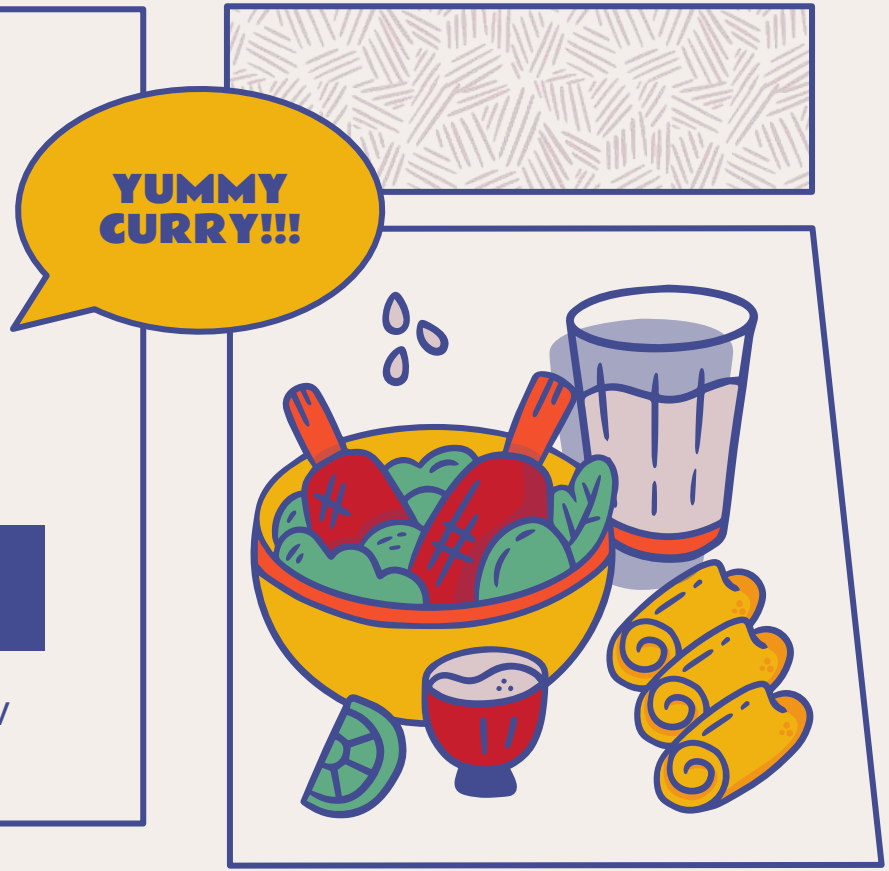
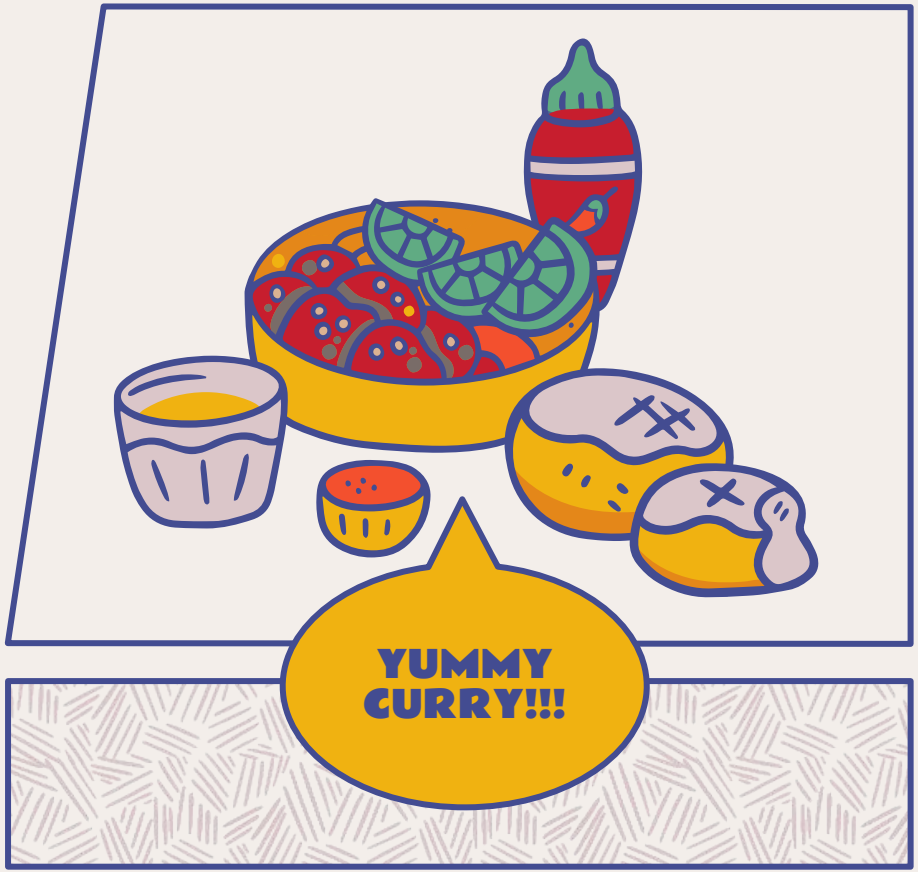
Yelp Dataset Project

DS3010

By: Ethan Falcão, Emre Sabaz, Maanav Iyengar, Nur Fateemah, Sarah Kogan





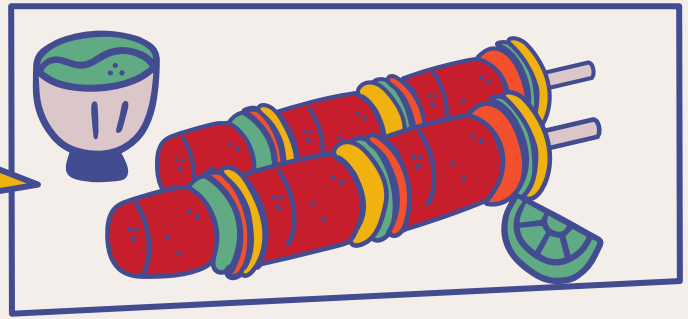
01

TASK 1:

Predicting the Business Attributes using Review and Tip textual information

Solution:

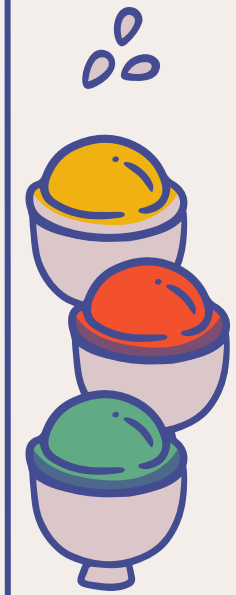
Use review and tip information to extract relevant features from the text data and then use these features to train a machine learning model



YUMMY CURRY!!!

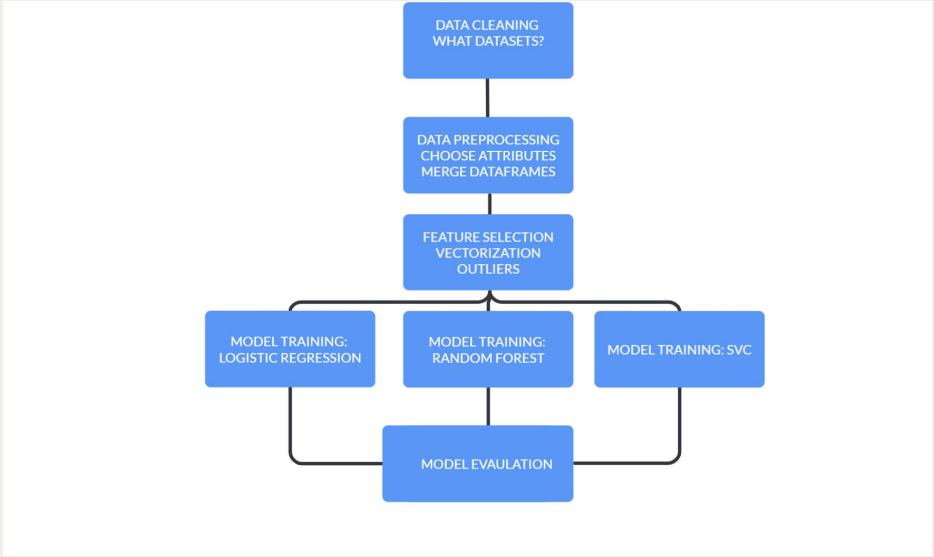
General Process:

- Clean and decrease data size/merge data together; Process attributes column
- Split into training and testing data
- Fit vectorizer with text data and transform
- Train different classifiers and test their accuracy rates



YUMMY CURRY!!!

Process Diagram:



Results (Accuracy & Split):

74% accuracy
(80% training vs. 20% testing)

YUMMY CURRY!!!

Results (Test Performance):

Accuracy: 0.7361950406334653

	precision	recall	f1-score	support
0	0.93	1.00	0.96	4453
1	0.84	0.99	0.90	3900
2	0.85	0.98	0.91	3646
micro avg	0.87	0.99	0.93	11999
macro avg	0.87	0.99	0.93	11999
weighted avg	0.87	0.99	0.93	11999
samples avg	0.86	0.94	0.89	11999

Random Forest Classifier Report

Test Accuracy: 0.7439049802042093

	precision	recall	f1-score	support
0	0.93	1.00	0.96	4453
1	0.84	1.00	0.91	3900
2	0.84	0.99	0.91	3646
micro avg	0.87	0.99	0.93	11999
macro avg	0.87	0.99	0.93	11999
weighted avg	0.87	0.99	0.93	11999
samples avg	0.86	0.95	0.89	11999

SVC Classifier Report

YUMMY CURRY!!!

Results (Confusion Matrix):

```
[[4716    0    0]
 [  80    0    0]
 [  40    0    0]]
```

YUMMY CURRY!!!

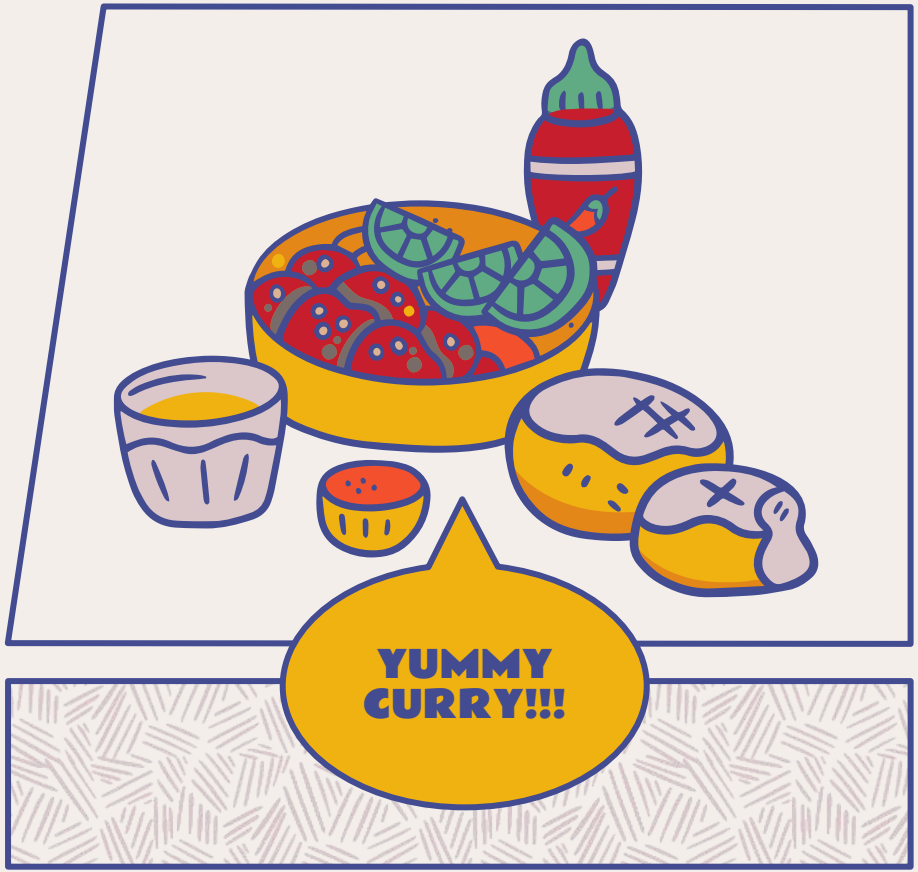
Future Work:

- Consider exploring different feature representation techniques
  - Experiment with alternative approaches such as word embeddings
    - (e.g., Word2Vec, GloVe, etc.)
- Model Selection: Try different models or algorithms to see if they yield better results

YUMMY CURRY!!!

Business Applications:

- By accurately categorizing businesses based on reviews, our model enhances Yelp's search algorithm, providing users with more relevant and tailored search results



02

TASK 2:

Identifying Fake Reviews

Solution:

Utilize natural language processing (NLP) techniques to extract relevant features from the text data, and then apply a supervised learning algorithm to classify reviews as fake or genuine based on these features

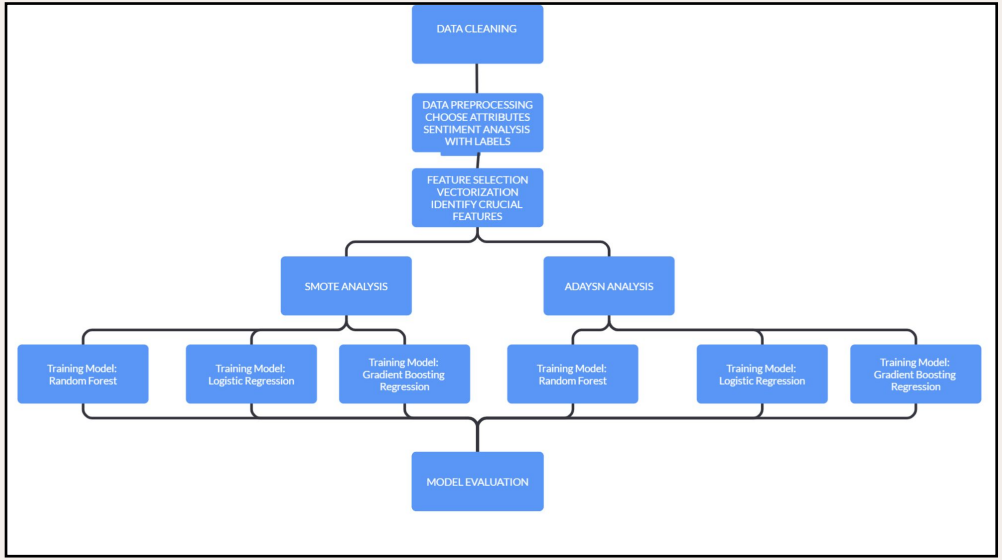
YUMMY CURRY!!!

General Process:

- Data preprocessing and feature engineering
- Machine Learning algorithms
- Model training and evaluating
- Supervised Learning

YUMMY CURRY!!!

Process Diagram:



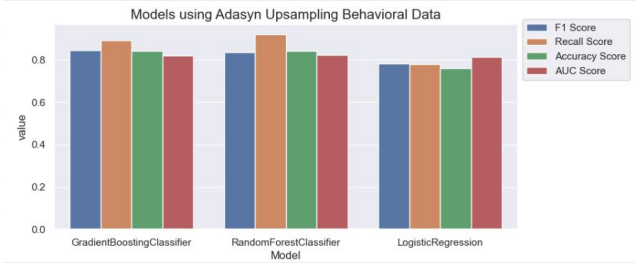
Results (Accuracy & Split):

84% accuracy
(30% training vs. 70% testing)

YUMMY CURRY!!!

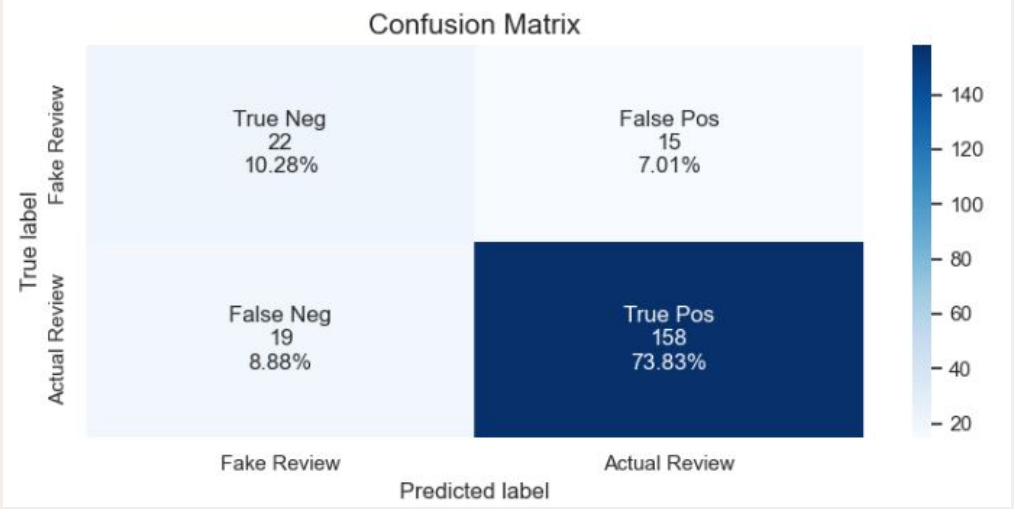
Results (Test Performance):

	Model	variable	value
0	GradientBoostingClassifier	F1 Score	0.844287
1	RandomForestClassifier	F1 Score	0.835436
2	LogisticRegression	F1 Score	0.783698
3	GradientBoostingClassifier	Recall Score	0.892655
4	RandomForestClassifier	Recall Score	0.920904
5	LogisticRegression	Recall Score	0.779661
6	GradientBoostingClassifier	Accuracy Score	0.841121
7	RandomForestClassifier	Accuracy Score	0.841121
8	LogisticRegression	Accuracy Score	0.761682
9	GradientBoostingClassifier	AUC Score	0.821347
10	RandomForestClassifier	AUC Score	0.821805
11	LogisticRegression	AUC Score	0.815086



YUMMY CURRY!!!

Results (Confusion Matrix):



YUMMY CURRY!!!

Future Work:

- Collecting additional data sources such as user profiles and activity history to supplement the review and tip data
- Utilizing more advanced anomaly detection techniques to identify unusual patterns or behaviors in the review data
- Considering the impact of external factors such as cultural nuances and language differences

YUMMY CURRY!!!

Business Applications:

- Fraud detection
- Reputation management
- Sentiment analysis

What We Learned:

- Data preprocessing is a crucial step in ML and can impact model performance greatly
- Model evaluation is a crucial step in data processing, as it is important to look at the way different models change the results
- It is important to consider different angles to approach a problem and to discuss solutions to find inaccuracies we would otherwise miss

YUMMY CURRY!!!

THANK YOU!!