Attribute Selection in Hedonic Pricing Modeling applied to the Portuguese Urban Housing Market

Paulo Batista[1], Gladys Castillo[2], João L. Marques[1] and Eduardo A. Castro[1]

[1] Department of Social Sciences, Law and Politics, University of Aveiro, 3810-192 Aveiro, Portugal
[2] Department of Mathematics, University of Aveiro, 3810-192 Aveiro, Portugal
{pauloricardolb, gladys, jjmarques, ecastro} @ua.pt

Abstract. One of the challenges associated with studying the housing market is related to the need to handle a high amount of variables. In this context, data mining techniques, and more specifically, feature selection methods allow the selection of relevant variables efficiently. Results from the application of eight different methodologies for feature selection with a real dataset on the urban housing market of Aveiro and Ílhavo municipalities show that we can build hedonic models with an acceptable explanatory power of housing prices while considerable reducing their complexity.

Keywords: feature selection, regression model, hedonic pricing modeling, data mining

1 Introduction

A hedonic pricing model [18] decomposes the price of an item into separate factors that determine that price. This econometric tool based on the Lancaster's theory of consumer demand is actively used in housing studies. In particular, valuation of access to central and local services and other housing attributes, and construction of price indices based on single sales data, have been addressed through hedonic specifications (see [14] for a classic and critical discussion and [17] for a recent review). *Multiple linear regression* is the statistical tool most widely used in hedonic price modeling due to its simplicity and easy implementation. A house is modeled as a set of multi-dimensional attributes (independent variables) that are combined together to give a certain price (the response variable). In this study, we use the price per square meter of listing prices, rather than selling prices. Nevertheless, we consider this information a reasonable approximation of the real housing price.

Business Intelligence (BI) and Knowledge Discovery in Databases (KDD) are two research areas that share the common objective of extracting unknown and useful information from data [6]. This article focuses on the modeling of residential house prices in two Portuguese municipalities, Aveiro and Ílhavo, using a hedonic approach. The price models are estimated from a large dataset that includes the information about properties put on the market in the period from 2002 to 2010 in these two municipalities. One of the major challenges in applying hedonic models is the identification of the factors that are crucial to the market value.

As pointed out in [1], two main tasks are behind the induction of models from data. First, we need to decide which attributes should be used. Next we need to choose the functional form of the model that combines these attributes. From that point of view, the selection of relevant features is one of the primary problems in any KDD process. Given the great diversity of indicators and information associated with the housing market, the identification of attributes determining the house price is a basic objective of real estate agents, researchers in urban planning, as well as, government institutions that, directly or indirectly, are involved in studying this problem. In urban planning, for instance, the identification of a set of key indicators in the housing price allows to offer solutions in order to promote those house amenities that are most valued by the target population. In the real estate market this selection is crucial for the constructions of more compact and interpretable house price indexes that can further be implemented in internet website or portals.

The main aim of this study is to evaluate eight different feature selection schemes, used to determine the hedonic pricing regression model with the best trade-off between the number of selected attributes and its explanatory power. All the implementation was carried out using RapidMiner® 5.0 [16], one of the world-wide leading open-source software for data mining. Results show that the use of feature selection techniques allows a considerable reduction of the number of attributes without a significant loss of the explanatory power. This reduction leads to the construction of simpler and more interpretable hedonic price models that can be used to better understand the housing market and its major determinants.

The work presented here is part of the research project *Drivers Of housiNg demand in portuguese Urban sysTem* (DONUT), supported by *Fundação para a Ciência e Tecnologia* (FCT).

The remainder of the paper is organized as follows: in Section 2, we describe the housing market dataset as well as the pre-processing and feature selection methods implemented, following the design stages of the CRISP-DM methodology [5] for BI; in Section 3, we present and analyze the results of the performance evaluation of the eight feature selection techniques; finally, in Section 4 we present the conclusions and some future work.

2 KDD Methodology

In this project we have applied the CRoss Industry Standard Process for DM (CRISP-DM) methodology [5]. CRISP-DM is an iterative and adaptive process based on a *life cycle* approach of a KDD project. The *life cycle* comprises six key stages as depicted in Figure 1. Each phase is composed of several tasks, described hierarchically by several levels of abstraction, allowing to make the process clearest to data analysts with different backgrounds.

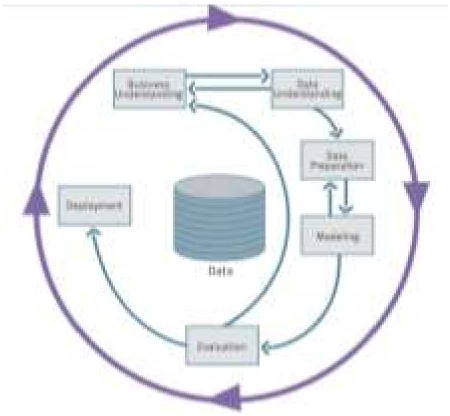

Fig. 1. Phases of the CRISP-DM model (extracted from the CRISP-DM Website)

In the following sections we will explain how we have implemented each of CRISP-DM stages in our study.

2.1 Business Understanding

First we need to understand the project goals from a business point of view, in order to formulate a data mining problem aimed at achieving those goals. The main objective of the current KDD process is to identify the determinants of the house price in Aveiro and Ílhavo municipalities using hedonic pricing models. This KDD process should be implemented taking into account the real limitations existing during the process of data collection and the selection of the initial attributes that allow us the description of the housing prices in the best possible way.

2.2 Data Understanding

In this phase we need to collect some data and identify data quality problems. To get information on the housing transaction market in Aveiro and Ílhavo it was essential the collaboration with the *Janela Digital S.A. company*[1], responsible for the *Casa Sapo* portal [4], the largest real estate website in Portugal. This portal concentrates a great amount of data and hence, valuable information for understanding the Portuguese real estate market. The house advertisings are mainly placed by real estate agencies with the help of the software *ImoGuia*, which allows the automatic publication of the ads in the portal. Less often, the ads can also be placed by the owners. The portal information can be free accessed by any potential buyer. However, the information collected in the portal has some inherent limitations. For instance, the lack of exclusivity in the housing market and the fact that the owners are not identified, due to privacy protection rules, can lead to the existence of

[1] http://www.janeladigital.com/

duplicate house records in the database. On the other hand, the seller may omit or modify some information of interest to the buyer that can lead to the devaluation of the house.

The raw data collected by *Janela Digital* includes housing property records published on the *Casa Sapo* from October 2000 to March 2010 for the municipalities of Aveiro and Ílhavo. Each property is described through three main type of attributes: *i*) *physical attributes:* representing the house basic physical structures (e.g., price per square meter, typology, area, level of preservation, etc.); *ii*) *location attributes:* describing the place where the house is located, defined by municipalities, parishes (administrative boundaries) and a disaggregation of parishes called *micro-zones,* that are homogeneous territories defined according to the housing characteristics; *iii*) a *free-text attribute,* where the clients can include descriptive information that they consider interesting for advertising (e.g. the existence of a balcony, fireplace, garage, etc.).

During a first assessment of the data quality some inconsistencies, that could become problematic for further data mining tasks, were identified. Aiming at the selection of a consistent sample of property records some cleaning processes were established based on the deletion of property records that: *i*) were not advertised for sale (some of the properties are advertised for renting purposes) ; *ii*) without location information (micro-zone or parish), *iii*) had data inconsistencies; *iv*) had missing values; and *v*) were duplicated.

2.3 Data Preparation (pre-processing phase)

In this phase we need to construct the final dataset from the initial raw data that will be used as input to modeling tools. The pre-processing tasks include cleaning and transformation of data as well as selection of attributes and examples. After the removal of records that satisfy the criteria above described and the addition of new attributes (described below in Section 2.3.1), the resulting database contains 19969 observations and 48 attributes. Among them, 28 attributes are numeric (real values), 4 are nominal and 16 are binominal. One of the 28 real attributes is the response variable, in our study taken as the house price per square meter (€/m^2). Finally, as suggested in regression tasks, the numeric attributes were normalized, the nominal variables were transformed into binary ones (obtaining 62 attributes) and highly correlated attributes were removed (4 attribute were deleted). As a result, for feature selection and modeling, we obtained a dataset with 19969 observations and 58 attributes.

2.3.1 Dataset Attributes

The resulting attributes are categorized into three types: *i*) *basic physical attributes; ii) descriptive physical attributes;* and *iii) spatial attributes.*

Basic Physical attributes:

Table 1 shows the name, type, a brief description and some statistics of these attributes. After applying binarization, a new binary attribute was generated for each value of the nominal attributes.

Table 1. Physical attributes characterization.

Name	Type	Description	Statistics	Range
PRICE_M2 **(target)**	*real*	*The house price in € / m².*	*avg = 1139.446 +/- 377.111*	[152.542 ; 714.286]
AREA	*real*	*The area in m²*	*avg = 151.130 +/- 81.057*	[20.000 ; 600.000]
PRESERVATION NEW	*nominal*	*Preservation level for a new house*	*mode = 0* *least = 2*	*0 - is not a new house,* *1- in construction* *2- constructed*
PRESERVATION USED	*nominal*	*Preservation level for a used house*	*mode = 0* *least = 4*	*0 – is not a used house* *1- 1 to 10 years old,* *2- 10 to 25 years old,* *3 - 25 years old* *4 - for demolition*
TYPE FLAT	*nominal*	*House type flat*	*mode = 2* *least = 1*	*0 – is not a flat* *1 - in apart. building with > 2 floors,* *2 - in apart. building with 2 floors.*
TYPE HOUSE	*nominal*	*House type dwelling*	*mode = 0* *least = 6*	*0 – is not a dwelling* *1 - isolated dwelling ,* *2 - twin dwelling,* *3 = 4 – old dwelling,* *5 - rural house,* *6 - typical dewelling-,*

Descriptive physical attributes

By applying some automated text processing on the *free-text* description of each advertised property registry, we could extract information about the existence of some important housing facilities in that property. As a result, 13 binary (dummy) variables were added to the dataset. Table 2 shows a list with the attribute names. If the attribute value is set to 1 then the particular equipment is in the property; 0, otherwise.

Table 2. List of Amenities represented in the descriptive physical attributes.

DOUBLE FLOOR	PLACE OF GARAGEM	REMODELED
JUNK ROOM	GARAGE	FIREPLACE
BALCONY	FLOOR	HYDROMASSAGE
ATTIC	WHIRLPOOL	
TERRACE	WC	

Spatial attributes

The spatial data processing has been subjected to various approaches (e.g. [10]). Among them, the simplest way is based on the construction of new variables that are used to reflect the distance of a given property to the central business district[2] (CBD), as well as distances to the set of amenities offered in its neighborhood. In our study we defined three territorial urban centers. As shown in the map in Figure 2, two of them are the hypothetical centers of Aveiro and Ílhavo. The third urban center is located in beach areas due to its relevance to the population from this region.

Following this approach, we added 36 new spatial variables of two types: *location attributes* and *neighborhood attributes.* The three new *location attributes* are the dummy (binary) variables: MACROZONA_CENTRO_AVEIRO (Aveiro CBD - true: 5368; false:14601); MACROZONA_CENTRO_ÍLHAVO (Ihavo CBD - true: 1880; false: 18089) and MACROZONA_PRAIAS (beaches - true: 1394; false: 18575). Because the property location provided in the raw dataset was not geo-referenced, each property was geocoded and assigned coordinates of the centroid location of its associated micro-zone. This process was performed manually with the help of a geographic information system software (ArcGIS / ESRI).

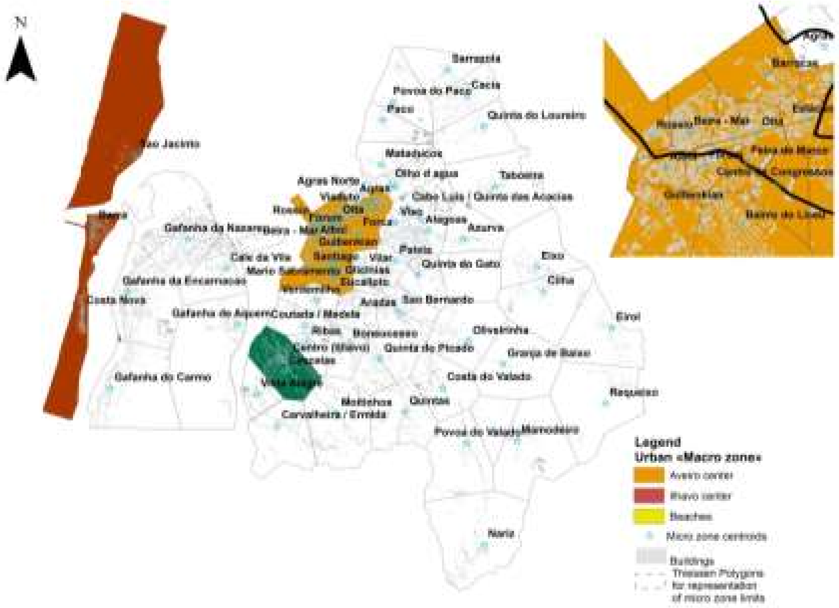

Fig. 2 Urban centers and micro-zones centroids for the Aveiro and Ílhavo municipalities.

[2] A CBD is a geographical area with highest density of amenities and services commonly used in the definition of location attributes in house price modeling.

To construct the *neighborhood attributes* we used the service provided by the website *Sapo Maps* [19] to extract information about the available urban amenities (equipments, services, place of interests, etc.) in the neighborhood areas of each advertised property. In spatial terms, the data available on *Sapo Maps* is disaggregated to the greatest possible level: a georeferenced point in the map. Then, with the collaboration of the *LabSapo* of the University of Aveiro [12], we obtained the coordinates of a set of amenities of the municipalities of Aveiro and Ílhavo. As a result, we obtained a total of 975 points classified into 11 pre-defined categories. Figure 3 depicts the distribution of the amenities by category.

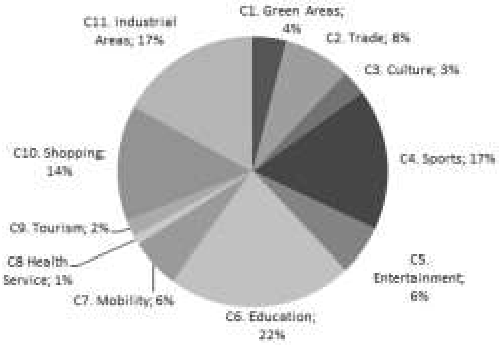

Fig. 3. The distribution of the amenities by category.

The neighborhood of a property is then determined by the distances (in *m*) to amenities offered in its surroundings within a comfortable walking distance. We also classified the amenities into three types of relevance:

- **[T1] -TYPE 1:** Amenities that serve the population daily (or at least weekly) located within an easy walking distance (within a 600m radius from the micro-zone centroid). For instance, local food markets: grocery shops, cafes, restaurants, bakeries, butcher shops, etc.
- **[T2]-TYPE 2**: Amenities with a sporadic use aimed at serving a small number of residents and situated within a 1200m radius. In this case, resident is willing to walk a little *bit* to reach this feature.
- **[T3]-TYPE 3**: Amenities exceptionally used by residents and located within an 1800m radius. They are serving a supra-local population (people outside Aveiro and Ílhavo). For example, the football stadium, the hospital, the tribunal, etc.

Figure 4 represents the spatial distribution of amenities of the Aveiro and Ílhavo municipalities obtained with the database service *Sapo Maps* indicating their relevance type. As a result, 33 new neighborhood attributes - three dummy variables for each of the 11 pre-defined categories were added to the dataset. For example, for the category CULTURE tree attributes were included: POT_CULTURE_T1, POT_CULTURE_T2 and POT_CULTURE_T3. Each neighborhood variable stores the distance (a real value) between the centroid of the micro-zone where the property is located and the amenities of a given category with a particular relevance type. The distance is estimated using the potential function described in [13]. This function is inspired by the concept of classical physics potential and adapted to the problem under study. Let us consider a set of amenities **A** of a given category *c* that belong to a circle with a radius set accordingly to the relevance type *t*. The potential of the micro-zone centroid *z* given the set **A** is defined as:

$$P(z, \mathbf{A}) = \frac{|\mathbf{A}|}{\min\limits_{a_i \in \mathbf{A}} \{d(z, a_i)\}}$$

where $d(z, a_i)$ represents the Euclidian distance between the centroid z and the amenity a_i. In other words, the potential is calculated as the ratio between the number of points in **A** and the distance between the micro-zone centroid and its nearest point in **A.**

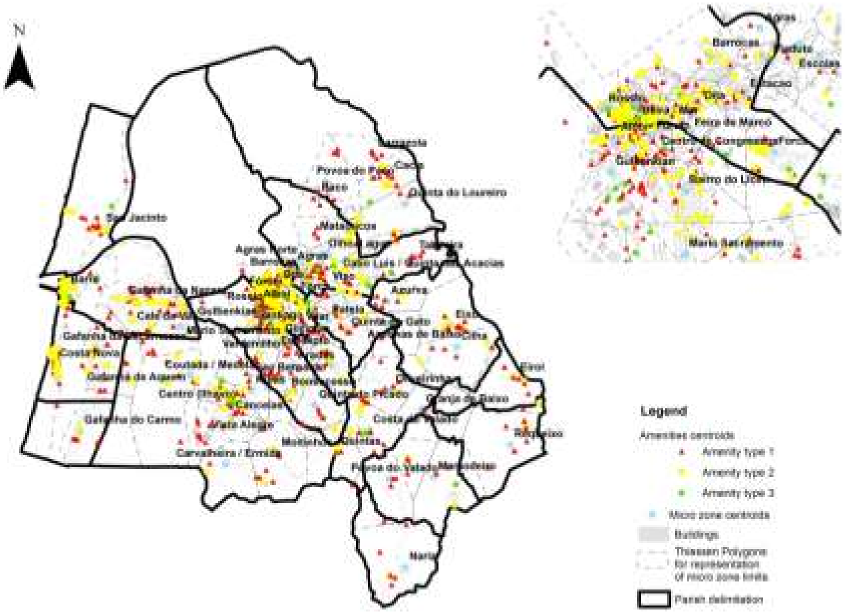

Fig. 4 Amenities provided by the portal *Sapo Maps* and their distribution according the relevance types T1, T2 and T3.

2.3.2 Selection of Relevant Attributes

Among several feature selection strategies [1] implemented in RapidMiner® 5.0, we chose those ones that led to the most interesting results in terms of the identification of relevant attributes for house price modeling, while considering the computational efforts made. In this work we compare eight different methodologies based on four approaches to feature selection: *embedded methods*, *principal component analysis* (PCA), *filters* and *weighting schemes*. Whereas the three last approaches are implemented during a separate process in the pre-processing phase before the induction algorithm run, in embedded methods [11] the feature selection and induction process cannot be separated. In this case the structure of the models under consideration (in our case, a multiple linear regression model) plays a key role. The feature selection methods we compared are:

- <u>Embedded methods</u>: we compared two selection methods embedded in the linear regression learning algorithm:
  - **M1.** *M5prime (M5') algorithm* [20]: an improvement of the Quinlan's model-tree inducer M5. Model trees combine a decision tree with linear regression functions at the leaves. M5' added a pruning process which allows a significant reduction of the tree size based on a small penalty in prediction performance.
  - **M2.** *forward greedy algorithm*: it starts with an empty set; then attributes are iteratively added according to a given quality measure (the AIC, in this case) until a stopping criterion is met.
- <u>Principal Component Analysis</u>: it builds higher-order attributes (*principal components*) from the original ones in terms of the variance they explain and selects those that capture the maximum amount of variation in the data. The new attributes are linear combinations of the original ones and orthogonal to each other. This method is unsupervised because it does not take into account the response variable. In this study the attribute selection is implemented in two ways:
  - **M3.** *new attributes are selected*: the principal components that explain at least 75% of the variance of the initial data are selected.
  - **M4.** *original attributes are selected*: the variables with the greatest *loading* (and value greater than 0.5) in each of the components with at least 75% of the cumulative variance are selected.
- <u>Filter Methods</u>: we implemented two approaches to feature subset selection (FSS) using two greedy hill climbing heuristics [3] for searching in combination with the *correlation-based feature selection* (CFS) measure [9] to evaluate the quality of each visited subset. According CFS, good feature subsets contains uncorrelated features but higly correlated with the target attribute. Two different search strategies to find an optimal or near optimal subset of features are compared:
  - **M5.** *backward elimination*: the search begins with the full set of attributes; then attributes are iteratively removed according to the CFS measure.
  - **M6.** *sequential forward selection*: the search begins with an empty set of features; then attributes are iteratively added according to the CFS measure.
- <u>Weighting Schemes</u>: for each attribute a weight that reflects its relevance is calculated; then attributes are ranked by their weights and a filter is applied in order to retain for further analysis attributes with weights greater than a given threshold (in our study we used 0.4). We applied two different methods to calculate the weights:

M7. *PCA weighting:* it uses the *loadings* of the first principal component as weights. This scheme is unsupervised (it does not consider the response variable.

M8. *SVM weighting*: it uses the coefficients of a linear SVM as weights. This weighting scheme has been very successfully applied in Bioinformatics [8] using microarray data.

After the application of feature selection, the resulting eight datasets were used as input to the modeling and evaluation phase of the KDD process.

2.4 Modeling

The hedonic method is a widely used technique to control the heterogeneous nature of house attributes when defining house price models [2]. Houses are seen as composite products and although their various characteristics are not sold separately, regressing the sale price of them yields the marginal contribution of each characteristic. *Multiple linear regression* (MLR) is pointed out as one of the most popular hedonic pricing models. Generally, hedonic pricing models are defined as

$$p(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon \qquad (1)$$

where $p(\mathbf{X})$ is the price of a house **X** characterized by its attributes $X_1, X_2, \ldots, X_n$; $\beta_1, \beta_2, \ldots, \beta_n$ are the estimated *regression coefficients*; β_0 is the *intercept parameter;* and ε is the *residual error component*.

In our study all the physical, descriptive and spatial attributes above described were considered for the inclusion in the hedonic price regression. In addition, the dependent variable and the independent numeric ones (*area* and *neighborhood attributes*) were logarithmized in order to get a better fit to the data (for economic justification see, for example, [15]).

2.5 Evaluation

There are two crucial aspects when evaluating a MLR model for hedonic price modeling: its *explanatory power* and its *predictive capability*. In this work we focus on the searching for attributes that have the greatest effect on the formation of housing prices. Therefore, in this context we are more interested in assessing the explanatory power. To this end, we use the *coefficient of determination,* R^2 - a statistical measure of the goodness of fit to data that shows how well the regression line approximates the real data. R^2 also gives us the level of the explained variability in the model [7]. In order to obtain the R^2 estimates for each induced model, we applied the simpler hold-out validation scheme (2/3 data was used as training set and 1/3 as testing set).

All the data mining processes were implemented in RapidMiner 5.0. This open-source data mining tool combines the power of development environments with the simplicity of visual programming. KDD processes in RapidMiner 5.0 are implemented through the use of more than 500 nestable operators (for input/output, pre-processing, modeling, evaluation, etc.). As depicted in Figure 8, all the operators are combined into a process graph (design) by means of a powerful but intuitive user interface. The operator *Linear Regression* was used to induce the regression models. This operator implements the well-known *least squares algorithm.*

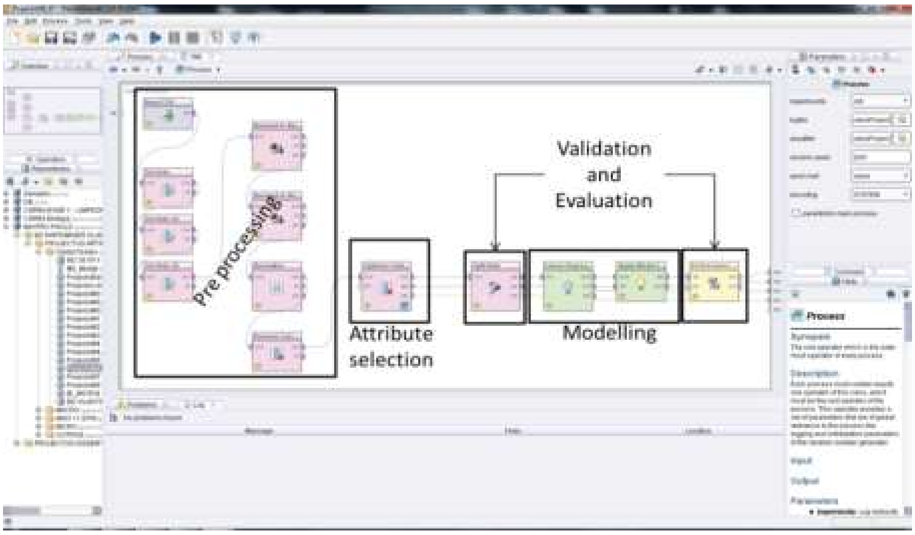

Fig. 5 One of the RapidMiner process graph implemented in our study

2.6 Deployment

The easiest way to construct a house price index is to use a summary measure, such as the mean or median price for a period [2]. The *Casa Sapo* portal currently estimates the price of a property using the average price for houses in the same location (municipality, parish and zone) and for the same type value and preservation level. At this moment, we are evaluating several house price indexes induced from real data gathered through the portal *Casa Sapo* with the aim to implement a hedonic price model and use it as the house price index in the portal.

3 Empirical Results and Analysis

Primarily, we are interested in evaluating how the selection of relevant attributes can affect the explanatory power of hedonic price models. Table 3 resumes the results after applying the eight feature selection methods explained in Section 2.3.2. The model labeled M0 is the baseline model when no feature selection was performed. The two last lines in this table depict the percentage of reduction in the number of attributes and the R² estimates, respectively. The best trade-off between the reduction in the number of attributes required to explain the house price and the dropping in explanatory power in relation to the baseline model is achieved by those MLRs models resulting after the application of a filter feature subset selection (FSS) method. The two MLR models obtained with FSS (M_5 and M_6) achieve a reduction of 83.5% in the number of attributes (10/59 attributes were selected). Nevertheless, the model M_6, obtained after applying the sequential forward search strategy for the greedy algorithm achieves a better trade–off resulting in an 11.18% drop in the explanatory power (R^2 drops from 0.635 to 0.546) against the 14.16% reduction obtained by M_5. The results indicate that FSS using greedy algorithms allows the induction of simpler and more interpretable hedonic price models without a great loss in their explanatory power. As observed, in the baseline models there are only 33 attributes that are significant in the formation of the price. On opposite, the models obtained after the application of FSS not only have the advantage of drastically reduce the number of attributes, but also all of the attributes selected are significant (with p-value=0).

Table 3 The hedonic pricing modeling results (columns with best values are shaded)

	Without Attribute Selection	**COM SELECÇÃO**							
		Embedded Approaches		**PCA**		**Filters**		**Weighting schemes**	
		M5 Prime	**Greedy [forward + AIC]**	**Using PCs**	**Using original attributes**	**FSS [backw. + CFS]**	**FSS [forward + CFS]**	**PCA**	**SVM**
	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
Execution Time	35"	22"	01'55	05"	06"	01:28'04"	46"	12"	14'39"
# Selected Attributes	59	56	46	8	5	10	10	20	4
# Significant Attributes[3]	33	31	43	8	4	10	10	15	4
R^2	0.635	0.635	0.635	0.420	0.449	0.546	0.564	0.319	0.477
Reduction in the # of attrib **(in %)**.		5.08	22.03	86.44	91.53	83.05	83.05	66.10	93.22
Reduction in explan. power **(in %)**		0.00	0.00	33.86	29.29	14.06	11.18	49.76	24.88

Analysis of Selected Attributes

Table 4 shows the coefficients of the most often selected attributes (a total of 6) in the resulting hedonic price models[4]. Among the 59 initial attributes, the most

[3] An attribute X_i is significant if the p-value for the t-test: H_0: $\beta_i = 0$ vs H_1: $\beta_i \neq 0$ is less than α=0.05 (5% significance level).

relevant one is the AREA (it appears in 7/8 models). Moreover, for all the defined models including this attribute, their coefficients show that it has a great effect on the formation of housing prices. From these results, we can also conclude that in the municipality of Aveiro and Ílhavo, the costs per square meter tend to go down as house size increases.

Table 4 Regression coefficients of the top 6 most often selected attributes

	Without Attrib. Selection	COM SELECÇÃO							
		Embedded Approaches		PCA		Filters		Weighting schemes	
		M5 Prime	Greedy [forward + AIC]	using PCs	using original attributes	FSS [backw. + CFS]	FSS [forward + CFS]	PCA	SVM
	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
R^2	0.635	0.635	0.635	0.420	0.449	0.564	0.564	0.319	0.477
ln(AREA)	-0.190	-0.190	-0.189		-0.148	-0.160	-0.153		-0.145
MACROZONA PRAIAS COD	0.390	0.391	0.402			0.216	0.365		0.457
ln(POT_AMENIDADE_DESPORTO_T2))	0.101	0.101	0.100					0.086	0.037
ln(POT_AMENIDADE_UTILIDADES_T1))	0.030	0.030	0.031		0.027			0.080	
MACROZONA CENTRO AVR COD	-0.386	-0.386	-0.390				0.150		0.113
ATRB_12_LAREIRA	-0.020	-0.020	-0.020			-0.013	-0.013		

At a second level of relevance we have the location variable Macrozona_Praias (beaches). Its importance is justified not only because appears 5/8 times but also because its regression coefficients present high values. A house located close to the beaches is more expensive than others located otherwise.

At third level of relevance we have four attributes: two neighborhood attributes that indicate the distance from a house to sport facilities and daily utilities services; one location attribute and one attribute associated with physical housing characteristics. Regarding the magnitude of its coefficients we can highlight the location attribute Macrozona_Centro_Avr that defines if a house is or not located in the center of Aveiro. It was expected that this coefficient would be positive in all models, meaning that the value of the property would increase when located in the

[4] The model M_3 is excluded of this analysis because it was induced after applying the PCA using the selected principal components instead of the original attributes.

center of Aveiro. However in models that use a large number of variables, namely the model 0, 1, and 2, the coefficient is negative. One possible explanation for this fact is that because we have so many variables, some of them, related with CBD, the centrality dimension is captured by other attributes included in the model; and this negative effect can be derived from other particularities (such as older and smaller houses, very typical of the center of the city) that are inversely correlated with the price of a house. Finally, a house that has a fireplace (Atrb_12_Lareira) presents a small negative impact suggesting that the addition of this facility can lead to a devaluation of its price. In fact, the fireplace is an amenity that was very relevant in the Aveiro-Ilhavo house market in recent past years. Hence, it can be associated with those older houses put in the market, but with less market value.

4 Conclusions and Future Work

The study described in this paper was carried out using real data gathered by the portal *Casa Sapo* for the municipalities of Aveiro and Ílhavo. After applying some pre-processing tasks in order to get more valuable information for feature selection and modeling, we obtained a dataset with 58 attributes mainly grouped into two types: *physical* attributes and *location* attributes. The main aim was the evaluation of eight attribute selection methods to determine the hedonic pricing model with the best trade-off between the number of selected attributes and its explanatory power. To conduct the study, the open-source data mining software RapidMiner® was employed. Results show that models induced after applying FSS filters using two different greedy search strategies include only a reduced number of attributes while their explanatory power are not substantially affected when compared with the baseline model. Results also indicate the importance of the location attributes. With the exception of the house area (the most relevant attribute), most of the attributes selected are related to the house location and its distance to urban centers, beaches and different amenities offered in its neighborhood.

On the other hand, the estimates of the R^2 obtained for all induced models could still be improved. MLR is not a very good way of representation for hedonic price modeling because it imposes a linear relationship on the data. Instead, we can use more sophisticated techniques: non-linear regression, support vector machines, neural networks, regression trees, etc. However, during our experiments we have tested some of these methods and found that none of them achieves better results that the simple MLR. For this reason, we believe that we could improve the explanatory power of induced models if we apply a stringent data cleansing process. With this aim, we are now working in the implementation and evaluation of other cleaning and pre-processing tasks that allow us to obtain a more valuable dataset for KDD processes.

Acknowledgments. The data used in this study were provided by Janela Digital S.A.

References

1. Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. J. Artificial Intelligence 97 Issue 1-2 (1997)
2. Bourassa*, S. C., Hoesli*, M., and Sun, J: *A* Simple Alternative *House Price Index Method*. Journal of Housing Economics, 15(1): 80-97 (2004)
3. Caruama, R. Freitag, D.: Greedy Attribute Selection, in Proceedings of the Eleventh International Conference on Machine Learning, p 28-36 (1994)
4. Casa Sapo Web Site, http://casa.sapo.pt/
5. CRISP-DM Web Site, http://www.crisp-dm.org
6. Feyad, U.: Data mining and knowledge discovery: making sense out of data. *IEEE Expert*, Vol 11 N.º 5, 20-2. (1996)
7. Freund, R.J., Wilson W.J., and Sa P.: Regression Analysis: Statistical Modeling of a Response Variable. 2nd ed. ed. Elsevier (2006)
8. Guyon I., Weston J., Barnhill S. and Vapnik V.: Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning, Vol. 46 No. 1, 389-422 (2002)
9. Hall, M. : Feature subset selection: a correlation based filter approach. In: International Conference on Neural Information Processing and Intelligent Information Systems, Springer Verlag, p.855-858 (1997)
10. Kiel, K.A., Zabel, J.E.: Location, location, location: The 3L Approach to house price determination. Journal of Housing Economics. Vol. 17, N.º 2.. p. 175-190 (2008)
11. Lal T. N., Chapelle O., Weston J., and Elisseeff A.: Embedded methods, in Feature Extraction, Foundations and Applications, Springer-Verlag (2006)
12. LabSapo, University of Aveiro http://labs.sapo.pt/ua/
13. Lopes António: Desenvolvimento regional: problemática, teoria, modelos. Fundação Calouste Gulbenkian. 406p. (2001)
14. Maclennan, D. Some thoughts on the nature and purpose of hedonic price functions. *Urban Studies* 14, 59-71 (1977)
15. Malpezzi, S.; Housing Economics and Public Policy. Housing Economics and Public Policy, O'Sullivan, T.; Gibb, K.; Blackwell Science, 327p (2008).
16. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M. and Euler, T.: Rapid Prototyping for Complex Data Mining Tasks, *Proceedings of the 12th ACM SIGKDD – International Conference on Knowledge Discovery and Data Mining,* (2006)
17. Palmquist, R.B: Property value models. Chapter 16, In: Maler, K. and Vincent, J. (Eds.) *Handbook of Environmental Economics Vol. 2*, North-Holland: Amsterdam, 763-813 (2005)
18. Rosen, S.: Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, Vol. 82 Nº 1, 34-55. (1974)
19. Sapo Mapas, http://mapas.sapo.pt/
20. Wang,Y., Witten, I.: Induction of model trees for predicting continuous classes. Working paper 96/23. University of Waikato, Department of Computer Science. 1996.