advantage) but also generates a loss when reselling the car after the leasing period. A conservative underestimation leads to high leasing rates and therefore generates an avoidable competitive disadvantage. With more than 150.000 recorded leasing contracts and resales over four years (2011-2014) from a major German car manufacturer, this is a typical regression problem where machine learning models can be used to explain the residual values based on features of the car like the age, the mileage or the fuel type. The trained models can be used for future contracts to determine a more accurate estimation of the expected residual value. The results affect different stakeholders, namely the car manufacturer, the leasing bank, the authorized dealers and the customers.

In such a practical application the experience and domain knowledge from managers and decisions makers about the leasing market and its special characteristics are an important factor in the whole analysis process. The goal is to achieve a closer cooperation between the data science experts and the many different experts who have in depth knowledge and experience in the field of leasing contracts and the leasing market. For this forecasting application we use ANNs which are suitable for noisy and nonlinear data [15-16]. The common criticism of ANNs is their "black-box" characteristic. After training the ANNs with data from the past (not the focus of this paper) the models are able to produce a forecast of previously unseen data (new contracts). ANNs in general and the training process in particular are highly complex mathematical optimization problems which are difficult to understand for professionals with fewer technical skills. General rejection and prejudices can be the result. But in the end, the decisions are made by humans who have to accept and trust the forecasts. On the other hand, the data scientists face the problem of incorporating the domain experts during the analysis to get the necessary domain knowledge about the data and reveal possible problems with the models which can only be identified by experts who have probably less mathematical and statistical skills (e.g. identifying omitted variables).

The general idea is to use heat maps, a familiar tool for most business people [17], to incorporate all different kinds of experts in the evaluation phase of the initial analysis, which supports iterative adjustments and improvements. The goal is to provide a method where domain experts with less technical skills, managers and data scientists can discuss the performance of the current approach on the same level of complexity. Problems with the analysis like missing data, omitted input variables or nonlinearities can be identified together, whereby all participants are actively involved in the process of deriving new or improved models. The next section describes the implementation of the proposed heat map method for model evaluation and illustrates its functionality.

3 Visual Model Evaluation with Heat Maps

Heat maps are widely used in research fields like geology or meteorology [18] and play an increasingly important role in business applications. Buehler and Pritsch use heat maps to present results of risk management models in a more intuitive and comprehensible way [17]. By providing easy to understand heat map visualizations of

different risk categories and business units they claim to make risk more transparent. Heatmaps could thus contribute to a dialogue between the board of directors, senior management and business unit leaders. Köpp et al. propose a heat map visualization technique for applications where forecast distributions of several future time steps are generated [19]. For an easier interpretation, forecast ensembles are often aggregated by the mean or median, which reduces the information to a single forecast line. The proposed heat map intuitively visualizes areas in the ensemble with high and low activity, which represents the uncertainty of the models and therefore facilitates decision making (for example the identification of the best point in time to buy a certain commodity). This is an example of how heat maps can be used to aggregate data without losing important information.

In this section, we propose a new technique for model evaluation using heat map visualizations of forecasting errors/residuals. In most data science oriented studies, model evaluation is based on some common performance measures like the Root Mean Square Error (RMSE) or the Mean Absolute Error (MAE) [20]. These performance measurements reduce the information about the model quality to a single number. These measures are usually used to compare different models or model specifications with each other in order to choose the best performing one. This common approach has the drawback that the information where exactly the model performs good or bad in the input and output space, cannot be observed. In addition, these rather technical evaluations of the analysis have no value for incorporating domain knowledge in the evaluation and are unable to generate trust in results of probably highly complex "black-box" methods.

Using heat maps for residual visualization is a new approach and has several advantages compared to conventional methods. With heat maps it is possible to incorporate a further dimension, because the information about the residuals can be visualized by a color scale. Therefore, it is possible to use a two-dimensional coordinate system with two variables of the input or output space on the axes. These "cuts" through the data enable the user to visualize the residuals in the same data space as a conventional scatterplot or density plot of the data points. This can help to detect hidden anomalies in the data/model when comparing the residual heat maps with scatterplots or density plots which are in the same range and size. These hidden anomalies can be, for example, a bad performance at the borders of the distribution or a special region like a data gap in the input space, which leads to bad performance in the results. This can give hints for an omitted variable bias or an insufficiently incorporated, nonlinear relationship between variables. The representation of the error terms in the form of a heat map remains clear and comprehensible even with a large number of data points, since the general appearance is independent of the number of data used to generate the visualization. Therefore, the method is also suitable for Big Data applications. However, in the case of large data sets, data aggregation takes place but without losing information about the performance in different regions of the data space, as is the case with single performance measures.

To illustrate the technique, let's consider 500 samples from a two-dimensional standard normal distribution. This stylized example serves as an artificial use case with known distribution and properties of the data points to validate the functionality of the

method within a controlled environment. The two dimensions (x and y) represent two continuous input variables (features) of a machine learning regression model. This information can be used to generate a scatterplot of the data points to get an idea about the distribution and possible correlations. Figure 1 shows a smoothed scatterplot of this example.

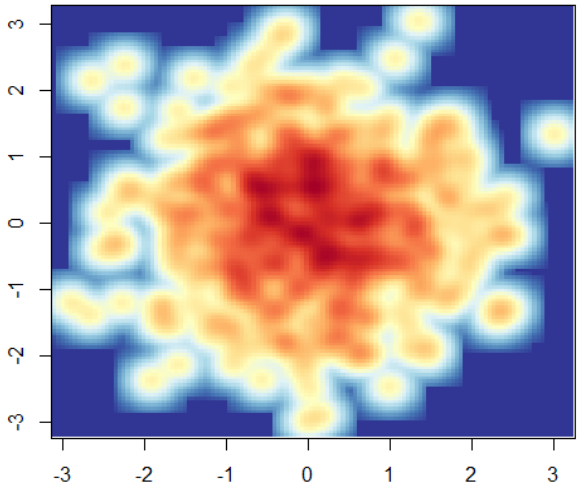

Figure 1. Smoothed scatterplot of 500 random samples of a two-dimensional standard normal distribution.

The residuals which correspond to these data points are now visualized in the exact same data space using a color scale. The assumption is, that each residual is representative for a specified region around its position. For example, it might be the case that a forecasting model predicts the realized values very well in one region of the data space (residuals in that area are rather low) and in another region the forecasting accuracy is less accurate (residuals in that area are rather high). Since the position of the data points are typically not on a regular grid, a method for weighting and smoothing is necessary. By using a kernel around the data points, each residual gets a specific range of influence. At the position of the data point, the influence of the residual should be high, while it decreases with increasing distance. The goal of using the kernel approach is to generate a regular grid with the weighted influence of the residuals on each grid point. A resulting discrete and regular grid is defined as a matrix of the dimension (m, n) within the same variable space. There are several other interpolation methods for irregular grids like kriging and inverse distance weighting [21], but in this use case, an actual weighting of the residuals has to be performed which must also work with data points at the exact same position (high density areas in the data). Each residual has a defined location in the variable space. In our approach, we use a two-dimensional Gaussian kernel with

$$\mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix} \quad (1)$$

where $\begin{pmatrix}\mu_A\\ \mu_B\end{pmatrix} \coloneqq \begin{pmatrix}x_i\\ y_i\end{pmatrix}$ for each new residual i. The covariance matrix Σ can be set by the user but for the default value we specify equation 2 with size s = 1000 as control parameter for the expansion of the kernel and zero correlation

$$\Sigma_{default} = \begin{pmatrix}\frac{(\max(X)-\min(X))^2}{s} & 0\\ 0 & \frac{(\max(Y)-\min(Y))^2}{s}\end{pmatrix} \quad (2)$$

The density function in the special two-dimensional case, for a specific residual i, at the grid points (m, n), can be defined as

$$f^i(m,n) = \frac{1}{2\pi\sigma_A\sigma_B\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(n-\mu_A)^2}{\sigma_A^2}+\frac{(m-\mu_B)^2}{\sigma_B^2}-\frac{2\rho(n-\mu_A)(m-\mu_B)}{\sigma_A\sigma_B}\right)\right) \quad (3)$$

For each residual, the position (x_i, y_i) is defined as the mean of the density function and each grid point is inserted into the generated function 3. This results in a list of weights (according to the two-dimensional normal density) for each residual, for each grid point. The threshold parameter t can be set to specify the minimal value of a weight, otherwise this residual has no influence on this specific grid point. The weight $w_{m,n}^i$ for residual i at grid point (m, n) is defined by

$$w_{m,n}^i = \begin{cases} f^i(m,n), & \text{if } f^i(m,n) \geq t \\ 0, & otherwise \end{cases} \quad (4)$$

If the distance between the residual and the grid point is too high (grid point inserted into the normal density kernel, with mean equals the position of the residual, results in a value below the threshold) the influence for this specific value is set to zero. Figure 2 illustrates the weighting procedure.

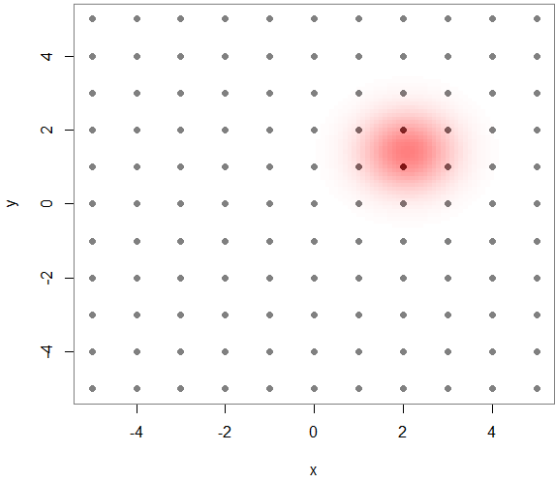

Figure 2. An example of one residual inserted into the regular grid of the heat map. The position of the residual in the data space is equal to the mean of the two-dimensional normal density.

The actual weighting for each residual r_i at each grid point is performed according to equation 5.

$$g_{m,n} = \begin{cases} \sum_{i=1} \left(r_i \frac{w_{m,n}^i}{\sum_{i=1} w_{m,n}^i} \right), & \text{if } \sum_{i=1} [w_{m,n}^i > 0] \geq c \\ NA, & \text{otherwise} \end{cases} \quad (5)$$

The cutting parameter c can be set to allow only those grid points to be colored that have at least a specified number of $w_{m,n}^i \geq c$ which means that more than one residual has an influence on that specific grid point. This can help to make the result more robust, because single value outliers are mitigated, but at the expense of greater information loss in low-density areas. After assigning a value $g_{m,n}$ to each point in the regular grid, the heat map can be produced. To illustrate the resulting heat maps, a third sample of 500 standard normal distributed values represents the artificial residuals corresponding to each (x, y) point in the data space. Figure 3 shows the result of this stylized example.

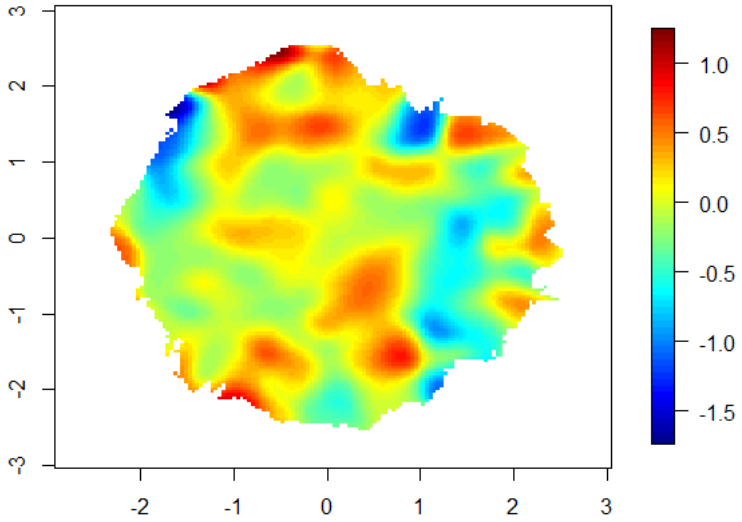

Figure 3. The figure shows a residual heat map which is generated according to the described procedure. In this case, the residuals are normally distributed which results in a heat map with no clear patters.

The red regions represent the high residual areas, while in the blue regions the residuals are rather small. In this example, there is no clear pattern apparent because the residuals are indeed normally distributed over the whole data space. This should be the case if the (forecasting) models work well in each area of the data space.

To illustrate a case when a specific region is biased, the residuals are now artificially manipulated by increasing the residuals located within the range of -0.1 and 0.1 on the x-axis according to equation 6.

$$r_i = \begin{cases} r_i + 2, & \text{if } -0.1 \leq x_i \leq 0.1 \\ r_i, & \text{otherwise} \end{cases} \quad (6)$$

Figure 4 shows the resulting heat map with the artificially biased region. A clear pattern is now visible, indicated by the red region. Such patterns would be expected if the model performs bad, especially in a particular region, for example because of an omitted variable or a nonlinearity which was not incorporated.

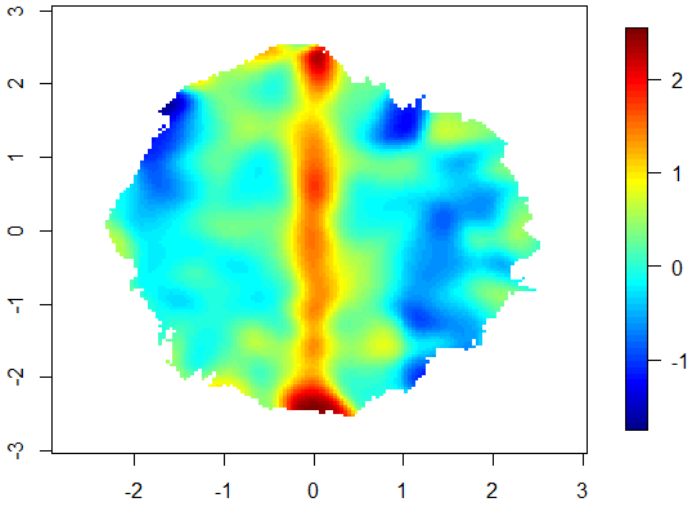

Figure 4. The residuals are now artificially manipulated by increasing the residuals located within the range of -0.1 and 0.1 on the x-axis. A clear pattern is visual now indicated by the red region. Such patterns would be expected if the model performs bad in a particular region for example because of an omitted variable or a nonlinearity which was not incorporated.

4 Discussion and Limitations

The method described in the previous section is now applied to the real world problem of forecasting the resale price of used cars. This example only serves as illustration purpose how visualization can be used in a business application and help to reveal previously unseen problems. After training the ANN models with the available input data, we visualize the resulting model errors on a validation dataset with 4500 hold out samples which were not represented in the training process. First, figure 5 presents a smoothed scatterplot of the observed residuals in the data space with the two variables “sale price” and “age” of the cars. The variable “age” is scaled and centered around zero (best practice for input variables of ANNs [22]), while the variable “sale price” is defined as the ratio of the initial list price and achieved market value. Figure 6 shows the resulting residual heat map (residuals as absolute values) of the same data set.

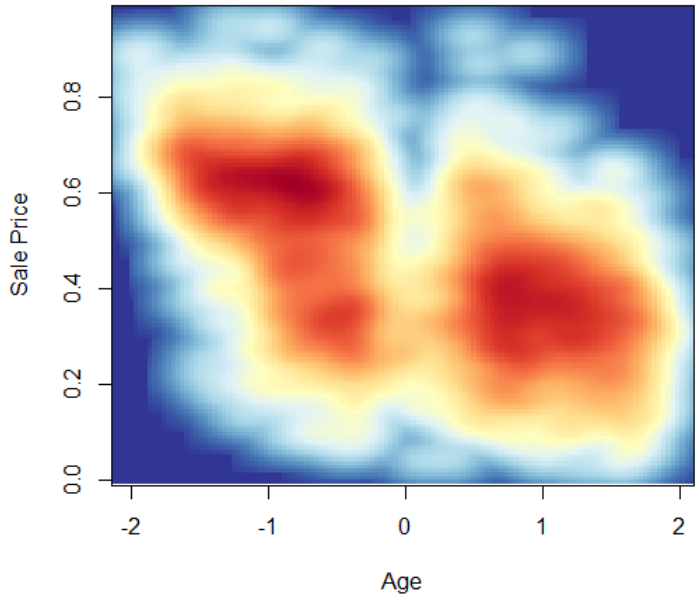

Figure 5. Smoothed scatterplot of the observed residuals in the data space with two variables: the sale price and age of the cars.

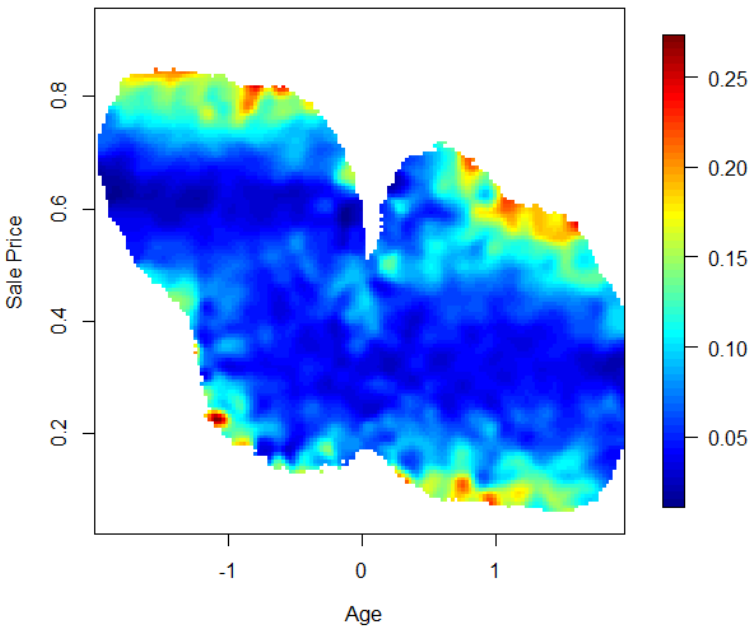

Figure 6. A systematic pattern in the residuals can be observed. In the upper and lower regions of the variable “sale price”, the residuals tend to be higher than in in the more centered parts.

A systematic pattern in the residuals of the ANN model can be observed. In the upper and lower regions of the variable "sale price" the residuals tend to be higher than in in the more centered parts. This can be a hint about information which are not incorporated in the model. Discussions with experts and managers about the observed results in this visualization reveal a plausible explanation at least for the upper regions of the sale price. In the data, no enhanced acquisition costs are reported. For example, a crane which was installed later on a used car increases the value of the whole car significantly, even if the residual value of the car decreases normally. When such cars are resold, no information about this extra equipment will be reported and the residual value of this car seems to be unusually high. The respective data points which would otherwise be marked as simple outliers can now be identified as patterns which are subject to a systematic bias induced by an omitted explanatory variable. Future data collection will incorporate information about enhanced acquisition costs.

From a manager perspective, this visualization technique enables the participation in the complex process of model building, evaluation and adjustment, even with lesser mathematical and statistical skills. They can assess the performance of the models within a familiar environment and actively discuss practice-oriented ideas about the problems of the advanced analytics models with the data science experts on the same complexity level. The participation in the "black-box" phase is also the basis for trust and acceptance of the results which can lead to more confident decision making.

From a data scientist perspective, this method can be used to uncover hidden problems in the data or algorithms based on the domain knowledge of experts in the field. This approach also offers a new way of visualizing the model error within the same data space as scatterplots or density plots. For example, comparing density plots and heat maps can provide a first insight about the performance in regions with less data or data gaps. This can give a hint if the problem is due to the data (possible omitted variables) or the used method (a nonlinearity which is not correctly incorporated). It provides much more valuable information than single performance measures like the RMSE. In contrast to conventional residual analytics methods like simple residual plots, the heat map can be used with an unlimited number of data points (applicable for Big Data machine learning models) in the evaluation. Residual plots are hardly interpretable if the number of data points increases. The color scale and the weighting function of the heat map allows to visualize many patterns without losing information about the model performance in different regions of the data space.

Like any approach, the heat map visualization faces some difficulties and drawbacks. Big Data machine learning applications often contain a large number of input factors, of which only two can be plotted simultaneously in one coordinate system. Scatterplot matrices can be used to represent all variables pairwise, but this also leads to a more complex visualization as the number of factors increases. In the described example, also only continuous variables are investigated within a regression problem. The information gain for one or even two discrete or binary variables on the axes is questionable. The method is also limited to regression problems. Classification problems can be incorporated by some adjustments on the heat map color scale (binary representation with two colors for true or false) for future tests. As mentioned in section 3, several parameters have to be set properly for achieving good results. These

parameters could have a tremendous influence on the heat map appearance. It is not a plug and play solution and users need substantial knowledge about the backend of the system. It is therefore once again the responsibility of the data scientists to generate meaningful presentations which can also lead to skepticism. Wrong parameter settings could have tremendous negative effects regarding trust and decision making performance. The advantage that decisions can be made with more confidence using visualizations can also be a drawback if these approaches lead to overconfidence in the models. The principle "trust but verify" is important with any model, but the visualizations do not help to understand the mathematical structure or the training process in general. An important problem is how to measure the actual benefits and the role of trust, acceptance and confidence [11-12] when using such visualizations in the analysis process. In a real world application, hardly any quantifiable indicator exists for the performance improvements induced by this visualization except a qualitative argumentation.

5 Conclusion

In this paper, a heat map visualization for model evaluation is introduced. The overall goal is a better integration of managers and decision makers in the model building and evaluation process of the data science pipeline in order to generate trust and acceptance and furthermore, to make use of the experts' domain knowledge in this phase. Based on a real world business example, the benefits of this method were discussed. The visualization technique allows domain experts to actively participate in the technical and complex model building and evaluation process which helps to reveal a systematic bias induced by an omitted explanatory variable. Data scientists are equipped with a tool for visualizing model errors in the same data space as two dimensional scatterplots to identify the exact regions where the models perform good or bad. The method is applicable even for Big Data machine learning models.

In further research we are interested in better understanding the actual benefits of visualization for data science in general. One approach is the laboratory experiment, in order to measure the results of an analysis process in a controlled environment by providing different visualization tools for control and treatment groups. This also allows to measure the differences in trust, acceptance and confidence of the subjects. In general, this paper aims to encourage the use of visualization in data science and hopefully initiates more research in this important area.

References

1. Conway, D.: The Data Science Venn Diagram, http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram (Accessed: 25.09.2016)
2. Shim, J.P., Warkentin, M., Courtney, J.F., Power, D.J., Sharda, R., Carlsson, C.: Past, present, and future of decision support technology. Decision Support Systems. 33, 111–126 (2002)

3. Power, D.J.: Decision Support Systems: A Historical Overview. In: Handbook on Decision Support Systems 1. pp. 121–140. Springer Berlin Heidelberg (2008)
4. Bresciani, S., Eppler, M.J.: The Benefits of Synchronous Collaborative Information Visualization: Evidence from an Experimental Evaluation. IEEE Transactions on Visualization and Computer Graphics. 15, 1073–1080 (2009)
5. Keim, D.A.: Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics. 8, 1–8 (2002)
6. Keim, D.A., Panse, C., Sips, M., North, S.C.: Visual data mining in large geospatial point sets. IEEE Computer Graphics and Applications. 24, 36–44 (2004)
7. Kelleher, C., Wagener, T.: Ten guidelines for effective data visualization in scientific publications. Environmental Modelling & Software. 26, 822–827 (2011)
8. Sun, G.-D., Wu, Y.-C., Liang, R.-H., Liu, S.-X.: A Survey of Visual Analytics Techniques and Applications: State-of-the-Art Research and Future Challenges. J. Comput. Sci. Technol. 28, 852–867 (2013)
9. Al-Kassab, J., Ouertani, Z.M., Schiuma, G., Neely, A.: Information visualization to support management decisions. Int. J. Info. Tech. Dec. Mak. 13, 407–428 (2014)
10. Franz, M., Scholz, M., Hinz, O.: 2D versus 3D Visualizations in Decision Support – The Impact of Decision Makers' Perceptions. ICIS 2015 Proceedings. (2015)
11. Becker, J., Heddier, M., Öksüz, A., Knackstedt, R.: The Effect of Providing Visualizations in Privacy Policies on Trust in Data Privacy and Security. In: 2014 47th Hawaii International Conference on System Sciences. pp. 3224–3233 (2014)
12. Sacha, D., Senaratne, H., Kwon, B.C., Ellis, G., Keim, D.A.: The Role of Uncertainty, Awareness, and Trust in Visual Analytics. IEEE Transactions on Visualization and Computer Graphics. 22, 240–249 (2016)
13. Wu, J.-D., Hsu, C.-C., Chen, H.-C.: An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications. 36, 7809–7817 (2009)
14. Lessmann, S., Listiani, M., Voß, S.: Decision Support in Car Leasing: A Forecasting Model for Residual Value Estimation. ICIS 2010 Proceedings. (2010)
15. Bishop, C.M.: Neural networks for pattern recognition. Oxford university press (1995)
16. Schocken, S., Ariav, G.: Neural networks for decision support:: Problems and opportunities. Decision Support Systems. 11, 393–414 (1994)
17. Buehler, K.S., Pritsch, G.: Running with risk. McKinsey Quarterly. 40–49 (2003)
18. Hagh-Shenas, H., Kim, S., Interrante, V., Healey, C.: Weaving Versus Blending: a quantitative assessment of the information carrying capacities of two alternative methods for conveying multivariate data with color. IEEE Transactions on Visualization and Computer Graphics. 13, 1270–1277 (2007)
19. Köpp, C., Mettenheim, H.-J., Breitner, M.H.: Decision Analytics with Heatmap Visualization for Multi-step Ensemble Data - An Application of Uncertainty Modeling to Historical Consistent Neural Network and Other Forecasts. Business & Information Systems Engineering. 6, 131–140 (2014)
20. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques, Third Edition. Morgan Kaufmann, Burlington, MA (2011)
21. JIN, G., LIU, Y., NIU, W.: Comparison between Inverse Distance Weighting Method and Kriging [J]. Journal of Changchun University of Technology. 3, (2003)
22. LeCun, Y.A., Bottou, L., Orr, G.B., Müller, K.-R.: Efficient BackProp. In: Montavon, G., Orr, G.B., and Müller, K.-R. (eds.) Neural Networks: Tricks of the Trade. pp. 9–48. Springer Berlin Heidelberg (2012)