Chapter 8
A Combined Data Analytics and Network Science Approach for Smart Real Estate Investment: Towards Affordable Housing

E. Sandeep Kumar and Viswanath Talasila

1 Introduction

Affordable housing is emerging as a major requirement due to the growth in the need for creating equality in the living standards of people in our society [1]. Government is extending the collaboration towards public–private partnership, especially realtors to construct houses and apartments for people who are below the median line of income. The locations for construction of such houses are dependent on numerous real estate attributes, which include social, cultural, economic, physical, and governmental [2]. Some of these attributes include public transportation facilities, availability of public schools and colleges, availability of water and supportive weather conditions, availability of hotels and restaurants, and so on [3]. As the attribute number list grows, so does the complexity of decision-making in location choice. Identification of best locations is an important requirement from the perspective of not only house construction but also purchase of the existing house and renting.

Hence, to understand the trends and solve the above mentioned problems in real estate, many methods and tools are used which are derived from various fields like data science, network science, statistics, probability theory, estimation theory, and so on. Today, the availability of huge volumes of data has paved the path for researchers to use concepts and tools of data analytics to discover goal-oriented knowledge from the existing data. One such important application is hedonic modeling, where the dependency of the various attributes on the real estate price is computed using basic regression methods and machine learning techniques. Use of linear and logistic regression, clustering techniques, support vector machines, and

E. Sandeep Kumar (✉) · V. Talasila
Department of Telecommunication Engineering, M.S Ramaiah Institute of Technology, Bengaluru, Karnataka, India
e-mail: viswanath.talasila@msrit.edu

N. V. M. Lopes (ed.), *Smart Governance for Cities: Perspectives and Experiences*, EAI/Springer Innovations in Communication and Computing,
https://doi.org/10.1007/978-3-030-22070-9_8

artificial neural networks [4–14] on the real estate multiple listing service data is often encountered in the literature. Additionally, predicting the house price using data analytic tools is also a well-studied research area. Regression techniques, neural networks, and gradient boosting [15–17] are few tools that are often used. The increasing availability of voluminous data has attracted network science practitioners to infer on the relational status and mutual dependencies among the various data ingredients. Use of social media like Facebook, Twitter, LinkedIn, and various search engines is generating digital footsteps which are used to develop knowledge graphs. These graphs (networks) make us understand the behavior of entities in a network [18–20].

To the best of the authors' knowledge, there has been no focused effort to use network science for the analysis of real estate networks and in specific for location identification. Majority of the software [21, 22] ask the user to enter an exact location, and based on the database query search it will identify suitable condominiums (apartment complexes) for investment. In addition, the existing literature on data analytics in real estate investment is based on the assumptions that a user already knows a location. There are many reasons why an investor may not know the specific location for investment. A simple reason may be that an investor is new to the city. A more involved reason is that even though an investor is native to the city, it is logically impossible to narrow down to a very specific location; at best a small geographical area can be identified. However, in big cities even a small area can easily compromise thousands of dwellings and commercial property; further, even the small area is often highly heterogeneous (in terms of people, establishments, facilities, etc.). Focusing only on price trends does not address the multiple concerns of an investor [15].

Choosing a good location for investment is very crucial since it is dependent on a large number of user's requirements. It may be based on job availability, economic status of people, availability of restaurants, low criminal activities and safety, public transportation facility, availability of schools and shopping malls, and many more. These multiple attributes make a user's decision to select a location more complex and difficult. Under the influence of this huge number of attributes, the location selection may tend towards suboptimal decisions. Hence, an intelligent way of choosing the locations is of greater need in real estate investment that also focuses on the selection of best attributes among that huge number of attributes. Moreover, the mutual influence of the attributes is not considered in the existing literature and attributes are assumed to be independent which is actually not true in general.

In this chapter, a novel algorithm has been presented that selects best attributes for a user and based on user's choice the algorithm identifies the best locations (throughout this chapter, authors refer to condominium complexes as locations) for investment in real estate. In the proposed work, nearly 200 real estate attributes were considered and the best attributes were selected based on the metric called χ which is computed mainly based on the Pearson correlation coefficient [23]. The obtained attributes form a source of choice for a user and based on his/her selections layers of machine learning techniques are activated. In the first layer roads and

streets are identified and in layer 2 condominiums (locations) in that street is spotted. For this purpose, statistical modeling with machine learning techniques is used which are taken from data science as analytic tools. However, applications of data analytics succeed well in identification and classification problems; a clearer inference on the relational status of the attributes cannot be obtained, especially when the entities are in huge number. Hence, a bipartite network is added as an extension to the machine learning layers that provides a relational status of the attributes and their influence on the various streets and roads (collectively we call them as landmarks hereafter), while selecting best location among the shortlisted locations by the machine learning layers, for investment. The network uses eigen centrality on a bipartite network of attributes and condominiums for selecting the best condominium for investment. The advantage of using the network layer is the selection of condominium based on the relationship status of all the other attributes with the condominiums in the network. An example simulation on the analysis of network dynamics is provided to leverage the advantages of network science by creating perturbations in the link weights of the network, to find the influential and stable attribute in the designed real estate bipartite network.

The location identification algorithm is tested on the data obtained from the official database called *TerraFly* [24] which is created and maintained by Florida International University (FIU). We have restricted this work to nine landmarks. Rest of this chapter is organized as follows: Sect. 2 deals with the applications of data and network science for smart governance; Sect. 3 discusses about the related works and the state-of-the-art comparison with the current work in the chapter; Sect. 4 deals with the data set used in this work and the assumptions on the work; Sect. 5 discusses the statistical modeling used to identify the best attributes, and discusses the use of decision trees and PCA and *K*-means clustering for location identification; Sect. 6 deals with the network science for location identification problem; Sect. 7 discusses the obtained results and discussions; Sect. 8 discusses the implications of the obtained result on the smart governance; and finally Sect. 9 discusses the conclusions of the work.

1.1 Scopes of This Work

- Application of statistical modeling to obtain the best attributes from nearly 200 real estate attributes based on the metric χ which is computed using Pearson's correlation coefficient.
- Use stacks of machine learning algorithms to identify locations (condominiums) for investment.
- Application of network science to obtain the best condominium based on centrality measures by constructing a real estate network.
- Simulation study on the most consistent and influential attribute in the presence of link weight perturbations in the designed real estate network.

2 Data and Network Science Applications in Smart Governance

Technology has already made its way into our daily activities. Shopping, transportation, communication, education, health, and so many other things rely on smart devices that are driven by advanced computing and techniques. Usage of such gadgets has paved the way to accumulate a large amount of user data which enabled the era of big data analytics [25]. Analysis of and drawing useful inferences based on various applications can help in the betterment of society. However, operating on such voluminous and versatile data has to be carried out using sophisticated methods. Data science is one such solution enabler [26] which is an interdisciplinary area comprising the tools for statistics, mathematics, probability theory, artificial intelligence, and machine learning to a larger extent and provides us better understandability of the data for various applications and one such application is *smart governance*. Applications in smart governance include consumer behavioral analytics, natural calamity predictions, crime predictions, social service-related analytics, healthcare analytics, data security and privacy, finance, and bank analytics.

Network science is another interdisciplinary research area that draws the theories and methods from graph theory, mathematics, statistics, physics, sociology, and data mining. The current-day computer networks, Internet, and various communication networks are analyzed using network analysis. Similarly, there are other few interesting applications of the network science (complex network analysis) that find their profound applications in smart governance including biological networks [27], transportation networks, epidemic networks, social networks, financial networks, and so on. Readers who are further interested to know about complex networks are directed to read [28]. Even though this chapter highlights the use of data and network science to solve the real estate location identification for housing application, there are still numerous other tasks in smart cities and governance that can be addressed using the able usage of data and network sciences separately or by the combination of both. Two specific examples of smart governance applications are discussed in detail in Sect. 8.

3 Related Works and State-of-the-Art Comparison

The application of data analytics and network science to solve the real estate location identification problem is novel. The existing works in data analytics focus much on the prediction and modeling of the real estate price. Authors in [29] propose a method to predict the house price index (HSI) using data analytic techniques like clustering and principal component analysis for Kookmin bank data. Work in [4] discusses the use of linear regression to find the relation between the real estate price and the attributes; for this purpose authors use real estate data of Harbin city. In [30], authors propose a linear regression hedonic model to find the spatial

dependency of the real estate price; for this purpose they have considered the real estate data of eight countries from the multiple listing service database.

In [31], authors discuss a framework of using fuzzy logic to find the selling price of a real estate property in the presence of incomplete information. They claim that the developed method helps in reducing the risk that arises from the uncertainties in the input. In [32], authors use fuzzy logic systems for hedonic house price modeling. In [33], authors use network science to determine the financial activities among different entities. This analysis would help in revealing financial crimes like terrorism, narcotic laundering, and so on.

Authors in [34] discuss the usage of network science to model the supply chain in the form of a network. Authors claim that such kind of structural studies helps in revealing the robustness of supply chains and reveal the topological behavior of the supply chains while overcoming the limitations of the existing supply chain networks. In [35], authors use complex network analysis in considering the micro- and macro-level attributes of revenue of UK stock market and forecast the stock value predictions using complex networks.

There have been limited applications of data science towards real estate location identification problem. The trend majorly includes hedonic price modeling that relates the real estate price with its attributes and prediction of real estate price using the attributes. There are no specific trends in network science applications for real estate investment-related problems.

By carefully traversing through the existing works, as quoted in Sect. 1 and in the current section, it is clear that all the existing works in data analytics focus more on the predictions of the price and hedonic modeling; the location identification for investment has not been explored much. In addition, there are no specific works of network science usage that focus on the issues in the real estate investment. There are few applications in the area of finance but there are no confined works that focus on the real estate. The existing software and companies of real estate investment [36–38] either focus on the construction of building, maintaining, and looking after the documentation works or suggest the investment locations based on some database queries. All these methods in the existing works do not apply any machine learning techniques that are based on the query approach. Hence, considering the real estate investment location identification problem the work proposed in this chapter is a novel attempt that combines data and network science under a single roof to find a robust solution for the given problem of location identification for affordable housing in real estate investment.

4 Data Set

The data is obtained from TerraFly database [24] managed and maintained by Florida International University (FIU) in collaboration with the US Government. The database is a big data platform and a query-based system with complete information regarding economic, social, physical, and governmental factors of selected countries.

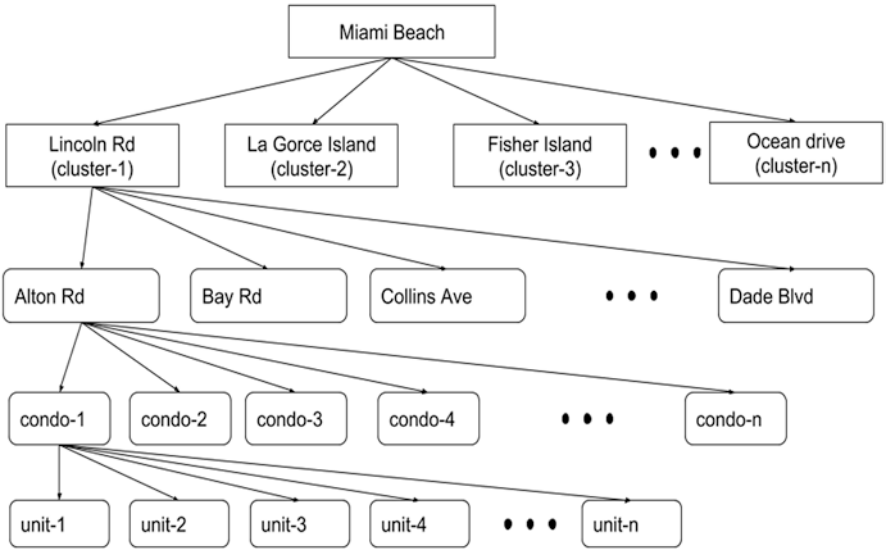

Fig. 8.1 Hierarchical view of the available data (clustered)

The landmarks (we call streets, roads, and so on as landmarks in this work) of Miami Beach are divided into clusters. Preference is given to nearby landmarks while clustering, however, can also be random. For our initial work, single cluster with nine landmarks and their associated real estate data (available as multiple listing service (MLS) data) is considered. They are condominium (also called as condos) data of Alton Rd, Bay Rd, Collins Ave, Dade Blvd, James Ave, Lincoln Rd, Lincoln CT, Washington Ave, and West Ave. These landmarks belong to Miami Beach City of Miami Dade County, FL, USA. The approximate count of condominiums was obtained from the official database of Miami Beach, i.e., for Alton Rd-7000 condominiums, Bay Rd-7000, Collins Ave-9000, Dade Blvd-1500, James Ave-2000, Lincoln Rd-2000, Lincoln CT-2000, Washington Ave-4000, and West Ave-2000, respectively. The hierarchical view of the data is shown in Fig. 8.1. For our analysis from every landmark, 500 condominium data were randomly picked for training and 500 for validation. The processes of training and validation were repeated in five sets and the average validation accuracy is quoted, which will be discussed in detail in the results section.

4.1 Assumptions in This Work

It is assumed that a user is not fully aware of the city location details. He/she has a very little idea about the locations, but do not know whether it is best or not for investment. In addition, a user has assumed a set of attributes; however, they need

not be optimal. However, a user should at least know which cluster of landmarks should be opted for his/her investment.

5 Identification of Location Using Data Analytics

In this section, a detailed analogy of the attribute selection, stacked layers of machine learning for location identification, and real estate network with its related analytics is discussed.

5.1 Attribute Selection

Real estate comprises a large list of attributes that can be broadly classified into economic, social, physical, and governmental [24]. The current framework is bound to the real estate attributes; however, the same method can be extended for other factors as well.

Out of a large number of attributes, the best set of attributes are selected based on the value χ, which is a representative of the strength of an attribute in a landmark in real estate investment. The calculation of χ is as follows:

$$\chi = w_1 C + w_2 N \tag{8.1}$$

where $w_1$and w_2 are constants and called the weights, C is the Pearson coefficient of an attribute with the real estate price, and N is the number of data sample points available in an attribute after cleansing. Here, χ is an identity number assigned to every attribute in a landmark, based on which a top attribute is selected. For this calculation, the Pearson coefficient is used as an initial choice. However, there are other correlation metrics such as Spearman and Kendall coefficients [39] that can also be used instead. The selection algorithm is as follows:

Algorithm 1

1. Start
2. Collect the condominium data of all landmarks in a cluster. Initialize w_1 and w_2.
3. For first landmark, find the parameter χ which determines the relation between the first attribute of a condominium and real estate price.
4. Repeat the experiment for all the attributes. Select the top $\boldsymbol{k}$ number attributes from every condominium. Select the top $\boldsymbol{u}$ number of attributes based on the number of occurrences in a landmark.
5. Repeat steps 2 and 3 for all landmarks.
6. Combine all $\boldsymbol{u}$ and select top $\boldsymbol{y}$ attributes based on number of occurrences. This set is the optimal attribute set for that cluster of landmarks.
7. Repeat this process for all clusters of landmarks.
8. End.

For simulations, the cluster of nine landmarks was chosen and the parameters were set to $k = 10$ and $u = 10$, i.e., choosing top ten attributes every time. However, $y = 9$ is the count of the number of attributes for the entire cluster. The simulation is repeated for five iterations with each time 500 condominiums selected in random from every landmark (training data set). The top nine attributes obtained are as follows:

- **Number of beds**: Number of bedrooms available in the unit of a condominium building.
- **Number of full baths:** Number of full bathrooms (tub, shower, sink, and toilet) available in the unit.
- **Living area in sq. ft.:** The space of the property where people are living.
- **Number of garage spaces:** Number of spaces available for parking vehicles.
- **List price:** Selling price of the property (land + assets) to the public.
- **Application fee:** Fee paid for owners' associations.
- **Year built:** Year in which the condominium/apartment complex is built.
- **Family limited property total value 1:** The property value accounted for taxation after all exemptions. This is for the district that does not contain schools and other facilities.
- **Tax amount:** The amount paid as tax for the property every year.
- When a user chooses this cluster of nine landmarks, the above attributes will be given to him/her as choices. A user can select and set the magnitudes to the attributes according to his/her wish. It is not mandatory that all attributes need to be filled. These attributes are passed onto two layers of machine learning: in the first layer decision trees and in the second layer principal component analysis with *K*-means clustering. In addition, the rationale behind the choice of a multilayer classification model is provided in [40]; interested readers are suggested to refer to that article.

5.2 Decision Trees in Layer 1

In this section, the use of decision trees [41, 42] to select the best landmark for investment is dealt with detail. The working of tree follows the naive ID3 algorithm [43].

The attribute set used in a tree may change depending upon the landmarks in that cluster selected. The top nine attributes of any cluster have χ values in every landmark obtained by averaging the χ values of all the condominiums in that landmark. These χ values are compared with the other landmarks and the highest χ value is retained with its respective landmark. The result in Table 8.1 is the output of Algorithm 1 on the training set (500 condominiums selected randomly from every landmark). The process was repeated for five iterations and in all iterations the landmarks and the attribute pair remained same likewise shown in Table 8.1. The χ values shown are the average of the five iterations. Same results were obtained for the validation set as well. This table serves as a backbone for the decision tree operation.

Table 8.1 χ values of attributes input to decision tree

Attribute	χ value	Landmark
Number of beds	1.338	Alton Rd
Number of full baths	1.380	Alton Rd
Year built	1.226	Lincoln CT
Application fee	1.235	James Ave
Number of garage spaces	1.233	Alton Rd
List price	1.894	James Ave
FLP total value	1.291	Washington Ave
Living area	1.375	Alton Rd
Tax amount	1.164	Bay Rd

A user selects a particular cluster followed by choosing attributes of his/her interest and setting suitable magnitudes for them. For a decision tree, only a user interest vector is used. For example, suppose a user is interested in the number of beds and number of full baths in the same order as in Sect. 5.1; then the user vector is 110000000. This is fed into the decision tree.

An example decision tree with three attributes which is symmetric and binary is as shown in Fig. 8.2. In the figure, every time when the tree traversal hops from one node to another, it considers the χ value (if it is "YES" case) of the current attribute and compares with the χ value of the previous and selects the landmark with the highest χ value. The process continues until one landmark is retained at the end. This technique is called the ***highest magnitude win*** approach. The obtained truth table is shown in Table 8.2.

Notice the last column in Table 8.2. The rows are the different cases (shown are just four cases) in which a user can enter attributes. The generated user vector is the row. For the case "011," which is the third row in Table 8.2, it means a user is interested in number of full baths and list price. The magnitude of χ for these attributes is 1.380 and 1.894, respectively. According to the highest magnitude win strategy, James Ave wins in this case, which is placed in the output landmark column in the corresponding row. This is the leaf of that branch in the tree. The tree and its traversal are shown in Fig. 8.2.

Even though the tree output does not change, the position of the tree node (order of the columns in Table 8.2) matters for the fact that the time taken for the tree to converge to a landmark depends on the node information richness. To identify the positions of the nodes, the ID3 algorithm is adapted and in turn helps us to choose an optimal root attribute. For a truth table that is binary in nature and having all possibilities in it, all attributes have the same information gain according to ID3 and any attribute can be a root. However, if the truth table has chosen possibilities of truth and false, ID3 will help to choose the best root attribute. A detailed procedure on the usage of ID3 for root node selection is provided in APPENDIX-B of [40]. In addition, a decision tree need not be always as shown in Fig. 8.2, where the children nodes are identical, but they can differ as well, based on the landmarks and the cluster considered.

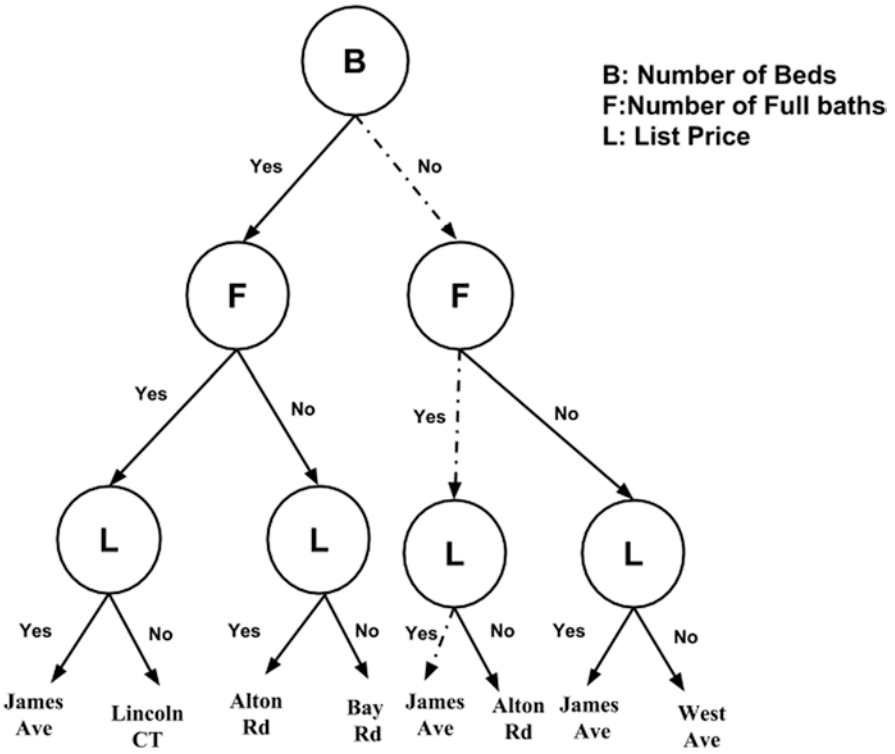

Fig. 8.2 Decision tree and its traversal

Table 8.2 Truth table for decision tree operation

Number of beds	Number of full baths	List price	Output landmark
0	0	1	James Ave
0	1	0	Alton Rd
0	1	1	James Ave
1	1	1	James Ave

5.3 Principal Component Analysis (PCA) and K*-Means Clustering in Layer 2*

The previous section discussed the use of decision trees to identify the best landmark among the chosen cluster of landmarks. In this section, layer 2 will be analyzed in detail, which helps to find locations in the landmark, output by layer 1. In the previous section, only the user's interest vector was considered, but in this layer the entered magnitudes are considered for location identification. The steps adapted in the layer 2 are discussed further in detail.

5.3.1 Finding Principal Components

Every landmark has condominiums which have the attributes mentioned as per Sect. 5 (in our study there are nine top attributes). Let us consider the algorithm below:

Algorithm 2

1. Start.
2. Consider first landmark from a cluster.
3. For every condominium in that landmark, find the principal components for the top attributes.
4. Consider the first principal component among the available components and average all the first components over a landmark.
5. Repeat steps 2 and 3 for all landmarks.
6. End.

Let the principal components [44] of a landmark be PC. Using PC of a particular landmark, principal scores of units in that condominium are calculated using (8.2). Every top attribute in that condominium is multiplied by PC:

$$\text{PCs} = \sum_{i=1}^{y} \text{attribute}\,(i) * \text{PC}\,(i). \tag{8.2}$$

where y is the count of top attributes of a cluster of landmarks. On averaging PC_S of all the condominium units, a PC score for a condominium is obtained. Repeat the process for all the landmarks and their associated condominiums. Averaging PC scores of units in a condominium will result in PC score for a condominium. Once the PC scores are available, *K*-means clustering [45] is applied to the scores of the condominiums of a landmark dividing the condominiums into K groups each having its own centroid.

Layer 2 mainly operates on the magnitude entered by a user which was not considered in layer 1. Since the landmark is already chosen by layer 1, the average PC of that landmark is known. Hence, the user-entered magnitude and the average PC are multiplied to get a PC score using (8.2). This score is compared with the centroids of the groups created by *K*-means clustering and the group with shortest Euclidean distance is selected as the best condominium group and its ingredient condominiums are rated as the best for a user according to his/her entered choices.

6 Network Science Approach for Location Identification

Data analytics uses systematic ways of modeling and learning the data, which is the digital trace in the ongoing world. However, to understand the network prospects of the data with their mutual influences, combining data with a network is very essential [46]. This interdisciplinary setup makes complex systems like real estate investment more understandable.

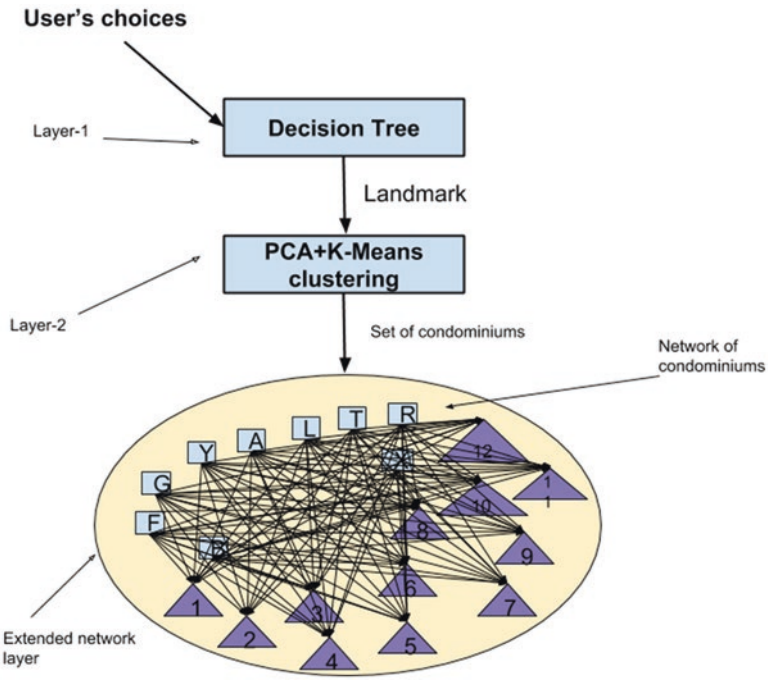

Fig. 8.3 Extended network layer

In this context, a network structure is rigged for the condominiums obtained from the layer 2 machine learning technique. This network view helps a user to decide the best condominium in the presence of mutual relationships among other condominiums with their attributes. The system model is shown in Fig. 8.3.

In Fig. 8.3, a network of condominiums is constructed. These condominiums were the list of condominiums output from layer 2 that comprised of PCA and *K*-means clustering. The obtained graph is a bipartite network which comprises two parties, viz. attributes and condominiums. It is to be observed that there is no link between the same party members and the link always flows from one party to the other. The attributes are shown in a square shape and the condominiums are shown in a triangular shape. On applying various centrality measures like eigen centrality [47], alpha centrality, closeness, and so on [48], it is able to draw various conclusions on the important condominiums in the network. In this work, eigen centrality is used to obtain the most influential condominium. In terms of real estate, eigenvalue depicts the amplification factor for the influence and the eigenvector infers the direction of the influence in the network.

Addition of network at the end of the machine learning layers is just a demonstration; however, layer 1 or layer 2 can also be visualized as a network, where suitable centrality measures will infer for landmark and location selection in that particular layer. As an example, in layer 1 eigen centrality can be used to select a landmark and in layer 2 alpha centrality to select the set of condominiums.

7 Results and Discussions

This section briefs the obtained results of the proposed methodology and its related discussions in detail.

The attribute selection (Algorithm 1) discussed in Sect. 5.1 was executed on five training and validation data sets. In each case, for the entire nine landmarks, unique attributes were listed. The attribute set was compared between the training and validation set and the number of mismatches was accounted. The obtained results for five datasets selected in random (i.e., five iterations) are available in Table. 8.3. It was observed that the system remains consistent with the training and validation result matching with an average of 96.86% accuracy.

The same process is repeated for the decision trees of layer 1. In the case of the tree, as it is explained already in Sect. 5.2, the tree traverses from one node to another considering the value of χ of every attribute. Hence, the attribute and landmark corresponding to the highest value of χ play a major role in decision tree outputting a landmark. The top attributes as per Sect. 5.1 are listed with its χ value averaged over a landmark in Table 8.4 for five iterations (both training and validation). It is observed that the winning landmarks with highest χ for an attribute remain the same in training and validation, which is the accuracy of 100% and in turn defines the accuracy of the decision tree.

From Table 8.4, the attribute with the landmark having highest χ value remains consistent throughout five iterations (as per both training and validation data sets), with 100% validation accuracy. The highest χ value is highlighted in bold in the table.

In layer 2, principal component analysis and *K*-means clustering were used to find the best condominiums. These techniques were applied to the training and the validation data set. The centroids of the groups of condominiums obtained after clustering were compared in both training and validation over five iteration data sets. The deviation error (mean absolute error) was noted in each case and the obtained results are available in Table 8.5.

The average validation accuracy of layer 2 from Table 8.5 is 90.25%.

The obtained condominiums from layer 2 were used to construct a bipartite network with attributes and the condominiums as the two parties and χ values linking them. From Table 8.4, it is evident that every attribute has χ value linking landmarks. Hence, a complex network of the attributes and condominiums is constructed

Table 8.3 Accuracy of best attributes selected

Iteration	No. of mismatches	Accuracy
1	1 out of 25	96%
2	0 out of 24	100%
3	1 out of 25	96%
4	1 out of 26	96.15%
5	1 out of 26	96.15%
Average		**96.86%**

Table 8.4 Attributes and their average χ values in every landmark

Iteration	Attributes	Alton Rd	Bay Rd	Collins Ave	Dade blvd	James Ave	Lincoln Rd	Lincoln CT	Washington Ave	West Ave
1 (Training)	Number of beds	**1.34**	1.29	1.22	1.22	1.20	1.20	1.20	1.16	1.24
	Number of full baths	**1.38**	1.30	1.28	1.25	1.14	1.21	1.20	1.27	1.27
	Year built	1.07	1.14	1.17	1.16	1.07	1.21	**1.23**	1.18	1.20
	Application fee	0.75	0.88	0.85	0.75	**1.22**	0.99	0.97	0.72	0.88
	Number of garage spaces	**1.24**	1.17	1.12	1.07	1.09	1.10	1.11	1.03	1.19
	List price	1.80	1.80	1.73	1.69	**1.89**	1.71	1.73	1.72	1.78
	FLP total value	1.28	1.27	1.24	1.09	0.99	1.09	1.17	**1.32**	1.26
	Living area	**1.37**	1.34	1.26	1.17	1.19	1.23	1.24	1.16	1.29
	Tax amount	1.09	**1.16**	0.93	0.99	0.12	0.84	0.88	1.08	0.99
1 (Validation)	Number of beds	**1.34**	1.30	1.21	1.22	1.19	1.20	1.20	1.16	1.21
	Number of full baths	**1.38**	1.31	1.27	1.26	1.13	1.20	1.20	1.27	1.27
	Year built	1.07	1.12	1.16	1.15	1.05	1.20	**1.23**	1.17	1.20
	Application fee	0.78	0.88	0.85	0.76	**1.24**	0.96	0.97	0.73	0.89
	Number of garage spaces	**1.24**	1.18	1.10	1.06	1.08	1.09	1.13	1.04	1.20
	List price	1.79	1.81	1.72	1.69	**1.89**	1.71	1.72	1.72	1.78
	FLP total value	1.27	1.28	1.23	1.11	0.98	1.08	1.17	**1.32**	1.27
	Living area	**1.38**	1.35	1.27	1.16	1.18	1.23	1.25	1.17	1.29
	Tax amount	1.08	**1.17**	0.88	1.01	0.08	0.82	0.89	1.08	1.00

2 (Training)	Number of beds	**1.34**	1.31	1.22	1.22	1.19	1.21	1.20	1.17	1.24
	Number of full baths	**1.38**	1.32	1.27	1.26	1.14	1.21	1.20	1.28	1.27
	Year built	1.07	1.11	1.18	1.16	1.06	1.2181	**1.2182**	1.19	1.20
	Application fee	0.76	0.86	0.82	0.75	**1.23**	0.95	0.97	0.72	0.90
	Number of garage spaces	**1.23**	1.16	1.11	1.06	1.08	1.10	1.11	1.03	1.20
	List price	1.80	1.80	1.73	1.69	**1.89**	1.71	1.71	1.73	1.78
	FLP total value	1.28	1.28	1.23	1.11	0.98	1.10	1.15	**1.34**	1.27
	Living area	**1.37**	1.35	1.26	1.17	1.18	1.24	1.24	1.18	1.29
	Tax amount	1.09	**1.15**	0.90	1.01	0.11	0.86	0.83	1.12	0.98
2 (Validation)	Number of beds	**1.33**	1.31	1.21	1.22	1.20	1.20	1.20	1.17	1.24
	Number of full baths	**1.37**	1.32	1.27	1.25	1.14	1.20	1.20	1.27	1.26
	Year built	1.08	1.11	1.19	1.15	1.07	1.20	**1.24**	1.18	1.21
	Application fee	0.76	0.87	0.83	0.75	**1.23**	0.93	0.97	0.72	0.89
	Number of garage spaces	**1.23**	1.16	1.11	1.06	1.09	1.10	1.12	1.04	1.18
	List price	1.80	1.80	1.72	1.69	**1.89**	1.70	1.73	1.72	1.78
	FLP total value	1.26	1.28	1.23	1.09	0.98	1.09	1.19	**1.32**	1.26
	Living area	**1.36**	1.35	1.27	1.16	1.19	1.23	1.25	1.17	1.28
	Tax amount	1.06	**1.15**	0.88	0.98	0.11	0.87	0.93	1.07	0.97
3 (Training)	Number of beds	**1.33**	1.30	1.06	1.22	1.19	1.21	1.20	1.17	1.24
	Number of full baths	**1.37**	1.31	1.11	1.25	1.14	1.21	1.20	1.27	1.26
	Year built	1.08	1.13	1.01	1.15	1.07	1.20	**1.23**	1.18	1.21
	Application fee	0.75	0.87	0.76	0.75	**1.24**	0.95	0.96	0.73	0.88
	Number of garage spaces	**1.22**	1.16	0.95	1.06	1.08	1.11	1.11	1.03	1.19
	List price	1.80	1.80	1.50	1.67	**1.89**	1.71	1.72	1.72	1.78
	FLP total value	1.25	1.28	1.07	1.09	0.97	1.09	1.17	**1.32**	1.26
	Living area	**1.36**	1.35	1.09	1.16	1.18	1.23	1.24	1.17	1.29
	Tax amount	1.07	**1.16**	0.77	0.99	0.06	0.82	0.90	1.07	0.97

(continued)

Table 8.4 (continued)

Iteration	Attributes	Alton Rd	Bay Rd	Collins Ave	Dade blvd	James Ave	Lincoln Rd	Lincoln CT	Washington Ave	West Ave
3 (Validation)	Number of beds	**1.34**	1.30	1.20	1.22	1.19	1.20	1.20	1.17	1.25
	Number of full baths	**1.38**	1.31	1.26	1.25	1.13	1.21	1.20	1.28	1.27
	Year built	1.07	1.12	1.16	1.15	1.06	1.20	**1.21**	1.18	1.20
	Application fee	0.78	0.87	0.85	0.76	**1.23**	0.94	0.97	0.72	0.89
	Number of garage spaces	**1.23**	1.17	1.10	1.07	1.08	1.11	1.13	1.04	1.21
	List price	1.80	1.80	1.72	1.69	**1.89**	1.69	1.71	1.73	1.79
	FLP total value	1.28	1.28	1.21	1.10	0.98	1.08	1.15	**1.32**	1.27
	Living area	**1.37**	1.35	1.25	1.16	1.18	1.23	1.24	1.17	1.31
	Tax amount	1.08	**1.16**	0.87	1.00	0.09	0.82	0.85	1.09	0.99
4 (Training)	Number of beds	**1.33**	1.30	0.99	1.22	1.20	1.20	1.20	1.17	1.24
	Number of full baths	**1.37**	1.31	1.03	1.25	1.15	1.20	1.20	1.27	1.26
	Year built	1.07	1.12	0.95	1.15	1.07	1.20	**1.23**	1.17	1.21
	Application fee	0.77	0.86	0.69	0.76	**1.21**	0.96	0.96	0.73	0.90
	Number of garage spaces	**1.23**	1.16	0.91	1.07	1.09	1.10	1.11	1.03	1.18
	List price	1.80	1.80	1.41	1.68	**1.89**	1.72	1.72	1.72	1.79
	FLP total value	1.26	1.28	1.01	1.08	0.99	1.09	1.17	**1.32**	1.27
	Living area	**1.37**	1.35	1.03	1.16	1.19	1.23	1.24	1.18	1.28
	Tax amount	1.08	**1.16**	0.72	0.98	0.15	0.83	0.90	1.07	0.95
4 (Validation)	Number of beds	**1.34**	1.30	1.22	1.22	1.19	1.20	1.20	1.16	1.24
	Number of full baths	**1.38**	1.31	1.27	1.25	1.13	1.21	1.20	1.26	1.27
	Year built	1.07	1.12	1.17	1.15	1.06	1.20	**1.21**	1.17	1.19
	Application fee	0.76	0.87	0.85	0.75	**1.24**	0.93	0.97	0.72	0.90
	Number of garage spaces	**1.24**	1.16	1.11	1.07	1.07	1.11	1.13	1.04	1.20
	List price	1.80	1.80	1.74	1.68	**1.89**	1.70	1.71	1.71	1.79
	FLP total value	1.27	1.27	1.23	1.09	0.97	1.08	1.15	**1.29**	1.26
	Living area	**1.38**	1.35	1.27	1.17	1.17	1.23	1.24	1.16	1.30
	Tax amount	1.08	**1.15**	0.90	0.99	0.06	0.82	0.85	1.04	0.98

5 (Training)	Number of beds	**1.34**	1.30	1.20	1.22	1.19	1.21	1.21	1.17	1.25
	Number of full baths	**1.38**	1.31	1.26	1.26	1.13	1.21	1.21	1.28	1.27
	Year built	1.07	1.12	1.16	1.16	1.06	1.20	**1.22**	1.18	1.20
	Application fee	0.76	0.88	0.84	0.76	**1.24**	0.95	0.97	0.72	0.88
	Number of garage spaces	**1.23**	1.18	1.10	1.07	1.07	1.10	1.12	1.04	1.20
	List price	1.79	1.80	1.72	1.69	**1.89**	1.71	1.71	1.72	1.77
	FLP total value	1.26	1.28	1.22	1.10	0.97	1.09	1.15	**1.32**	1.27
	Living area	**1.37**	1.35	1.25	1.16	1.17	1.23	1.24	1.17	1.30
	Tax amount	1.09	**1.16**	0.88	1.00	0.06	0.84	0.85	1.09	1.00
5 (Validation)	Number of beds	**1.34**	1.30	1.20	1.22	1.20	1.20	1.20	1.16	1.23
	Number of full baths	**1.38**	1.31	1.26	1.25	1.14	1.21	1.20	1.27	1.26
	Year built	1.07	1.13	1.15	1.15	1.07	1.203	**1.24**	1.18	1.22
	Application fee	0.76	0.87	0.85	0.75	**1.23**	0.946	0.966	0.72	0.90
	Number of garage spaces	**1.23**	1.18	1.07	1.07	1.09	1.11	1.13	1.20	1.19
	List price	1.80	1.80	1.71	1.68	**1.89**	1.70	1.72	1.73	1.78
	FLP total value	1.27	1.28	1.20	1.09	0.98	1.07	1.19	**1.34**	1.26
	Living area	**1.38**	1.35	1.23	1.17	1.19	1.23	1.25	1.17	1.28
	Tax amount	1.09	**1.16**	0.87	1.00	0.10	0.82	0.94	1.10	0.97

Table 8.5 Deviation error of centroids

Iteration	Alton Rd	Bay Rd	Collins Ave	Dade Blvd	James Ave	Lincoln Rd	Lincoln CT	Washington Ave	West Ave
1	13.11%	16.88%	7.9%	10.6%	5.8%	11.3%	6.4%	4.6%	17.0%
2	11.46%	12.72%	14.1%	11.0%	1.9%	7.1%	6.6%	6.7%	8.1%
3	10.12%	7.79%	10.0%	11.3%	18.3%	7.4%	10.7%	5.1%	7.3%
4	11.02%	7.69%	1.9%	7.1%	10.4%	12.0%	11.5%	5.3%	15.3%
5	5.215%	11.60%	6.9%	7.7%	10.3%	11.5%	10.0%	3.9%	26.1%
Avg. error	10.18%	11.3%	8.2%	9.6%	9.3%	9.9%	9.0%	5.1%	14.8%
Correct clustering	89.8%	88.6%	91.7%	90.3%	90.6%	90.1%	90.9%	94.8%	85.2%

and eigen centrality is applied to check the best condominium.Let us consider following values for the attributes: number of garage spaces = 3, application fee = 400, number of full bathrooms = 3, number of bedrooms = 2, built year = 1986, taxable property value = 1,942,446, living area = 1007 Sq. ft, tax amount = 8633, and list price = 2,000,000. It was observed that the landmark selected by layer 1 was James Ave, and 401 condominiums were selected by layer 2 in that landmark. After applying eigen centrality, out of 401 condominiums condominium-1701 was selected as the most central condominium. The obtained complex network of real estate scenario including attributes and landmarks is shown in Fig. 8.4, in which the blue-colored circles are the attributes and the red-colored triangles are the condominiums. The links are having varying colors based on their weights (χ magnitude). If the link weight is more than 1 then the color is green; else it is blue.

One of the important purposes of using network science is to study the relationship between the condominiums and the attributes, and their influences on each other. Consider another example, where the attributes are set for simulations like the following: number of beds = 2; number of garage spaces = 2; number of full bathrooms = 2; and application fee = 126. According to the proposed algorithm, Alton Rd was the result of the decision tree and 169 condominiums were selected by layer 2. According to eigen centrality condominium-6487 was selected as the best condominium. The obtained network is as shown in Fig. 8.5.

To know which attribute remains consistent under the link weight variations, the link weights of the network in the Fig. 8.5 are varied (such that percent of the link weight was added to the weight as random noise) and each time the eigen centrality values were noted. At 0% variation (or wihtout variations) of link weights, the obtained centrality values are as shown in Table 8.6. The link weights were added in steps of 10% of their existing weights and the changes in the centrality measures of the network nodes were observed.

It was observed that as the link weight increases, *F*'s centrality value also increases and at a point of 30% increase *L* loses its central position and *F* becomes the more central attribute (Fig. 8.6). Hence, it was concluded that adjusting the weight of the links controls the centrality of the nodes in the network. This also implies that the more the correlation of an attribute with the real estate price of a condominium in a landmark, the more that attribute will become central in the

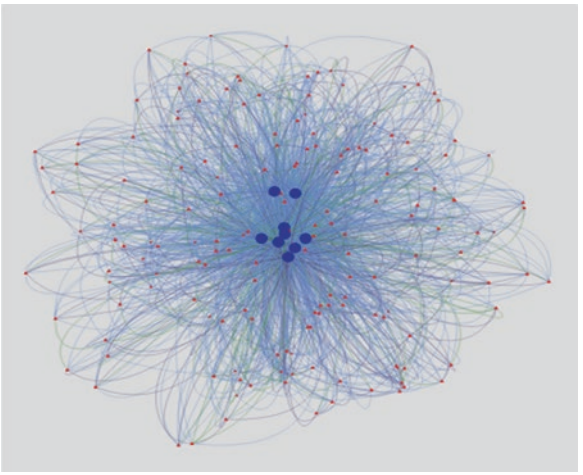

Fig. 8.4 Complex real estate network of condominiums of James Ave

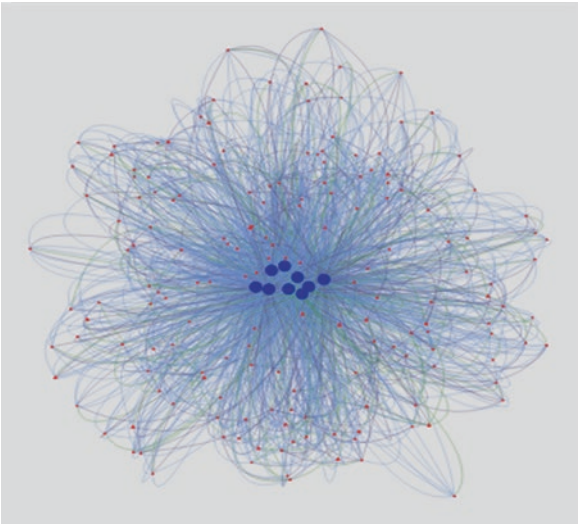

Fig. 8.5 Complex real estate network of condominiums of Alton Rd

Table 8.6 Eigen centrality values at 0% change in the link weight

Attribute	Eigen centrality value
Year built (Y)	0.4426
Number of garage spaces (G)	0.5621
Tax amount (X)	0.6645
Application fee (A)	0.6921
Living area (R)	0.7502
FLP total value 1 (T)	0.7648
Number of bedrooms (B)	7651
Number of full baths (F)	0.7787
List price (L)	1.0000

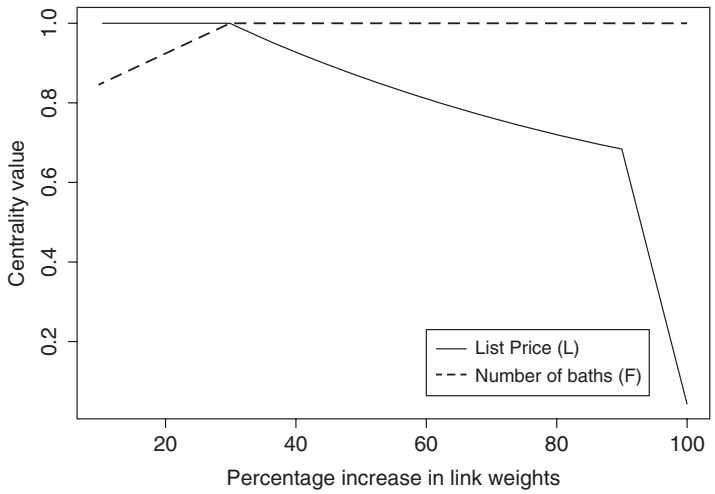

Fig. 8.6 Effect of the link weight change on the centrality values of the top two attributes

network. The same explanation holds true for the case when the link weights are decreased. In another experiment, the weights associated with all attributes were increased to a maximum 10% of their values' existing weights randomly, to check the most stable attribute. This helps us to understand the most consistent attribute due to sudden uncertain inflations. This simulation indicated that during sudden changes in the correlation between an attribute and real estate price, i.e., χ value, which may be due to natural calamities, inflation, and so on, list price attribute remains stable attribute by being most influential in the real estate investment. The same analogy can be drawn on condominiums as well and the most consistent condominium can be found.

8 Implications of Results on Smart Governance

In this section, two major applications of the results obtained due to this research study are highlighted. Firstly, the impact of the various factors on major public utility centers in a city (Fig. 8.7): The public utilities may be bus stations, airports, train stations, shopping malls, highways, and harbors and the factors include weather, ongoing protest/strikes, and road traffic. It is clear that in this network, there are two parties: the public utilities and the factors influencing them. Hence, a bipartite network will give a relationship between these parties. The thickness of the edges indicates the intensity of impact on the entity by that respective factor. From the sample network it is clear that the shopping malls in the city are highly impacted by all these factors and the harbor is least affected. A detailed analysis of how the various shopping malls are affected by the factors will give a detailed analysis of the

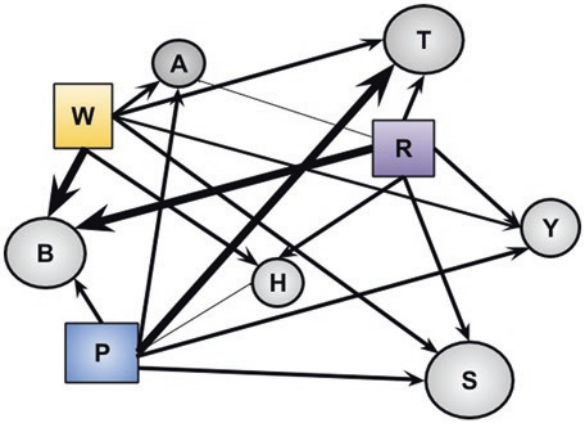

Fig. 8.7 Impact network (*W* weather, *A* airport, *T* train station, *B* bus station, *P* ongoing protest, *H* harbor, *S* shopping mall, *Y* highway, *R* road traffic)

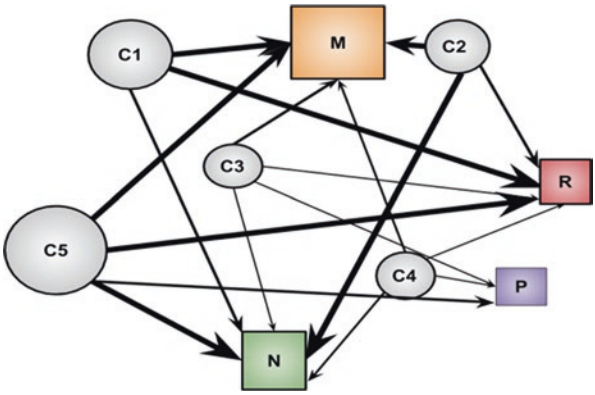

Fig. 8.8 Crime network (*N* molestation, *M* murder, *R* rape, *P* robbery)

relationship. These kind of inferences can be drawn in the same lines as that of the real estate network analysis. All factors influencing the entities may be considered or just top factors/attributes depending on the requirements of the study. If there are many attributes influencing the entities, the best attributes can be selected using the technique mentioned in the above sections.

Another interesting application is the crime studies. In Fig. 8.8, there are five criminals namely C1–C5 and the criminal activities in which they were indulged include murder, robbery, rape, and molestation. The network studies on the graph inference that murder is the most central activity performed by these criminals in that city and among all C5 turns out to be central. Knowing the inference we can go a step deeper and analyze the murder activity in the city in more detail. These are two among numerous scenarios that could be studied using a combination of data and network science.

9 Conclusion

Identification of location has always been a complex task for a user in real estate investment for affordable housing. In this context, the work proposed in this chapter uses the concepts of data science like statistical modeling which uses the Pearson correlation as the means to identify best attributes, and stacked machine learning techniques like decision trees, PCA, and *K*-means clustering for identification of location. Use of these machine learning algorithms is just a demonstration use case; however other techniques like artificial neural networks, deep learning networks, support vector machines, and so on can also be used. For the locations obtained from the machine learning layers, a bipartite network is constructed and the best location is selected in the presence of the influence of other attributes using eigen centrality. Combining of data and network analytics to obtain more insight into the location identification problem has not been explored much in the existing literature. Hence, this combination provides a more comprehensive approach to affordable housing in real estate investment. In addition, the methodology and the results obtained can be adapted for solving other issues in smart governance.

Acknowledgements Authors would like to thank Dr. Naphtali Rishe and Dr. S.S. Iyengar of School of Computing and Information Sciences, Florida International University, Miami, Florida, for providing the database and valuable suggestions throughout this work.

References

1. Internet document—"Affordable Housing in India", An Inclusive Approach to Sheltering the Bottom of the Pyramid
2. J.F. Schram, *Real Estate Appraisal* (Rockwell, Bellevue, 2006)
3. D.H. Carr, J. Lawson, J. Schultz, Dearborn Real Estate Education, *Mastering Estate Appraisal* (Dearborn Financial Publications, Chicago, 2003)
4. Y. Zhang, S. Liu, S. He, Z. Fang, Forecasting research on real-estate prices in Shanghai, in: *2009 International Conference on Grey Systems and Intelligent Services (GSIS 2009)*, Nanjing, 2009, pp. 625-629
5. W. Wei, T. Guang-ji, Z. Hong-rui, Empirical analysis on the housing price in Harbin City based on hedonic model, in: *2010 International conference on Management Science and Engineering 17th Annual Conference Proceeding*, Melbourne, VIC, 2010, pp. 1659–1664
6. B. park, J.K. Bae, Using machine learning algorithms for housing price prediction: "The case of Fairfax County", Virginia housing data. Expert Syst. Appl. **42**(6), 2928–2934 (2015). ISSN: 0957-4174
7. H. Xue, The prediction on residential real estate price based on BPNN, in: *2015 8th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, Nanchang, 2015, pp. 1008–1013
8. B. Liu, B. Mavrin, D. Niu, L. Kong, House price modeling over heterogeneous regions with hierarchical spatial functional analysis, in: *2016 IEEE 16th International Conference on Data Mining (ICDM)*, Barcelona, 2016, pp. 1047–1052
9. C. Cheng, X. Cheng, M. Yuan, K. Chao, S. Zhou, J. Gao, L. Xu, T. Zhang, A novel architecture and machine learning algorithm for real estate. Signal Inf. Process. Netw. Comput. **473**, 491–499 (2017). Springer, Singapore. Lecture Notes in Electrical Engineering

10. Kecheng Zhao, Wei Shen, Spatial characteristic with individual house properties and multi-level approach to hedonic models, in: *2011 International Conference on Computer Science and Service System (CSSS)*, Nanjing, 2011, pp. 2579–2582
11. T. Oladunni, S. Sharma, Hedonic housing theory—a machine learning investigation, in: *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Anaheim, CA, 2016, pp. 522–527. doi: https://doi.org/10.1109/ICMLA.2016.0092
12. B. Park, J.K. Bae, Using machine learning algorithms for housing price prediction. Expert Syst. Appl. **42**, 2928–2934 (2015)
13. I.D. Wilson, S.D. Paris, J.A. Ware, D.H. Jenkins, Residential property price time series forecasting with neural networks, in: *The Twenty-First SGES International Conference on Knowledge Based Systems and Applied Artificial Intelligence*, Cambridge, December 2001, pp. 17–28, Springer Publications
14. H. Xu, A. Gade, Smart real estate assessments using structured deep neural networks, in: *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, San Francisco, CA, 2017, pp. 1-7
15. S. Lu, Z. Li, Z. Qin, X. Yang, R.S.M. Goh, A hybrid regression technique for house prices prediction, in: *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, Singapore, 2017, pp. 319–323
16. D. Sangani, K. Erickson, M. A. Hasan, Predicting zillow estimation error using linear regression and gradient boosting, in: *2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, Orlando, FL, 2017, pp. 530–534
17. W.T. Lim, L. Wang, Y. Wang, Q. Chang, Housing price prediction using neural networks, in: *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Changsha, 2016, pp. 518–522
18. J. Demongeot, H. Pempelfort, J.M. Martinez, R. Vallejos, M. Barria, C. Taramasco, Information design of biological networks: application to genetic, immunologic, metabolic and social networks, in: *2013 27th International Conference on Advanced Information Networking and Applications Workshops*, Barcelona, 2013, pp. 1533–1540
19. D.P. Cheung, M.H. Gunes, A complex network analysis of the United States air transportation, in: *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Istanbul, 2012, pp. 699–701
20. Q. Xuan, Z.Y. Zhang, C. Fu, H.X. Hu, V. Filkov, Social synchrony on complex networks. IEEE Trans. Cybernetics **48**(5), 1420–1431 (2018)
21. ESRI—Real estate website, https://www.esri.com/en-us/industries/real-estate/overview
22. Black stone, https://www.blackstone.com/the-firm/asset-management/real-estate
23. J. Wang, Pearson correlation coefficient, in *Encyclopedia of Systems Biology*, ed. by W. Dubitzky, O. Wolkenhauer, K. H. Cho, H. Yokota, (Springer, New York, NY, 2013)
24. The data for our work was taken from: www.terrafly.com/
25. G. Skourletopoulos et al., Big data and cloud computing: a survey of the state-of-the-art and research challenges, in *Advances in Mobile Cloud Computing and Big Data in the 5G Era. Studies in Big Data*, ed. by C. Mavromoustakis, G. Mastorakis, C. Dobre, vol. 22, (Springer, Cham, 2017)
26. Alan Said, Data Science in Practice (Springer Publications, 2019)
27. E. Hart, Biological networks, in *Encyclopedia of Astrobiology*, ed. by R. Amils et al., (Springer, Berlin, Heidelberg, 2014)
28. V. Latora, V. Nicosia, G. Russo, *Complex Networks: Principles, Methods and Applications* (Cambridge University Press, Cambridge, UK, 2017)
29. S. Han, Y. Ko, S. Kim, D.H. Shin, Home sales index prediction model based on cluster and principal component statistical approaches in a big data analytic concept. KSCE J. Civil Eng. **21**(1), 67–75 (2017). Springer publications
30. T. Oladunni, S. Sharma, Spatial dependency and hedonic housing regression model, in: *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Anaheim, CA, 2016, pp. 553–558

31. V. Del Giudice, P. De Paola, G.B. Cantisani, Valuation of real estate investments through fuzzy logic. Buildings **7**, 26 (2017)
32. C. Bagnoli, H.C. Smith, The theory of fuzzy logic and its application to real estate valuation. J. Real Estate Res. **16**(2), 169–200 (1998). American Real Estate Society
33. W. Didimo, G. Liotta, F. Montecchiani, Network visualization for financial crime detection. J. Vis. Lang. Comput. **25**(4), 433–451 (2014). https://doi.org/10.1016/j.jvlc.2014.01.002
34. S. Perera, M.G.H. Bell, M.C.J. Bliemer, Network science approach to modelling the topology and robustness of supply chain networks: a review and perspective. Appl. Netw. Sci. **2**, 2–35 (2017). https://doi.org/10.1007/s41109-017-0053-0
35. Z. Wang, J. Han, Visualization of the UK stock market based on complex networks for company's revenue forecast, in *Information and Knowledge Management in Complex Systems. ICISO 2015. IFIP Advances in Information and Communication Technology*, ed. by K. Liu, K. Nakata, W. Li, D. Galarreta, vol. 449, (Springer, Cham, 2015)
36. Realdata, https://www.realdata.com/
37. CREmodel, https://www.cremodel.com/
38. Proapod, http://www.proapod.com/
39. Y. Dong, Value ranges of Spearman's Rho and Kendall's Tau of a class of copulas, in: *2010 International Conference on Computational and Information Sciences*, Chengdu, 2010, pp. 182-185. doi: https://doi.org/10.1109/ICCIS.2010.335
40. S.E. Kumar, V. Talasila, N. Rishe, T.V.S. Kumar, S.S. Iyengar, Location identification for real estate investment using data analytics. Int. J. Data Sci. Analytics, 1–25 (2019)
41. M.J. Moshkov, Time complexity of decision trees, in *Transactions on Rough Sets III*, ed. by J. F. Peters, A. Skowron, (Springer, Berlin, 2005), pp. 244–459
42. D. Hu, Q. Liu, Q. Yan, Decision tree merging branches algorithm based on equal predictability, in: *2009 International Conference on Artificial Intelligence and Computational Intelligence*, Shanghai, 2009, pp. 214–218
43. O.Z. Maimon, R. Lior, *Data Mining with Decision Trees: Theory and Applications*, 2nd edn. (World Scientific, Singapore, 2015)
44. S. Sehgal, H. Singh, M. Agarwal, V. Bhasker and Shantanu, Data analysis using principal component analysis, in: *2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom)*, Greater Noida, 2014, pp. 45–48
45. G.A. Wilkin, X. Huang, K-means clustering algorithms: implementation and comparison, in: *Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)*, Iowa City, IA, 2007, pp. 133–136
46. I. Scholtes, Understanding complex systems: when big data meets network science. IT—Information Technology **57**(4), 252–256 (2015). https://doi.org/10.1515/itit-2015-0012
47. A. Bihari, M. K. Pandia, Eigenvector centrality and its application in research professionals' relationship network, in: *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, Noida, 2015, pp. 510–514
48. F. Grando, D. Noble, L. C. Lamb, An analysis of centrality measures for complex and social networks, in: *2016 IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, 2016, pp. 1–6