International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-6 Issue-2, May 2017

Data Mining Approach To Find the Interest of People in Purchasing Real Estate

Abhishek Yadav, Rajeswari C

Abstract: ***Data mining is extraction of information and consolidates it to helpful data which can be utilized for future calculation and prediction of an event. Subsequently by utilizing data mining methods we are anticipating the interest of individuals in different types of real estate and we are also defining certain pattern can be helpful in purchasing them. In this research we are going to gather individual's enthusiasm for the kind of properties, and other different kinds of information and transform them into data chains utilizing certain data-mining algorithms by means of which we can predict people's interest about what type of property they are likely going to buy and at what locations they are most likely to buy property. The data has been collected from websites such as OLX.com. In this research two data mining techniques that have been used to classify the data on the basis of certain attributes are Classification (zero classifier) and clustering (simple k means) and on the basis of results several graphs and bar charts are drawn.***

Problem Statement: Understanding the problems of client in real estate field about what properties to buy, what locations to select and how to utilize those Properties and defining different approaches to resolve it using certain data mining concepts and algorithms.

I. INTRODUCTION

Today real estate sector is one of the quickest developing markets in the world. The Indian real estate sector has witnessed high growth in recent times with the rise in demand for office as well as residential spaces for example the demand for office space has been increasing in some of the major cities of India like Delhi, Mumbai, Pune, Chennai, Kolkata etc. Meanwhile the demand for residential space is also increasing in various cities like Bangalore, Delhi, etc.

Therefore this research paper is going to help people in predicting their interest in buying different properties in different cities according to their convenience. In this research paper we will be collecting data regarding various real estates, cities and facilities available in the respective properties. And after collecting and refining all the datasets, we will be using certain data mining techniques such as Classification (zero classifier) and clustering (simple k means) to determine certain attributes through the help of which people can be able to determine which city best suits them for buying different kind of properties that will be either for residential and renting purposes or for office use.

Revised Manuscript Received on 30 May 2017.
***** Correspondence Author
Abhishek Yadav, Department of Information Technology & Engineering, VIT University, Vellore, Tamil Nadu 632014, India.
Rajeswari C, Department of Information Technology & Engineering, VIT University, Vellore, Tamil Nadu 632014, India.

Through this research paper people might come to know what type of property is available, area of property in terms of square feet, facilities available in those properties and also the price of the property.

The goal of the data mining process is to extract information from a data set and transform it into understandable structure for further use. Data mining tools helps in predicting future trends so that organizations are able to make knowledge based decisions. Hence using data mining techniques we can predict the interest of people in real estate and their pattern of purchasing them.

We are going to follow an iterative process for predicting the result. Knowledge discovery as a process consists of an iterative sequence of the following steps:

1. Data cleaning: In this step, noise or irrelevant data is removed.
2. Data integration: In this step, multiple data sources may be combined.
3. Data selection: data relevant to the analysis task are retrieved from the database.
4. Data transformation: In this step, data are transformed into forms appropriate for mining by performing summary or aggregation operations.
5. Data mining: An essential process where intelligent methods are applied in order to extract data patterns.
6. Pattern evaluation: In this step we have to identify the truly interesting patterns representing knowledge based on some interestingness measures.
7. Knowledge presentation: visualization and knowledge representation techniques are used to present the mined knowledge to the user.

ZeroR classifier classification technique predicts the values which are more likely to be available or use for example ZeroR predicts class value flat from types of property available attribute and class value Delhi from attribute locations which means these values are more likely to be chosen while looking for properties in different cities.

Simple k means clustering technique predicts the most suitable clusters on the basis of different attributes. For example using class location two clusters are formed that are Delhi and Greater Noida. Using class reason two clusters are formed that are buying and residential.

II. LITERATURE REVIEW

This section explains about data mining and its techniques. Also, it describes the used data mining tool in the research. The discussion of previous work in using data mining techniques for predicting the interest of people in purchasing real estate properties is included in this section as well.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

A. Data-Mining

The main target of data mining is to discover previously unknown relationships between several features and attributes of data. Data mining is also called knowledge discovery and data mining (KDD). It is the extraction of useful patterns from data sources,
for example databases, texts, web, images, etc. Patterns must be valid, novel, potentially useful, and understandable. There are several techniques that can be used in data-mining, for example: classification, clustering, association rules, etc.

The following section gives brief information about the used data mining techniques in this project, which are as follows:-

B. Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms.

Types of classification techniques are:

: - Classification by trees such as Decision Stump and Random Forrest.
: - Bayes Classification such as Bayes Net and NativeBayes.
: - Functions such as Simple Logistics and SMO.
: - Classification based on rules such ZeroR, OneR and DecisionTable.

C. Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes.

Types of clustering methods

: - Simple K Means
: - Hierarchical Clusterer
: - Make DensityBased Clusterer
: - Filtered Clusterer

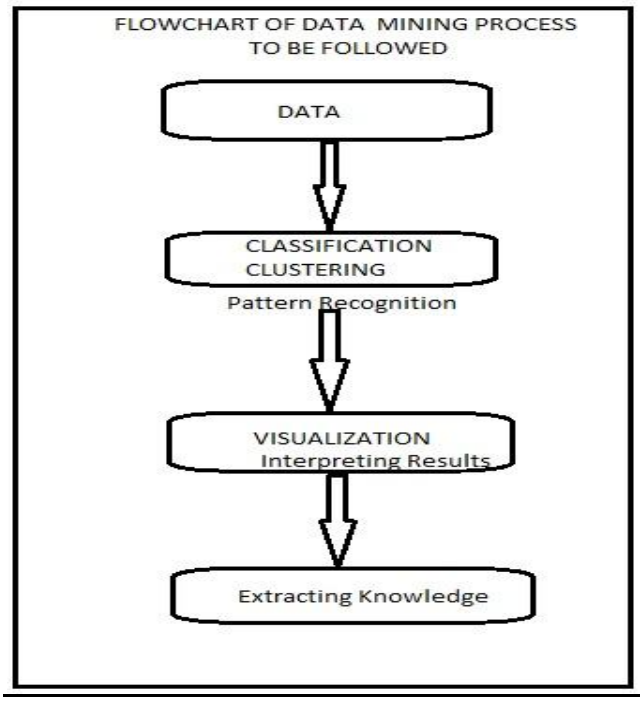
III. RELATED WORK

There are several research papers that worked on examining the attributes that contribute to the finding the interest of people in purchasing real estate properties using different data mining techniques. Based on dataset available, Vishal Venkat Raman used 2 data mining modeling techniques namely, linear regression and rule induction for predicting the area for a real estate customer using Rapid Miner tool [1]. Data mining techniques are used in two different phases, where the first one deals with ranking of areas in each of the 5 zones of Delhi and identify an ideal location for real estate customers using the Info Gain Attribute Eval function of WEKA tool. The second phase is about predicting the most suitable area for any given customer based on his/her interest. We predict using a classical technique called linear regression and try to give an analysis of the results obtained. A total of 23 attributes are listed along with area in which the real estate is located and the number of bedrooms available in it. Vishal collected about 500 data sets for this process.

In the first phase, WEKA could rank the attributes which influence the pricing of a real estate. The most influencing attribute as calculated by WEKA was bus stops, followed schools, colleges and so on. In second phase, a suitable area for any given real estate customer has been predicted. According to the linear regression technique used in this paper, the most suitable and preferred residential area has been identified.

In another research paper, [2]Swati Singh developed a model by using various existing data mining techniques that defines the pattern or interest of people during purchasing real estate. She used two data mining techniques that are classification (Zeror algorithm) and clustering (simple k means) on the collected data. She prepared a questionnaire, which contains 25 multiple choice questions based *on real estate* and a survey has been done based on this questionnaire, as a part of data collection**.** Here the sample of questions distributed among 300 people and the information was gathered. This questionnaire contains questions on real estate that generally asked by a common man while he/she is interested in purchasing a property. The queries are basically related to for e.g. Reason behind purchasing the property, their interest in type of property, their budget, the area they need for their property, their desired location, and many more. On the basis of dataset different patterns and bar charts had been drawn based on their result. According to the patterns, 22.1% of people are interested in purchasing office and they want their deal must be done by agent (Dealer). The choices for the location are Delhi, Gurgaon, Gr. Noida but they give the preference to Gurgaon.

A. Theoretical Background



Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

B. Clustering

Finding groups of objects (clusters)
: - Objects similar to one another in the same group
: - Objects different from the objects in other groups

C. Clustering Approaches

Partitioning approach: Construct various partitions and then evaluate them by some criterion. Typical methods: k-means, k-medoids, CLARANS.

Simple K Means: - Each cluster is represented by the center of the cluster. The main goal is to classify data into groups of information. Consists on the separation of all information observations of a specific dataset into k different clusters which aggregate each one of the data entries. For this are defined *k* center points, one for each cluster, called centroids. Each piece of data is connected to the partition with the nearest mean from that point, that is, the nearest centroid.

As a result of this operation we'll get *k* sets of data entries, each one related with one specific centroid. Within these sets, the distance to the respective centroid is minimized.

Algorithm

The process to achieve the result sets of classified data is quite simple. It basically consists on **several iterations** of a specific process, designed to get a **optimal minimum solution** for all data points.

Let's look this process in detail.

First, we need to establish a **function** of what we want to minimize, in our case the distance between every data point and the correspondent centroid.

So, what we want is:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} \|x_j - c_i\|^2$$

With this function well defined, we can split the process in **several steps**, in order to achieve the wanted result. Our starting point is a **large set of data entries** and a ***k*** defining the number of centers.

1 – The first step is to choose randomly ***k*** of our points as partition centers.

2 – Next, we **compute** the **distance** between every data point on the set and those centers and store that information.

3 – Supported by the last step calculations, we **assign** each point to the **nearest cluster center**. This is, we get the minimum distance calculated for each point, and we add that point to the specific partition set.

4 – Update de cluster center positions by using the following formula:

$$c_i = \frac{1}{|k_i|} \sum_{x_{j \in k}}^{k} x_j$$

5 – If the cluster centers change, repeat the process from **2**. Otherwise you have **successfully computed** the k means clustering algorithm and got the **partition's members** and **centroids**.

The achieved result is the **minimum configuration** for the selected start points. It is possible that this output isn't the optimal minimum of the selected set of data, but instead a **local minimum** of the Function. To mitigate this problem, we can run process more than one time in order to get the **optimal solution**.

It is important for you to know that there are some **variations of the initial center choice method**. Depending on the problem you want to solve, some initial processes might benefit your implementations.

IV. CLASSIFICATION

Classification is a data mining technique that assigns categories to a collection of data in order to aide in more accurate predictions and analysis. Also called sometimes called a Decision Tree, classification is one of several methods intended to make the analysis of very large data sets effective. Data classification is the process of organizing data into categories for its most effective and efficient use.

A. Classification Approaches

i. ZeroR Classifier

ZeroR is the simplest classification method which relies on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods. Predicts the mean (for numeric class) or mode (for nominal class).

ii. Visualization

We can visualize the available dataset in weka by plotting x-y axis graph which shows different points of different colours depicting the relationship between attributes. We can also distinguish between points that are in fact very close together using jitter slider. Another thing we can do is to select bits of the given dataset using select instance option available exclude other points. So that can be the way of cleaning up outliers in our dataset by selecting instances and saving the new datasets.

We can also visualize the result of the classifier. Here we can visualize classifier errors by getting the class plotting against the predicted class. Here square boxes represents errors and if we click on these boxes we can see where the errors are. We can visualize any number of instances where predicted class is different from existing classes. Basically the entries in the confusion matrix are represented by different instances. We can look at the plot and find out where the miss classifications are actually occurring, the errors in the confusion matrix.

B. Visualizing the Data

: -We can visualize it.
: -Clean it up by deleting outliers.
: - look at classification errors.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

V. EXPERIMENT AND RESULT

A dataset has been created based on the characteristics of real estate. This dataset contains 16 different attributes. These attributes are reason for purchasing the property, people's interest in type of property, location of different properties, area of property and many more. The dataset is then saved into.

CSV file format so that it can be used in Weka tool for performing different tasks. The dataset is then loaded into the Weka tool and as soon as the tool receives the dataset, it will pre-process the data.

During pre-processing of data, the Weka will calculate the number of instances, attributes and sum of weights of attributes. For example in the current relation the number of attributes are 16 and instances are 19. The Weka also calculates the count of different labels of attributes and on the basis of that bar charts are drawn. The next task is to classify the given dataset and for classification ZeroR classifier has been used. ZeroR predicts the majority category (class). Therefore ZeroR classifier predicts the class value: buying with an accuracy of 47.3684% that is correctly classified instances are 9 and incorrectly classified instances are 10. Therefor people's reason for purchasing property is to be buying it for future investments.

Zeror predicts class value "Buying", when classification is done by using class "Reason".

```
=== Run information ===

Scheme:       weka.classifiers.rules.ZeroR
Relation:     real estate
Instances:    19
Attributes:   16
              reason
              type of property
              deal through
              location
              budget
              type of flats
              furnished
              area of flat
              area of office
              area of plot
              farm house
              loan facility
              way of payment
              wifi
              AC
              priority of floor
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: buying
```

The confusion matrix shows that ZeroR predicts the majority class correctly:-

```
=== Confusion Matrix ===

 a b c d   <-- classified as
 9 0 0 0 | a = buying
 3 0 0 0 | b = investment
 6 0 0 0 | c = residential
 1 0 0 0 | d = renting
```

The confusion matrix is also known as contingency table and in our case we have 4 classes therefore 4*4 confusion matrix will be there. The number of correctly classified instances is the sum of diagonal elements and all other are incorrectly classified instances. Here a=buying is classified as one category which is having value 9 in confusion matrix.

Detailed Accuracy by Class:-

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	1.000	1.000	0.474	1.000	0.643	buying
	0.000	0.000	0.000	0.000	0.000	investment
	0.000	0.000	0.000	0.000	0.000	residential
	0.000	0.000	0.000	0.000	0.000	renting
Weighted Avg.	0.474	0.474	0.224	0.474	0.305	

The ***True Positive*** *(TP)* rate is the proportion of examples which are classified as class *buying*, among all examples which are truly having class *buying*, i.e. how much part of the class was captured. In confusion matrix, it is the diagonal element divided by the sum over the relevant row, i.e. 9/ (9+0+0+0) =1.0 for class *buying* in our example dataset.

The ***False Positive*** (FP) rate is the proportion of cases which were classified as class buying, but actually belong to an alternate class, among all cases which are not of class buying. In confusion matrix, it is the column sum of class *buying* minus the diagonal element, divided by the rows sums of all other classes, i.e. (19-9)/(3+6+1) = 1.0.

Precision is TP / predicted Positive

What fraction of those predicted positive are actually positive. Precision is also referred to as Positive predictive value (PPV)

Here precision is 1/(1+1.109) = 0.474.

Here 1.109 is false positive(FP) value.

Recall is TP / actual positives (also referred to as sensitivity) what fraction of those that are actually positive were predicted positive.

Here Recall is 1/ (1+0) =1.0.

Here 0 is false negative (FN) i.e. there are no examples which are predicted negative that are actually positive.

F-measure is the harmonic mean of precision and recall for buying attribute it is calculated as: - 2*[0.474/ (0.474+1)] = 0.643.

A. Clustering

Clustered Output:

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

```
Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       9 ( 47%)
1      10 ( 53%)


Class attribute: location
Classes to Clusters:

 0 1  <-- assigned to cluster
 0 3 | noida
 3 0 | gr noida
 0 1 | gurgaon
 1 0 | meerut
 4 4 | delhi
 1 2 | ghaziabad

Cluster 0 <-- gr noida
Cluster 1 <-- delhi

Incorrectly clustered instances :       12.0      63.1579 %
```

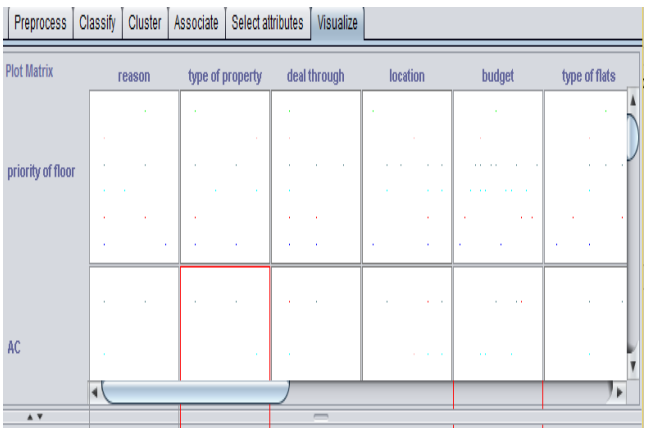
Clustering through Simple k means using class "location", that forms 2 clusters that are cluster 0: gr Noida, cluster 1: Delhi

Clustered Output:

```
Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       9 ( 47%)
1      10 ( 53%)


Class attribute: reason
Classes to Clusters:

 0 1  <-- assigned to cluster
 6 3 | buying
 3 0 | investment
 0 6 | residential
 0 1 | renting

Cluster 0 <-- buying
Cluster 1 <-- residential

Incorrectly clustered instances :       7.0       36.8421 %
```

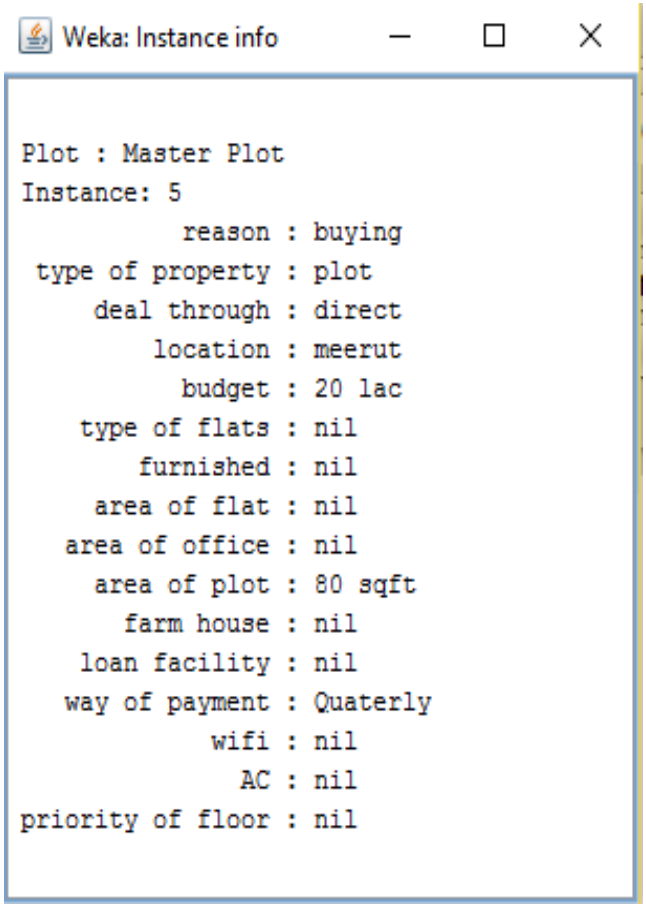
Clustering through Simple k means using class "Reason", that forms 2 clusters that are cluster 0: buying, cluster 1: residential.

VI. VISUALIZATION OF DATA



Here is the matrix of two dimensional plots, a 16 by 16 matrix of plots, after selecting one of these plots, we can look at a plot of location at x-axis and priority of floor at y-axis.

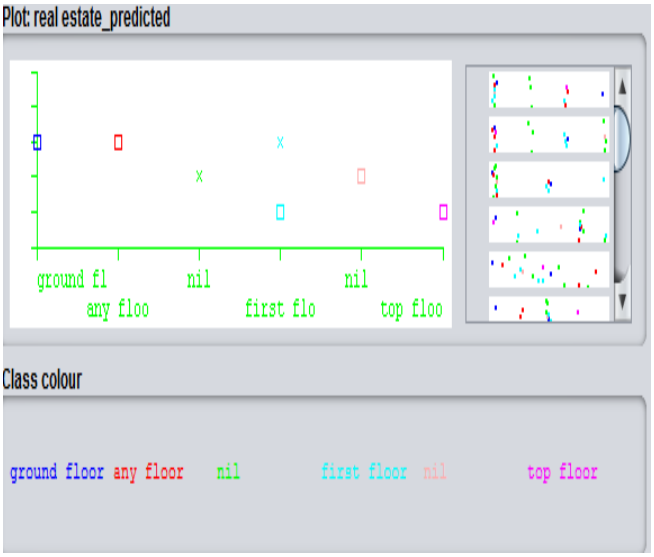
That's the plot of the data, the colour corresponds to the six classes. We can look at individual data points by clicking on them.



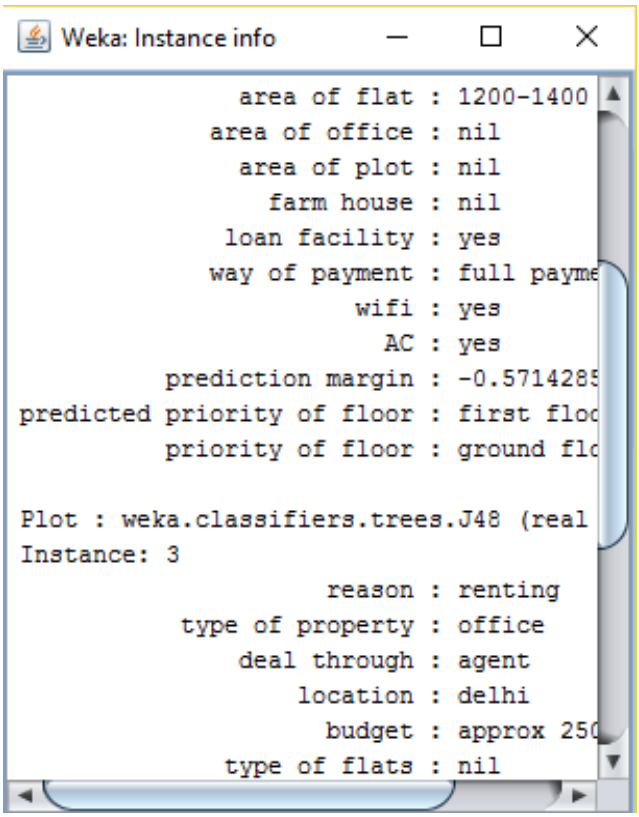
This data point is talking about instance number 5 with type of property as plot and location as Meerut and so on.

Here jitter slider helps us distinguish between points that are in fact very close together. We can also visualize the result of the classifier by selecting the J48 classifier under Tree and then choosing option visualize classifier errors.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)



Here we got the class plotted against the predicted class. The square boxes represent errors and if we click on one of these we can see where the errors are.



So there are two instances where predicted class is first floor and actual class is ground floor. So I can look at this plot and find out where the misclassifications are actually occurring in the confusion matrix.

REFERENCES

1. Vishal Venkat Raman, Swapnil Vijay, Sharmila Banu K Identifying Customer Interest in Real Estate Using Data Mining Techniques.
2. Swati Singh, Gaurav Dubey Finding interest of people in purchasing real estate by using data mining techniques.
3. Xian Guang LI, Qi Ming LI The Application of Data Mining Technology in real estate market prediction.
4. Itedal Sabri Hashim Bahia, Ministry of Higher Education and Scientific Research, Baghdad, Iraq A Data Mining Model by Using ANN for Predicting Real Estate Market: Comparative Study.
5. Ruben D. Jaen, Florida International University Park, PC236 Miami, Fl 33199Data Mining: An Empirical Application in Real Estate Valuation.