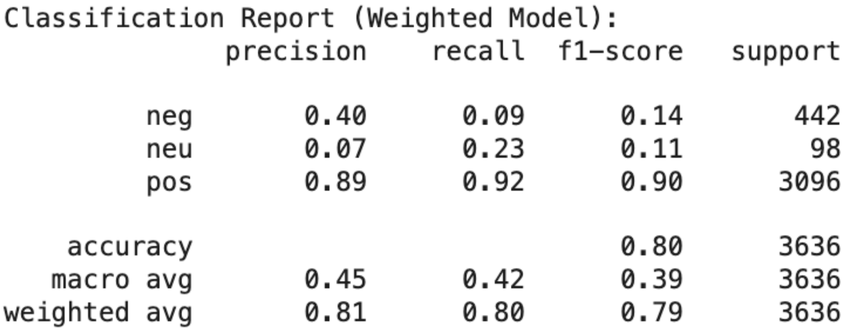


Figure A8



```
Classification Report (Weighted Model):
              precision    recall  f1-score   support

         neg       0.40      0.09      0.14       442
         neu       0.07      0.23      0.11        98
         pos       0.89      0.92      0.90      3096

    accuracy                           0.80      3636
   macro avg       0.45      0.42      0.39      3636
weighted avg       0.81      0.80      0.79      3636
```

Table A13

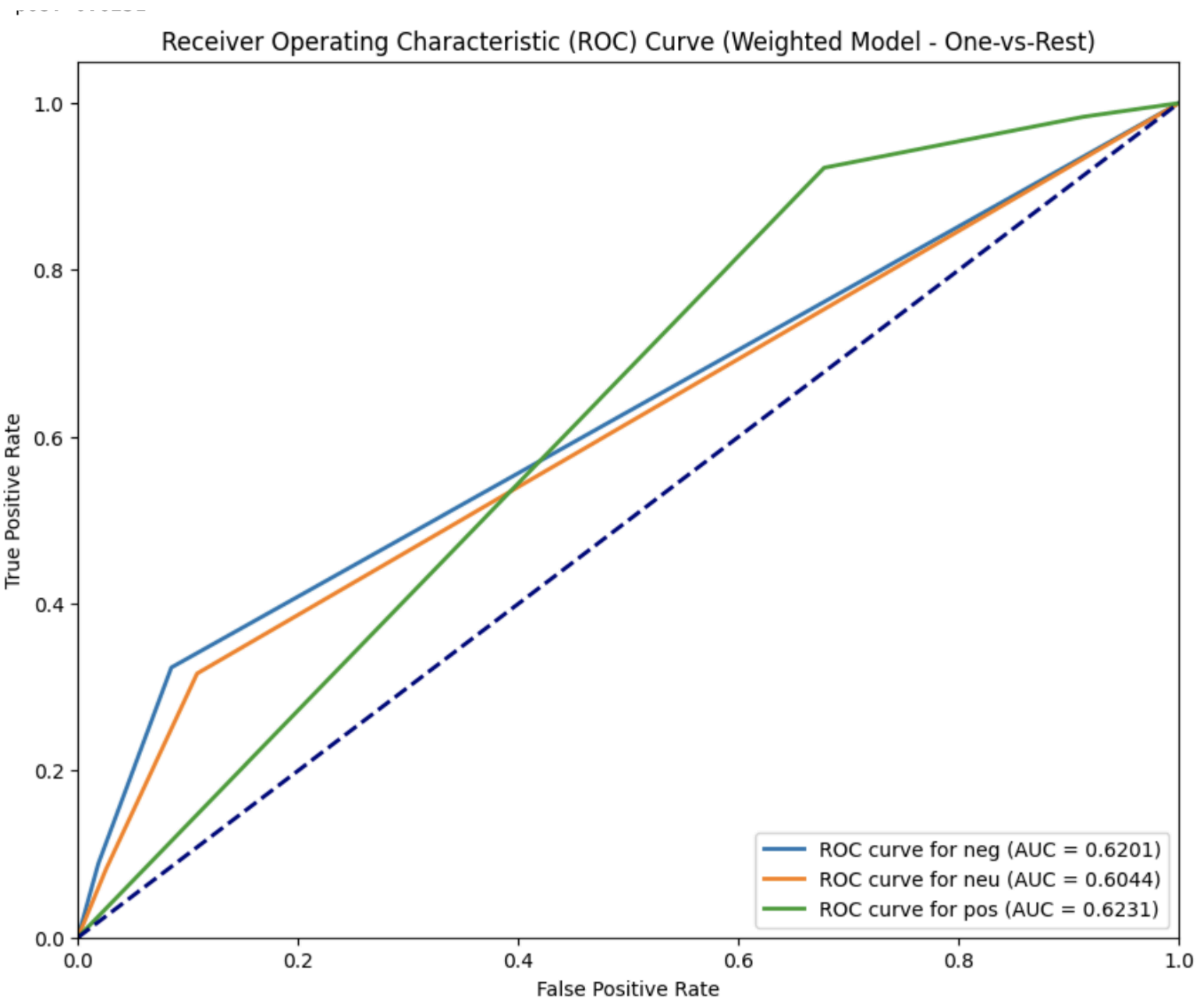


Figure A9

4.5 Predicting sentiment alignment from upvotes (binary model)

This new model is a simple binary classifier that uses only the thumbs_up count to predict whether a review's text sentiment is aligned with its star rating (1 = aligned, 0 = misaligned). The intent is to test the minimal hypothesis that higher upvotes alone signal alignment. From there, we created Figure A10, a confusion matrix which predicts all the reviews as aligned vs. misaligned, Table A14 is a classification report on the alignment classes, and Figure A11 is a ROC curve visualization for upvotes and alignment. These figures are shown below.

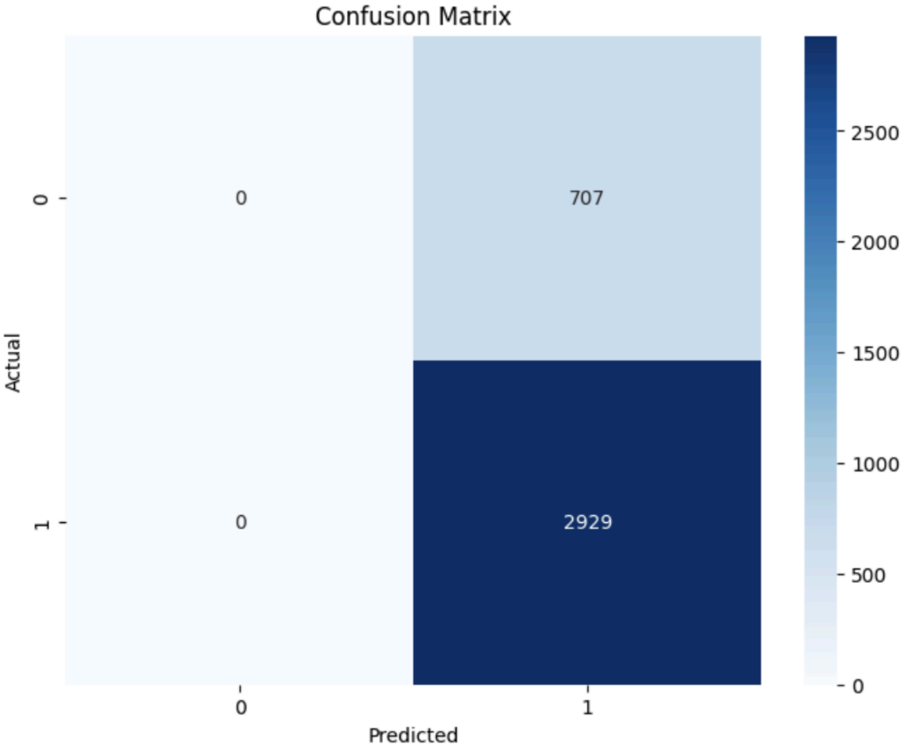
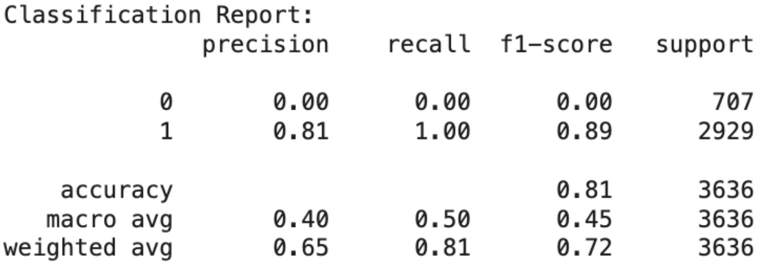


Figure A10



```
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       707
           1       0.81      1.00      0.89      2929

    accuracy                           0.81      3636
   macro avg       0.40      0.50      0.45      3636
weighted avg       0.65      0.81      0.72      3636
```

Table A14

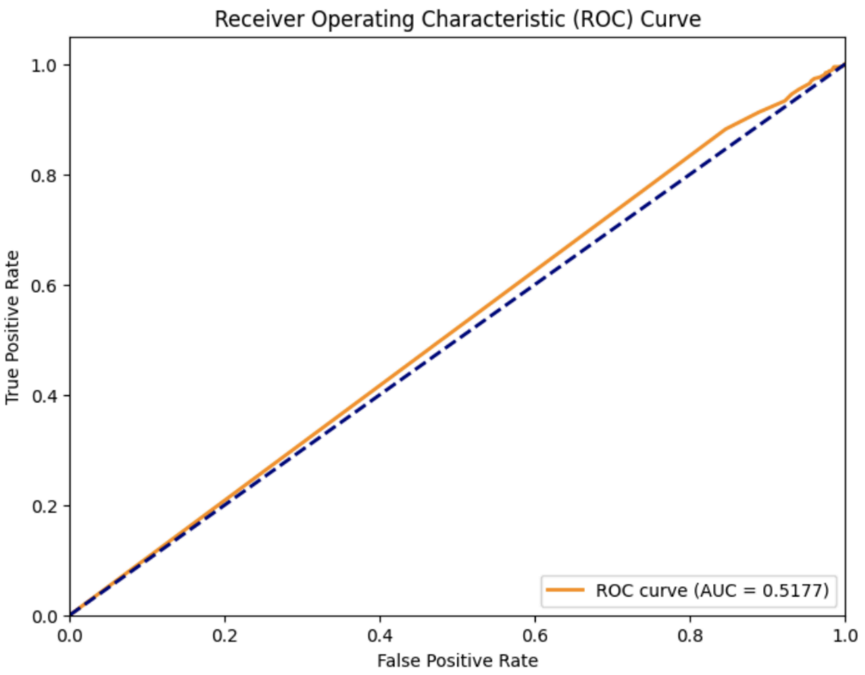


Figure A11

5. Deviations & Assumptions

This project uses a complete-case analysis, which simplifies documentation and guarantees zero missingness among kept variables. This choice can introduce bias if the data are not missing completely at random. The star_bin categorization uses fixed thresholds selected for interpretability; alternative thresholds would change category shares but would not change the underlying stars distribution.