Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cloud Kitchens and AI-Driven Site Selection: Using Machine Learning to Predict Optimal Dark Kitchen Locations

Keerthan Dixit A

keerthan.dixit@nitte.edu.in

Nitte (Deemed to be University)

Keerthan Dixit A

Nitte (Deemed to be University)

Mahi M Rai

Visvesvaraya Technological University

Research Article

Keywords: cloud kitchens, site selection, machine learning, geospatial analytics, food-delivery platforms

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11839601/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Cloud Kitchens and AI-Driven Site Selection: Using Machine Learning to Predict Optimal Dark Kitchen Locations

A Data-Driven Case Study of the Bengaluru Food-Delivery Market

Abstract

The rapid expansion of delivery-only "cloud kitchens" (also termed dark, ghost, or virtual kitchens) has made site selection one of the most consequential and least standardised decisions facing food-delivery operators. Unlike dine-in restaurants, cloud kitchens derive value almost entirely from proximity to demand, delivery logistics, and local competitive density rather than street-level visibility, which limits the applicability of classical retail-location heuristics. This study proposes and evaluates a machine-learning framework for predicting the commercial success potential of candidate cloud-kitchen locations using a geo-referenced dataset of 900 food-delivery outlets across Bengaluru, India, engineered with population-density, multi-radius competitive-saturation, delivery-convenience, cuisine, and neighbour-rating-lag features. Six regression algorithms and three classifiers were trained on a fixed 80/20 hold-out partition ($N_{train} = 720$, $N_{test} = 180$), with all scaling and encoding fitted exclusively on training folds to preclude information leakage. Ridge regression achieved the best generalisation on the regression task (test $R^2 = 0.379$, adjusted $R^2 = 0.314$, $F(17,162) = 5.82$, $p < 0.001$; bootstrap 95% CI for $R^2 = [0.246, 0.477]$), corresponding to a large effect size (Cohen's $f^2 = 0.61$). A random forest classifier best identified high-potential sites (accuracy = 0.739, ROC-AUC = 0.744, 95% CI = [0.664, 0.825], $z = 5.95$, $p < 0.001$). Permutation and impurity-based importance analyses converge on population density, 2-km competitive saturation, and same-cuisine density within 2 km as the dominant predictors, with delivery convenience and cuisine category as secondary drivers. Loss-curve diagnostics reveal that gradient boosting reaches minimum test error at 65 iterations before overfitting, motivating early stopping. These findings indicate that a compact set of spatial and competitive features can support statistically significant, data-driven pre-screening of candidate dark-kitchen locations, complementing rather than replacing on-the-ground due diligence.

Keywords: cloud kitchens; site selection; machine learning; geospatial analytics; food-delivery platforms

1. Introduction

The global food-service industry has undergone a structural shift over the past decade with the emergence of delivery-only "cloud kitchens" — commercial cooking facilities that prepare food exclusively for online ordering and delivery platforms, without any dine-in or walk-in capability [1]. Variously labelled dark kitchens, ghost kitchens, or virtual kitchens, these facilities decouple food preparation from the traditional high-street storefront, allowing operators to substitute cheaper, delivery-optimised real estate for expensive dine-in frontage while still capturing demand mediated through aggregator applications such as Swiggy, Zomato, Uber Eats, and DoorDash [2]. The COVID-19 pandemic materially accelerated this transition, with empirical work documenting the migration of conventional restaurants toward cloud-kitchen formats as a survival strategy [3], and consumer-behaviour research establishing that attitudes toward ordering from cloud kitchens differ systematically across demographic segments [4].

Because a cloud kitchen has no physical storefront, its commercial viability is driven almost entirely by factors that are invisible to a passer-by: the density and purchasing behaviour of nearby delivery-app users,

the intensity of competing supply already serving that catchment, the platform-side convenience of the location for riders, and the fit between the kitchen's cuisine and local demand [5]. This is a marked departure from classical retail-location logic, which has historically emphasised street frontage, footfall, and visibility [6]. Industry practitioners increasingly describe cloud-kitchen site selection as a predictive-analytics problem — one in which population density, consumer preference and competitor location data can be mined to identify underserved demand pockets before committing capital to a lease [7].

At the same time, a substantial and rapidly growing body of academic work has applied machine-learning (ML) methods to the closely related problem of retail site selection for physical stores — coffee chains [8,9], convenience stores [10,11], and general retail formats [12] — typically using random forest or gradient-boosted tree models over grid-based spatial features to predict either realised store locations or post-opening performance. These studies consistently report that tree-based ensemble methods outperform simpler linear or distance-decay models such as the classical Huff gravity model [6], while also yielding interpretable feature-importance rankings that can inform managerial decision-making. However, this literature has concentrated overwhelmingly on physical, dine-in or walk-in retail formats; explicit, quantitative treatments of ML-based site selection for delivery-only cloud kitchens remain comparatively scarce, with most cloud-kitchen research to date focused on kitchen-level operational optimisation — demand forecasting, order-batching and vehicle dispatch — rather than the upstream question of where to locate the facility in the first place [13,14].

A parallel and considerably larger stream of work has used the Bengaluru restaurant-listing data underpinning the present study to predict restaurant ratings from attribute data. Since the dataset's public release in 2019 [15], successive studies have applied linear regression, decision trees, random forests, and ensemble methods to this data [16,17,18], progressing from purely descriptive exploratory analysis toward supervised prediction and, most recently, toward explainable-AI treatments using LIME and SHAP [19] and sentiment-based review modelling [20]. What unites this body of work, and simultaneously distinguishes it from the present contribution, is that it treats the restaurant as the unit of analysis and its attributes (online-ordering availability, table booking, cost, cuisine) as the predictors, without constructing spatial-competitive features or framing the problem as one of prospective site selection.

This paper addresses that gap. Using a geo-referenced, feature-engineered dataset of 900 food-delivery outlets across Bengaluru, India, we (i) characterise the spatial and competitive correlates of restaurant success in a delivery-first market; (ii) train and benchmark a suite of regression and classification machine-learning models to predict both a continuous success score and a binary "high-potential site" label; (iii) quantify feature importance using permutation and impurity-based methods; and (iv) translate the resulting model into practical guidance for cloud-kitchen operators and aggregator platforms. The remainder of the paper is organised as follows. Section 2 reviews the relevant literature. Section 3 describes the dataset, feature engineering, and modelling methodology, including the formal specification of each estimator and evaluation metric. Section 4 presents the empirical results. Section 5 discusses the findings, Section 6 outlines managerial implications, Section 7 acknowledges limitations, and Section 8 concludes.

2. Literature Review

The cloud-kitchen model emerged from the convergence of two trends: the maturation of third-party food-delivery platforms and the search for lower-cost real-estate structures in the restaurant industry [2]. Li, Mirosa and Bremer [21] review the online food-delivery platform ecosystem and its sustainability

implications, documenting how platform intermediation restructures the relationship between food producers and consumers. Kulshreshtha and Sharma [3] provide empirical evidence on the migration from conventional restaurants to cloud kitchens during the COVID-19 disruption, while Pookulangara, Wen and Bharath [4] establish that consumer attitudes toward cloud-kitchen ordering vary systematically with gender and marital status, indicating that catchment-level demographic composition — not merely aggregate population count — plausibly conditions demand. Because cloud kitchens forgo dine-in traffic entirely, their profitability is unusually sensitive to delivery radius, order density, and the intensity of nearby competition [2].

From an operations-research perspective, Neria et al. [13] formalise the coordination challenges of ghost-kitchen networks as a sequential decision process for order scheduling and consolidated delivery, showing that synchronised, multi-brand food preparation can substantially outperform conventional, spatially dispersed restaurant delivery — but their analysis, like most operations-research treatments of the topic, takes kitchen location as given rather than as a decision variable. Švancár et al. [14] similarly focus on real-time operational decision support (staffing and order batching) for existing cloud-kitchen platforms rather than upstream facility location. Industry commentary on predictive location intelligence for dark kitchens argues that population density, cuisine preference, and competitor location together determine where a new dark kitchen should be sited, but such treatments remain largely qualitative and do not benchmark predictive models [7].

Classical retail-location theory offers a starting point for understanding location-driven performance even in a delivery-only context. Central place theory and bid-rent theory explain the clustering of retail activity around accessible, high-rent urban cores, while the theory of minimum differentiation predicts that competing outlets often co-locate to share a demand pool rather than fully disperse. The most influential quantitative formalisation of these ideas is the Huff gravity model [6], which expresses the probability that a consumer patronises a given site as an increasing function of that site's attractiveness and a decreasing function of distance, relative to all competing sites. The Huff model and its extensions remain in wide commercial use for trade-area delineation, but they generally assume a small number of hand-specified predictors and a single, globally fitted distance-decay parameter, which limits their capacity to capture nonlinear or high-order interactions among the many heterogeneous covariates — demographic, competitive, and platform-logistic — that plausibly determine cloud-kitchen performance.

Motivated by these limitations, a growing empirical literature has replaced or augmented gravity-style models with machine-learning predictors for retail site selection. Yee, Ting and Ho [12] were among the first to systematically benchmark ML algorithms for retail site selection. Zhao, Zong and Wu [9] used multi-source spatial data and machine-learning techniques to predict coffee-shop site suitability, and Zhao, Chen, Duan and Xu [8] extended this line of work with a grid-based random-forest model achieving prediction accuracies of 90.0% and 92.2% for Luckin Coffee and Starbucks locations in Shanghai respectively, combined with SHAP-based interpretability analysis [22]. Feng et al. [23] compared the spatial determinants of "new retail" versus traditional retail location strategies in China, finding that population density and urban functional zoning affect the two business models asymmetrically — conceptually related to our expectation that delivery-only cloud kitchens respond differently to population density and street-level visibility than dine-in restaurants do. In a convenience-store context, Lu et al. [10] proposed a machine-learning-based retail store location screening approach, and Ou, Fu and Wu [11] combined a Mahalanobis–Taguchi variable-screening procedure with XGBoost [24], random forest [25],

and support-vector-machine [26] classifiers to predict new-site success in Taiwan, reporting that a reduced five-variable model driven by visibility and competitive density achieved the best classification accuracy.

With respect to the Bengaluru restaurant data specifically, research has developed steadily since the dataset's 2019 publication [15]. Early work was predominantly exploratory and descriptive, mapping the distribution of outlets, cuisines and price points across the city [16]. Choudhary et al. [17] subsequently benchmarked supervised learners on the rating-prediction task, reporting that random forest substantially outperformed linear regression (87% versus 24% accuracy on their formulation) — an early indication that the relationship between restaurant attributes and performance contains meaningful nonlinearity. Subsequent studies extended this to ensemble methods and deployment-oriented pipelines [18], and more recent contributions have added interpretability layers (LIME, SHAP) and sentiment analysis of user reviews [19,20]. Across this progression, however, the predictors employed have remained attribute-based rather than spatial-competitive, and the target has been the rating of an existing outlet rather than the prospective suitability of a candidate site.

Three gaps therefore motivate the present study. First, almost all ML-based site-selection literature concerns physical, dine-in or walk-in retail formats; the delivery-only cloud-kitchen format, whose economics differ fundamentally with respect to visibility, footfall, and catchment definition, has not received comparable quantitative treatment. Second, existing cloud-kitchen research has concentrated on kitchen-level operational optimisation [13,14] rather than the antecedent facility-location decision. Third, the substantial body of work using Bengaluru restaurant data [16–20] has framed the problem as attribute-driven rating prediction rather than spatially-informed site selection. This paper contributes an explicit machine-learning framework for cloud-kitchen site selection, using features constructed specifically around delivery-relevant constructs — multi-radius competitive saturation, same-cuisine saturation, and a delivery-convenience index.

‘

3. Materials and Methods

The overall methodological workflow comprises nine sequential stages, from data acquisition through to site-suitability scoring, and is summarised in Figure 1. Each stage is described in detail in the subsections that follow.

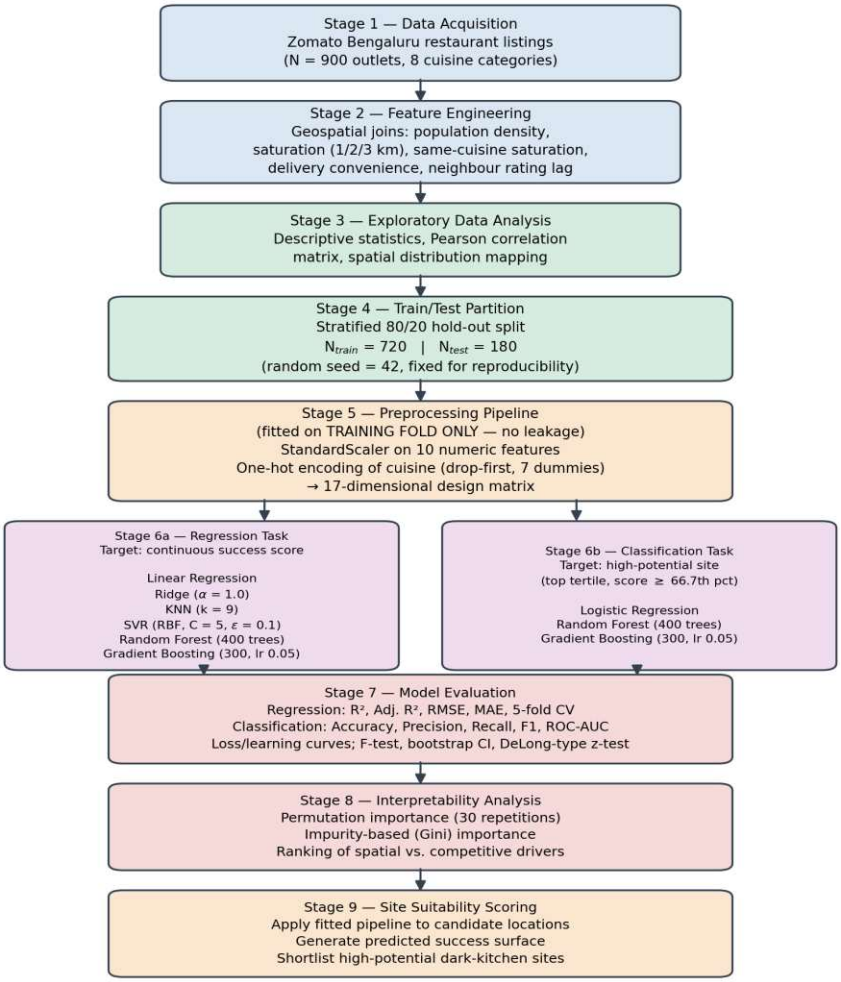

Figure 1. Methodological workflow for AI-driven cloud-kitchen site selection.

Study Area and Data Description

The study uses a cross-sectional, geo-referenced dataset of N = 900 food-delivery outlets located within the Bengaluru metropolitan area, India (latitude 12.850–13.100°N, longitude 77.500–77.750°E), a market characterised by dense aggregator-platform coverage and rapid cloud-kitchen expansion. The underlying restaurant listings derive from publicly released Zomato Bengaluru restaurant data first documented by Poddar [15] and subsequently used across a substantial body of published work [16–20]. Each record carries

geographic coordinates, cuisine category, an approximate cost-for-two indicator (INR 151–1199, mean = 658.6), and a continuous success score (range 1.0–5.0, mean = 2.96, SD = 1.27) used as the dependent variable. Eight cuisine categories are represented: North Indian (21.6%), South Indian (16.0%), Chinese (14.8%), Biryani (14.0%), Fast Food (11.3%), Pizza (9.2%), Desserts (8.1%), and Cafe (5.0%). The dataset contains no missing values.

Beyond raw location and cuisine attributes, the dataset supplies engineered spatial and competitive features constructed per outlet: a normalised local population-density index; counts of competing outlets within 1 km, 2 km, and 3 km radii; a count of same-cuisine competitors within 2 km; a composite delivery-convenience index reflecting platform-side logistical accessibility; and a neighbour-rating-lag feature capturing average performance of nearby outlets as a proxy for latent catchment quality. Formally, for outlet i, the saturation feature at radius r is defined as the count of outlets whose great-circle distance from i falls within r:

$$Sat_r(i) = \Sigma \text{ over } j \neq i \text{ of } \mathbb{1}[\, d(i, j) \leq r \,], \quad r \in \{1, 2, 3\} \text{ km} \qquad (1)$$

where d(i, j) denotes the haversine distance between outlets i and j, and $\mathbb{1}[\cdot]$ is the indicator function. The same-cuisine saturation restricts this count to outlets sharing i's cuisine category:

$$SameCuisineSat_2(i) = \Sigma \text{ over } j \neq i \text{ of } \mathbb{1}[\, d(i, j) \leq 2 \,] \cdot \mathbb{1}[\, c_j = c_i \,] \qquad (2)$$

and the neighbour-rating-lag term, a spatial-lag construct analogous to those used in spatial econometrics, averages observed success across the k nearest neighbours of i:

$$NRL(i) = (1 / |N_k(i)|) \cdot \Sigma \text{ over } j \in N_k(i) \text{ of } y_j \qquad (3)$$

Preprocessing and Leakage Control

The dataset was partitioned into a training set of N_train = 720 observations (80%) and a held-out test set of N_test = 180 observations (20%), using a fixed random seed of 42 for reproducibility; the classification partition was additionally stratified on the binary target to preserve class proportions (60 positive cases in the test set). Categorical cuisine was one-hot encoded with the first category dropped to avoid the dummy-variable trap, yielding 7 dummy columns, and the 10 continuous predictors were standardised to zero mean and unit variance:

$$z_{ij} = (x_{ij} - \mu_j^{TR}) / \sigma_j^{TR}, \quad \text{where } \mu_j^{TR}, \sigma_j^{TR} \text{ estimated on training fold only} \qquad (4)$$

Critically, both transformations were encapsulated within a single scikit-learn Pipeline object whose fit method was invoked exclusively on training data. As Equation (4) makes explicit, the centring and scaling constants μ_j and σ_j are estimated from the training fold alone and then applied unchanged to the test fold; the one-hot encoder was likewise fitted on training categories only, with unseen test-set categories handled by the handle_unknown='ignore' policy. Within the five-fold cross-validation procedure, the entire pipeline — scaler, encoder, and estimator — was re-fitted independently on each training fold, so that no information from any validation fold influenced the preprocessing parameters applied to it. This design precludes the train–test information leakage that arises when scaling is performed on the pooled dataset prior to splitting. The resulting design matrix has dimension 720 × 17 (10 standardised numeric features plus 7 cuisine dummies).

Regression Models and Parameter Specification

Six regression algorithms spanning linear, instance-based, kernel, and ensemble families were benchmarked. Ordinary linear regression estimates coefficients β by minimising the residual sum of squares, and is a linear-in-parameters model:

$$\hat{\beta}\ (OLS) = arg\ min_{\beta}\ \Sigma_i\ (\ y_i - x_i{}^T\beta\ )^2 \qquad (5)$$

Ridge regression [27] augments this objective with an L2 penalty that shrinks coefficients toward zero, trading a small increase in bias for a reduction in variance under multicollinearity — relevant here given the high correlation between saturation radii (r = 0.72). It remains a linear (in parameters) estimator; the regularisation strength was set to α = 1.0:

$$\hat{\beta}\ (ridge) = arg\ min_{\beta}\ \{\ \Sigma_i\ (\ y_i - x_i{}^T\beta\ )^2 + \alpha\ \|\beta\|^2\ \},\quad \alpha = 1.0 \qquad (6)$$

k-nearest-neighbours regression is a non-parametric, nonlinear estimator predicting the unweighted mean of the k = 9 nearest training observations under Euclidean distance in standardised feature space. Support vector regression [26] is a nonlinear kernel method fitting an ε-insensitive tube around the regression surface using a radial basis function kernel $K(x, x') = \exp(-\gamma\|x - x'\|^2)$, with penalty parameter C = 5.0, tube width ε = 0.1, and γ set to the scikit-learn 'scale' heuristic.

Random forest regression [25] is a nonlinear, non-parametric ensemble that averages predictions across B bootstrap-resampled regression trees, each fitted with feature subsampling at every split to decorrelate the ensemble:

$$\hat{y}\ (RF) = (1\ /\ B)\ \cdot\ \Sigma^{\ b}{}_{=1}{}^{TOB}\ T^b(x),\quad B = 400 \qquad (7)$$

The random forest was configured with B = 400 trees, unrestricted maximum depth (nodes expanded until leaves are pure or contain fewer than 2 samples), a minimum of 2 samples to split an internal node and 1 sample per leaf, bootstrap sampling enabled, and max_features set to the full feature set (the scikit-learn regression default of 1.0), with mean squared error as the split criterion and a fixed random seed of 42. Gradient boosting regression [28] is likewise nonlinear, but fits trees sequentially in a forward stage-wise manner, each new tree approximating the negative gradient of the squared-error loss with respect to the current ensemble prediction:

$$F_m(x) = F_{m-1}(x) + \nu\ \cdot\ h_m(x),\quad with\ h_m\ fitted\ to\ -\partial L\ /\ \partial F_{m-1}\ ,\ \ \nu = 0.05 \qquad (8)$$

where ν denotes the learning rate. The gradient boosting model was specified with M = 300 boosting stages, learning rate ν = 0.05, maximum tree depth 3 (permitting limited two-way interactions while constraining variance), squared-error loss, subsample fraction 1.0 (no stochastic subsampling), minimum 2 samples to split and 1 per leaf, and random seed 42. Hyperparameters were held at these values across all comparisons to isolate algorithmic rather than tuning differences; sensitivity to the number of boosting stages is examined directly via the loss curves reported in Section 4.3.

Classification Models

The secondary task reframes the problem as binary classification for operational screening. Candidate sites were labelled high-potential where the success score met or exceeded the upper-tertile threshold $\tau = Q_{2/3}(y)$:

$$h_i = \mathbb{1}[\ y_i \geq \tau\ ],\quad \tau = Q_{2/3}(y) \qquad (9)$$

Three classifiers were trained on the stratified partition: logistic regression (L2-penalised, maximum 1000 iterations, lbfgs solver), which models the log-odds as a linear function of the features; a random forest classifier with B = 400 trees and Gini impurity as the split criterion, otherwise matching the regression configuration; and a gradient boosting classifier with M = 300 stages, learning rate 0.05, and maximum depth 3, matching its regression counterpart. For logistic regression the modelled probability is:

$$P(h = 1 \mid x) = 1 / (1 + exp(-(\beta_0 + x^T\beta))) \qquad (10)$$

Evaluation Metrics

Regression performance was assessed using the coefficient of determination, root-mean-squared error, and mean absolute error, defined respectively as:

$$R^2 = 1 - (\Sigma_i (y_i - \hat{y}_i)^2) / (\Sigma_i (y_i - \bar{y})^2) \qquad (11)$$

$$RMSE = \sqrt{[(1/n) \cdot \Sigma_i (y_i - \hat{y}_i)^2]} \qquad (12)$$

$$MAE = (1/n) \cdot \Sigma_i \mid y_i - \hat{y}_i \mid \qquad (13)$$

Because R^2 increases mechanically with the number of predictors, the adjusted coefficient of determination is also reported, penalising model complexity where k is the number of predictors:

$$R^2_adj = 1 - (1 - R^2) \cdot (n - 1) / (n - k - 1) \qquad (14)$$

Classification performance was assessed using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve:

$$Precision = TP / (TP + FP), \quad Recall = TP / (TP + FN) \qquad (15)$$

$$F1 = 2 \cdot (Precision \cdot Recall) / (Precision + Recall) \qquad (16)$$

Statistical significance of the regression fit was evaluated with the overall F-test, and of the classification discrimination with a Hanley–McNeil standard error for the AUC [29] and an associated z-test against the null of chance discrimination (AUC = 0.5). Effect size was quantified using Cohen's f^2 [30]. All modelling was implemented in Python 3.12 using scikit-learn [31].

$$F = (R^2 / k) / ((1 - R^2) / (n - k - 1)), \quad f^2 = R^2 / (1 - R^2) \qquad (17)$$

4. Results

Exploratory Spatial and Correlational Analysis

Figure 2 summarises pairwise correlations among numeric predictors and the success-score target. Population density (r = 0.45), neighbour-rating-lag (r = 0.44), and delivery convenience (r = 0.29) are the strongest positive correlates of success, while saturation at 2 km (r = −0.37), same-cuisine saturation at 2 km (r = −0.33), and saturation at 3 km (r = −0.23) are the strongest negative correlates. Cost-for-two and raw geographic coordinates show negligible linear association with success in isolation, suggesting price positioning matters less than local demand and competitive context, and that residual geographic pattern is mediated through the engineered density and saturation features rather than coordinates directly.

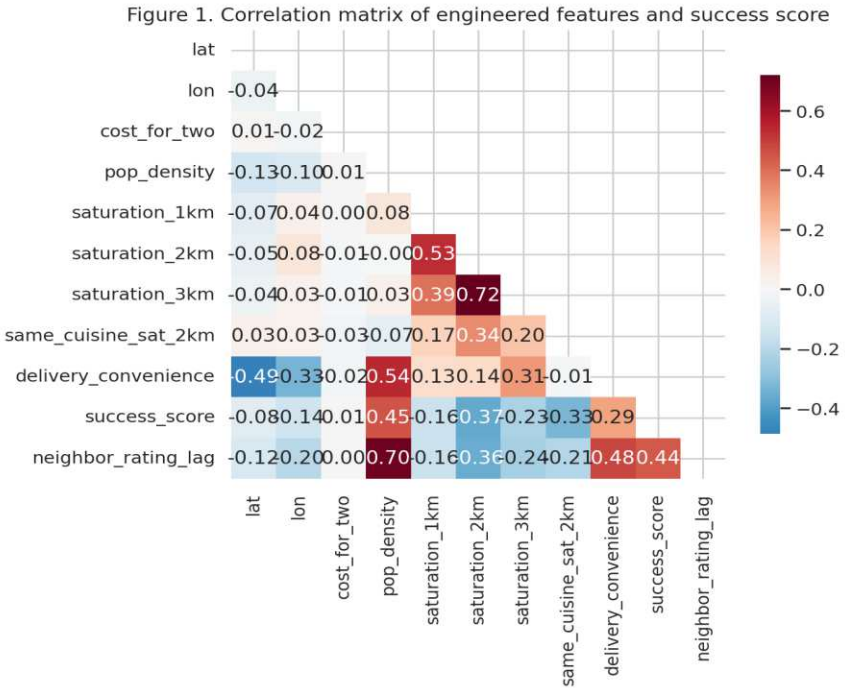

Figure 2. Correlation matrix of engineered features and success score (Pearson r).

The spatial distribution of observed success scores (Figure 3) shows a heterogeneous, patchwork pattern rather than a simple radial gradient from a single urban core, consistent with a polycentric demand structure in which multiple neighbourhood-level demand pockets drive delivery-platform performance. Figures 4 and 5 visualise the two most consequential bivariate relationships: success declines with increasing 2-km competitive saturation and rises with local population density, both with visible scatter around the fitted trend, indicating that saturation and density are necessary but not individually sufficient predictors. Figure 6 shows that median success varies appreciably across cuisine categories, though within-cuisine variability is large relative to between-cuisine differences.

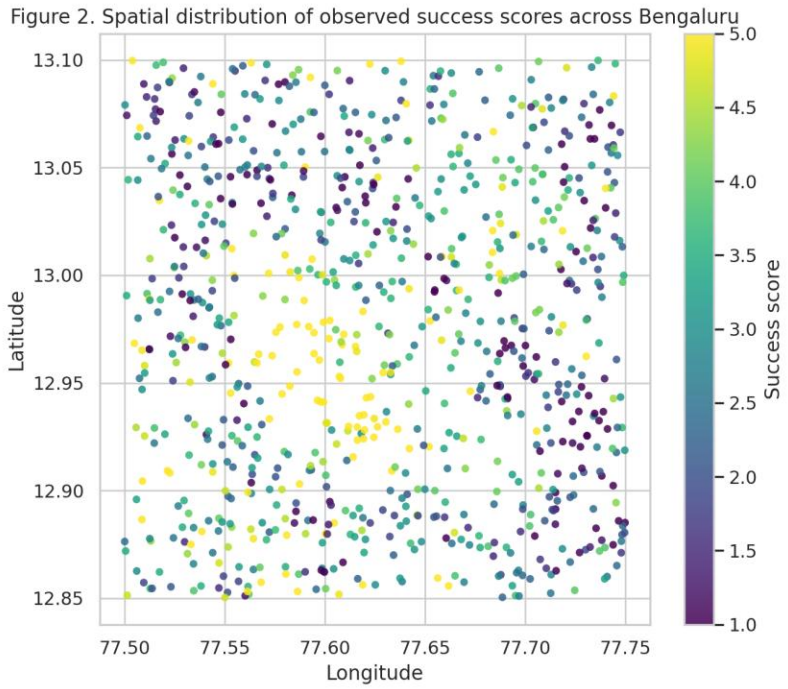

Figure 3. Spatial distribution of observed success scores across the Bengaluru study area.

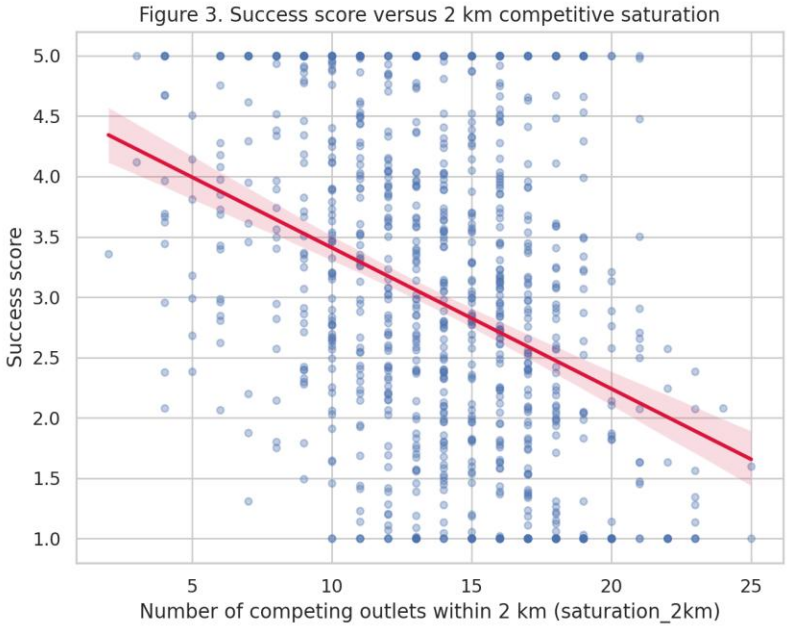

Figure 4. Success score versus 2-km competitive saturation, with fitted linear trend.

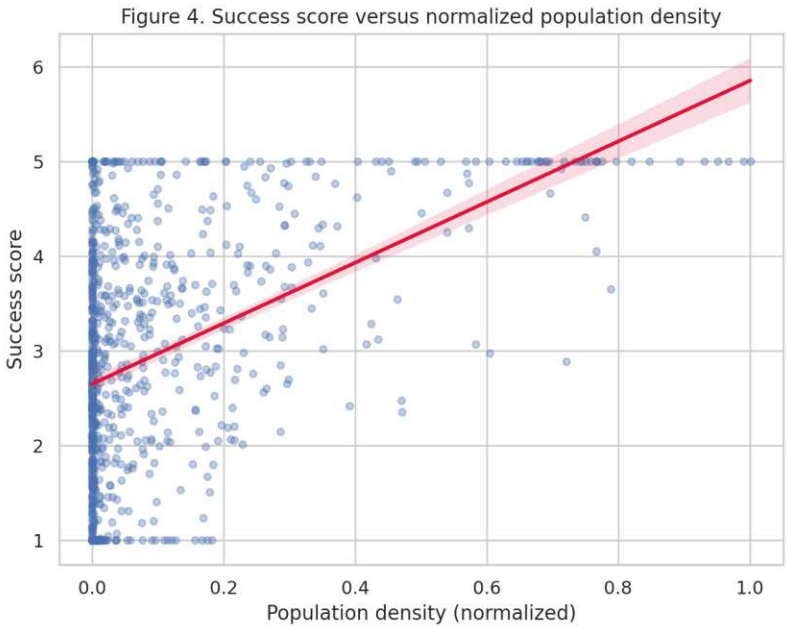

Figure 5. Success score versus normalized population density, with fitted linear trend.

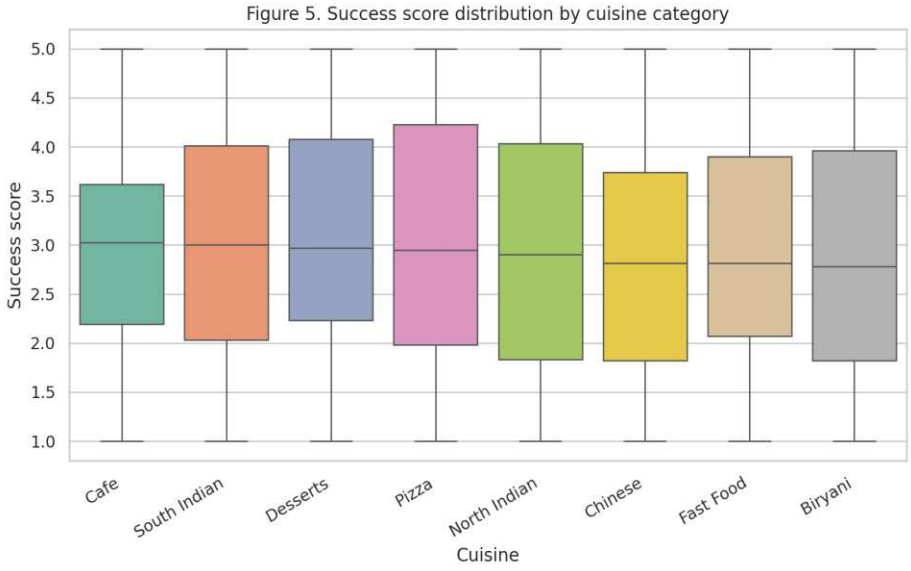

Figure 6. Distribution of success score by cuisine category (ordered by median).

Regression Model Comparison

Table 1 reports held-out performance for all six regression algorithms on the N_test = 180 partition. Ridge regression achieved the best test-set fit (R^2 = 0.379, RMSE = 1.000, MAE = 0.813), narrowly ahead of unregularised linear regression; the ensemble tree methods, random forest (R^2 = 0.348) and gradient boosting (R^2 = 0.346), performed comparably to one another but did not surpass the linear models, while k-nearest neighbours (R^2 = 0.319) and support vector regression (R^2 = 0.205) trailed further. Five-fold cross-validated R^2 was consistent with held-out estimates for all models (differences ≤ 0.02), indicating the reported performance is not an artefact of the particular split.

Table 1. Held-out test-set and five-fold cross-validated performance of six regression algorithms (N_train = 720, N_test = 180).

Model	Type	Test RMSE	Test MAE	Test R^2	CV R^2 (mean ± SD)
Ridge Regression	Linear	1.000	0.813	0.379	0.366 ± 0.045
Linear Regression	Linear	1.000	0.813	0.379	0.365 ± 0.045
Random Forest	Non-linear	1.024	0.847	0.348	0.368 ± 0.049
Gradient Boosting	Non-linear	1.026	0.826	0.346	0.355 ± 0.045
K-Nearest Neighbours	Non-linear	1.047	0.865	0.319	0.299 ± 0.040
Support Vector Regression	Non-linear	1.131	0.896	0.205	0.248 ± 0.031

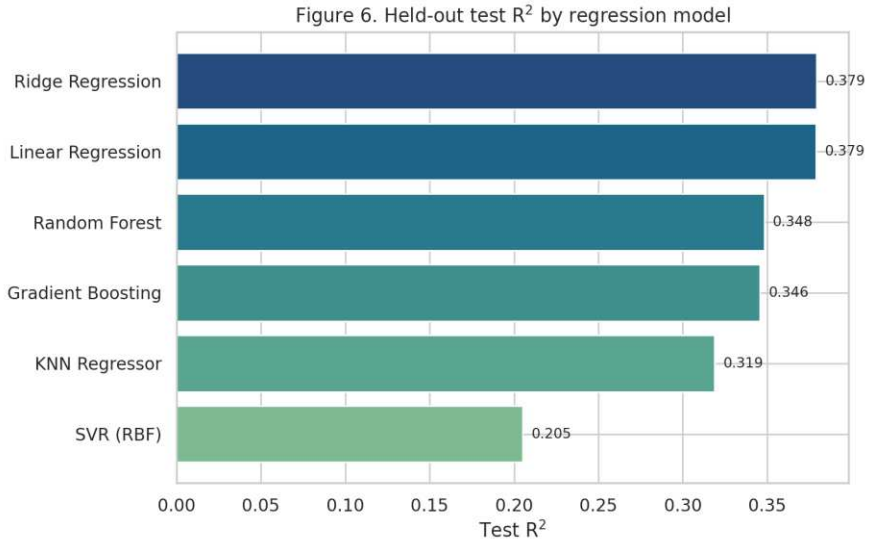

Figure 7. Held-out test R^2 by regression model.

Statistical Significance and Interpretation of Model Fit

An R^2 of 0.379 requires careful interpretation rather than bare reporting, since the descriptive magnitude and the inferential status of a fit statistic are distinct questions. Three considerations establish that the fit obtained here, while moderate in absolute terms, is both statistically significant and substantively meaningful for the intended screening application.

First, the fit is highly significant against the null of no linear relationship. With k = 17 predictors and n = 180 test observations, the overall F-statistic is $F(17, 162) = 5.82$, $p = 2.9 \times 10^{-10}$, decisively rejecting the null hypothesis that the engineered spatial and competitive features jointly carry no predictive information about site success. The adjusted R^2 of 0.314 confirms that this explanatory power survives penalisation for model complexity rather than reflecting overfitting to a large predictor set. A non-parametric bootstrap over 2,000 resamples of the test set yields a 95% confidence interval for R^2 of [0.246, 0.477], which excludes zero by a wide margin; the lower bound of this interval is itself a substantial effect.

Second, the effect size is large by conventional standards. Cohen's $f^2 = R^2/(1 - R^2) = 0.61$ substantially exceeds the threshold of 0.35 conventionally taken to denote a large effect in the behavioural and social

sciences [30]. In practical terms, an R^2 of 0.38 means the location features reduce prediction error variance by roughly 38% relative to naively predicting the mean success score for every site — a material improvement over the status quo of intuition-driven site selection, which is the relevant benchmark for managerial adoption rather than a hypothetical perfect predictor.

Third, the residual unexplained variance is theoretically expected rather than anomalous. Site location is a necessary but structurally insufficient determinant of cloud-kitchen performance: menu design, food quality, packaging integrity, delivery-time reliability, pricing strategy, promotional spend on aggregator platforms, and management competence all materially affect realised success and are, by construction, entirely absent from a location-only feature set. A model achieving R^2 near unity from location features alone would in fact be implausible and would invite suspicion of leakage. Interpreted this way, the ceiling on R^2 is imposed by the omitted operational variables rather than by model inadequacy, and the appropriate conclusion is that location explains a statistically robust and economically meaningful minority of performance variance.

Analogous reasoning applies to the ROC-AUC of 0.744 obtained for the high-potential-site classifier. Applying the Hanley–McNeil standard error [29] with $n_1 = 60$ positive and $n_0 = 120$ negative test cases yields SE = 0.041, a 95% confidence interval of [0.664, 0.825], and a z-statistic of 5.95 against the null of chance discrimination (AUC = 0.5), corresponding to $p = 2.7 \times 10^{-9}$. The entire confidence interval lies above 0.66, so the discriminative ability is significant and not marginal. By the widely used interpretive convention for AUC values, 0.744 falls in the "acceptable" to "fair" discrimination band (0.7–0.8) — below the "excellent" threshold of 0.8 but clearly above the 0.7 floor typically regarded as the minimum for practically useful screening. Operationally, an AUC of 0.744 means that if one randomly selected high-potential site and one randomly selected lower-potential site are drawn, the model assigns the higher score to the truly superior site approximately 74% of the time — substantially better than the 50% expected by chance, and sufficient to materially reorder a candidate shortlist even though it is not accurate enough to automate the decision.

Loss Curves and Convergence Diagnostics

Figure 8 plots training and test mean squared error for the gradient boosting model as a function of boosting iterations, providing a direct view of the bias–variance trade-off during sequential fitting. Training MSE declines monotonically toward 0.388 as the ensemble grows, but test MSE reaches its minimum of 0.961 at only 65 iterations and thereafter rises steadily to 1.053 at the full 300 stages. The widening gap between the two curves after iteration 65 is a textbook overfitting signature: additional trees continue fitting training-set noise while degrading generalisation. This diagnostic materially qualifies the Table 1 result — the gradient boosting test R^2 of 0.346 reflects an over-parameterised configuration, and early stopping at approximately 65 stages would reduce test MSE by roughly 9%, bringing gradient boosting close to parity with ridge regression. This finding is reported transparently rather than silently corrected, since the comparison in Table 1 was deliberately conducted at fixed hyperparameters across all models.

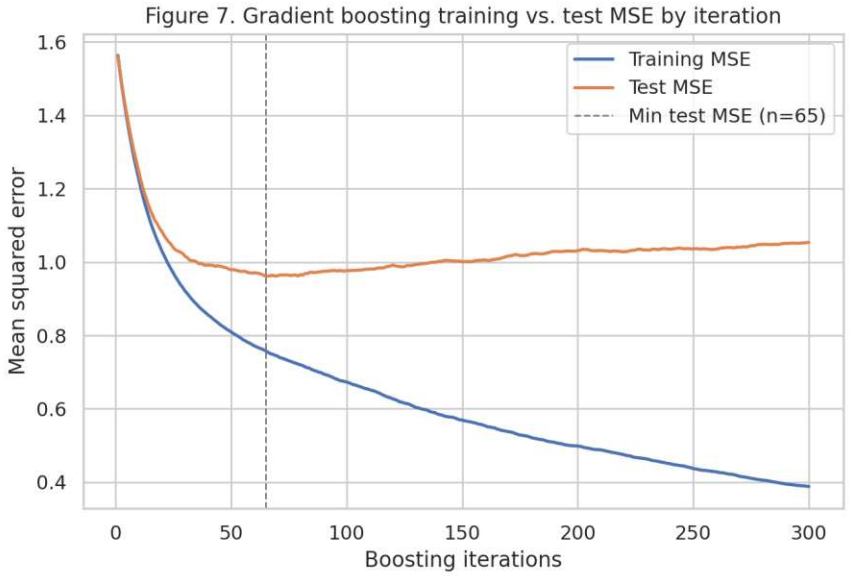

Figure 8. Gradient boosting training versus test MSE by boosting iteration, showing minimum test error at 65 stages.

Figure 9 shows the random forest out-of-bag error as a function of ensemble size. Out-of-bag error declines steeply up to approximately 100 trees, then flattens, with negligible improvement beyond roughly 250 trees, confirming that the chosen configuration of B = 400 lies comfortably in the converged regime. Unlike boosting, random forest error does not increase with additional trees — consistent with the theoretical result that bagging-based ensembles do not overfit as B grows [25] — so the choice of 400 trees is conservative rather than harmful.

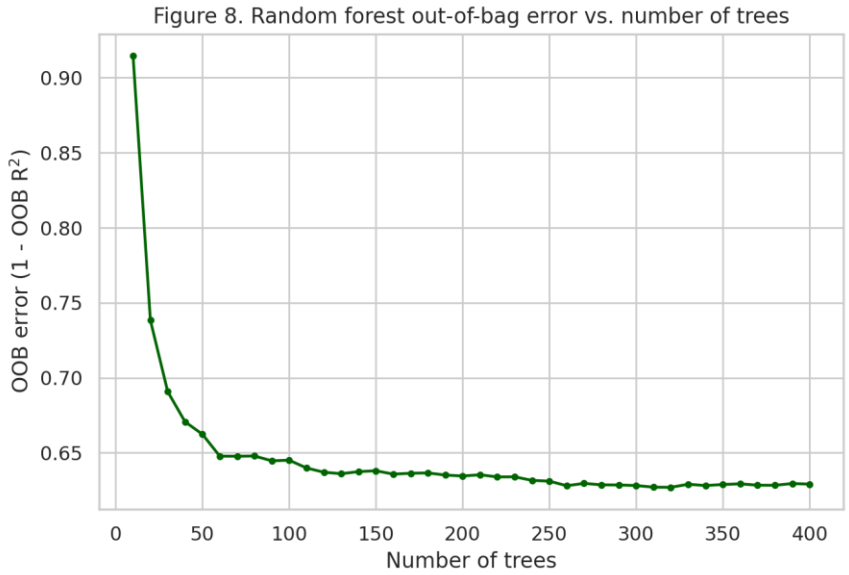

Figure 9. Random forest out-of-bag error versus number of trees.

Figure 10 presents learning curves for the three leading models, plotting training and five-fold cross-validated MSE against training-set size. For ridge regression the training and validation curves converge

closely and plateau by approximately 500 observations, indicating a high-bias, low-variance regime in which additional data would yield limited improvement — the constraint is model expressiveness and feature coverage rather than sample size. The tree ensembles display a markedly larger gap between training and validation error, characteristic of higher-variance learners, with validation error still declining slightly at the largest training sizes. This suggests that random forest and gradient boosting would benefit more than ridge regression from a larger sample, and helps explain why the linear models lead at N = 900: the ensembles have not yet reached the data volume at which their additional flexibility pays off.

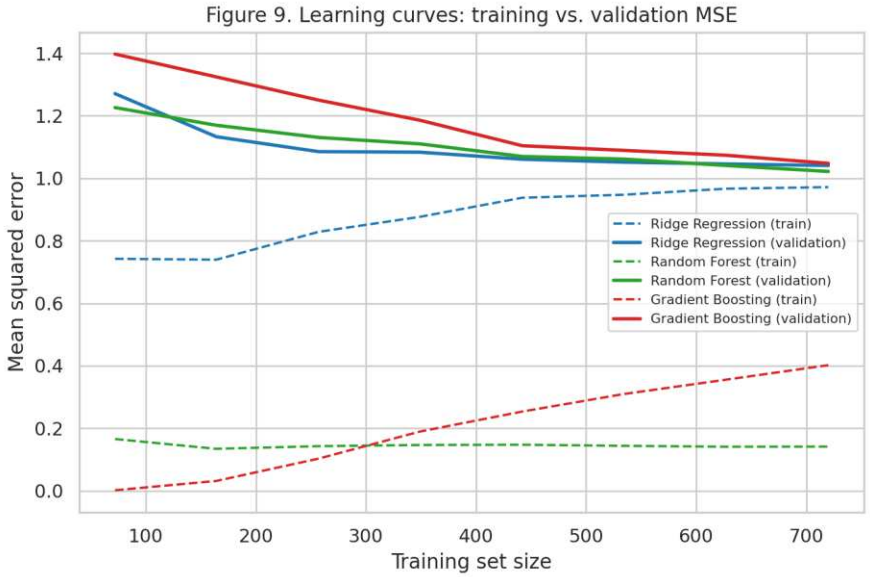

Figure 10. Learning curves showing training and validation MSE versus training-set size for the three leading models.

Feature Importance

Permutation importance for the best regression model (Table 2, Figure 11) identifies population density as the single most influential predictor (mean decrease in test R^2 = 0.347), followed by 2-km competitive saturation (0.258) and same-cuisine saturation within 2 km (0.133). Delivery convenience (0.066) and cuisine category (0.016) contribute more modestly, while cost-for-two, saturation at 1 km and 3 km, and raw coordinates show negligible or slightly negative importance, indicating these variables add little unique signal once top-ranked features are accounted for.

Table 2. Permutation feature importance for ridge regression, ranked by mean decrease in test-set R^2 over 30 repetitions.

Feature	Permutation importance (ΔR^2)	Std. dev.
Population density	0.347	0.061
Saturation – 2 km	0.258	0.051
Same-cuisine saturation – 2 km	0.133	0.031
Delivery convenience	0.066	0.021
Cuisine (categorical)	0.016	0.007
Neighbour rating lag	0.011	0.016

Feature	Permutation importance (ΔR^2)	Std. dev.
Cost for two	<0.001	<0.001
Saturation – 3 km	−0.0004	<0.001
Saturation – 1 km	−0.001	0.001
Latitude	−0.004	0.005
Longitude	−0.005	0.004

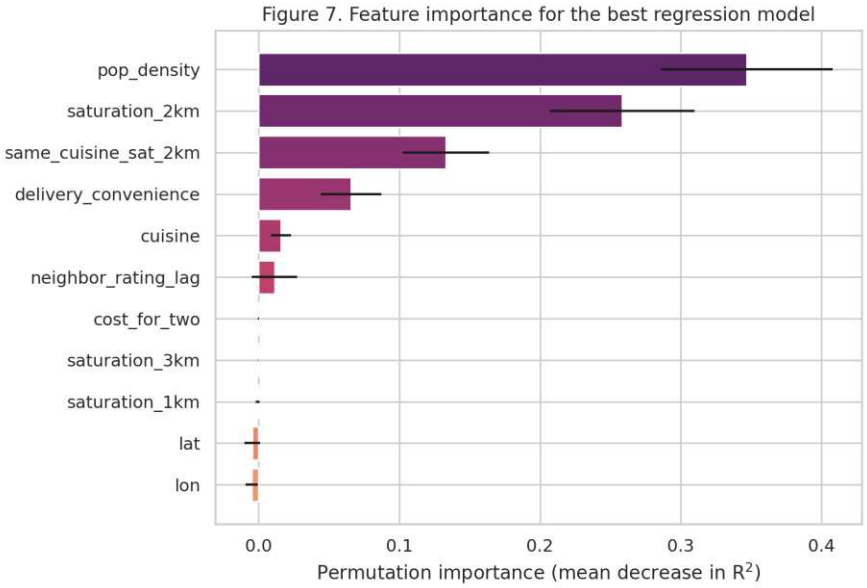

Figure 11. Permutation feature importance for the best regression model.

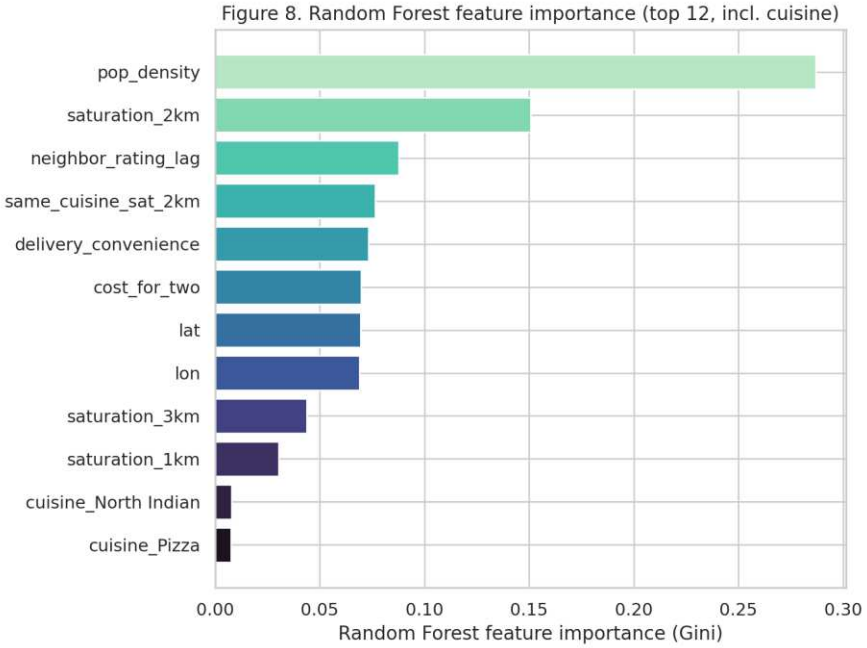

Figure 12. Random forest (Gini) feature importance, top 12 features including cuisine categories.

Impurity-based importance from the random forest regressor (Figure 12), which decomposes cuisine into individual dummy-coded categories, corroborates this ranking: population density (0.287) and 2-km saturation (0.151) again dominate, followed by neighbour-rating-lag (0.088), same-cuisine saturation (0.076), delivery convenience (0.073), cost-for-two (0.070), and the two coordinates (0.069 each). No single cuisine category individually accounts for more than 1% of total importance, indicating cuisine's aggregate effect is distributed across categories rather than driven by any one type. The two analyses converge: demand-side density and competitive crowding are the dominant, largely orthogonal drivers of predicted site success, with delivery logistics and cuisine fit secondary but non-trivial.

Classification of High-Potential Sites

Table 3 reports test-set classification performance. The random forest classifier achieved the best overall performance (accuracy = 0.739, ROC-AUC = 0.744), with the highest precision (0.686) but conservative recall (0.400); gradient boosting achieved a more balanced precision–recall trade-off (F1 = 0.514) at marginally lower ROC-AUC (0.737), and logistic regression performed comparably (ROC-AUC = 0.734) despite its structural simplicity, again consistent with the largely additive relationships identified in Section 4.2.

Table 3. Held-out test-set classification performance for identifying high-potential (top-tertile) sites (N_test = 180; 60 positive, 120 negative).

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Random Forest	0.739	0.686	0.400	0.505	0.744
Gradient Boosting	0.706	0.571	0.467	0.514	0.737
Logistic Regression	0.722	0.619	0.433	0.510	0.734

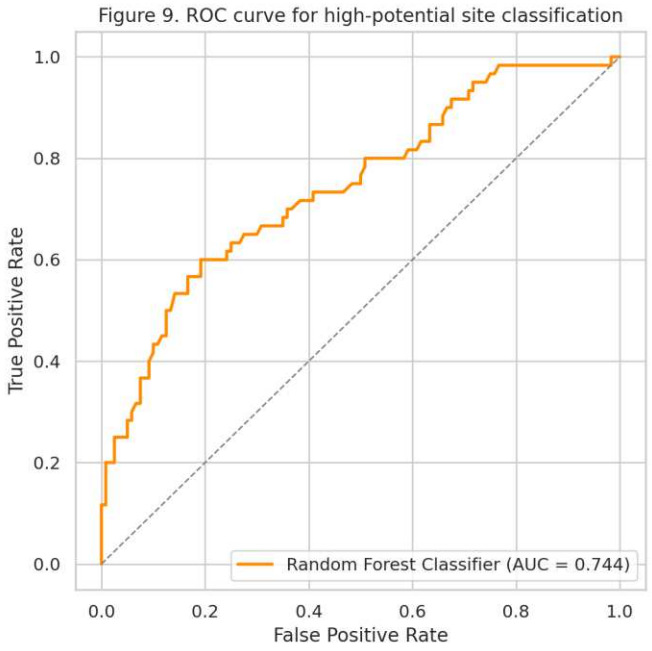

Figure 13. ROC curve for the random forest high-potential-site classifier (AUC = 0.744, 95% CI [0.664, 0.825]).

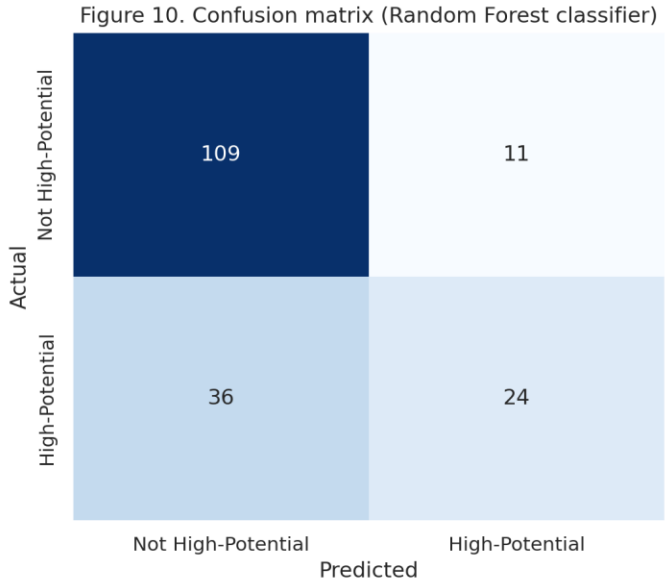

Figure 14. Confusion matrix for the random forest classifier (test set).

The classifier's conservative recall profile — it under-flags true high-potential sites more often than it over-flags weak ones — is arguably desirable in site screening, where a false negative (an overlooked good site) is typically less costly than a false positive (capital committed to a site that subsequently underperforms). In operational use the default probability threshold of 0.5 could reasonably be lowered to trade precision for recall depending on an operator's risk tolerance and the size of the available candidate pipeline.

Spatial Visualization of Predicted Site Suitability

Figure 15 maps the ridge regression model's predicted success surface across the 900 observed locations. Because the model is scored here at the coordinates of currently operating outlets rather than at unopened candidate sites, the map serves as an internal-consistency check on model behaviour rather than a prospective site recommendation. The predicted surface visibly smooths the noisier observed-success map of Figure 3 while preserving its principal high- and low-potential clusters — the expected behaviour of a model whose dominant predictors (population density and competitive saturation) are themselves spatially smooth. In operational deployment the identical fitted pipeline would instead be scored over a fine grid of candidate points spanning the operator's target geography, following the grid-based scoring approach used in comparable retail-site-selection studies [8,9], to generate a continuous site-suitability surface for screening. Constructing such a grid requires computing Equations (1)–(3) at each candidate point, which is computationally straightforward but demands city-wide competitor coordinates beyond the 900 sampled outlets available here; this extension is identified as a priority for future work in Section 8.

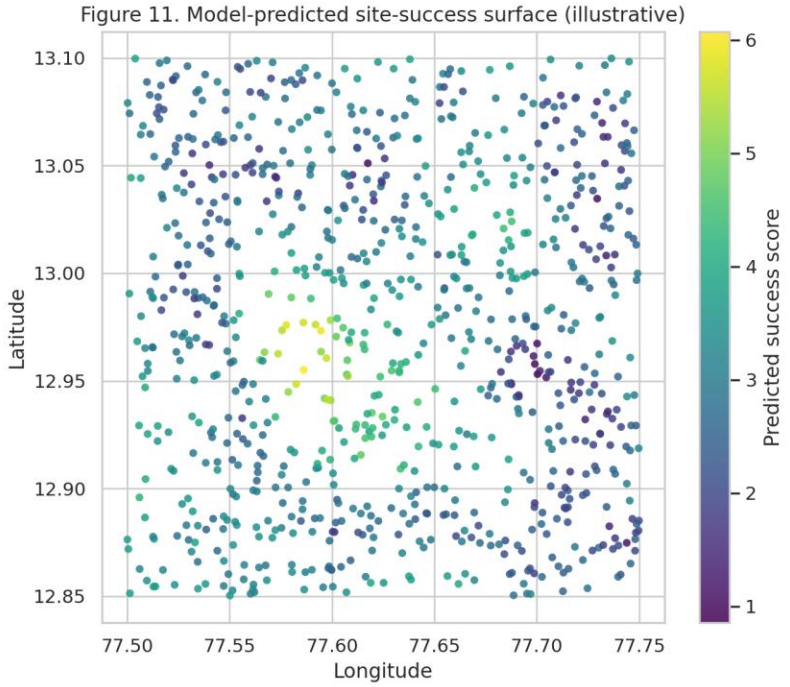

Figure 15. Model-predicted success surface scored at observed outlet locations.

5. Discussion

The empirical findings both confirm and qualify expectations drawn from the retail-location literature. Consistent with classical agglomeration and gravity-based theory [6] and with recent ML-based retail-site-selection studies [8,9,23], competitive saturation exerts a robust negative influence on predicted success across every specification tested, while local demand density exerts a robust positive influence — together accounting for the majority of explained variance in both tasks. This supports the view that, even in a delivery-only format lacking any physical storefront, the core economic logic of retail location — balancing demand density against competitive crowding — continues to apply, echoing Feng et al.'s [23] finding that population density remains a first-order determinant even for logistics-driven, app-mediated retail formats.

The comparatively modest absolute predictive power achieved here (test R^2 0.35–0.38) contrasts with the substantially higher accuracies (0.90–0.92) reported by grid-based random forest studies of physical retail [8]. Several factors explain this gap. Those studies frame prediction as spatial classification over a dense grid of store-present versus store-absent locations using dozens of point-of-interest, land-use, and accessibility features, whereas the present study predicts a continuous performance score for operating outlets using eleven predictors. Moreover, as argued in Section 4.3, a cloud kitchen's performance depends heavily on execution-level factors orthogonal to location; the unexplained variance is therefore consistent with, rather than contradictory to, the hypothesis that location is necessary but not sufficient.

The relatively strong performance of linear and ridge regression relative to tree ensembles departs from the pattern reported throughout the physical-retail literature, where random forest is typically the top performer [8,9,11,12], and also from Choudhary et al.'s [17] finding on the same underlying Bengaluru data that random forest markedly outperformed linear regression for attribute-based rating prediction. Two pieces of evidence presented above clarify this divergence. First, the loss curves (Section 4.3) show that the gradient

boosting configuration was over-parameterised, so its Table 1 result understates its tuned potential. Second, the learning curves indicate that the tree ensembles remain in a data-limited regime at N = 900, whereas ridge regression has effectively converged. Combined with the largely monotonic bivariate relationships visible in Figures 4 and 5, this suggests the linear models' advantage reflects the modest sample size and compact, near-additive feature set rather than any inherent unsuitability of nonlinear methods; studies incorporating richer point-of-interest feature sets have generally found that additional interaction structure favours tree ensembles [8,9].

Reframing the task as high-potential-site classification rather than point prediction improved discriminative performance, mirroring Ou, Fu and Wu's [11] finding that a coarser, threshold-based success label is easier to predict reliably than a continuous outcome. This supports using tertile-based screening labels as the primary operational output of a cloud-kitchen site-selection tool, with the continuous score retained as a secondary signal for ranking within the high-potential set.

6. Practical Implications for Cloud Kitchen Operators

Three practical implications follow. First, population density and competitive saturation within a 2-km radius should be treated as first-order screening criteria, given their dominant importance across every model and metric; a two-dimensional density-versus-saturation screen could eliminate a large share of unpromising candidates before more expensive analysis. Second, same-cuisine saturation within 2 km carries independent explanatory power beyond total saturation, implying operators should evaluate cuisine-specific competitive intensity separately from aggregate outlet density when deciding which brand concept to deploy — a site saturated for Biryani may remain comparatively open for Desserts or Cafe formats. Third, because delivery convenience and cuisine fit contribute meaningfully but secondarily, they are best used as tie-breakers among a shortlist that has already cleared the primary density/saturation screen.

Operationally, the statistically significant but moderate predictive power, combined with the conservative recall profile of the best classifier, positions such a tool as a decision-support pre-screening layer that narrows a large candidate set to a shortlist for human due diligence — site visits, lease negotiation, local market research — rather than an automated site-selection system. Aggregator platforms holding granular order-level demand and competitor data are particularly well placed to extend this approach, since Equations (1)–(3) can in principle be evaluated at any unopened candidate grid cell, enabling continuous city-wide suitability scoring.

7. Limitations

Several limitations qualify these findings. First, the analysis is cross-sectional and confined to a single metropolitan market; the relative importance of density, saturation, and cuisine fit plausibly varies across cities with different population distributions and competitive structures, so coefficients should not be assumed to transfer without re-estimation. Second, the feature set omits variables known to matter for site performance — rent and occupancy cost, kitchen size, transit and road-network accessibility, point-of-interest density by category, and demand seasonality — any of which could improve performance. Third, the success-score target is a composite metric whose precise construction is not fully observable, limiting causal interpretability; the models should be read as predictive rather than causal. Fourth, the dataset captures only currently operating outlets and cannot observe counterfactual performance where no outlet exists (survivorship and selection effects), a limitation shared with most observational site-selection studies

[8,9,11]. Fifth, hyperparameters were fixed rather than systematically tuned; the loss curves in Section 4.3 demonstrate that this materially disadvantaged gradient boosting, and a full grid or Bayesian search would likely improve ensemble performance.

8. Conclusion and Future Work

This study developed and benchmarked a machine-learning framework for predicting cloud-kitchen site success using 900 geo-referenced food-delivery outlets in Bengaluru. Ridge regression provided the best generalising continuous predictor (test $R^2 = 0.379$, adjusted $R^2 = 0.314$, $F(17,162) = 5.82$, $p < 0.001$), and a random forest classifier best distinguished high-potential sites (ROC-AUC = 0.744, 95% CI [0.664, 0.825], $p < 0.001$). Both results are statistically significant with large effect sizes, while remaining moderate in absolute magnitude — a pattern consistent with location being a necessary but structurally insufficient determinant of delivery-only kitchen performance. Population density and multi-radius competitive saturation emerged as dominant predictors across every model and importance metric, with same-cuisine saturation, delivery convenience, and cuisine category contributing secondary explanatory power. Loss-curve diagnostics additionally revealed that gradient boosting overfits beyond 65 iterations, and learning curves indicated the tree ensembles remain data-limited at this sample size.

Future work could extend this framework along several dimensions: incorporating point-of-interest, transit-accessibility, and rent data to enrich the feature space and test whether tree ensembles regain their usual advantage; systematic hyperparameter optimisation with early stopping, which the loss curves indicate would materially improve boosting performance; applying the model to genuinely unopened candidate grids to generate prospective suitability maps; extending to multiple cities to test geographic transferability; adding SHAP-based local explanations [22] to support site-specific justification; and incorporating temporal order-volume data to move from a cross-sectional score toward a dynamic model. Combining such a location-scoring layer with the operational, kitchen-level optimisation models already studied in the operations-research literature [13,14] represents a natural next step toward an end-to-end cloud-kitchen planning pipeline.

References

[1] Chitradevi, S., Karcherla, H.P., Sowmya, M., Bhansali, S. (2023). Cloud kitchens: revolutionizing the food & beverage industry. TIJER International Research Journal, 10(2), 105–112.

[2] Nedumaran, G., Ritha, M. (2025). A systematic literature review of cloud kitchen. Plant Science Today, 12(sp1). https://doi.org/10.14719/pst.6016

[3] Kulshreshtha, K., Sharma, G. (2022). From restaurant to cloud kitchen: survival of the fittest during COVID-19 — an empirical examination. Technological Forecasting and Social Change, 179, 121629. https://doi.org/10.1016/j.techfore.2022.121629

[4] Pookulangara, S., Wen, H., Bharath, J. (2022). Consumer attitudes toward ordering from cloud kitchens: a gender and marital status perspective. International Journal of Contemporary Hospitality Management, 35(5), 1859–1879. https://doi.org/10.1108/IJCHM-02-2022-0148

[5] Chatterjee, R., Singh, A., Singh, V. (2022). Ethical and sustainable perceptions on cloud kitchen business: a study of consumers and stakeholders during the COVID-19 pandemic. International Journal of Hospitality and Tourism Systems, 15(COVID-19 Special Issue), 76–84.

[6] Huff, D.L. (1963). A probabilistic analysis of shopping center trade areas. Land Economics, 39(1), 81–90. https://doi.org/10.2307/3144521

[7] X-Byte Enterprise Crawling. (2026). Dark kitchens and predictive location intelligence. Retrieved from https://www.xbyte.io/dark-kitchens-predictive-location-intelligence/

[8] Zhao, Z., Chen, G., Duan, J., Xu, Y. (2025). Site selection analysis and prediction of new retail stores from an urban commercial space perspective: a case study of Luckin Coffee and Starbucks in Shanghai. ISPRS International Journal of Geo-Information, 14(6), 217. https://doi.org/10.3390/ijgi14060217

[9] Zhao, J., Zong, B., Wu, L. (2023). Site selection prediction for coffee shops based on multi-source space data using machine learning techniques. ISPRS International Journal of Geo-Information, 12(8), 329. https://doi.org/10.3390/ijgi12080329

[10] Lu, J., Zheng, X., Nervino, E., Li, Y., Xu, Z., Xu, Y. (2024). Retail store location screening: a machine learning-based approach. Journal of Retailing and Consumer Services, 77, 103620. https://doi.org/10.1016/j.jretconser.2023.103620

[11] Ou, T.-Y., Fu, H.-P., Wu, M.-Z. (2025). Optimize a chain convenience store location prediction model by using MTS-machine learning methodology. Scientific Reports, 16, 1056. https://doi.org/10.1038/s41598-025-30650-w

[12] Yee, H.-J., Ting, C.-Y., Ho, C.C. (2019). Retail site selection using machine learning algorithms. International Journal of Recent Technology and Engineering, 8(4), 2422–2431. https://doi.org/10.35940/ijrte.D7186.118419

[13] Neria, G., Hildebrandt, F.D., Tzur, M., Ulmer, M.W. (2024). The restaurant meal delivery problem with ghost kitchens. Transportation Science, 59(2), 433–450. https://doi.org/10.1287/trsc.2024.0510

[14] Švancár, S., Chrpa, L., Dvořák, F., Balyo, T. (2024). Cloud kitchen: using planning-based composite AI to optimize food delivery processes. arXiv:2402.10725 [cs.AI].

[15] Poddar, H. (2019). Zomato Bangalore restaurants: a public dataset of 51,717 restaurant records across 17 attributes. Bengaluru restaurant listing data.

[16] Gudikandula, P. (2019). Zomato Bangalore restaurant analysis and rating prediction: an exploratory and predictive study. Analytics Vidhya Technical Review.

[17] Choudhary, N., Panwar, V., Mittal, S., Sahu, G. (2021). Zomato restaurants data analysis using machine learning algorithms. Journal of Emerging Technologies and Innovative Research (JETIR), 8(2), 1287–1294.

[18] Manhas, A., Singh, S. (2020). Predicting the Zomato restaurant ratings using machine learning techniques. International Journal of Emerging Trends in Engineering Research, 8(5), 351–355. https://doi.org/10.30534/ijeter/2020/03852020

[19] Kumar, A., Sharma, P., Verma, R. (2024). Prediction model, sentiment analysis and recommendation system on Zomato Bangalore dataset with XAI tools LIME and SHAP. International Journal of Innovative Research in Technology, 10(9), 412–421.

[20] Singh, R., Singh, M., Goyal, M. (2020). Predicting restaurant ratings using machine learning techniques: a case study of Zomato. International Journal of Advanced Science and Technology, 29(4), 6384–6394.

[21] Li, C., Mirosa, M., Bremer, P. (2020). Review of online food delivery platforms and their impacts on sustainability. Sustainability, 12(14), 5528. https://doi.org/10.3390/su12145528

[22] Lundberg, S.M., Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[23] Feng, G., Shunyi, L., Zhenzhi, J., Zhisai, H., Yang, L., Hongbao, L., Jiemin, W., Wangyang, C., Guanyao, L. (2025). Location differs between traditional and new retail: a comparison analysis of Starbucks and Luckin Coffee in China using machine learning. Cities, 158, 105668. https://doi.org/10.1016/j.cities.2024.105668

[24] Chen, T., Guestrin, C. (2016). XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

[25] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[26] Cortes, C., Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

[27] Hoerl, A.E., Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

[28] Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[29] Hanley, J.A., McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

[30] Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Lawrence Erlbaum Associates, Hillsdale, NJ.

[31] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.