Predicting Restaurant Rating Star

Presentation by **Hoda Shoghi**

Project Objective

Goal:

Develop a model to predict restaurant star ratings.

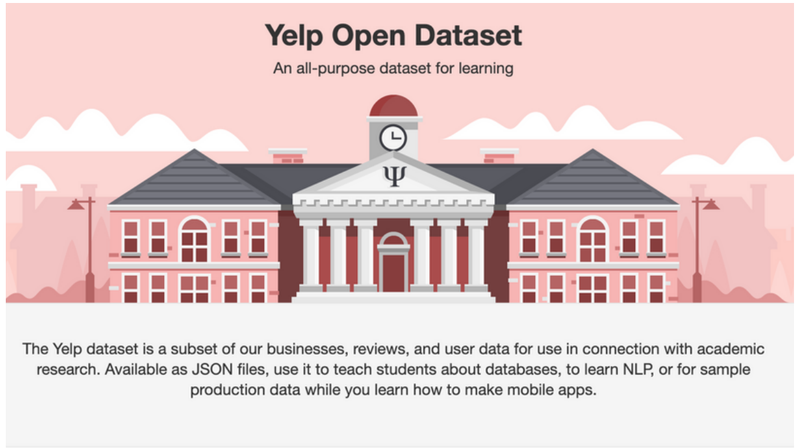
Why?:

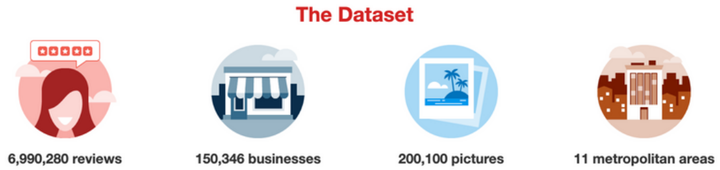
To understand what influences ratings and help future restaurant owners succeed.

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)

Pre-Processing

Handpick the best ingredients

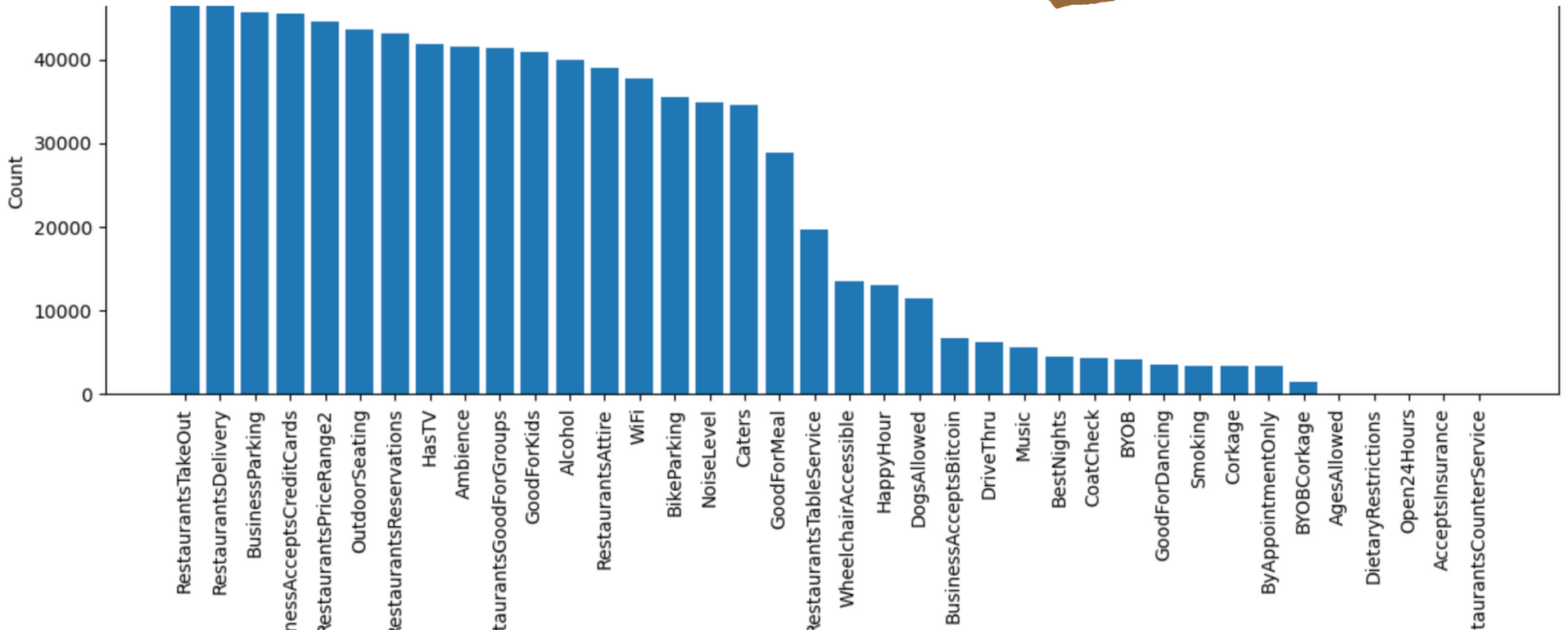
Refine them

Mix them in just the right PROPORTIONS

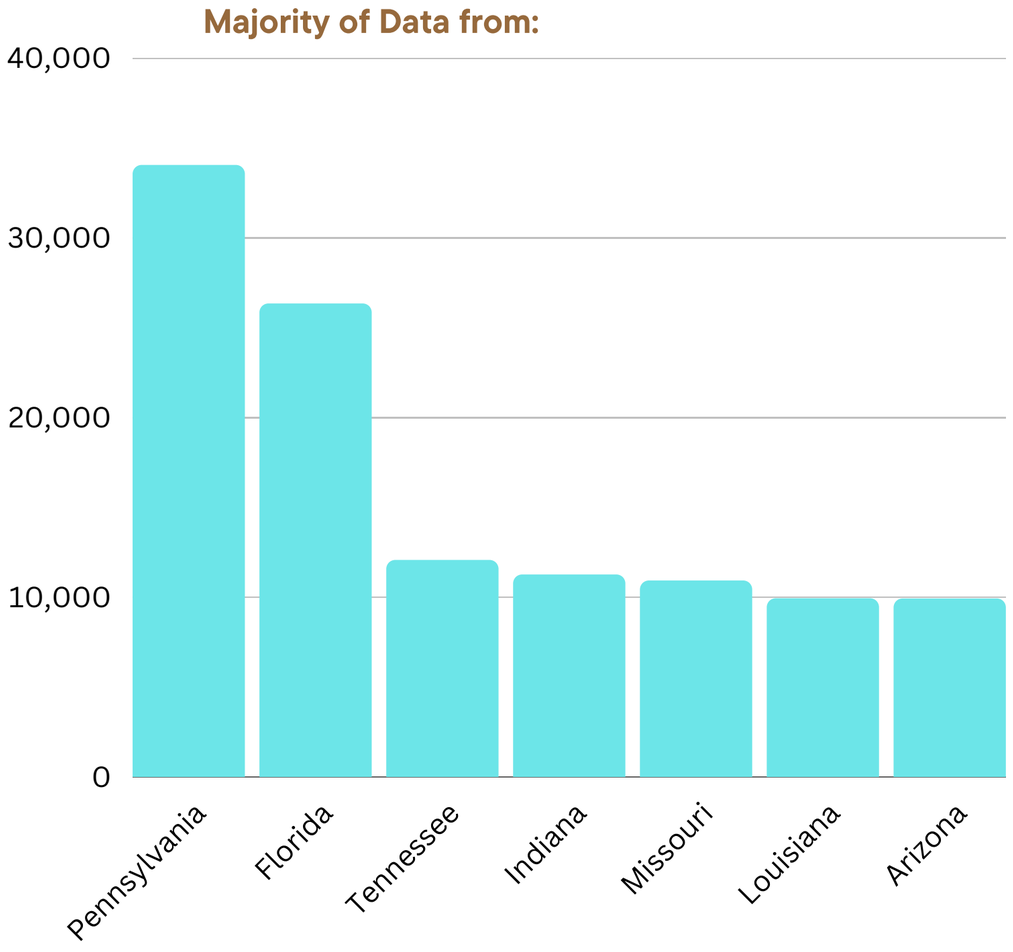
Getting the foundation right!!

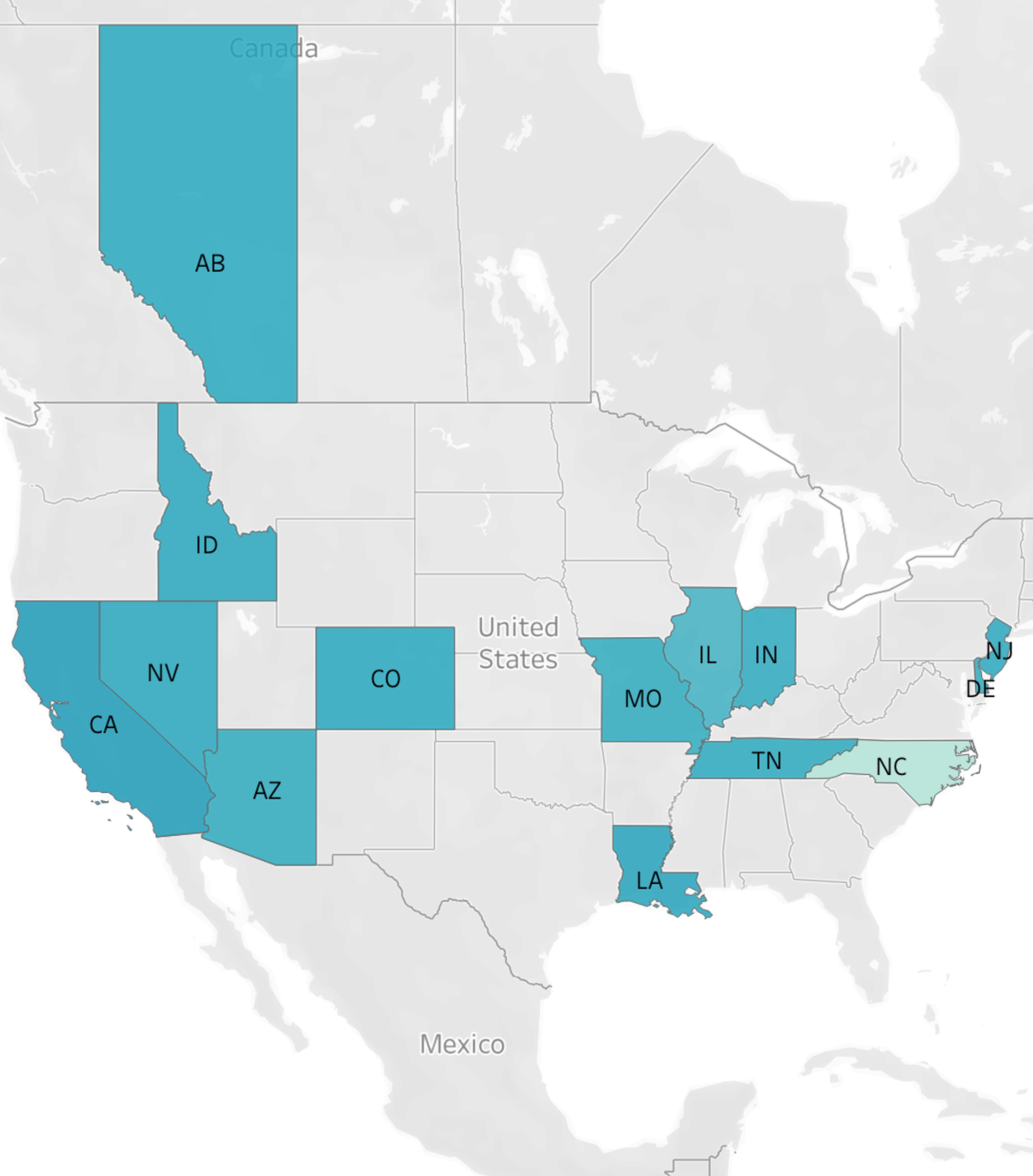
Data Cleaning and Pre-processing

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

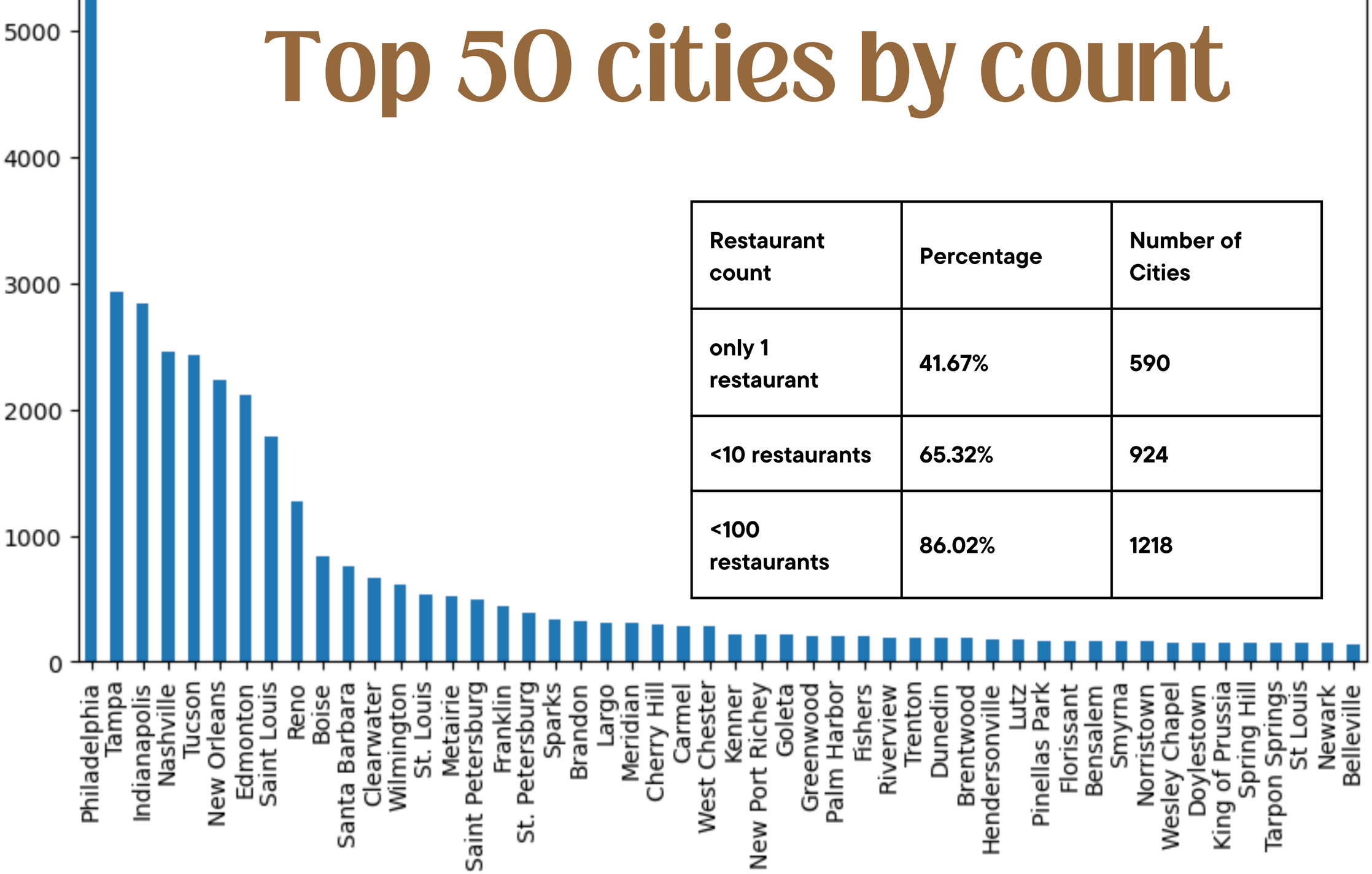


Data Geography



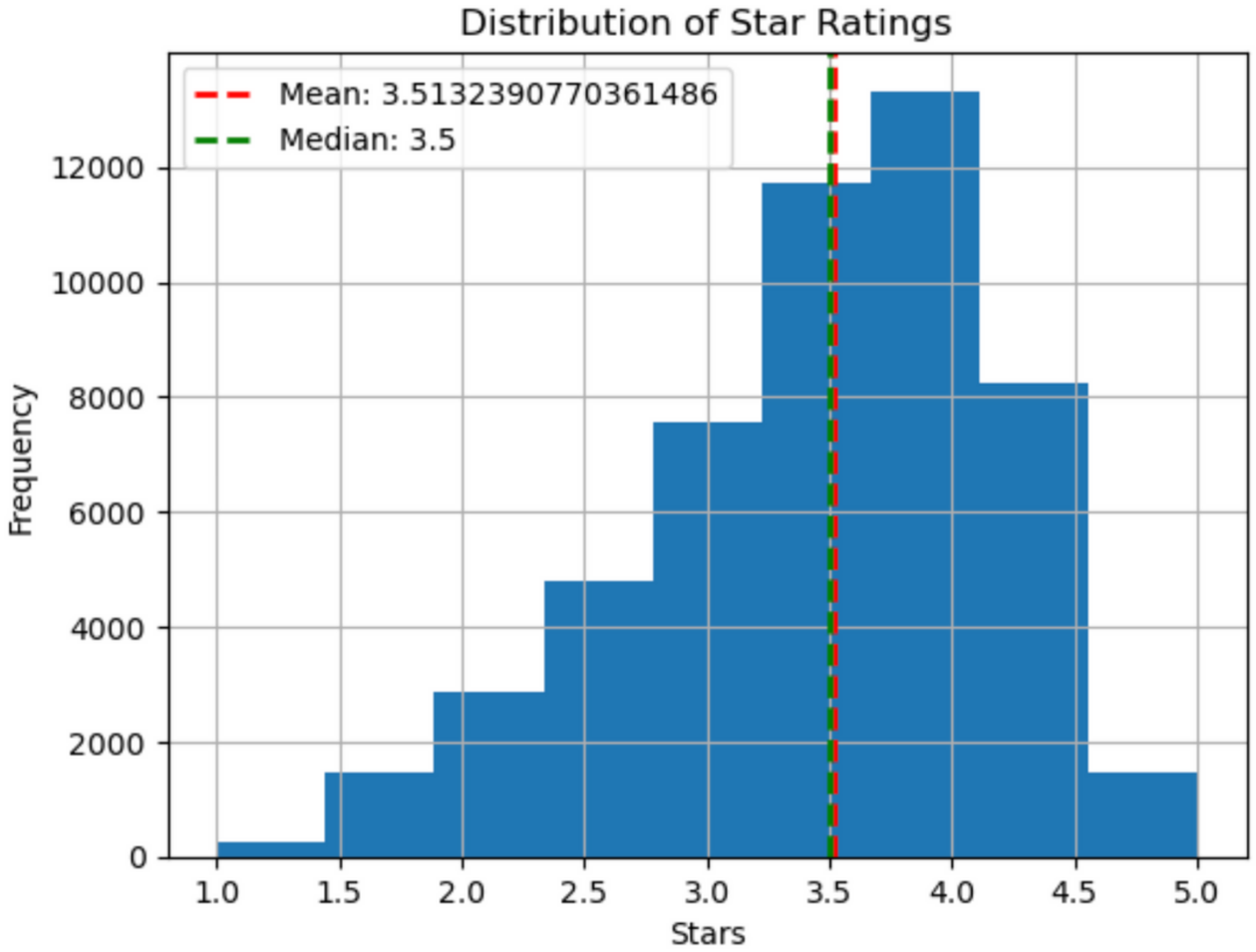


Top 50 cities by count

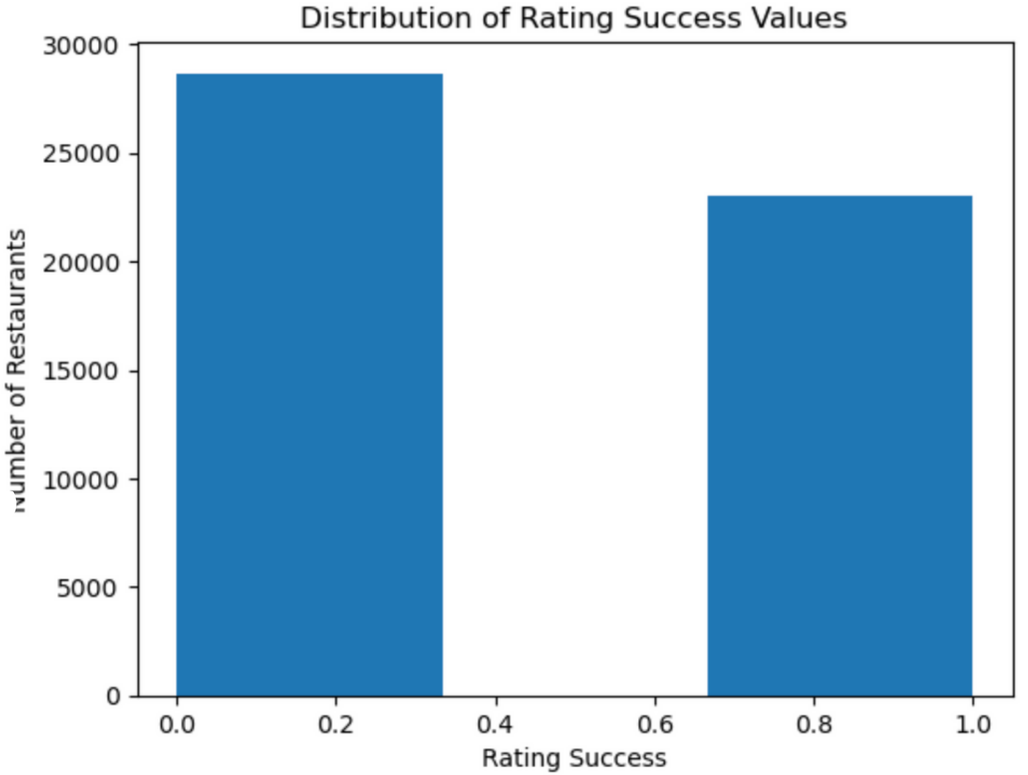


Restaurant count	Percentage	Number of Cities
only 1 restaurant	41.67%	590
<10 restaurants	65.32%	924
<100 restaurants	86.02%	1218

Target Response

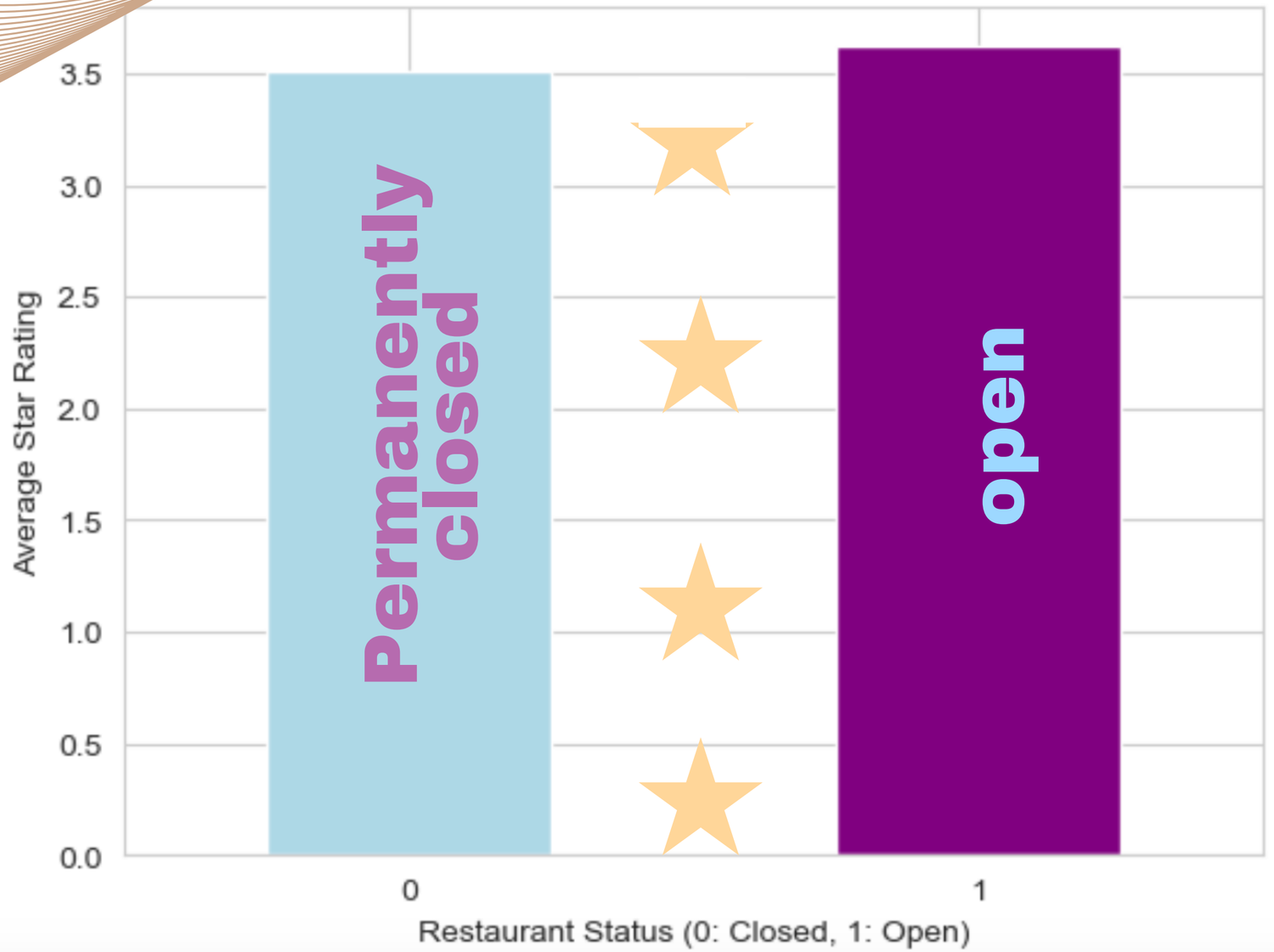


converted to binary target variable

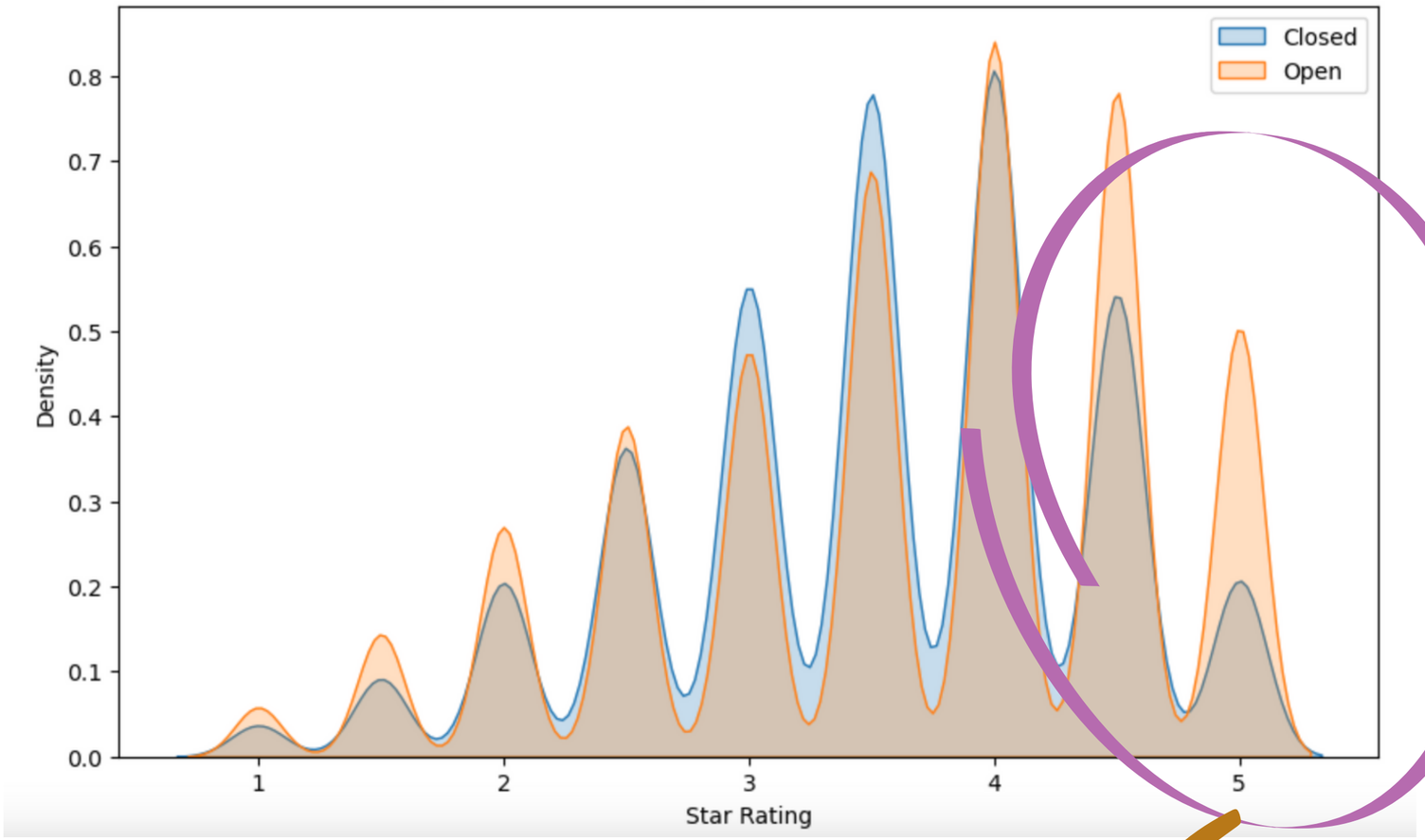


Permanently Closed!?

Average rating, Open Vs Cloesd



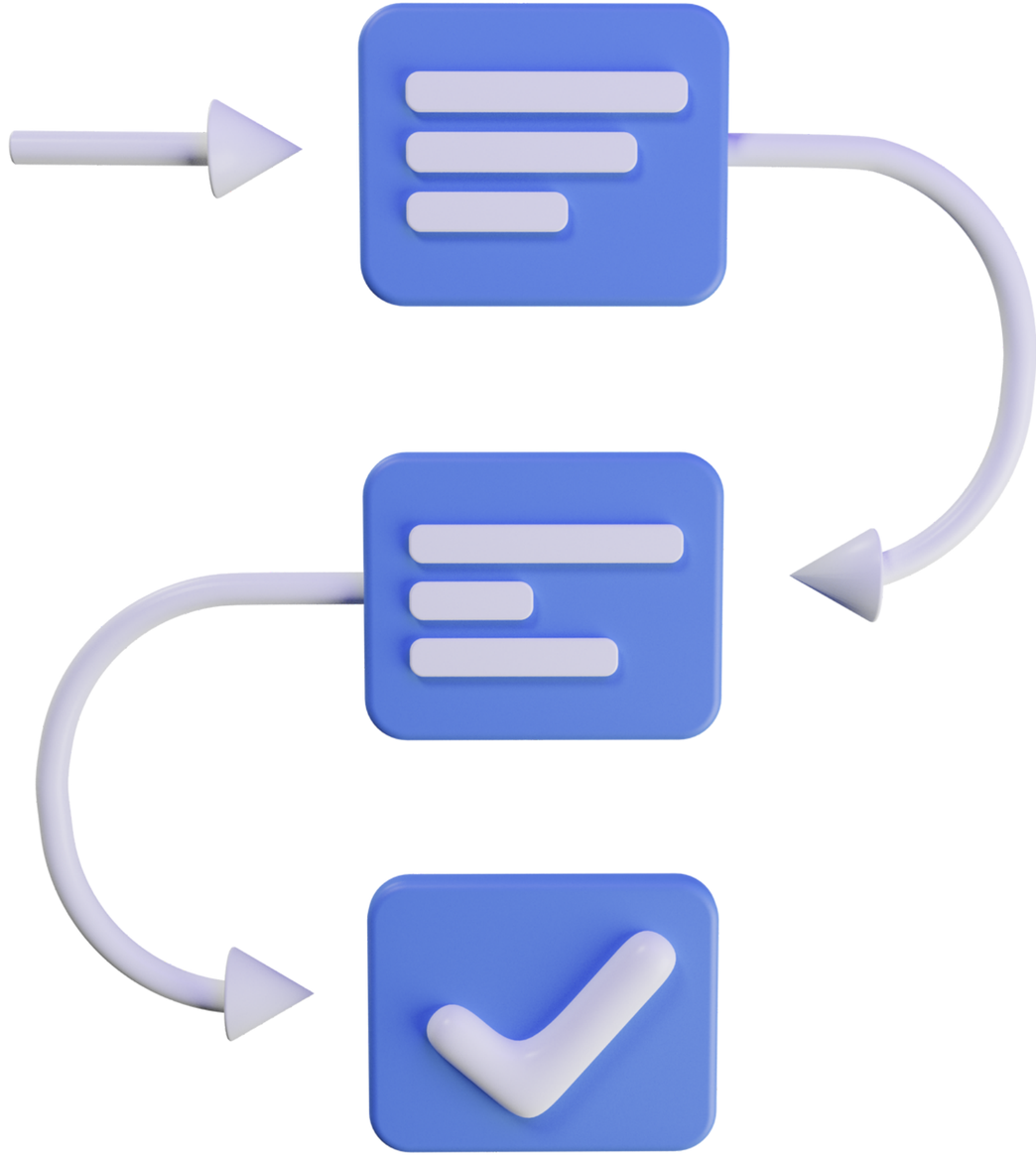
Distribution of ratings for open and closed restaurants.



Rating stars	is_open = 0 (%)	is_open = 1 (%)
1.0	1.03	1.40
1.5	2.55	3.47
2.0	5.70	6.50
2.5	10.13	9.37
3.0	15.43	11.47
3.5	21.73	16.59
4.0	22.50	20.24
4.5	15.16	18.83
5.0	5.78	12.14

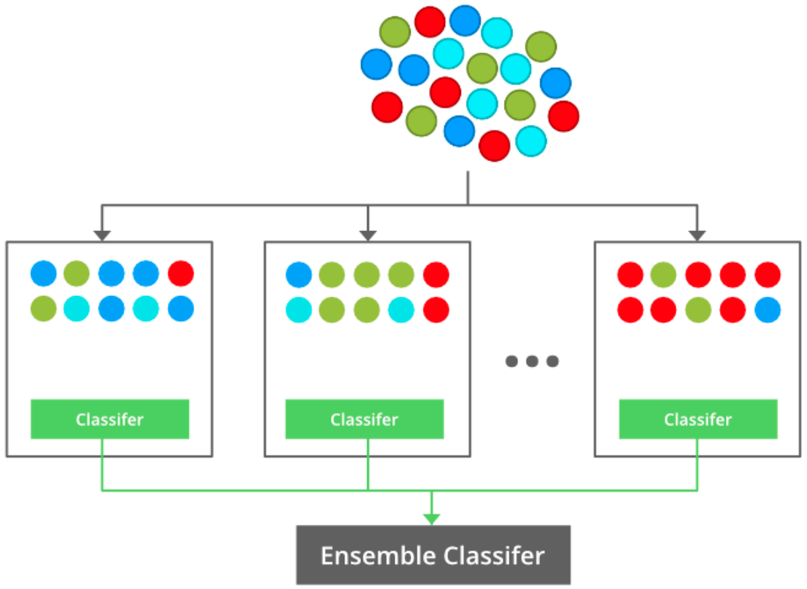
Machine Learning Models

Logistic Regression	70 %
Random Forest	70 %
Random Forest with Feature reduction	69 %
Gradient Boosting	70 %
Gradient Boosting in Pipeline	71%
XGboost	71%
XGboost in Pipeline	71%



XGBoost

XGBoost is the winner



Original Data

Bootstrapping

Aggregating

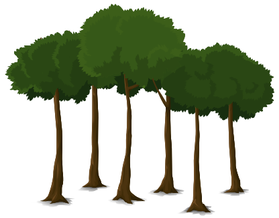
Bagging

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

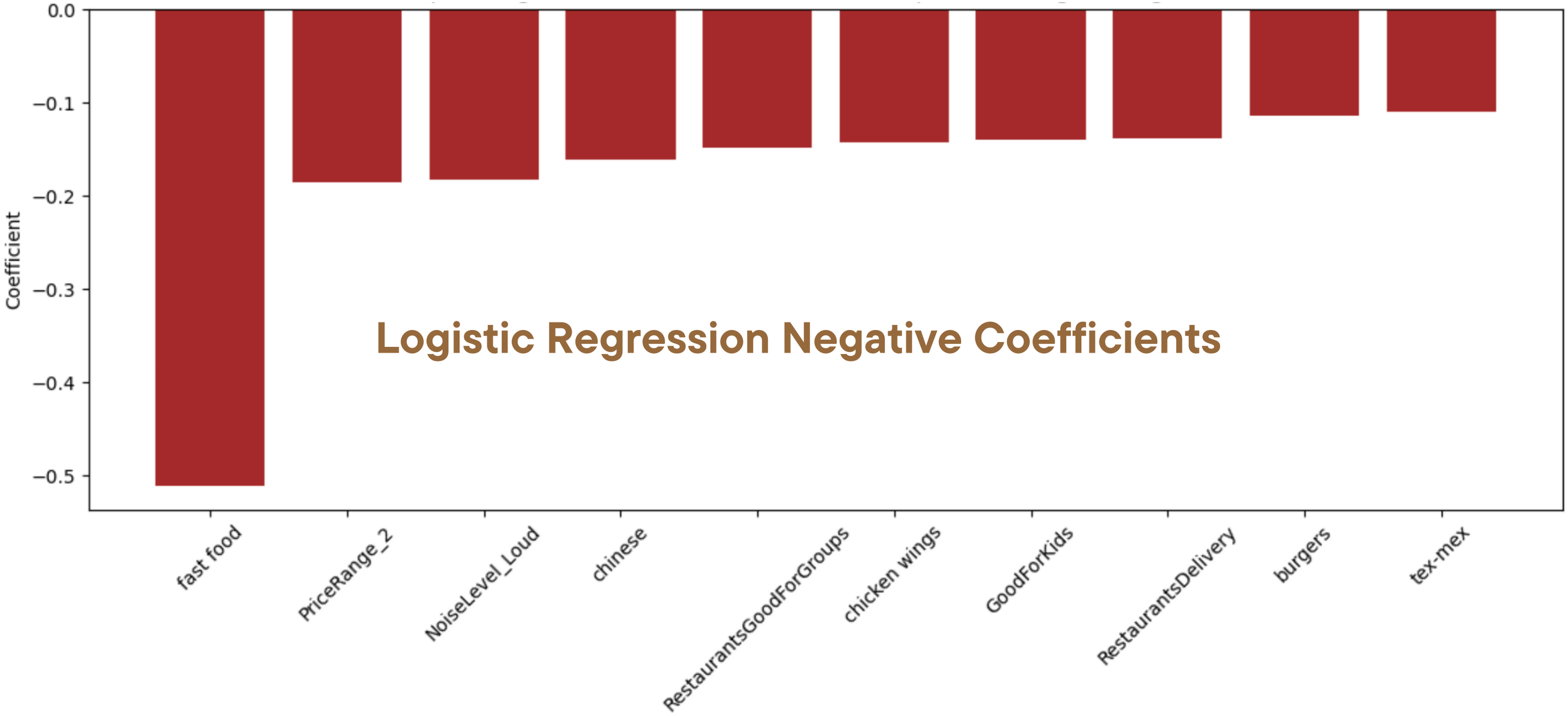
- Logistic Regression Coefficients



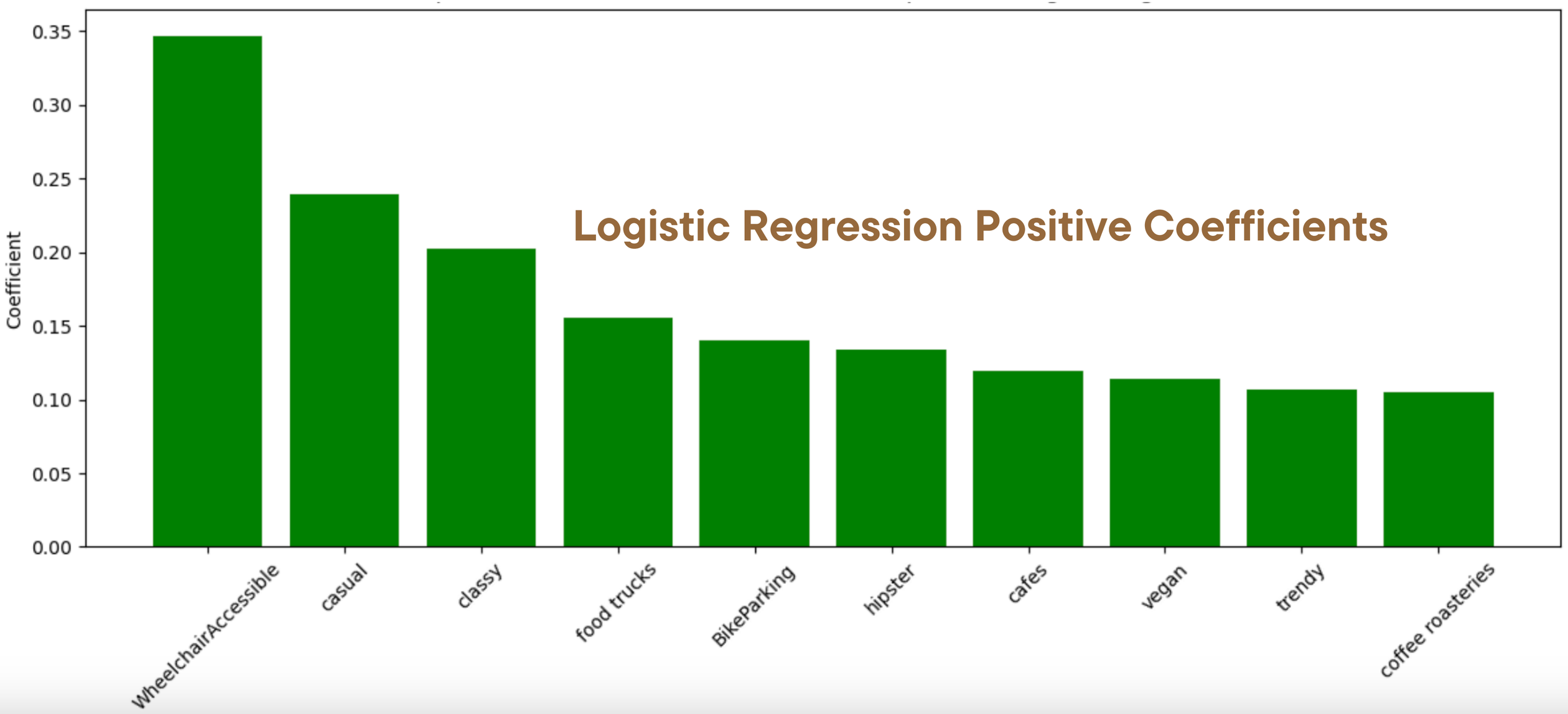
- Random Forest Feature Importance

- SHAP Analysis with XGBoost

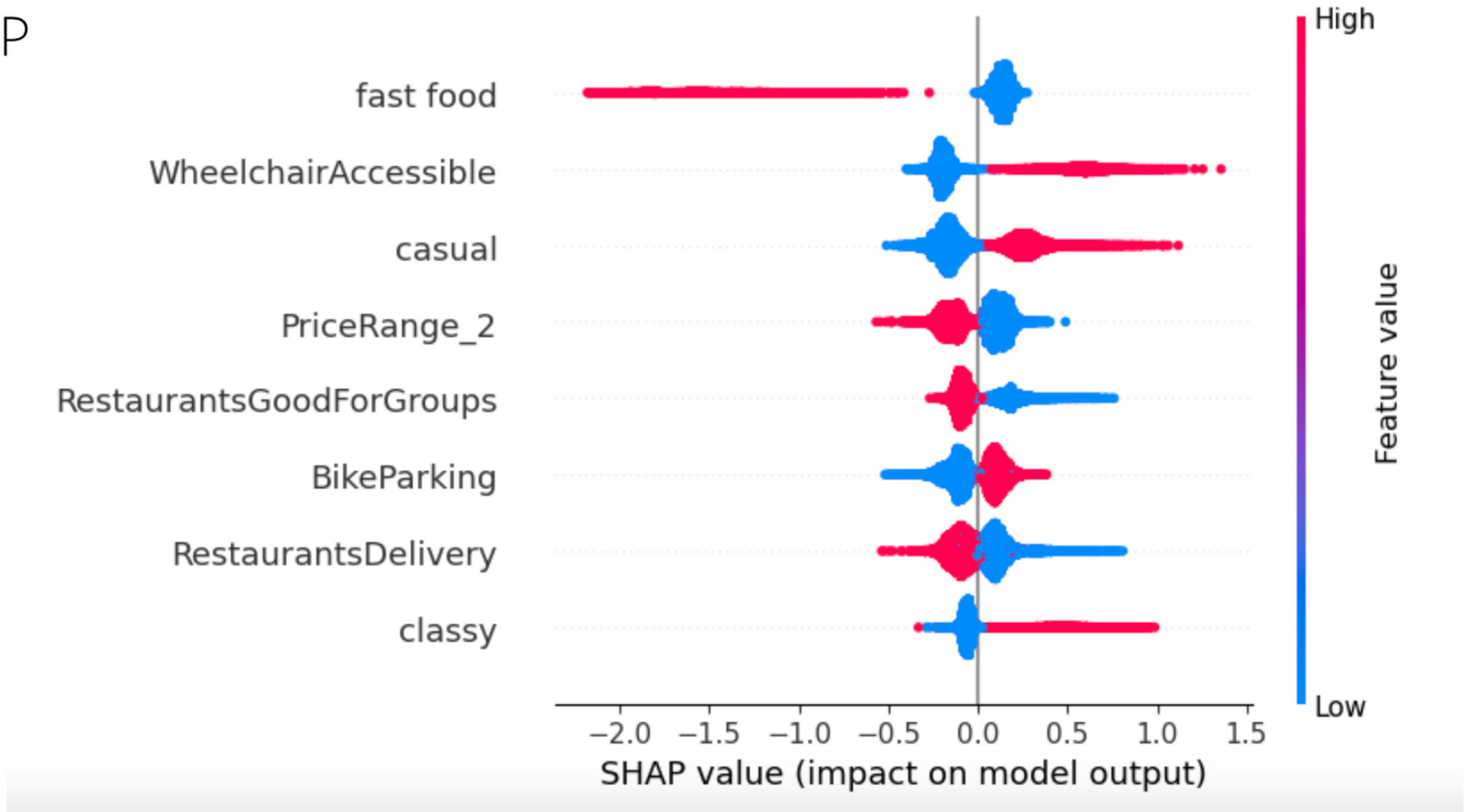
What Diners Might Not Prefer



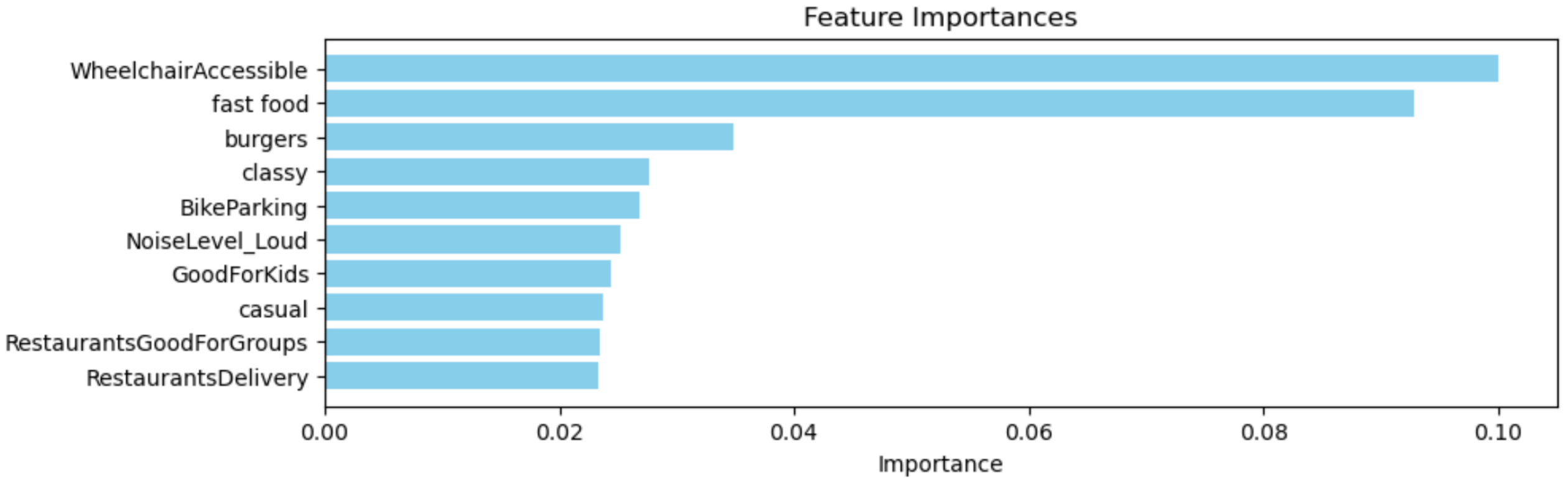
Factors That Boost Restaurant Ratings



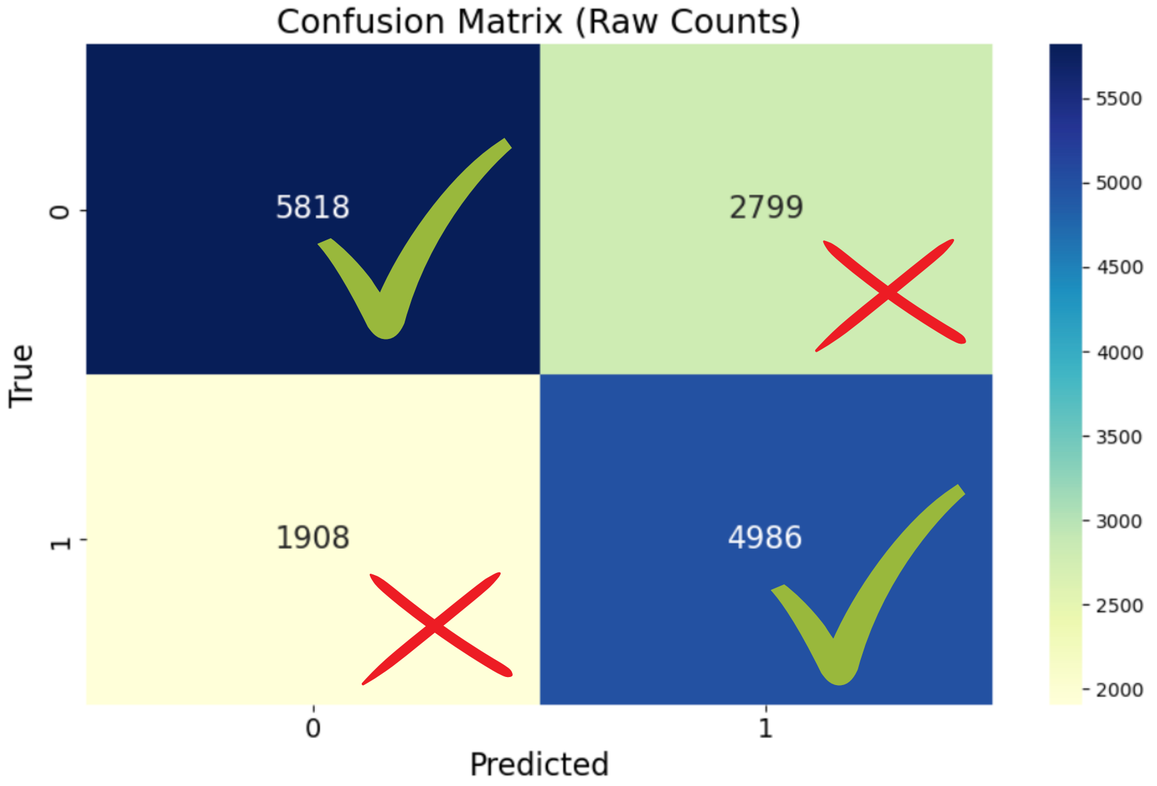
SHAP Analysis with XGBoost

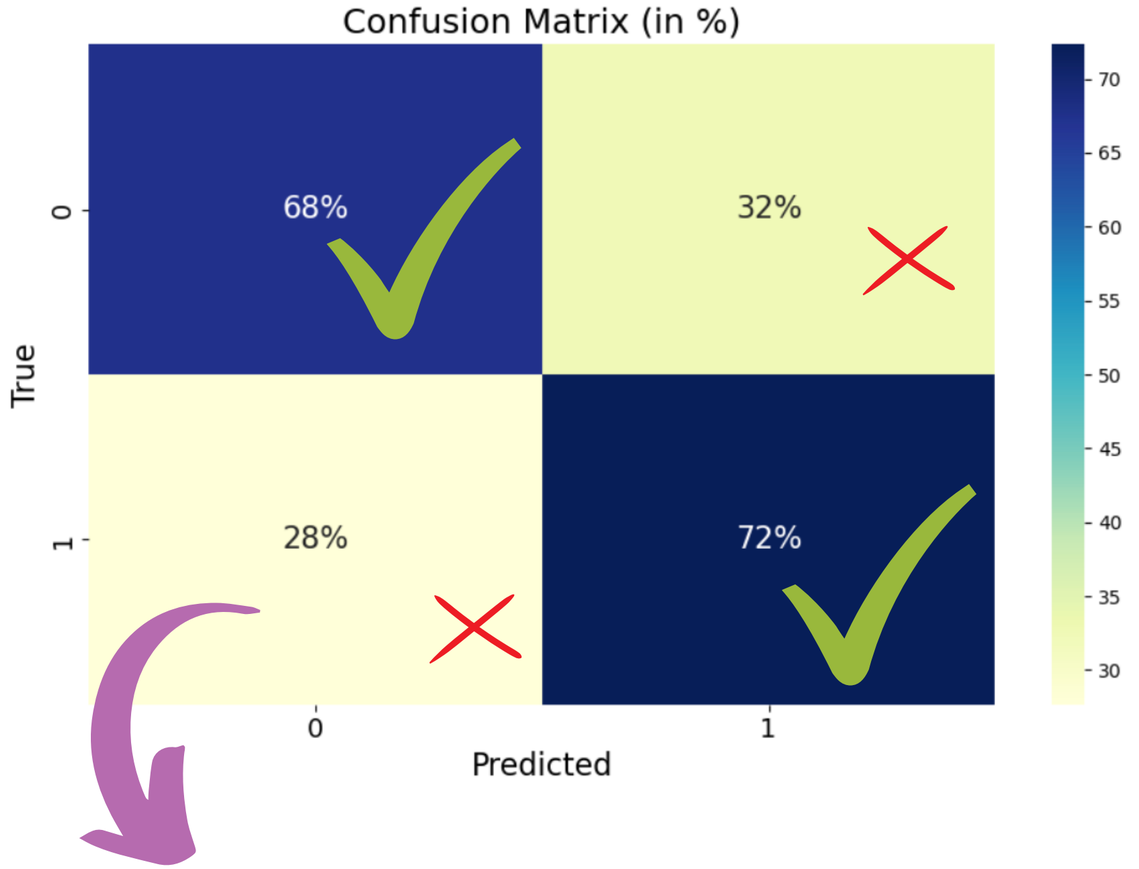


Random Forest Feature Importance



Evaluating Our Model's Predictions

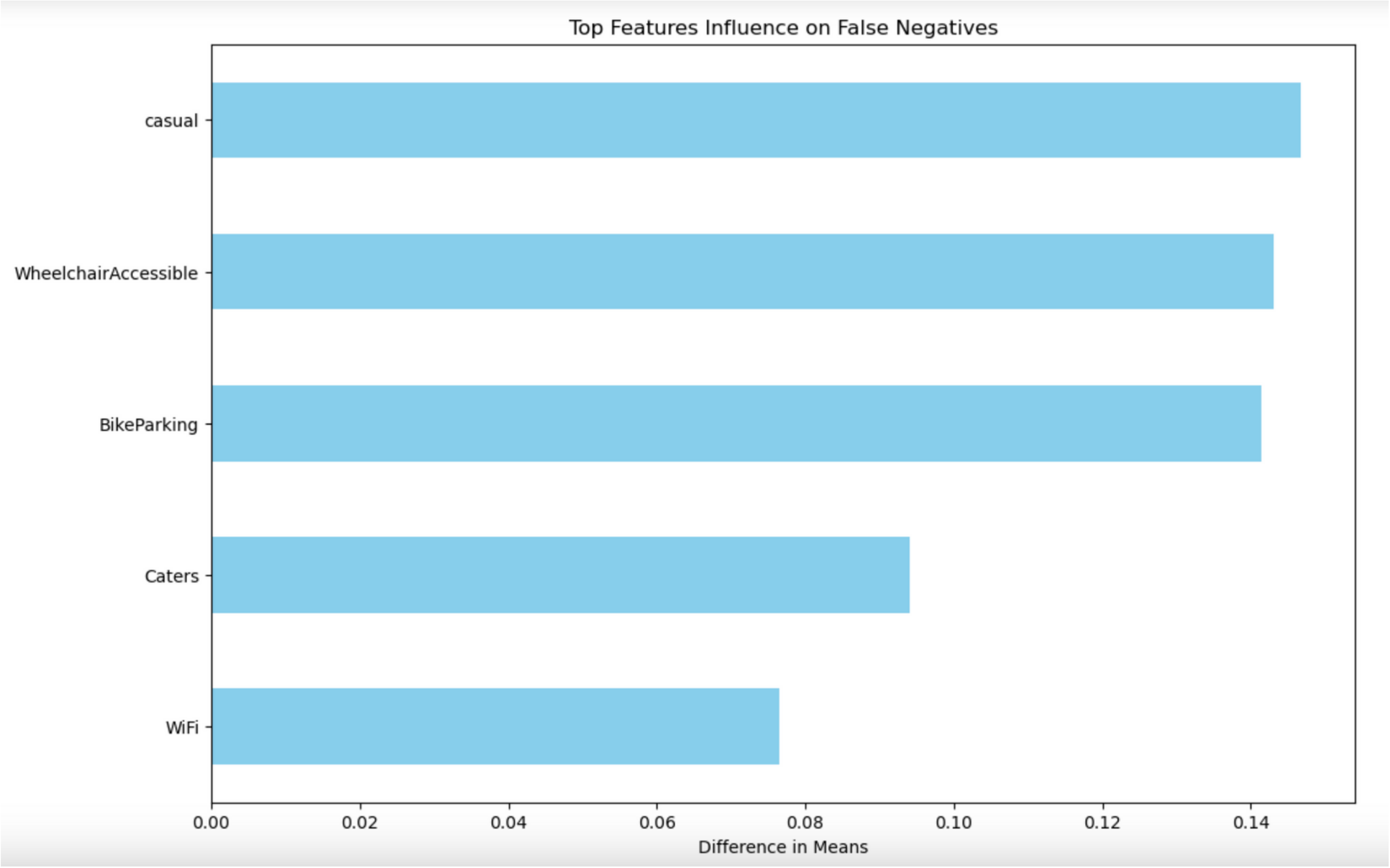




WHERE PREDICTIONS AND REALITY DON'T MATCH

Model bias

Feature Mean Differences Analysis



Forging Ahead

Potential Improvements

Recommendations:

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

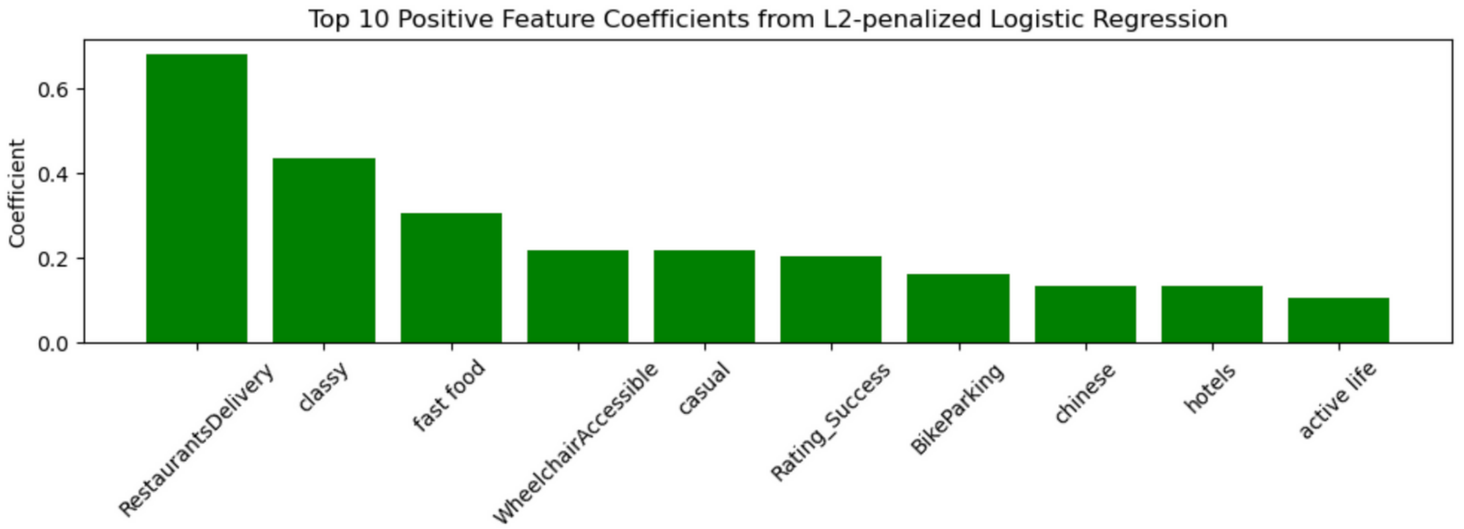
b. **Resampling:** Achieve balanced representation for wheelchair accessible venues.

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

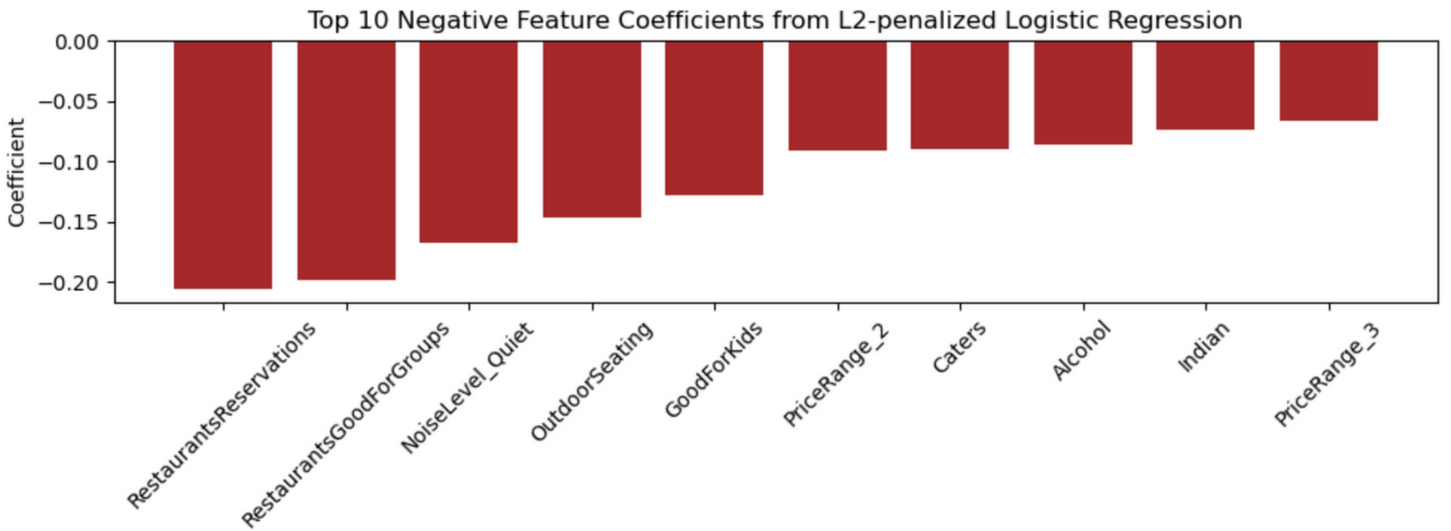
Thank You

Appendix

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