Exceptional Model Mining

Duivesteijn, W.

Chapter 4

Deviating Interactions – Correlation Model

An Exceptional Model Mining instance strives to find subgroups, for which a particular kind of interaction between multiple target attributes is unusual, when compared to that same interaction between the same attributes on the entire dataset. Possibly the simplest such interaction is the correlation model. In this correlation model, we consider two numeric targets, ℓ_1 and ℓ_2. Within this model class, we will refer to them as $x = \ell_1$ and $y = \ell_2$. We are interested in their linear association as measured by the correlation coefficient ρ, estimated by the sample correlation coefficient

$$\hat{r} = \frac{\sum \left(x^i - \bar{x}\right)\left(y^i - \bar{y}\right)}{\sqrt{\sum \left(x^i - \bar{x}\right)^2 \sum \left(y^i - \bar{y}\right)^2}}$$

where x^i denotes the i^{th} observation on x, and $\bar{x}$ denotes its mean. We let ρ^G and ρ^{G^C} denote the population coefficients of correlation for G and G^C, respectively, and let $\hat{r}^G$ and $\hat{r}^{G^C}$ denote their sample estimates.

4.1 Quality Measure $\varphi_{\mathbf{scd}}$

To find descriptions with a substantial coverage and deviating correlation coefficient, we develop a statistically-oriented quality measure, based on the test

$$H_0 : \rho^G = \rho^{G^C} \quad \text{against} \quad H_1 : \rho^G \neq \rho^{G^C}$$

Generally, the sampling distribution of $\hat{r}$ is unknown. If x and y follow a bivariate normal distribution, we can apply the Fisher z transformation

$$z' = \frac{1}{2} \ln \left(\frac{1+\hat{r}}{1-\hat{r}} \right)$$

The sampling distribution of z' is approximately normal [84]. Its standard error is given by

$$\frac{1}{\sqrt{\xi - 3}}$$

where ξ is the size of the sample. As a consequence

$$z^* = \frac{z' - z^{C\prime}}{\sqrt{\frac{1}{n-3} + \frac{1}{n^C-3}}}$$

approximately follows a standard normal distribution under H_0. Here z' and $z^{C\prime}$ are the z-scores obtained through the Fisher z transformation for G and G^C, respectively. If both n and n^C are greater than 25, then the normal approximation is quite accurate, and can safely be used to compute the p-values. As quality measure φ_{scd} (acronym for Significance of Correlation Difference) we take 1 minus the computed p-value. Because we have to introduce the normality assumption to be able to compute the p-values, φ_{scd} should be viewed as a heuristic measure. Transformation of the original data (for example, taking their logarithm) may make the normality assumption more reasonable.

4.2 Experiments

4.2.1 Datasets

The *Windsor Housing* dataset [2] concerns 546 houses that were sold in Windsor, Canada in the summer of 1987. The information for each house includes the two attributes of interest, $\ell_1 = x = lot\_size$ and $\ell_2 = y = sales\_price$. An additional 10 attributes are available as descriptive attributes, including the number of bedrooms and bathrooms and whether the house is located at a desirable location. The correlation between lot size and sale price is 0.536, which implies that a larger size of the lot

Table 4.1: Statistics concerning the datasets used in the Correlation model (this chapter), Classification model (Chapter 5), and alternative Regression model (Section 7.4) experiments. Here, N is the total number of records, k is the number of descriptive attributes, and m is the number of targets on which the model is fitted.

Dataset	Domain	N	k	m
Affymetrix	Bioinformatics	63	311	2
Windsor Housing	Residential property value	546	10	2

coincides with a higher sales price. The fitted regression function is $y = 34136 + 6.60 \cdot x$, showing that on average one extra square meter corresponds to a sales price increase of \$6.60.

The *Affymetrix* dataset comes from the domain of bioinformatics. In genetics, genes are organised in so-called *gene regulatory networks*. This means that the expression (its effective activity) of a gene may be influenced by the expression of other genes. Hence, if one gene is regulated by another, one can expect a linear correlation between the associated expression-levels. In many diseases, specifically cancer, this interaction between genes may be disturbed. The *Affymetrix* dataset shows the expression-levels of 313 genes as measured by an Affymetrix microarray, for 63 patients that suffer from a cancer known as neuroblastoma [64]. Additionally, the dataset contains clinical information about the patients, including age, sex, stage of the disease, etc. As targets, we consider the expressions of the two genes *ZHX3* ('Zinc fingers and homeoboxes 2') and *NAV3* ('Neuron navigator 3'), showing a slightly positive overall correlation of 0.218.

4.2.2 Experimental Results

On the *Windsor Housing* dataset, we run an experiment with φ_{scd}. As discussed in Section 4.1, in order to be confident about the test results for this quality measure, the coverage of a description has to be over 25. This number was used as minimum support threshold for a run of Cortana using φ_{scd}. The following description (and its complement) was found to show the most significant difference in correlation ($\varphi_{\text{scd}}(D_1) = 0.9993$)

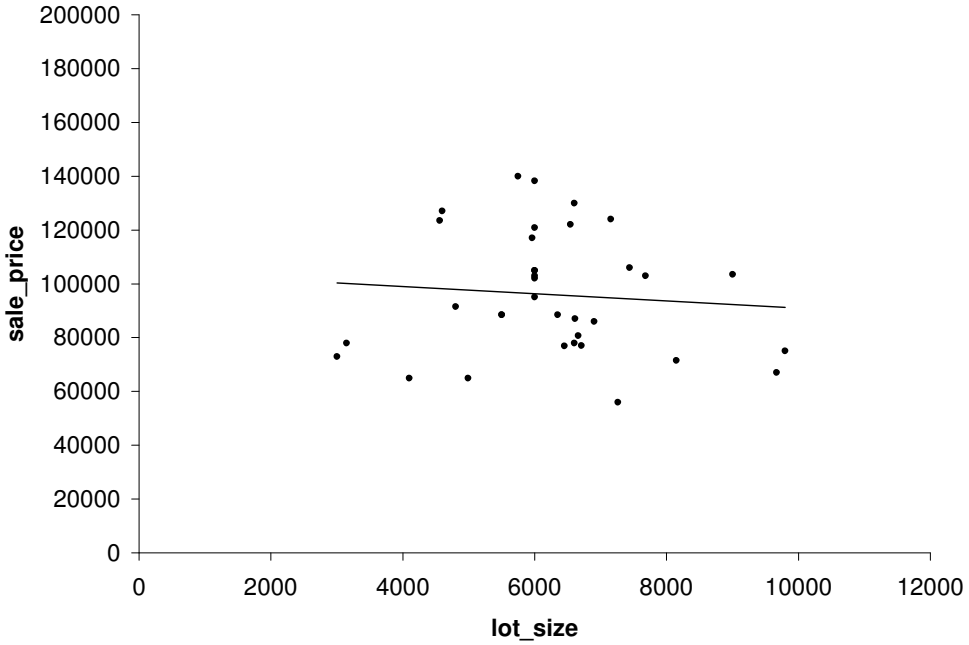
$$D_1 : drive = 1 \wedge rec\_room = 1 \wedge nbath \geq 2$$

This is the group of 35 houses (covering 6.4% of the dataset) that have a driveway, a recreation room and at least two bathrooms. The scatter plots for the D_1 and D_1^C are given in Figure 4.1. The subgroup shows a correlation of $\hat{r}^{G_1} = -0.090$ compared to $\hat{r}^{G_1^C} = 0.549$ for the remaining 511 houses. A tentative interpretation could be that D_1 describes houses in the higher segments of the market where the price of a house is mostly determined by its location and facilities. The desirable location may provide a natural limit on the lot size, such that this is not a factor in the pricing. Figure 4.1 supports this hypothesis: houses in D_1 tend to have a higher price ($\$95,947$ on average, versus $\$68,122$ on the whole dataset).

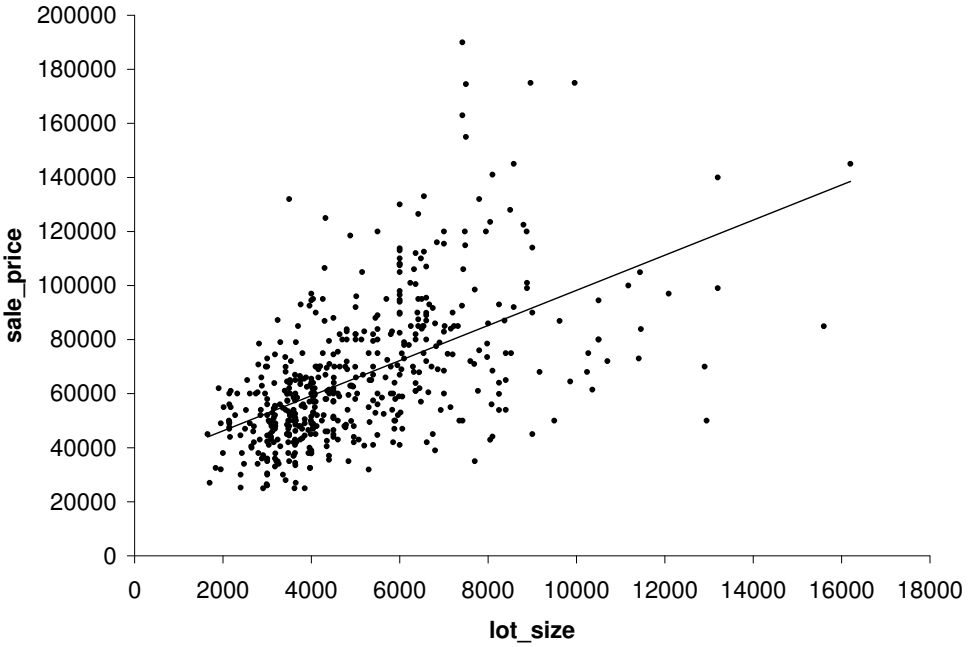
In general *sales_price* and *lot_size* are positively correlated, but EMM discovers a description with a slightly negative correlation. However, this value is not significantly different from zero: a test of

$$H_0 : \hat{r}^{G_1} = 0 \quad \text{against} \quad H_1 : \hat{r}^{G_1} \neq 0$$

yields a p-value of 0.61. The scatter plot confirms our impression that *sales_price* and *lot_size* are uncorrelated within the description. For purposes of interpretation, it is interesting to perform some post-processing. In Table 4.2 we give an overview of the correlations within different descriptions whose intersection produces the final result, as given in the last row. It is interesting to see that the condition $nbath \geq 2$ in itself actually leads to a slight increase in correlation compared to the whole database, but the combination with the presence of a recreation room leads to a substantial drop to $\hat{r} = 0.129$. When we add the condition that the house should also have a driveway we arrive at the final result with $\hat{r} = -0.090$. Note that adding this last condition only eliminates 3 records (the size of the subgroup goes from 38 to 35) and that the correlation between sales price and lot size in these three records (defined by the condition $nbath \geq 2 \wedge \neg drive = 1 \wedge rec\_room = 1$) is -0.894. We witness a phenomenon similar to Simpson's paradox: splitting up a description with positive correlation (0.129) produces two descriptions both with a negative correlation (-0.090 and -0.894, respectively). This is a real-life occurrence of an effect similar to the one we witnessed in the artificial dataset of Figure 3.1, used in Chapter 3 for the sake of argument.



(a) G_1, with $\hat{r} = -0.090$.



(b) G_1^C, with $\hat{r} = 0.549$.

Figure 4.1: *Windsor Housing* - φ_{scd}: Scatter plot of *lot_size* and *sales_price* for the subgroup G_1 corresponding to description D_1 : $drive = 1 \wedge rec\_room = 1 \wedge nbath \geq 2$ and its complement.

Table 4.2: Descriptions on the housing data, and their sample correlation coefficients and supports.

D	$\hat{r}^{G_D}$	$\lvert G_D \rvert$
Whole dataset	0.536	546
$nbath \geq 2$	0.564	144
$drive = 1$	0.502	469
$rec\_room = 1$	0.375	97
$nbath \geq 2 \wedge drive = 1$	0.509	128
$nbath \geq 2 \wedge rec\_room = 1$	0.129	38
$drive = 1 \wedge rec\_room = 1$	0.304	90
$nbath \geq 2 \wedge rec\_room = 1 \wedge \neg drive = 1$	−0.894	3
$nbath \geq 2 \wedge rec\_room = 1 \wedge drive = 1$	−0.090	35

4.3 Alternatives

A logical consideration for a quality measure would be the absolute difference of the correlation for the description D and its complement, i.e.

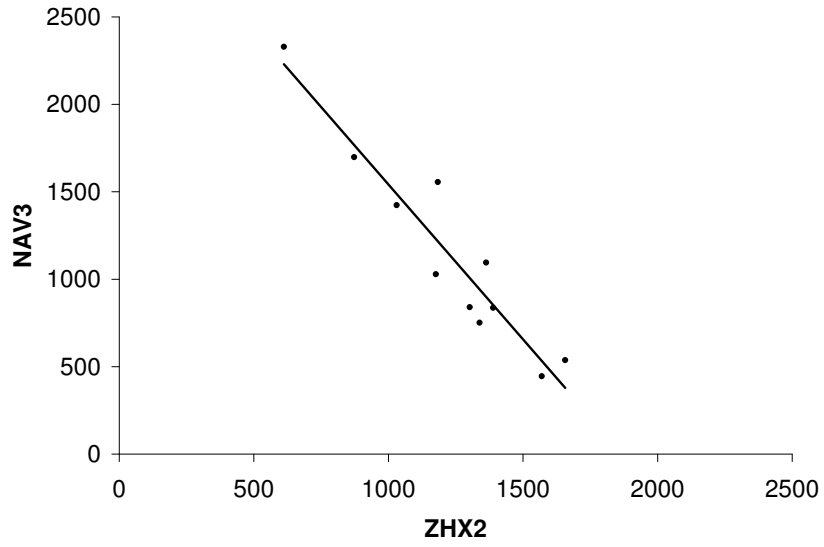
$$\varphi_{abs}(D) = \left| \hat{r}^{G_D} - \hat{r}^{G_D^C} \right|$$

Unfortunately, this measure does not take into account the coverage of the descriptions, and hence does not do anything to prevent overfitting.

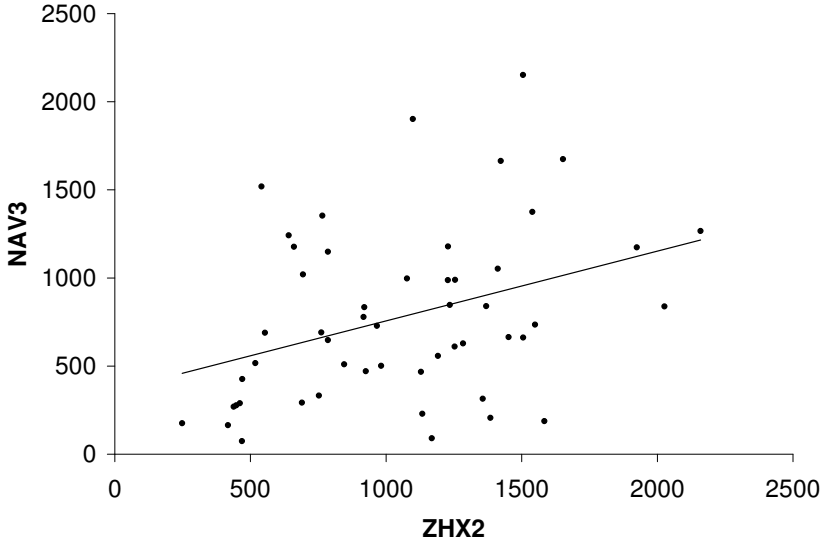
On the *Affymetrix* dataset, recall that we analyse the correlation between *ZHX3* and *NAV3*, showing a very slight correlation ($\hat{r} = 0.218$) on the whole dataset. We analyze this dataset in terms of the absolute difference of correlations φ_{abs}, allowing the use of all remaining attributes (both gene expression and clinical information) for building descriptions. As the φ_{abs} measure does not have any provisions for promoting larger subgroups, we use a minimum support threshold of 10 (15% of the patients). The largest distance ($\varphi_{abs}(D_2) = 1.313$) was found with the following description covering 11 records (17.5%) of the dataset

$D_2 : 11\_band = \text{'no deletion'} \wedge survival\ time \leq 1919 \wedge XP\_498569.1 \leq 57$

Figure 4.2 shows the plot for this description and its complement with the regression lines drawn in. The correlation for the description is $\hat{r}^{G_2} = -0.95$



(a) G_2, with $\hat{r} = -0.950$.



(b) G_2^C, with $\hat{r} = 0.363$.

Figure 4.2: *Affymetrix* - φ_{abs}: Scatter plot of the subgroup corresponding to description D_2 : *11_band = 'no deletion'* $\wedge$ *survival time* $\leq 1919 \wedge$ *XP_498569.1* ≤ 57 and its complement.

and the correlation in the remaining data is $\hat{r}^{G_2^C} = 0.363$. Note that the description displays a very "stable" behavior: all points are quite close to the regression line, with $R^2 \approx 0.9$.

As an improvement of φ_{abs}, the following quality function weighs the absolute difference between the correlations with the *entropy function* of the split between the description and its complement, as introduced in Section 3.2.1. Hence, when we find descriptions with φ_{abs}, but we find their coverage not substantial enough, we can solve this problem by running EMM with the alternative quality measure φ_{ent}, defined as

$$\varphi_{ent}(D) = \varphi_{ef}(D) \cdot \left| \hat{r}^{G} - \hat{r}^{G^C} \right|$$

4.4 Conclusions

In this chapter, we propose to use the correlation between two numeric targets as a measure of exceptionality for descriptions. This is probably the simplest form of target interplay for which Exceptional Model Mining can find deviating descriptions. As such, a domain expert should be able to easily interpret not only a found description, but also the associated model. As we have seen, particularly in discussing description D_1 found on the *Windsor Housing* dataset, a rationale for a subgroup can relatively easily be given based on the domain-specific interpretation of attributes on which the description is defined. This rationale can be fortified straightforwardly by inspecting the corresponding sample correlation coefficients. The statistical test, yielding the impression that the targets are uncorrelated within D_1, gives us confidence that the rationale makes sense. Also, a domain expert could learn a lot from observations such as the Simpson's paradox observed in Table 4.2. Thus, having only one parameter of interest in gauging the interesting interplay between targets, even though it restricts the EMM framework to relatively simple models, can enhance the analysis of the experimental results.