YELP DATA EXPLORATION

Pradip Hayu
Gabe Campese
Data Science Capstone

Agenda

- Problem & Data Collection
- Data Preprocessing & EDA
- Model Construction
- Predicting Star Ratings
- Review Analysis
- Conclusion

Problem Statement & Data Collection

- Problem Statement
- Opening a business is extremely susceptible to risk as that is what comes with being an entrepreneur. Whether it be a restaurant or shopping center, important aspects such as location, demographics, and expectations must be evaluated to determine where and what makes a successful business. More specifically, initial decisions can have everlasting effects because they put performance and investment at stake.

- Yelp Dataset
- Very large dataset
  - *6 separate datasets*
    - Business
    - Review
    - Checkin
    - Tip
    - User
    - Photo
- json to csv format
- Data read into Jupyter Notebook

Literature Review

- Dataset featured in the public Yelp Dataset Challenge
- Plentiful Natural Language Processing (NLP) projects
- Markov Chains
  - *Finishing reviews based on the inputted text*
- Text classification
  - *Return predicted review category based on inputted text*
- Sentiment analysis
  - *Focused efforts on review text*

Data Preprocessing

Pandas dataframe operations

Focusing on restaurant data

- *Merged restaurants data from business with reviews data to narrow in on restaurants*

Business attributes (X) (restaurants)

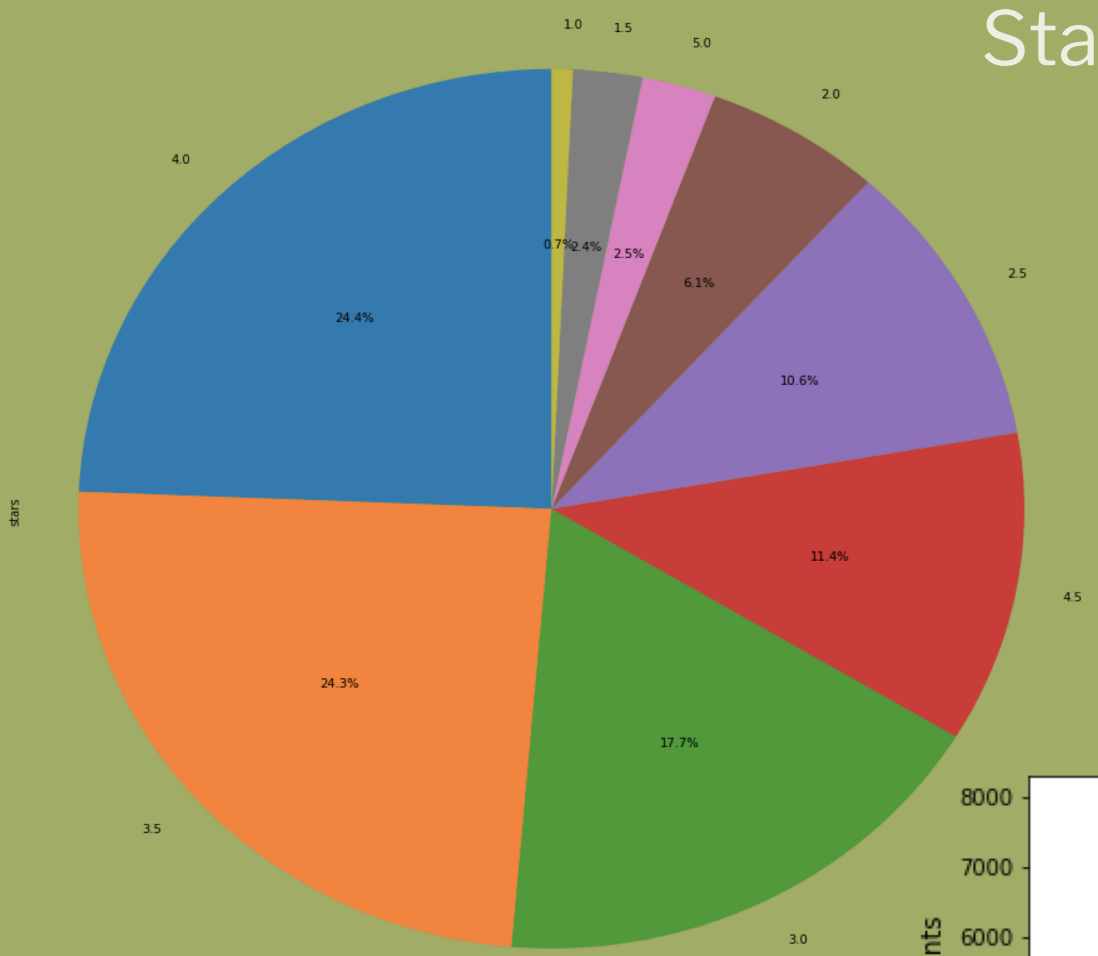
- *Ratings of 4 & above is 1, ratings 3.5 & less are set to 0 (Target Y)*
- *Opens door to many models to predict ratings*

Decision to disregard 4 datasets including: User, Tip, Hours, and Check-In data

Assessment of NaN data

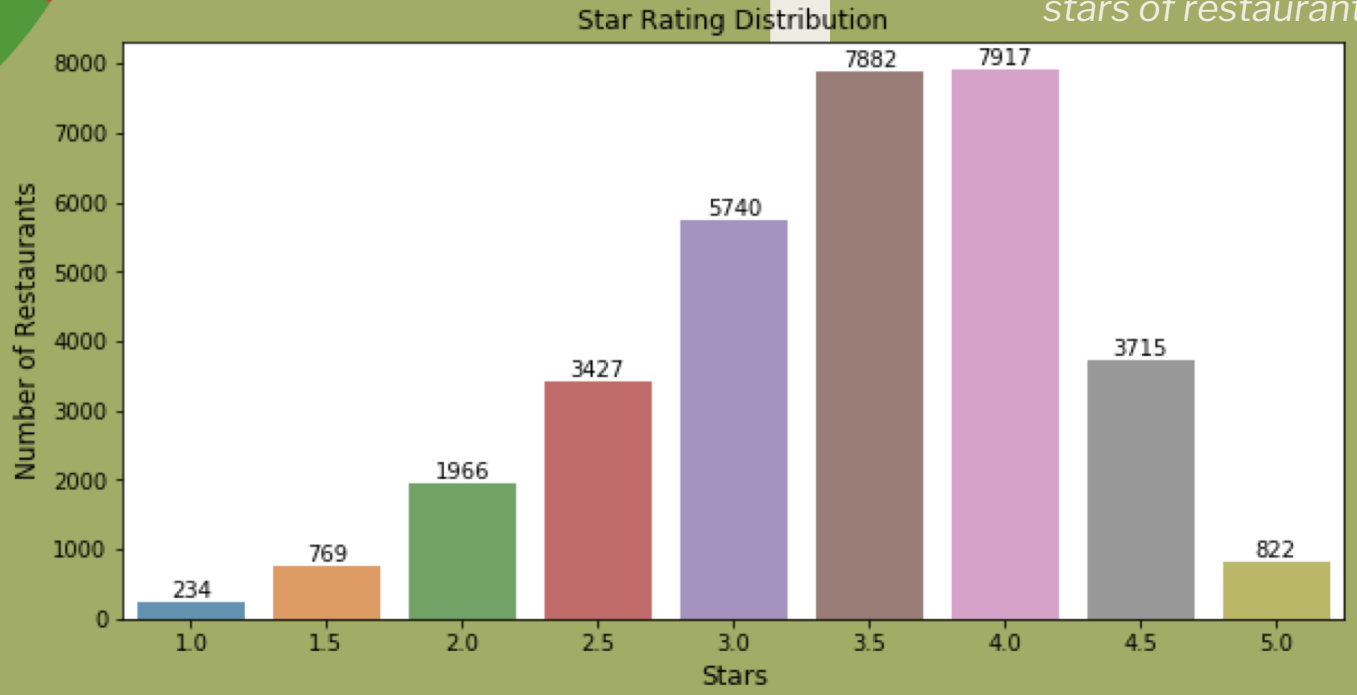
- *Columns with more than 70% NaNs were removed*

Star Rating Distribution

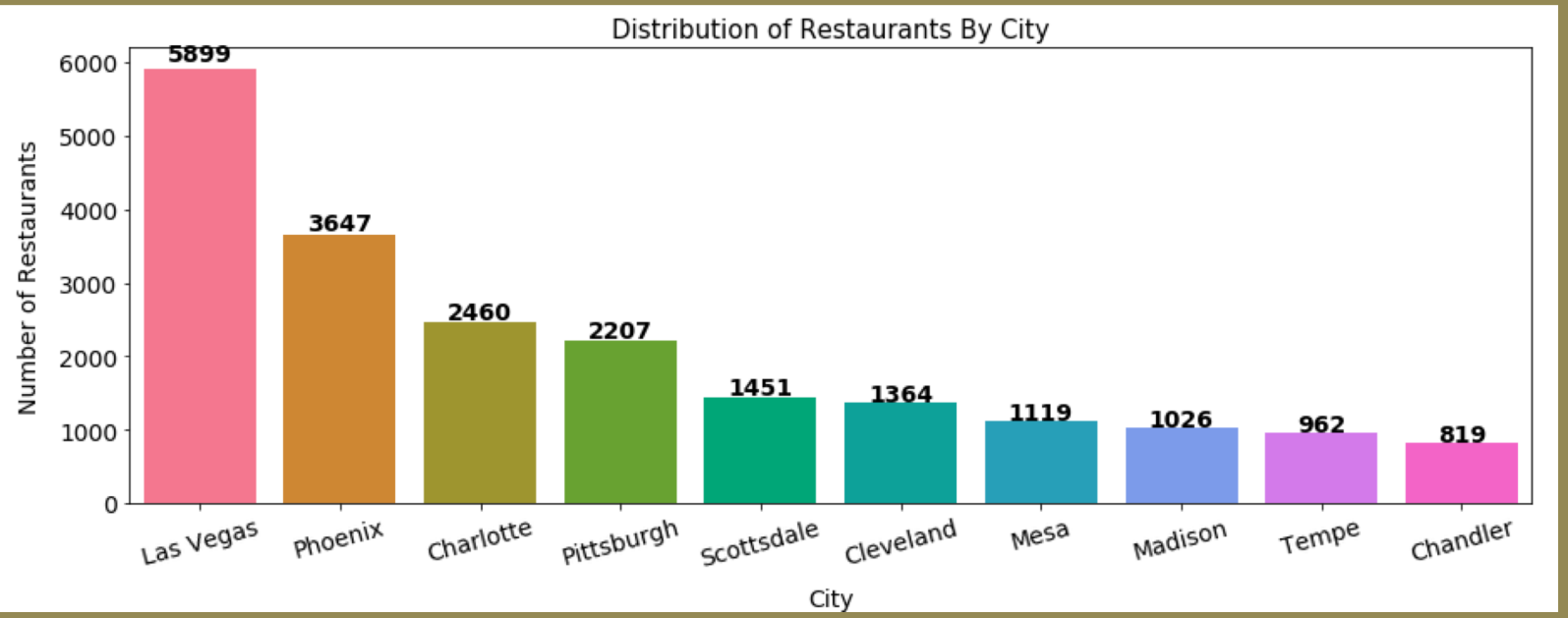


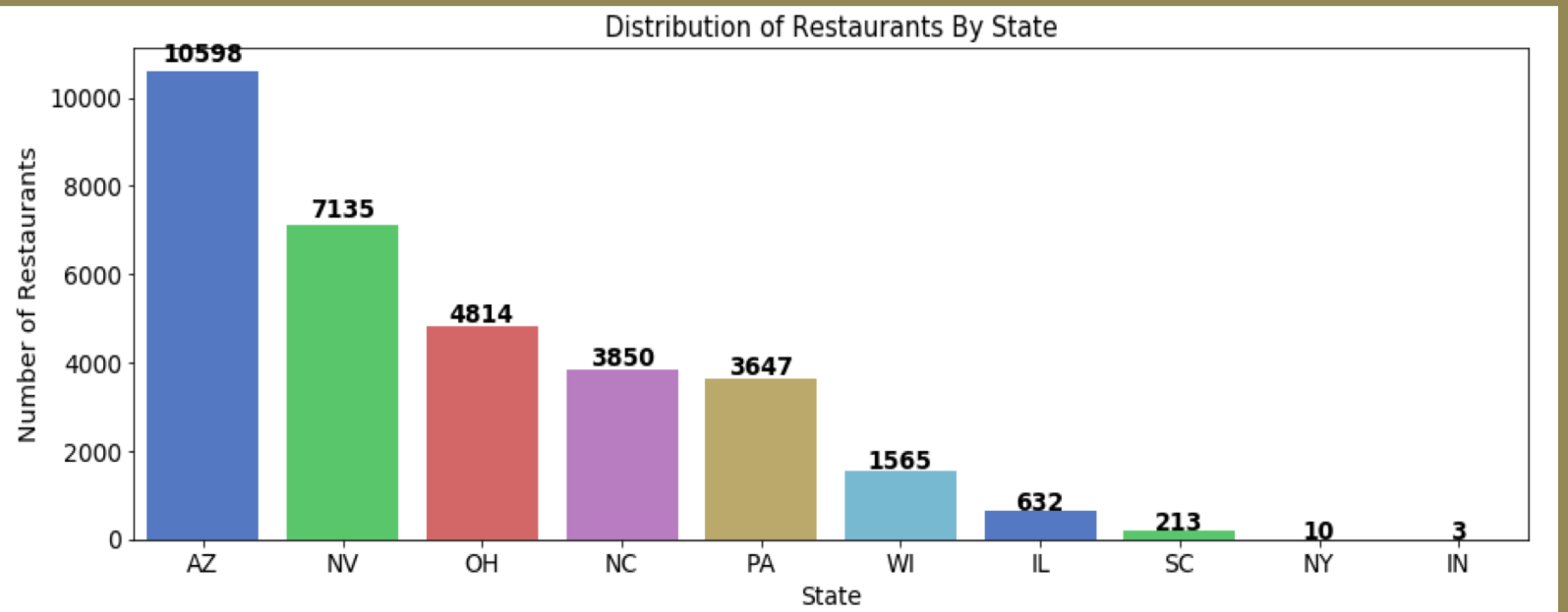
Exploratory Data Analysis (EDA)

- Pandas and numpy operations
  - *Seaborn and matplotlib for viz*
- Star distribution
  - *Gives foundation*
  - *General distribution of stars of restaurants*



Distribution of Restaurants By City and State

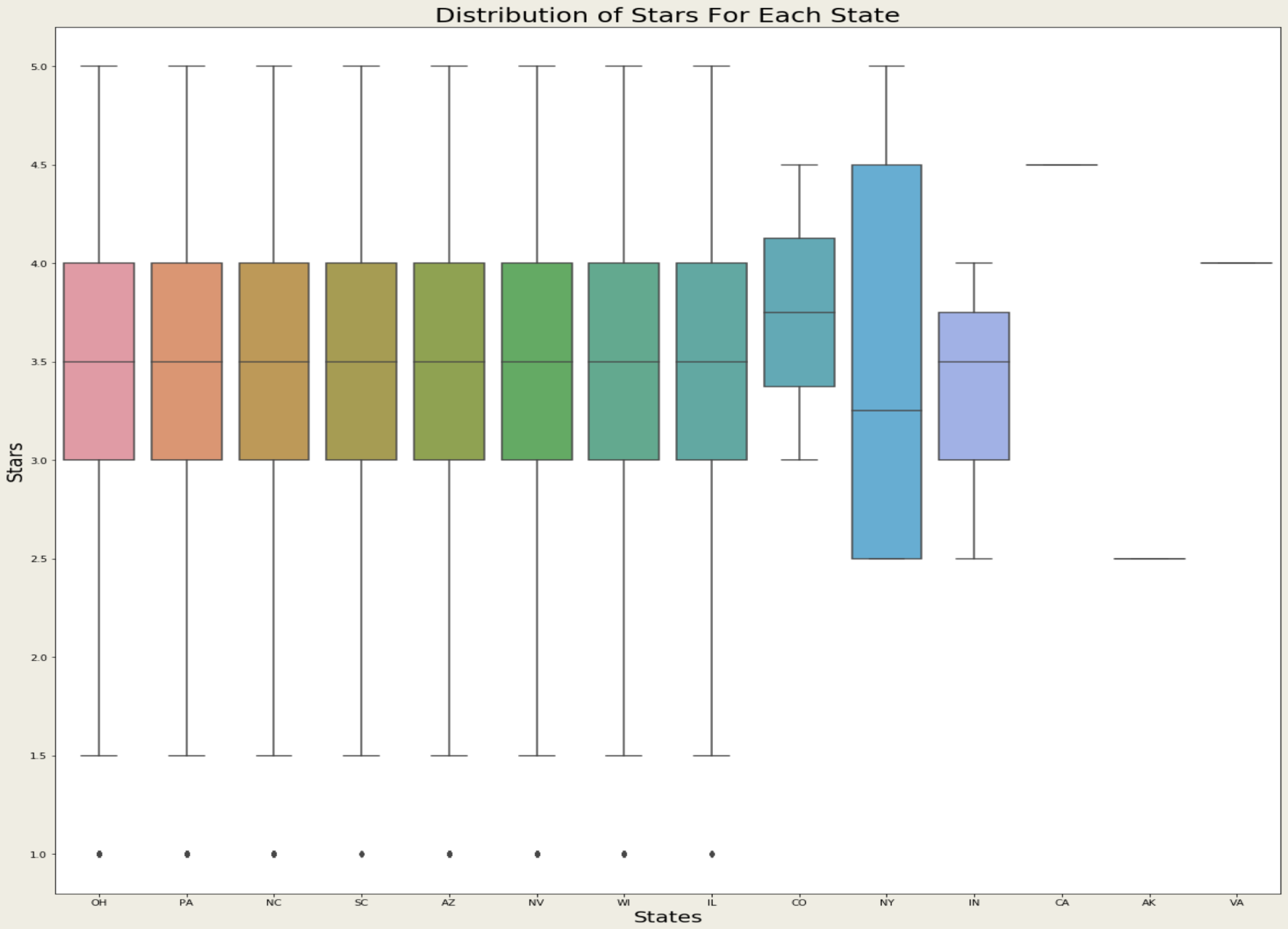




Exploratory Data Analysis (EDA)

- Pandas and numpy operations
  - *Seaborn and matplotlib for viz*
- Summarize main ideas of data to get ball rolling
- Great initial landscape assessment
  - *Distribution of where the restaurants are*
  - *City*
  - *State*

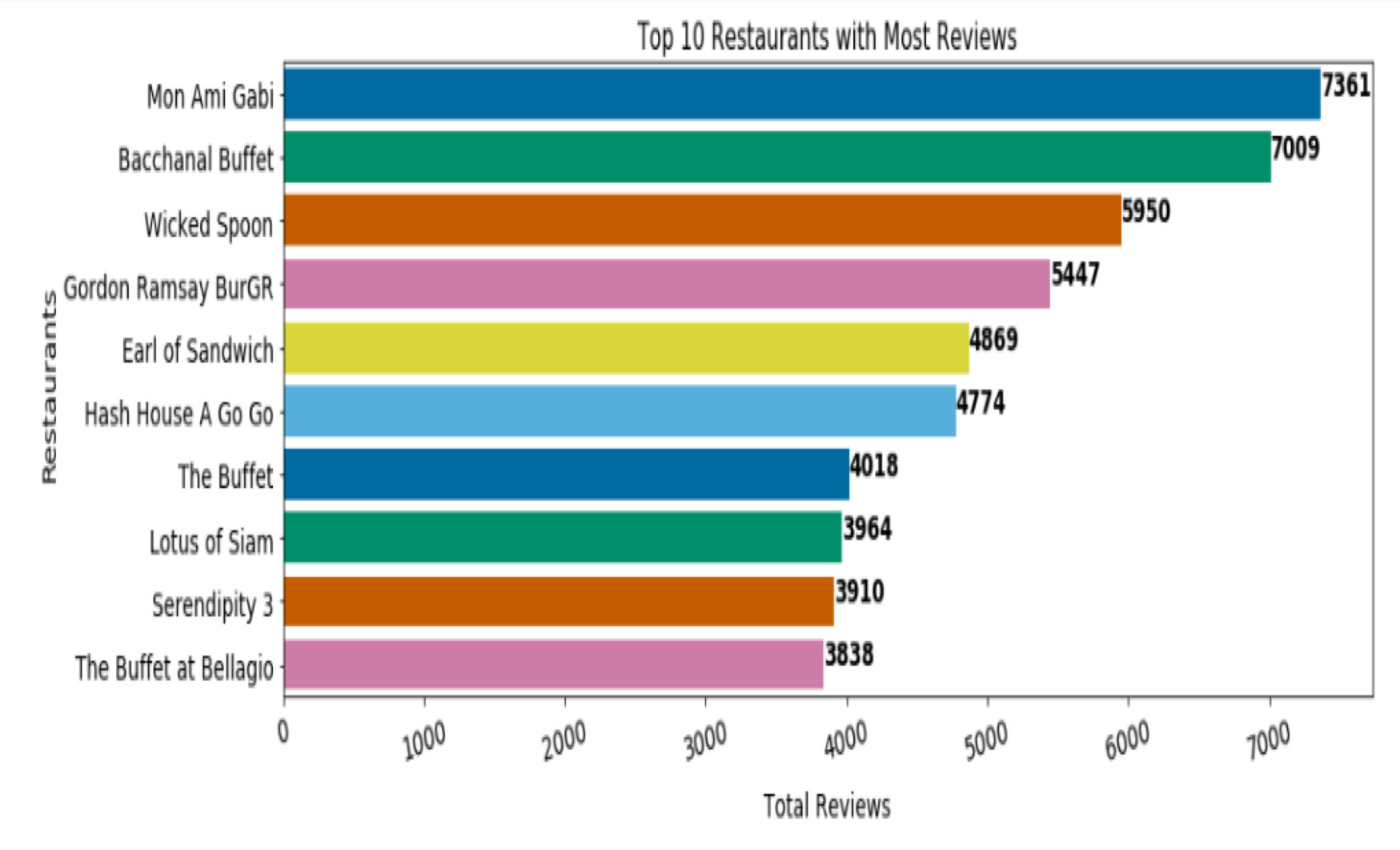
Distribution of Stars For Each State



Exploratory Data Analysis (EDA)

- Pandas and numpy operations
  - *Seaborn and matplotlib for viz*
- Star rating distribution for each state's restaurants

Most Reviewed Restaurants



Exploratory Data Analysis (EDA)

- Pandas and numpy operations
  - *Seaborn and matplotlib for viz*
- General distribution of reviews of restaurants

Predicting Star Rating

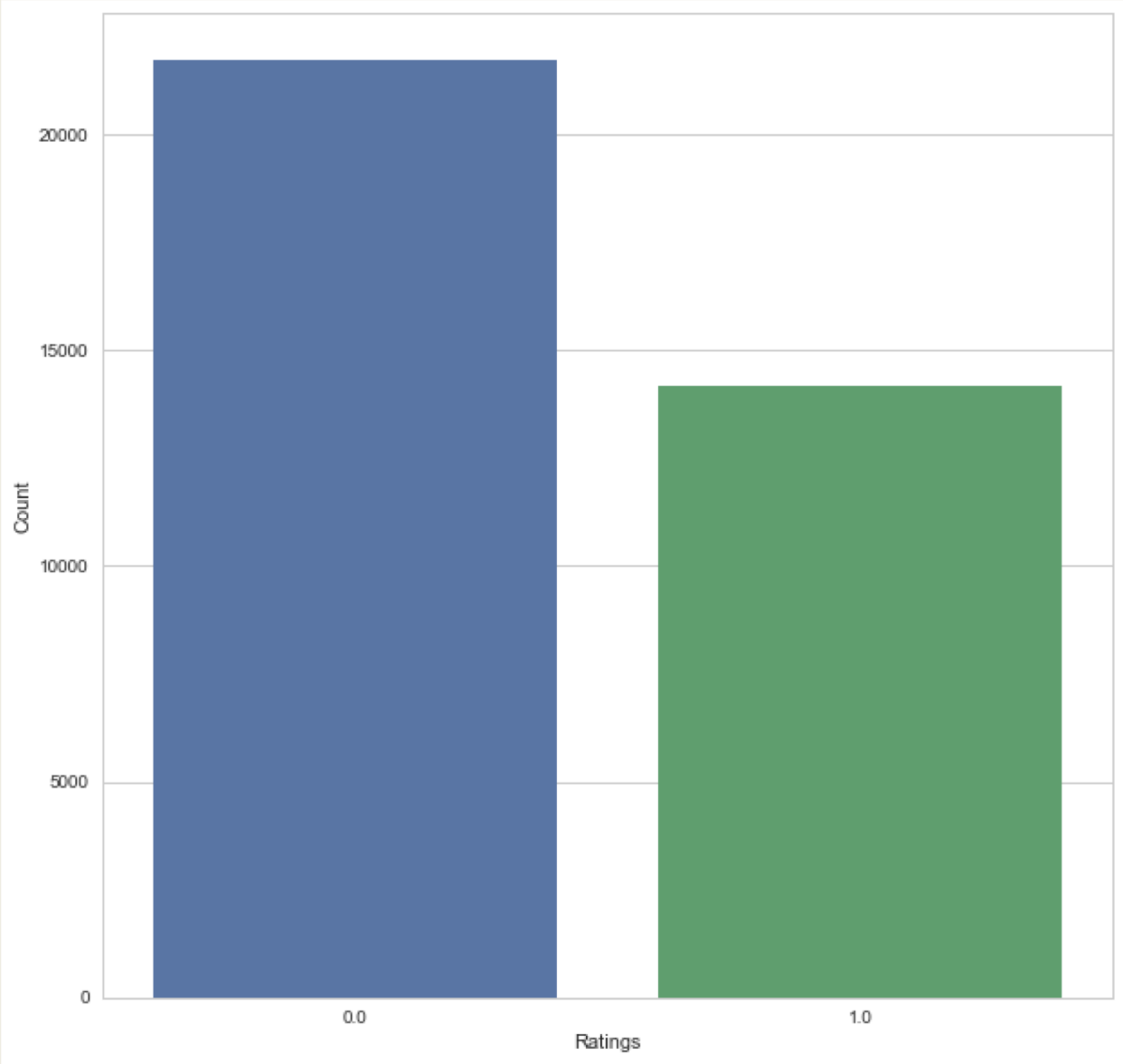
Building of various models

Comparison of models & their scores

Hyperparameter Tuning

Model Selection

Model Selection



- ❑ Features and Target
- ❑ Train, Test, Split
- ❑ OverSampling
- ❑ Classifiers:
  - ➢ Logistic Regression
  - ➢ Decision Tree
  - ➢ Random Forest
  - ➢ K-Nearest Neighbors
  - ➢ Gaussian Naive Bayes
- ❑ Model Evaluation: Accuracy, Precision and Recall

Random Forest

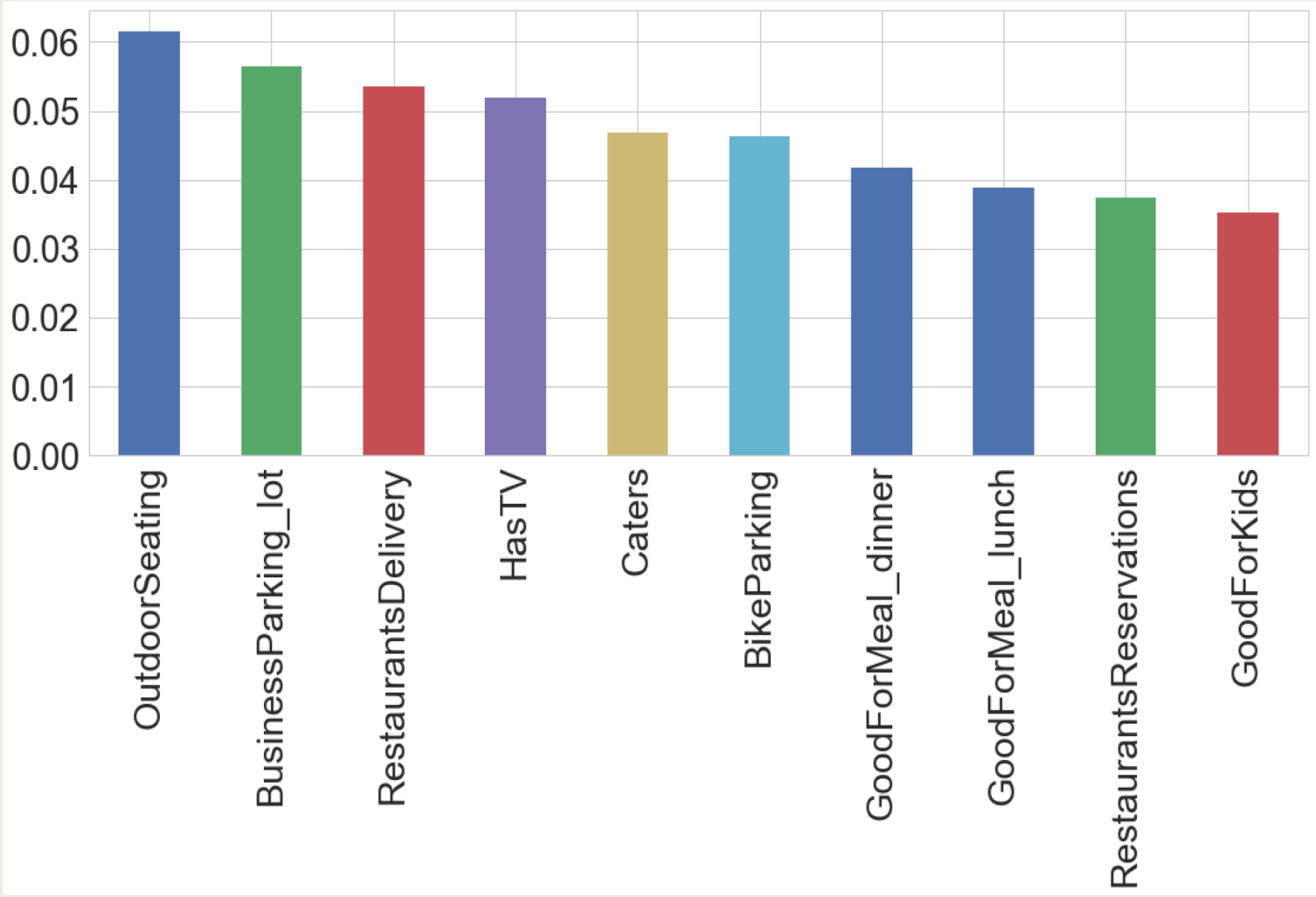
Confusion Matrix

	Predicted[0]	Predicted[1]
[Actual]0	4237	2280
[Actual]1	1935	2312

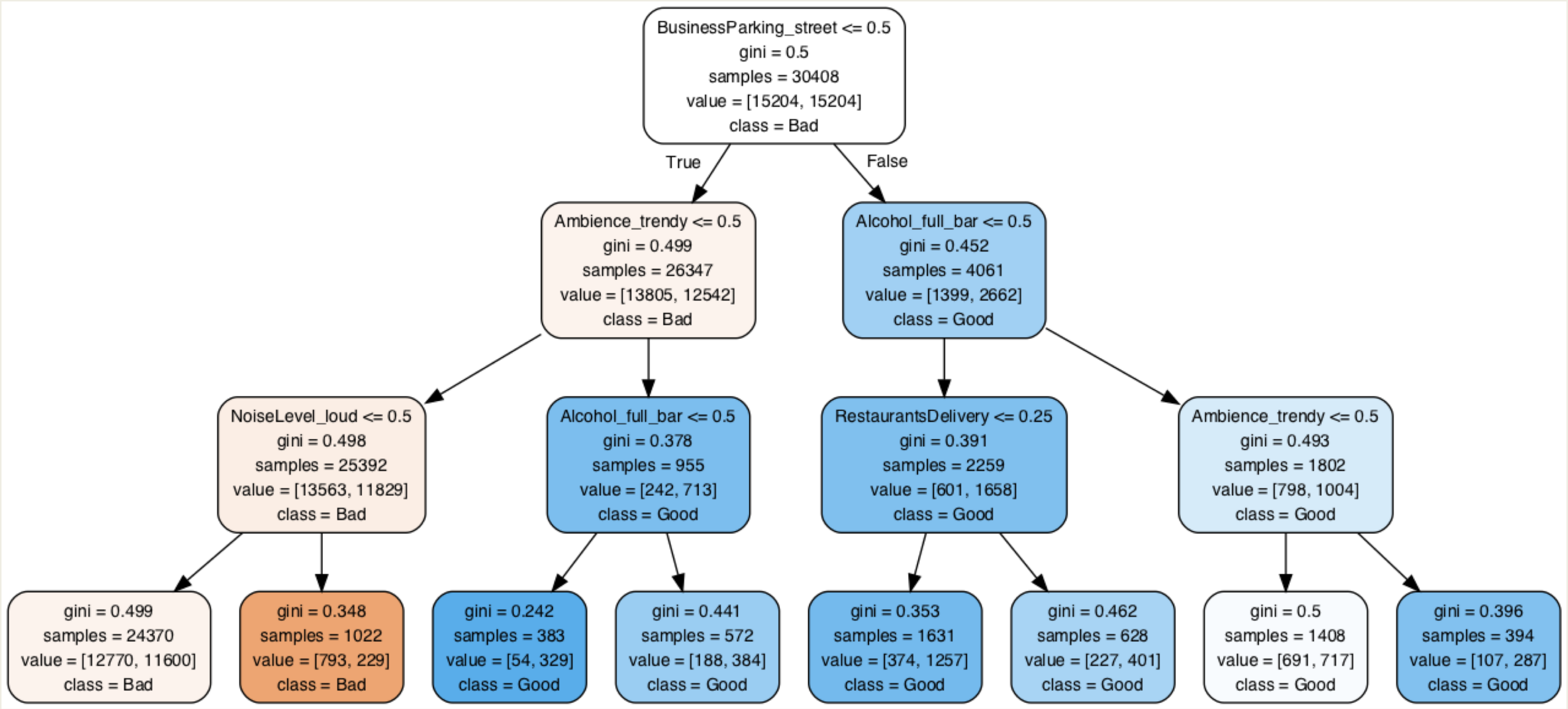
Classification Report

	precision	recall	f1-score	support
0.0	0.69	0.65	0.67	6517
1.0	0.50	0.54	0.52	4247

Feature Importance



Decision Tree



Hyperparameter tuning & Model Selection

Model	**Hyperparameters Used**	**Time Taken For Execution(sec)**	**Accuracy(%)**
Random Forest(rf)	n_estimators, min_samples_split and min_samples_leaf.	84.54	68.03
Decision Tree(dt)	min_samples_split and min_samples_leaf.	14.86	66.29
KNN(knn)	n_neighbors	5166.92	64.27
Logistic Regression(lr)	multi_class, solver and C	550.60	61.83
Gaussian Naive Bayes(gnb)	var_smoothing	2.46	53.21

Review Analysis

- From the US restaurants data, 7 different cuisines were selected:
  Chinese, French, Italian, Japanese, Korean, Thai, & Vietnamese

- Labelled reviews as positive or negative:
  Individual Review Star equal to or greater than 4 as ‘Positive’
  Individual Review Star 3 or less than 3 as ‘Negative’

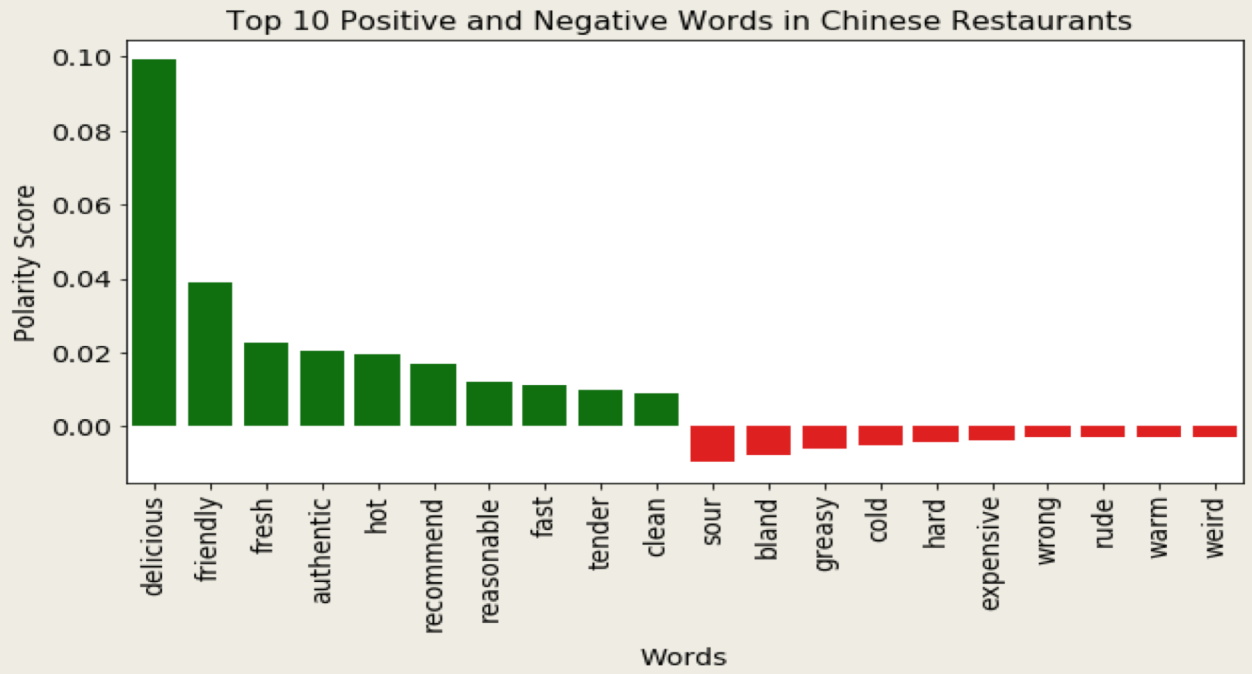
- Features = ‘bag of words’ which is the frequencies of words in each review

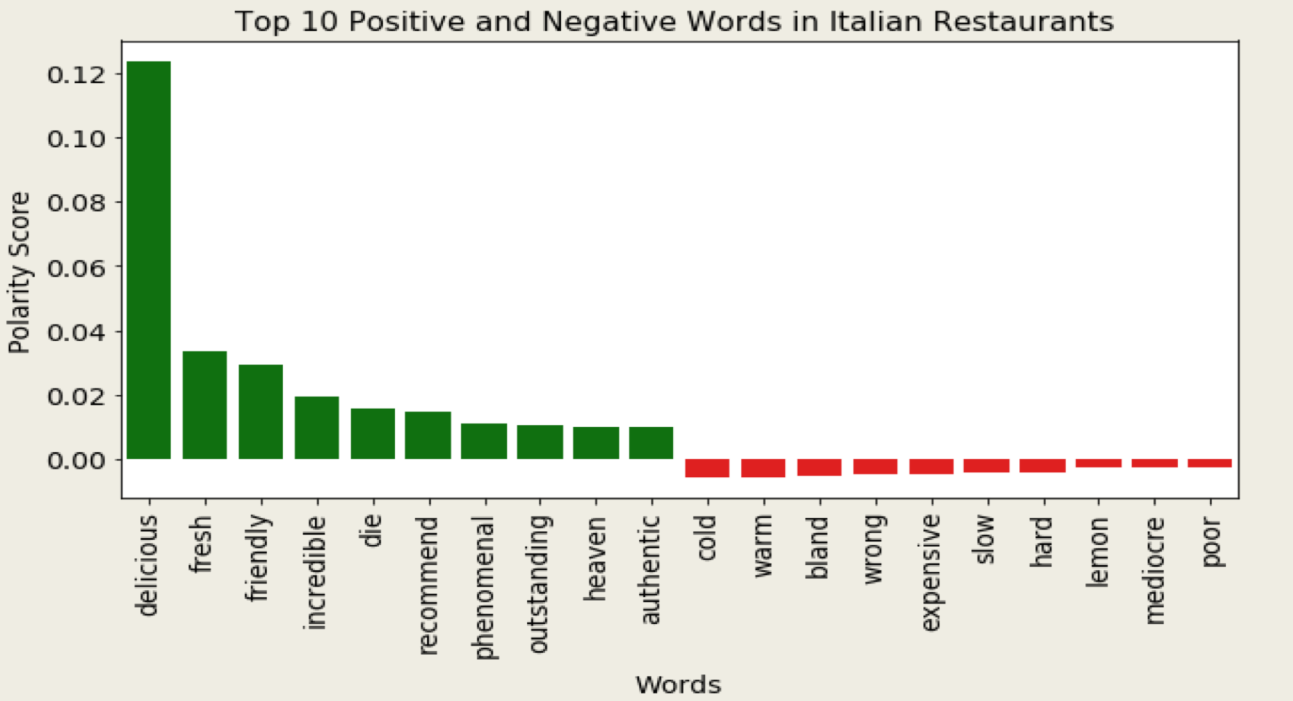
- Support Vector Machine(SVM) model was implemented to get relatively positive and negative words and get score of each word

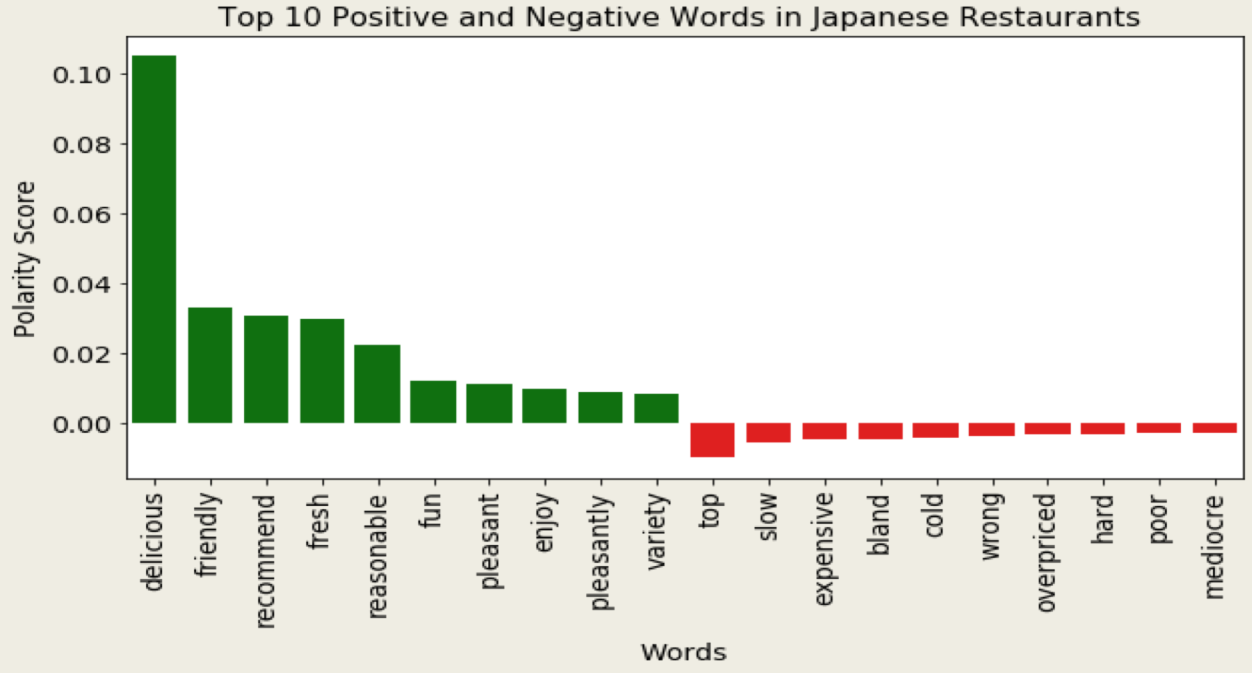
- Polarity score of each word

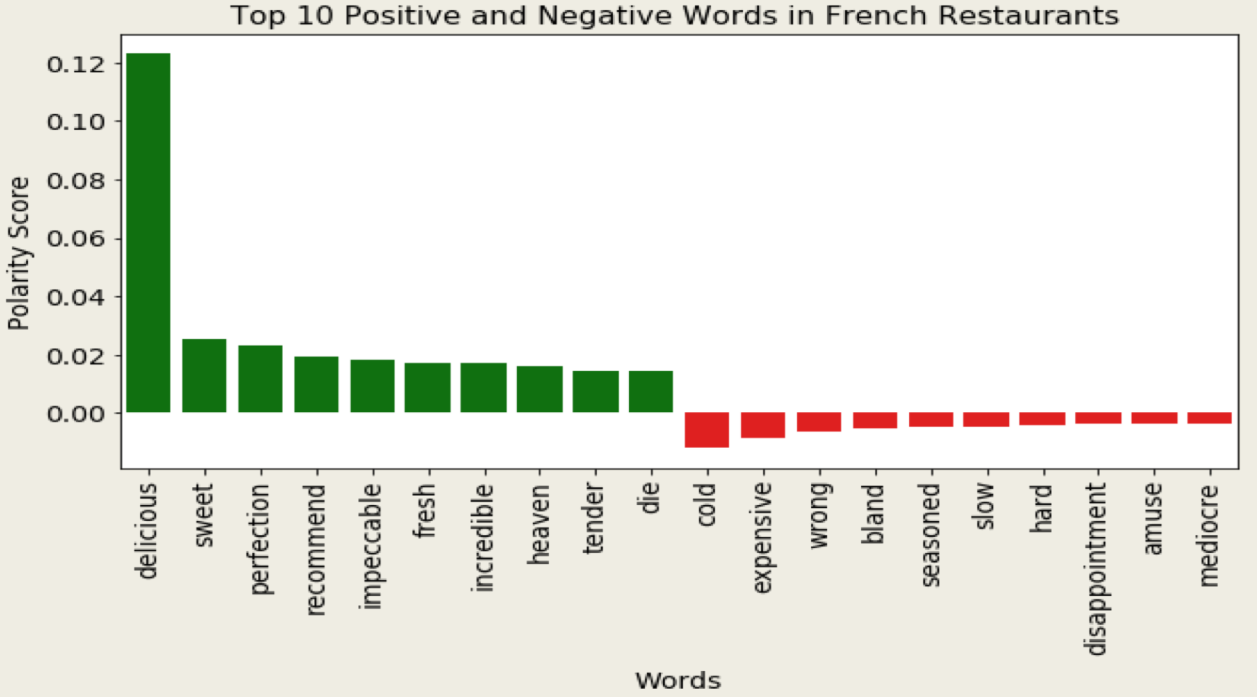
- Top 10 positive & top 10 negative words for each cuisine based on polarity score

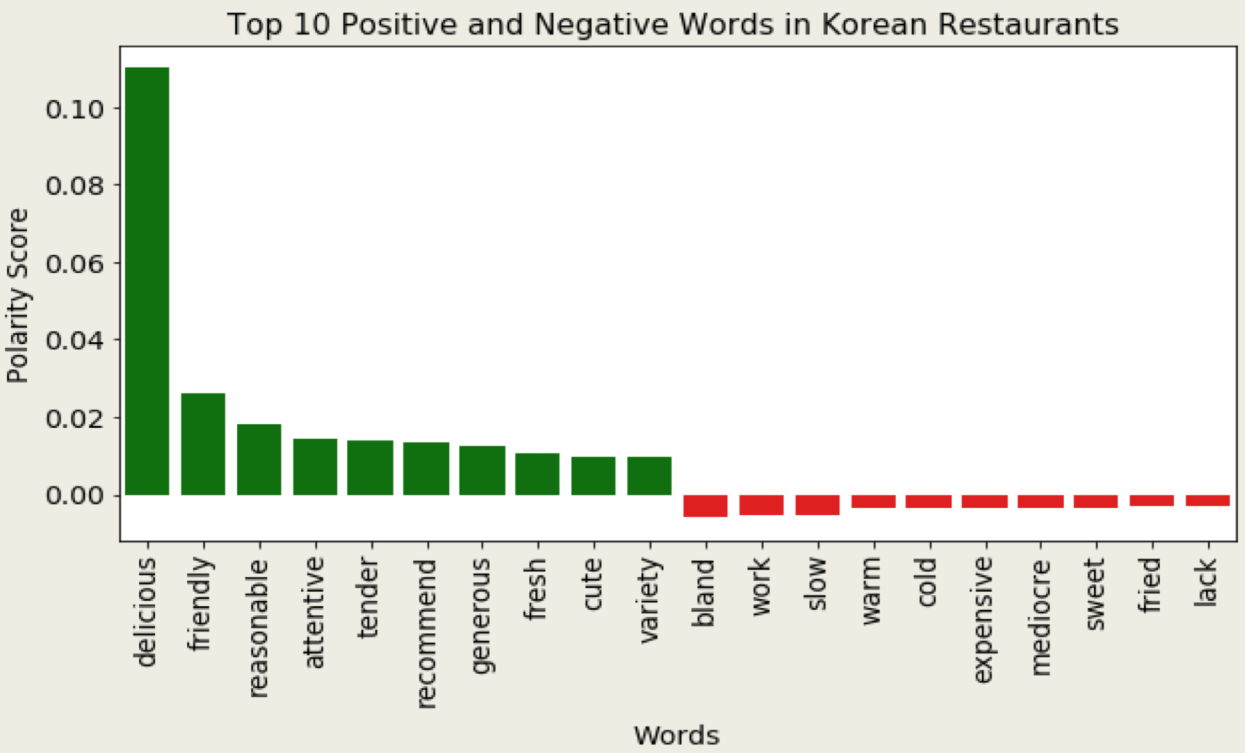
- Make recommendations based on those top words

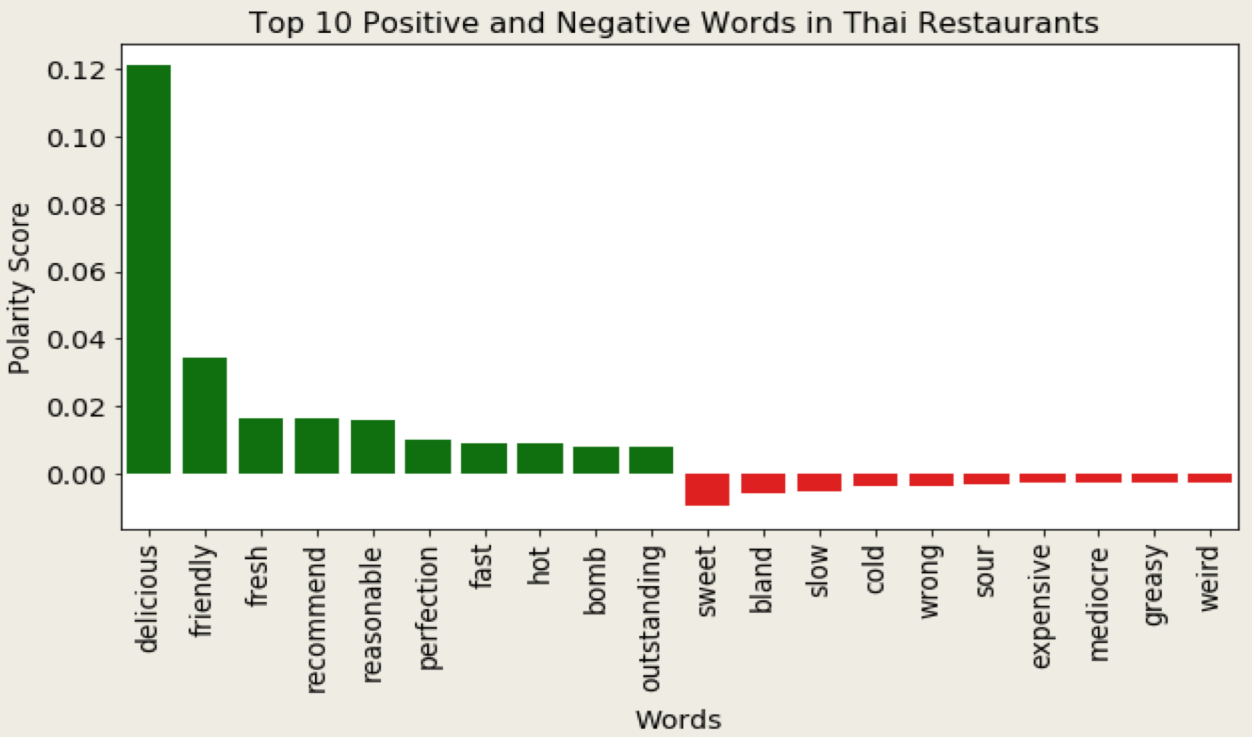

Top 10 Positive and Negative Words in Chinese Restaurants
Polarity Score
0.10
0.08
0.06
0.04
0.02
0.00
delicious
friendly
fresh
authentic
hot
recommend
reasonable
fast
tender
clean
sour
bland
greasy
cold
hard
expensive
wrong
rude
warm
weird
Words

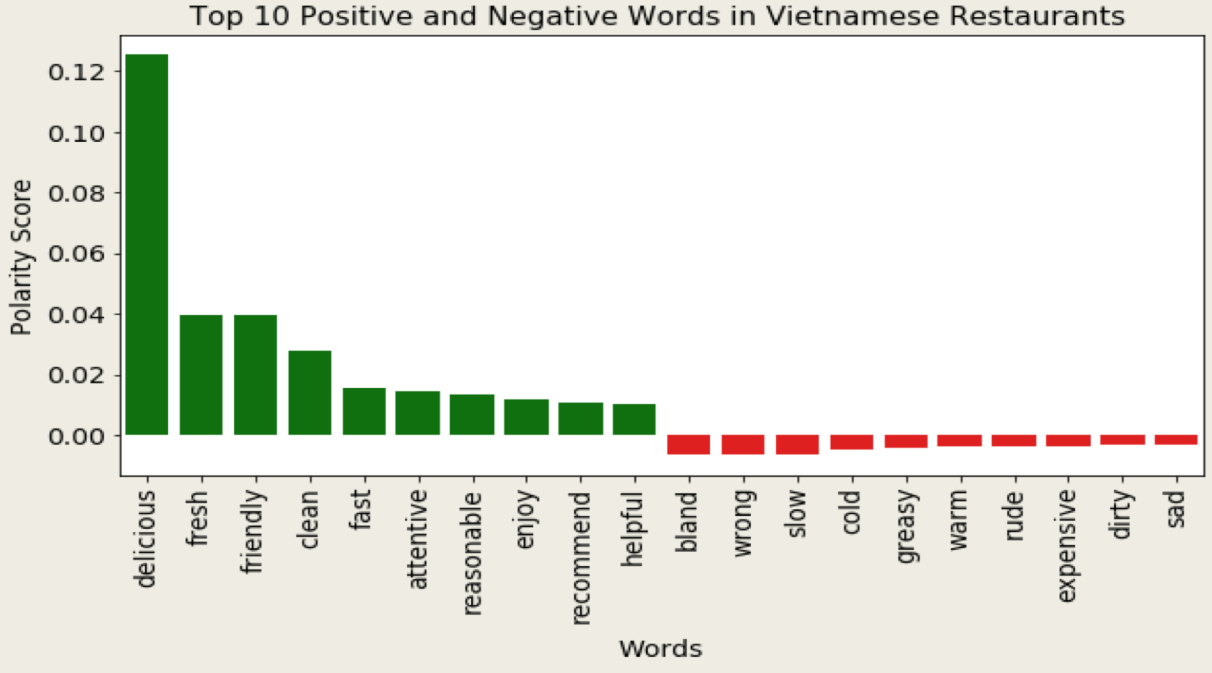

Top 10 Positive and Negative Words in Italian Restaurants
Polarity Score
0.12
0.10
0.08
0.06
0.04
0.02
0.00
delicious
fresh
friendly
incredible
die
recommend
phenomenal
outstanding
heaven
authentic
cold
warm
bland
wrong
expensive
slow
hard
lemon
mediocre
poor
Words


Top 10 Positive and Negative Words in Japanese Restaurants
Polarity Score
0.10
0.08
0.06
0.04
0.02
0.00
delicious
friendly
recommend
fresh
reasonable
fun
pleasant
enjoy
pleasantly
variety
top
slow
expensive
bland
cold
wrong
overpriced
hard
poor
mediocre
Words


Top 10 Positive and Negative Words in French Restaurants
Polarity Score
0.12
0.10
0.08
0.06
0.04
0.02
0.00
delicious
sweet
perfection
recommend
impeccable
fresh
incredible
heaven
tender
die
cold
expensive
wrong
bland
seasoned
slow
hard
disappointment
amuse
mediocre
Words


Top 10 Positive and Negative Words in Korean Restaurants
Polarity Score
0.10
0.08
0.06
0.04
0.02
0.00
delicious
friendly
reasonable
attentive
tender
recommend
generous
fresh
cute
variety
bland
work
slow
warm
cold
expensive
mediocre
sweet
fried
lack
Words


Top 10 Positive and Negative Words in Thai Restaurants
Polarity Score
0.12
0.10
0.08
0.06
0.04
0.02
0.00
delicious
friendly
fresh
recommend
reasonable
perfection
fast
hot
bomb
outstanding
sweet
bland
slow
cold
wrong
sour
expensive
mediocre
greasy
weird
Words


Top 10 Positive and Negative Words in Vietnamese Restaurants
Polarity Score
0.12
0.10
0.08
0.06
0.04
0.02
0.00
delicious
fresh
friendly
clean
fast
attentive
reasonable
enjoy
recommend
helpful
bland
wrong
slow
cold
greasy
warm
rude
expensive
dirty
sad
Words

cuisines										
Chinese	delicious	friendly	fresh	authentic	hot	recommend	reasonable	fast	tender	clean
French	delicious	sweet	perfection	recommend	impeccable	fresh	incredible	heaven	tender	die
Italian	delicious	fresh	friendly	incredible	die	recommend	phenomenal	outstanding	heaven	authentic
Japanese	delicious	friendly	recommend	fresh	reasonable	fun	pleasant	enjoy	pleasantly	variety
Korean	delicious	friendly	reasonable	attentive	tender	recommend	generous	fresh	cute	variety
Thai	delicious	friendly	fresh	recommend	reasonable	perfection	fast	hot	bomb	outstanding
Vietnamese	delicious	fresh	friendly	clean	fast	attentive	reasonable	enjoy	recommend	helpful

cuisines										
Chinese	sour	bland	greasy	cold	hard	expensive	wrong	rude	warm	weird
French	cold	expensive	wrong	bland	seasoned	slow	hard	disappointment	amuse	mediocre
Italian	cold	warm	bland	wrong	expensive	slow	hard	lemon	mediocre	poor
Japanese	top	slow	expensive	bland	cold	wrong	overpriced	hard	poor	mediocre
Korean	bland	work	slow	warm	cold	expensive	mediocre	sweet	fried	lack
Thai	sweet	bland	slow	cold	wrong	sour	expensive	mediocre	greasy	weird
Vietnamese	bland	wrong	slow	cold	greasy	warm	rude	expensive	dirty	sad

OBSERVATIONS

-Delicious, friendly, fresh

- Friendly before reasonable

- French sweet as positive, Korean sweet as negative

- Fun : Japanese food

- Variety: Japanese/Korean

-Bland/cold/expensive: Major issue (7 out of 7 cuisines) followed by slow(6 out of 7)

Example Recommendations

Cuisines	Recommendations
Chinese	• Increase flavor • Control sour and greasiness • Monitor food temperature • Monitor price & improve service
French	• Monitor food temperature • Increase flavor • Monitor price & improve service
Japanese	• Monitor food temperature • Increase flavor • Monitor price & improve service
Italian	• Monitor food temperature • Increase flavor • Monitor price & improve service • Control use of lemon in foods

Cuisines	Recommendations
Korean	• Monitor food temperature • Increase flavor • Monitor price & improve service • Too sweet or fried foods are not too popular
Thai	• Monitor food temperature • Increase flavor • watch out on price • Control sweetness & greasiness • Improve service
Vietnamese	• Increase flavor • Control greasiness • Monitor price • Improve service and cleanliness

Conclusion

- ★ The 1st objective of this project was to create various classification models to predict star ratings of restaurants
- ★ After hyperparameter tuning, the Random forest model emerged as the best model with about 68% accuracy
- ★ The 2nd objective of this project was to conduct the review sentiment analysis and figure out the top 10 positive words and top 10 negative words for 7 different cuisines
- ★ Appropriate recommendations are made to the restaurants based on the top positive and negative words

Limitations & Future Research

- We only created 5 different models for this project to predict star ratings of restaurants. In the future, we can generate more models or encoding methods to improve our results
- We set a threshold of 70%. If the NaN values were 70% or more in the attributes, we simply deleted them. We can try setting up different thresholds to see other possible changes
- We can fine tune more hyperparameters in the future
- We can apply similar sentiment analysis techniques for reviews in other areas like movie, social media posts, etc

References

Cao, Y., Yu, B., Zhang, Y., & Zhou, J. (n.d.). Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews. Retrieved October 15, 2019, from https://arxiv.org/pdf/1709.08698.pdf

Hu, M., & Liu, B. (2004, May 15). Opinion Mining, Sentiment Analysis, and Opinion Spam Detection. Retrieved October 28, 2019, from https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.

Kong, A., Nyugen, V., & Xu, K. (n.d.). Predicting International Restaurant Success with Yelp. Retrieved September 20, 2019, from http://cs229.stanford.edu/proj2016spr/report/062.pdf

Yelp Open Dataset. (n.d.). Retrieved September 15, 2019, from https://www.yelp.com/dataset.