What affects the lifetime of a business?

1. Project description and abstract

Why do some businesses survive for years, while others with similar characteristics close quickly? Understanding factors that affect the duration of a business is an interesting question to research for a person with statistics and machine learning knowledge. Restaurants and small businesses operate in a highly competitive environment, where customer attention, local market conditions, and other not obvious characteristics jointly influence the long-term development story of the business. Knowing which characteristics are correlated with longer business lifetimes can help owners, investors, and policymakers make more informed decisions.

Our project conducts research using restaurant information on Yelp, which is one of the most comprehensive public enterprise datasets available. The Yelp open dataset provides rich high-dimensional information, enabling us to examine the business performance from traditional structured features (ratings, number of reviews, location, category) as well as behavioral and time signals (timestamped check-ins and early customer activities). From this data, we construct measures of early popularity, chain status, local competition, ratings, and review volume, and define business lifetime in months. Our goal is to combine unsupervised structure discovery, interpretable linear modeling, and non-linear machine learning methods to understand the relationship between different dimensions of business behavior and lifetime. We also implement random forests, spline-based regressions, and gradient boosting machines to study how closure risk varies with competition across different price levels.

In this project, we will research three questions: (i) whether dimension-reduction and clustering can create interpretable higher-level features for business survival models, (ii) how strongly our engineered features differ from linear behavior results, and (iii) how closure risk is associated with local competitive density across different price levels. While answering these questions, we find that PCA and K-means mainly help us interpret the structure of the data, separating “high-review, low-rating” businesses from “high-competition” ones without materially changing linear-model AUC (≈0.56 in all sparse linear variants). When we move to nonlinear models, test AUC rises from about 0.68 for logistic and spline models to roughly 0.74 for gradient boosting, so extra flexibility yields clear but moderate gains. We also document that the probability of closure increases with city-level restaurant counts at all price levels and is noticeably higher for mid- and high-priced restaurants than for the cheapest ones.

The main feature-engineering idea is **second-layer structure**: instead of stopping at raw variables like review volume or competition, we compress them into principal components and cluster labels that act as “business archetypes.” For example, PC1 loads heavily on local competition and early review volume, PC2 on early ratings, and our K-means clusters isolate patterns such as “high competition, high volume” versus “high rating, low volume.” This lets us talk about survival not just in terms of single variables but in terms of recognizable business profiles.

Beyond the core regression and classification tools from class, we use **LOESS smoothing** to visualize raw relationships between engineered features and lifetime, and **partial dependence plots** with tree ensembles to study nonlinear effects such as the diminishing-returns pattern of early popularity and the near-linear effect of competition. Lastly, our analysis on local competition showed that it had a positive impact on the closure of restaurants but the shape of this relationship varied across different price levels. Using various machine learning models we concluded that more expensive restaurants were marked sensitive to competitive saturation whereas less expensive restaurants remained comparatively resilient to local competition.

AI Usage Statement. We used generative AI tools (ChatGPT) for assistance with code debugging and wording/grammar suggestions in this report. All authors take full responsibility for the content of this work and for verifying the accuracy and integrity of any material generated or assisted by AI tools.

Literature Review

1. Interpretable Business Survival Prediction (Vallapuram et al., 2021)

Vallapuram et al. (2021) used Yelp data to predict whether a business will survive over a fixed horizon (open in 2017, alive or not by the end of 2019). They engineered four feature families from location-based social networks: geography, user mobility, business attributes, and linguistic features from reviews. They trained several ML classifiers across these feature sets: gradient-boosted decision trees, multilayer perceptrons, LSTMs; and then applied LIME (Local Interpretable Model-Agnostic Explanations) to generate human-readable explanations for survival predictions.
This paper's emphasis on qualitative, interpretable features (business attributes and linguistic features) as the strongest predictors gave us ideas for building intuitive, business-level features such as early popularity, chain status, and competition indices rather than relying only on raw counts or black-box embeddings. We also took a similar logic by prioritizing interpretable engineered features and using tools like partial dependence plots over our core features (early popularity, competition) so that we can discuss business profiles and thresholds rather than just reporting AUC.

Main findings:
1. Business attributes and linguistic features are the most predictive, with AUC ≈ 0.72 and 0.67 respectively, outperforming geography and mobility-only models.
2. Text models achieve very high performance on **sentiment** (AUC ≈ 0.98) but more modest gains on survival, showing that reviews analyze customer perceptions better, but survival is harder to predict.

2. Principal Component Analysis and Interpretable Reduction

Jolliffe and Cadima (2016) review PCA as a tool for reducing high-dimensional, correlated data into a small number of interpretable components. They emphasize that PCA is most valuable for revealing latent structure, because loadings often correspond to intuitive dimensions such as popularity, quality, or scale (rather than for improving predictive accuracy).
This directly informed our RQ1 approach. Since Yelp features like review volume, ratings, and competition are strongly correlated, PCA allowed us to uncover clear business-level dimensions (e.g., competition–popularity, quality). However, consistent with Jolliffe and Cadima's findings, including these components in LASSO did **not** improve AUC. PCA maximizes variance, not survival relevance, and it cannot capture nonlinear effects, which we later observed to be central to business closure patterns.

Main findings:
1. PCA improves interpretability through meaningful latent structure.
2. PCA is not designed to improve linear predictive accuracy, particularly when outcomes are nonlinear.
3. This explains why our PCA + sparse linear modeling pipeline yielded better understanding but not better prediction.

Citations:
Vallapuram, A. K., Nanda, N., Kwon, Y. D., & Hui, P. (2021). Interpretable business survival prediction. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 99–106). Association for Computing Machinery. https://doi.org/10.1145/3487351.3488353

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Data Processing & Summary Statistics

The Yelp business-level dataset is a large collection of information on individual local businesses, primarily restaurants. Each observation corresponds to a single business ID and links it to its location (city, ZIP code, coordinates), categories, rich attributes, opening hours, and a time-stamped sequence of customer check-ins and reviews.

business_id	name	city	state	stars	review_count	is_open	main_category	same_cat_in_city	is_chain	first_checkin	last_checkin	lifetime_months	long_lived	n_checkins
id_001	Business 1	SampleCity	SC	3.5	10	1	Category1	5	0	2010-01-01	2010-12-31	12.0	0	50
id_002	Business 2	SampleCity	SC	4.5	20	1	Category2	10	1	2011-01-01	2011-12-31	13.5	0	53
id_003	Business 3	SampleCity	SC	5.5	30	1	Category3	15	0	2012-01-01	2012-12-31	15.0	0	56
id_004	Business 4	SampleCity	SC	3.5	40	1	Category4	20	1	2013-01-01	2013-12-31	16.5	0	59
id_005	Business 5	SampleCity	SC	4.5	50	1	Category1	25	0	2014-01-01	2014-12-31	18.0	0	62
id_006	Business 6	SampleCity	SC	5.5	60	1	Category2	30	1	2015-01-01	2015-12-31	19.5	0	65
id_007	Business 7	SampleCity	SC	3.5	70	1	Category3	35	0	2016-01-01	2016-12-31	21.0	0	68
id_008	Business 8	SampleCity	SC	4.5	80	1	Category4	40	1	2017-01-01	2017-12-31	22.5	1	71
id_009	Business 9	SampleCity	SC	5.5	90	1	Category1	45	0	2018-01-01	2018-12-31	24.0	1	74
id_010	Business 10	SampleCity	SC	3.5	100	1	Category2	50	1	2019-01-01	2019-12-31	25.5	1	77

From this raw information, we construct key variables such as business lifetime (in months), average rating, review volume, early popularity, chain status, and several measures of local competition at the city and category level. Together, these features allow us to study how reputation, early traction, and competitive density relate to whether a business survives or closes.

1. Business Metadata

- Location: city, state, postal code, precise latitude/longitude
- Categories: cuisine type, restaurant style, food specialization
- Attributes: service options, ambiance, pricing indicators, accessibility features
- Operational status: the variable is_open flags whether the business is currently active

2. Customer Feedback Signals

- Average star rating (stars) representing perceived quality
- Total number of reviews (review_count) indicating consumer engagement
- Textual reviews and lengths reflecting customer depth of experience

These feedback variables correlate strongly with survival in both linear and nonlinear models:

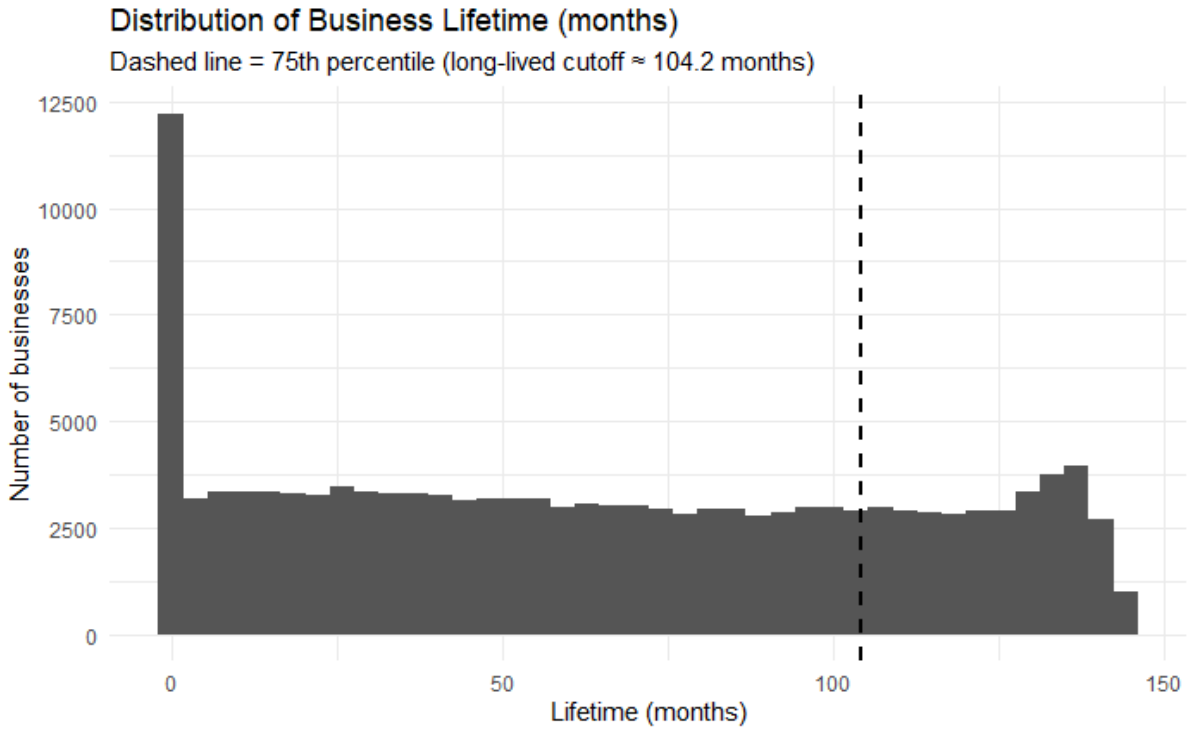
- Review count drives PC1 in PCA and Rating drives PC2
- Review volume and rating appear prominently in clustering centroids

3. Behavioral Time-Series Data

- Check-in timestamps: records of customer visits
- Early popularity: number of check-ins in the first 6 months
- Review timestamps: used to reconstruct observed lifetime

Lifetime (months)=Last Recorded Check-in or Review - First Recorded Activity

Figure 1 displays the distribution of business lifetimes. The histogram is right-skewed, with many businesses having relatively short observed lifetimes and a long tail of establishments that remain active for several years. The dashed vertical line marks the **75th percentile of the lifetime distribution**, which we use as the cut-off for defining “long-lived” businesses in the classification models below: businesses at or above this threshold are coded as “long_lived=1”, the remainder as “long_lived=0”. This definition concentrates on explaining the factors associated with being in the upper quartile of the lifetime distribution, rather than modelling lifetime as a continuous response.



For modeling the “long_lived” was recoded as a factor ("no", "yes"), “is_chain” and “is_open” were treated as categorical predictors. A **70/30 train–test split** (stratified by long_lived) is used.

To research the effect of closure status to local competitive density, only entries that contained the keyword **“Restaurants”** in their list of categories were used. This ensured that all observations were directly relevant to understanding restaurant-level closure behavior. After subsetting, a group of variables that carried no predictive value for closure status were removed.

Table: Dropped Variables

Reason	Variable
ID variable (cause overfitting)	business_id
No predictive power	name, address
Redundant (replaced by engineered features)	is_open
High Cardinality (too many levels)	hours

To obtain a usable measure of price level, the nested attribute **RestaurantsPriceRange2** was extracted from the business attributes field. This variable encodes pricing levels on a 1-4 scale and was recoded as an ordered categorical variable representing market positioning (from cheap to expensive restaurants). Because price level plays a central role in moderating competitive effects, careful extraction and cleaning of this variable was essential. We restricted the analysis to businesses with non-missing latitude/longitude, stars, review_count, and RestaurantsPriceRange2; observations missing any of these core variables were dropped (0.17% of total observations).

Extremely large values of review_count and early_checkins_6m were handled by a square-root transformation to reduce the influence of outliers. Although Yelp provides full text reviews, we did not focus on them; instead we used stars and review_count as aggregate “text-derived” signals. Categorical predictors (is_open, is_chain, price_level) were encoded as dummy variables via R’s factor handling.

1.Unsupervised Learning

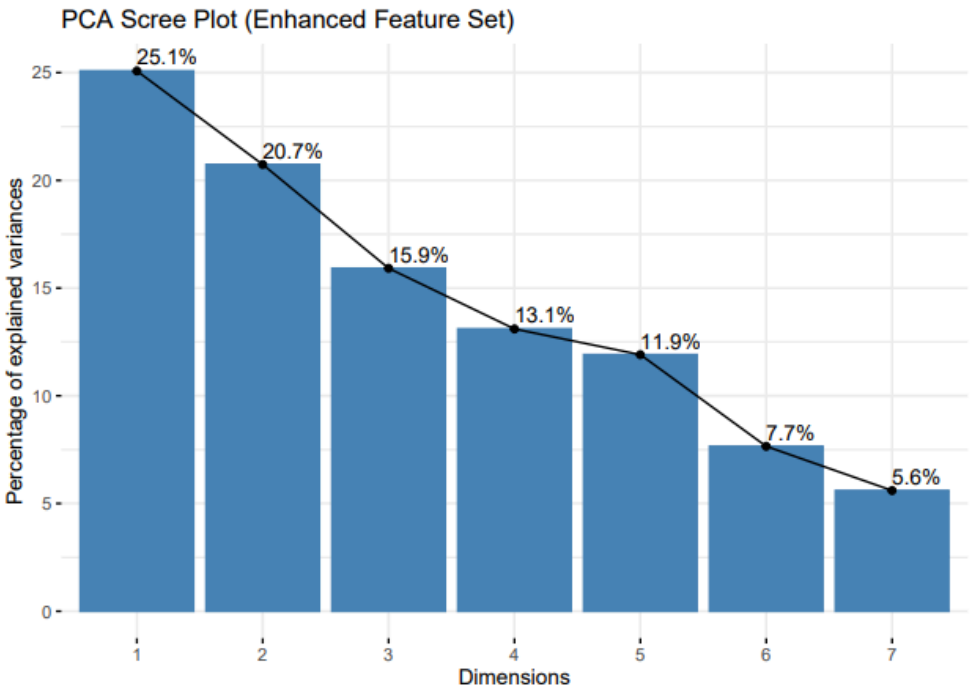
1.1. Preprocessing and Feature Construction

1.1.1. Principal Component Analysis

Continuous variables (ratings, review totals, early traction, competition counts, longevity, and category breadth) are **standardized** prior to PCA and clustering.
Categorical variables are not one-hot–encoded into PCA (to avoid distorting variance structure). Instead:

- Business categories → transformed into primary category and competition features (comp_citycat, comp_zipcat).
- Price level → used later to interpret clusters.
- City/ZIP → reflected through competitive-density features.

This ensures meaningful mixing of categorical information without overwhelming the PCA space.



- PC1: Market exposure & competition density (driven by comp_citycat, comp_zipcat)
- PC2: Business activity / customer engagement (driven by rev_total, rev6)

- PC3: Reputation / customer satisfaction (driven by rating6)

These orthogonal PCs provide clean, high-level features that summarize Yelp's multi-dimensional business environment and enable stable clustering. Together, the PCs provide **interpretable higher-level features** that summarize Yelp business environment structure much better than raw variables alone, enabling clearer clustering and more stable LASSO modeling.

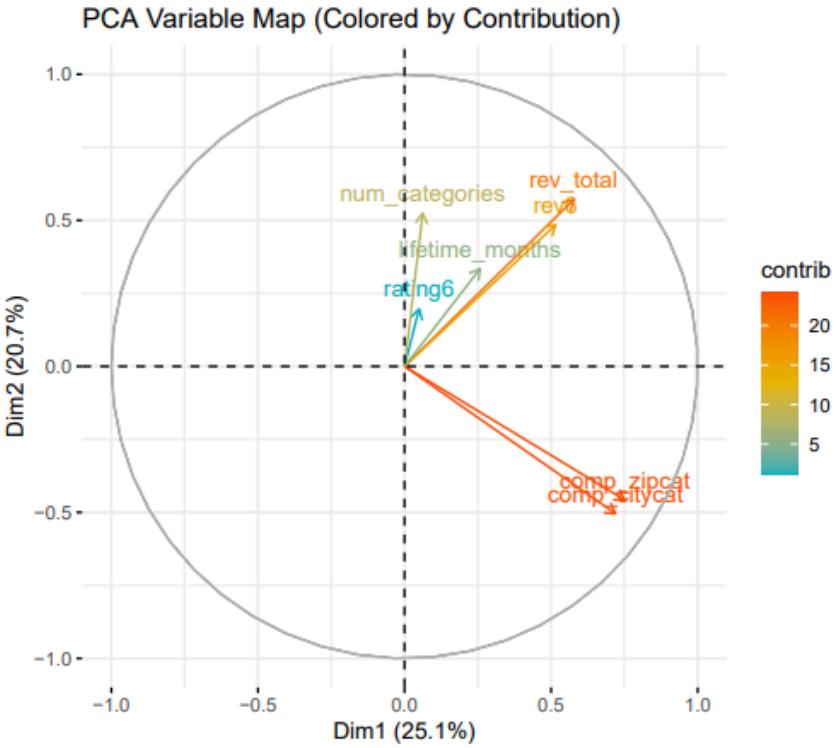


Figure 2: Variable Contributions to PCs

Each arrow represents an original variable (such as rating6, rev_total, comp_zipcat, lifetime_months, etc.). Arrow length = contribution of the variable to the PCA explanation; angle = correlation with other variables.

- Dimension 1 direction: The competition characteristics (comp_citycat, comp_zipcat) are the most distinct → PC1 represents "market size/competition density".
- Dimension 2 direction: rev_total and rev6 show a strong correlation → PC2 captures "evaluation intensity / visitor flow".
- Rating 6 has the highest contribution in Dim3 → PC3 captures the "reputation dimension".

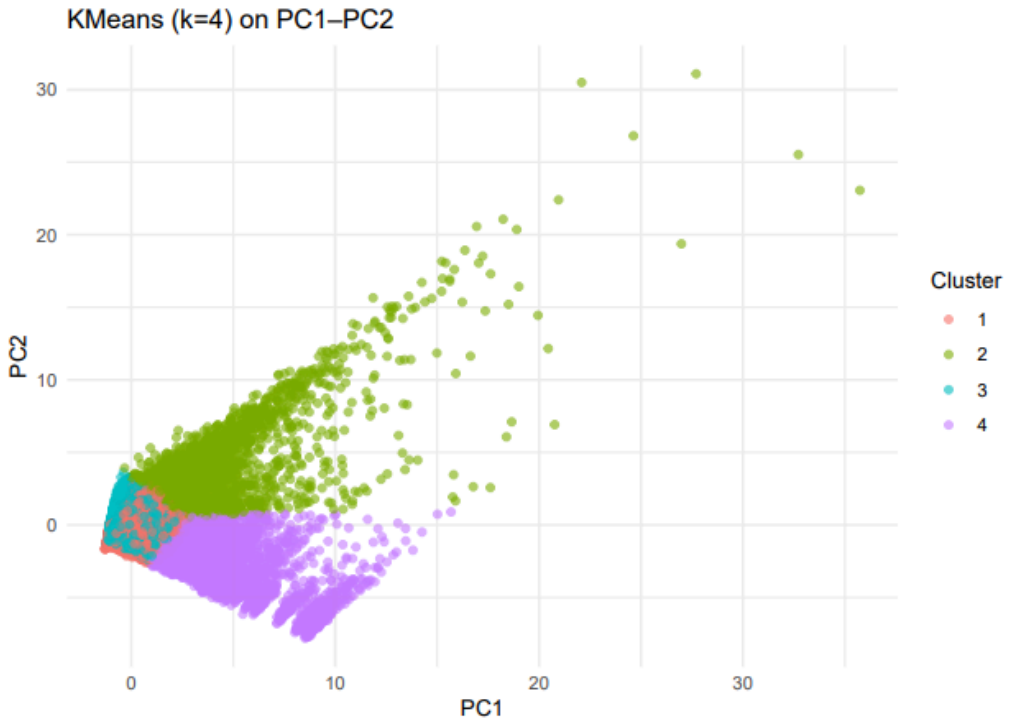
The graph shows that PCA successfully compressed the original features into three interpretable directions: **competition, activity, and reputation**.

1.2 Clustering Methodology

1.2.1. KMeans Clustering

The KMeans algorithm was used as a baseline because of its speed and suitability for spherical clusters in PCA space. Using k = 4, the algorithm converged to stable centroids after multiple initializations. Examination of centroid positions revealed four interpretable business profiles:

Cluster 1: High-rating (average ≈ 4.5 stars), moderate review count, and low competition: Representing niche, quality-focused businesses.
Cluster 2: Moderate ratings (~3.7) with very high review counts: large-volume, mainstream businesses concentrated in dense ZIP codes.
Cluster 3: Low ratings (~2.8) and low early activity: Underperforming or declining establishments.
Cluster 4: Mid-rating but high competition intensity - Businesses in saturated markets.

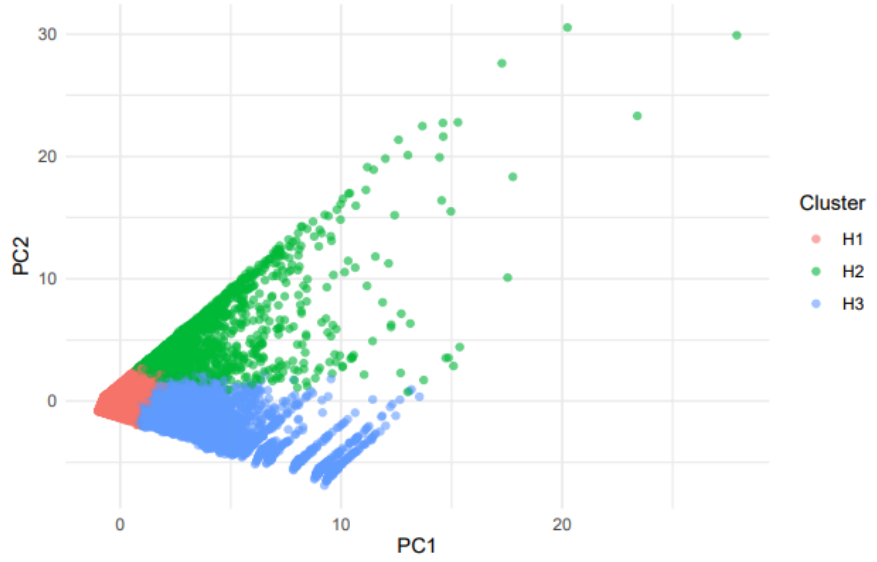


These clusters reveal clear behavioral differences across Yelp businesses, showing segmentation along both performance (ratings, review volume) and environmental (competition) dimensions.

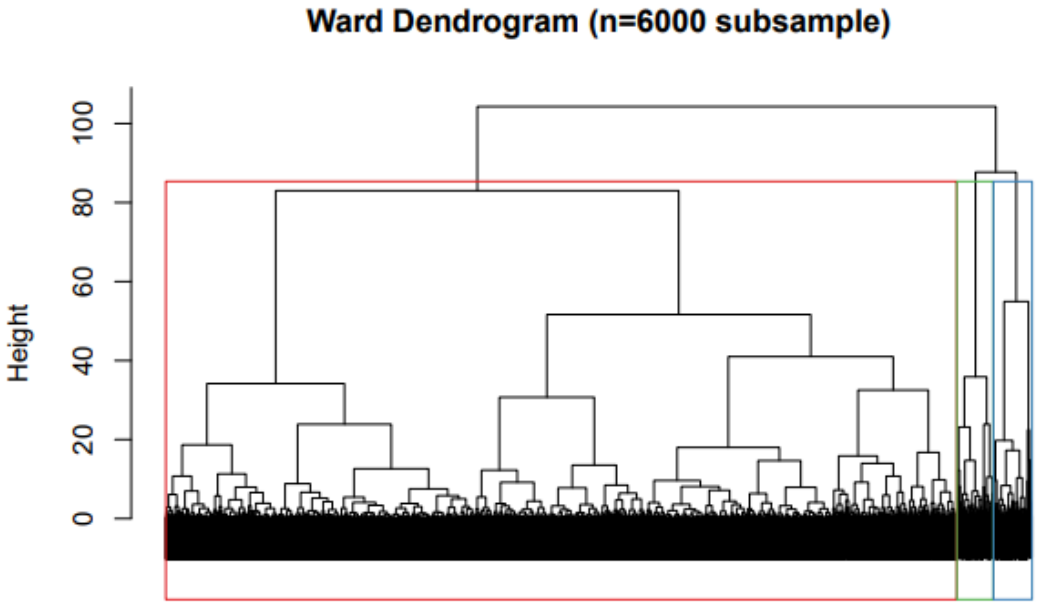
1.2.2. Ward Hierarchical Clustering

GMM was used to capture clusters with possible elliptical shapes and overlapping boundaries that KMeans or Ward may not fully detect. The probabilistic framework of GMM identifies soft boundaries between high-competition, high-volume businesses and lower-activity businesses, especially those that fall into transitional regions of PCA space.

In previous iterations of the analysis, GMM with G = 3 components identified the following:



- A large central component representing moderate ratings and review activity.
- A high-performance component characterized by strong early activity and competitive ZIP codes.
- A lower-engagement component with reduced visibility and weaker customer traction.



In our subsampled PCA dataset, the dendrogram shows three major branches, highlighted by the red, green, and blue bounding boxes. These three high-level clusters correspond to:

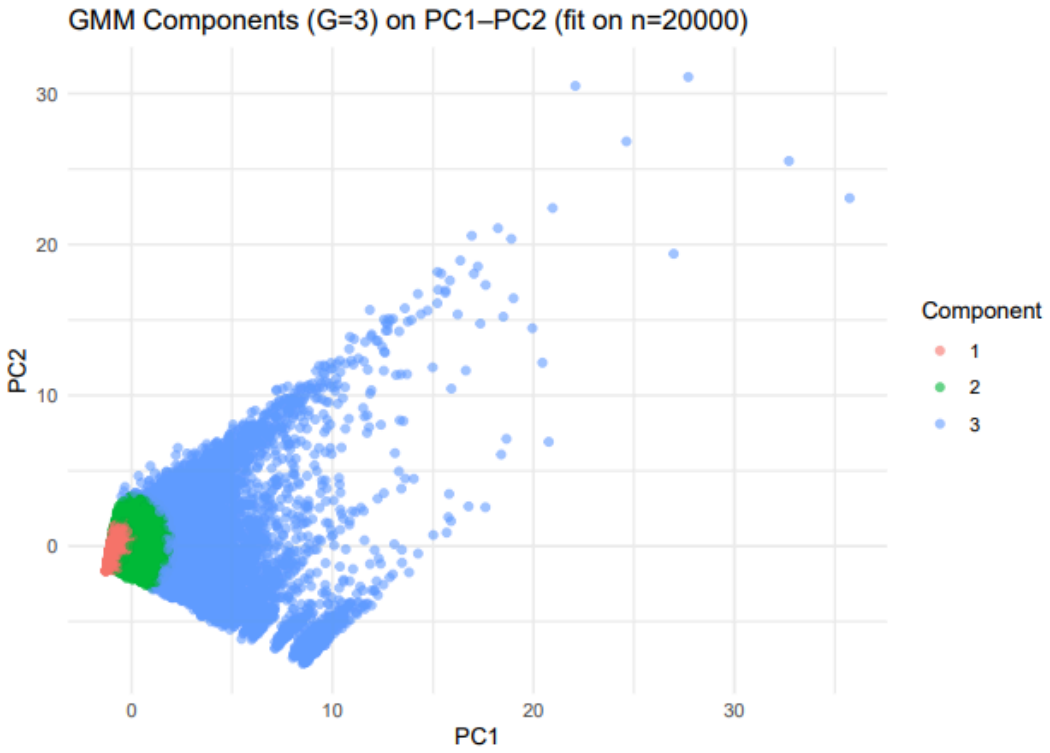
- a large and diverse group of lower-activity or less competitive businesses (**red region**)
- a medium-sized cluster of businesses with stronger early engagement or higher reputation (**green region**)
- a small but tightly grouped cluster of businesses located in highly competitive ZIP areas (**blue region**).

The fact that multiple large branches form at comparable heights is consistent with our PCA results, where PC1, PC2, and PC3 explain roughly one-third of the variance each.

1.2.3. Gaussian Mixture Models (GMM)

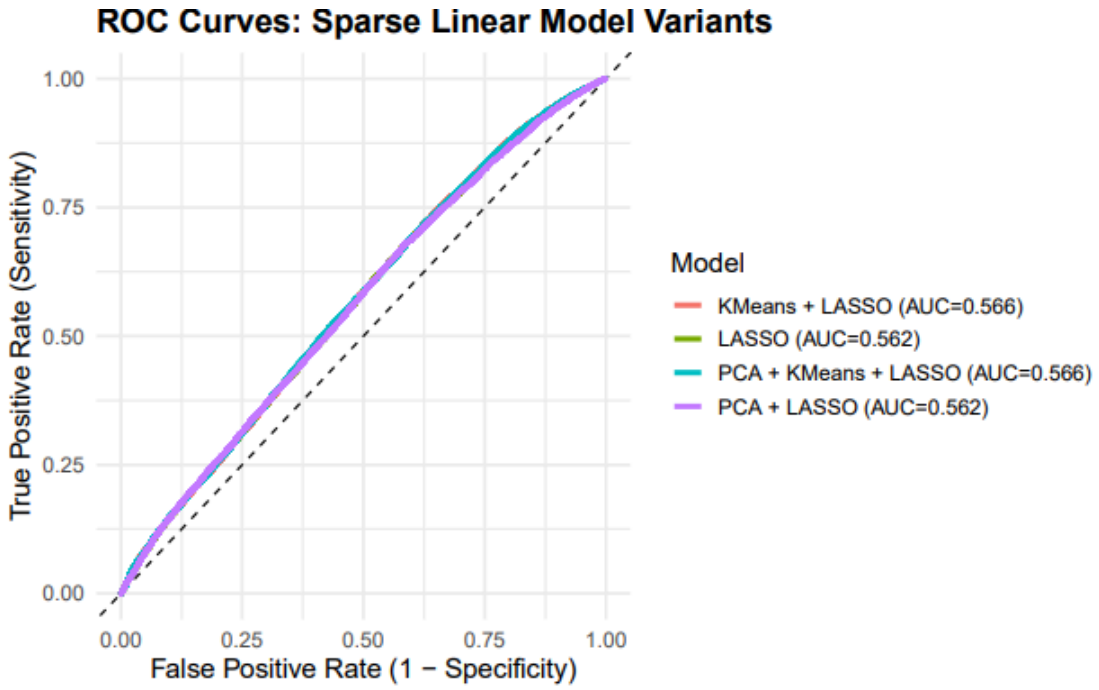
To allow for elliptical and overlapping clusters, a three-component GMM was fitted to the same PCA features. The mixture model identified three components (G1-G3) broadly consistent with the Ward results:

- **G3** (blue): the dominant component that spans the main cloud along PC1. These businesses show typical behavior with moderate–high activity and competition; most observations fall here.
- **G1** (red): a compact group near the origin (low PC1, low PC2), representing low-activity firms in less competitive settings.
- **G2** (green): another small group near the origin, slightly offset toward higher PC2 than G1, suggestive of low-to-moderate activity in somewhat denser competitive environments.



The probabilistic nature of GMM highlights transitional cases between G2 and G3, reflecting real-world overlap between competitive but successful businesses and quiet, low-volume ones.

1.3 Model Improvement



Among all the sparse linear model variants we tested, the predictive performance remained largely unchanged. As shown in the ROC curve comparison, the AUC values of all models were between 0.562 and 0.566, regardless of whether PCA or clustering was added.

This behavior is expected for the following reasons:

- The potential predictive signals in the Yelp business data are very weak.
- PCA and clustering reorganize the existing information, but they do not create new signals.

Although PCA and clustering improve interpretability and reveal the structure of the data, they **do not enhance the predictive accuracy** because the Yelp dataset itself does not contain strong early warning signals for business closures.

2. Feature Engineering

As part of the data pre-processing steps, other features, especially the local competition (number of restaurants and the ones having the same category) were engineered.

Variable	Type	Description	Formula/Logic
n_city_restaurants	Numeric	Number of restaurants in the same city	city_count = table(df$city) as.numeric(city_count[df$city])
n_same_cat	Numeric	Businesses having same category	Extracted from data
closed	Numeric	Closure status of the restaurant	ifelse(df$is_open==0, 1, 0)

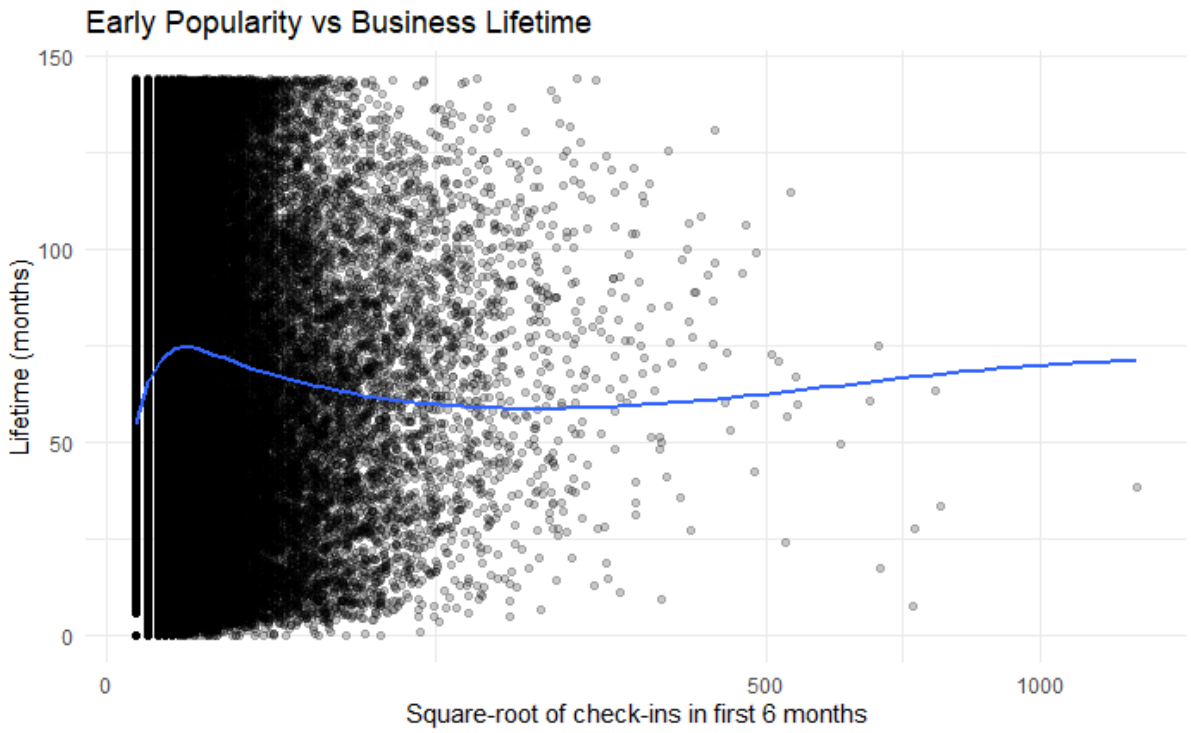
Local competition was only restricted to a particular region (city) which limited the power to capture more density within the data. More in-depth density estimation would have been possible with the help of zip codes; however, the dataset merges U.S. ZIP codes with Canadian postal codes and treats them inconsistently, making them unreliable for geographic segmentation. As a result, city-level competition was the most usable and stable measure available for the analysis.

2.1.1. Early Popularity as a Survival Signal

A central engineered feature in this study is **early popularity**, defined as the number of customer check-ins recorded during the first six months after a business's initial check-in (early_checkins_6m). This variable is extremely skewed: many businesses have few early check-ins, while a small fraction receive very high early traffic. To visualise its relationship with lifetime without being dominated by the heavy right tail, the x-axis was applied a **square-root transformation** to early_checkins_6m.

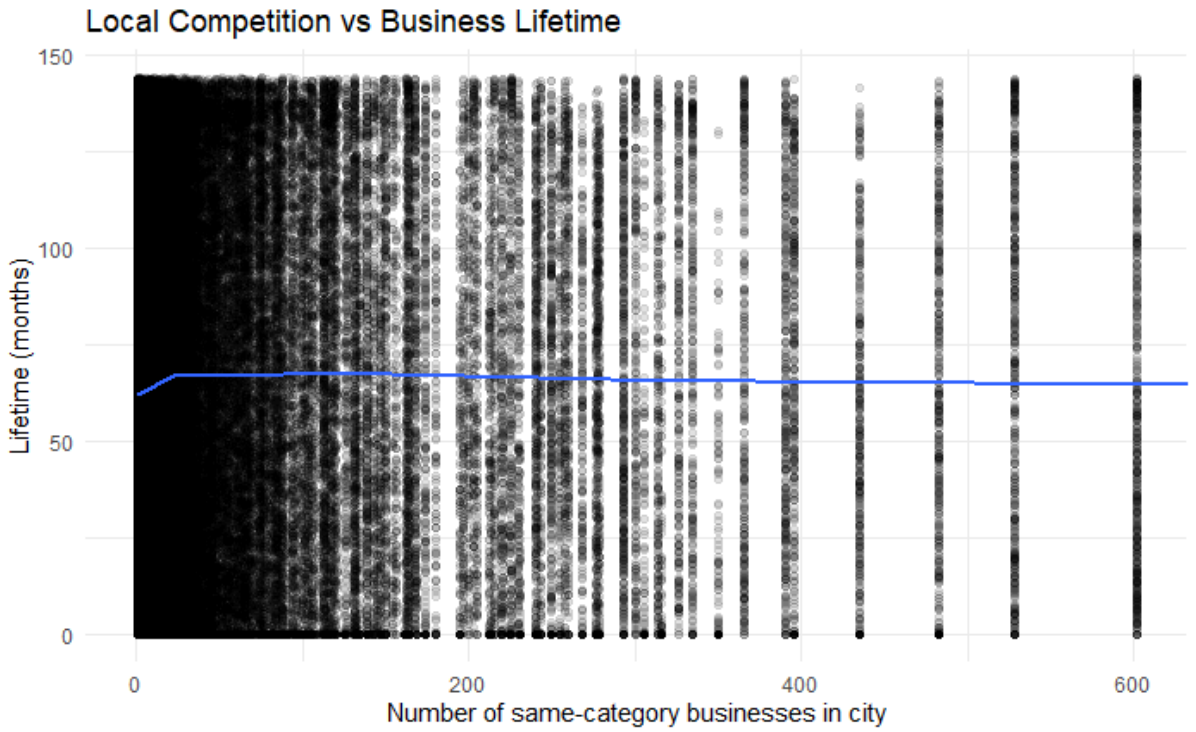
Each point represents a business, with its lifetime on the y-axis and √(early_checkins_6m) on the x-axis; a LOESS curve summarises the local trend. The curve rises steeply from about 50 months at very low early popularity to around 70-75 months as √(check-ins) increases, indicating that **some early demand is associated with longer survival**. Beyond moderate levels of early popularity, the curve flattens and then gently increases, suggesting **decreasing returns**: once a business is already popular, additional early check-ins are associated with only modest further gains in lifetime. This pattern is clearly nonlinear and would be poorly captured by a single linear slope.

A LOESS fit is clearly **not well approximated by a single linear slope**, so a purely linear model would be a poor summary.



2.1.2. Local competition density

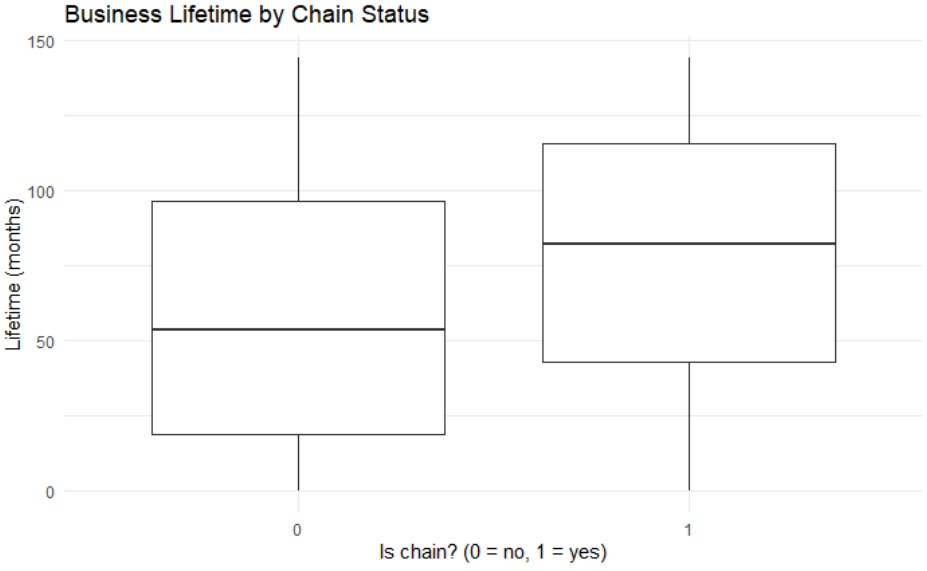
In contrast to early popularity, **lifetime vs. same_cat_in_city** plot shows a very different picture: the LOESS curve is **almost flat with only a slight slope**; any nonlinearity is small relative to the noise cloud.



This suggests that **ZIP-level or city-level competition on its own does not strongly determine lifetime**; instead it may act in interaction with other features (e.g., price, chain status) explored elsewhere in the project.

2.1.3. Chain effect

Chains (businesses whose name appears at least twice in the dataset) have higher median lifetimes and a visibly higher upper quartile than independent businesses. The “chain effect” is not just noise in the random forest, it is visible directly in the raw data.



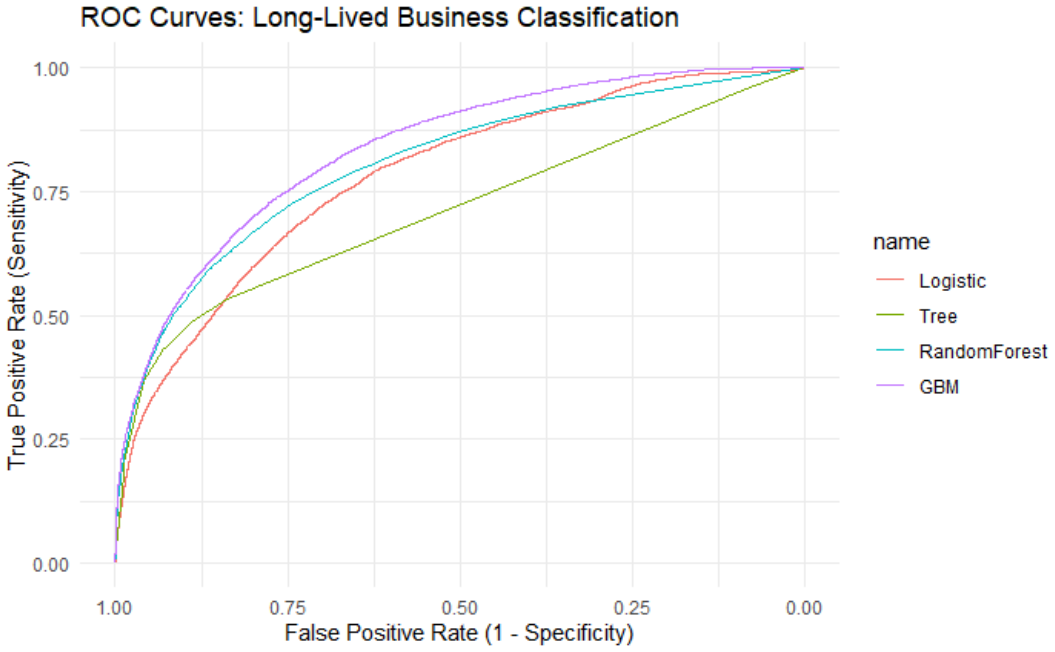
It can be summarized that nonlinearity is **feature specific**. Its value is strong for some engineered features (especially early popularity) but weak for others (competition), and some effects (chain status) are more about level shifts than curvature.

3. Prediction models

To quantify how much these nonlinear patterns matter for prediction, the binary outcome long_lived is modelled using four supervised learning algorithms. The data are randomly split into a 70% training set and a 30% test set, stratified by the outcome. The models are:

1. **Logistic regression** (baseline linear model)
2. **CART decision tree**
3. **Random forest**
4. **Gradient boosting machine** (GBM)

Hyperparameters for tree-based models were selected by 5-fold cross-validation on the training data using ROC-AUC as the tuning metric, and final performance is reported as AUC on the held-out test set. Tree-based models are tuned using cross-validation with the ROC-AUC as the optimization metric. For the CART model we tuned the maximum tree depth and minimum node size to prevent very deep, overfitted trees. For the random forest we varied the number of trees and the “mtry” (number of predictors tried at each split), using many trees with a relatively small “mtry" to keep individual trees de-correlated. For the GBM we tuned the learning rate, number of trees and interaction depth, using a small learning rate and shallow trees so that the ensemble grows slowly and remains well regularised.



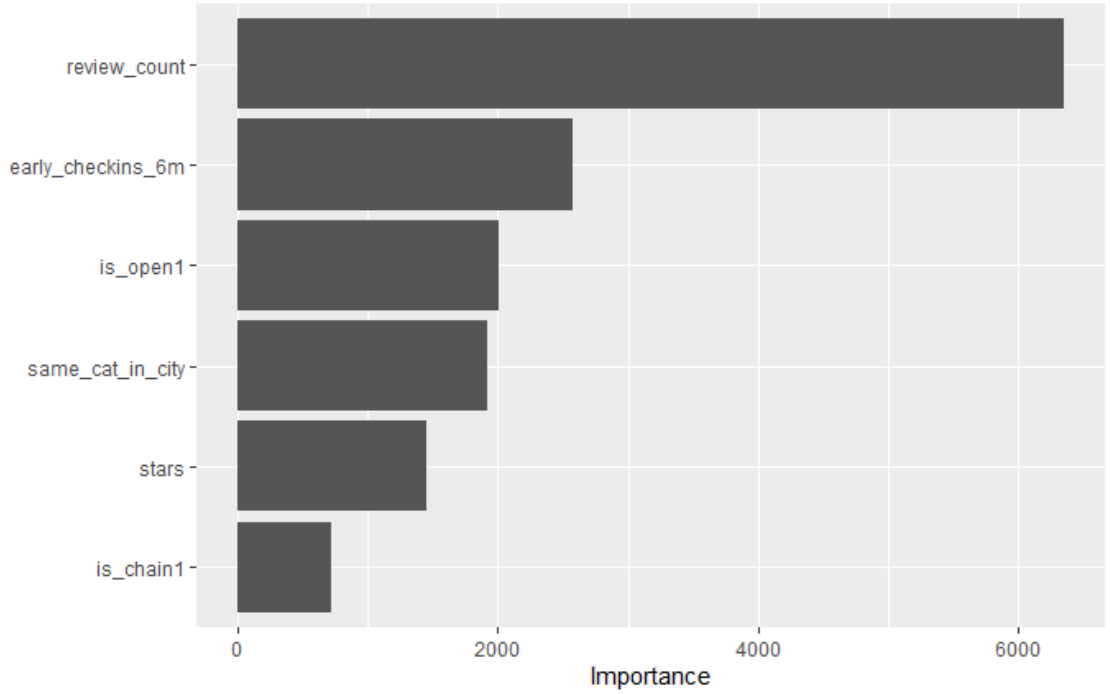
The linear logistic model already performs reasonably well, achieving an AUC of 0.781. Both tree ensembles improve on this baseline: the random forest attains an AUC of 0.805 and the GBM 0.836. The single decision tree achieves relatively high accuracy but clearly lower AUC, reflecting a model that is more unstable at distinguishing between long-lived and non-long-lived businesses across thresholds.

The gain from nonlinear modelling is **moderate but non-trivial**. ROC curves demonstrate that the random forest and GBM dominate the logistic and single tree across most operating points, indicating that **nonlinear effects and interactions among the engineered features carry predictive information beyond a linear specification**.

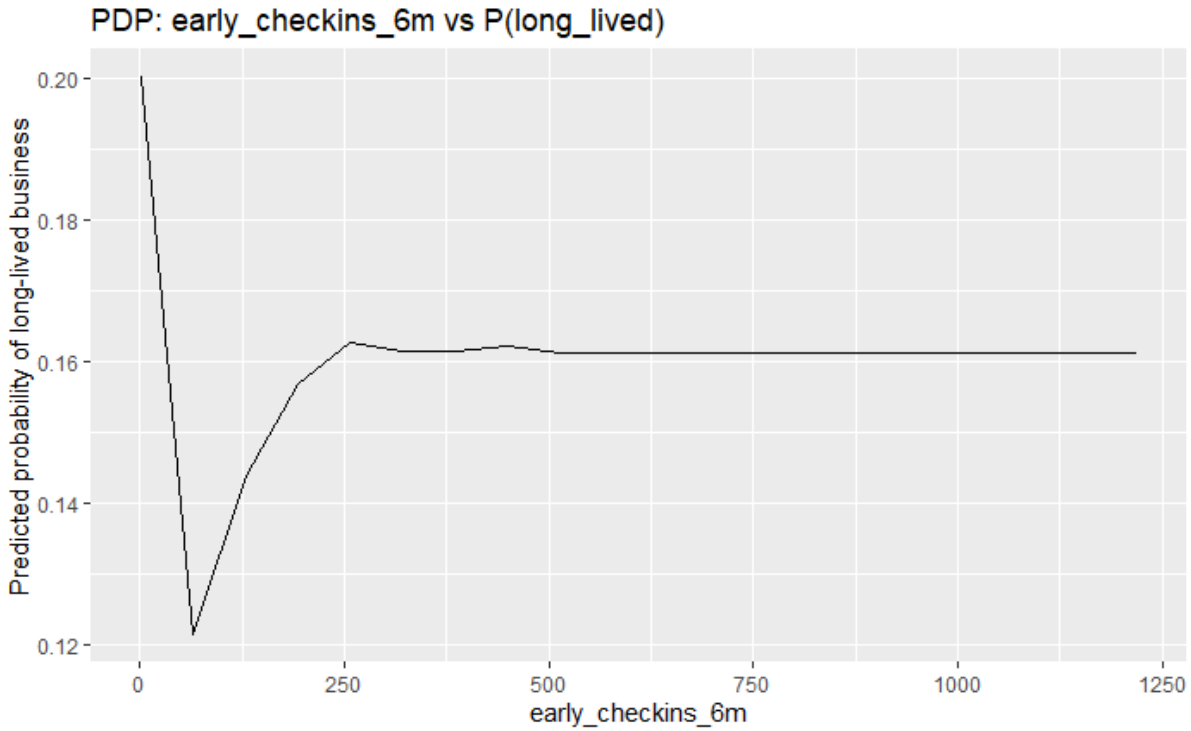
3.1. Model interpretation: variable importance, partial dependence

To understand which features drive these predictions and how the nonlinear models use them, the random forest was used and both **variable importance** and **partial dependence plots (PDPs)** were extracted.

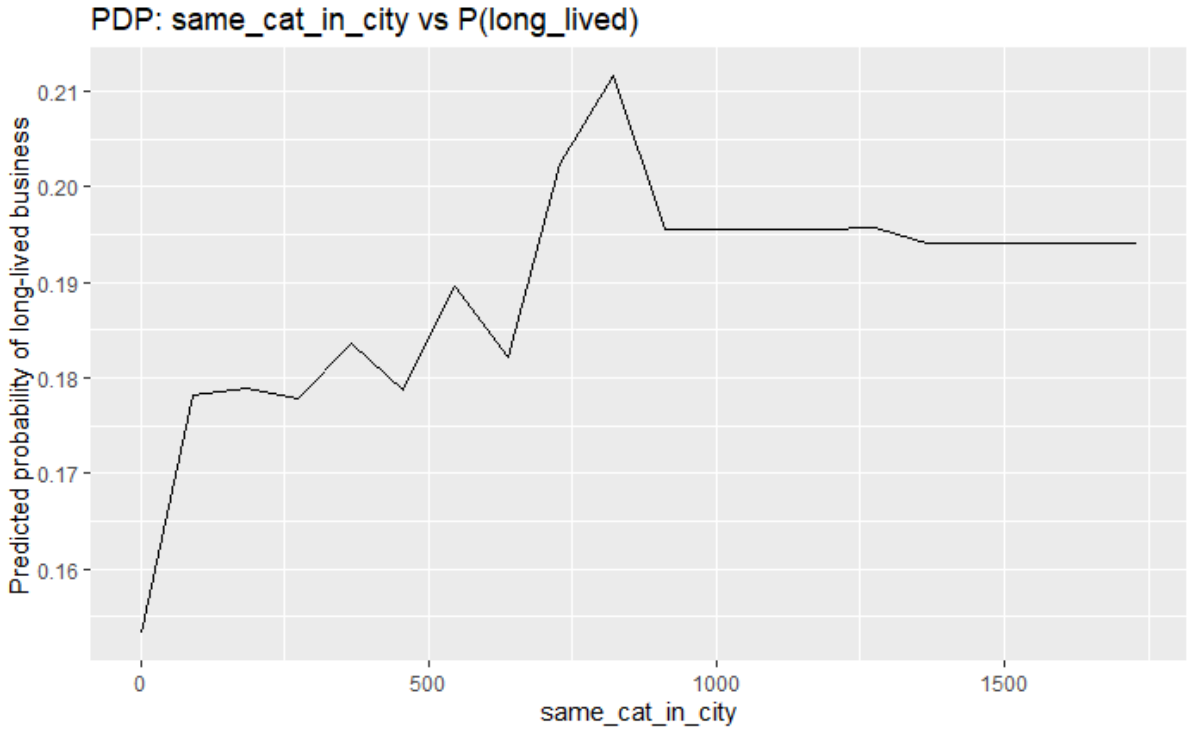
This ordering in variable importance is intuitive. Businesses with more reviews and more early check-ins tend to be those that attract and retain customers, which naturally correlates with longevity. Being currently open and being part of a chain both capture elements of **organisational resilience** and brand recognition. Local competition and star rating still matter, but less than direct measures of engagement and scale.



The PDP for early popularity closely mirrors the LOESS curve in Figure 3: the predicted probability of being long-lived increases sharply from low to moderate early check-in counts, then plateaus and exhibits diminishing marginal gains at high values. This confirms that the random forest is indeed exploiting the nonlinear structure seen in the raw data.



The PDP for local competition shows only mild curvature, consistent with Figure 4. The predicted probability of being long-lived changes gradually as the number of same-category businesses in the city increases, but there are no sharp turning points. This suggests that competition contributes primarily as a **small adjustment** to the baseline probability, rather than as a strongly nonlinear driver.



The RQ2 results show that:

1. Some of the features encode **strong nonlinear relationships** with survival (especially early popularity), which linear models cannot fully capture.
2. Nonlinear models (RF, GBM) can leverage these patterns to deliver **moderately better predictive performance**, without sacrificing interpretability thanks to PDPs and simple raw plots.
3. Engineered features **do exhibit nonlinear patterns**, but the **extent of nonlinearity is heterogeneous** across features, and nonlinear models turn those **feature-specific** patterns into moderate but meaningful gains in predicting which businesses survive the longest.

From a practical point of view, business owners or investors could use these findings to:

- Flag businesses with very low early popularity as at-risk and potentially offer support or targeted advertising.
- Recognize that being in a highly competitive area is not automatically bad; in this dataset, competition on its own has weak nonlinear effects.
- Treat chain status as a genuine structural advantage when assessing long-term survival prospects.

To understand how competitive density influences restaurant closure and how this effect varies across price levels, a set of supervised learning models using both linear and nonlinear techniques were developed. The main predictors included the engineered competition features (n_city_restaurants and n_same_cat), the extracted price_level, and key control variables such as review count and average rating.

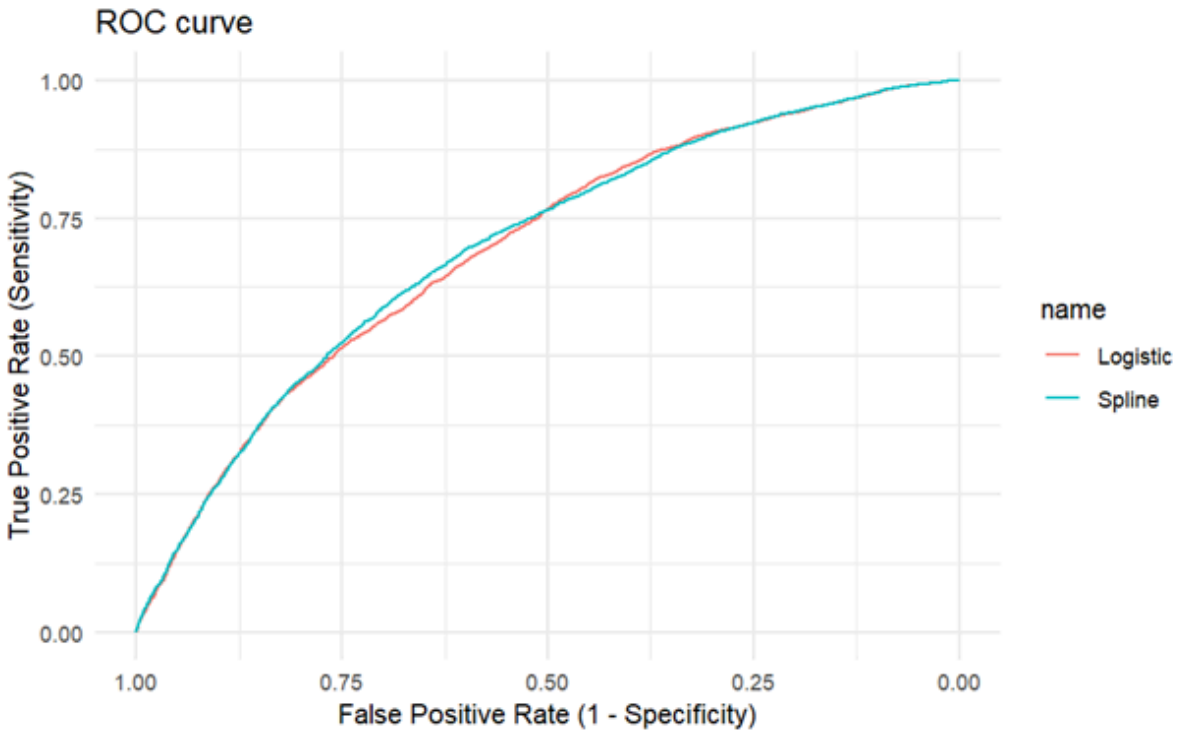
1. *Logistic Regression*

Interaction terms between competition variables and price level were included to capture the possibility that restaurants in different pricing segments face distinct competitive pressures. The fitted model showed a **positive association** between competition and the probability of closure, with stronger effects for mid- and high-priced restaurants compared to budget restaurants. On the held-out test set, the logistic model achieved an accuracy of about **0.68** and an AUC of **0.69**. This indicates that, while the model does better than random guessing, there is still substantial unexplained variation in closure outcomes.

2. *Spline Regression*

Exploratory work suggested that the effect of competition might not be purely linear, especially for higher-end restaurants. To allow for curvature, we fit a logistic regression with **natural splines** on n_city_restaurants, while still interacting competition with price_level.

The spline model produced slightly better test performance than the linear logistic model, with accuracy around **0.68** and AUC increasing to roughly **0.70**. More importantly, the spline-based partial effects showed that the predicted probability of closure rises as city-level restaurant density increases across all price levels, but the slope is much steeper for expensive. Cheap restaurants (price level 1) showed relatively flatter curves, suggesting that they are less sensitive to competitive saturation.



The ROC curves in the above plot compare the predictive performance of the logistic and spline logistic models. Both models perform meaningfully better than random guessing, but the spline logistic regression shows a consistent, though modest, improvement across most thresholds.

Summary:

Across all modeling approaches, the results consistently showed that competitive density is a significant predictor of closure, but its impact differs substantially across price tiers. Budget restaurants (Price Level 1) appeared resilient even in highly competitive environments, while mid-priced (Level 2-3) and high-end restaurants (Level 4) face increasing closure risk as competition intensifies.

Nonlinear models offered clearer insight into these dynamics, outperforming simple logistic regression and highlighting complex patterns not captured by linear specification. The combination of interaction terms, spline functions, and tree-based models provided a robust understanding of how market competition and pricing strategy jointly influence business survival.

Limitations

While the logistic and spline logistic models provided useful insight into how competition and price level relate to restaurant closure, several limitations remain. First, although we limited the main analysis to parametric models for interpretability, more flexible machine learning methods such as Random Forests and Gradient Boosting Machines showed stronger predictive performance in our exploratory work. These models were able to capture nonlinear interactions and complex feature relationships that the logistic family could only approximate. Incorporating them directly into the main analysis could provide deeper insight into the structure of closure risk.