EARTH OBSERVATION AND GEOMATICS ENGINEERING
website: https://eoge.ut.ac.ir

Utilizing location-based social network data for optimal retail store placement

Hoorsana Damavandi[1], Neda Abdolvand[1]*, Farid Karimipour[2]

[1] Department of Management, Faculty of Social Science and Economics, Alzahra University, Tehran, Iran
[2] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Iran

Article history:
Received: 14 January 2019, Received in revised form: 29 August 2019, Accepted: 16 September 2019

ABSTRACT

Finding an optimized place is undeniably a momentous subject in establishing the marketing strategies of a retail store. Based on the existing literature, the process of selecting an optimized location for a business can be defined as a ranking problem that compares and rates existing or potential sites based on their ability to attract customers. Consequently, this article is concentrated on the evaluation of machine learning ranking methods in ranking existing retail stores based on the data derived from LBSNs. Using feature engineering techniques, we defined and calculated a set of features for 239 retail store branches in Tehran, from the venue data obtained from the Foursquare API. Additionally, we derived a rank for each store representing store popularity via user-generated data from Foursquare, Dunro, and Google Maps. Next, we implemented a number of classification and "learn-to-rank" algorithms to rate these stores. Finally, by evaluating the prediction precision and ranking precision of the algorithms used, we analyzed the fit and prediction power of all ranking algorithms. The outcomes of this research suggest that most algorithms used are, in fact, reliable methods for ranking retail store sites. Therefore, such algorithms can be used as a technique for retail store site selection, given a list of existing or potential sites for a store. Additionally, our results clearly suggest a superiority in the ranking precision of "learn-to-rank" algorithms for retail store placement. Out of all algorithms used, with a ranking precision of 0.854, MART is the most powerful algorithm for ranking retail store sites.

KEYWORDS
Retailing
LBSN
Geomarketing
Feature Selection
Machine Learning

1. Introduction

Selecting an optimal place for a store is one of the important aspects of strategic business planning (Aboulola, 2018). When aiming for appealing to the target market, a store's prosperity can be enhanced through careful planning of the marketing mix elements; product, price, promotion, and place (Kotler & Armstrong, 1989). Planning for the place element is particularly crucial for a retail store. *"No matter how good it is offering, merchandising, or customer service, every retail company still has to contend with three critical elements of success: location, location, and location"* (Taneja, 1999, P:136-137). Considering its geographical nature, the problem can be solved by the application of geospatial analytics concepts. Geospatial analytics is the intersection of geographic analysis, business intelligence, and data visualization (Ting et al., 2018).

By investigating the existing literature, the computational techniques for store location selection can be classified into "Traditional" and "Modern data-driven" approaches (Damavandi et al., 2018). While traditional techniques have been around and widely used for the past century, more practical approaches have been introduced in the past few years. The extension of Wi-Fi communications and GPS-equipped mobile devices, along with the emergence and acceptance of location-aware services and the increasing popularity of social networks, resulted in the introduction of location-based social networks (LBSNs) (Kheiri et al., 2016). Consequently, a myriad of easily accessible

* Corresponding author
E-mail addresses: hdamavandi.phd@ivey.ca (H. Damavandi); n.abdolvand@alzahra.ac.ir (N. Abdolvand); fkarimipour@ut.ac.ir (F. Karimipour)
DOI: 10.22059/eoge.2020.271740.1041

geospatial big data became available to researchers, leading to a new age of purely data-driven research methodologies in Geography and its related topics (Miller & Goodchild, 2016). Voluminous, fast, and various data that can be transformed into maps and offer information about the location of shapes and their geographic characteristics are referred to as geospatial big data. Of all various sources generating such data, by recording the social and spatial preferences of users, location-based social networks (LBSNs) can be deemed as one of the richest options available. Drawing on the special characteristics of LBSN data and the analytical potential of machine learning algorithms, researchers have proposed new solutions for store placement in the past few years (Yu et al., 2013). On the other hand, the question of finding an optimal place for a store can be defined as ranking a set of existing or potential places. By exploiting the ranking power of machine learning algorithms, past researchers have tried to rank stores based on a number of features pertaining to a business's popularity.

To sum up, it is obvious that researchers have shown an inclination towards machine learning algorithms as a solution for store placement problems. However, to the best of our knowledge, a machine learning approach has never been used in order to assess the popularity of retail stores based on their location. Hence, the primary goal of this paper is to tackle the retail store placement question via data mining techniques. In addition, none of the modern data-driven methods for store placement have ever been used in Iran. Seeing that the popularity of a retail store based on its location may have some correlations with the structure of a city, and cities as complex networks are affected by local variables including cultural and religious norms, the other goal of this study is testing the reliability of these modern techniques in the city of Tehran. Thereby, this paper begins with a review of various techniques used for store placement and an exploration of the properties of LBSN data and their academic applications. It proceeds by explaining the proposed model and methodology for retail store placement via feature selection and a description of the extraction, preprocessing, and preparation of the required data. Finally, by applying a set of machine learning algorithms including regression, classification, and learn-to-rank methods, the relative rank is predicted for each store, and by evaluating the precision of prediction and ranking of each algorithm, a comprehensive framework for retail location selection is presented.

2. Literature review

The question of placing retail stores across the network of a city in a way that optimizes the overall sales and customer attraction has been of interest to researchers, managers, and other planning authorities for many years (Damavandi et al., 2018). Central Place Theory (Christaller, 1933), Spatial Interaction Theory (Reilly, 1929-1931), and the theory of Minimum Differentiation (Hotelling, 1929) are considered as the main pillars of traditional methods for retail placement (Brown, 1993; Litz, 2008). The *Central Place Theory* is mainly concerned with attempts to characterize the regionalization of urban space in a hierarchical manner. It originates from Christaller's theory indicating that there is a reverse relationship between the distance from the source of supply and the demand for a product (Arcaute et al., 2015; Brown, 1993). Many scientists have since tried to exploit the main concept of this theory for retail location selection (Bacon, 1991; R. Johnston, 1968; R. J. Johnston, 1966; Nakamura, 2014; Potter, 1981). Arguing that distance from rivals is more important than distance from customers, the *Theory of Minimum Differentiation* originates from the claims of Hotelling (1929). The Space Syntax theory (Hillier & Hanson, 1984), Natural Movement (Hillier et al., 1993), and Multiple Centrality Assessment (Porta et al., 2009) are examples of research domains influenced by the theory of Minimum Differentiation. Also known as Gravity Models, the theories that are based on the *Spatial Interaction Theory* are greatly distinguished among spatial analysts. This theory arose from the assertions of Reilly (1927), emphasizing the importance of a customer's perception of accessibility and availability of retail stores. Wilson's entropy-maximization model and Huff's probabilistic potential model are the most accepted modified versions of the Spatial Interaction theory (Damavandi et al., 2018; Litz, 2008). However, despite their widespread use for nearly a decade, shortcomings such as being time-consuming, relying on traditional sources of data, and unrealistic assumptions have propelled scientists to look for better solutions. The immense granularity and easy access to spatial big data are among the reasons that make these methods perfect candidates for superseding these traditional techniques. The literature from the past few years, clearly suggests that relying on ranking methods has been one of the main approaches adopted by researchers for location selection. Karamshuk et al. (2013), looked at this problem from a Feature Selection perspective. For the first time, they defined a number of features based on the unique attributes of LBSN data to assess the popularity of food chains. They used Support Vector Regression (SVR), Decision Trees, Linear Regression, and RankNet learn-to-rank algorithm for feature selection and ranking. Xu et al. (2016), used a similar approach to rank a number of food chains and tool ware stores. They applied Linear Regression, Kernel Regression, SVR, Random Forests, Gradient Boosting Regression Trees (GBRT), and LambdaMART on data retrieved from LBSNs. Wang et al. (2016) used Ridge Regression along with SVR and GBRT to rank a number of restaurants based on their predicted popularity. Yu et al. (2016) defined the same problem as choosing a store from a list of candidates while trying to maximize the number of customers. They utilized Matrix Factorization, Logistic Regression, Bayesian

Classification, Decision Trees, and Support Vector Classification (SVC) to rank hypothetical stores. Rahman and Nayeem (2017), implemented SVM in order to find suitable places for live campaigns.

Reviewing the past literature clearly suggests that there is a gap in assessing the power of ranking techniques for retail store site selection. Since selecting an optimal place for a retail store is of especial importance and has been a question of interest for nearly a decade in the academic Marketing literature, this paper is focused on evaluating this new technique for location selection – the use of machine learning algorithms – for retail store site selection. Additionally, the results of past literature are not consistent in terms of introducing the most powerful algorithm for store placement because they all chose different algorithms to use. Therefore, we use all the classification and learn-to-rank algorithms used in the past literature to compare their ranking power. Table 1 and Table 2 demonstrate all the features and algorithms used in the mentioned articles and in this research. Some of the features used in the literature are defined based on the information retrieved from check-in data. The Foursquare API does not offer public access to such data, so the only way to obtain them is by using the Twitter API. Since Twitter is blocked in Iran, the extracted data from Twitter would not provide accurate information in terms of location, as Iranian users tend to use VPNs to access this website. Therefore, we only used the features from previous literature that could be calculated using the venue data offered directly by Foursquare. Additionally, as we redefined the question of retail store placement as a ranking problem, we were only interested in the classification algorithms used in the literature due to their ranking nature, and not the regression algorithms. We also used three learn-to-rank algorithms that are among the most popular and used information retrieval algorithms (Urbon-Bayes, 2017).

Table 1. An overview of all the features used in the related literature and this paper.

Feature \ Article	(Rahman & Nayeem, 2017)	(Xu et al., 2016)	(Wang & Chen, 2016)	(Yu et al., 2016)	(Karamshuk et al., 2013)	This Paper
Area Density	✓		✓	✓	✓	
Area Entropy	✓		✓	✓	✓	✓
Jensen's Quality[1]			✓		✓	
Competition		✓	✓	✓	✓	✓
Area Popularity	✓	✓			✓	✓
Transition Density					✓	
Incoming Flow					✓	
Transition Quality					✓	
Market Attractiveness			✓			
Market Competitiveness			✓			
Temporal Signal	✓					
Distance from Downtown		✓		✓		✓
Traffic Accessibility		✓		✓		✓
Complementarity				✓		✓

[1] According to Jensen (2006), there is a correlation between the presence of certain types of stores near one another. By calculating the effects of such occurances, the overall quality of an area is determined.

Table 2. An overview of all the algorithms used in the related literature and this paper.

Algorithm \ Article		(Rahman & Nayeem, 2017)	(Xu et al., 2016)	(Wang & Chen, 2016)	(Yu et al., 2016)	(Karamshuk et al., 2013)	This Paper
Traditional Classification/ Regression Algorithms	Linear Regression		✓			✓	
	SVR	✓		✓	✓	✓	
	SVC						✓
	Decision Trees				✓	✓	✓
	Logistic Regression				✓		✓
	Bayesian Classification				✓		✓
	Ridge Regression			✓			
	K-Nearest Neighbor		✓				✓
	Random Forests		✓				✓
	GBRT		✓				
Learn-to-rank Algorithms	RankNet					✓	✓
	LambdaMART		✓				✓
	MART						✓

3. Proposed model and methodology

Since the methodology used in this paper is a data mining approach, we propose a research model based on the widely used and accepted CRISP-DM[2] model for data mining and problem-solving. The proposed model in this paper is demonstrated in Figure 1. The datasets used in this paper were retrieved from LBSNs and LBSs as they are among the richest sources of spatial data. To this end, the venue data for each retail store was first extracted from the Foursquare API. With monthly over 55 active users in 2018, Foursquare has been the most popular LBSN around the globe. In general, Foursquare has a lot of similarities to other types of social networks. For instance, users are able to create personal profiles, share their preferences, and stay connected to their friends and acquaintances. With a closer look, it is apparent that its main purpose is offering unique services and value through enabling the sharing of location. Since many branches of the retail stores under study do not exist in Foursquare, a combination of data retrieved from Foursquare, Google Maps, and Dunro was used to extract store location, surrounding venues, and corresponding ranks.

Dunro can probably be considered as the Iranian version of Foursquare. Introduced by a team of 48 active developers of the Iranian startup ecosystem, Dunro has gained 300 thousand users and has been registered over 530 thousand local businesses since 2016 (Khajegiri, 2018). As the most used location-based service all over the world with over one billion users worldwide, Google Maps offers surveying services such as satellite imagery, street maps, real-time traffic information, navigation for pedestrians, etc. To obtain enough data for meaningful results, we considered all available branches of Shahrvand, Refah, Sepah, Etka, Ofogh Kourosh, Yaran Daryan, Canbo, Yas, Hyperstar, and HyperMe retail stores. Altogether, these retailers have approximately 354 branches in Tehran, from which 115 were eliminated for not having a ranking in any of the sources used for ranking extraction. Therefore, a total of 239 branches were left to be ranked. The distribution of these branches across Tehran is illustrated in Figure 2. Since Foursquare data are presented as JSON files, changing their format to Python is essential in this step.

[2] Cross-industry Standard Process for Data Mining

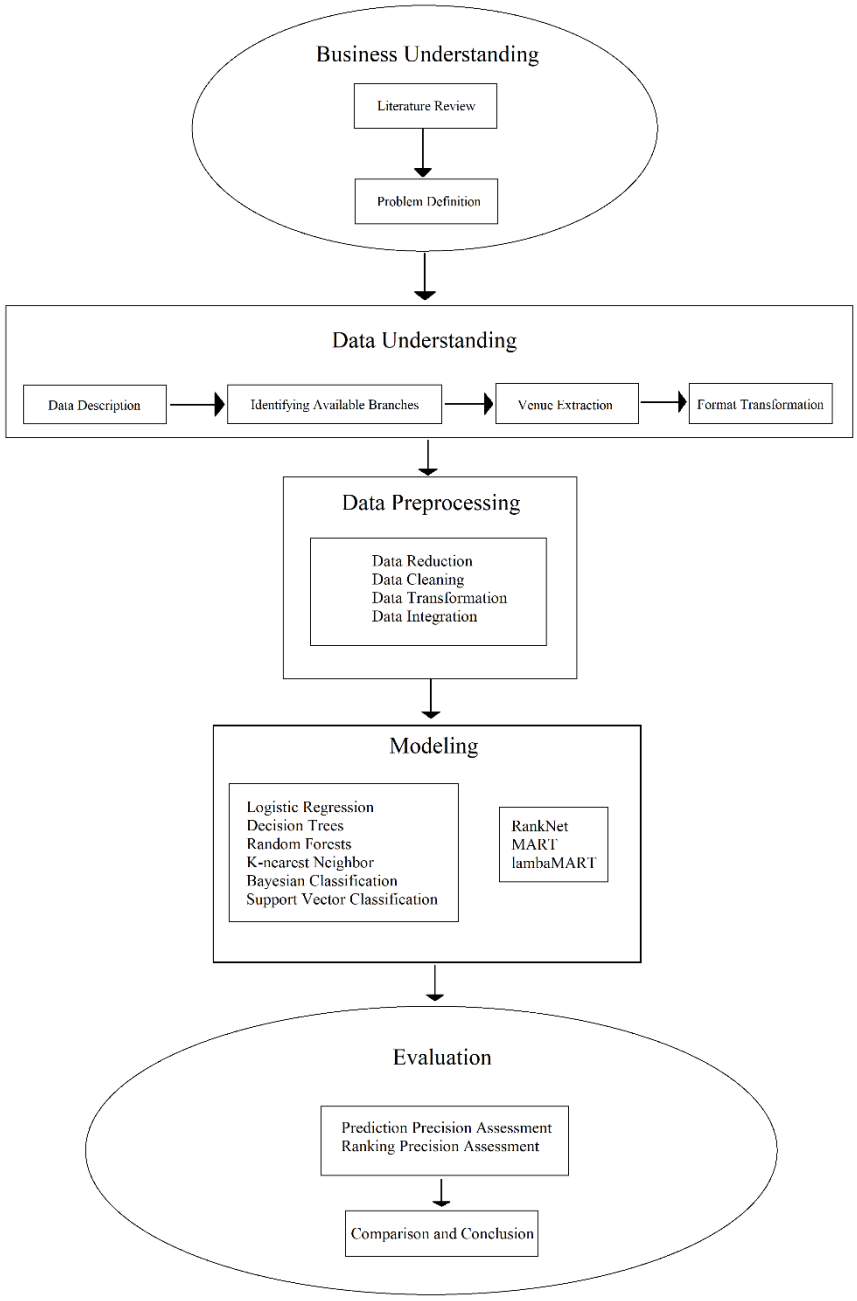

Figure 1. Research Flowchart Based On the CRISP-DM Model for Analytical Problems.

3.1. Data understanding

The datasets used in this paper were retrieved from LBSNs and LBSs as they are among the richest sources of spatial data. To this end, the venue data for each retail store was first extracted from the Foursquare API. With monthly over 55 active users in 2018, Foursquare has been the most popular LBSN around the globe. In general, Foursquare has a lot of similarities to other types of social networks. For instance, users are able to create personal profiles, share their preferences, and stay connected to their friends and acquaintances. With a closer look, it is apparent that its main purpose is offering unique services and value through enabling the sharing of location. Since many branches of the retail stores under study do not exist in Foursquare, a combination of data retrieved from Foursquare, Google Maps, and Dunro was used to extract store location, surrounding venues, and corresponding ranks. Dunro can probably be considered as the Iranian version of Foursquare. Introduced by a team of 48 active developers of the Iranian startup ecosystem, Dunro has gained 300 thousand users and has been registered over 530 thousand local businesses since 2016 (Khajegiri, 2018). As the most used location-based service all over the world with over one billion users worldwide, Google Maps offers surveying services such as satellite imagery, street maps, real-time traffic information, navigation for pedestrians, etc. To obtain enough data for

meaningful results, we considered all available branches of Shahrvand, Refah, Sepah, Etka, Ofogh Kourosh, Yaran Daryan, Canbo, Yas, Hyperstar, and HyperMe retail stores. Altogether, these retailers have approximately 354 branches in Tehran, from which 115 were eliminated for not having a ranking in any of the sources used for ranking extraction. Therefore, a total of 239 branches were left to be ranked. The distribution of these branches across Tehran is illustrated in Figure 2. Since Foursquare data are presented as JSON files, changing their format to Python is essential in this step.

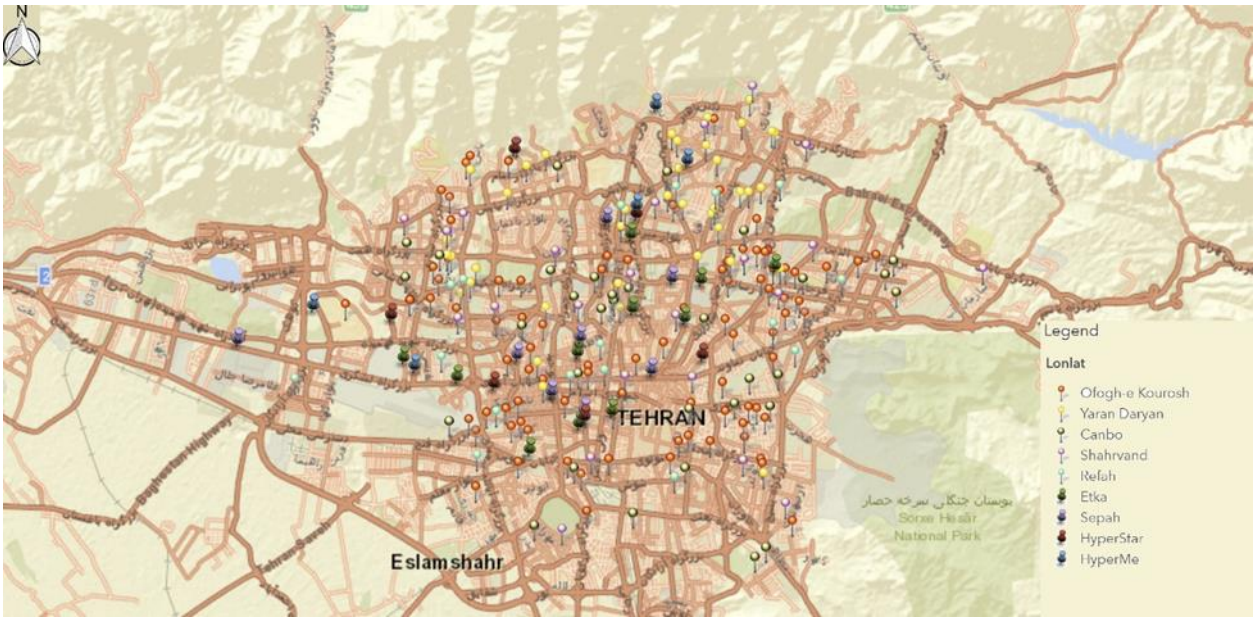

Figure 2- Distribution of branches of Retail Stores across Tehran.

3.2. Data preprocessing

Preprocessing is probably the most consequential step in any data mining effort since great algorithms are not able to effectively predict while applied to unprepared data. This step includes data reduction, data cleaning, data integration, and data transformation (Malley et al., 2016).

Data reduction includes decreasing the volume of data by separating the fields that are necessary for prediction and tossing the rest away (Malley et al., 2016). By doing so, calculations will require less computational power and, therefore, will be less time-consuming. In this paper, data retrieved from the Foursquare API included several fields of information, some of which were not needed in the process of feature engineering. Hence in this step, inessential fields were discarded.

The process of handling the disorder in data is referred to as data cleaning. The occurrence of missing values, outliers, and duplications are examples of the disorder. Such incidents can occur for a number of reasons pertaining to technical or human errors and blunders due to limitations in data collection tools (Malley et al., 2016). In Foursquare venue data, missing values can occur as undefined venue categories; noises occur when information in any of the fields is assigned incorrectly; outliers are occurrences of abnormal data that are not in line with the distribution of the rest of the data like places with extreme numbers of check-ins and duplications are incidents of multiple venues pertaining to the same place. To handle such inconveniences, we used an elimination strategy for missing values, the drop_duplicates command in python for duplicates, and relied on the built-in capabilities of machine learning algorithms to identify and deal with noises and outliers.

3.3. Data transformation

Based on our proposed model, data transformation includes normalization and feature engineering. Feature engineering, which is one of the most important steps in this article, includes defining a number of features based on the primary features extracted, in order to enhance the process of prediction.

Area popularity: the number of active users in a given area can be considered as an indicator of the popularity of that area (Xu et al., 2016) and can be calculated as follows where m is the number of users and C is a set containing all the users (Xu et al., 2016):

$$X_l(r) = \{m \in C : dist(m,l) < r\} \quad (1)$$

The data obtained from the Foursquare API include the total number of users that checked in at every single venue in defined proximity of the venue of interest. We used the summation of the number of users that checked in at each

venue in a 200-meter radius of each retail store, as an indicator of area popularity for each store (Xu et al., 2016).

Neighbor's entropy: the heterogeneity of the type of venues can affect the popularity of a store. In order to assess the effects of heterogeneity, the entropy formula of Information Theory can be used. The more the entropy of an area, the more various the types of its stores are (Karamshuk et al., 2013; Rahman & Nayeem, 2017; Wang & Chen, 2016; Xu et al., 2016; Yu et al., 2016). Entropy can be defined as follows where N(l,r) is the number of neighbors of a venue in an area with the radius of r, $N_\gamma(l,r)$ is the number of neighbors from type ɣ and τ is a set containing all the neighbors (Karamshuk et al., 2013):

$$X_l(r) = -\sum_{\gamma \in \tau} \frac{N_\gamma(l,r)}{N(l,r)} * \log\left(\frac{N_\gamma(l,r)}{N(l,r)}\right) \quad (2)$$

The venue data specifies the category of each venue as a separate feature. To extract area entropy for each retail store, we used the category feature to calculate the formula mentioned above.

Competition: to define this feature, we calculated the ratio of the number of retail stores, department stores, supermarkets and grocery stores to the number of all the venues in a given area (Karamshuk et al., 2013; Wang & Chen, 2016; Xu et al., 2016; Yu et al., 2016). This feature is defined as follows (Karamshuk et al., 2013):

$$X_l(r) = -\frac{N_\gamma(l,r)}{N(l,r)} \quad (3)$$

Based on the same category feature that was used to extract area entropy, we counted the total number of department stores, supermarkets and grocery stores in a 200-meter radius of each retail store, and divided the resulting number by the total number of venues around a given retail store to obtain area competition.

Distance from the city center: more popular places are often less distant from city centers (Xu et al., 2016; Yu et al., 2016). As a result, the reciprocal of the distance from the city center can be an indicator of popularity. As this paper tries to assess the popularity of retail stores in the city of Tehran, the reciprocal of the distance from the Grand Bazar of Tehran was calculated to account for the effects of distance from the city center, calculated as follows (Yu et al., 2016):

$$X_l(s) = \frac{1}{\log(d_s)} \quad (4)$$

The venue data also contained the exact longitude and latitude of the venue of interest. Therefore, we used the Haversine formula, each venue's coordinates, and the coordinates of the center of the Grand Bazar of Tehran to calculate the distance from each retail store to Tehran's business city center.

Traffic accessibility: one thing that is common between all approaches for location selection is that accessibility is undoubtedly considered as one of the major factors for customers in decision making. The number of transportation stations in a given area can be a good indicator of accessibility (Xu et al., 2016; Yu et al., 2016). Traffic accessibility can be calculated as follows (Yu et al., 2016):

$$X_l(s) = \frac{\log_2(N_{bus}(s,r)+1)}{\log_2(d_{bus})} + \frac{\log_2(N_{sub}(s,r)+1)}{\log_2(d_{sub})} \quad (5)$$

in which $N_{bus}(s,r)$ is the number of bus stations, $N_{sub}(s,r)$ is the number of metro stations, d_{bus} is the minimum distance from the nearest bus station and d_{sub} is the minimum distance from the nearest metro station. Using the category feature included in the venue data obtained for each retail store, we counted the number of bus stations and subway stations in a 200-meter radius of each store to extract the traffic accessibility feature.

Complementarity: the concept of complementary products and services is one of the widely used concepts in Economics and Marketing. Based on that concept, complementary businesses are defined as "Businesses that are not offering the exact products or services as ours but are offering a product/service that is related to ours and can be used by our customers." The presence of such businesses near a retail store may increase its attractiveness (Xu et al., 2016; Yu et al., 2016). According to Singh (2011), the presence of a parking lot near a retail store may have a complementary impact on its sales. Therefore, the number of dual selections of parking lots and retail stores in a given area is calculated to account for this feature. This relationship is measured as follows (Yu et al., 2016):

$$P_{t \to t^*} = \frac{2.N_{set}(t,t^*)}{N_T.(N_T - 1)} \quad (6)$$

where $N_{set}(t,t^*)$ is the number of possible dual selections of t and t*, and $N_T.\frac{N_T - 1}{2}$ is the number of all dual selections of venues. We used the category of venues within a 200-meter radius of each retail store, to calculate the number of existing pairs of parking lots and department stores close to the stores of interest.

By calculating all aforementioned features for each store, a set of features was constructed, which needed to be

normalized in order to prevent features with bigger numbers from having greater effects on the prediction outcome.

3.4. Data integration

Table 3 demonstrates the final results of the preprocessing steps, in which the dataset includes all the normalized features, ranks, and the transformation of ranks in terms of relevance scores.

Data integration includes the integration of data extracted from different sources and dealing with differences in labeling strategies and standards (Malley et al., 2016). Since all features were extracted from Foursquare, there was no need for the integration of features. On the other hand, the labels were retrieved from three different sources with different ranking ranges and weights and needed to be integrated properly. To tackle this problem, we first normalized all weights and ranks using the Maximin formula, and then obtained the final rank of each store by calculating the weighted average of the ranks obtained from the Foursquare application, Dunro and Google Maps. The integration of store ranks was executed through the following steps:

- Normalization of the number of people voting for each store based on the maximin formula.
- Normalization of the rate for each store based on the maximin formula.
- Calculating the rank of each store as follows:

$$Finalrank = W_{normal_G}.Rank_G + W_{normal_D}.Rank_D + W_{normal_F}.Rank_F \quad (7)$$

- Sorting stores ascendingly based on the calculated ranks.
- Splitting the sorted list into four equal parts.
- Assigning 0-3 relevance scores accordingly.

3.5. Validation

To evaluate the performance of the algorithms, we used two different sets of metrics, one for the evaluation of the prediction power of algorithms and one for the evaluation of their ranking power. To evaluate the prediction power, we used Precision, Recall, and F-measure for the classification algorithms, which are all considered as standard validation metrics for data mining techniques. For learn-to-rank algorithms, a precision@k metric is usually used to evaluate the prediction precision and is equivalent to the precision metric defined for classification algorithms.

Since a combination of different types of algorithms is being used for ranking, a uniform metric is needed to allow for general comparisons among the final predictions. Therefore, nDCG@k[3] was used for the evaluation of ranking precision, which is a standard Information Retrieval metric that measures the gain of each item by comparing its relative position in the predicted list to its actual rank (Xu et al., 2016). In order to calculate this metric, a list of relevance judgments is required. A relevance judgment list determines the relationships between items by assigning a relative rank of 0 3 to each item, ranking items as irrelevant, marginally relevant, fairly relevant, and highly relevant (Jarvelin & Kekalainen, 2002). The relative ranks of stores extracted from Foursquare, Dunro, and Google Maps were used as the relevance judgment list in this paper. DCG@k evaluated the precision by examining the first k items on the prediction list and specifying the percentage of these k items that were ranked correctly and thereby calculated the cumulative gain (Jarvelin & Kekalainen, 2002). DCG@k is calculated using Eq. 8 (Karamshuk et al., 2013)

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel(l_i)}}{\log_2^{i+1}} \quad (8)$$

After having a uniform and normalized set of features and labels, the last step before implementing the machine learning algorithms would be splitting the data into two sets for training and testing. Randomly sampling a subset of the dataset and setting it aside for testing would decrease the chance of overfitting (Gupta, 2017). There are a number of approaches for splitting the datasets. In this paper, we utilized the k fold cross validation technique for maximum precision. By splitting the dataset into k equal subsets, the validation iterated for k times, and each time a new subset was used for testing, minimizing variances, and decreased the probability of overlooking possible hidden patterns by utilizing all the data simultaneously for training and testing (Gupta, 2017). Afterward, the machine learning algorithms were implemented with the use of two prevalent tools; Ranklib and Sci kit learn libraries. Finally, the precision of prediction and precision of ranking metrics were measured and compared to obtain the final results.

4. Implementation of algorithms

In order to implement the classification and learn to rank algorithms discussed in the methodology section, we used the following prevalent coding tools.

Sci kit learn library: Sci kit is an open source software library for coding in python. This library contains many algorithms, such as prevalent regression, classification, and clustering algorithms, and is written in a way that is compatible with other python libraries such as NumPy and

[3] Normal Discounted Cumulative Gain

SciPy. Logistic Regression, Random Forests, Support Vector Classification, Naïve Bayesian, K Nearest Neighbor, and Decision Trees are the traditional classification algorithms we chose to use for ranking in this paper.

Table 3. A sample of the final dataset that includes features and relevance scores for all retail store branches. The original dataset contained 239 rows.

Store Name	Area Popularity	Area Entropy	Competition	Distance from Downtown	Accessibility	Complementarity	Rank	Relevance Score
17 Shahrivar Canbo Store	8.30E+03	4.70E+00	6.25E-02	3.14E-04	8.87E-04	0.00E+00	1.15E-01	3.00E+00
Hakimiyeh Shahrvand Store	8.27E+03	6.97E+00	2.08E-02	5.83E-05	0.00E+00	1.77E-03	1.98E-01	3.00E+00
Shariati Jonoob Canbo Store	1.35E+03	6.35E+00	4.26E-02	1.50E-04	0.00E+00	0.00E+00	9.58E-05	0.00E+00
Javadiyeh Canbo Store	8.38E+03	5.95E+00	2.08E-02	2.84E-04	2.66E-03	1.77E-03	3.46E-01	3.00E+00
Zanjan Sepah Store	4.77E+03	5.96E+00	2.08E-02	1.53E-04	8.87E-04	0.00E+00	4.36E-03	1.00E+00
Azadi Refah Store	5.80E+03	6.07E+00	2.08E-02	1.08E-04	1.77E-03	0.00E+00	5.42E-02	2.00E+00
Lashgar HyperStar Store	8.47E+03	5.07E+00	8.51E-02	3.84E-04	1.85E-03	0.00E+00	1.31E-01	3.00E+00
Milad Shahrvand Store	9.98E+03	5.88E+00	0.00E+00	1.15E-04	8.87E-04	0.00E+00	1.74E-01	3.00E+00
Chamran Etka Store	8.48E+03	6.25E+00	2.08E-02	5.98E-04	2.66E-03	1.77E-03	3.86E-01	3.00E+00
Ansar Canbo Store	1.62E+03	4.77E+00	2.08E-02	1.71E-04	0.00E+00	0.00E+00	2.18E-03	0.00E+00

Ranklib library: Ranklib is a byproduct of the lemur project. The lemur project was dedicated to creating search engines, search tools, text analysis tools, and data resources used for research and development in information retrieval and text mining. Ranklib provides 8 Learn-to-Rank algorithms written in JAVA. MART, RankNet, and LambaMART are among these algorithms (Dang, 2012). Therefore, we utilized this library to implement these three algorithms in order to rank the retail stores based on the features.

The approach used in this paper for location optimization was based on the concept of feature selection. Choosing a subset of the most effective features in prediction is referred to as feature selection and is considered one of the principal techniques in pattern recognition, machine intelligence, and data mining (Prati, 2012). By screening out the features with little predictive abilities, irrelevant and redundant data was eliminated, and the performance of learning algorithms was enhanced (Karamshuk et al., 2013). Decreasing the space needed for recording, cognition cost and time, as well as increasing the precision of classification, simplification of the understanding and visualization, and dealing with the curse of dimensionality are among the advantages offered by feature selection (Jain & Zongker, 1997; Guyon, 2003; Karabegovic & Ponjavic, 2012, Guan et al., 2017). Since the problem of store selection has been considered as a ranking problem by many researchers (Xu et al., 2016), learn-to-rank algorithms usually used for information retrieval purposes can be utilized for solving this problem as well. Learn-to-rank algorithms are based on the integration of the main concepts in ranking and machine learning. Additionally, by enhancing learn-to-rank algorithms through the utilization of ensemble learning techniques, a number of algorithms have been introduced and classified into two groups; feature selection based on ensemble learning (FSen) and ensemble learning based on feature selection (ENfs). FSen algorithms use a number of models for training and combine their results, yielding ranking precision higher than any of the models separately (Prati, 2012; Yang et al., 2010).

Depending on the nature of the available data, feature selection can be employed in a supervised, semi-supervised, or an unsupervised manner (Huang, 2015). Supervised feature selection, which was used in this paper, is the process of training algorithms for prediction based on pre-determined labels. Logistic Regression, SVC, Decision Trees, Random Forests, Bayesian classification, and the K Nearest Neighbor algorithms are all considered as feature selection algorithms and have been used for retail store ranking in this paper. MART and RankNet are examples of learn-to-rank algorithms used in this paper, and LambaMART, which is one of the most infamous examples of FSen algorithms, was used as a representative of FSens. By applying these algorithms, we technically compared the ranking power of some of the more widely used feature selection algorithms with learn-to-rank and FSen algorithms in ranking retail stores based on their popularity. The coding procedure used for the implementation and validation of these algorithms is shown in the form of pseudocode in **Error! Reference source not found.**. Figure 3-8 demonstrates the results of the implementation of classification algorithms.

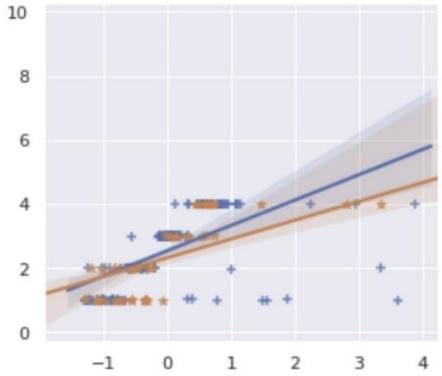

Figure 3. The results of the implementation of Logistic Regression.

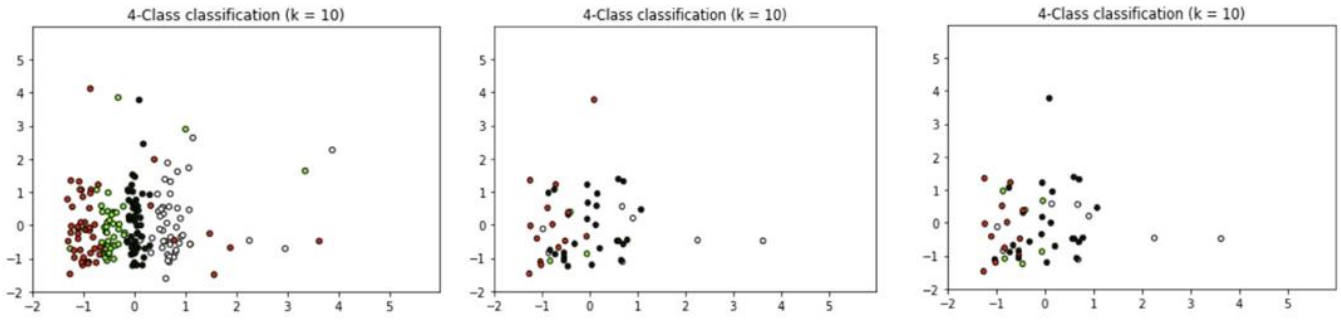

Figure 4. The results of the implementation of K-nearest Neighbors.

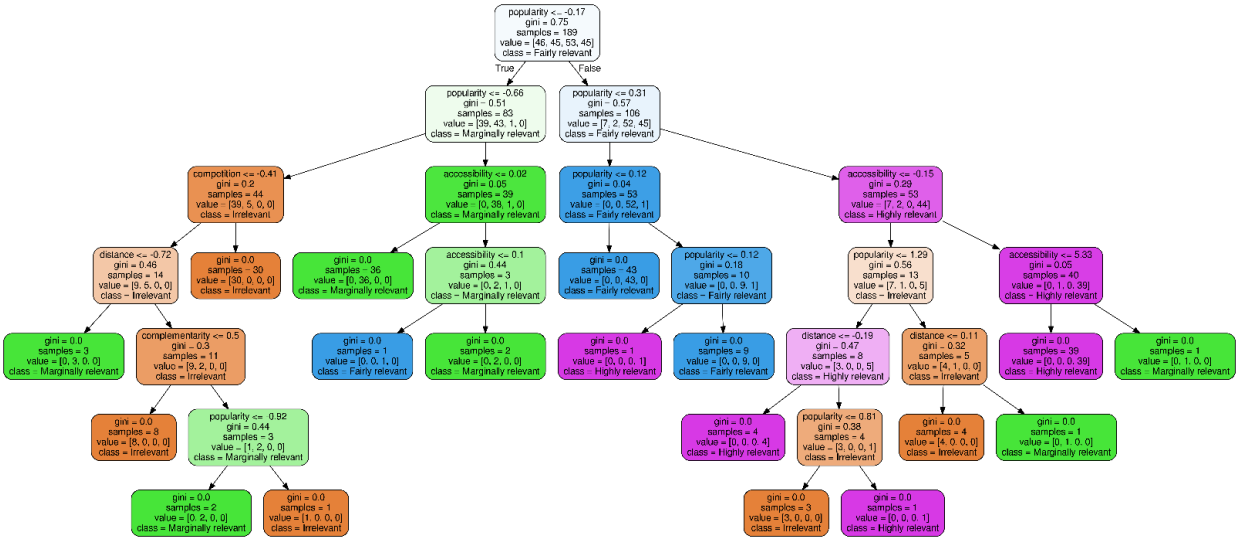

Figure 5. The results of the implementation of Decision Trees.

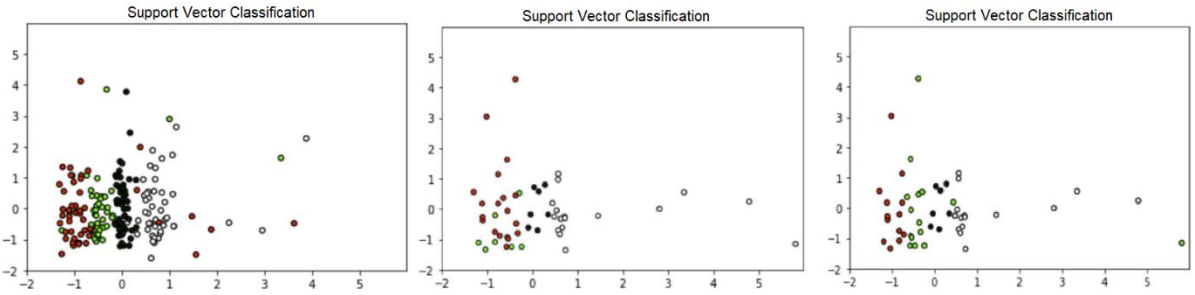

Figure 6. The results of the implementation of Support Vector Classification.

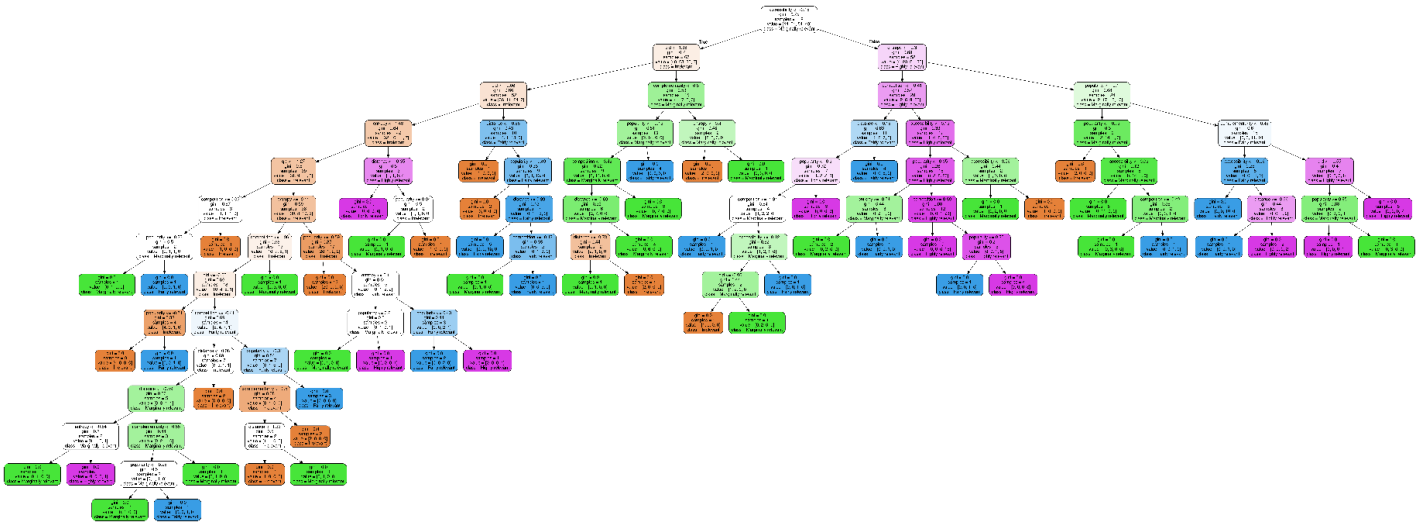

Figure 7. The results of the implementation of Random Forests.

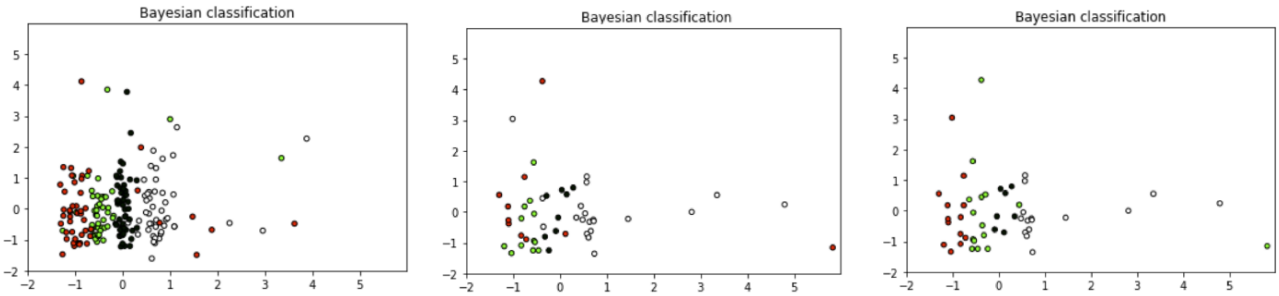

Figure 8. The results of the implementation of Naïve Bayesian Classification.

```
#Normalize features
  scaler = preprocessing.StandardScaler().fit(Featuress)
  x = scaler.transform(Featuress)
#Split dataset to train and test randomly
  nr.seed(9922)
  indx = range(Featuress.shape[0])
  indx = ms.train_test_split(indx, test_size = 50)
  x_train = x[indx[0],:]
  y_train = np.ravel(Labelss[indx[0]])
  x_test = x[indx[1],:]
  y_test = np.ravel(Labelss[indx[1]])
#Measure Precision, recall and Fmeasure for each algorithm
  LR = linear_model.LogisticRegression()
  RF = RandomForestClassifier()
  DTC = DecisionTreeClassifier()
  KNC = KNeighborsClassifier()
  SVM = svm.SVC()
  GNB = GaussianNB()
  Labels = Labelss.reshape(Labelss.shape[0],)
  scoring = ['precision_macro', 'recall_macro', 'f1_macro']
  def precision_recall_fmeasure(selector):
    scores = ms.cross_validate(selector, Featuress, Labels, cv=10, scoring=scoring, return_train_score=False)
    return scores
#Implement algorithms and measure nDCG@k
  def implementation(algorithm)
   algorithm_fit = algorithm.fit(x_train, y_train)
   y_predicted = algorithm_fit.predict(x_test)
    def ndcg_metrics(algorithm):
     ndcg_selector = ndcg.ndcg_at_k(y_predicted, 20)
     accuracy = sklm.accuracy_score(y_test, y_predicted)
     return (ndcg_selector)
Print results
End
```

Figure 9. Pseudocode for the algorithm implementation and evaluation steps.

5. Evaluation and comparison

Evaluating the performance of the proposed model involves (1) computing the precision, recall, and F-measure of prediction for feature selection algorithms and (2) prediction precision for learn-to-rank algorithms and computing the ranking precision by calculating the nDCG@k index for every algorithm. Such calculations are in need of a set of pre-determined labels referred to as the ground truth. These labels are the ranks we extracted, unified, and normalized before. The results of computing Precision, Recall, and F-measure for classification algorithms are demonstrated in Table 4. For learn-to-rank algorithms, a precision@k metric is usually used to evaluate the prediction precision of an algorithm and is equivalent to the precision metric defined for feature selection algorithms. Precision@20 was calculated for all three learn-to-rank algorithms and is presented in Table 4. In terms of ranking precision, nDCG@20 was calculated for every algorithm, and the results are presented in and Table 5 and Figure 10. By comparing the precision of ranking for all algorithms, it is clear that MART has the most precision in ranking retail stores. Additionally, all of the learn-to-rank algorithms used – RankNet, MART, and LambdaMART – have superiority in terms of precision of ranking and prediction compared to traditional classification algorithms. Within the classification algorithms used, SVC yielded the least precise results, which is consistent with the results of the prediction precision metrics, and Bayesian Classification was the most precise in terms of ranking retail stores.

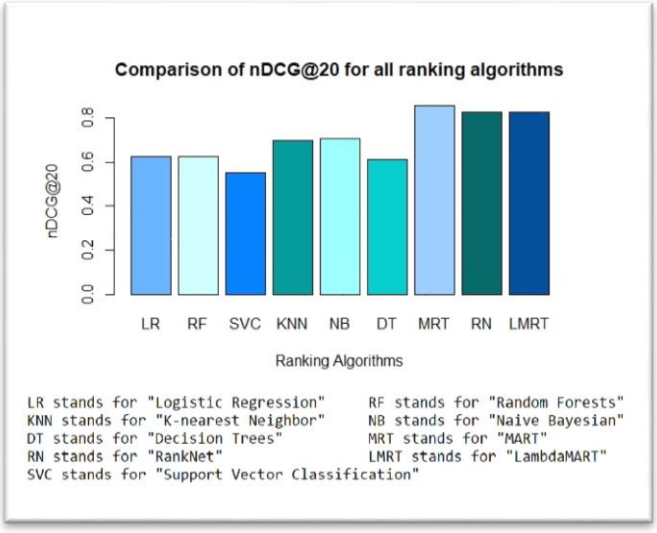

Figure 10. A Schematic comparison of nDCG@20 for all algorithms.

Table 4. Mean Precision, Recall and F-measure for classification and regression algorithms

Classification Algorithms	Mean F-measure	Mean Recall	Mean Precision
Logistic Regression	0.630	0.670	0.664
Random Forests	0.873	0.878	0.894
SVC	0.289	0.363	0.514
K Nearest Neighbor	0.899	0.899	0.911
Decision Trees	0.886	0.870	0.880
Naïve Bayesian Classification	0.758	0.741	0.729
Learn-to-rank Algorithms	Precision@20		
MART	0.95		
RankNet	0.90		
LambdaMART	0.95		

Table 5. The comparison of nDCG@20 for all algorithms.

Algorithm	**nDCG@20**
Logistic Regression	0.622
Random Forests	0.623
Support Vector Classification	0.552
K Nearest Neighbor	0.696
Naïve Bayesian	0.706
Decision Trees	0.611
MART	0.854
RankNet	0.823
LambdaMART	0.8275

6. Discussion and conclusion

The foremost goal of this paper was to utilize machine learning algorithms for ranking retail stores. To the best of our knowledge, this approach has never been used for assessing the popularity of retail stores. To do so, we used a number of prevalent classification and learn to rank. algorithms. Additionally, since factors such as economics and cultural characteristics of a country can influence the structure of its cities and the distribution of stores across it,

evaluating the practicality of using feature selection for store ranking in the city of Tehran was the secondary goal of this paper. Two other papers, as well, have used learn-to-rank algorithms for location selection. Karamshuk et al. (2013) implemented RankNet, and their attained results indicated that SVR would yield more precise predictions than RankNet for location selection in the food industry. Xu et al. (2016) compared the results of LambdaMART with a number of traditional machine learning techniques and concluded that while RF was better suited for ranking coffee shops, LambdaMART would present higher precision in ranking appliance stores. Based on the results of the evaluation metrics used in this paper, it can be concluded that learn-to-rank algorithms deliver more precise results for retail store placement and among the three learn-to-rank algorithms used in this paper, MART had the most overall precision in ranking and prediction.

However, since all the regression and classification algorithms, except for SVC, yielded reliable results, in terms of precision of prediction and ranking, if simpler calculations were one of the objectives of analysts, these algorithms might seem more appealing. Consequently, in today's world, with its ever-changing markets and intense rivalry, having access to real-time user-generated spatial data can help retail store owners and marketing analysts to make more informed decisions and plan more accurately for the place factor of the marketing mix. On the other hand, considering the features used in ranking, it can be inferred that a combination of these features can be used in optimal retail store placement. Hence, whenever a store is placed in an area with more popularity, better accessibility, less distance to the city center, with a greater variety of venues, more complementary businesses and less competition, it is in an overall better geographic position and can be expected to have relatively more customers than an alternative with an inferior location.

With a focus on utilizing spatial analytics for store placement, we used LBSN data to predict retail store popularity and ranking. Since all the features here were of static nature and extracted from the venue data available on Foursquare API, it can be concluded that using static data alone can lead to more precise planning for the place of a retail store. However, by retrieving Foursquare check-in data, which is available on the public API of Twitter, the movement of potential customers can be studied in regard to the selected area as well. Since Twitter is blocked in Iran, we could not exploit the characteristics of the check-in data due to the use of VPN applications by the Iranian users, which makes the obtained locations meaningless and unreliable. Therefore, by focusing on a city from another country in which the Twitter data are accurate in terms of their longitudes and latitudes, we would conduct another research and consider features such as the quality and density of customer transition to an area as well. Moreover, LBSNs offer a variety of datasets, one of which is User-Generated Reviews (UGR). As UGRs are usually provided by real consumers with no profits to gain, they are more likely to impact the perception and decisions of other customers. As a result, utilizing standard text mining algorithms, including Natural Language Processing and Sentiment Analysis, could lead to invaluable information concerning the popularity of a retail store and might be a promising approach for enhancing the framework used in this paper.

Reference

Aboulola, O. I. (2018). *GIS Spatial Analysis: A New Approach to Site Selection and Decision Making for SmallRetail Facilities*. Clement Graduate University.

Arcaute, E., Molinero, C., Hatna, E., Murcio, R., Vargas-ruiz, C., Masucci, A. P. P. & Batty, M. (2016). Cities and Regions in Britain Through Hierarchical Percolation. *The Royal Society*, *3*(4), 20.

Bacon, R. W. (1991). Consumer Shopping and Equilibrium Market Areas in the Presence of Demands. *Environment and Planning A*, *23*(9), 1361–1374.

Brown, S. (1993). Retail Location Theory : Evolution and Evaluation. *The International Review of Retail, Distribution and Consumer Research*, *3*(2), 185–229.

Damavandi, H., Abdolvand, N. & Karimipour, F. (2018). *The Computational Techniques for Optimal Store Placement : A Review*. *Computational Science and Its Applications – ICCSA 2018* (Vol. 1). Melbourne, Australia: Springer International Publishing.

Dang. (2012). Introduction to Ranklib. Retrieved September 26, 2018, from https://sourceforge.net/p/lemur/wiki/Home/

Guan, D., Yuan, W., Lee, Y. & Najeebullah, K. (2018). A Review of Ensemble Learning Based Feature Selection. *IETE Technical Review*, *31*(3), 190–198.

Gupta, P. (2017). Cross Validation in Machine Learning. Retrieved November 18, 2018, from https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f

Guyon, I. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning*, *3*(3/1/2003), 1157–1182.

Hillier, B. & Hanson, J. (1984). *The Social Logic Of Space*. England, Cambridge: Cambridge University Press.

Hillier, B., Perm, A., Hanson, J., Grajewski, T. & Xu, J. (1993). Natural movement or Configuration and Attraction in Urban Pedestrian Movement. *Environment and Planning B: Planning and Design*, *20*(1), 29–66.

Hotelling, H. (1929). Stability in Competition. *The Economic Journal*, *39*(153), 41–57.

Huang, S. H. (2015). Supervised Feature Selection : A Tutorial. *Artificial Intelligence Research*, *4*(2), 22–37. https://doi.org/10.5430/air.v4n2p22

Jain, A. & Zongker, D. (1997). Feature Selection: Evaluation, Application, and Small Sample Performance. *Transactions on Pattern Analysis and Machine Intelligence*, *19*(2), 153–158.

Jarvelin & Kekalainen (2002). Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446.

Jensen, P. (2006). A network-based prediction of retail stores commercial categories and optimal locations. *American Physical Society*, *74*(3), 035101.

Johnston, R. (1968). Urban Growth and Central Place Patterns. *Geographical Research*, *59*(2), 33–41.

Johnston, R. J. (1966). The Distribution of an Intra-metropolitan Central Place Hierarchy. *Geographical Research*, *39*(24), 391–399.

Karabegovic, A. & Ponjavic, M. (2012). Geoportal as Decision Support System with Spatial Data Warehouse. In *Proceedings of the Federated Conference on Computer Science and Information Systems, 9-12 Sept., 2012* (pp. 915–918). Wroclaw, Poland.

Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V. & Mascolo, C. (2013). Geo-Spotting : Mining Online Location-based Services for. In *19th ACM SIGKDD international conference on Knowledge discovery and data mining, 11-14 August, 2013* (pp. 793–801). United States, Chicago.

Khajegiri, T. (2018). An Outlook On Dunro's Future, In Persian. Retrieved November 1, 2018, from http://www.bartarinha.ir/fa/news/766661/

Kheiri, A., Karimipour, F. & Forghani, M. (2016). Intra-urban Movement Pattern Estimation Based on Location Based Social Networking Data. *Journal of Geomatics Science*, *6*(1), 141–158.

Kotler, P. & Armstrong, G. (1989). *Principles of Marketing*. England, London: Pearson Education.

Litz, R. A. (2008). Does Small Store Location Matter ? A Test of Three Classic Theories of Retail Location. *Journal of Small Business & Entrepreneurship*, *21*(4), 477–492. https://doi.org/10.1080/08276331.2008.10593436

Malley, Ramazzotti, & Wu. (2016). Data Pre-preprocessing. In *Secondary Analysis of Electronic Health Records* (pp. 115–141). Switzerland, Cham: Springer International Publishing.

Miller, H. & Goodchild. (2016). Data-driven Geography. *GeoJournal*, *80*(4), 449–461. https://doi.org/10.1007/s10708-014-9602-6

Nakamura, D. (2014). Social participation and social capital with equity and ef fi ciency : An approach from central-place theory q. *Applied Geography*, *49*, 54–57. https://doi.org/10.1016/j.apgeog.2013.09.008

Porta, S., Strano, E., Iacoviello, V., Messora, R., Latora, V., Cardillo, A. & Scellato, S. (2009). Street centrality and densities of retail and services in Bologna , Italy. *Environment and Planning B: Planning and Design*, *36*(3), 450–466.

Potter, R. (1981). The multivariate functional structure of the urban retailing system : a British case study. *Transactions of the Institute of British Geographers*, *6*(2), 188–213.

Prati, R. C. (2012). Combining feature ranking algorithms through rank aggregation. In *The 2012 International Joint Conference on Neural Networks (IJCNN), 10-15 June, 2012*. Brisbane, QLD, Australia.

Rahman, K. & Nayeem, M. A. (2017). Finding suitable places for live campaigns using location-based services. In *Proceedings of the Fourth International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data, 14 - 19 May, 2017* (pp. 1–6). Chicago, Illinois.

Reilly. (1927). *Methods for the Study of Retail Relationships*. Austin, Texas: University of Texas Bulletins and Publications.

Singh, H. (2011). *Retail Management: A Global Perspective*. India, New Delhi: S Chand & Co Ltd.

Taneja, S. (1999). Technology Moves In. *Chain Store Age*, *75*(5), 136–137.

Ting, Ho, Yee & Mastah. (2018). Geospatial Analytics in Retail Site Selection and Sales Prediction. *Big Data*, *6*(1), 42–52. https://doi.org/10.1089/big.2017.0085

Urbon-Bayes, P. (2017). Learn to Rank 101. Retrieved from https://medium.com/@purbon/learning-to-rank-101-5755f2797a3a

Wang, F. & Chen, L. (2016). Where to Place Your Next Restaurant ? Optimal Restaurant Placement via Leveraging User-Generated Reviews. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016* (pp. 2371–2376). Indianapolis, IN, USA.

Xu, M., Wang, T., Wu, Z., Zhou, J., Li, J. & Wu, H. (2016). Store Location Selection via Mining Search Query Logs of Baidu Maps. *Computing Research Repository*, *arXiv:1606*(4), 17–28.

Yang, P., Liu, W., Zhou, B. B., Chawla, S. & Albert, Y. (2010). Ensemble-based Wrapper Methods for Feature Selection and Class Imbalance Learning. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining, April 14-17, 2019* (pp. 1–12). Hyderabad, India.

Yu, Tian, M., Wang, Z. H. U. & Guo, B. I. N. (2016). Shop-Type Recommendation Leveraging the Data

from Social Media and Location-Based Services. *ACM Transactions on Knowledge Discovery from Data*, *11*(1), 1–21.

Yu, Z., Zhang & Yang. (2013). Where is the Largest Market : Ranking Areas by Popularity from Location Based Social Networks. In *2013 IEEE 10th International Conference on Autonomic and Trusted Computing, 18-20 December, 2013* (pp. 157–162). Milan, Italy.