Predicting Restaurant Rating Star

Presentation by **Hoda Shoghi**

Project Objective

Goal:

Develop a model to predict restaurant star ratings.

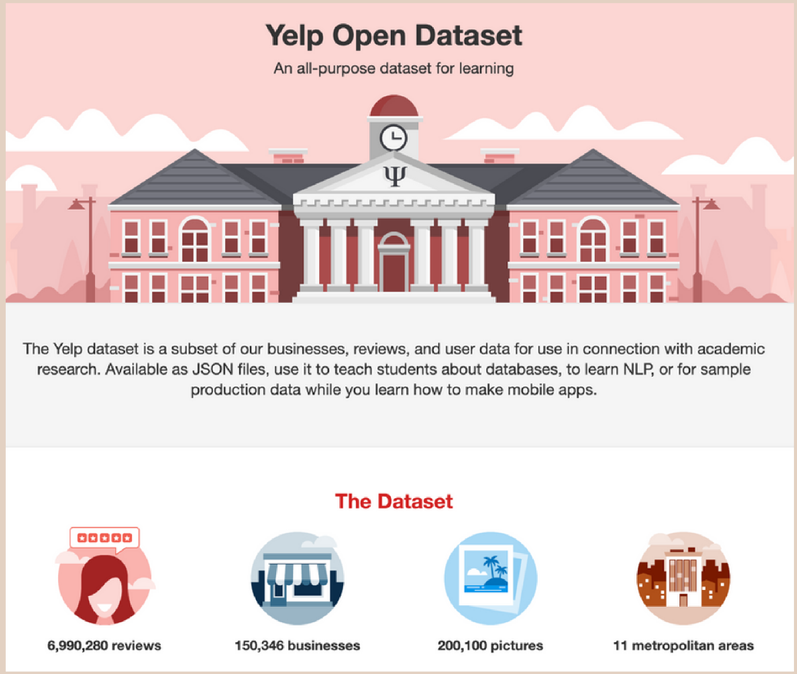
Why?:

To understand what influences ratings and help future restaurant owners succeed.

Purpose:

Offer valuable insights to potential restaurant owners and stakeholders about key factors that can influence a restaurant's success.

Dataset OverView



Source:	Yelp's Public Dataset
Size:	150346 Businesses
Characteristics	Examples: Price Range, Cuisine Type, Location,Casual Dining, Ambience, etc.
Target Response:	Restaurant's Star Rating (1 to 5 stars)

Data Pre-Processing

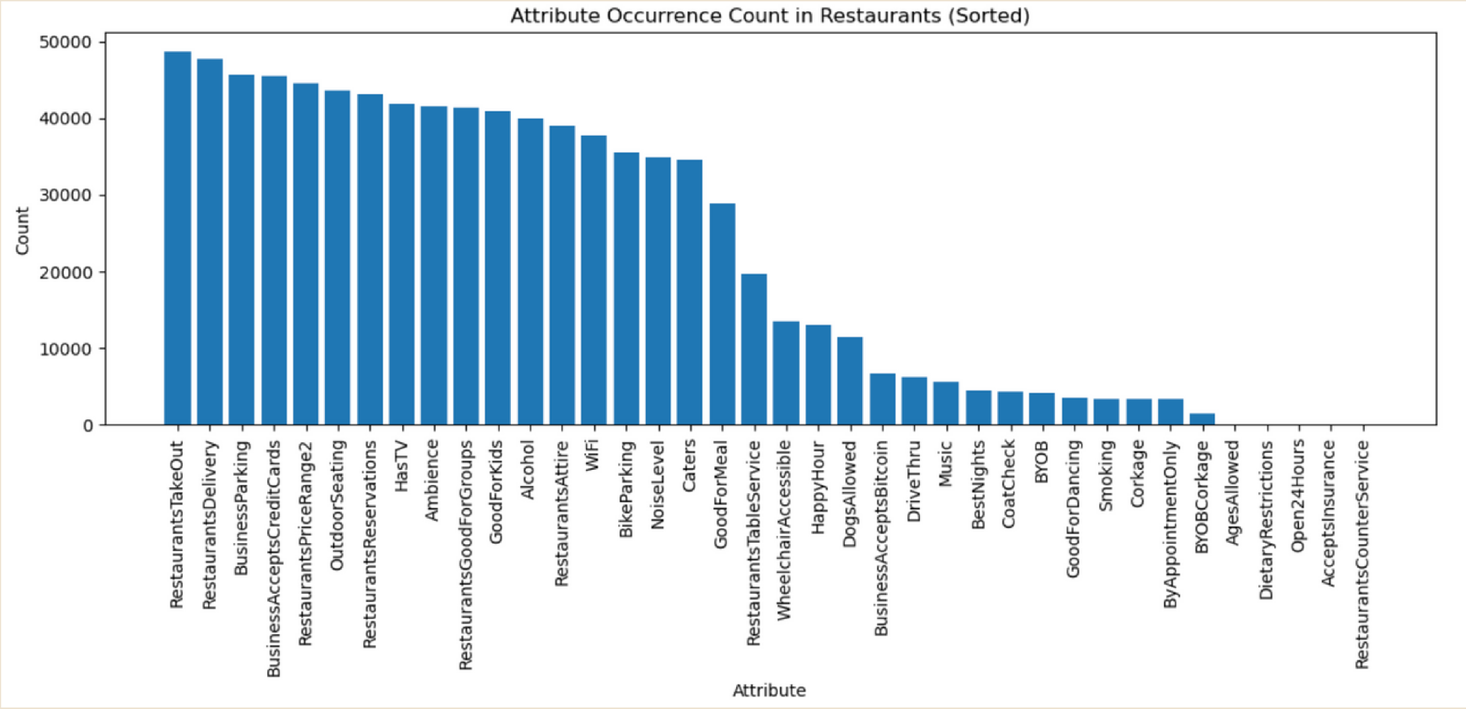
Getting the foundation right!!

- **Handling Missing Values**:
- **Feature Engineering**:
- **Data Transformation**:

Data Cleaning and Pre-processing

attributes	categories
{'RestaurantsDelivery': 'False', 'OutdoorSeati...	restaurants, food, bubble tea, coffee & tea, b...
{'BusinessParking': 'None', 'BusinessAcceptsCr...	burgers, fast food, sandwiches, food, ice crea...
{'Caters': 'True', 'Alcohol': 'u'full_bar'', '...	pubs, restaurants, italian, bars, american (tr...
{'RestaurantsAttire': ''casual'', 'Restaurants...	ice cream & frozen yogurt, fast food, burgers,...
{'Alcohol': ''none'', 'OutdoorSeating': 'None'...	vietnamese, food, restaurants, food trucks

- Remove unrellevent data
- Impute some missing data
- Explode & unpack data
- One hot encoding data
- Combine data
- Optimizing data
- Removing the low count
- Scaling.



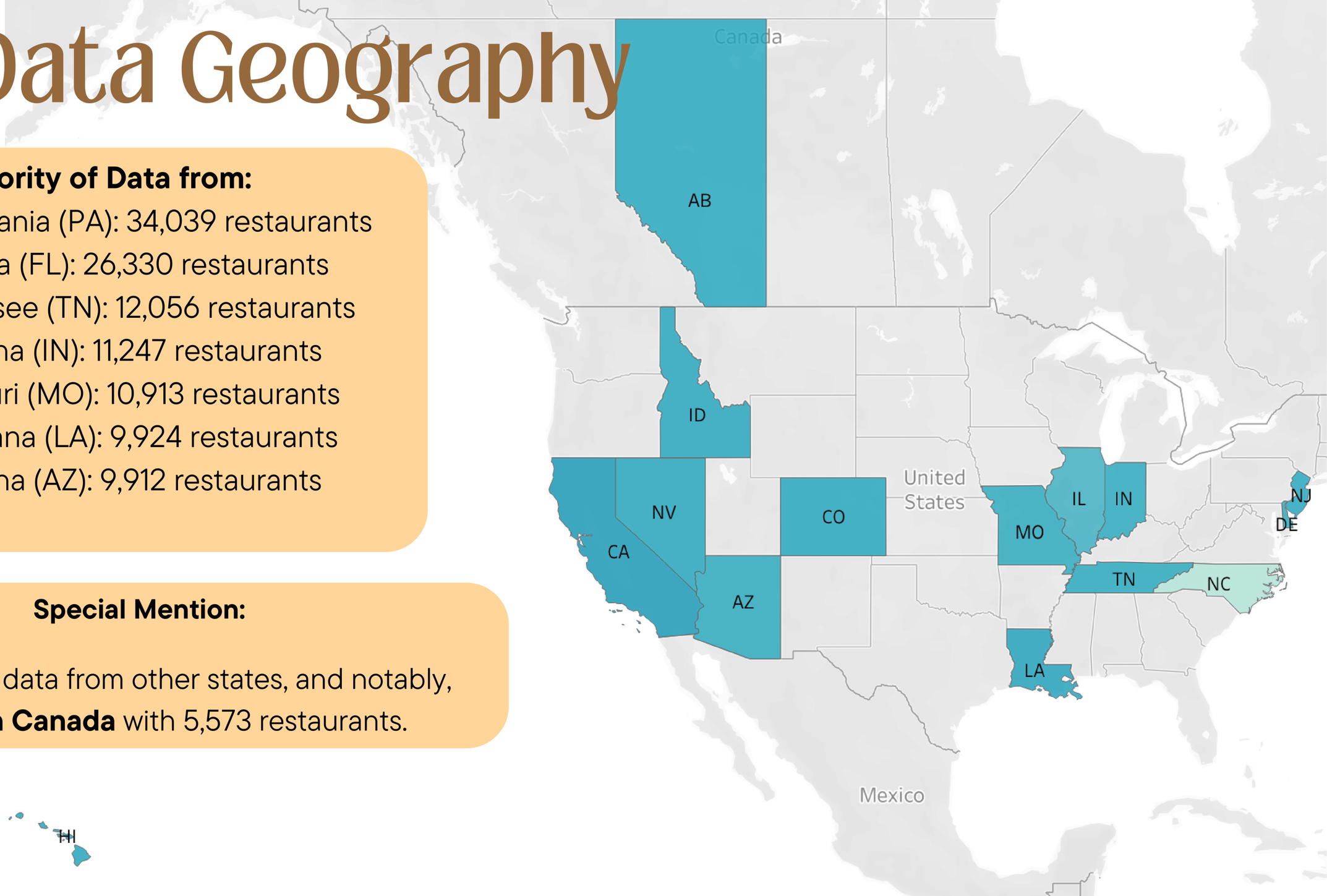
EDA-Data Geography

Majority of Data from:

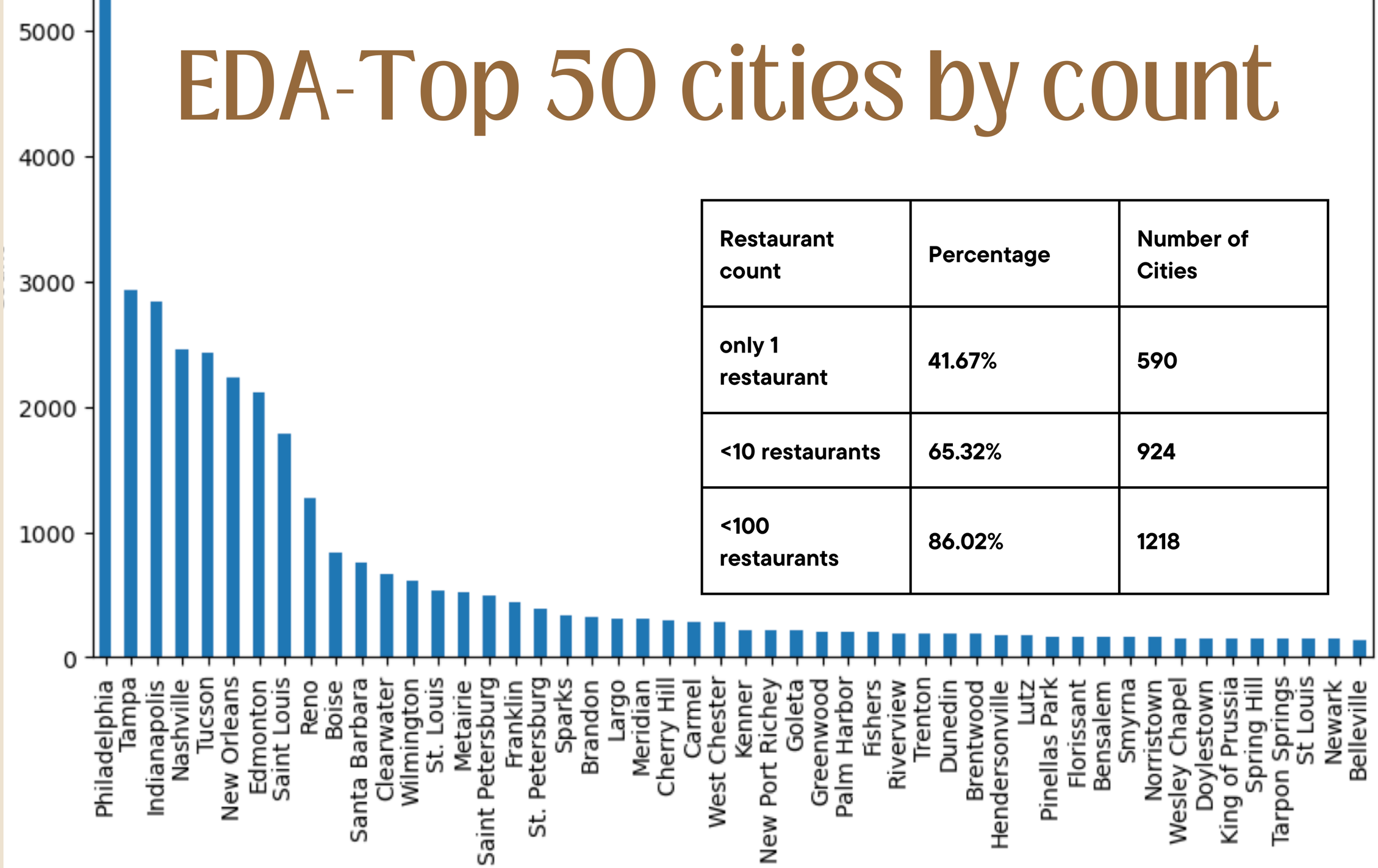
- Pennsylvania (PA): 34,039 restaurants
- Florida (FL): 26,330 restaurants
- Tennessee (TN): 12,056 restaurants
- Indiana (IN): 11,247 restaurants
- Missouri (MO): 10,913 restaurants
- Louisiana (LA): 9,924 restaurants
- Arizona (AZ): 9,912 restaurants

Special Mention:

- We also have data from other states, and notably, **Alberta in Canada** with 5,573 restaurants.

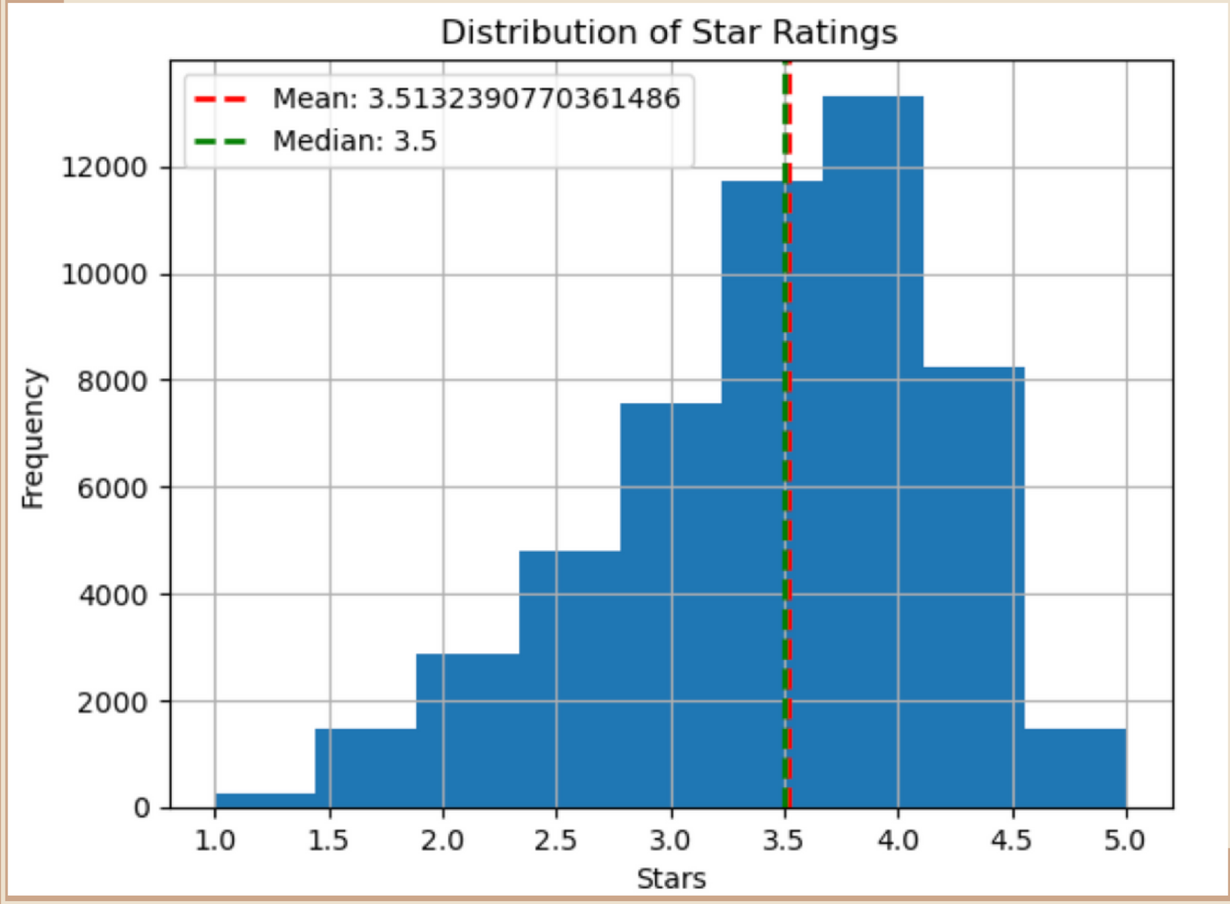


EDA-Top 50 cities by count

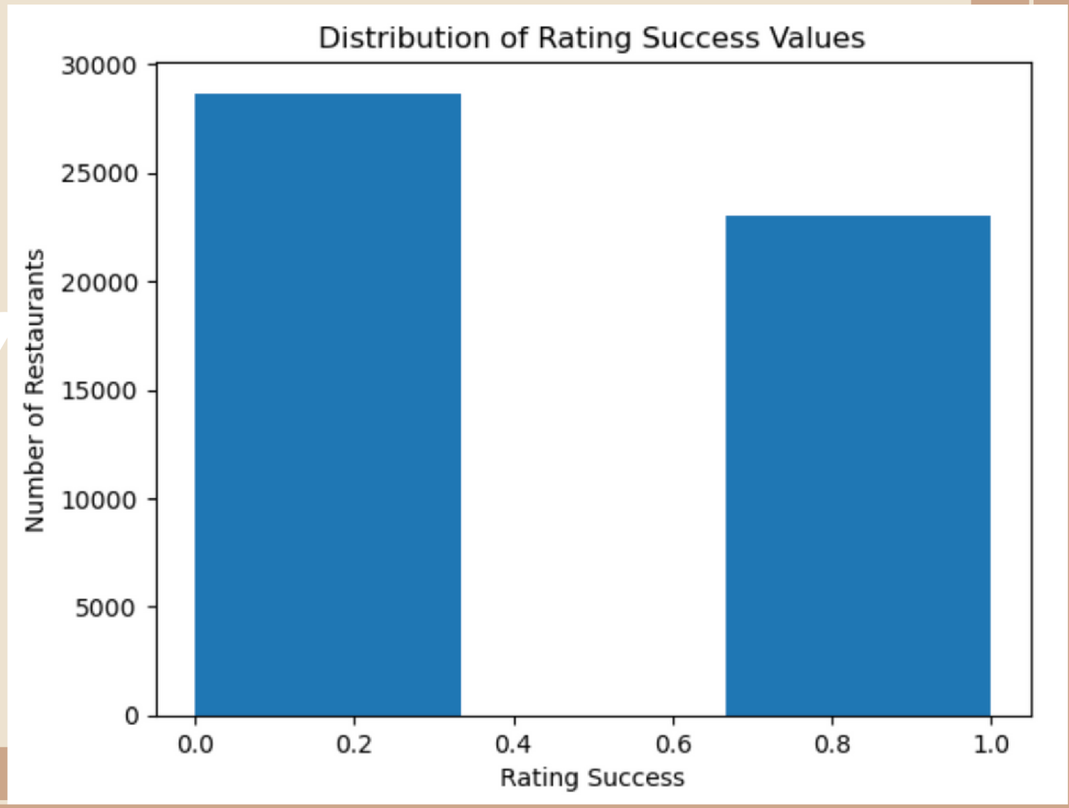


Restaurant count	Percentage	Number of Cities
only 1 restaurant	41.67%	590
<10 restaurants	65.32%	924
<100 restaurants	86.02%	1218

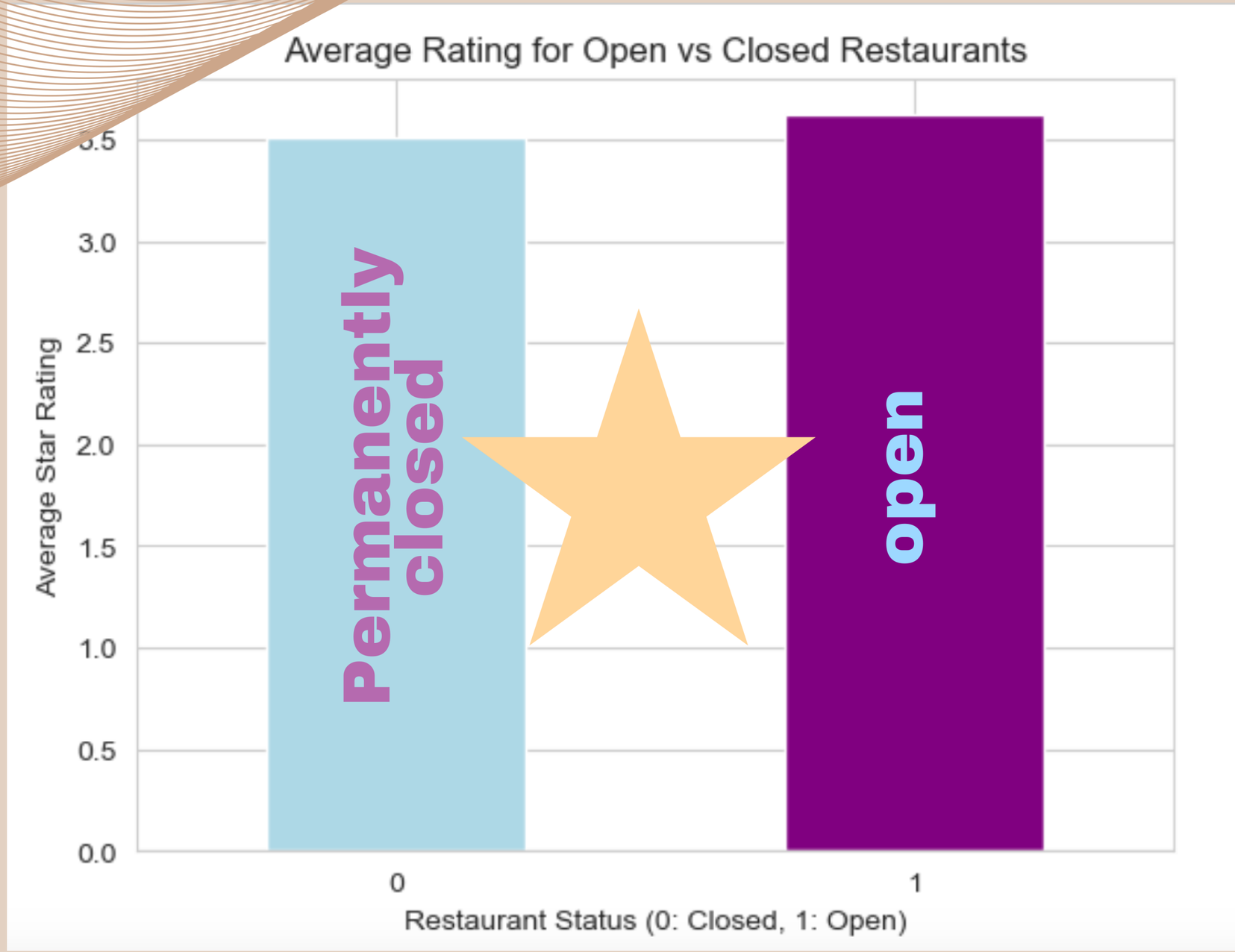
EDA -Target Variable



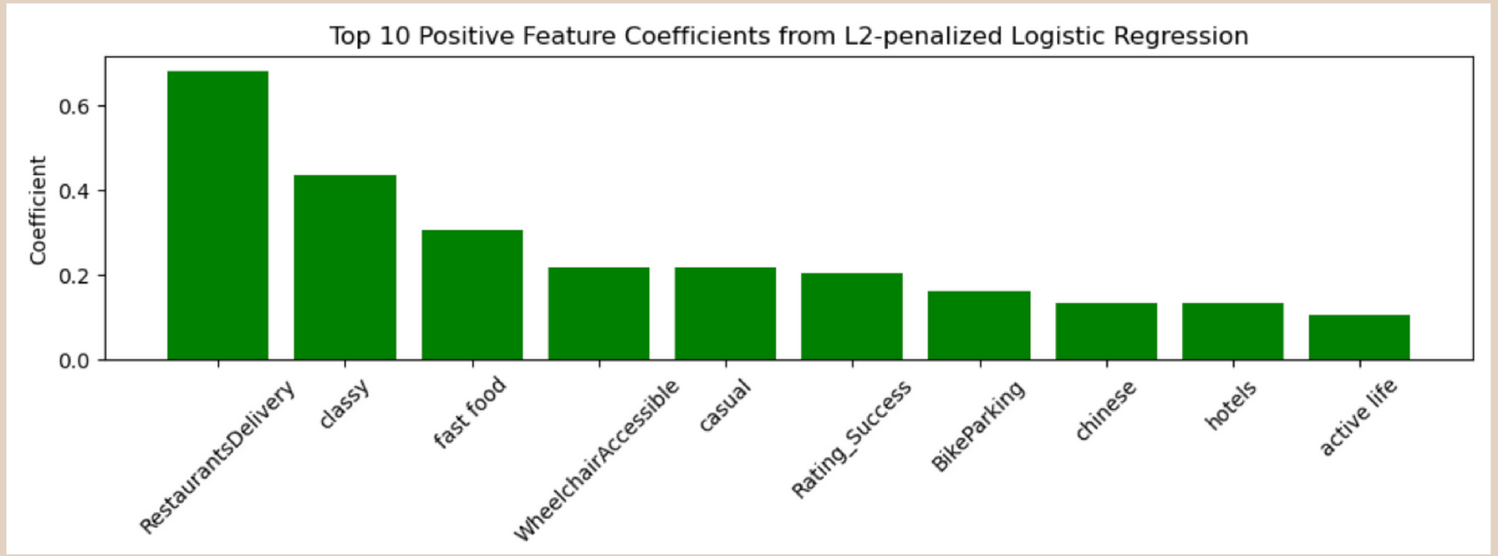
converted to binary target variable



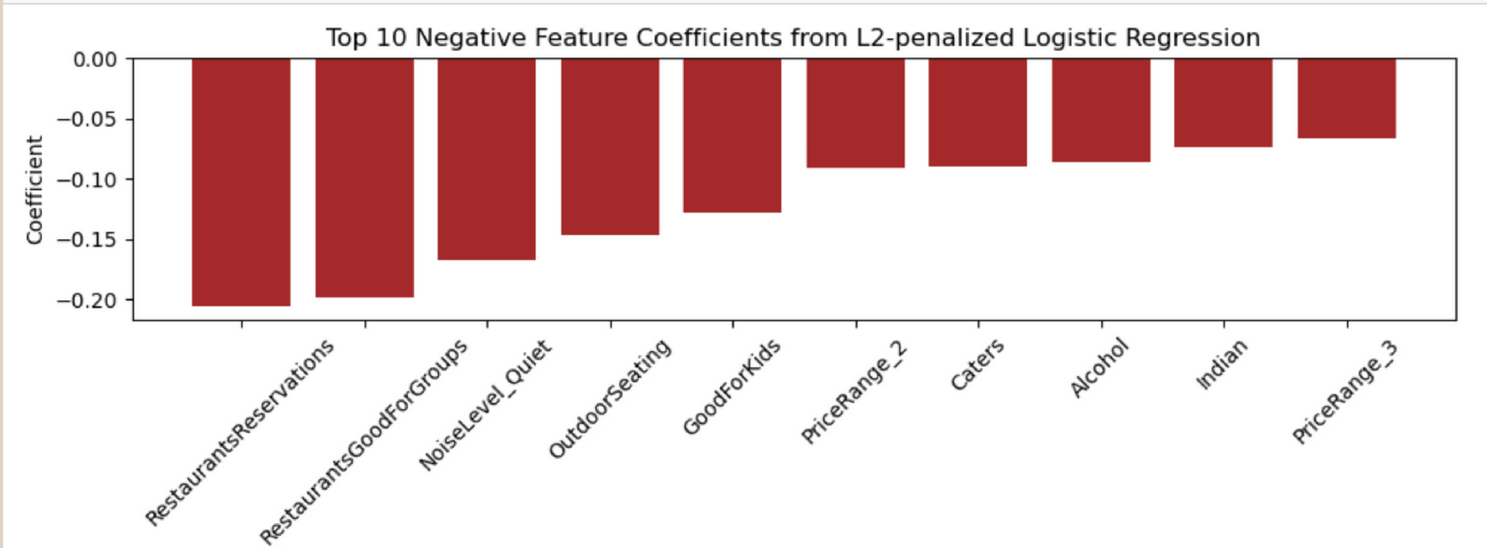
Permanently Closed!?


Average Rating for Open vs Closed Restaurants
Average Star Rating
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Permanently closed
open
0
1
Restaurant Status (0: Closed, 1: Open)

Determinants of Closure



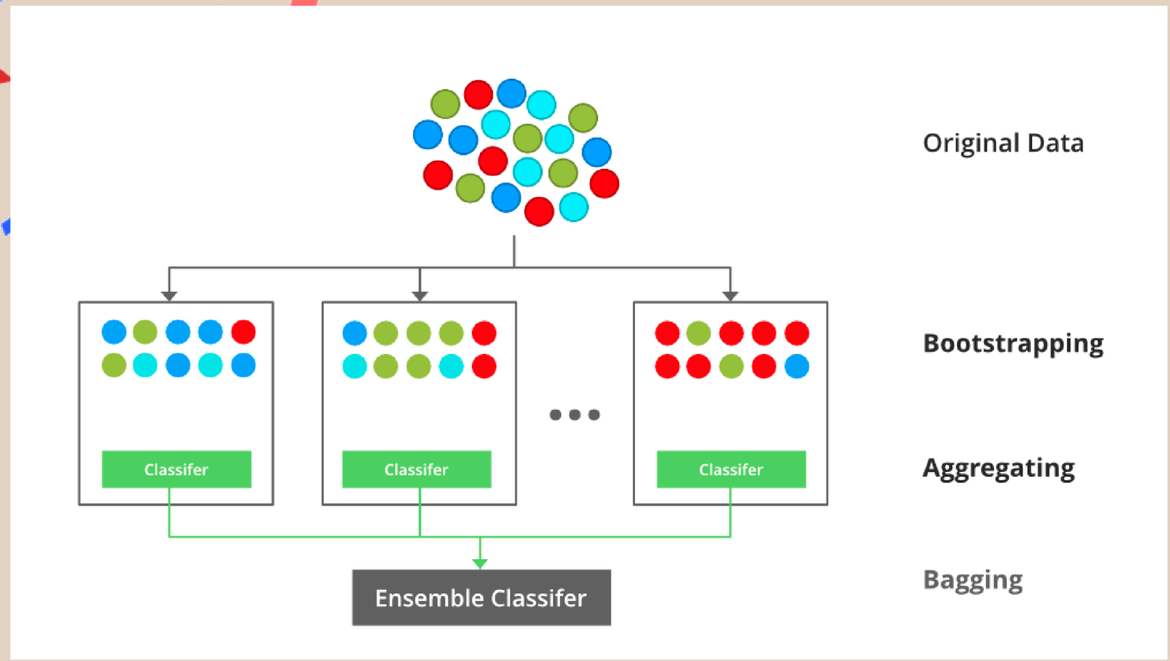
Logistic Regression Model Performance: **Train Accuracy: 74.83%** **Test Accuracy: 74.80%**



Model	Parameters	Accuracy	Precision Low Rating	Recall Low Rating	F1-score Low Rating	Precision High Rating	Recall High Rating	F1-score High Rating
Logistic Regression	penalty='l2', solver='saga', C=100 ,test_size=0.30, random_state=42	0.70	0.71	0.77	0.74	0.68	0.60	0.64
Random Forest	bootstrap=False, max_depth=17, min_samples_leaf=2, min_samples_split=10, n_estimators=100	0.70	0.70	0.80	0.75	0.70	0.58	0.63
Random Forest Refined Model	feature Importance , Refined Model with Feature reduction	0.69	0.71	0.76	0.73	0.67	0.60	0.63
Gradient Boosting	n_estimators=250, random_state=42	0.70	0.71	0.78	0.74	0.68	0.61	0.64
PipeLine	feature extraction (PCA)	0.71	0.70	0.76	0.73	0.70	0.60	0.63
XGBoost	default parameters	0.71	0.73	0.77	0.75	0.69	0.64	0.67
PipeLine	classifier_colsample_bytree': 0.8, 'classifier_gamma': 0, 'classifier_learning_rate': 0.1, 'classifier_max_depth': 5, 'classifier_n_estimators': 250, 'classifier_subsample': 0.8	0.71	0.73	0.77	0.75	0.69	0.64	0.66

XGBoost

XGBoost is the winner



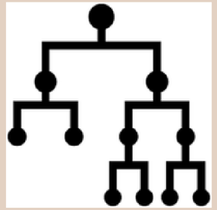
Classification Report

Metric	Class 0	Class 1
Precision	0.73	0.69
Recall	0.77	0.64
F1-Score	0.75	0.66
Accuracy	0.71	

Key Drivers of Restaurant Ratings

- Logistic Regression Coefficients

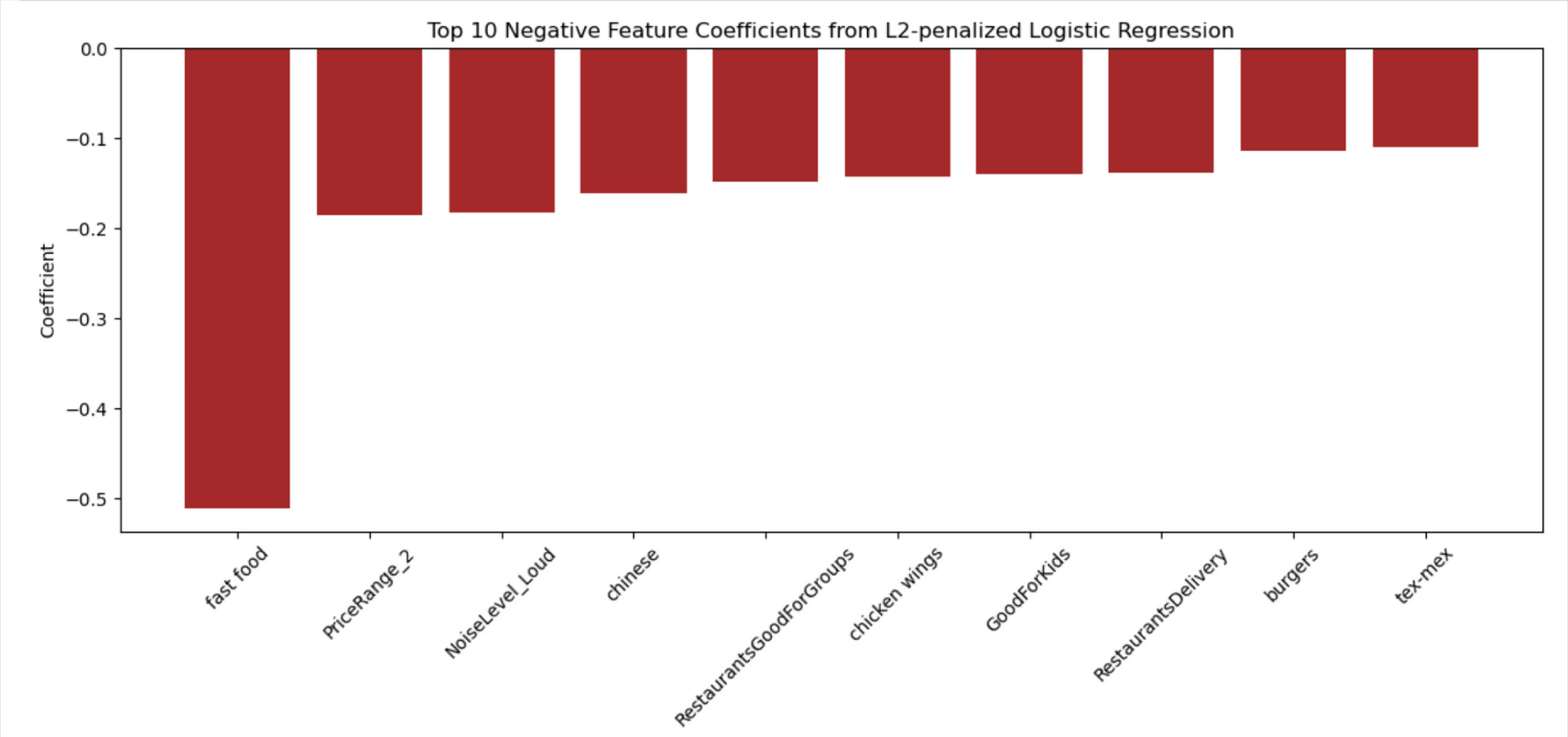
- Random Forest Feature Importance



- SHAP Analysis with XGBoost

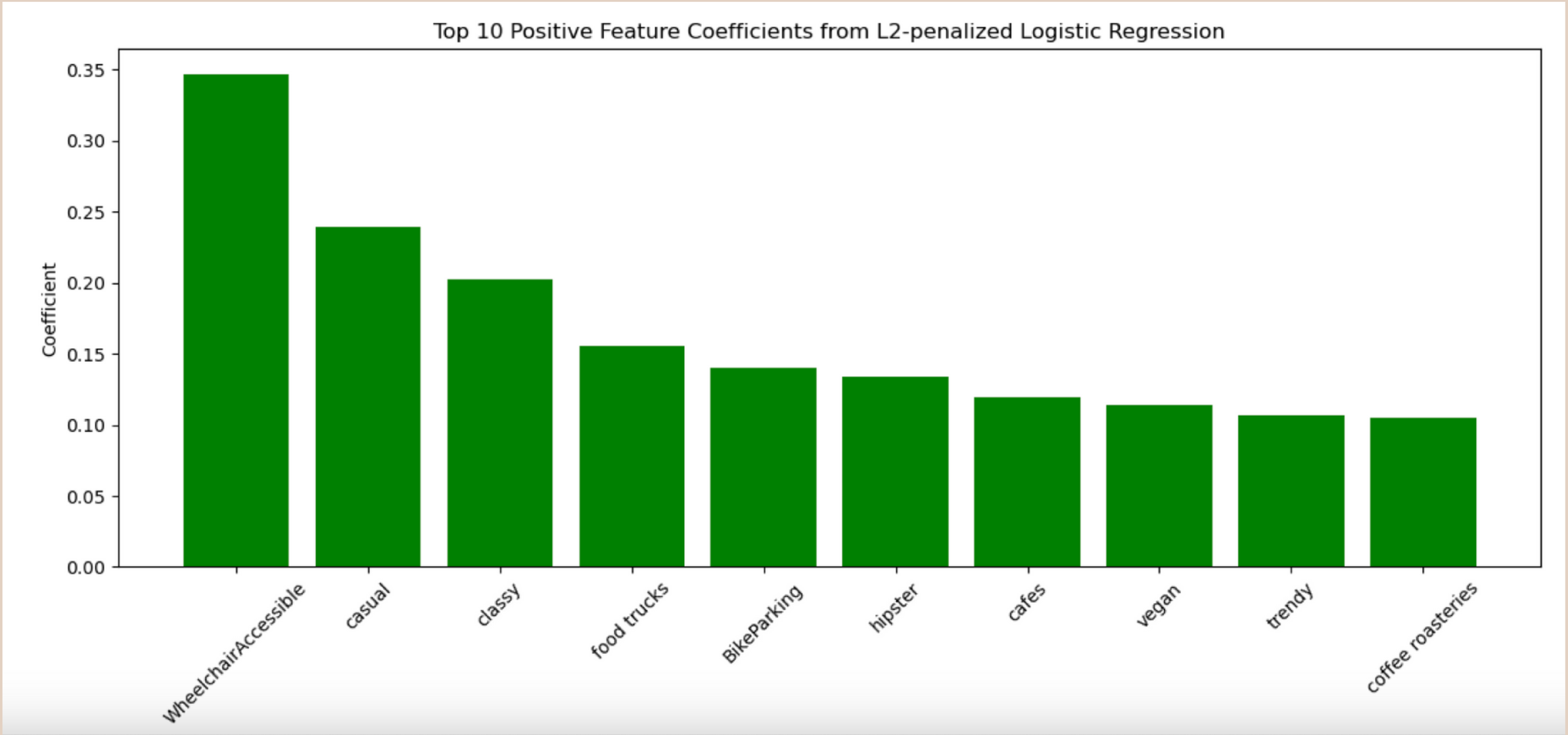
What Diners Might Not Prefer

Logistic Regression Negative Coefficients



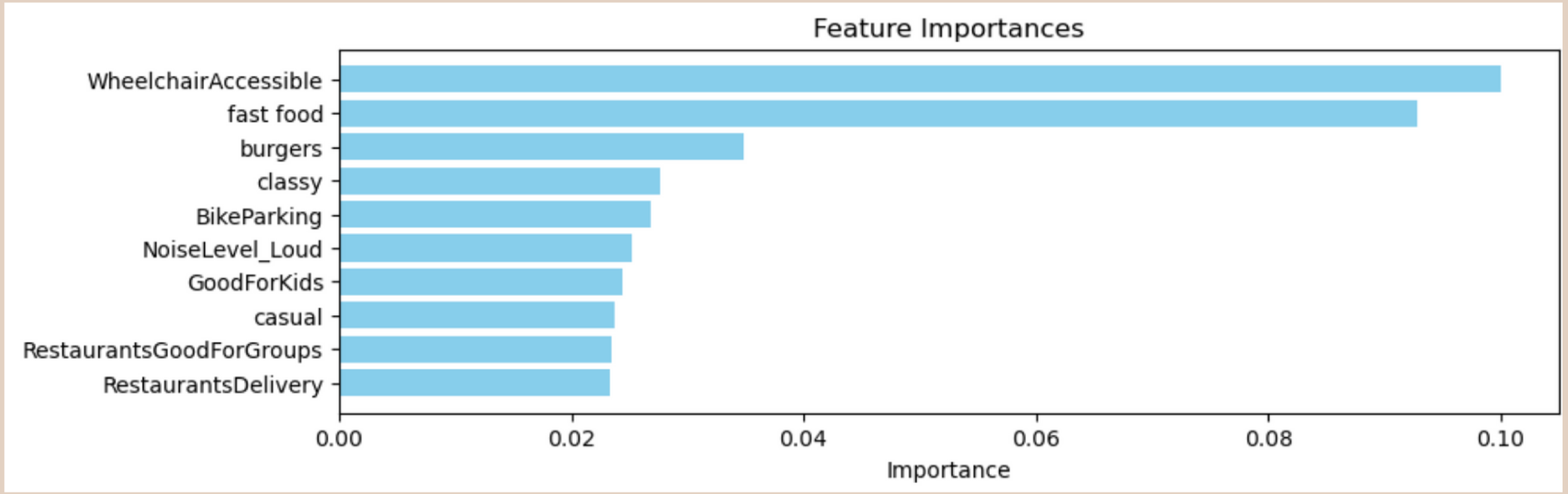
Factors That Boost Restaurant Ratings

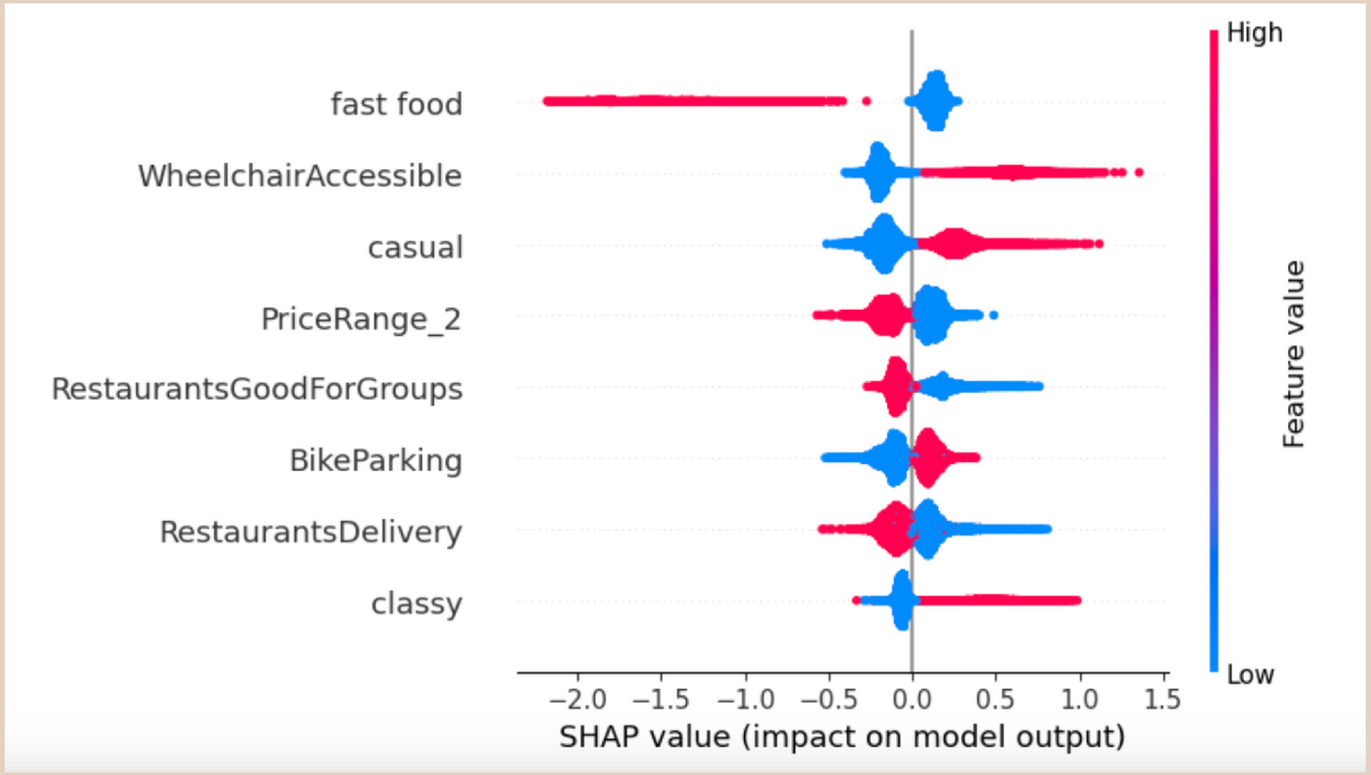
Logistic Regression Positive Coefficients

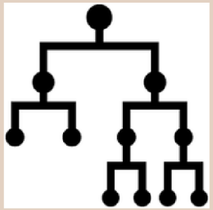


Random Forest

Random Forest
Feature Importance

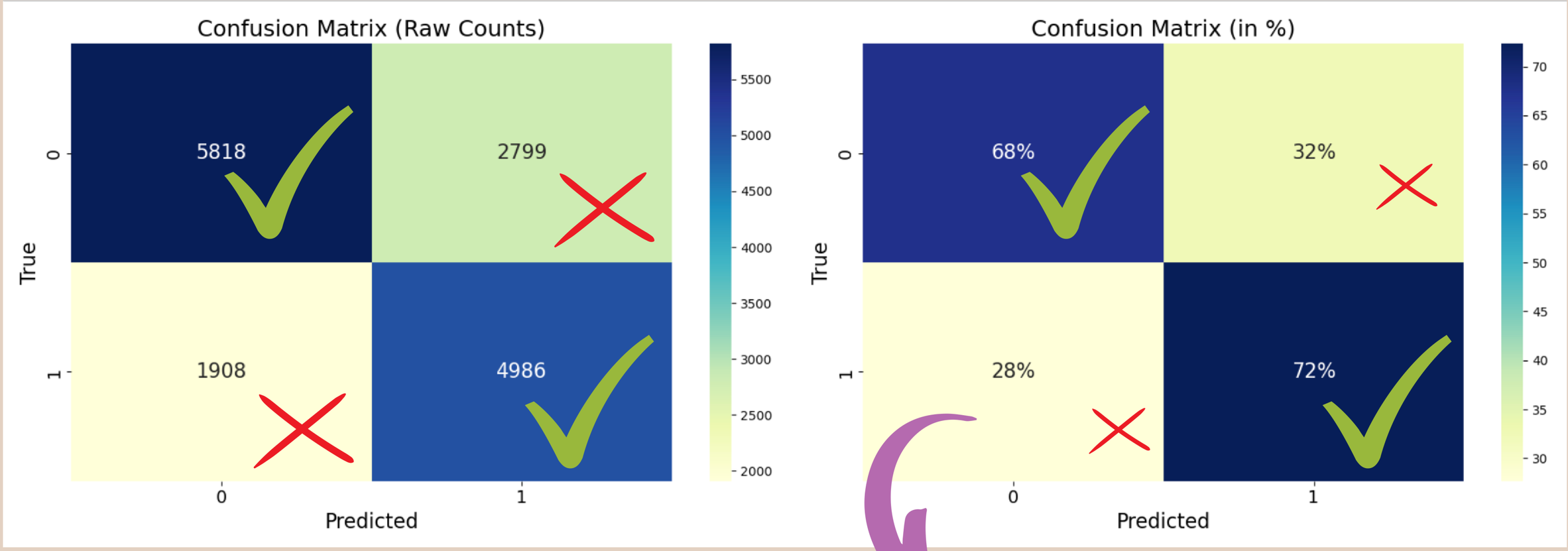






- **SHAP Analysis with XGBoost**

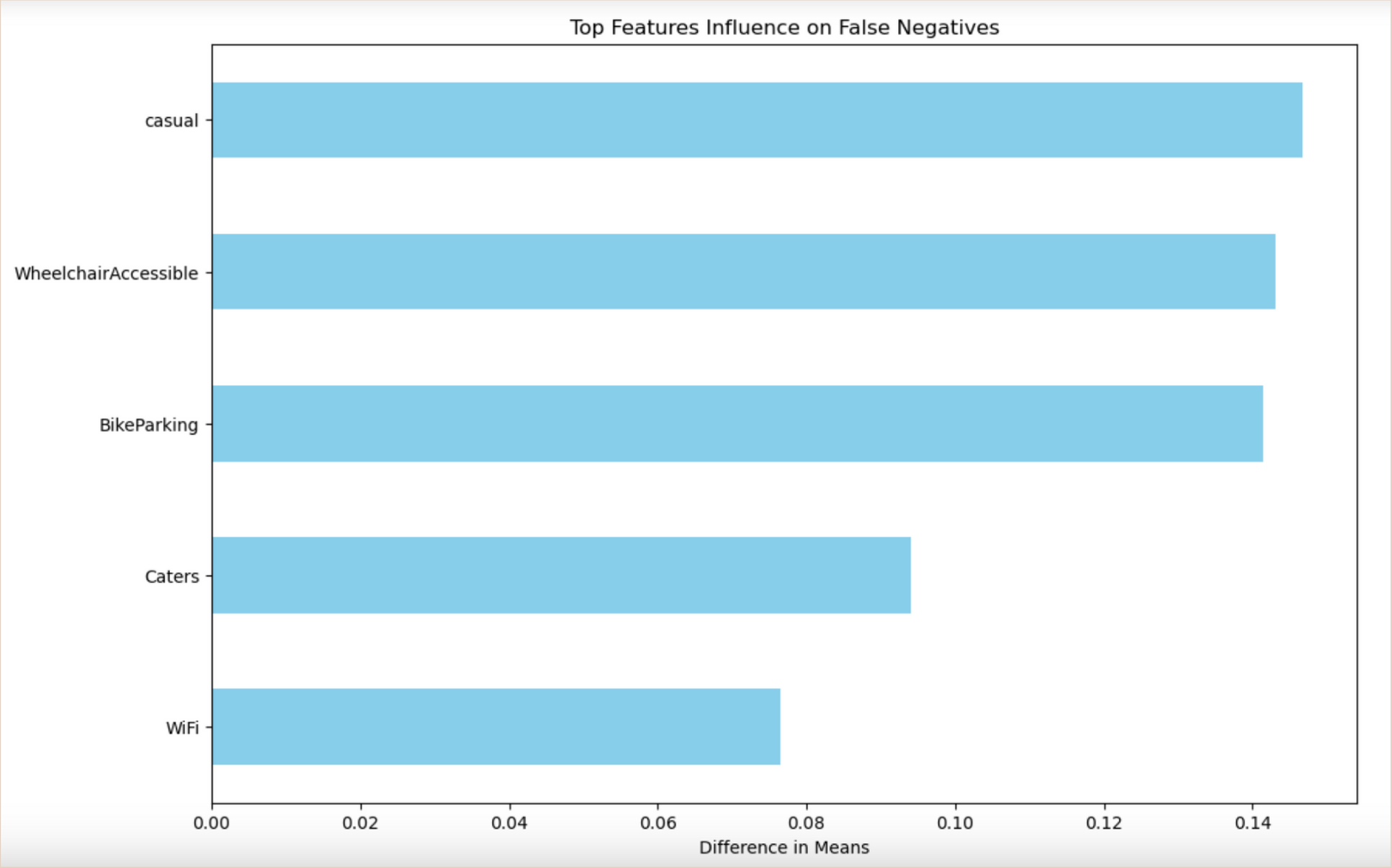
Evaluating Our Model's Predictions



WHERE PREDICTIONS AND REALITY DON'T MATCH

Model bias

Feature Mean Differences Analysis



Forging Ahead

Potential Improvements

Recommendations:

a. **Collect More Data:** Especially on under-represented features.

b. **Resampling:** Achieve balanced representation for wheelchair accessible venues.

c. **Feature Weight Adjustment:** Refine importance of standout features for better prediction.

Thank You