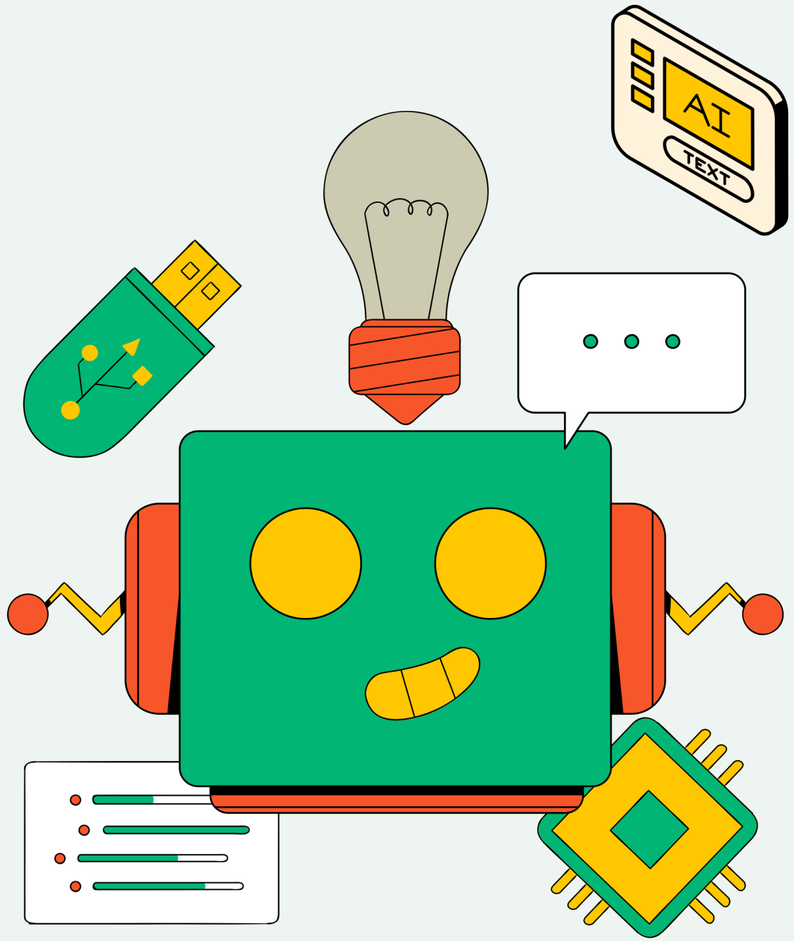
THYNK UNLIMITED

WE LEARN FOR THE FUTURE

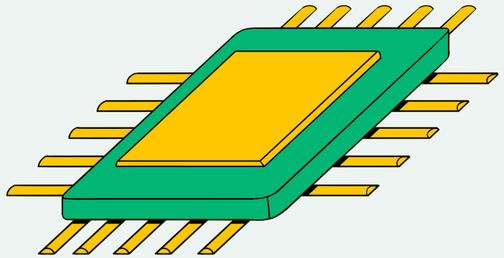


MACHINE LEARNING WITH WEKA

BUSINESS INTELLIGENCE

Cloud

SCENARIO 1 AND 3



REPORT OUTLINE

- Business motivations
- Data preprocessing
- Scenario 1 – Overall review
  - Exploratory data analysis (EDA)
  - Model building
  - Results
    - Model selections
    - Limitation and recommendation
- Scenario 3 – Visit more than 15 cities
  - Exploratory data analysis (EDA)
  - Model building
  - Results
    - Model selections
    - Limitation and recommendation
- Real-world application

BUSINESS MOTIVATION

01

CUSTOMER INSIGHTS

Understand customer sentiments and preferences from reviews to inform marketing strategies and product development.

02

REPUTATION MANAGEMENT

Quickly identify and address negative feedback to maintain customer trust and reputation.

03

COMPETITIVE ANALYSIS

Benchmark customer satisfaction levels to identify competitive strengths and weaknesses.

DATA PREPROCESSING

Used **Web scrapping**, unspecified countries were categorized as "Others", encompassing overseas territories where residency doesn't align with the passport country

New columns were created for all rating attributes, assigning "high" for ratings of 4 and 5, and "low" for ratings of 1, 2, and 3.

Regular expressions: remove numbers and non-word characters from the text. Concatenated into a single 'review_concat' column

Langdetect library: text in languages other than English are filtered out to minimize noise in the model.

Spacy library: text was converted to **lowercase**, **tokenized**, and **Part of Speech tagging** (POS-tagging) was applied to isolate adjectives (ADJ) and nouns (NN). The ultimate objective is to extract emotional adjectives to express a “high” or “low” rating review.

Nltk library: removed **stop-words** and added additional ones to filter out terms that don't contribute to sentiment analysis effectively. Utilize **TfidfVectorize()** to perform **TF-IDF** string-to-word vector transformation

REVIEW CLEANING RESULT

Original text: 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.

Cleaned text: fabulous excellent rich modern friendly

Original text: Fabulous historic Chicago hotel This is a must see and stay. Experience the opulence and historic grandeur of the gift Potter Palmer gave to his bride as a wedding gift, The Palmer House. The lobby ceilings were restored by Lido Lippi. The same Lido Lippi who restored the Sistine Chapel.The winged angels outside the the entrance to Empire Room were designed by non other than Louis Comfort Tiffany.Be sure to visit the Lakewood dining room as have one of Bertha Palmer's famous Chocolate Fudge Brownies. It is worth the trip just for the brownie and people watch in the lobby. If you have the opportunity take the tour History is Hott with Ken Price.This Chicago landmark is one gracious, elegant lady.

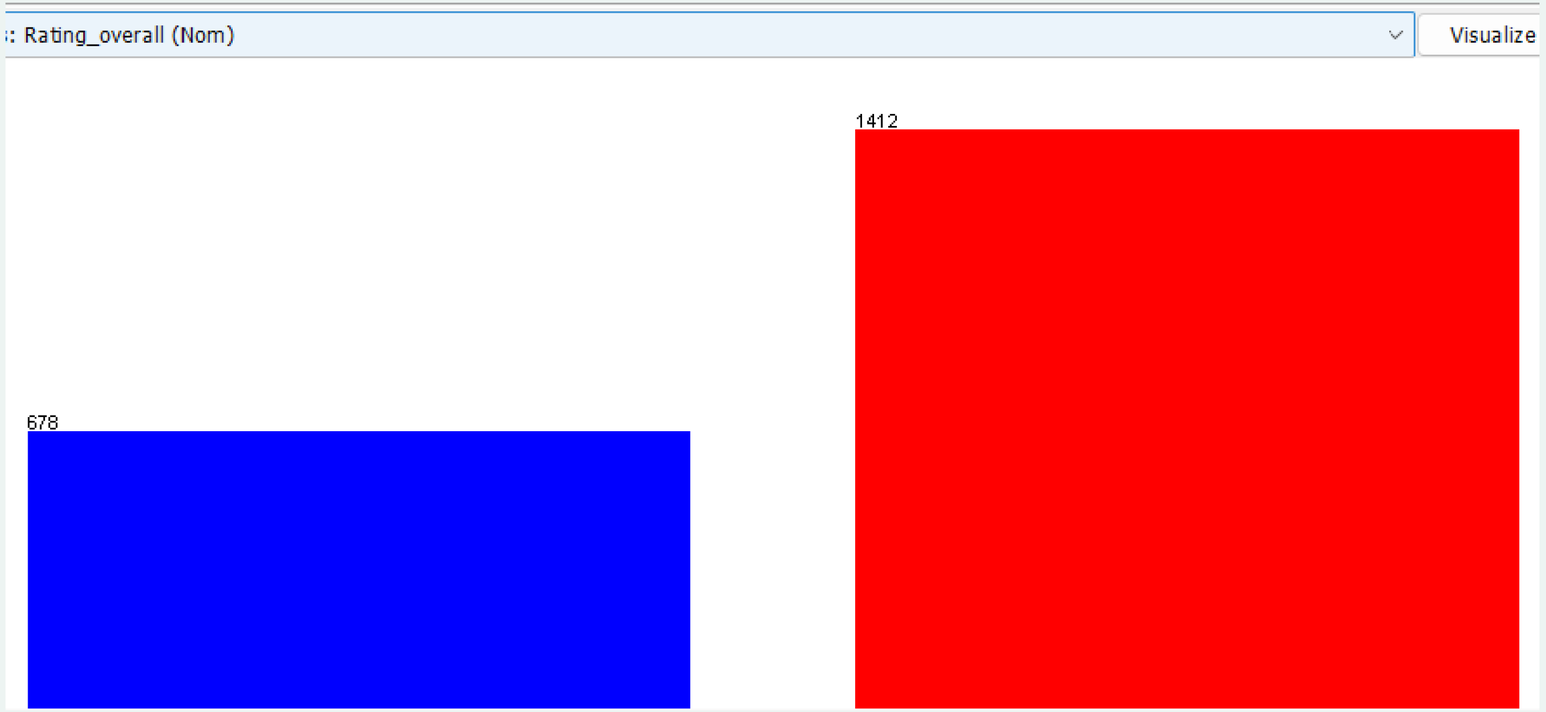
Cleaned text: fabulous famous worth gracious elegant

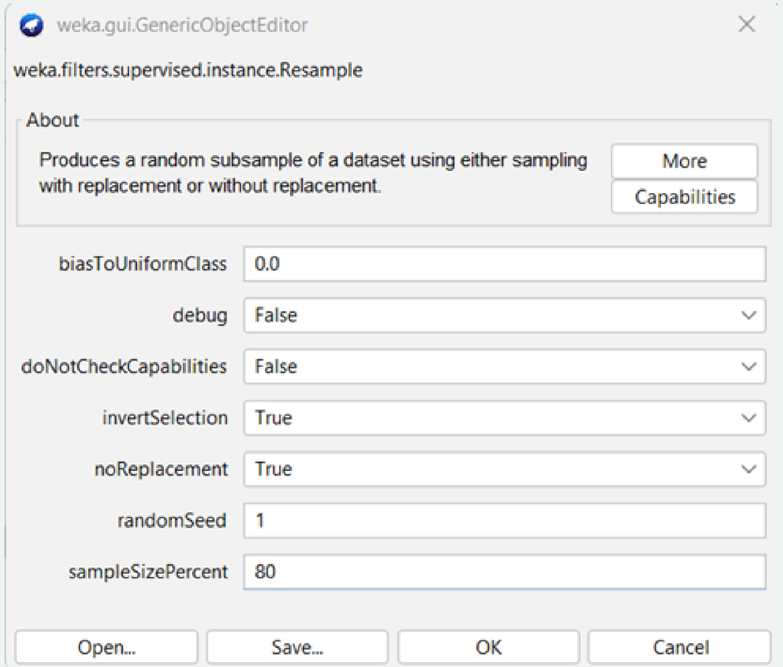
SCENARIO 1 - EDA

After cleaning, the final word cloud distribution is as followed:

SCENARIO 1 - TRAIN TEST SPLIT

- NumericToNominalwith “Rating_overall” and set attribute as Class
- **Resample:** split the dataset into 3 parts with a ratio of 60 – 20 – 20 for the training dataset, validation dataset, and test dataset, respectively.
- Since this is a text classification problem, using traditional sample method such as SMOTE is not recommeded.





SCENARIO 1 - MODEL BUILDING VALIDATION SET

Model	Precision	Recall	f1-score	Accuracy	ROC Area
Naive Bayes (Baseline model)	0.851	0.810	0.830	77.59%	0.880
Naive Bayes Mutinominal Updatable (Chosen for cross validation)	0.794	0.966	0.872	80.84%	0.875
MLPClassifier (SoftPlus, Squared Error)	0.837	0.847	0.828	78.54%	0.828
Voted perceptron (Chosen for cross validation)	0.854	0.895	0.874	82.57%	0.791
Mutilayer Perceptron (MLP) (Chosen for cross validation)	0.843	0.898	0.869	81.80%	0.872
Logistic regression with Ridge estimator (Chosen for corss validation)	0.842	0.892	0.866	81.42%	0.861
Random Forest (max depth = 20) (Chosen for cross validation)	0.790	0.972	0.871	80.65%	0.871
Sequential Minimal Optimization (SMO) (Logistic, RBF kennel) (Chosen for cross validation)	0.839	0.932	0.883	83.33%	0.879

Training different basic models using different classifying methods with Naive Bayes as the Baseline model. The best models then get Cross-validated to check the it's performance against variation.

SCENARIO 1 - MODEL BUILDING VALIDATION SET

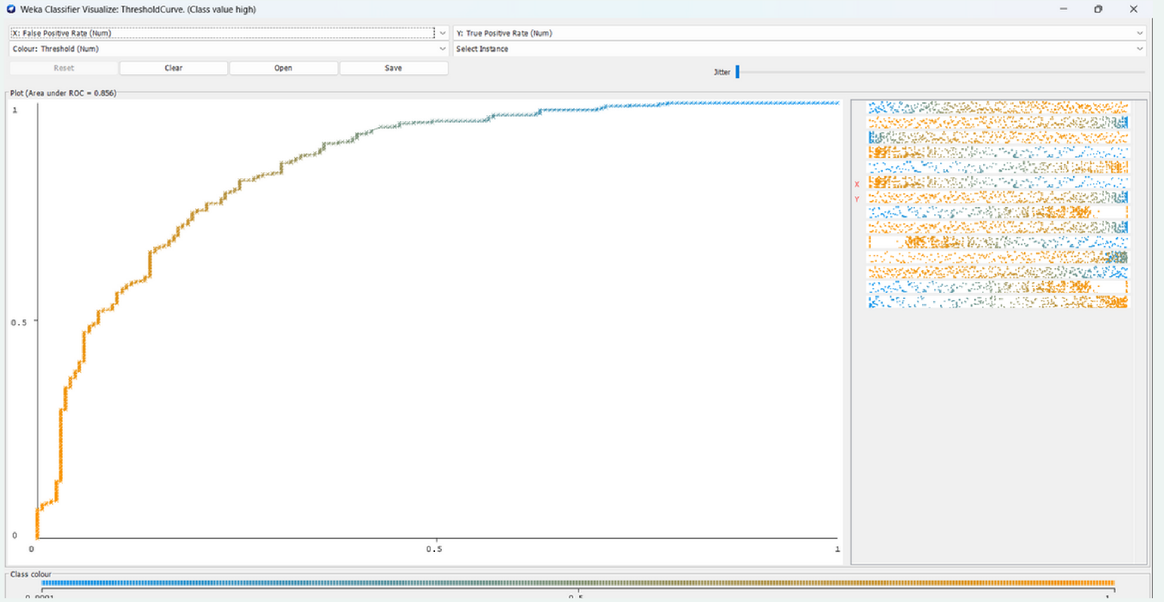
Model	Precision	Recall	f1-score	Accuracy	ROC Area
Naive Bayes Mutinominal Updatable	0.796	0.965	0.872	80.91%	0.873
Voted perceptron	0.836	0.863	0.849	79.28%	0.755
Mutilayer Perceptron (MLP)	0.841	0.896	0.868	81.53%	0.862
Logistic regression with Ridge estimator	0.840	0.890	0.864	81.10%	0.859
Random Forest (max depth = 20)	0.786	0.960	0.865	79.67%	0.849
Sequential Minimal Optimization (SMO) (Logistic, RBF kennel)	0.843	0.872	0.857	80.34%	0.853

Based on the cross-validation results these models are chosen to test on the final test-set

1. Naive Bayes Mutinominal Updatable
2. Multilayer Perceptron (MLP)
3. Logistic regression Ridge
4. SMO (Logistic, RBF kennel)

SCENARIO 1 - RESULTS

Model	Precision	Recall	f1-score	Accuracy	ROC Area
Mutilayer Perceptron (MLP)	0.837	0.924	0.873	81.84%	0.857
Logistic with Ridge	0.836	0.907	0.870	81.65%	0.856
SMO	0.820	0.915	0.865	80.69%	0.851
Naive Bayes Multinominal Updatable	0.773	0.972	0.861	78.78%	0.859



- The best model is selected based on the combination of metrics including precision, recall, F1-score, accuracy, and AUC
- The objective is to predict whether a review is considered a high rating or a low rating (i.e., classified as True). Therefore, the metrics for the True class were considered, with priority given to precision rate and accuracy to determine the best model.
- Based on these criteria, the Logistic model emerges as the optimal choice for prediction since it takes less computation power and achieve almost the same accuracy as complex models such as MLP.

SCENARIO 1 - LIMITATION & RECOMMENDATION

With the limited amount of time, data, and computational resources (personal computer), building a more complex model can take a lot more development time. However, with an 80%+ accuracy, the model can be used with a high degree of confidence to solve real-world business problems.

WORD EMBEDDING

- Instead of low dimension TF-IDF transformation utilize Word embedding method on the entire text using Pytorch library to model a more complex relationship of NLP

USE RNN AND LTSM

- Recurrent neural network and Long-term short memory model are the two advanced NLP models developed to capture language complexity

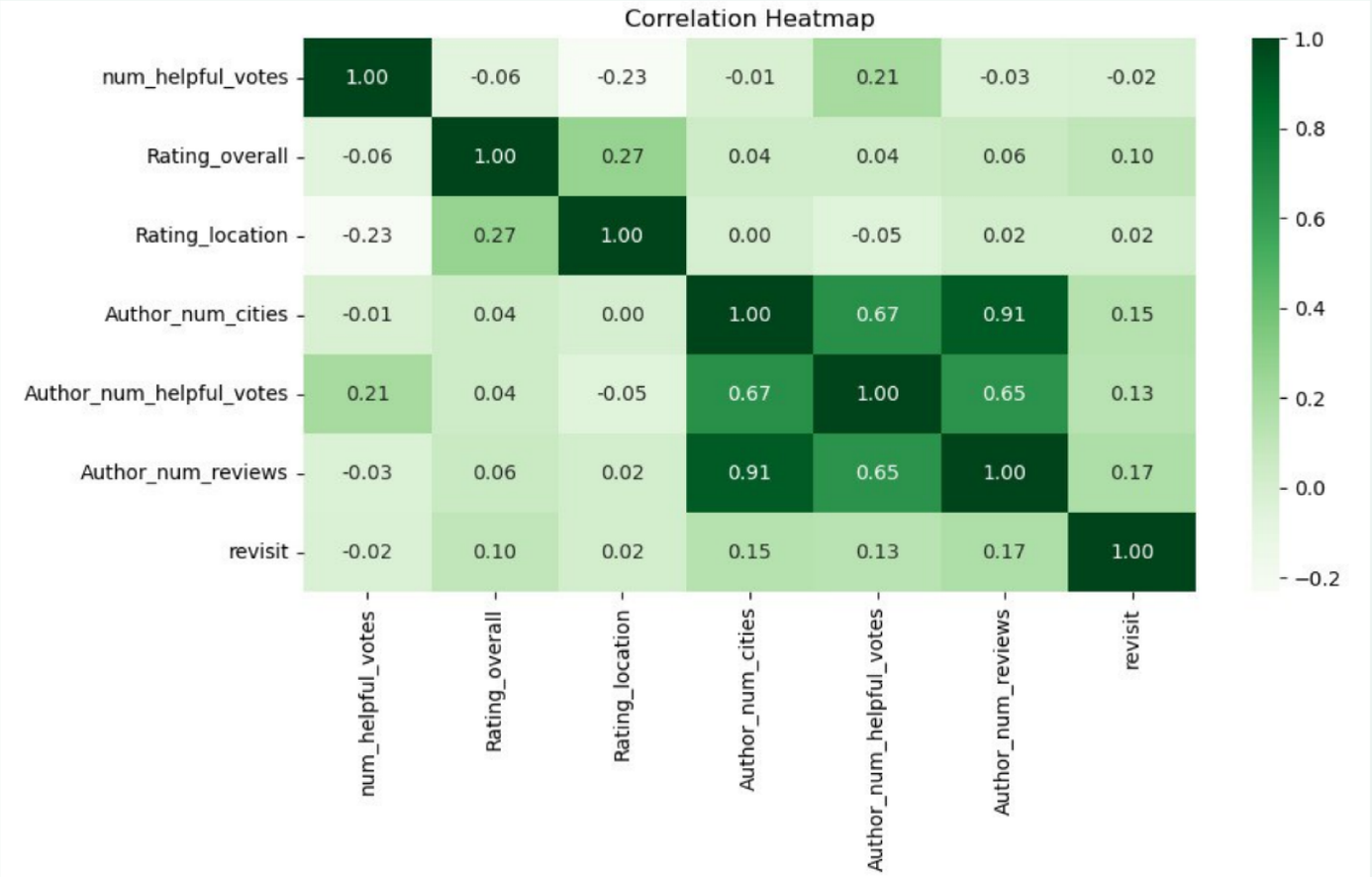
DATA ENRICHMENT

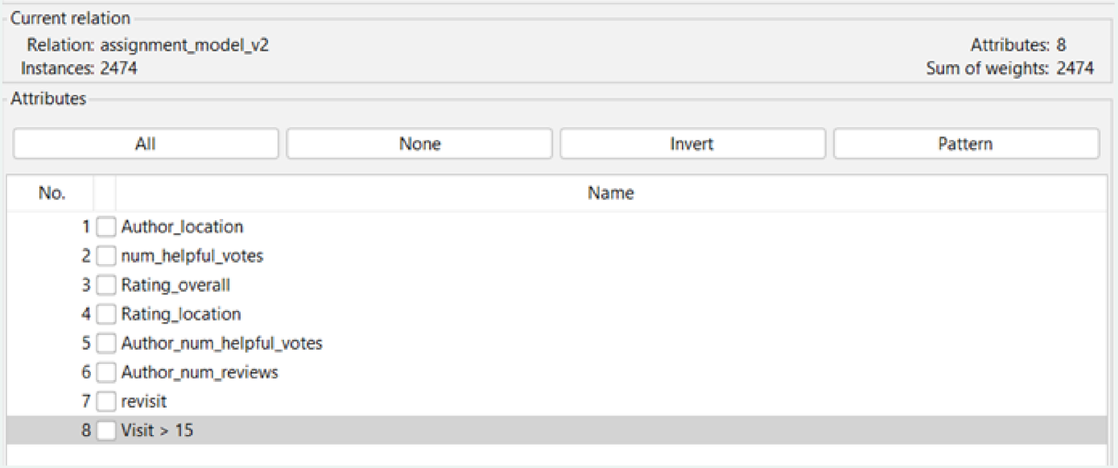
- Synonym replacement (SR)
- Random insertion (RI)
- Random substitution (SR)
- Random delete (RD)
- Reverse translation

SCENARIO 3 - EDA

- The dependent variable “Author_num_cities” exhibits a strong correlation with “Author_num_helpful_votes” and “Author_num_reviews”. It is hypothesized that these two variables will significantly contribute to predicting authors who have visited more than 15 cities. For example, they could serve as key nodes in the decision tree model, as we will later test during the model building with Weka.

- There are strongly correlated features in the dataset, including “Rating_overall”, “Rating_service”, “Rating_cleanliness”, “Rating_value”, “Rating_sleep_quality”, and “Rating_rooms”. Hence, only the “Rating_overall” attribute will be included in the model. The remaining attributes need to be removed to mitigate multicollinearity issues.





SCENARIO 3 - TRAIN TEST SPLIT

- **NumericToNominal**with “Rating_overall”, “Rating_location”, “Visit > 15”, and “Author_num_helpful_votes”: convert these numeric data to categorical data.

- **StringToNominal** and **NominalToBinary**with “Author_location”: convert string data to categorical data, then to dummy variable for machine learning models.

- **NominalToBinary**with “Visit > 15”: convert categorical (True/False) values to binary (0-1).

- **NumericToBinary** with “revisit”: convert numeric (0-1 values) to binary values for machine learning models

- **Removed** “Author_location=Others” attribute to mitigate multicollinearity after creating dummy variables.

- **Resample:** split the dataset into 3 parts with a ratio of 60 – 20 – 20 for the training dataset, validation dataset, and test dataset, respectively.

- **SMOTE**: rebalance the **training data**! The true (real-life) distribution of Author_num to "non-churn" is imbalanced.

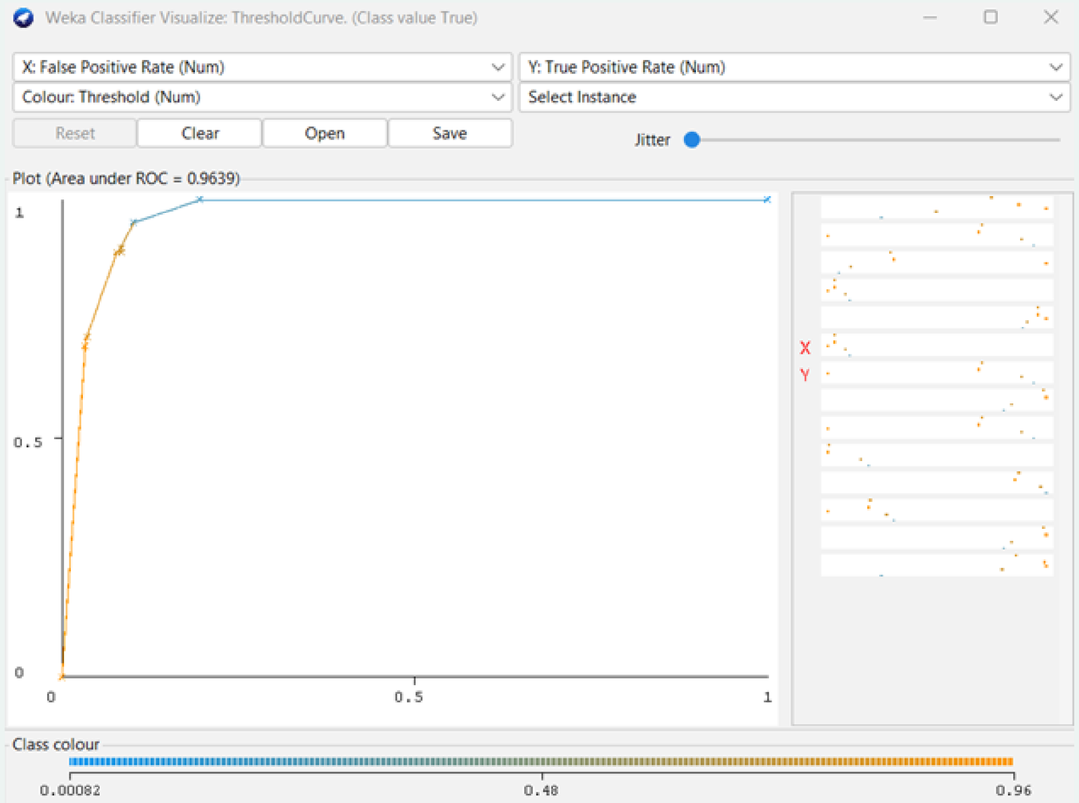
SCENARIO 3 - MODEL BUILDING USING CV

Model	Precision	Recall	f1-score	Accuracy	ROC Area
OneR (Baseline model)	0.936	0.906	0.921	93.23%	0.929
Logistic Regression (Choose to test with test dataset)	0.932	0.925	0.928	93.79%	0.982
Naïve Bayes	0.933	0.807	0.866	89.12%	0.966
Decision Tree - Min 2 instances per leaf	0.912	0.955	0.933	94.06%	0.962
Decision Tree - Min 10 instances per leaf (Choose to test with test dataset)	0.908	0.959	0.933	93.98%	0.972
Random Forest (Choose to test with test dataset)	0.912	0.960	0.935	94.21%	0.985

- Fit models using the training dataset and applied Cross-Validation to mitigate overfitting.
- Based on the Cross-Validation results, Logistic Regression, Decision Tree (with a minimum of 10 instances per leaf), and Random Forest to test with the test dataset.

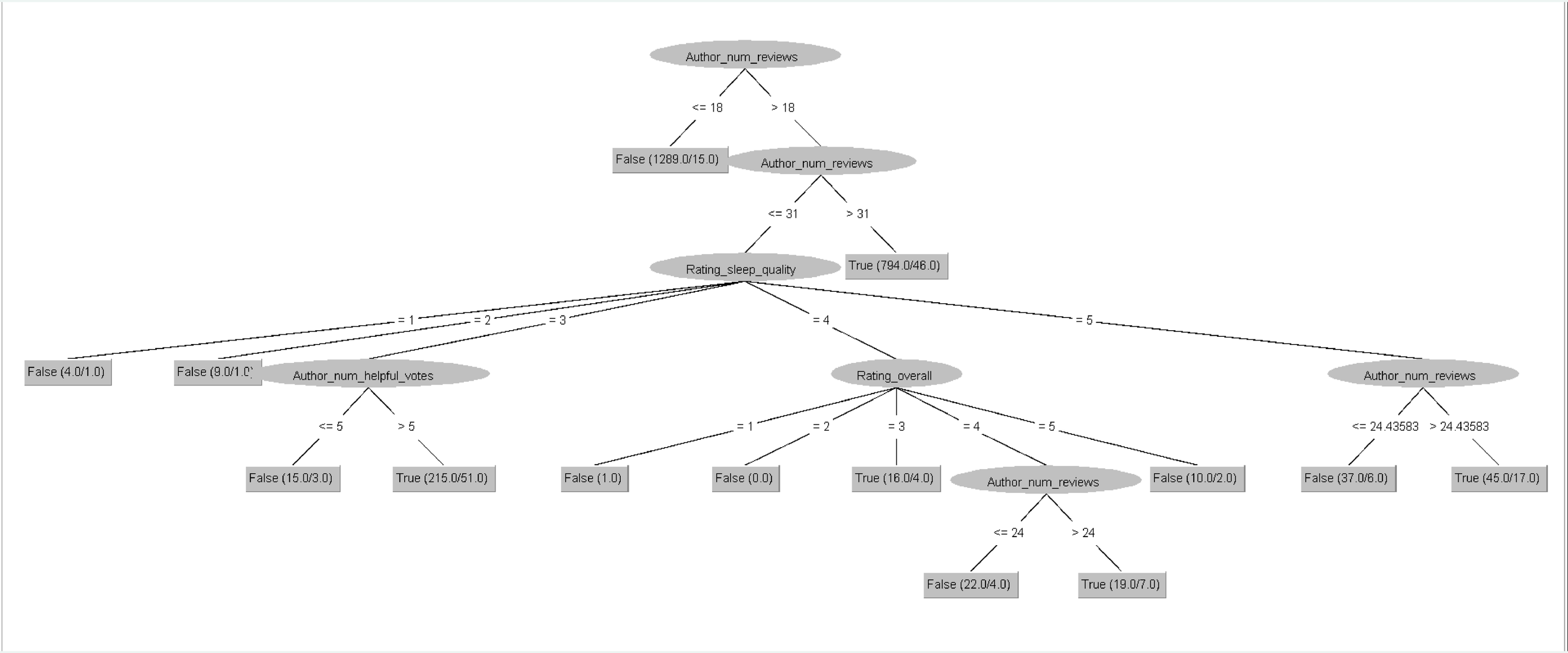
SCENARIO 3 - RESULTS

Model	Precision	Recall	f1-score	Accuracy	ROC Area
Decision Tree – Min 10 instances per leaf	0.766	0.897	0.827	91.13%	0.964
Logistic Regression	0.712	0.949	0.813	89.72%	0.967
Random Forest	0.752	0.906	0.822	90.73%	0.963



- The best model is selected based on the combination of metrics including precision, recall, F1-score and accuracy.
- 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, with priority given to precision rate and accuracy to determine the best model.
- Based on these criteria, the Decision Tree model emerges as the optimal choice for prediction..

SCENARIO 3 - TREE VISUALIZATION



SCENARIO 3 - LIMITATIONS AND RECOMMENDATIONS

Using simple metrics such as passport strength and total number of reviews, we can have a good prediction of those who have traveled to more than 15 countries. With an accuracy of 91%, the model can be trusted be used in real word problems. However, there are still areas to be improved on:

COST-SENSITIVE ANALYSIS

- Using cost matrix to penalize False positive predictions and maximize True positive predictions

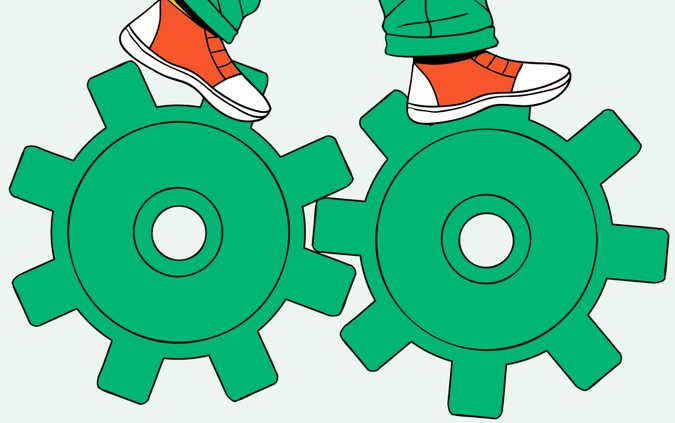
TRAINING DATA

- Increase training data can potential help the model to deal with variations in classifier which is currently not accounted for

BOOSTING AND BAGGING

- Utilize boosting, bagging to further improve the model accuracy

REAL WORLD APPLICATIONS



HOTEL CHAIN

Hotel chain management can utilize this model to identify problematic areas in the chain, which can be a particular hotel in the chain, the hotel processes or services, and take immediate actions against it.

COMPETITORS

Using the model to extract the perceived strengths and weaknesses of a hotel (chain) and use that to benchmark against its competitors to find its **Unique selling points** for increasing competitive advantage

SEGMENTATION

The company can use this model for customer segmentation where they could have a better understanding of different customer personas for better products and service offerings.