Network-based quality index aggregation in the retail location problem. A supervised learning approach

Virginia Ahedo, José Ignacio Santos, José Manuel Galán

https://doi.org/10.37610/0njk0c03

Abstract

In retailing, the location problem is a fundamental strategic aspect. It is usually formalized as a multi-criteria optimization problem to choose the most appropriate spot. A relevant element in the selection is the adequacy of the commercial ecosystem in the vicinity of the location. To account for this criterion, there are different primary indices based on networks that formalize the quality of the available options with regard to the surrounding ecosystem. Previous research suggests that aggregating the different indices using a classifier can improve the quality of these metrics. In this paper, we compare different classifiers to assess their performance in that respect. The analysis has been performed in a context of transfer knowledge and information fusion using data from all the cities in Castile and Leon, Spain. Our results show that the random forest and generalized linear models obtain results significantly superior to other alternatives.

Keywords

Complex networks, retail location problem, prediction, knowledge transfer, classification, pattern recognition.

1. Introduction

Economic activity is not spatially homogeneously distributed (Krugman 1991). Understanding the pattern of this geographical distribution and the reasons behind it is a crucial issue in economics and management. This knowledge facilitates the optimization of location choices and the creation of suitable institutional policies by political and economic decision-makers.

While this problem is relevant to companies of all sizes and industries, it is particularly significant in the case of retailing. In retail stores, the choice of location is deemed the most crucial strategic decision, as competitors cannot imitate it exactly (Zentes, Morschett & Schramm-Klein 2012). Furthermore, despite the growing importance of other distribution channels, in-store sales continue to be their primary source of revenue (Berman, Evans & Chatterjee 2018).

In retailing, there are numerous factors to consider when selecting the most suitable location from the available options. Some important considerations include, but are not limited to, the socioeconomic characteristics of the local population, population density, accessibility, and the size of the store.

Due to the multitude of factors affecting the decision, the location problem is commonly formalized as a Multi-Criteria Decision-Making (MCDM) problem (Çoban 2020; Shaikh, Memon, Prokop & Kim 2020). A significant determinant of the decision is the presence/absence of competitors and/or of complementary commercial activities in the neighboring area. In planned shopping centers, an adequate balanced tenancy is a fundamental element from the very initial design. However, in unplanned primary and secondary areas, it is also crucial to quantify the suitability of specific locations in terms of the commercial ecosystem, as it will also play a critical role.

This assessment of the surrounding ecosystem and its effect on a particular retail store is not straightforward, since there are several mechanisms with opposing effects that can take place. For instance, a neighborhood with many competitors can decrease the local market power of an individual store; yet, at the same time, the presence of several competitors in the area may increase its overall appeal to potential customers, allowing them to compare and make more efficient choices (Konishi 2005). This would only account for the effect of businesses within the same industry. The impact of complementary businesses can be even more challenging to assess and quantify.

✉ Virginia Ahedo (1)
vahedo@ubu.es
https://orcid.org/0000-0002-9812-388X

José Ignacio Santos (1)
jisantos@ubu.es
https://orcid.org/0000-0002-6653-043X

José Manuel Galán (1)
jmgalan@ubu.es
https://orcid.org/0000-0003-3360-7602

(1) Universidad de Burgos, Departamento de Ingeniería de Organización, Escuela Politécnica Superior, Ed. A1, Avda. Cantabria s/n 09006, Spain

In the literature, several quality indices have been proposed to quantify the interactions between business categories based on networks of the commercial structure (Jensen 2006; Sánchez-Saiz et al. 2022). These indices make various assumptions about the global and local commercial structure of a city, capturing complementary elements between them. Their use as inputs of different aggregation strategies has recently been shown to significantly improve predictive quality and to outperform the individual use of the different indices in location recommender systems (Ahedo, Santos & Galan 2021). However, the influence of the specific aggregation algorithm or classifier used has not yet been comparatively analyzed. Although there are numerous algorithms and out-of-the-box predictive strategies available for a wide range of problems, determining which one is most appropriate depends on the nature of the specific problem at hand (Wolpert 2002).

In this paper, we comparatively assess the predictive capacity of various classification algorithms to determine the business category of a retail store based on the network-based quality indices of the business in its neighborhood. The prediction of a high-performing classifier for different business categories can serve as an indicator of the attractiveness of the location for each of these activities, providing a quantitative decision-making tool. In particular, we have focused on solving the problem from the perspective of knowledge transfer and information fusion, by leveraging knowledge of the data and commercial structure of several cities and aggregating it through network consensus techniques. This approach generates a training dataset, from which the algorithms learn the patterns, which are then used to evaluate predictions on a test set from a different city.

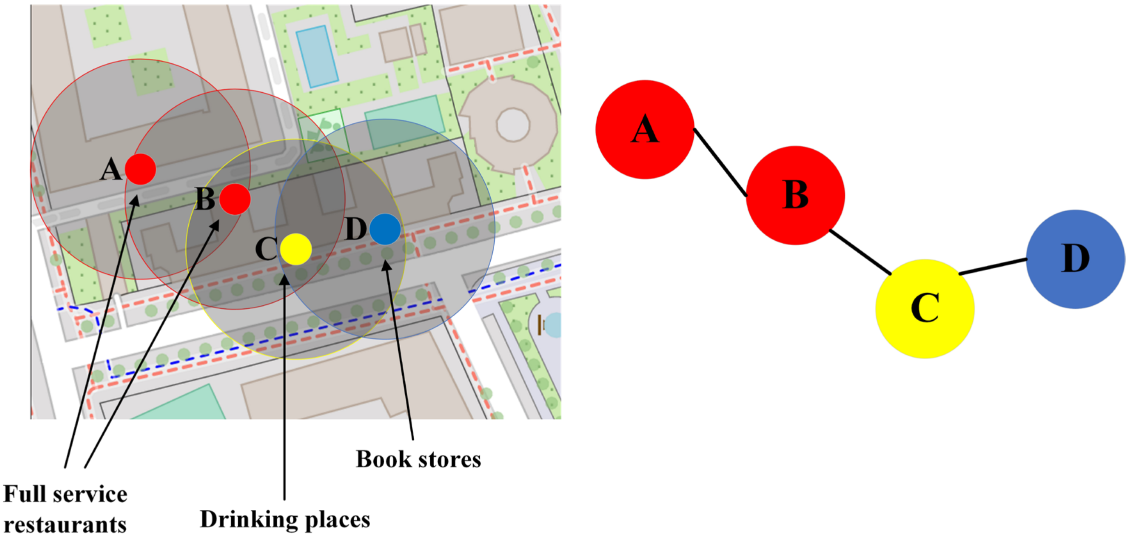
The structure of this paper unfolds as follows: In the subsequent section, we impart a theoretical background pertaining to the retail location problem, examining it through the lens of a complex network perspective based on suitable commercial environments, and elucidate the various approaches in a comprehensive manner. Following this, we present the design of a computational experiment through which we empirically analyze whether aggregating different primary metrics, using various machine learning algorithms, enhances the capture of localization patterns. The results and discussions are subsequently provided, and ultimately, conclusions are drawn in the final section.

2. Theoretical background

2.1. Primary metrics description

The basic methodologies for identifying the suitability of locations based on the interactions between businesses are network-based and include Jensen, permutation, and rewiring (Jensen 2006, 2009; Gómez, Jensen & Arenas 2009; Sánchez-Saiz et al. 2022). These methodologies construct an interaction network of retail stores in which the nodes are the stores in the city and links are created between them if they are within a given proximity radius (typically 100 meters in previous research) (see Fig. 1).

Figure 1 Interaction network between retail stores. The network is generated from the geolocation of the different shops. Each store is represented in this network as a node. An undirected link is created between two stores if they are both within a distance of less than a certain radius (typically 100 m). Each node keeps as a label the commercial category to which it belongs.



Then, a second interaction network, in this case between the different business categories (bakeries, pharmacies, restaurants, etc.) is obtained from the first network. This second network is a weighted and signed graph, with high positive weights for categories that are likely to be nearby and negative weights for categories that do not typically coexist in the same neighborhood (see Fig. 2). The different methodologies (Jensen, permutation and rewiring) differ in their approach to deriving the second network from the first one.

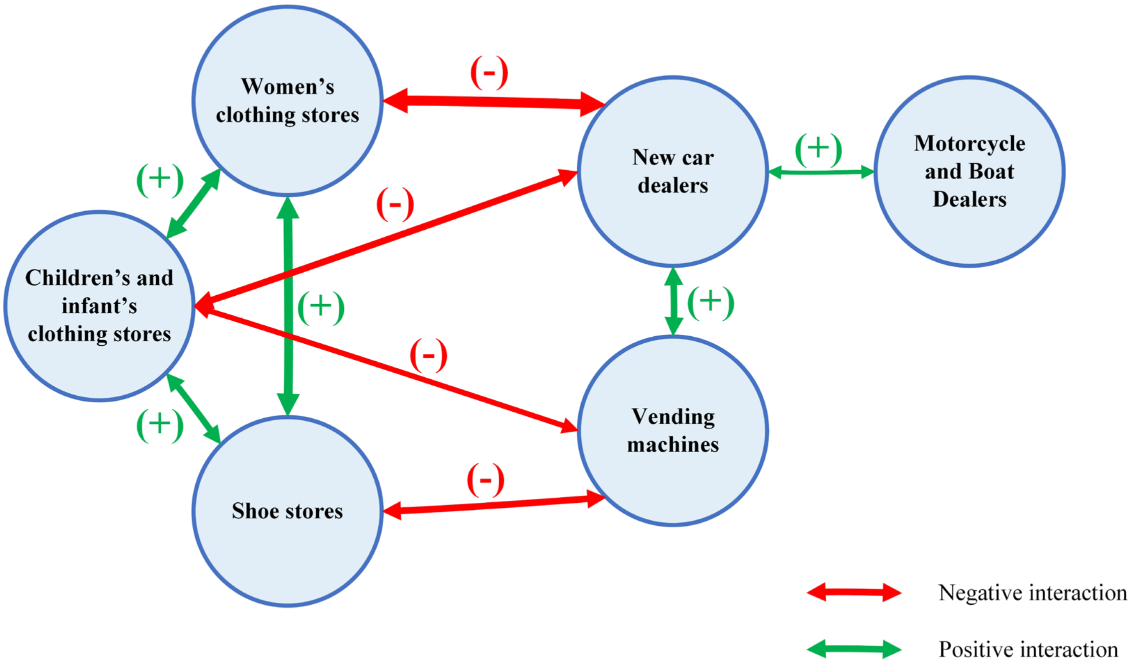


Figure 2 Interaction network between business categories. It is used to calculate the attraction or repulsion between different commercial categories. These relationships can be either positive (green links in this simplified example) or negative (red links in the figure). The links representing the relationships are not only signed but weighted (thickness in the figure), and the weight determines the strength of the relationship (attraction or repulsion). Depending on the method used, these relationships may or may not be symmetrical. Note that only statistically significant relationships between categories are included in the network. In the cases where no relationship exists between the nodes, neither a positive nor a negative link is created.

In Jensen's approach (Jensen 2006, 2009; Gómez, Jensen & Arenas 2009), the relationship between two categories is calculated differently depending on whether the commercial typology is analyzed with respect to itself or to a different category. In the first case, the intra-category coefficient defined according to (1) is calculated:

$$M_{AA} = \frac{|T|-1}{|A|(|A|-1)} \sum_{a \in A} \frac{N_A(a,r)}{N_T(a,r)} \quad [1]$$

Where T is the set of all the stores in a given city, A is the set of the stores that belong to category A, and $N_s(p,r)$ represents the number of stores in set S within a radius r from shop p.

In the case of the inter-category coefficient, equation (1) is slightly modified to obtain (2). This second coefficient summarizes the relationship between business typologies A and B.

$$M_{AB} = \frac{|T|-|A|}{|A||B|} \sum_{a \in A} \frac{N_B(a,r)}{N_T(a,r)-N_A(a,r)} \quad [2]$$

The coefficients previously defined to measure the interactions between different business categories have values higher than 1 when the empirical interaction is higher than expected, and values lower than 1 otherwise. Before constructing the final network, these inter- and intra-category coefficients are transformed by taking logarithms, thus converting values greater than one into positive interactions and values less than one into negative ones. The strength of the empirical interactions is then compared to different null models through Monte Carlo simulation to determine their statistical significance. For those interactions that are not statistically significant, the links are removed from the category network, which is equivalent to setting to zero their value in the matrix representation of the network.

Jensen's work (Jensen 2006, 2009; Gómez, Jensen & Arenas 2009) inspired the rewiring and permutation approaches (Sánchez-Saiz et al. 2022). Notwithstanding, they differ fundamentally in three elements: (i) in both rewiring and permutation, the calculation of the interaction is based on the sum of the number of links joining the different categories in the primary network (stores network); (ii) the sign of the interaction is not obtained by taking logarithms, but from the value of the Z-score function (3); and (iii) the null model used to determine the statistical significance of the empirical relationships differs in each case. In the permutation method, the business structure of the city is fixed, and the business categories are permuted among the different locations; recall that the number of stores from each category is maintained. On the other hand, in the rewiring method, the local environment of each business (the number of neighbors it has, i.e., its degree) is maintained by disconnecting the links of each node of the primary network and rewiring them randomly (Sánchez-Saiz et al. 2022) (see Fig. 3).

In equation (3) x_{AB} represents the empirical number of links obtained between retail stores from category A and retail stores from category B, and $x_{AB}^{null}{}_{model}$ and $s_{AB}^{null}{}_{model}$ correspond to the mean and the standard deviation of the null distribution of the number of links between these two categories.

$$Z_{AB} = \frac{x_{AB} - \overline{x_{AB}^{null\_model}}}{s_{AB}^{null\_model}} \qquad [3]$$

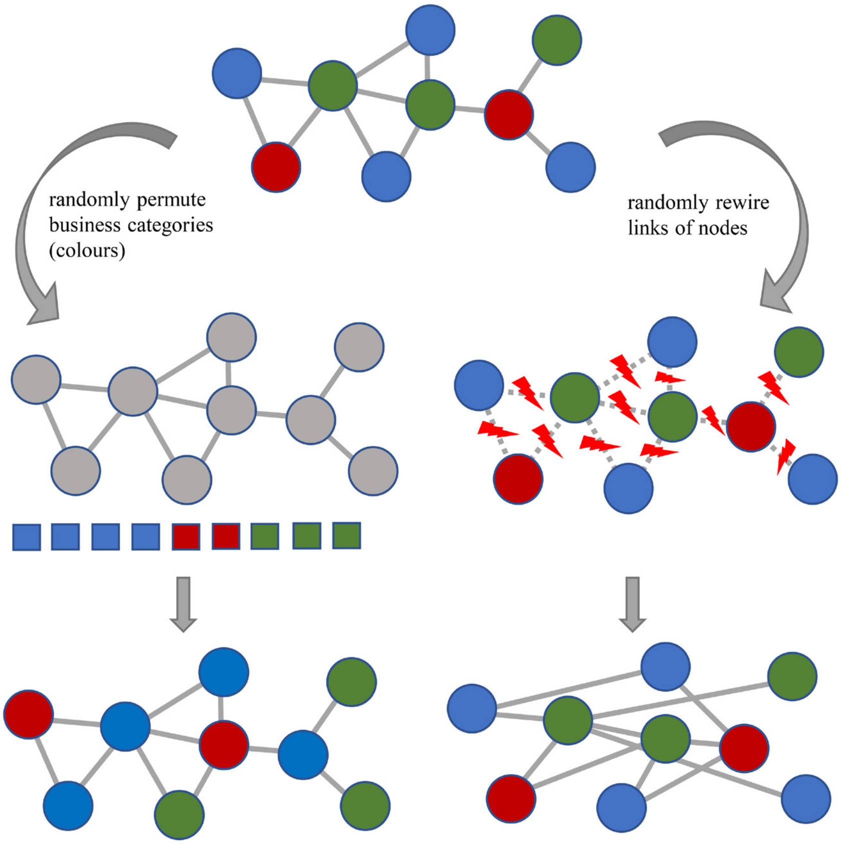

randomly permute business categories (colours)
randomly rewire links of nodes

Figure 3 To determine the statistical significance of the relationships, the empirical values are compared with different null models. In the permutation model (left), the commercial network is fixed, and it is the different commercial categories that are randomized by permuting their labels among the different nodes of the network. In the rewiring model (right), the number of links of each node (its degree) is maintained but each link is cut in two halves and the different ends are randomly paired.

Once the different relationships between commercial categories are known for any of the three methods, two alternative types of quality indices can be calculated to comparatively assess the suitability of different locations.

In Jensen's original formulation, it is assumed that the quality of a given location depends on how closely the proportion of stores in the neighborhood resembles the ideal distribution obtained in the business categories network. From this assumption, Jensen's Quality Index of a particular location (x,y) for activity i is defined as follows (Jensen 2006)

$$Q_{JENSEN_i}(x,y) \equiv \sum_{j=1}^{N} a_{ij}\left(nei_{ij}(x,y) - \overline{nei_{ij}}\right) \qquad [4]$$

In equation (4) N denotes the total number of different business categories, nei_{ij} (x,y) indicates the number of neighbor stores from category j that exist around (x,y) (it is assumed that (x,y) belongs to category i), $\overline{nei_{ij}}$ is the average number of neighbors of category j that the stores of type i have, and $a_{ij}=\log(M_{AB})$ is the corresponding value of the Jensen's matrix of interactions between commercial categories. This quality index can be generalized directly to the permutation and rewiring methods by simply changing the source of the weighting factor a_{ij}. Instead of being taken from Jensen's matrix of interactions between business categories, the weighting factor can be taken from the permutation and rewiring matrices of interactions between categories, respectively.

A different set of quality indices are the so-called Raw Quality Indices, which again can be calculated for Jensen, permutation and rewiring. The calculation of raw indices assumes that the quality of a given location is not given by its similarity to the ratio empirically found, but by the number of neighboring stores with which there is a positive interaction, and the absence of stores with which there is an empirically negative relationship. Formally:

$$Q_{JENSEN-RAW_i}(x,y) \equiv \sum_{j=1}^{N} a_{ij}\left(nei_{ij}(x,y)\right) \qquad [5]$$

Again, the calculation of the different raw quality indices only requires taking the weighting factors a_ij from the corresponding matrix of each method.

2.1. Succinct literature review and problem statement

Research into the spatial organization of retail commercial activities has wielded significant influence over the location problem. The initial approach, focused on employing location data, business categories, and network analysis to determine quality indices regarding the optimal location for stores, has been successfully implemented in recommendation systems and decision-support tools (Jensen 2006, 2009). This methodology, coupled with the reformulation of modularity to facilitate the analysis of community structures in correlated and signed data networks (Gómez, Jensen & Arenas 2009), has been used in conjunction with metaheuristic optimization techniques as a recommendation system (Sánchez-Saiz, Galán & Santos 2014). Furthermore, it has supported decision-making tools for locating well-known food chains in New York, such as Starbucks, McDonald's, and Dunkin' Donuts (Karamshuk, Noulas, Scellato, Nicosia & Mascolo 2013). Proving its efficacy, it has positioned itself as a metric with a high predictive capacity, sometimes used together with other features obtained from Foursquare or other additional social networks such as Facebook (Lin et al. 2016) or Baidu (Xu et al. 2016; Chen, Chen & Chen 2020).

The work of Sánchez-Saiz et al. (2022) enlarges the set of metrics and indices designed to capture varying patterns within a robust core of business relationships, employing consensus techniques. In more recent research (Ahedo, Santos & Galán 2021; Ahedo, Santos & Galán 2023), various studies have demonstrated improvement in predictive accuracy by combining several indices via supervised learning models. However, the empirical identification of the best algorithms from a systematic predictive perspective has yet to be addressed.

In this research, we systematically analyze the predictive capacities of the most popular machine learning algorithms within scientific discourse, employing them as information fusion tools. This study is accomplished by cohesively integrating information gleaned from all established primary metrics in the literature above. Our methodology is empirically rooted, obtaining the results across the datasets of several cities.

3. Computational experiment design

3.1. Dataset

The experiments were conducted using retail store location data from the nine provincial capitals of Castile and Leon, an autonomous community located in northwestern Spain. The size of the cities ranged from the largest, with almost 300,000 inhabitants, to the smallest, with approximately 40,000 inhabitants, according to data from 2017, when the dataset was collected. The dataset used for the study was created from the Yellow Pages, taking business category and address information, and was subsequently georeferenced using the MapQuest Application, Open Street Map data, and Google Maps API. Importantly, it is publicly available at: (Sánchez-Saiz et al. 2021). In order to create the primary networks of interactions between retail stores for each city, a 100-meter radius was established. Then, to create the networks between retail categories, the different business typologies were categorized using the North American Industry Classification for Small Business (NAICS) to enable comparisons with previous research (Jensen 2006, 2009; Gómez, Jensen & Arenas 2009; Ahedo, Santos & Galan 2021; Sánchez-Saiz et al. 2022). Note that the NAICS classification system includes 68 different business categories.

3.2. Performance metric and information fusion

To calculate the comparative performance of the different algorithms, we use the Mean Reciprocal Rank (MRR) (Voorhees 1999; Radev, Qi, Wu & Fan 2002), which corresponds to the average of the reciprocal ranks of the results obtained for a number of instances Q:

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i} \qquad [6]$$

The choice of the MRR as the performance metric in our experiment is based on the nature of the problem, as the predicted label for each location represents suitability but is not a binary answer. To clarify this idea, consider that the presence of a certain type of business at a given location shows that the site is suitable for that category, but it does not necessarily mean it is not suitable for other types of businesses. Hence, to evaluate the algorithms, we aim for the empirically observed categories to be at the top of the recommendation ranking provided by the algorithm, but not necessarily the first (which would be measured by metrics such as accuracy).

In our analysis, given that the training data come from different cities, each with potential interaction particularities, a decision had to be made regarding how to combine the information coming from several sources. Specifically, to solve this information fusion problem we have used consensus network techniques, which were proposed precisely in the context of retailing networks. In particular, we use the consensus networks of relationships (Sánchez-Saiz et al. 2022). In this type of aggregation, from the matrices/networks of significant interactions obtained in each city, a new matrix is created by combining all of them in the following way: for each pair of business relationships, the relationship in each city is analyzed, adding +1 for each city in which the relationship is found to be positive and statistically significant, -1 if the relationship is negative and statistically significant, and 0 otherwise. The resulting matrix (according to each of the three different methods: Jensen, rewiring, and permutation) is then used as a reference for calculating the quality indices.

Also relevant in consensus techniques is whether to use some kind of threshold such that, if the number of significant relationships does not reach a certain value, they are made zero in the consensus matrix. Importantly, although these threshold techniques can unveil the core of the network and, hence, the most important commercial relationships, previous work has shown that their application is counterproductive from a predictive perspective (Ahedo, Santos & Galan 2021). Consequently, we have not used any thresholds in our analysis.

3.3. Classifiers and hyperparameter optimization

The calculation of the different quality indices yields a dataset with six quality indices for each commercial category (6 indices/category x 68 categories = 408 quality indices). Since, although preliminarily, it has been shown that the combined use of all of them allows taking advantage of their complementarity (Ahedo, Santos & Galan 2021), in this paper, we comparatively analyze the capacity of different supervised learning algorithms to combine in an aggregate form the information provided by all the quality indices taken together. Nowadays, it is not known which is the best classifier for the specific application context of our study, that is, for predictive use in general, and for knowledge transfer in particular. (Recall that by knowledge transfer we mean that data obtained in some cities is used to make predictions in other cities).

Our experiment design involves using data from eight cities as training set to make predictions (knowledge transfer) about the ninth city (test set). This structure is then rotated to obtain nine results for each classifier.

To select the appropriate hyperparameters for each model, which can have a significant impact on the algorithm's performance, we use 5-fold cross-validation. This method involves dividing the training data into five roughly equal parts, using four of them to train the algorithm with different combinations of the parameters, and evaluating the performance of the classifier on the data from the fifth part (validation set). This process is repeated five times, rotating the training and validation parts each time. The performance attained with each set of hyperparameters is then compared using the aggregated data from the five evaluations, and, eventually, the optimal hyperparameters are selected. Finally, the model is retrained on the entire training dataset and evaluated on the test set (see Annex 1 for more details).

Due to the exponential growth in computational time with the number of hyperparameters in a grid search strategy, we adopted a random grid search strategy. This search method has been shown to be more efficient, as similar optimization results are often obtained in much less computational time (Bergstra & Bengio 2012).

The classifiers used include most of the state-of-the-art algorithms for tabular datasets. Specifically, the following algorithms have been included in the analysis:

- Random Forest: The random forest algorithm is an ensemble technique based on the bagging strategy —bootstrap aggregation— (Breiman 2001). This technique combines the results of multiple weak classifiers, typically deep and unpruned decision trees, which are trained on different bootstrapped samples. In addition, the algorithm also uses the random subspace method in the training of each tree to decorrelate as much as possible the different weak learners. This algorithm often performs well because it reduces variance, is resistant to overfitting and correlation between regressors, and can work with nonlinear patterns and interactions (James, Witten, Hastie & Tibshirani 2013). The hyper-parameters optimized in the cross-validation stage of our experiments have been the number of predictors considered at each split of the trees and the total number of trees.

- Naive Bayes: this is a probabilistic classifier based on Bayes' theorem. Despite being a very basic algorithm that makes the restrictive assumption that, given a class, each feature is independent of any other, it has been shown to generate good classification results (Kupervasser 2014). Moreover, given its training simplicity, it is often used as a benchmark algorithm. In this algorithm, the only hyperparameter tuned in the cross-validation process has been the Laplace parameter.

- Deep Learning: We used a classifier based on a multi-layer feedforward artificial neural network and trained with stochastic gradient descent using back-propagation. This neural network is suitable for tabular data as opposed to other types of deep neural network algorithms such as Convolutional Neural Networks and Recurrent Neural Networks more oriented to image processing or temporal data. However, despite its extraordinary performance, one of the problems of this type of classifier is the number of optimization parameters to adjust, which can make the search much more intensive than in other algorithms. In our case, to make the results comparable in training time with the rest of the algorithms, we have focused on optimizing just the number of layers and neurons in each layer, testing with configurations between 2 and 5 layers and a number of neurons between 5 and 200.

- Gradient boosting machine (GBM): this is a classifier that, like the random forest, uses an ensemble technique, in this case, boosting (Friedman 2001). Specifically, the idea is to improve the classification of a weak learner (also often regression and/or classification trees) by using successive additional classifiers that refine the errors produced by previous learners. There are different versions of the technique. In our case, we have used the algorithm described in Hastie et al.(2009) and implemented in the h2o R package (H2O. ai 2020). Although this algorithm generally produces outstanding results with tabular data, it has quite a few hyperparameters to optimize, which renders it difficult to use. In the process, the maximum depth of the trees, the number of trees, the learning rate, and the learning rate annealing have been tuned in the validation set.

- Generalized Linear Model (GLM): this is a family of regression and classification techniques that generalize linear regression models for outcomes following exponential distributions. To do so, it relates the target variable to the model through a link function and allows the magnitude of the variance of each observation to be a function of its predicted value (Hastie, Tibshirani & Friedman 2009). In our case, the response variable is modeled as a multinomial distribution. The optimization process of the hyperparameters using the validation set is performed employing the regularization variables. The alpha values have been analyzed in the range 0-1, hence considering ridge regression and the Lasso in the extreme values, and elastic nets in the intermediate cases (Tibshirani 1996; Zou & Hastie 2005). As for the lambda value, it is optimized with the training data for each alpha value. GLMs have been used previously to try to combine indices with good results (Ahedo, Santos & Galán 2023).

- Stacking: apart from boosting and bagging, there is a third ensemble technique for combining different classifiers known as stacked generalization or stacking (Wolpert 1992). Unlike the other strategies, which use the same type of classifier for combination, stacking consists of combining different classifiers to generate an aggregate classifier as accurate as possible. The process is organized in two hierarchical levels, a first level in which different base classifiers are trained, and a second level in which another classifier acts as a meta-learner trying to learn when to use one classifier or another, or when to combine several classifiers to make the final classification (in our case the meta-learner is a GLM model). This type of strategy is often quasi-optimal (Matlock, De Niz, Rahman, Ghosh & Pal 2018; Ghasemian, Hosseinmardi, Galstyan, Airoldi & Clauset 2020; Martin, Ahedo, Santos & Galan 2022). However, depending on the number and type of base classifiers, its computational cost can be very high. To make the analysis comparable and fair with the other techniques, we have equalized the computation time (one hour of parallel execution on each of the eight cores for each dataset) and limited the number of base classifiers to fifteen.

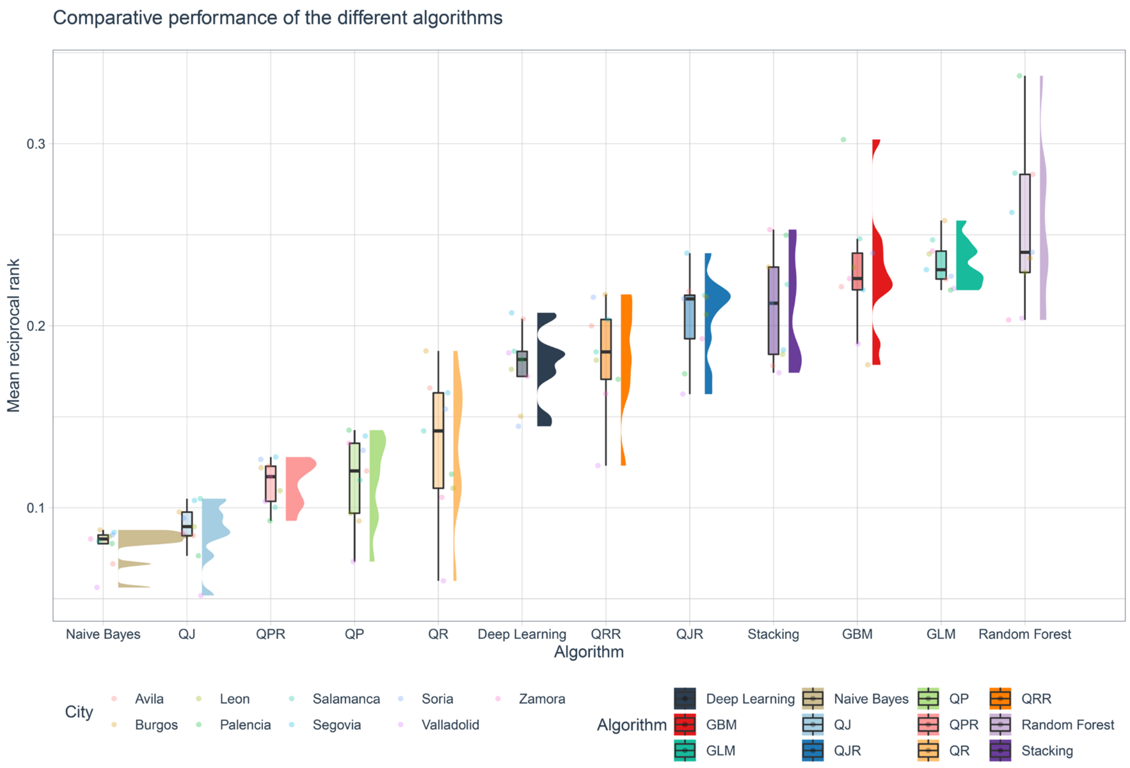
- Along with the classification algorithms, we have also used the six quality indices based on non-aggregated networks separately to quantify the improvement obtained, if any, when using a classifier that combines them together. These are Quality Jensen (QJ), Quality Permutation (QP), Quality Rewiring (QR), Quality Jensen Raw (QJR), Quality Permutation Raw (QPR) and Quality Rewiring Raw (QRR).

4. Results and discussion

4.1. Computational results

The results of the analysis are presented in Fig. 4. The data shows a significant degree of variation depending on the city analyzed. This is not unexpected, as the predictability of each city is influenced by factors such as the commercial organization and geographical specialization of that city, which, in turn, are often determined by the city's size. Smaller cities, for example, tend to be less commercially specialized and, as a result, tend to be less predictable (Sánchez-Saiz et al. 2022).

Figure 4 Comparative performance of the different algorithms. Each dot represents the MRR obtained for each city. The algorithms and metrics analyzed are Deep Learning, Random Forest, Naive Bayes, Stacking, Generalized Linear Model (GLM), Gradient boosting machine (GBM), Quality Jensen (QJ), Quality Permutation (QP), Quality Rewiring (QR), Quality Jensen Raw (QJR), Quality Permutation Raw (QPR) and Quality Rewiring Raw (QRR).



It is also observed that there are several algorithms — especially the random forest, the generalized linear model, and the gradient boosting machine— that improve the results obtained by the primary quality indices. However, the results in Fig. 4 may be misleading since the observations in each case are not independent but paired (the same cities are used for each algorithm). This may cause significant differences between the algorithms to be visually blurred as the overlap in the performance ranges may not be due to the algorithm itself but to the influence of the particular city. Therefore, we used the Friedman test to better analyze whether there are significant differences between the algorithms. This test is a non-parametric alternative to ANOVA when the data are paired. The results are shown in the following table:

Table 1 Friedman test results.

Friedman χ^2	Degrees of freedom	*p*-value
87.838	11	4.417e-14

From Table 1, it can be concluded that there are significant differences between at least two of the algorithms. However, these results are incomplete without trying to identify the algorithms between which significant differences in performance exist. To do so, we have performed post-hoc comparisons between the different algorithms using pairwise Wilcoxon rank-sum tests to compare between group levels, with corrections for multiple testing and paired data. Given the relatively small number of cities, we used the correction proposed by Benjamini & Yekutieli (2001) based on the false discovery rate, a less restrictive condition than the corrections based on the family-wise error rate. The results are presented in Fig. 5. The p-value has been rounded to the second decimal place. Cells colored in grey are comparison values that show no significant differences. Cells colored in the red-white-green range are p-values that show differences at the 0.1, 0.05, and lower significance levels.

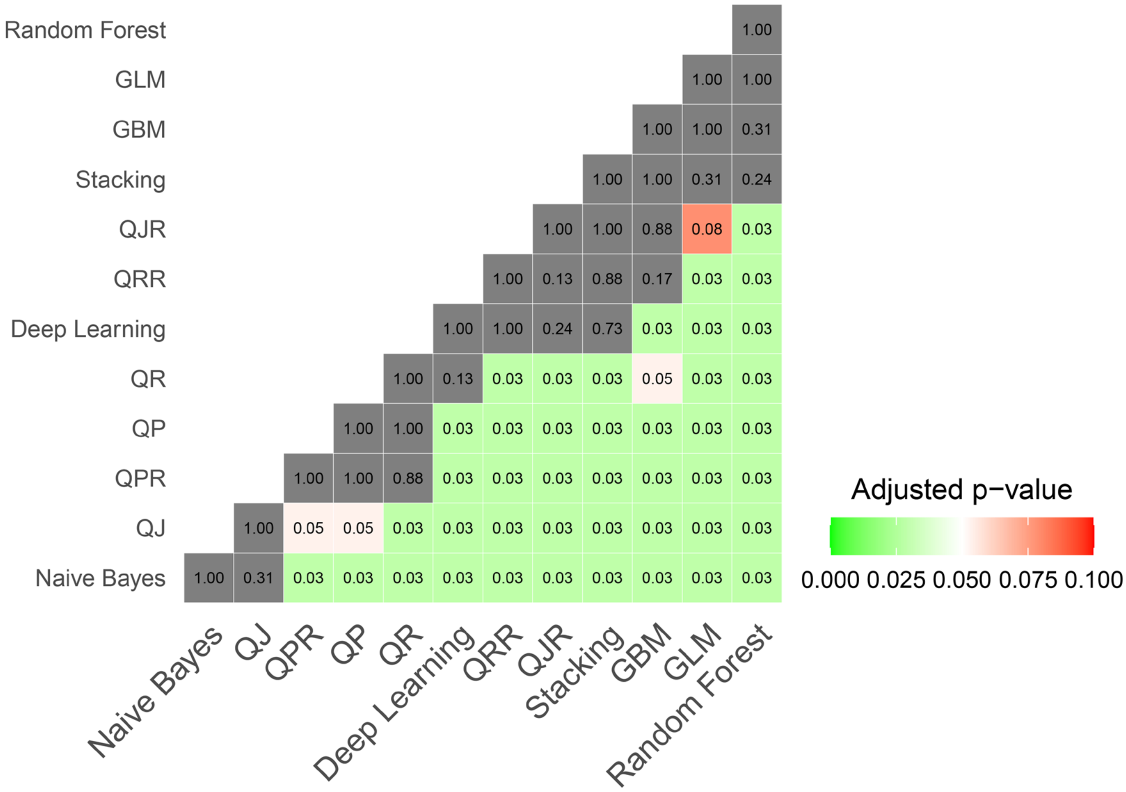


Figure 5 P-values of the pairwise Wilcoxon rank-sum tests comparing the performance of each pair of the algorithms. The algorithms and metrics analyzed are Deep Learning, Random Forest, Naive Bayes, Stacking, Generalized Linear Model (GLM), Gradient boosting machine (GBM), Quality Jensen (QJ), Quality Permutation (QP), Quality Rewiring (QR), Quality Jensen Raw (QJR), Quality Permutation Raw (QPR) and Quality Rewiring Raw (QRR).

The results reveal significant differences between the random forest and all the primary metrics, clearly demonstrating that its use, which allows to aggregate all of them, is an effective strategy to improve prediction. In the case of GLMs, the difference found with the best of the primary techniques is only significant at the 0.1 level. The Naive Bayes results are poor since the independence hypothesis is violated in this problem. The results for stacking, Deep learning, and GBM are relatively modest in our analysis. However, this may be due to the experimental setup in which training time limits have been imposed on the problem.

4.2. Discussion and Managerial Implications of the Results

Navigating the retail location decision-making process requires a comprehensive, multi-criteria approach. One relevant aspect is adequately quantifying the commercial ecosystem, without forgetting many others, such as population density, the size of the location, access, the socio-economic level of the neighborhood, etc. This research presents important advancements in extracting more accurate quantitative insights, thereby capturing the complexities of this commercial dimension with greater precision.

Managers should consider several key takeaways from our findings when dealing with location decisions. Firstly, engaging deeply with data on commercial interactions and utilizing existing and new quality indices to evaluate this crucial dimension is vital to better capture commercial patterns. The integration and information fusion of these quality indices through algorithms, such as Random Forest and GLM, stands out as a viable strategy to enhance spatial suitability assessments, particularly regarding balanced tenancy. Utilizing models capable of effectively integrating information allows for significantly improving metrics without the need for additional data, which is normally costly to obtain, by simply exploiting information more efficiently.

However, while our results are promising, managers must exercise caution when applying these insights to different contexts or cities. It is imperative to validate the relevance and accuracy of the quality indices and geolocated data for their specific region or city. It is ideal to utilize data from the city in which the evaluation will be addressed to capture the cultural and commercial specificities that the location might have. But if this is not possible, and data transferred from other cities are used, the transferability of data and insights between cities needs to be carefully scrutinized, ensuring alignment in critical aspects such as culture, demographics, consumption patterns, and market dynamics.

This research not only represents a significant step towards more accurately quantifying and evaluating the commercial ecosystem in location problems but also underscores the necessity for managers to adopt a data-informed, quantitative approach when maneuvering through complex, multi-criteria decision-making processes. It offers a foundational roadmap for managers, enabling more informed strategic planning and decision-making in retail location.

5. Conclusions

The location problem is complex, as it has multiple dimensions. One of the dimensions that is not easy to quantify, but is considered relevant in the decision, is the adequacy of the existing commercial ecosystem to the activity to be located. Different proposals and quality indices based on networks of commercial interactions try to evaluate this aspect. Previous work suggests that it is possible to improve the performance of the different primary indices by using them together as inputs in a classifier. In addition, since obtaining and processing data for new cities is usually more expensive than running computational experiments on previously obtained data, in this contribution we have explored the combination of quality indices in the context of transfer knowledge, i.e., using data from other cities to evaluate the effect on another city, which is assumed to be unknown only at the global level.

Our results comparing different algorithms show significant differences between the random forest and the GLM with the rest of the primary indices, thus making their results a better proxy of the suitability of the different location alternatives. Hence, the use of different quality indices aggregated by means of GLM or random forests in location recommendation systems and/or multi-criteria decision tools improves the evaluation of spatial fitness with respect to balanced tenancy.

However, our work has also some limitations. Although there are very few papers using extensive geolocated data from other cities, the number of observations in our datasets is relatively low for the power of some tests to capture statistically significant differences. In addition, in our experiments, by design, computational limits have been set to the training and optimization of the hyperparameters of the algorithms. This may favor algorithms with simpler optimization processes such as random forest and GLM. Other algorithms such as DL, stacking, or GBM may improve their performance with extended training and optimization times.

Notwithstanding the above, our results are relevant and show that using specific algorithms to aggregate individual quality indices results in significant improvements, as very high mean reciprocal ranks are obtained in a robust and straightforward manner.

A pivotal future research path for this work involves exploring the relationship between model accuracy and computational cost to understand the efficiency of each aggregation approach for this problem. The objective is to scrutinize not only the precision of different algorithms and combinations of quality indices but also to evaluate the efficiency ratio, considering the improvement in accuracy relative to computational cost. Such an analysis could unveil crucial insights into whether marginal improvements in accuracy justify potentially high computational costs.

Moreover, exploring model interpretability and recommendation explainability, alongside investigating the dynamics and evolution of commercial ecosystems within cities over time, could be foundational in enhancing the practical application of location recommendation systems. Implementing dynamic machine learning approaches that adapt to changes in input data and developing methods that allow users to comprehend the rationale behind model recommendations might be vital for maintaining model relevance and accuracy over time and ensuring the practical adoption of these tools in decision-making. Additionally, a deeper analysis of knowledge transfer between divergent cities could provide insights into the limits and opportunities of knowledge transfer approaches in this domain, investigating how city characteristics impact transfer efficacy and developing methods to adapt models to new cities efficiently.

Acknowledgements The authors acknowledge financial support from the Spanish Ministry of Science, Innovation and Universities (Excellence Network RED2018-102518-T), the Spanish State Research Agency (PID2020-118906GB-I00/AEI/10.13039/501100011033) and the Fundación Bancaria Caixa D. Estalvis I Pensions de Barcelona, La Caixa (2020/00062/001). In addition, we acknowledge support from the Santander Supercomputación group (University of Cantabria), that provided access to the Altamira Supercomputer —located at the Institute of Physics of Cantabria (IFCA-CSIC) and member of the Spanish Supercomputing Network— to perform the different simulations/analyses.

References

AHEDO, V., SANTOS, J. I. & GALAN, J. M. (2021). «Knowledge Transfer in Commercial Feature Extraction for the Retail Store Location Problem». IEEE Access, 9, pp. 132967–132979, doi: 10.1109/ACCESS.2021.3115712.

AHEDO, V., SANTOS, J. I. & GALÁN, J. M. (2023). «Combining Quality Indexes in the Retail Location Problem Using Generalized Linear Models». In: Lecture Notes on Data Engineering and Communications Technologies. Springer, pp. 47–52, doi: 10.1007/978-3-031-27915-7_9

BENJAMİNİ, Y. & YEKUTİELİ, D. (2001). «The Control of the False Discovery Rate in Multiple Testing under Dependency». The Annals of Statistics, 29(4), pp. 1165–1188.

BERGSTRA, J. & BENGİO, Y. (2012). «Random Search for Hyper-Parameter Optimization». J. Mach. Learn. Res., 13(null), pp. 281–305.

BERMAN, B. R., EVANS, J. R. & CHATTERJEE, P. M. (2018). Retail Management. A Strategic Approach. Pearson.

BREİMAN, L. (2001). «Random Forests». Machine Learning, 45(1), pp. 5–32, doi: 10.1023/A:1010933404324.

CHEN, Y. M., CHEN, T. Y. & CHEN, L. C. (2020). «On a method for location and mobility analytics using location-based services: a case study of retail store recommendation». Online Information Review, doi: 10.1108/OIR-10-2017-0292.

ÇOBAN, V. (2020). «Solar energy plant project selection with AHP decision-making method based on hesitant fuzzy linguistic evaluation». Complex & Intelligent Systems, 6(3), pp. 507–529, doi: 10.1007/s40747-020-00152-5.

FRİEDMAN, J. H. (2001). «Greedy Function Approximation: A Gradient Boosting Machine». Annals of Statistics, 29(5), pp. 1189–1232.

GHASEMİAN, A., HOSSEİNMARDİ, H., GALSTYAN, A., AİROLDİ, E. M. & CLAUSET, A. (2020). «Stacking models for nearly optimal link prediction in complex networks». Proceedings of the National Academy of Sciences of the United States of America, 117(38), pp. 23393–23400, doi: 10.1073/pnas.1914950117.

GÓMEZ, S., JENSEN, P. & ARENAS, A. (2009). «Analysis of community structure in networks of correlated data». Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, 80(1), p. 16114, doi: 10.1103/PhysRevE.80.016114.

H2O.Aİ (2020). h2o: R Interface for H2O. R package version 3.30.0.6. https://github.com/h2oai/h2o-3.

HASTİE, T., TİBSHİRANİ, R. & FRİEDMAN, J. (2009). The Elements of Statistical Learning. 2nd ed. New York, NY: Springer.

JAMES, G., WİTTEN, D., HASTİE, T. & TİBSHİRANİ, R. (2013). An Introduction to Statistical Learning. First. Springer Science & Business Media.

JENSEN, P. (2006). «Network-based predictions of retail store commercial categories and optimal locations». Physical Review E, 74(3), p. 035101, doi: 10.1103/PhysRevE.74.035101.

JENSEN, P. (2009). «Analyzing the Localization of Retail Stores with Complex Systems Tools». In: Adams, N. M., Robardet, C., Siebes, A., & Boulicaut, J.-F. (eds.) Advances in Intelligent Data Analysis VIII. Springer Berlin Heidelberg, pp. 10–20, doi: 10.1007/978-3-642-03915-7_2.

KARAMSHUK, D., NOULAS, A., SCELLATO, S., NİCOSİA, V. & MASCOLO, C. (2013). «Geo-Spotting: Mining Online Location-based Services for Optimal Retail Store Placement». In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 793–801, doi: 10.1145/2487575.2487616.

KONİSHİ, H. (2005). «Concentration of competing retail stores». Journal of Urban Economics, 58(3), pp. 488–512, doi: 10.1016/j.jue.2005.08.005.

KRUGMAN, P. (1991). Geography and Trade. London, UK: MIT Press.

KUPERVASSER, O. (2014). «The mysterious optimality of Naive Bayes: Estimation of the probability in the system of "classifiers"». Pattern Recognition and Image Analysis, 24(1), pp. 1–10, doi: 10.1134/S1054661814010088.

LİN, J., OENTARYO, R., LİM, E.-P., VU, C., VU, A. & KWEE, A. (2016). « Where is the Goldmine?: Finding Promising Business Locations through Facebook Data Analytics». In: Proceedings of the 27th ACM Conference on Hypertext and Social Media - HT '16. New York, New York, USA: ACM Press, pp. 93–102, doi: 10.1145/2914586.2914588.

MARTİN, O., AHEDO, V., SANTOS, J. I. & GALAN, J. M. (2022). «Comparative study of classification algorithms for quality assessment of resistance spot welding joints from pre- and post-welding inputs». IEEE Access, pp. 1–1, doi: 10.1109/ACCESS.2022.3142515.

MATLOCK, K., DE NİZ, C., RAHMAN, R., GHOSH, S. & PAL, R. (2018). «Investigation of model stacking for drug sensitivity prediction». BMC Bioinformatics, 19(Suppl 3), p. 71, doi: 10.1186/s12859-018-2060-2.

RADEV, D. R., Qİ, H., WU, H. & FAN, W. (2002). « Evaluating Web-based Question Answering Systems». In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02). Las Palmas, Spain: European Language Resources Association.

SÁNCHEZ-SAİZ, R.M., GALÁN, J. M. & SANTOS, J. I. (2014). «Localization Based on Business Interactions Through a Simulated Annealing Algorithm». In: Managing Complexity. Springer, pp. 325–331, doi: 10.1007/978-3-319-04705-8_38.

SÁNCHEZ-SAİZ, R. M., AHEDO, V., SANTOS, J. I., GÓMEZ, S. & GALÁN, J. M. (2021). «Dataset of the retailing location networks in the cities of Castile-Leon, Madrid and Barcelona», doi: 10.36443/10259/5585.

SÁNCHEZ-SAİZ, R. M., AHEDO, V., SANTOS, J. I., GÓMEZ, S. & GALÁN, J. M. (2022). «Identification of robust retailing location patterns with complex network approaches». Complex & Intelligent Systems, 8(1), pp. 83–106, doi: 10.1007/s40747-021-00335-8.

SHAİKH, S. A., MEMON, M. A., PROKOP, M. & KİM, K. S. (2020). «An AHP/TOPSIS-based approach for an optimal site selection of a commercial opening utilizing geospatial data». Proceedings - 2020 IEEE International Conference on Big Data and Smart Computing, BigComp 2020, pp. 295–302, doi: 10.1109/BigComp48618.2020.00-58.

TİBSHİRANİ, R. (1996). «Regression Selection and Shrinkage via the Lasso». Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58(1), pp. 267–288, doi: 10.2307/2346178.

VOORHEES, E. M. (1999). «TREC-8 Question Answering Track Report». In: Proceedings of the 8th text retrieval conference. National Institute of Standards and Technology, pp. 77–82.

WOLPERT, D. (1992). «Stacked Generalization». Neural Networks, 5, pp. 241–259.

WOLPERT, D. H. (2002). «The Supervised Learning No-Free-Lunch Theorems». In: Roy, R., Köppen, M., Ovaska, S., Furuhashi, T., & Hoffman, F. (eds.) Soft Computing and Industry. London: Springer London, pp. 25–42.

XU, M., WANG, T., WU, Z., ZHOU, J., Lİ, J. & WU, H. (2016). «Store Location Selection via Mining Search Query Logs of Baidu Maps», doi: 10.48550/arXiv.1606.03662.

ZENTES, J., MORSCHETT, D. & SCHRAMM-KLEİN, H. (2012). Strategic Retail Management. Wiesbaden: Gabler Verlag.

ZOU, H. & HASTİE, T. (2005). «Regularization and variable selection via the elastic net». Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), pp. 301–320, doi: 10.1111/j.1467-9868.2005.00503.x.

Annex 1

Algorithm	Grid Search strategy	Other parameters
Deep Learning	Architectures for the 'hidden' layers were tested, including [5, 5, 5, 5, 5], [10, 10, 10, 10], [50, 50, 50], [100, 100, 100], and [200, 200]	epochs=10; rho=0.99; rate=0.005;
GBM	The 'max_depth' of the trees was tested at various levels: 4, 6, 8, 12, 16, and 20, while the number of trees ('ntrees') was explored at 3000 and 10000. The learning rate ('learn_rate') was evaluated at 0.02 and 0.05,	learn_rate_annealing = 0.99; stopping_rounds = 5; stopping_tolerance = 1e-4; stopping_metric = "AUC";
GLM	Optimizing the alpha parameter, which controls the balance between L1 and L2 regularization. The grid was defined to explore a sequence of alpha values ranging from 0 to 1, incremented by 0.01, thus ensuring a detailed search across the entire plausible spectrum of the elastic net mixing parameter.	
Random Forest	The 'ntrees' parameter was tested at 200, 500, 750, and 1000 to identify the optimal number of trees in the forest. 'mtries' was varied from 10 to 100 in increments of 10 to determine the best number of variables randomly sampled as candidates at each split. 'max_depth' was explored from 10 to 30 (inclusive) with a step of 10, while 'min_rows' was adjusted from 1 to 3 to optimize the minimum number of observations per leaf. Additionally, 'nbins', representing the number of bins for the histogram to build, was varied between 20 and 30 in steps of 10. Lastly, 'sample_rate' was evaluated at 0.55, 0.632, and 0.75 to assess different fractions of the training data to be used for learning.	
Naive Bayes	The laplace parameter was crafted to explore a sequence of laplace values spanning from 0 to 10, discretized into 50 equidistant steps	