CIS-445 Knowledge Discovery in Databases

Project #4

Due: See Blackboard

Individual or Group Project (maximum two students in a group)

Worth 100 points

Objective: Learn Data Mining/Analysis Tool: SAS Enterprise Miner 13.1/14.1

Task: Predicting the value of a continuous variable, namely, the sale price of residential real estate properties
Nodes: File Import, StatGraph or StatExplore, Variable Selection, Interactive Binning, Transform Variables, Filter, Data Partition, Neural Network, Regression, MBR (Memory-Based Reasoning), Model Comparison
Software: SAS EM 13.1/14.1

This is an individual or group project (maximum two students in a group). Each member of the group has to make meaningful contribution to the project. In the project report you must briefly elaborate on the contribution of each member and assign percentage contribution to each member.

In this project you will build and test several models to predict the sale prices of residential real estate properties. The data is stored in an Excel file named *RealEstate.xlsx* and contains the features of low-end residential properties in two neighborhoods of a medium size city in the Mid-West of the U.S. The data set contains 363 samples and 17 attributes: 16 input attributes and 1 output (target attribute). The attributes and the values they take are summarized in Table 1.

Table 1. Description of the variables

	Name of variable	Level	Description
1	Neighborhood	Nominal	Location of the property. Values: 1 – Neighborhood-1; 2 – Neighborhood-2
2	Age	Interval	Age of the property in [Years] computed as Year Sold minus Year Built
3	LotSize	Ordinal/ Nominal	Size of the lot. Values: 1 – ≤0.25 acre (small); 2 – (0.25,0.5] acre (Medium)
4	ConstructionType	Nominal	Type of the construction. Values: 1 – 1 story; 2 – 1.5 story; 3 – 2 story
5	WallType	Nominal	Type of the walls. Values: 1 – Frame; 2 – Brick; 3 – Other
6	Basement	Interval	Size of the basement in [Square foot]
7	FirstFloor	Interval	Size of the first floor [Square foot]
8	SecondFloor	Interval	Size of the second floor [Square foot]
9	UpperArea	Interval	Size of the upper area in [Square foot]
10	TotalArea	Interval	Total area in [Square foot]
11	Baths	Interval	Number of bathrooms. Values: 0 – substandard; 1 – 1 bath, 2 – 1½ baths; 3 – 2 baths; 4 – 2½ baths; 5 – 3 baths; 6 – more than 3 baths
12	CentralAir	Binary	Central air conditioning. Values: 0 – None; 1 – Present
13	FirePlace	Interval	Number of fireplaces
14	BasementType	Nominal/ Ordinal	Type of the basement. Values: 0 – None; 1 – Partial; 2 – Full or Complete
15	GarageType	Nominal	Type of the garage. Values: 0 – None; 1 – Carport; 2 – Detached; 3 – Attached; 4 – Garage in basement; 5 – Built in garage
16	GarageSize	Interval	Number of cars in the garage. Values: 0 – No garage; 1 – 1-car garage; 2 – 2-car garage, etc.
17	SalePrice	Interval	[$], Target/Output variable.

When you open the Excel file and examine the range of values that the variables take, you will notice that there are a number of outliers and obvious errors resulting from typos while the numbers were being

entered into the database. I marked them in red and bold in the Excel cells. There are probably other outliers or data entry errors as well. Do not just erase the rows in Excel to get read of the outliers, but use SAS EM to remove them as suggested below.

As all variables contain numerical values, SAS EM will assign the Interval Level to all of them when you import the data set. Also, SAS EM will assign the Role Input to all of them. Table 1 clearly shows that several variables must be coded as the Ordinal or Nominal Level as there is no distance clearly defined between the values/levels that these variables take. Also, do not forget to define SalePrice as the Target variable. If needed, perform the changes to the measurement level of each variable according to the specifications shown in Table 1.

Examine the distribution of all variables for any unreasonable or low frequency values as well as outliers. In particular, look at variables Age, LotSize, GarageType, and SalePrice. Some values of the input variables may require binning (assigning them to an adjacent, more frequent group), while others may require variable transformation (such as logarithm). Some variables such as SalePrice have unusually low values and others such as Age have negative values. From a few variables, including the target variable, the outliers will have to be removed by using the Filter node. The call on how to perform this data cleaning and transformation of the variables is entirely yours. Also, it appears that variable TotalArea is likely to be a linear combination of variables FirstFloor + SecondFloor + UpperArea. Do you need to use all four variables in the model, or just two or three of them? You may run the Variable Selection node or the Person correlation analysis to find this out. The input variables should be highly correlated with the target variable and weakly correlated with each other, especially for linear regression. Perhaps the Filter node will take care of the most of these data preparation issues, including removing outliers? Removing outliers will obviously lead to the smaller data set which will have less than 363 samples for building and testing the models. In the Filter node go to the Properties and click on **…** for Class Variables and click on **…** for Interval variables to explore the options for filtering.

In the class examples we have evaluated the performance of predictive models, which predict the value of a continuous variable (here SalePrice), by several simple measures. These are the mean absolute error (MAE), the root mean squared error (RMSE), coefficient of correlation (R), and the coefficient of determination (also called the coefficient of fitness), R^2. When you browse the Results from the model nodes you will see that SAS EM 13.1/14.1 allows one to evaluate the quality of the models by a myriad of other measures as well. These measure are, for example, Akaike Information Criterion (AIC) and/or Schwarz Bayesian Criterion (SBC). While the discussion of these two criteria is far beyond the scope of this course, smaller values for AIC and SBC are preferred. The best models can actually be selected based on the smallest value of AIC or SBC. Look at the defaults in the model selection criterion in the Properties panel for the models.

Import the Excel file. Use one or two data exploration nodes to examine the distribution of all variables to identify outliers, find out how to transform variables, and group variables into appropriate bins, if needed. You may use the Variable Selection node to find out the predictive power of each input variable. Variables which do not have sufficient predictive power will be removed from further analysis. Partition the data appropriately for the training set and the validation set only. Do not use the test set as the data set is too small. Use a neural network (NN), MBR with 10 neighbors (see Appendix), and Linear Regression as predictive models. You may create two paths: one with variable transformation and selection as well as filtering outliers, and the other without. What are the most significant variables determined by the Regression node? Look at the effects and *p* values. Are there any missing values in the data set? Do you need to use the Impute node and/or Replacement node? Compare the results generated by SAS on the validation data set using RMSE and MAE of the models. [The models in SAS EM generate the maximum absolute error (MAE), not the mean absolute error.] You may also look at R^2 for the linear regression model (the NN and MBR models do not generate the R^2 coefficient). Which of the models would you choose to predict the sale prices? Do the transformation of variables and removing outliers help?

You should **not** expect to obtain "great" results in terms of low RMSE, MAE and high R^2 (close to 1) as the older houses in low-end neighborhoods are very hard to appraise accurately due to their lack of homogeneity. There may be also other variables not collected and not included in the models which have more predictive power than those used. In future projects you may want to use the Cluster node to group similar properties together and then build models on each cluster. Actually, real estate market clustering (segmentation) is the key to creating the models which yield better prediction results (smaller errors).

Use the Help menu to learn more about nodes and the models' evaluation criteria to the extent you can understand.

Write a 2 page report to describe the project and provide the summary of your results. State the contribution of each member. Create a table with the values for the measures you have used to compare the models. Use font 12, Times Roman, and 1" top, bottom, right, and left margins. Merge your report, the final workflow diagram, and any screenshots with the results that you find relevant into a single file, save it as *Project4_LastName1_LastName2.pdf*, and e-mail it to jozef.zurada@louisville.edu on the due date.

Appendix on MBR

Memory- or case-based reasoning (MBR) is a type of case-based reasoning. Broadly construed, it is the process of solving new problems based on the solutions of similar past cases. The MBR method requires no model to be fitted, or function to be estimated. Instead it requires all cases with their known solutions to be maintained in memory, and when a prediction is required, the method recalls items from memory and predicts the value of the dependent/target variable. In solving a new case, the MBR approach retrieves a case (or cases) it deems sufficiently similar and uses that case (or cases) as a basis for solving the new case.

MBR employs a *k*-nearest neighbor algorithm to predict cases. The *k*-nearest neighbor algorithm takes a data set of existing cases and a new case whose target value is to be predicted, where each existing case in the data set is composed of a set of variables and the new case has one value for each variable. The normalized Euclidean distance between each existing case is computed. The *k* existing cases that have the smallest distances to the new case are the *k*-nearest neighbors to that case. The average or the median over the *k* nearest cases is computed and it yields the predicted value. The example below outlines the process and the numbers are made up.

Table2. *k*=10 nearest neighbors (sorted by case number).

Case id	Actual Sale Price in [$]	Case ranking based on the normalized distance to the instance whose value is predicted: (1 – closest, 10 – farthest)
6	67000	3
15	72500	8
35	71300	1
68	69000	9
123	75000	10
178	79000	7
190	65000	2
205	77800	4
255	63000	6
269	60900	5

Table 3. Examples of average sale price determination.

k	Case id of the nearest neighbor(s)	Target value of nearest neighbor(s)	The mean predicted value (sale price) of the instance
3	35, 190, 6	71300, 65000, 67000	(71300+65000+ 67000)/3=**67767**
5	35, 190, 6, 205, 269	71300, 65000, 67000, 77800, 60900	(71300+65000+ 67000+77800+ 60900)/5=**68400**
10	35, 190, 6, 205, 269, 255, 178, 15, 68, 123	71300, 65000, 67000, 77800, 60900, 63000, 79000, 72500, 69000, 75000	(71300+65000+ 67000+77800+ 60900+63000+79000+ 72500+69000+75000)/10=**70050**

One can easily notice that there are two critical choices in the nearest neighbor method, namely, the distance function and the cardinality *k* of the neighborhood. In this project, SAS EM will use the normalized Euclidean distance for numeric variables and the Hamming distance for categorical variables to calculate the similarity between cases in 16-dimensional space (16 input variables). Normalization is required to ensure that features with larger values do not overweight features with lower values.