TripAdvisor Hotel Review Analysis

Business Intelligence
Spring 2024

Table of contents

Data Processing

Using Weka, R, and Python, we process the data such that it is usable for the analysis.

Attribute	Task	Details
Author_location	**Extract country name**	Used "**Web scrapping**" to categorize locations into countries.
Rating location	**Fill missing values**	Considering that all missing ratings are considered neutral, the missing values were filled with a rating of 3.
Author_num_helpful_votes	**Fill missing values**	All missing values are filled with 0.
Revisit	**Convert to a binary variable**	Convert 'Trigger_revisit' to 1 and 'No_revisit' to 0.
Author_num_cities	**Convert to a binary variable**	Created a new column "Visit > 15" in the dataset to classify authors visiting more than 15 cities (value True) or not (value False)
'title' and 'text'	**Text mining**	Concatenated into a single 'review_concat' column Remove numbers and non-word characters, non-English text from the column. ⇨Extract sentimental adjectives to express "good" or "bad" review. Example: "*Splendor and Elegance The beauty of the lobby is only exceeded by the fabulous attention to excellent service. It is rich in history and yet modern-day amenities. Superb on every level. And pet friendly.*" ⇨"*fabulous excellent rich modern friendly*"

01

Scenario 1

Can we model the hotel's overall ratings using reviews?

Scenario 1 – Business Motivation

Customer Insights

Understand customer sentiments and preferences from reviews to making informed marketing strategies

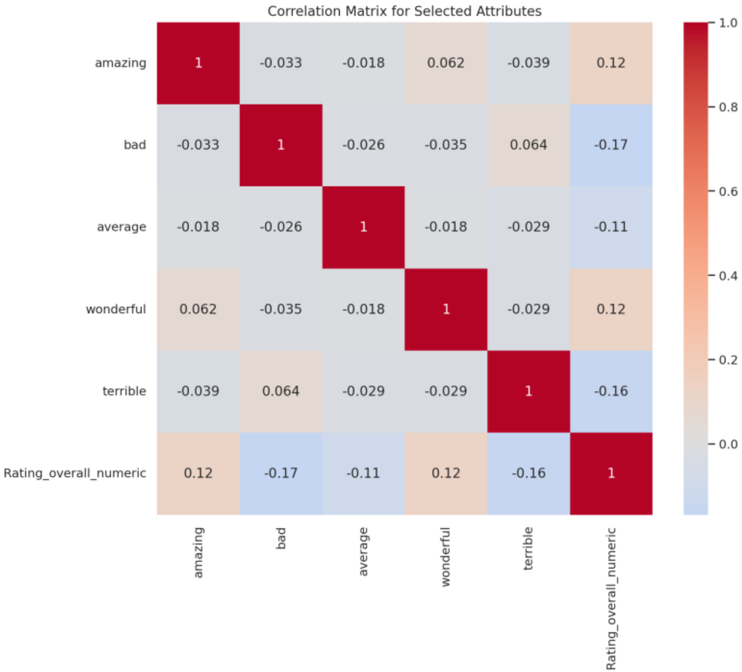
Reputation Management

Promptly recognize and tackle adverse feedback to uphold customer confidence and safeguard your reputation

Competitivity Study

Evaluate customer satisfaction levels through competitive benchmarking to uncover both strengths and weaknesses relative to competitors.

Scenario 1 – Explanatory Data Analysis



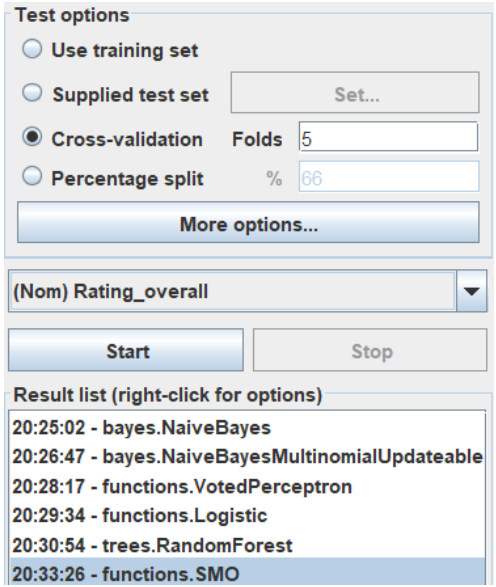
The **correlation matrix** for some selected attributes.

The **word cloud distribution** after processing text columns

Scenario 1 – Model Validation and Selection

Model	Precision	Recall	F1-score	Accuracy	ROC Area
Naïve Bayes (Benchmark)	0,627	0,686	0,655	76.555%	0,793
Naïve Bayes Multinominal Updatable	0,874	0,481	0,620	80.909%	0,871
Voted perceptron	0,749	0,597	0,664	80.430%	0,754
Logistic regression with Ridge Regulation (alpha = 1.0E-6)	0,735	0,633	0,680	80.669%	0,782
Random Forest (default config)	0,769	0,569	0,654	80.478%	0,758
SMO (Support Vector Machine – Logistic, PolyKernel)	0,763	0,606	0,675	81.101%	0,758

Best model is selected based on the combination of metrics including precision, recall, F1-score, and accuracy.



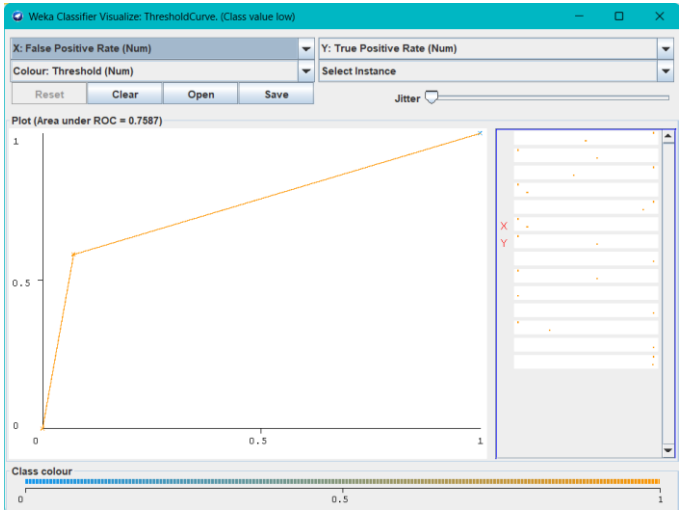
The models are trained using the training set (2090 data points/80% of the total data set) and validated by cross-validation with the number of folds equal to 5.
Based on the metrics score, SMO is the best model and is chosen to be tested with the testing set.

Scenario 1 – Model Performance Testing

Model	Precision	Recall	F1-score	Accuracy	ROC Area
SMO (Support Vector Machine – Logistic, PolyKernel)	0,800	0,588	0,678	81.835%	0,759

The main goal is to point out whether an overall rating is a bad review (i.e., classified as "low"). Therefore, the metrics for the "low" class were considered.

```
=== Summary ===

Correctly Classified Instances         428               81.8356 %
Incorrectly Classified Instances        95               18.1644 %
Kappa statistic                          0.5555
Mean absolute error                      0.1816
Root mean squared error                  0.4262
Total Number of Instances              523

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0,588     0,071     0,800      0,588     0,678     0,568     0,759     0,604     low
                 0,929     0,412     0,824      0,929     0,874     0,568     0,759     0,814     high
Weighted Avg.    0,818     0,301     0,816      0,818     0,810     0,568     0,759     0,746

=== Confusion Matrix ===

   a   b   <-- classified as
 100  70 |   a = low
  25 328 |   b = high
```



The models are tested with the testing dataset (523 data points/20% of the total data set) and achieved the outputs similar to the validation outputs, indicating that the results are reliable.

Conclusion

It is recommended for the company to use the SMO model to predict the sentiment of a review.

Just by taking into account the review title and its text, we can accurately predict whether the attitude of the reviewer is positive or negative.

The final model achieves a predictive accuracy of 81.8%, which is fairly similar to the accuracy achieved with cross-validation on the training set, indicating a stable performance.

Recommendation

By utilizing this forecast, company leadership can identify problem areas in their customer service and promptly address any issues that might impact the company's reputation.

Employing the model to analyze the company's perceived strengths and weaknesses allows for benchmarking against competitors, leading to the discovery of unique selling points and enhanced competitive advantages.

Scenario 3

Can we model the influence of the review author?

Scenario 3 – Business Motivation

Marketing Strategy

Companies might target authors having visited many cities with promotions or information about hotels in cities they haven't visited yet.

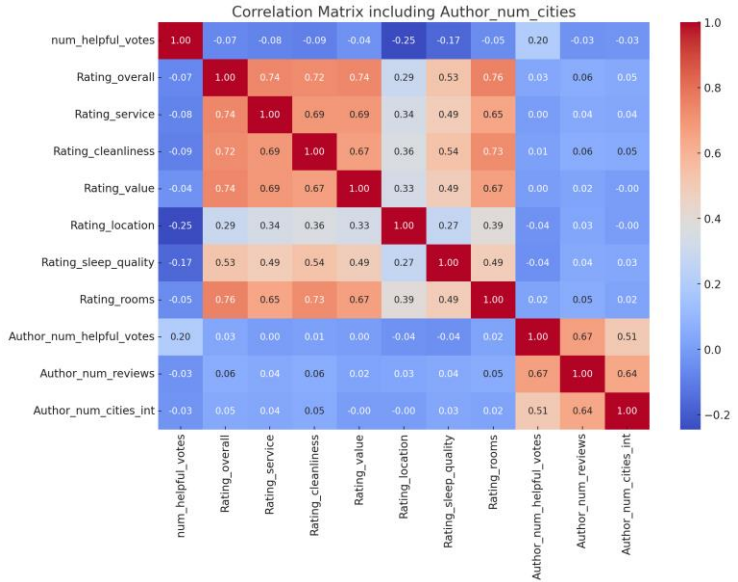
Customer Segmentation

Companies can segment customers into different categories (e.g., frequent travelers, occasional travelers) for more effective service and product offerings.

Influencer Identification

Authors visiting many cities may be considered influencers. Their opinions might have a broader impact, making them valuable partners for promotions or brand ambassadorships.

Scenario 3 – Explanatory Data Analysis

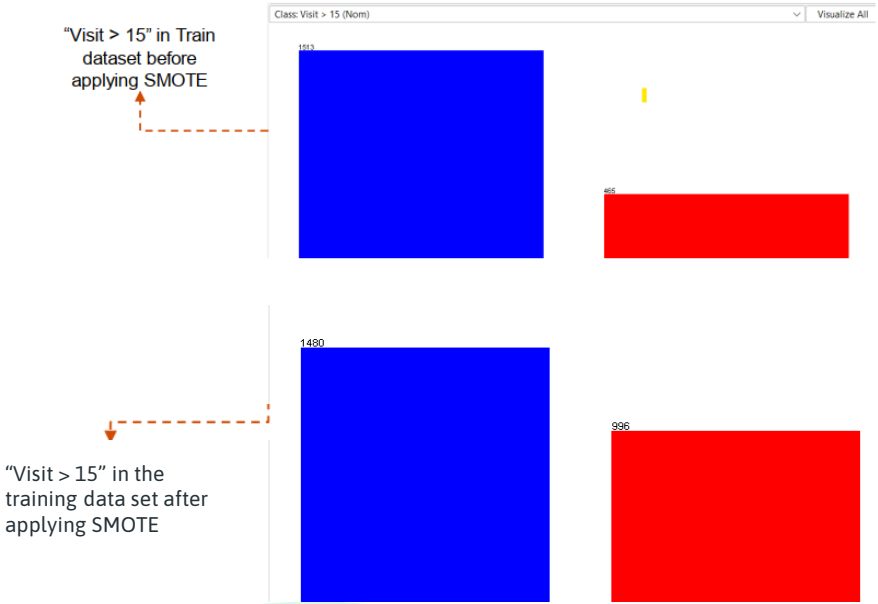


The **correlation matrix** for selecting features.

- The target variable '**Author_num_cities**' exhibits a strong correlation with '**Author_num_helpful_votes**' and '**Author_num_reviews**'. However, these two columns are also strongly correlated.
- The other columns seem to not help predict whether an author has visited more than 15 cities.

Scenario 3 – Feature Engineer and Data Processing

- Constructed a new attribute called 'Visit 15' as an indicator of whether the corresponding author has visited more than 15 countries.
- SMOTE: We had to **rebalance** the training data! The distribution of Author_num_cities (Visit > 15 cities) is imbalanced.



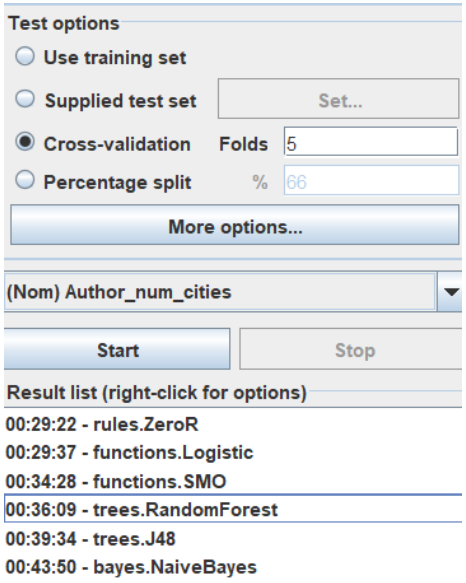
70: Author_num_helpful_votes_per_review Numeric
0.4
0.875
1.6
0.636364
0.625
3.5
0.25
2.294118
0.923077
0.647059
1.75
0.5
0.25
0.272727
1.2
0.133333
0.0
0.0
1.666667
0.368421
1.0
0.5
0.307692
0.454545
1.227273
1.0
0.72973
0.214286

We aggregate two features **'Author_num_helpful_votes** ' and '**Author_num_reviews**' by constructing a new one called '**Author_num_helpful_votes _per_review**' by taking **'Author_num_helpful_votes** ' divided by '**Author_num_reviews**' to resolve the multicollinearity.

Scenario 3 – Model Validation and Selection

Model	Precision	Recall	F1-score	Accuracy	ROC Area
ZeroR (Baseline model)	NA	0	NA	59.8%	0,499
Logistic regression with Ridge Regulation (alpha = 1.0E-6)	0,909	0,885	0,897	91.8%	0,947
SMO (Support Vector Machine – Logistic, PolyKernel)	0,929	0,852	0,889	91.4%	0,851
Random Forest (default config)	0,906	0,933	0,919	93.4%	0,974
DecisionTree (min 10 instances per leaf)	0,877	0,940	0,907	92.3%	0,964
Naïve Bayes	0,908	0,775	0,836	87.8%	0,924

Best model is selected based on the combination of metrics including precision, recall, F1-score, and accuracy.



The models are trained using the training (2476 instances) and validated by cross-validation with the number of folds equal to 5.

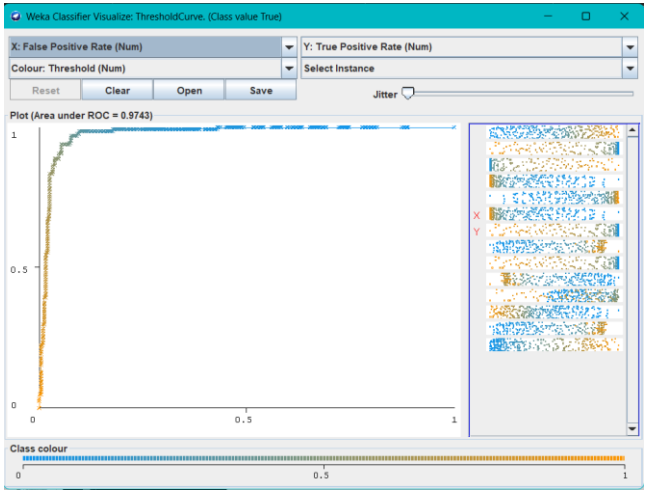
Based on the metrics score, Random Forest is chosen to be tested with the testing set since they have the best performances.

Scenario 3 – Model Performance Testing

Model	Precision	Recall	F1-score	Accuracy	ROC Area
Random Forest (default config)	0,863	0,904	0,883	94%	0,974

The objective is to predict whether an author has visited more than 15 cities (i.e., classified as True).
Therefore, the metrics for the True class were considered.

```
Correctly Classified Instances         466               93.9516 %
Incorrectly Classified Instances        30                6.0484 %
Kappa statistic                          0.8421
Mean absolute error                      0.1644
Root mean squared error                  0.2442
Total Number of Instances              496

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0,951     0,096     0,967       0,951   0,959       0,843    0,974     0,992     False
                 0,904     0,049     0,863       0,904   0,883       0,843    0,974     0,889     True
Weighted Avg.    0,940     0,084     0,941       0,940   0,940       0,843    0,974     0,966

=== Confusion Matrix ===

   a   b   <-- classified as
 353  18 |   a = False
  12 113 |   b = True
```



The models are tested with the testing dataset (496 data points) and achieved outputs similar to the validation outputs, indicating that the results are reliable.

Conclusion

Research indicates that authors garnering a high volume of reviews and positive votes are more likely to visit over 15 countries, with prediction models demonstrating up to 94% accuracy in identifying such behavior.

The correlation seems to stem from authors receiving more reviews as they explore various cities, with their vast travel experiences likely enhancing the quality of their reviews.

Recommendation

The company is advised to implement a Random Forest and data rebalancing for predicting authors who visit more than 15 countries.

Leveraging these predictions, companies can uncover insights into customer travel behaviors, enabling the development of tailored travel products or targeted marketing strategies. These initiatives are particularly useful in catering to the demands and preferences of frequent travelers, aligning with the company's objectives for its key customer segments.