HOW TO MODEL? DATA MINING CONCEPT AND TOOLS

K. Haralampiev

Department of Sociology, Sofia University, Sofia 1113, Tsarigradsko Shose blvd. 125, bl. 4, tel. +359885046256, e-mail: k_haralampiev@hotmail.com

> *"Historical data does become critical in decision-making when it is combined with current data and continuously refreshed with new information. This approach is the basis of predictive decision-making; models are based on what has happened and why, and also on what is happening now and what will likely happen next."*
> *[IBM predictive analytics]*

Abstract: In the paper we present the data mining concept about modelling as well as it is realized in IBM SPSS Modeler. First, we describe the four common research tasks which need modelling. Second, we introduce the IBM SPSS Modeler algorithm for automatic modelling which include automatic model selection and automatic predictors' selection. And third, we examine three examples of application of IBM SPSS Modeler for investigation of utility company consumers' behaviour.

Key words: modelling, data mining, automatic model selection, automatic predictors' selection, predictor importance, utility company consumers' behaviour

INTRODUCTION

The "traditional" modelling approach follows the next few steps: first, we choose the appropriate model, then we choose the appropriate independent variables (also called factors or predictors) and finally we estimate the parameters of the chosen model. This approach often requires a mass of preliminary qualitative analyses and very good statistical skills. However, what if we put all models and all possible predictors together? And then rely on some software to choose which model is most appropriate and which predictors really influence on the dependent variable (also called target). Generally this is what we call "data mining". This approach is more convenient for managers who usually have only a slight idea for the possible predictors and almost never have an idea about the possible statistical models.

The goal of this paper is to show how we could prepare data mining for different purposes using statistical tool IBM SPSS Modeler.

METHODOLOGY AND RESEARCH DESIGN

Generally our methodology is based on a CRISP-DM research methodology using IBM SPSS Modeller predictive models and tools – fig. 1. [4, 5, 6].

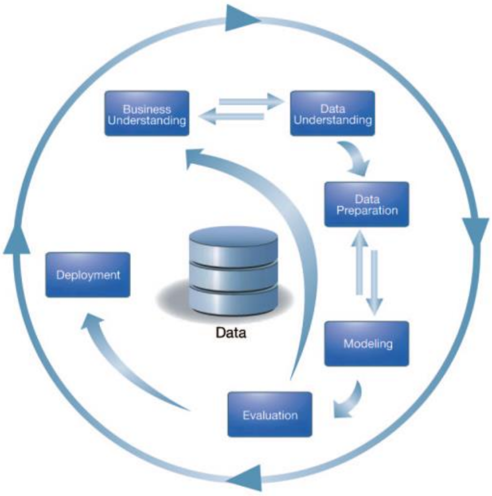


Fig. 1. CRISP – DM methodology

In this paper we will pay attention only to the step called "modelling". When modelling, we could solve different research tasks, broadly divided into four groups:
- Classification – we want to distribute all cases into preliminary defined groups. Usually these groups are based on some qualitative (nominal or ordinal) variable and as predictors we could use mix of qualitative and quantitative (scale) variables. This research task is typical for decision making under risk.
- Segmentation – we want to distribute all cases into "natural" groups which are not preliminary defined and we expect our analysis to reveal them. Usually as predictors we could use only scale variables. This research task is typical for market segmentation and targeting.
- Association – we want to determine whether there is a relationship between two or more variables. The dependent variable could be both qualitative and quantitative. The

predictors also could be both qualitative and quantitative variables. This research task is typical for the science where we want to explain the hidden patterns of origin and progress of a specific phenomenon.
- Time series analysis – we want to trace out the dynamics of a specific phenomenon. This research task is typical for stock traders for instance.

However, for all research tasks we follow the same algorithm using IBM SPSS Modeler:

1. We choose one of the Auto nodes. There are four Auto nodes:
- Auto Classifier Node which could be used both for classification and association;
- Auto Numeric Node which also could be used both for classification and association;
- Auto Cluster Node which could be used for segmentation;
- Time Series Node which could be used for time series analysis.

2. We choose all possible predictors.

3. We choose all possible models:
- Supported model types in Auto Classifier Node include C5, Logistic Regression, Bayesian Networks, Discriminant Analysis, KNN Algorithm, SVM, C&R Tree, QUEST, CHAID and Neural Net.
- Supported model types in Auto Numeric Node include Regression, Generalized Linear, KNN Algorithm, SVM, C&R Tree, CHAID, Neural Net and Linear.
- Supported model types in Auto Cluster Node include Kohonen, K-Means and TwoStep.

4. The software evaluates all models and chooses the best of them. The criteria about the best model(s) are different for the different research tasks:
- For the classification the criterion is the probability for correct classification;
- For the segmentation the criterion is the measure of cohesion and separation;
- For the association the criterion is goodness-of-fit measure;
- For the time series analysis the criterion is also goodness-of-fit measure.

5. For the chosen model(s) software automatically removes all predictors whose influence is not significant and then ranks remaining predictors by their importance.

6. Software adds new variables in the database:
- Auto Classifier Node adds predicted group membership and corresponding probability for correct classification;
- Auto Numeric Node adds the predicted values obtained through the model;
- Auto Cluster Node adds cluster membership;
- Time Series Node adds predicted values obtained through the model and confidence interval boundaries.
These new variables could be useful in next analyses.

SOME EXAMPLES

In our previous works we have presented application of IBM SPSS Modeler with different databases and for different purposes [1,2,7,8,9].

Now we will illustrate one Auto Classifier Node, one Auto Numeric Node and one Auto Cluster Node.
Our sample is drawn from a consumption database provided by the Operations department of "CEZ Electro Bulgaria" AD [3], combined with approximately 850 confirmed cases involving illegal electricity consumption and theft protocols for a period of 16 months from January 2011 to April 2012. The consumption is divided in three different tariffs – day, night and complete.

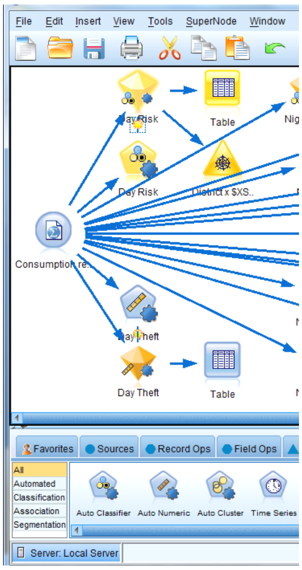


Fig. 2. Fragment of entire model including Auto Classifier Nodes (top) and Auto Numeric Nodes (bottom).

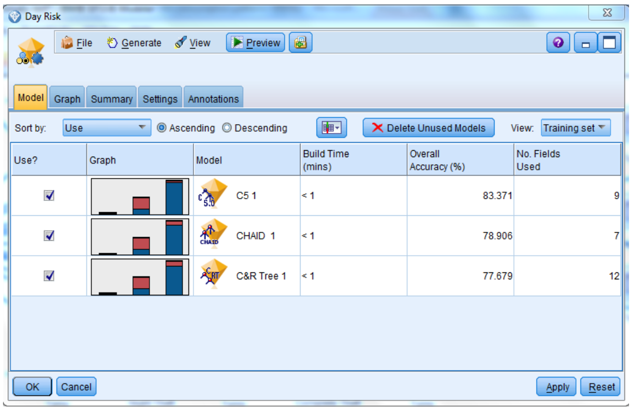
To illustrate the Auto Classify Node we apply the data for daily electricity consumption tariff. The target variable, which indicates the risk of electricity theft, has three categories coded as follows – Table 1.

Table 1. Target variables code

Code	***Target variable***
0	No thefts detected
1	Thefts detected
2	Not inspected

For research purposes we considered only the thefts in April 2012, which is the last month where inspection protocols exist. The predictors which we have used to build an early warning model are the residential area (Postal Code), district and monthly electricity consumption since January 2011 to March 2012.

As a result of the application of the Auto Classify Node we get three best models – C5, CHAID and C&R Tree – Fig. 3.



Sort by: Use ● Ascending ○ Descending View: Training set

Use?	Graph	Model	Build Time (mins)	Overall Accuracy (%)	No. Fields Used
☑		C5 1	< 1	83.371	9
☑		CHAID 1	< 1	78.906	7
☑		C&R Tree 1	< 1	77.679	12

Fig. 3. The three best models obtained by Auto Classify Node.

These three models are versions of the so-called Classification Trees, so we will examine in detail only the first (the best) tree.

Figures in column *Graph* shows that about two-thirds of consumers with detected thefts (red color) are classified correctly, and the majority of consumers who have not yet inspected (blue color) also are classified correctly. Of particular interest to us is that little part of not inspected users which is classified as customers with detected thefts. These are users who are identified as risky by the "early warning system". They could be recommended to CEZ for future inspection.

Since this Classification Tree is very large, we will show only a fragment of it – Fig. 4.

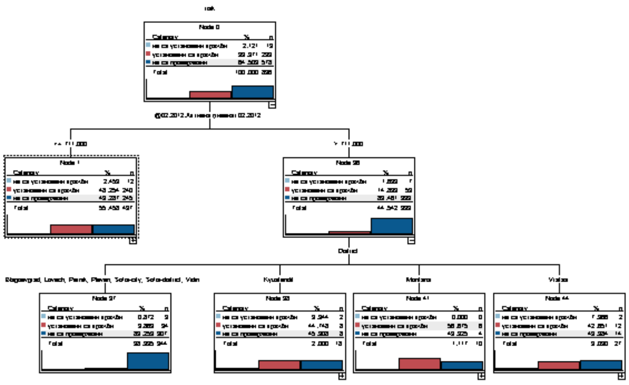


Fig. 4. C5 Classification Tree (fragment).

Figure 4 shows that the most important predictor is consumption in February 2012. There are two alternative explanations:
- For the "early warning system" it is very important what the consumption in February was;
- Since February is two months away from April (month during which the thefts are detected), for the "early warning system" is very important what the consumption was two months before the month in which we inspect for thefts.
Verification of these hypotheses is the subject of future work.

Thefts are detected for almost half of the examined consumers whose daily consumption during the month of February 2012 was under 711 KWh. For the users with a daily consumption of more than 711 KWh, the percentage of detected thefts is very low, but there are significant differences by districts. In the districts of Montana, Kyustendil and Vratsa, the theft rate is much higher than in other districts of the country.

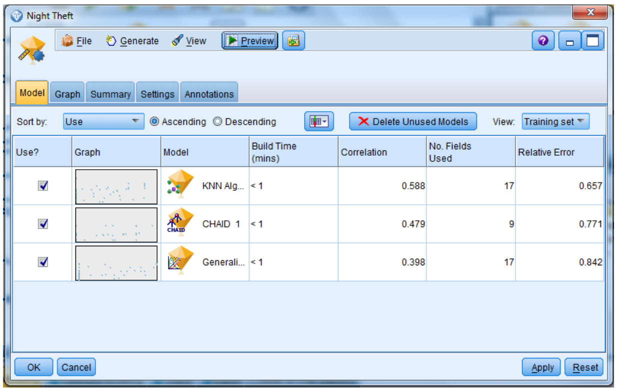
This analysis could be extended by unfolding of the tree and the successive study of all of its nodes.

Also, we could add Table Node to the model. It will allow us to identify risky customers.

To illustrate Auto Numeric Node we use the data for night consumption. The target variable, which indicates the risk of electricity theft, is the amount of stolen electricity.

The predictors, which we use to build an early warning model, which directs our attention to the potential risk users, are the same as in the Auto Classify Node.

As a result of the application of Auto Numeric Node we get three best models – KNN Algorithm, CHAID and Generalized Linear – Fig. 5.



Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
☑		KNN Alg...	< 1	0.588	17	0.657
☑		CHAID 1	< 1	0.479	9	0.771
☑		Generali...	< 1	0.398	17	0.842

Fig. 5. The three best models obtained by Auto Numeric Node.

Since in this case the three best models are independent, and they are not versions of one common model, here we only summarize the results instead of showing particular tables and graphs.

The most important predictors for potential risk users are the energy consumption in October 2011, followed by consumption in February 2012 and July 2011. Again, various explanations, similar to the previous case, are possible. The definition of these hypotheses and their verification is a subject of future work.

The night consumption in October 2011 and the amount of stolen electricity in April 2012 are positively correlated – the increase of the night consumption leads to increase of the stolen energy, i.e. the higher night consumption during this month is more risky. The night consumption in July 2011 and February 2012 and the stolen energy in April 2012 are negatively correlated – if the night consumption in these two months is bigger, then the amount of stolen electricity is less, i.e. the low night consumption in these two months is risky.

There is a common geographical pattern related to users from the districts of Montana, Kyustendil and Vratsa being most risky in the above context.

We use Auto Cluster node for identification and definition of end users consumption patterns.

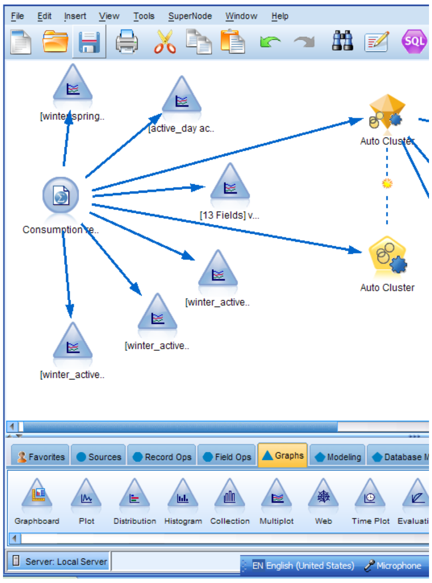
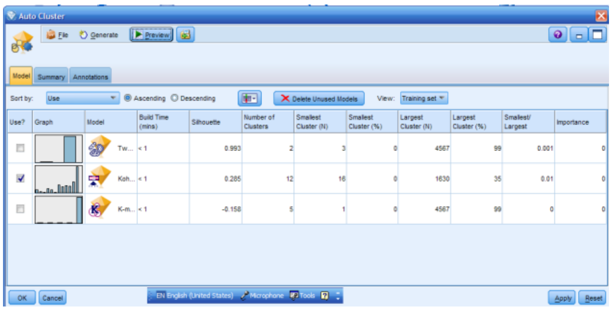


Fig. 6. Fragment of entire model including Auto Cluster Nodes (right).

As variables for Auto Cluster Node we have used consumption distinguished by seasons and total consumption.

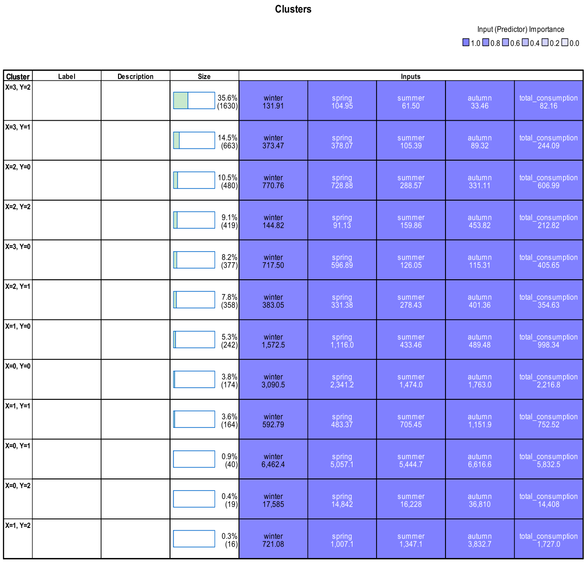
In the figure 7 are shown all three models ranked by Silhouette measure of cohesion and separation.



Use?	Graph	Model	Build Time (mins)	Silhouette	Number of Clusters	Smallest Cluster (N)	Smallest Cluster (%)	Largest Cluster (N)	Largest Cluster (%)	Smallest/ Largest	Importance
□		Tw...	< 1	0.993	2	3	0	4567	99	0.001	0
☑		Koh...	< 1	0.285	12	16	0	1630	35	0.01	0
□		K-m...	< 1	-0.158	5	1	0	4567	99	0	0

Fig. 7. The three best Auto cluster models

We choose to explain the second model – Kohonen – because the first and the third models are quite disproportional – there is one big cluster which includes over 99.5% of cases. In figure 8 are shown seasonal consumption patterns of all 12 clusters obtained by Kohonen method.



Clusters

Input (Predictor) Importance

■1.0 ■0.8 ■0.6 ■0.4 □0.2 □0.0

Cluster	Label	Description	Size	Inputs				
X=3, Y=2			35.6% (1630)	winter 131.91	spring 104.95	summer 61.50	autumn 33.46	total_consumption 82.16
X=3, Y=1			14.5% (663)	winter 373.47	spring 378.07	summer 105.39	autumn 89.32	total_consumption 244.09
X=2, Y=0			10.5% (480)	winter 770.76	spring 728.88	summer 288.57	autumn 331.11	total_consumption 606.99
X=2, Y=2			9.1% (419)	winter 144.82	spring 91.13	summer 159.86	autumn 453.82	total_consumption 212.82
X=3, Y=0			8.2% (377)	winter 717.50	spring 596.89	summer 126.05	autumn 115.31	total_consumption 405.65
X=2, Y=1			7.8% (358)	winter 383.05	spring 331.38	summer 278.43	autumn 401.36	total_consumption 354.63
X=1, Y=0			5.3% (242)	winter 1,572.5	spring 1,116.0	summer 433.46	autumn 489.48	total_consumption 998.34
X=0, Y=0			3.8% (174)	winter 3,090.5	spring 2,341.2	summer 1,474.0	autumn 1,763.0	total_consumption 2,216.8
X=1, Y=1			3.6% (164)	winter 592.79	spring 483.37	summer 705.45	autumn 1,151.9	total_consumption 752.52
X=0, Y=1			0.9% (40)	winter 6,462.4	spring 5,057.1	summer 5,444.7	autumn 6,616.6	total_consumption 5,832.5
X=0, Y=2			0.4% (19)	winter 17,585	spring 14,842	summer 16,228	autumn 36,810	total_consumption 14,408
X=1, Y=2			0.3% (16)	winter 721.08	spring 1,007.1	summer 1,347.1	autumn 3,832.7	total_consumption 1,727.0

Fig. 8. Kohonen Clusters' description

In the figure we can see the descriptive statistics of obtained clusters. In the fourth column there are clusters' sizes. In the next five columns there are mean consumptions of all clusters.

The biggest cluster (X=3, Y=2) includes 35.6% of cases, but its mean consumption are relatively small. On the other hand the cluster with largest consumption (X=0, Y=2) includes only 0.4% of cases – it is next to the last according to the clusters' size.

The next step after the clusters' description is to determine cluster membership of each case. It could be done by adding Table Node to the model.

CONCLUSION AND FUTURE WORK

In this research we have investigated how patterns of consumption in utility companies can be modeled and how to show customer's risk and theft behavior. The results include electricity consumption patterns based on a CRISP-DM research methodology using IBM SPSS Modeler predictive models and tools.

The generated models, crucial patterns recognition and results can be applied by utility companies, their partners and consultants for future energy efficiency processes innovation and business models transformation.

Our future work plans are:
- To integrate into the models other relevant predictors such as climate data, demographic data, specific social and economic parameters and to discover more inside patterns;
- To reengineer and improve the mechanism aimed to identify risky consumers and forecast prevention of thefts based on identified patterns;
- To innovate and create effective business processes of the utility company;
- To use business analytics for new innovative business model transformation.

ACKNOWLEDGEMENT

The authors are grateful to the Marketing & Sales support department of "CEZ Electro Bulgaria" AD for the explicit sample data base and helpful information.

The work on this paper has been sponsored by IBS Bulgaria, IBM Premier Business partner

REFERENCES

1. Bachvarov A., P. Ruskov, K. Haralampiev, Electricity end users' consumption patterns in different west Bulgarian locations. International scientific conference UNITECH'12, Gabrovo, 2012, vol. I, p. I-86-91.
2. Bachvarov, A., K. Haralampiev, P. Ruskov. Business Analytics for Household Electricity Consumptions. IV International scientific conference "E-governance", Sozopol, 2012, p. 103-110.
3. CEZ Electro Bulgaria AD: http://www.cez.bg/en/home.html, http://www.cez.bg/en/customer-service/ways-of-payment.html.
4. IBM Corp., CRISP-DM 1.0, Step-by-step data mining guide, © Copyright IBM Corporation 2010.
5. IBM Corp., Switching perspectives, creating new business models for a changing world of energy, 2010.
6. Morelli T., Shearer C., Buecker A., IBM SPSS predictive analytics: Optimizing decisions at the point of impact, IBM redpaper 4710, Copyright IBM Corp. 2010.
7. Ruskov, P., K. Haralampiev, I. Milosheva, E. Efremova. Practical Educational Challenges with Business Analytics. Jubilee Scientific Conference "Statistics, Information Technology and Communications", Sofia, 2011, p. 567-577.
8. Ruskov, P., K. Haralampiev, L. Georgiev. Online Investigation of SMEs Competitive Advantage. Management, Enterprise and Benchmarking (MEB) International Conference, Budapest, 2012, p.143-159.
9. Бъчваров, А., К. Хараламписв, П. Русков. Изследване на аномалиите в месечната консумация на електричество в зависимост от температурата. XX юбилеен международен симпозиум „Управление на топлоенергийни обекти и системи", Банкя, 2012, с. 31-34