Predicting Restaurant Ratings Using Yelp Business Metadata and Engagement

Team Members:
Pilar de Haro - pdeharo0898@sdsu.edu
Aichu Tan - dtan3697@sdsu.edu
Swikriti Joshi - sjoshi2639@sdsu.edu
Eduardo Rosas - erosas9172@sdsu.edu

Github Link : https://github.com/AichuTan/BDA602_final_yelp_analysis
Video Link : https://www.youtube.com/watch?feature=shared&v=70_A0i9r2HI

Abstract

Online reviews heavily influence consumer choices, and Yelp ratings affect restaurant visibility and success. This study explores whether structured metadata : price range, amenities, hours, and location can predict restaurant ratings without text reviews. Using the 2024 Yelp Open Dataset, we analyzed 36,261 restaurants with 61 engineered features, framing prediction as a binary classification (≥4★ vs <4★) using Logistic Regression, Elastic Net, Random Forest, and XGBoost within a unified Scikit-learn pipeline (5-fold CV, MLflow). Ensemble models outperformed linear ones, revealing that nonlinear feature interactions drive customer satisfaction. XGBoost achieved the highest accuracy (0.733) and ROC-AUC (0.79), while Random Forest had the best F1-score (0.659) and guided feature interpretation. Logistic regression provided interpretability via coefficient signs and odds ratios. Both tree-based importance and logistic coefficients highlighted review count, price range, operating hours, and amenities (e.g., wheelchair access, table service) as key predictors. K-Prototypes clustering (k = 3) identified family-casual, mid-range social, and quick-service archetypes, each linked to distinct operational and rating profiles. Overall, structured business metadata effectively predicts Yelp ratings and offers actionable insights for small restaurants to enhance service quality, accessibility, and customer satisfaction.

1. Introduction

Online reviews significantly influence consumer choices, and Yelp has become a leading platform for evaluating restaurants. A restaurant’s average star rating is particularly important, shaping customer trust, search visibility, and business performance. For small businesses, understanding what drives higher ratings can inform strategic decisions and improve competitiveness.

While most previous research emphasizes sentiment analysis of review text, structured metadata including location, categories, amenities, and business hours remains underexplored. These features are often easier to interpret and directly actionable for small business owners.

This study leverages the 2024 Yelp Open Dataset to predict restaurant ratings using only metadata, offering a lightweight, interpretable alternative to text-based approaches. We frame this as a binary classification problem (≥4 vs <4 stars) using four supervised algorithms: Logistic Regression, Elastic Net, Random Forest, and Gradient Boosting.

Additionally, we apply unsupervised clustering (K-Means and K-prototype methods) to identify latent groupings among restaurants. Our contributions are threefold: demonstrating the predictive power of metadata, identifying which business attributes correlate most with ratings, and providing practical guidance for small businesses and marketing teams.

2. Data Sources

2.1 Dataset Overview

This study uses the **2024 Yelp Open Dataset**, which includes millions of business listings, user reviews, and structured metadata from across the United States and other regions. We focused on two core components: (1) the business.json file, which provides metadata such as categories, average ratings, price range, and amenities, and (2) a subset of reviews (2019-2022) from the review.json file. Restaurants were identified using the categories field within the business data.

The raw JSON-lines files contain nested dictionaries for attributes, categories, and hours. To prepare the data for analysis, these nested structures were normalized into a tabular format, producing a single row per restaurant. Our final sample contains **36,261 restaurants** with **61 engineered features**, representing a comprehensive cross-section of the U.S. restaurant landscape.

2.2 Data Cleaning and Preprocessing

Metadata Transformation

We extracted key fields from business.json, including stars, review_count, attributes, categories, hours, city, state, and is_open. The review_count variable was converted to numeric and log-transformed (log1p) to stabilize skew and reduce the impact of extreme values. To quantify operating patterns, we engineered several operational features from daily hour strings, such as **total weekly hours**, **days open**, **weekend hours**, and **average daily hours**. After these features were computed, the original seven daily-hour columns were dropped to avoid redundancy and sparsity.

The nested attributes dictionary was expanded into **22 engineered features**, capturing key business characteristics such as *ByAppointmentOnly*, *GoodForKids*, *OutdoorSeating*, and *WheelchairAccessible*. Multi-class categorical fields were standardized for consistency:

- **Wi-Fi**: {no, free, paid}
- **Alcohol service**: {none, beer_and_wine, full_bar}
- **Attire**: {casual, dressy, formal}
- **Noise level**: {quiet, average, loud, very_loud}
- **Price range**: ordinal 1-4 (inexpensive → very expensive)

Boolean-like fields (e.g., *TakesReservations*, *Delivery*, *Takeout*, *HasTV*, *DogsAllowed*) were coerced into consistent categorical values while preserving missingness. To prevent **label leakage**, we removed the true stars column (target variable) and all textual identifiers such as name, address, and postal_code from the feature set.

Categorical Feature Engineering

Cuisine and venue type were represented through the **Top 20 most frequent restaurant categories**, including *American (Traditional)*, *Fast Food*, *Italian*, *Mexican*, and *Chinese*. Category strings were parsed, cleaned, and deduplicated to remove umbrella and non-food tags. Each selected category was encoded as a **tri-state indicator** (1 = present, 0 = absent, NaN = missing) to preserve information about uncertainty or missingness.

Geographic data were standardized by normalizing city names and state codes. To control sparsity, we **one-hot encoded** only the Top 20 U.S. cities, collapsing all others into an *“Other”* category. In later modeling stages, we excluded these one-hot city indicators and instead retained **latitude** and **longitude** as continuous covariates, which preserved geographic signal while reducing high-dimensional redundancy.

Review Aggregation and Label Generation

For labels and auxiliary signals, we aggregated reviews from 2019-2022 to the business level. For each restaurant, we computed the **average star rating**, **total review count**, and **first and last review dates**, filtering for businesses with at least **three reviews (MIN_REV ≥ 3)** to ensure label reliability. In addition to star ratings, we engineered two review-style proxies:

- **Mean review word length** (to represent verbosity), and
- **Share of short reviews (≤ 24 words)** (to capture review brevity or frequency patterns).

These review-derived aggregates were merged with business metadata using business_id as a unique key, producing a unified dataset for supervised learning. Numeric fields were stored as continuous variables, and missing values were deliberately retained to allow consistent imputation across folds during cross-validation.

2.3 Final Preparation for Modeling

All numeric features were scaled using a **MinMaxScaler**, while categorical variables were one-hot encoded through a **ColumnTransformer** within the Scikit-learn pipeline. The final dataset was split into **80% training** and **20% testing** sets. Finally, to maintain reproducibility, the preprocessed dataset and modeling-ready feature matrix were exported in both **CSV** and **Pickle (PKL)** formats, enabling version control and compatibility with MLflow logging and deployment workflows.

3. Methods

This study employed four **supervised learning algorithms** and one **unsupervised clustering method** to predict and interpret restaurant ratings using structured metadata from the 2024 Yelp Open Dataset. The supervised models included **Logistic Regression**, **Elastic Net**, **Random Forest**, and **XGBoost**, while **K-Prototypes** was used for unsupervised segmentation.

Logistic Regression served as an interpretable baseline, and Elastic Net extended it with combined L1–L2 regularization to handle correlated predictors and improve generalization. The ensemble models, Random Forest and XGBoost, were selected for their capacity to model nonlinear relationships and capture higher-order feature interactions among mixed numerical and categorical attributes.

All analyses were implemented in **Python**, using **Scikit-learn** for preprocessing and model pipelines, **XGBoost** for gradient boosting, and **MLflow** for automated experiment tracking and reproducibility. A unified **ColumnTransformer** managed data preprocessing, incorporating median imputation for numeric variables, one-hot and ordinal encoding for categorical features, and Min–Max scaling to normalize feature ranges.

Model training and hyperparameter tuning were conducted using **Randomized Search** with **five-fold stratified cross-validation** to ensure robust model selection. Ensemble models were optimized for parameters such as the number of trees, maximum depth, and learning rate, while linear models were tuned for regularization strength and Elastic Net mixing ratio. All configurations, metrics, and artifacts were systematically logged in MLflow for transparent comparison and reproducibility.

Model performance was assessed using **accuracy**, **precision**, **recall**, **F1-score**, and **ROC-AUC**. ROC-AUC served as the primary optimization metric due to its robustness to class imbalance, while F1-score guided final threshold calibration. Each model was retrained on the full training dataset and validated on a holdout test set, with **ROC** and **precision–recall curves** visualizing discriminative performance.

For **unsupervised analysis**, the **K-Prototypes algorithm** was applied to uncover latent groupings among restaurants based on metadata features such as amenities, price range, and operating hours. K-Prototypes was chosen over K-Means for its ability to handle mixed data types, combining Euclidean distance for numeric attributes with dissimilarity matching for categorical variables. The optimal number of clusters was determined using the **elbow method** and **within-cluster dissimilarity (WCC)** to balance interpretability and model fit.

Overall, this integrated framework combined **predictive modeling** and **descriptive clustering**: the supervised models quantified how well business metadata predicted rating categories (≥4.0 vs. <4.0 stars), while the unsupervised approach revealed underlying **restaurant typologies** within the Yelp dataset. Together, these methods provided both **predictive accuracy** and **actionable segmentation insights** into the operational and service characteristics that shape restaurant performance.

4. Analysis & Results

4.1 Exploratory Data Analysis (EDA)

Before applying predictive models, an exploratory data analysis (EDA) was conducted to examine relationships between restaurant metadata features and average Yelp ratings. Pearson correlation coefficients were computed between all numeric and encoded attributes and the target variable (`avg_stars_2019`), revealing several operational and service-related features associated with rating outcomes.

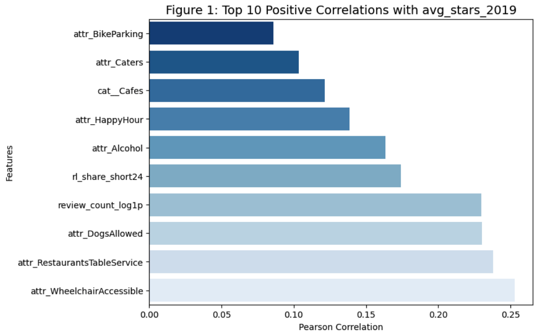


Figure 1: Top 10 Positive Correlations with avg_stars_2019

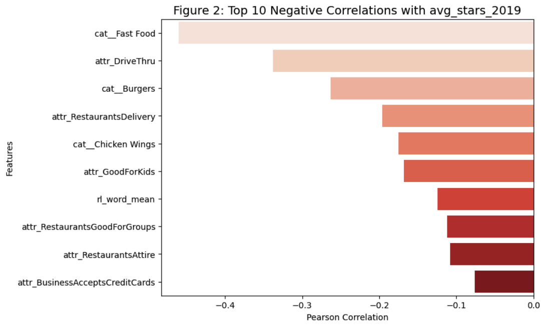


Figure 2: Top 10 Negative Correlations with avg_stars_2019

Figure 1 presents the top ten features most positively correlated with higher Yelp ratings. The strongest positive relationships were found for **Wheelchair Accessibility**, **Table Service**, **Dog-Friendly policies**, **review count**, **Alcohol or Happy Hour availability**, and **Bike Parking**—amenities that enhance comfort, accessibility, and inclusivity. Additionally, restaurants categorized as **Cafés** and **Caterers** tended to receive higher ratings, suggesting that relaxed, service-oriented venues are more favorably reviewed.

In contrast, **Figure 2** highlights the top ten features negatively correlated with ratings. Attributes such as **Fast Food**, **Drive-Thru**, **Delivery**, **Chicken Wings**, and **Burgers** were associated with lower average scores. These are typical of **quick-service business models**, which prioritize convenience but may compromise personalized service or ambiance, resulting in more moderate customer evaluations. Other features, including **Good for Kids**, **RestaurantsGoodForGroups**, **Restaurant Attire**, and **average review word length (word_mean)**, also showed weak negative correlations—indicating that family-oriented or highly casual restaurants may receive slightly lower ratings on average.

Overall, the EDA suggests that **amenities, accessibility, and service quality** are key drivers of customer satisfaction on Yelp, while **fast-service and high-volume restaurant types** tend to receive lower ratings. These insights provide a valuable foundation for subsequent predictive modeling and feature importance analysis.

4.2 Overview of the ML Pipeline

The end-to-end machine learning workflow consisted of four main stages:

1. **Exploratory Data Analysis (EDA):** Examined the distribution of star ratings, review counts, and price ranges, and filtered out restaurants with fewer than three reviews to ensure label reliability.
2. **Preprocessing**: Continuous features (e.g., total weekly hours, latitude, and log-transformed review counts) were **median-imputed** and **Min-Max scaled**, while categorical and Boolean attributes were processed through **One-Hot** and **Ordinal Encoding**. Additional engineered features captured operational and review-based patterns relevant to restaurant performance.
3. **Modeling and Hyperparameter Tuning:** Four **supervised models -** Logistic Regression, Elastic Net, Random Forest, and XGBoost-and two **unsupervised algorithms -** K-Means, and K-Prototypes, were trained using a unified **Scikit-learn pipeline** with **five-fold stratified cross-validation**. Hyperparameter tuning was guided by grid/randomized search and performance visualizations such as heatmaps and 3D plots.
4. **Evaluation and Experiment Tracking:** All models were logged and compared using **MLflow**, which automatically recorded parameters, metrics, and visual results. This enabled transparent performance tracking, reproducibility, and easy comparison across different model runs.

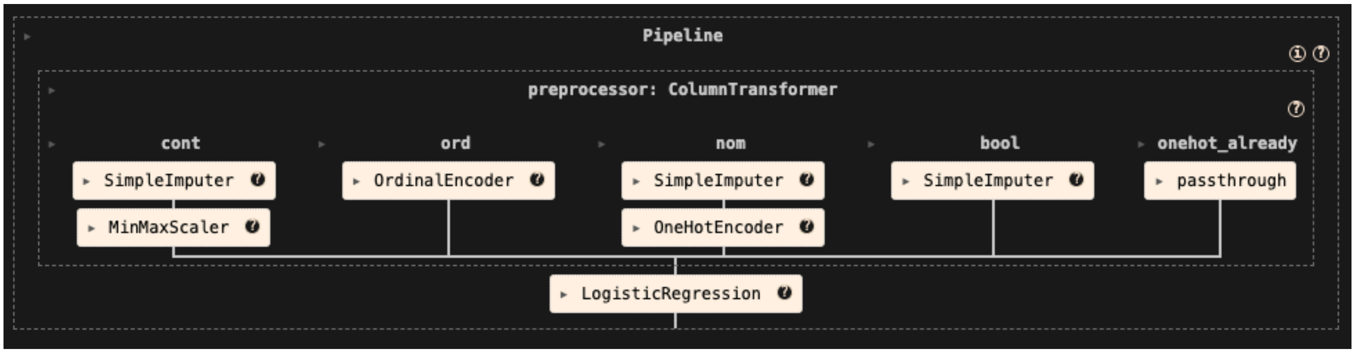


Figure 3. *Scikit-learn pipeline for Yelp rating prediction. The ColumnTransformer preprocesses continuous, ordinal, nominal, and Boolean features using appropriate imputers and encoders before passing the transformed data to a LogisticRegression classifier. This unified pipeline ensures consistent preprocessing across folds, prevents data leakage, and supports reproducible model tracking in MLflow.*

4.3 Model Evaluation and Comparison

Four supervised classification models - Logistic Regression, Elastic Net, Random Forest, and XGBoost, were trained to predict whether a restaurant's average Yelp rating was high (≥4★) or low (<4★). Each model was implemented within the same preprocessing pipeline and evaluated using Stratified 5-Fold Cross-Validation to ensure class balance. Performance was measured using Accuracy, F1-score, and ROC-AUC, which together capture precision, recall, and overall discriminative ability.

Table 1. Summarizes the results across all algorithms. Both ensemble tree-based models (Random Forest and XGBoost) outperformed the linear baselines, confirming that non-linear relationships among business attributes, such as amenities, pricing, and hours carry predictive value.

Model	Accuracy	F1 Score	ROC-AUC
Logistic Regression	0.69	0.648	0.760
Elasticnet	0.677	0.638	0.760
Random Forest	0.723	0.659	0.787
XGBoost	0.733	0.638	0.787

While XGBoost achieved slightly higher accuracy (0.733), the Random Forest model yielded the best F1-score (0.659) and matched XGBoost's ROC-AUC (0.79), indicating stronger class balance and generalization. Logistic Regression and Elastic Net performed moderately well, establishing interpretable baselines and confirming that metadata alone provides substantial predictive signal.

These results demonstrate that even without text-based sentiment features, structured business metadata, capturing price range, hours, location, and amenities, can effectively predict Yelp star ratings. The Random Forest model was therefore selected as the primary model for interpretability and feature analysis, balancing predictive performance with explanatory power.

4.4 Feature Importance and Interpretability

Two complementary analyses were conducted to interpret the predictive model outputs and identify which business attributes most strongly influenced restaurant ratings. The **Random Forest model** (Figure 4) ranked features by their contribution to predictive accuracy using **Gini importance**, which quantifies how much each feature reduces node impurity across all trees in the ensemble. For a given feature *j*, its importance I*j* is computed as:

$$I_j = \frac{1}{T} \sum_{t=1}^{T} \sum_{n \in \mathcal{N}_t(j)} p(n)\, \Delta i(n)$$

where T is the total number of trees, *Nt(j)* is the set of nodes in tree ttt where feature *j* is used for splitting, *p(n)* is the proportion of samples reaching node *n*, and Δ*i(n)*is the reduction in Gini impurity resulting from that split. This measure was chosen because it provides an interpretable, model-based estimate of each variable's global importance, reflecting how consistently a feature contributes to improving classification performance across the ensemble.

Figure 5: Logistic Regression — Direction & Strength of Feature Effects

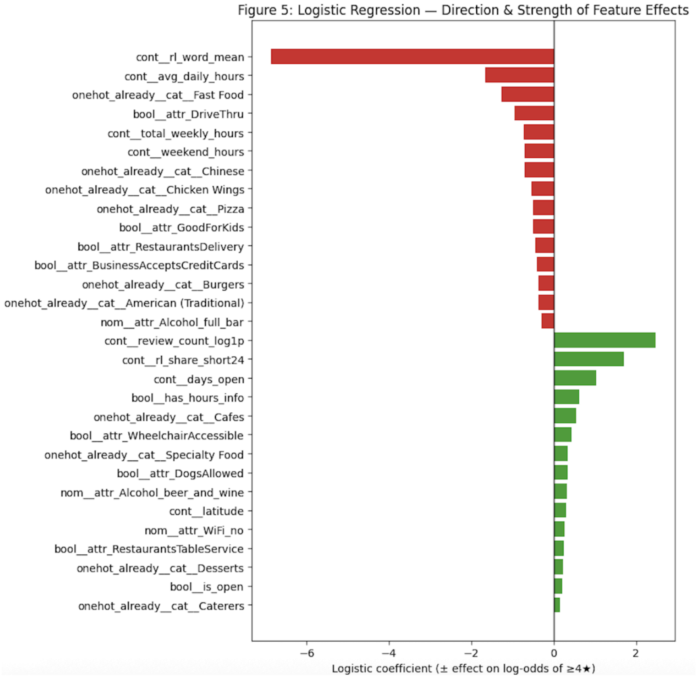
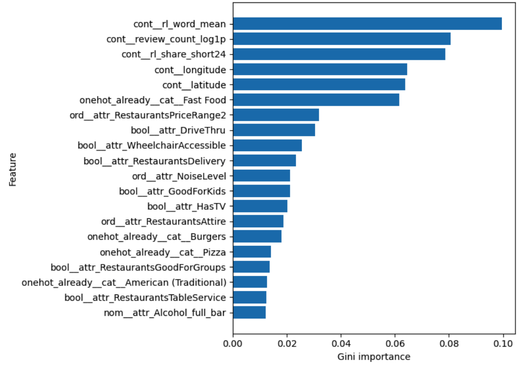


Figure 4:Random Forest Feature Importance (Top 20)



The most influential variables were **average review length (rl_word_mean)**, **log-transformed review count**, and **share of short reviews**, suggesting that patterns in review behavior indirectly reflect customer satisfaction even without text sentiment. Among **categorical and operational features**, restaurant type (e.g., *Fast Food, Burgers, Pizza, American*), **price range**, **Drive-Thru**, **Wheelchair Accessibility**, **Delivery**, **Noise Level**, **Good for Kids**, **Good for Groups**, **Table Service**, **HasTV**, and **Restaurant Attire** also played key roles in differentiating high- from low-rated restaurants.

To further interpret model behavior, a **Logistic Regression model** (Figure 5) was used to examine the **direction and strength** of feature effects. Each coefficient represents how a one-unit increase in a feature changes the likelihood of achieving a high rating (≥ 4★). Positive coefficients increase that likelihood, while negative coefficients decrease it.

Among **review behavior features**, longer reviews were negatively associated with high ratings, while **shorter reviews** and **higher review counts** showed positive effects—indicating that concise, frequent feedback patterns align with better-rated restaurants. For **operational and service attributes**, **Drive-Thru**, **Delivery**, and **longer operating hours** had negative coefficients, suggesting that convenience-oriented formats may trade off perceived quality. In contrast, having **hours of information**, **more days open**, and amenities such as **Wheelchair Accessibility**, **Table Service**, **Dog-Friendly policies**, and **Cafés** increased the odds of higher ratings.

Finally, **cuisine-related features** such as *Fast Food, Pizza,* and *American (Traditional)* were associated with lower ratings, whereas *Dessert* and *Café* categories showed positive effects. Overall, these findings align with the Random Forest results which are **accessible, service-oriented restaurants** tend to earn higher Yelp ratings, while fast-service models receive more moderate evaluations.

Overall, the model-based findings closely mirror the EDA results, confirming consistent relationships between restaurant characteristics and Yelp ratings. Features such as **Wheelchair Accessibility**, **Table Service**, **Dog-Friendly policies**, and **review count**, which showed strong positive correlations in EDA, also emerged as important predictors in both the Random Forest and Logistic Regression models. These attributes reflect comfort, accessibility, and inclusivity—key qualities linked to customer satisfaction. Conversely, **Fast Food**, **Drive-Thru**, and **Delivery** consistently appeared as negative predictors across all analyses, suggesting that quick-service business models, while convenient, may sacrifice perceived service quality or atmosphere. Logistic Regression provided additional directional insight, revealing that **shorter reviews**, **more days open**, and **amenities like Cafés and Wheelchair Accessibility** increase the odds of higher ratings, while **longer reviews**, **extended hours**, and **cuisine types such as Burgers, Pizza, and American (Traditional)** reduce them. Collectively, these findings demonstrate that **service-oriented, accessible, and experience-focused restaurants** tend to achieve higher Yelp ratings, whereas **fast-service or high-volume formats** receive more moderate evaluations. The consistency across EDA and both predictive models reinforces the robustness of these patterns and validates the importance of operational and amenity features in shaping customer satisfaction.

4.5 Unsupervised Clustering Analysis

To complement the predictive modeling, an **unsupervised clustering analysis** was conducted using the **K-Prototypes algorithm** to uncover natural groupings of restaurants based on mixed metadata. K-Prototypes was selected because it effectively handles both numerical (e.g., hours, review volume, coordinates) and categorical variables (e.g., amenities, price range, and cuisine type).

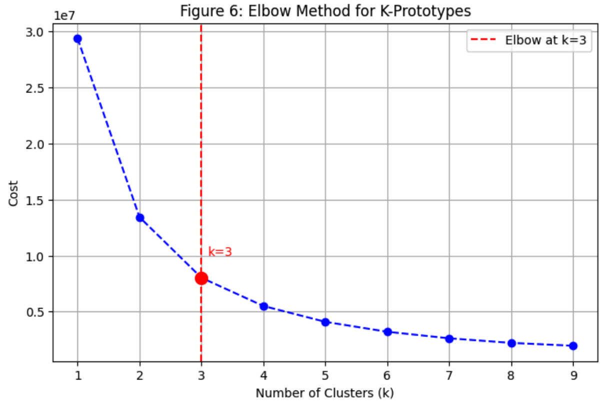


Figure 6: Elbow Method for K-Prototypes

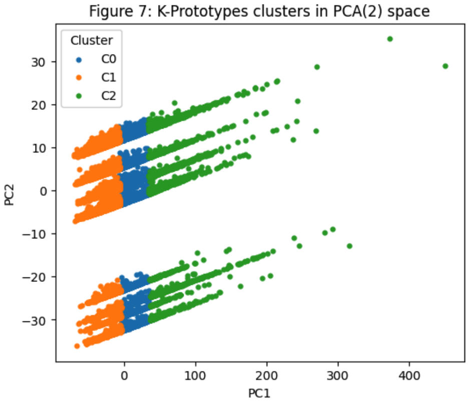


Figure 7: K-Prototypes clusters in PCA(2) space

The **elbow method** (Figure 6) identified **three optimal clusters (k = 3)**, which were confirmed through a two-dimensional **PCA projection** (Figure 7) showing clear group separation. The **cluster size distribution** (Figure 8) indicated that Clusters 0 and 1 represented the majority of restaurants (≈16,000 each), while Cluster 2 formed a smaller subset (≈3,000).

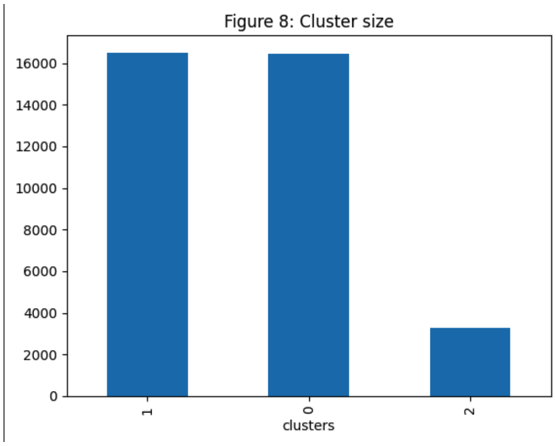


Figure 8: Cluster size

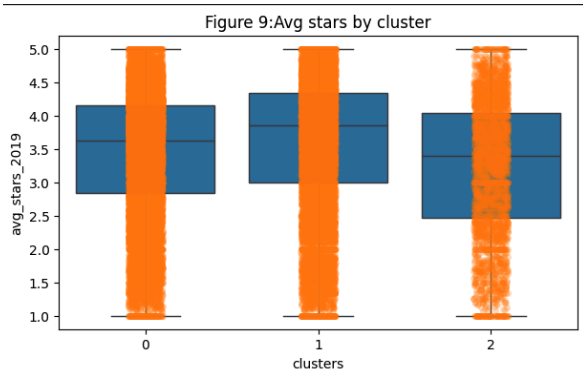


Figure 9:Avg stars by cluster

Average rating distributions (Figure 9) revealed meaningful differentiation among clusters. **Cluster 1** achieved the highest mean rating (**3.61 ★**), followed by **Cluster 0** (**3.43 ★**) and **Cluster 2** (**3.24 ★**). Lift profiles clarify the operational character of each group (*See Appendix E-G*). **Cluster 1** (highest-rated) reflects **value-oriented, accessible, family-casual venues** efficient formats that would still benefit from a few comfort upgrades (e.g., Wi-Fi, clearer hours, basic accessibility). **Cluster 0** (mid-rated) represents **social, mid-priced casual dining** with alcohol/happy hour, table service, and moderate noise, delivering solid but not top-tier satisfaction. **Cluster 2** (lowest-rated) corresponds to **mid-to-upscale dine-in** which higher prices, reservations, quieter rooms, suggesting that premium signals alone do not guarantee higher ratings and may suffer from expectation value gaps (e.g., wait times, service consistency).

These segments reinforce the supervised findings: **comfort and service accessibility** (Wi-Fi, table service, reservations, moderate noise) and **clear operational choices** (price positioning, reliable hours) align with higher ratings, whereas **pure speed/convenience formats** tend to rate lower unless balanced by comfort signals. Across clusters, **small, targeted operational upgrades**, improving accessibility, adding Wi-Fi, maintaining consistent hours, and managing ambiance which offer practical, low-cost levers to move restaurants toward the higher-rated profiles.

5. Discussion and Conclusion

This study demonstrates that **structured business metadata alone** can effectively predict Yelp restaurant ratings without relying on text-based sentiment analysis. The analysis integrated multiple supervised learning models and unsupervised clustering to provide both **predictive performance** and **descriptive insights** into the operational factors driving customer satisfaction.

From a predictive standpoint, ensemble tree-based models such as **Random Forest** and **XGBoost** substantially outperformed linear baselines, achieving **ROC-AUC ≈ 0.79**. **XGBoost** yielded the highest accuracy (0.733), while **Random Forest** achieved the best F1-score (0.659), indicating stronger class balance and generalization. Although slightly less accurate, **Logistic Regression** offered valuable interpretability, its coefficients directly revealed whether business attributes increased or decreased the likelihood of earning a high Yelp rating. Together, these results confirm that **non-linear metadata interactions**—especially among price range, amenities, and operating hours—carry strong predictive power.

Feature analysis across models revealed several consistent drivers of restaurant success. Amenities and accessibility features such as **Wi-Fi**, **table service**, **wheelchair access**, and **dog-friendly policies** were positively associated with higher ratings, emphasizing the importance of comfort and inclusivity. In contrast, features characterizing **quick-service formats** (e.g., fast food, drive-thru, and extended operating hours) correlated with lower ratings, suggesting that convenience-oriented models may trade speed for perceived quality or atmosphere. Notably, **review behavior features**—such as shorter, more frequent reviews and higher review counts—also signaled stronger customer engagement and satisfaction.

The **unsupervised clustering analysis** provided additional insight into how restaurants naturally group based on their metadata. Using **K-Prototypes (k = 3)**, three distinct operational archetypes emerged:

- **Cluster 1 (highest-rated)** — Value-oriented, accessible, family-casual restaurants emphasizing affordability and convenience.
- **Cluster 0 (mid-rated)** — Social, mid-priced casual dining venues offering alcohol, happy hour, and table service, reflecting a balanced customer experience.
- **Cluster 2 (lowest-rated)** — Mid-to-upscale dine-in restaurants with higher prices and reservations but lower ratings, suggesting a potential mismatch between customer expectations and perceived value.

These clusters reinforce the supervised model results: **service accessibility, comfort, and clear operational strategy** are strong predictors of customer satisfaction. Conversely, even upscale establishments may underperform if customer experience does not match price expectations.

Overall, the findings show that **metadata-only models can serve as practical, interpretable tools** for restaurant owners and marketing teams. Small, targeted operational changes such as adding Wi-Fi, improving accessibility, ensuring consistent hours, and managing ambiance, which represent **low-cost, high-impact levers** to enhance both customer satisfaction and online visibility.

Limitations and Future Work

While metadata provides a robust foundation for prediction, it cannot capture subjective nuances such as service tone, food quality, or emotional sentiment expressed in reviews. Future research could enhance this framework by incorporating **text embeddings** from review content, **temporal features** reflecting rating trends, and **geospatial context** (e.g., neighborhood affluence or competition density). Integrating these additional data sources could improve both predictive performance and the depth of business insights.

Conclusion

In conclusion, this project establishes that **structured Yelp metadata alone can accurately predict restaurant ratings and uncover actionable operational patterns**. By combining supervised learning and clustering, the analysis delivers both **quantitative validation** and **strategic guidance**, helping small and mid-sized businesses identify which operational, pricing, and amenity factors most influence customer satisfaction. These insights demonstrate the real-world potential of **data-driven decision-making** in the restaurant industry and lay the groundwork for future research integrating richer, multi-modal data sources.

Author Bios

Pilar de Haro - Group Leader. Pilar is responsible for cleaning the dataset, training, testing, and validating the machine learning models. She also leads the analysis of results and compiles the final project paper. Contact: pdeharo0898@sdsu.edu

Aichu Tan - Modeling and Machine Learning. Aichu leads data preprocessing, feature engineering, and model development, using MLflow for reproducible experiment tracking and performance optimization. Contact: dtan3697@sdsu.edu

Swikriti Joshi - Data Preparation and Visualization. Swikriti focuses on cleaning and formatting the data and creating visual dashboards to effectively present insights. Contact: sjoshi2639@sdsu.edu

Eddie Rosas - Data Analysis and Visualization. Eddie contributes to cleaning the data, building machine learning models, and creating visualizations to communicate findings clearly. Contact: erosas9172@sdsu.edu

References (APA Style)

Guo, Y., Lu, A., & Wang, Z. (2017). Predicting restaurants' rating and popularity based on Yelp dataset (CS 229 Final Project). Stanford University. https://cs229.stanford.edu/proj2017/final-reports/5244334.pdf

Luo, Y., & Xu, X. (2019). Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: A case study of Yelp. Sustainability, 11(19), 5254. https://doi.org/10.3390/su11195254

Starakiewicz, T., & Wójcik, P. (2025). Predicting restaurant survival using nationwide Google Maps data. Knowledge-Based Systems, 313, 113198. https://doi.org/10.1016/j.knosys.2025.113198

Yu, M. (2015). Restaurants review star prediction for Yelp dataset (CSE 258 Project). UC San Diego. https://jmcauley.ucsd.edu/cse258/projects/fa15/017.pdf

Appendix:

Github Link : https://github.com/AichuTan/BDA602_final_yelp_analysis
Video Link : https://www.youtube.com/watch?feature=shared&v=70_A0i9r2HI

Appendix A: Metadata Report:

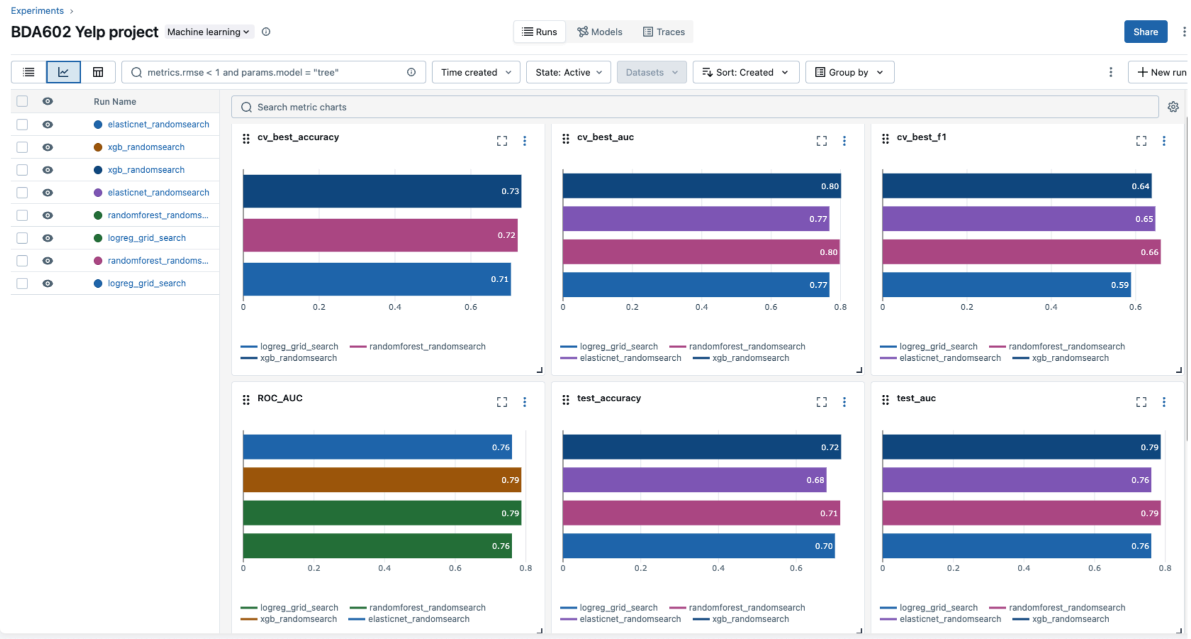
This table summarizes the key metadata features derived from the ***2024 Yelp Open Dataset*** *after preprocessing and feature engineering. These variables—including business attributes,* *operational details, location data, and review behavior metrics—served as input features for the restaurant rating prediction models.*

	dtype	n_unique	missing_sum	missing_pct	non_null	example
business_id	object	36261	0	0.0	36261	MTSW4McQd7CbVtyjqoe9mw
city	object	839	0	0.0	36261	Philadelphia
state	object	15	0	0.0	36261	PA
latitude	float64	34934	0	0.0	36261	39.9555052
longitude	float64	34500	0	0.0	36261	-75.1555641
review_count	float64	1127	0	0.0	36261	80.0
is_open	boolean	2	0	0.0	36261	TRUE
review_count_log1p	float64	1127	0	0.0	36261	4.394449
attr_ByAppointmentOnly	boolean	2	33122	0.9134	3139	FALSE
attr_BusinessAcceptsCreditCards	boolean	2	4889	0.1348	31372	FALSE
attr_BikeParking	boolean	2	9408	0.2595	26853	TRUE
attr_RestaurantsPriceRange2	category	4	6589	0.1817	29672	1.0
attr_RestaurantsTakeOut	boolean	2	3074	0.0848	33187	TRUE
attr_RestaurantsDelivery	boolean	2	4898	0.1351	31363	FALSE
attr_Caters	boolean	2	10665	0.2941	25596	TRUE
attr_WiFi	category	3	8736	0.2409	27525	free
attr_WheelchairAccessible	boolean	2	24474	0.6749	11787	TRUE
attr_HappyHour	boolean	2	25391	0.7002	10870	FALSE
attr_OutdoorSeating	boolean	2	7938	0.2189	28323	FALSE
attr_HasTV	boolean	2	6543	0.1804	29718	TRUE
attr_RestaurantsReservations	boolean	2	6953	0.1917	29308	FALSE
attr_DogsAllowed	boolean	2	25718	0.7092	10543	FALSE
attr_Alcohol	category	2	22886	0.6311	13375	full_bar
attr_GoodForKids	boolean	2	9158	0.2526	27103	TRUE
attr_RestaurantsAttire	category	3	10843	0.299	25418	casual
attr_RestaurantsTableService	boolean	2	19729	0.5441	16532	FALSE
attr_RestaurantsGoodForGroups	boolean	2	8730	0.2408	27531	TRUE
attr_DriveThru	boolean	2	31014	0.8553	5247	TRUE
attr_NoiseLevel	category	4	12587	0.3471	23674	average
attr_Smoking	category	0	36261	1.0	0	
total_weekly_hours	float64	178	0	0.0	36261	0.0
days_open	float64	8	0	0.0	36261	0.0

weekend_hours	float64	81	0	0.0	36261	0.0
avg_daily_hours	float64	175	0	0.0	36261	0.0
has_hours_info	boolean	2	0	0.0	36261	FALSE
cat__Sandwiches	Int8	2	0	0.0	36261	0
cat__American (Traditional)	Int8	2	0	0.0	36261	0
cat__Pizza	Int8	2	0	0.0	36261	0
cat__Fast Food	Int8	2	0	0.0	36261	0
cat__Breakfast & Brunch	Int8	2	0	0.0	36261	0
cat__American (New)	Int8	2	0	0.0	36261	0
cat__Burgers	Int8	2	0	0.0	36261	0
cat__Mexican	Int8	2	0	0.0	36261	0
cat__Italian	Int8	2	0	0.0	36261	0
cat__Coffee & Tea	Int8	2	0	0.0	36261	1
cat__Seafood	Int8	2	0	0.0	36261	0
cat__Chinese	Int8	2	0	0.0	36261	0
cat__Salad	Int8	2	0	0.0	36261	0
cat__Chicken Wings	Int8	2	0	0.0	36261	0
cat__Cafes	Int8	2	0	0.0	36261	0
cat__Delis	Int8	2	0	0.0	36261	0
cat__Caterers	Int8	2	0	0.0	36261	0
cat__Specialty Food	Int8	2	0	0.0	36261	0
cat__Bakeries	Int8	2	0	0.0	36261	1
cat__Desserts	Int8	2	0	0.0	36261	0
rev_count_2019	Int64	547	0	0.0	36261	20
avg_stars_2019	float64	6240	0	0.0	36261	4.55
first_review_2019	datetime64[ns]	36141	0	0.0	36261	2019-03-12 17:04:09
last_review_2019	datetime64[ns]	36191	0	0.0	36261	2021-11-01 18:22:07
rl_word_mean	float64	19355	0	0.0	36261	81.45
rl_share_short24	float64	2106	0	0.0	36261	0.05

Appendix B: MLflow Experiment Tracking Dashboard

*This dashboard displays the MLflow interface used to monitor and compare model performance for the BDA602 Yelp Project. Experiments included **Elastic Net**, **Random Forest**, **XGBoost**, and **Logistic Regression**, evaluated using metrics such as **Accuracy**, **F1-score**, and **ROC-AUC**. The results show that **XGBoost** and **Elastic Net** achieved the highest ROC-AUC values (≈ 0.80), indicating superior predictive* *performance. MLflow enabled systematic experiment logging, performance visualization, and transparent model comparison, facilitating data-driven model selection and reproducibility.*



Experiments ›

BDA602 Yelp project Machine learning

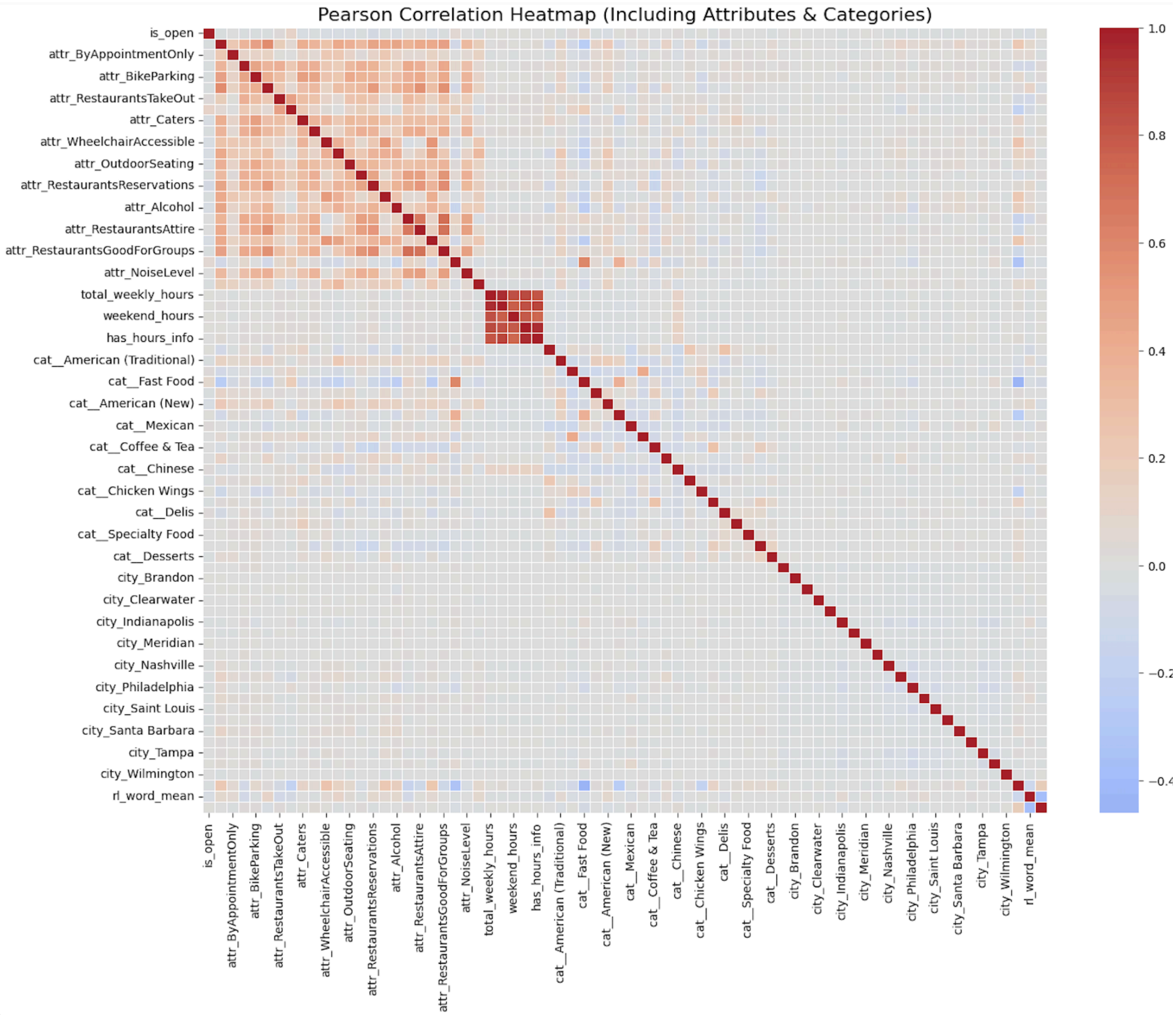
metrics.rmse < 1 and params.model = "tree"

Time created | State: Active | Datasets | Sort: Created | Group by

Run Name
elasticnet_randomsearch
xgb_randomsearch
xgb_randomsearch
elasticnet_randomsearch
randomforest_randoms...
logreg_grid_search
randomforest_randoms...
logreg_grid_search

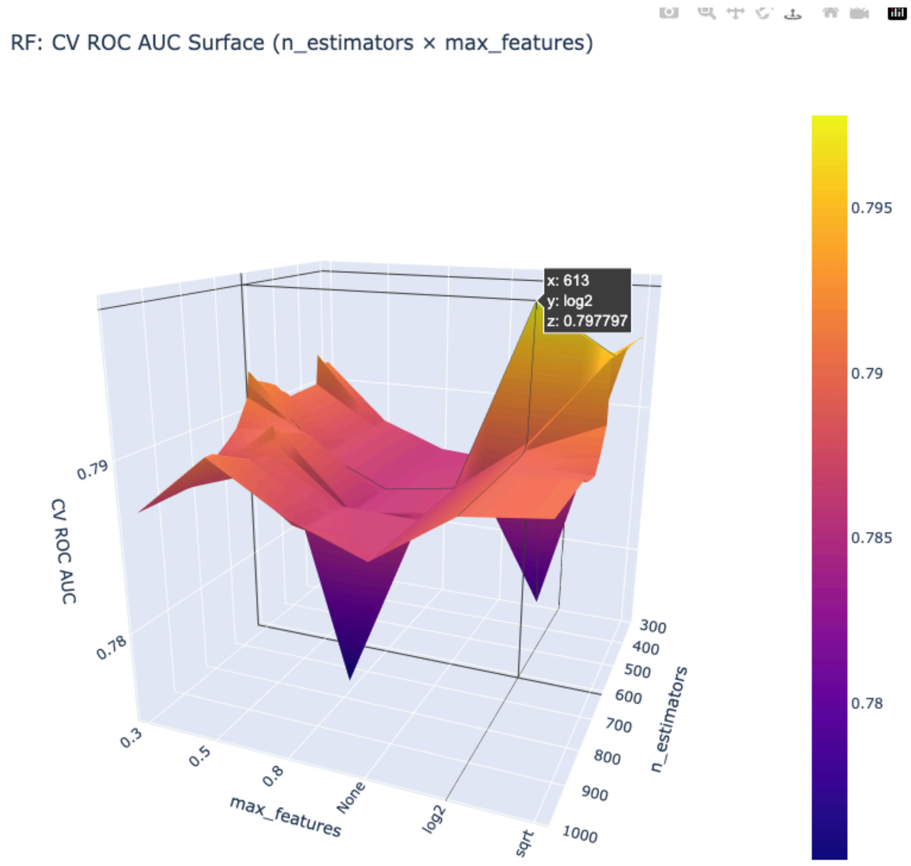
Appendix C: Pearson Correlation Heatmap

This heatmap illustrates the pairwise Pearson correlations among business attributes and restaurant categories in the BDA602 Yelp Project dataset. Strong positive correlations are observed among related service features such as ***Outdoor Seating****,* ***Good for Groups****, and* ***Restaurant Reservations****, reflecting logical operational relationships. Most variables exhibit weak or negligible correlations, indicating* ***low multicollinearity*** *and supporting the inclusion of a diverse set of features for robust model training.*



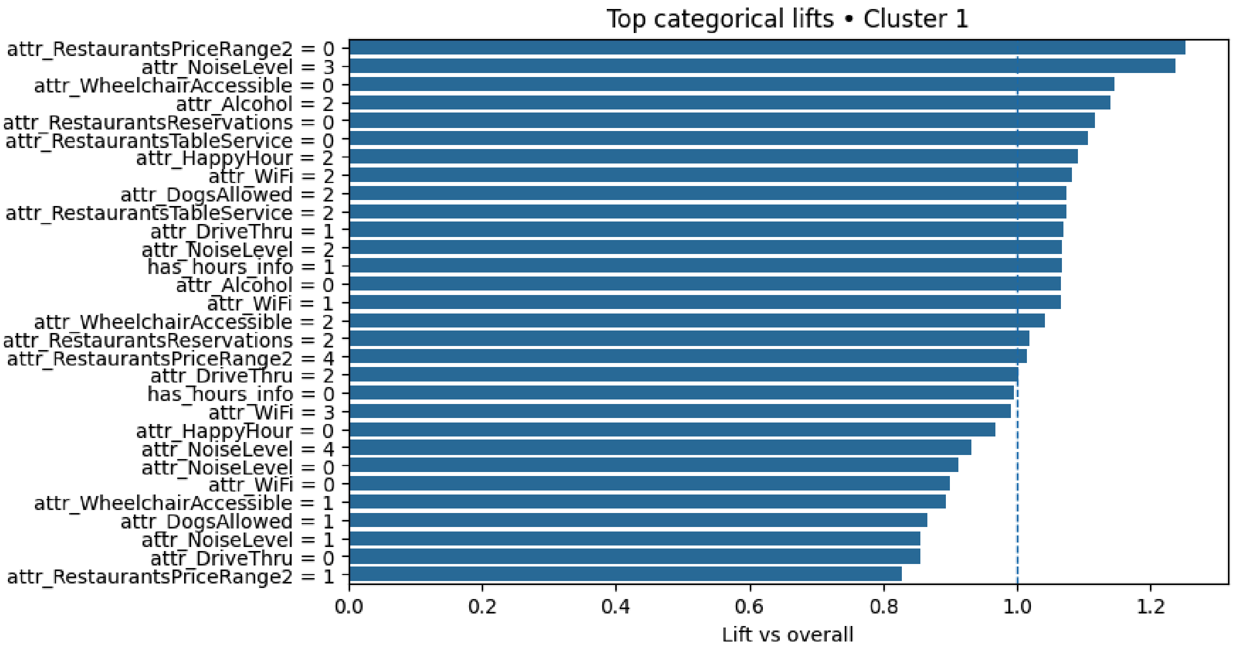
Appendix D: Random Forest Hyperparameter Tuning

This 3D surface plot visualizes the cross-validated ROC-AUC scores of the Random Forest model across varying values of `n_estimators` *and* `max_features`*. Model performance peaked at approximately* ***n_estimators ≈ 600*** *and* ***max_features = log2****, achieving a cross-validated* ***ROC-AUC of 0.798****, indicating an optimal balance between model complexity and generalization.*



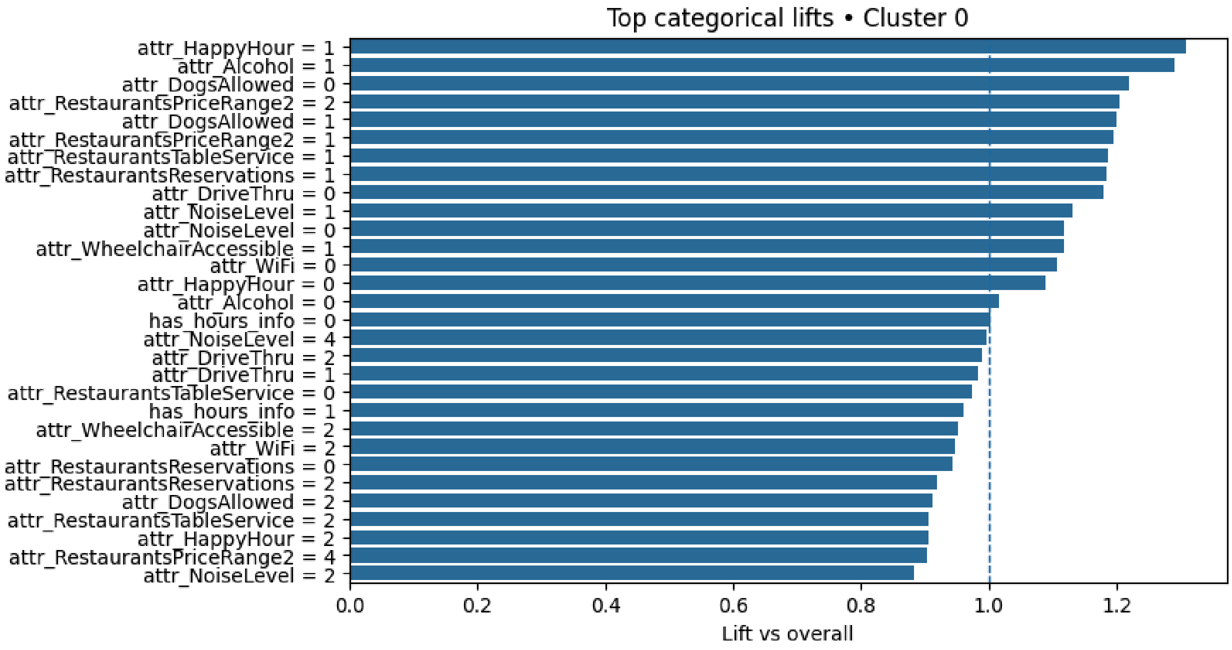
Appendix E. Top Categorical Lifts for Cluster 1 (K-Prototypes Clustering)

This figure displays the categorical attributes most overrepresented in Cluster 1 relative to the overall dataset. Restaurants in this cluster are characterized by low price range (PriceRange = 0–1), louder environments, limited accessibility and table service, and casual, family-oriented operational styles. These features align with the cluster's profile as budget-friendly, quick-service, or family-casual restaurants that prioritize value and convenience over premium amenities.



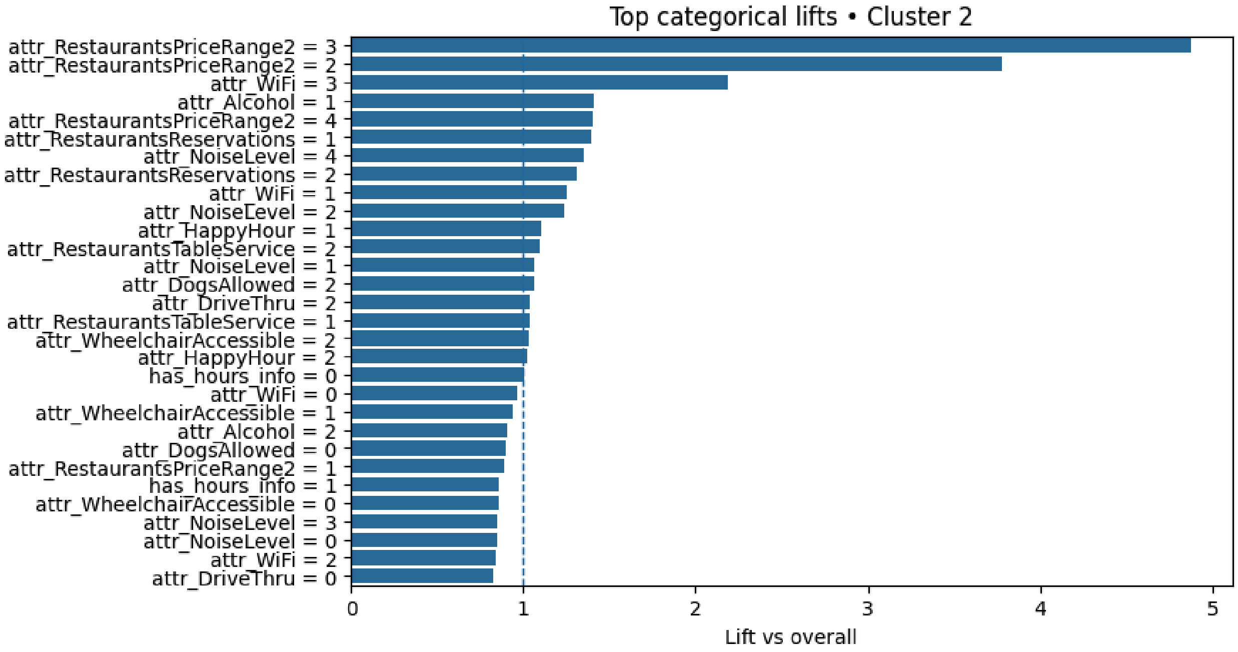
Appendix F. Top Categorical Lifts for Cluster 0 (K-Prototypes Clustering)

This figure illustrates the categorical attributes most overrepresented in Cluster 0 relative to the full dataset. Restaurants in this group are primarily ***mid-priced, social-dining establishments*** *offering* ***alcohol****,* ***happy-hour promotions****,* ***table service****, and* ***moderate noise levels****. These characteristics correspond to* ***casual, dine-in environments*** *that emphasize comfort and social interaction, aligning with the cluster's mid-range average rating performance.*



Appendix G. Top Categorical Lifts for Cluster 2 (K-Prototypes Clustering)

This figure displays the categorical attributes most overrepresented in Cluster 2 relative to the overall dataset. The cluster is dominated by ***mid- to high-priced, dine-in restaurants*** *that frequently provide* ***alcohol service****,* ***Wi-Fi****, and* ***reservations*** *within* ***quiet or moderately quiet settings****. Despite these upscale attributes, Cluster 2 exhibited the lowest mean ratings, suggesting that* ***premium features alone do not guarantee higher satisfaction*** *and that factors such as service consistency and perceived value may be more influential.*


Top categorical lifts • Cluster 2
attr_RestaurantsPriceRange2 = 3
attr_RestaurantsPriceRange2 = 2
attr_WiFi = 3
attr_Alcohol = 1
attr_RestaurantsPriceRange2 = 4
attr_RestaurantsReservations = 1
attr_NoiseLevel = 4
attr_RestaurantsReservations = 2
attr_WiFi = 1
attr_NoiseLevel = 2
attr_HappyHour = 1
attr_RestaurantsTableService = 2
attr_NoiseLevel = 1
attr_DogsAllowed = 2
attr_DriveThru = 2
attr_RestaurantsTableService = 1
attr_WheelchairAccessible = 2
attr_HappyHour = 2
has_hours_info = 0
attr_WiFi = 0
attr_WheelchairAccessible = 1
attr_Alcohol = 2
attr_DogsAllowed = 0
attr_RestaurantsPriceRange2 = 1
has_hours_info = 1
attr_WheelchairAccessible = 0
attr_NoiseLevel = 3
attr_NoiseLevel = 0
attr_WiFi = 2
attr_DriveThru = 0
Lift vs overall