It's not a Bug, it's a Feature: How Visual Model Evaluation can help to incorporate Human Domain Knowledge in Data Science

Completed Research Paper

Dennis Eilers
Leibniz Universität Hannover
Königsworther Platz 1
30167 Hannover, Germany
eilers@iwi.uni-hannover.de

Cornelius Köpp
edicos Group
Hindenburgstraße 28/29
30175 Hannover, Germany
cornelius.koepp@edicos.de

Christoph Gleue
Leibniz Universität Hannover
Königsworther Platz 1
30167 Hannover, Germany
gleue@iwi.uni-hannover.de

Michael H. Breitner
Leibniz Universität Hannover
Königsworther Platz 1
30167 Hannover, Germany
breitner@iwi.uni-hannover.de

Abstract

The question of how to incorporate human domain knowledge in practical data science projects is still a major challenge. While machine learning tasks are usually carried out by technically skilled data scientists, these analysts do not necessarily have the required domain knowledge concerning a particular business problem to explain certain phenomena. In real-world data science applications, this may result in models that do not adequately reflect relationships in the data. We address this issue by introducing a heat map based technique for model error visualization to facilitate discussions of the results between data scientists and domain experts. By discussing model errors with domain experts during the iterative analysis process, the generated insights can be used for engineering new features (explanatory variables) which better represent the problem and therefore improve the results. We demonstrate the visualization approach based on artificial data and in the context of a real-world industry example.

Keywords: Feature Engineering, Domain Knowledge, Data Science, Visualization, Heat Maps, Model Evaluation

Introduction and Motivation

Data-driven decision-making is a steadily growing field in information systems research (Agarwal and Dhar 2014). The underlying concepts can be summarized under the term data science, which is often defined as a combination of statistics, programming and domain knowledge (Conway 2010). While statistics and programming knowledge are obvious requirements for valid and reproducible analyzes, the role of domain knowledge is more unclear in this discipline (Dumbill et al. 2013). In the literature, domain knowledge is often seen as part of an iterative feature engineering process for machine learning tasks (Domingos 2012). Features (independent variables), define the input space for an algorithm which learns the dependencies in

the data in order to extract important insights from it or support future decisions based on new, previously unseen patterns. Feature engineering, as a process of constructing a proper representation of the input space for a machine learning algorithm, is a challenging task and requires knowledge relating to the analyzed problem statement and the available data. Besides the development of improved learning algorithms, feature engineering and the general question of how to present data to a machine learning algorithm are receiving increasing attention. Many recent studies focus on the importance of input representation and justify their success on the basis of intelligent feature engineering (Lash and Zhao 2016; Ghiassi et al. 2016). In practical applications this is usually an iterative procedure comprising model construction and evaluation as well as adjustments based on the performance of the current input space. It is shown that finding the right features for a given problem statement can outperform the most sophisticated algorithm optimizations (Yu et al. 2010; Bengio et al. 2013; Heaton 2016).

In contrast to the available concepts regarding feature engineering, we identify a general problem, especially in real-world industry applications. While more straightforward analytic business tasks are tackled more and more directly by decision-makers using self-service business intelligence systems (Alpar and Schulz 2016), the increasing complexity of machine learning tasks involving large amounts of data is often handled exclusively by highly specialized teams of data scientists. While data scientists have the necessary technical skills, the most valuable domain knowledge lies with employees, managers and decision-makers with many years of experience in their field. This can result in a gap between data scientists, who need domain knowledge to explain certain phenomena in data with the correct features, and domain experts, who lack technical skills to understand the information that data scientists actually require to solve certain problems by means of their models. This represents a gap between understanding/explaining the problem and the available solution in the form of domain knowledge. The consequences can be two-fold. On the one hand, data scientists lack important information for building proper machine learning models, while on the other hand, domain experts who use the models as decision support aids may not accept or trust the results of black-box models constructed without their participation. Both can lead to non-optimal decision-making. Integrating these expert groups to follow a common objective is still a major challenge today for a successful data science project in the industry and therefore a suitable field for information systems research. A collaborative analysis system addressing this issue should therefore focus on both aspects. It is important to most efficiently support human decision-makers with data-driven expert systems, and much research has been carried out in this area (Shim et al. 2002; Power 2008). But it is equally important that domain experts are also part of the system itself, e.g. by supporting data scientists with their domain knowledge when constructing the underlying models.

A key success factor for this purpose is communication between different groups. Communication is shown to be essential for overall team performance and can be achieved by proper visualization techniques for a given task (Bresciani and Eppler 2009). Visualizations traditionally focus on decision support systems (Al-Kassab et al. 2014; Franz et al. 2015), exploratory analysis (Keim 2002; Keim et al. 2004), and communication of results for better comprehensibility (Kelleher and Wagener 2011; Sun et al. 2013). Standard approaches for model evaluation such as QQ-plots or residual plots are often unintuitive for most non-technical personnel. In this study we therefore address the aforementioned issue by introducing a new visualization technique for machine learning model evaluation based on the idea of heat maps, which is a familiar concept for many business people and analysts alike (Buehler and Pritsch 2003; Köpp et al. 2014). The idea is to visualize model residuals using a color scale dependent on the features derived from the underlying data. This can help to intuitively reveal location-dependent differences of model performance in the constructed two-dimensional data space. This is intended to provide an intuitive view of model performance and yields context-based insights into the behavior of model residuals.

The paper is motivated by a real-world industry example in cooperation with a large car manufacturer. The data analysis problem is based on the question of how to forecast the resale prices of leased cars. This is a typical supervised learning regression problem in which features such as the age and the mileage of a car determine the dependent resale price. These forecasts have a critical impact on the leasing business performance of the company because the expected resale price of a car is the most important factor for determining the necessary monthly leasing rates. Based on more than 250.000 data sets from completed leasing contracts during 2011/01 to 2016/12, we illustrate our analysis approach, how we try to incorporate the domain knowledge of different stakeholders, and especially how the visualization technique can be used to iteratively improve the results. The contribution of this paper can be summarized as follows:

- From a management perspective: The proposed heat map visualization of machine learning model performance provides an intuitive and familiar view on the analysis results. This can facilitate discussions with data scientists about model performance on the same level of complexity. This makes it easier for domain experts to incorporate their knowledge to find and possibly solve problems with the available data or analysis approach. The participation can lead to a greater trust in the results, and therefore to more confident decision-making.
- From a data science perspective: The heat map approach helps the data scientist to pose the right questions about their own models, which can be answered with the available domain knowledge. Data scientists are thus equipped with an intuitive tool for assessing the performance of their models, which helps to identify regions of poor performance in the data space. The technique can also be applied for big data model diagnostics.

The following section provides a theoretical background including related concepts and research. Based on these ideas, we then introduce our visualization approach. The applicability is first demonstrated using artificial data. We then introduce the industry example and illustrate the use of the technique in a real-world data analytics project. This is followed by a discussion outlining the limitations of the approach and possible directions for further research. The paper concludes with a short summary.

Research Background

Starting from the problem statement, we are interested in how domain experts can be better integrated into the data science process. Domain knowledge plays an important role in different analytics applications and many studies try to make use of human perception and creativity by way of visualizations. Most of the approaches use visual representations of data to synthesize information and derive insights from it (Keim et al. 2008). Tam et al. (2011), for example, use interactive visualizations of a high-dimensional feature space, which enable humans to generate a decision tree for detecting facial expressions and emotions. They show that decision trees generated by humans outperform machine learning approaches. Domain knowledge is critical in this application, especially in the sparse data space. This application is useful for dealing with noise and uncertainty as well as for mitigating underfitting or overfitting problems. The use of visualization methods in an analytics context can be summarized under the term visual analytics. For a comprehensive survey, we refer to Sun et al. (2013) and Endert et al. (2014). A first definition of visual analytics is given by Thomas and Cook (2006) as the science of analytical reasoning facilitated by interactive visual interfaces. Keim et al. (2008) set the focus of visual analytics on the integration of humans and machines, and state that tackling analysis problems in combination lays the foundation for significantly improved solutions. Even though recent studies (Tam et al. 2011; Tam et al. 2017) show that visual analytics approaches can outperform pure machine learning approaches, the overarching idea is the integration of human knowledge in the analysis process for joint reasoning of humans and machines in order to achieve better results (Endert et al. 2014; Keim et al. 2008; Krause et al. 2016).

From the perspective of machine learning, domain knowledge represents an important component in the iterative process of feature engineering. Domingos (2012) points out that the key success factor for a machine learning application are not the algorithms themselves but rather engineering the correct features. The reason for this is that most algorithms are of general purpose whereas the presented features are domain-specific and pose a challenge in each new case. An active field of research is the automation of feature engineering (Markovitch and Rosenstein 2002; Harvey and Todd 2015). Most of the studies in this field also highlight the importance of incorporating human domain knowledge. Recent studies try to combine the ideas presented in the broad visual analytics literature with ideas from feature engineering. These focus on the development of visual analytics techniques and tools to facilitate the feature engineering process. Brooks et al. (2015) apply a visual summarized representation of errors in text classification problems. They show that it is possible to significantly increase classifier performance using their visual analytics approach. They also show that a closer look at model errors can be beneficial regarding the iterative process of feature engineering. They recommend further research in this field aimed at the way in which new features should be considered.

The main objective of these approaches is to enable domain experts to work directly on the various models with easy-to-use tools. Because visualizations can greatly improve the communication necessary for successful team performance (Bresciani and Eppler 2009), our goal is rather to facilitate a closer collaboration between domain experts and highly-specialized data scientists in order to combine the best

of both worlds. The data scientist should be able to ask comprehensible questions about model performance and the domain expert should be able to assess the models intuitively to recognize possible problems without having to understand the exact machine learning methods in detail. A better mutual understanding of the perception of the other can then also help to increasingly blend the roles across organization teams. Krause et al. (2016) introduce an interesting new approach which is not directly related to feature engineering but provides an inspiring insight into how to improve an interpretation of the outcome of machine learning models by means of visual analytics. The authors use interactive visualizations of input-output relationships from classifier models to make the reasoning behind the black-box model more comprehensible for users. They claim that for model evaluation in particular, a better interpretation of the models using this technique can be beneficial. The core idea is to understand model behavior rather than trying to understand the black-box itself. Adopting this approach, they make models more comprehensible without reducing predictive performance. They also recommend further research on the development of new visualizations for this purpose. This idea is especially useful for applications aimed at incorporating domain experts in the feature engineering process because it demonstrates that regarding model quality, it is more important to understand the reasons underlying model behavior rather than understanding the complex model in detail. The level of model behavior is important as well as intuitive, and is therefore suitable as a basis for discussions between different expert groups. Our goal in this study is hence to visualize the behavior of complex models, especially to enable domain experts to participate in discussions about the performance based on a behavior level. Therefore, we adopt the ideas of Brooks et al. (2015) to focus on the errors of the generated model and try to combine these with an understanding of the input-output relationship according to Krause et al. (2016) to connect "model behavior understanding" with "error understanding".

In order to achieve a meaningful integration of different expert groups, the visualization technique must be comprehensible for all. The goal must be to establish a common language between groups with different proficiencies to enable a transparent discussion about the model quality with all stakeholders. Developing and promoting visual literacy as a bridge between data literacy and domain literacy can be one way to address this. Through the creative use of vision-competencies, people can be enabled to communicate with each other on the same level (Fransecky and Debes 1972). The implementation or development of a visualization technique that fulfills all the above ideas and defined requirements represents the challenge for our research. In this paper we propose the concept of heat maps and argue by means of examples from the literature that their general idea is familiar for domain experts and data scientists alike, which makes them suitable for our purposes.

Buehler and Pritsch (2003) have already presented heat maps in a practical risk management application as a tool to illustrate the results of their models in a more intuitive and comprehensible way. They provide easily comprehensible heat maps of different risk categories and business units and justify their work by way of commonly assessable and more transparent representation of risks. In this case, the heat maps contribute towards a dialogue between the board of directors, senior management and business unit leaders. Köpp et al. (2014) propose a heat map visualization technique for applications in which forecast distributions of several future time steps are generated. In order to facilitate an interpretation of the results, 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. This represents the uncertainty of the models, thereby facilitating management decision-making (for example the identification of the best point in time to buy a certain commodity). In a recent study, Klemm et al. (2016) use 2D and 3D heat maps to visualize quality-of-fit measures such as the R^2 of regression models. The authors use all possible feature input combinations for constructing their heat maps in order to investigate the underlying relationships between input and target features. In their case study they report that domain experts greatly appreciate their approach for exploring dependencies between the presented factors and a specific disease. These studies demonstrate the general applicability of heat maps for comprehensible and intuitive data/model representations. In this study we apply these ideas and our own experience from practice to focus on communication between data scientists and domain experts about the model quality in order to improve feature engineering by means of visualization. The following section introduces the heat map approach and illustrates functionality based on artificial examples.

Heat Map Visualization of Model Errors in Feature Space

In many data science oriented studies, the evaluation of presented model types is carried out by reporting performance measures such as the Root Mean Square Error (RMSE) in the case of regression analysis or accuracy assessment in classification problems. These measures aggregate the overall performance to a single number which is used for comparing different models. "Horse races" of many different benchmark models often neglect individual model properties and individual optimization potentials. Information aggregation makes it impossible to gain an insight into model performance in different regions of the data space and these rather technical performance measures are of limited value for investigating possible improvements in an iterative feature engineering process. There already exist several ideas to visualize residuals for model diagnostics (Kuhn and Johnson 2013). Residual plots are frequently applied for linear regression analyses to identify non-linearities or heteroscedasticity but they do not allow an identification of possible reasons for bad performance in the context of different feature combinations. A check for remaining patterns/information or outliers in the residuals can also be done with histograms or QQ-plots which would reveal deviations from a normal distribution but also no proper connection to the features and thus to the problem domain can be established. This also applies to more general visual model evaluation techniques like actual vs. predicted scatter plots or lift charts and ROC curves for classification problems.

In this study we therefore focus on the feature space and try to analyze and assess the behavior of individual models in different regions of the underlying data. For this purpose a new technique for visualizing model errors based on the concept of heat maps is introduced which makes it possible to represent model errors in the same dimensions as familiar scatter plots. This extends the possibilities of residual plots by a further dimension and makes errors visible in an intuitive and familiar form. In a pilot study we illustrated a first approach regarding the application of heat maps for residual analytics and incorporated feedback from scientists and practicing employees (Eilers and Breitner 2017). We have further developed these ideas and now combine the above-mentioned findings from previous research dealing with visual analytics and visual model evaluation. The visualization technique presented in this study now explicitly addresses the identified business problem of a gap between data scientists and domain experts by providing a communication instrument for discussing model performance and possible ideas for feature engineering. The code will be available on GitHub (https://github.com/eilersde/residualheatmap) to enable further extensions and evaluations of the technique in different contexts. In this section, functionality is illustrated by way of artificial data examples with known statistical properties. We then go on to apply the technique to a real-world industry example in subsequent sections.

Firstly, in order to gain an insight into the feature space in which the heat maps can be applied, scatter plots are a commonly used method to visualize pairwise distributions of the data, and provide an overview of regions with available information. As shown by Carr et al. (1987), a meaningful representation can be achieved by incorporating the density of the distribution even for large sample sizes. Figure 1 shows a smoothed scatter plot of 500 samples from a standard normal distribution. In this first example, x and y serve as two continuous features which explain a dependent variable in an artificial regression problem. Darker regions represent a higher density, indicating a greater availability of information.

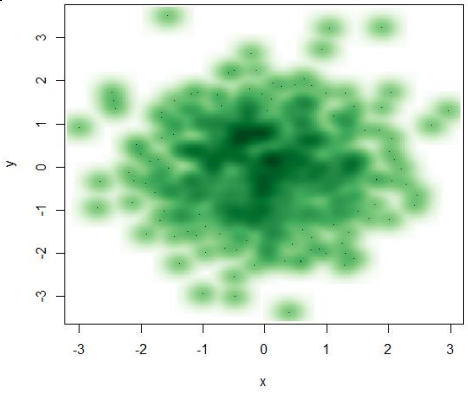


Figure 1. Scatter plot with density information

The residuals r are calculated as the difference between the real value of the dependent variable and the model prediction. In order to construct the residual heat map in the same two-dimensional feature space, each data point is assigned its corresponding residual. It is assumed that each residual represents the local model performance. Since the positions of the data points are typically not on a regular grid, a method for weighting the influence is necessary. By using a kernel around the data points, each residual is assigned a specific range of influence. At the position of the data point, the influence of the residual should be high, while decreasing with the distance. The aim of using the kernel approach is to generate a regular grid with the weighted influence of the residuals at each grid point. In contrast to a discrete binning approach, the kernel guarantees a stable representation independent of the selected grid dimensioning. The result is defined as a matrix (regular grid) of dimension (M, N) which serves as a basis for generating the residual heat map. In this study we use a Gaussian kernel with

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

where $\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \coloneqq \begin{pmatrix} x_i \\ y_i \end{pmatrix}$ for each new residual r_i. The covariance matrix Σ of the kernel can be specified depending on the scale of the features. As a default, we set

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

where s controls the kernel expansion and thus the range of influence of each residual. In the general case, the kernel function for a residual r_i at the grid points (m, n) is then defined as

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

For each residual, the position (x_i, y_i) is defined as the mean (center) of the kernel and each grid point is inserted into the generated function (3). This procedure maps residual weights onto the corresponding grid point. The weight $w_{m,n}^i$ for residual r_i at grid point (m, n) is defined by

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

The threshold parameter t can be set to specify the minimal required value of a weight. At lower values, it is assumed that this residual has no influence on this specific grid point. Hence, if the distance between the residual and the grid point is too large, the weight for this specific value is set to zero in order to ensure a merely local influence. This prevents assignment of performance values in regions without any information. The actual value $h_{m,n}$ assigned to a specific grid point is calculated by the weighted sum of all residuals. The weight $w_{m,n}^i$ for residual r_i is divided by the sum of all weights at grid point (m, n) which ensures standardized values in the interval [0,1].

$$h_{m,n} = \begin{cases} \sum_i \left(r_i \frac{w_{m,n}^i}{\sum_j w_{m,n}^j} \right), & \text{if } \sum_i \left[w_{m,n}^i > 0 \right] \geq c \\ NA, & \text{otherwise} \end{cases} \tag{5}$$

The parameter $c \in \mathbb{N}$ can be set to allow only those grid points to be colored that have at least a specified number of $w_{m,n}^i > 0$, which means that a minimum of c residuals must have a relevant influence on that specific grid point. A larger c mitigates the effect of outliers but limits the representability of regions with lower density. After assigning a value $h_{m,n}$ to each point in the regular grid, the heat map can be generated by means of a defined color palette. To illustrate this approach, a third sample of 500 standard normally-distributed values represents the artificial residuals r (white noise) corresponding to each point (x, y) in the feature space. Figure 2 shows the result of this example.

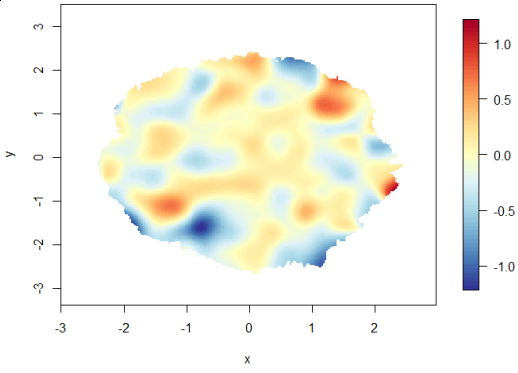


Figure 2. Heat map of artificial model residuals

The values of the residuals are visualized by means of a diverging color palette with 64 divisions. Red regions in the heat map represent high positive residuals while blue regions represent high negative residuals. The use of different color palettes depends on the application of the heat map and can be customized individually. However, it is not recommended to use a default rainbow palette as this can lead to misinterpretations or wrong conclusions (Borland and Taylor 2007) and color blind people would not be able to distinguish certain patterns. Wong (2011) provides a palette of eight colors, which is optimized for color blind individuals.

In this example no clear pattern is apparent because the residuals are indeed normally distributed over the whole data space. This should also be the case if the models work well in each area of the data space. In order to illustrate a case in which a specific region is biased, the previously defined residuals are now artificially manipulated by increasing the values lying within the range of -0.2 and 0.2 on the x-axis according to

$$r_i' = \begin{cases} r_i + 2, & \text{if } x_i \in [-0.2, +0.2] \\ r_i, & \text{otherwise} \end{cases}. \tag{6}$$

Figure 3 shows the resulting heat map with the artificially biased region. A clear pattern is now visible, as indicated by the red region, which serves as an indicator of possible problems regarding the ability of the model to explain certain phenomena. It is important to note that although the errors are positively biased, we have intentionally chosen a balanced color scale which is centered around zero to avoid an illusion of the error magnitude.

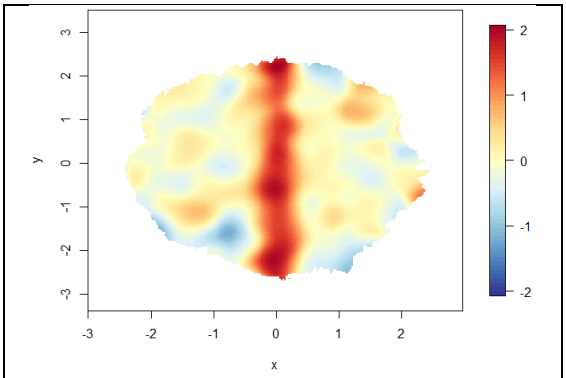


Figure 3. Bias in model residuals

The heat map approach permits a visualization of residuals in the same data space as two-dimensional scatter plots. Comparing scatter plots and heat maps can help to intuitively identify incorrectly specified models if the residuals follow a clear pattern depending on the underlying data. The following trivial but

comprehensible example introduces this idea, while subsequent sections discuss the benefits using real data. Fitting a linear regression to a polynomial function of the form

$$f(x) = x^2 + x + \varepsilon \tag{7}$$

where ε is a normally-distributed disturbance term illustrates how the heat map visualizes regions of good and bad performance. Figure 4 shows an example plot of this function. The fitted regression line is specified without a quadratic term and therefore unable to adequately explain the data. The corresponding residuals are visualized by the color scale.

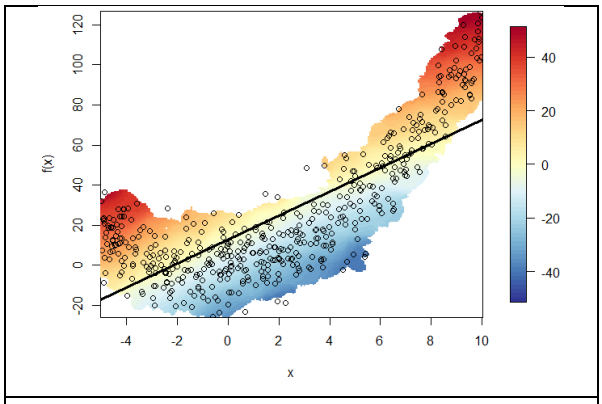


Figure 4. Residual heat map for linear regression

In order to demonstrate how this approach can help to identify and solve problems relating to machine learning models connected with feature engineering, we reproduce an artificial example provided by Heaton (2016) and evaluate the results based on the residual heat map. Heaton (2016) uses a variety of machine learning models to test their ability to learn the correct features from the presented data. All models seem to have problems (high RMSE) learning ratio-difference functions of the form

$$f(x_1, x_2, x_3, x_4) = \frac{x_1 - x_2}{x_3 - x_4}. \tag{8}$$

We use the same sampling approach as Heaton (2016) for each x_i to generate 2000 random numbers from a uniform distribution in the range [1,10] and train a feed-forward Artificial Neural Network (ANN) with two hidden layers containing 50 and 25 hidden neurons, respectively, based on the four features x_1 to x_4. ANNs serve as a machine-learning technique for classification and regression problems. They can efficiently learn non-linear patterns and also work well with noisy data (Schocken and Ariav 1994). Since the learning process and optimization are not the focus of this work, we refer to Bishop (1995) for further details.

The experiment confirms unusually high values of the RMSE. Figure 5 provides an overview of the evaluation results based on heat maps using the absolute values of residuals visualized by means of a sequential color palette with 64 divisions. Figure 5a clearly reveals that the model errors are not randomly distributed over the whole input data space but rather depend on the relationship between the features x_3 and x_4. The heat map approach is not solely limited to the use of existing features but also permits a visualization of the impact of new engineered features on the error distribution, thereby supporting problem identification. Figure 5b shows a plot of a new feature "$x_3 - x_4$" against an independent random variable, which backs up the obvious inference that the model fails to synthesize the function if the denominator tends towards zero (undefined). Other feature combinations shown in Figure 5c and 5d reveal no clear pattern. Even without knowing the underlying function, this technique enables the user to intuitively identify the local problem. This example shows that the high error is not driven by the capabilities of the machine learning model itself but is rather a question of the provided data (which in this case concerns a sampling approach). If obvious location-dependent patterns are evident, a context-based explanation of the problem and possible features that are worth to be included in the model are much easier to find. For the purpose of demonstrating possible insights using real-world data, an industry example based on a typical regression problem in practice is presented in the following section.

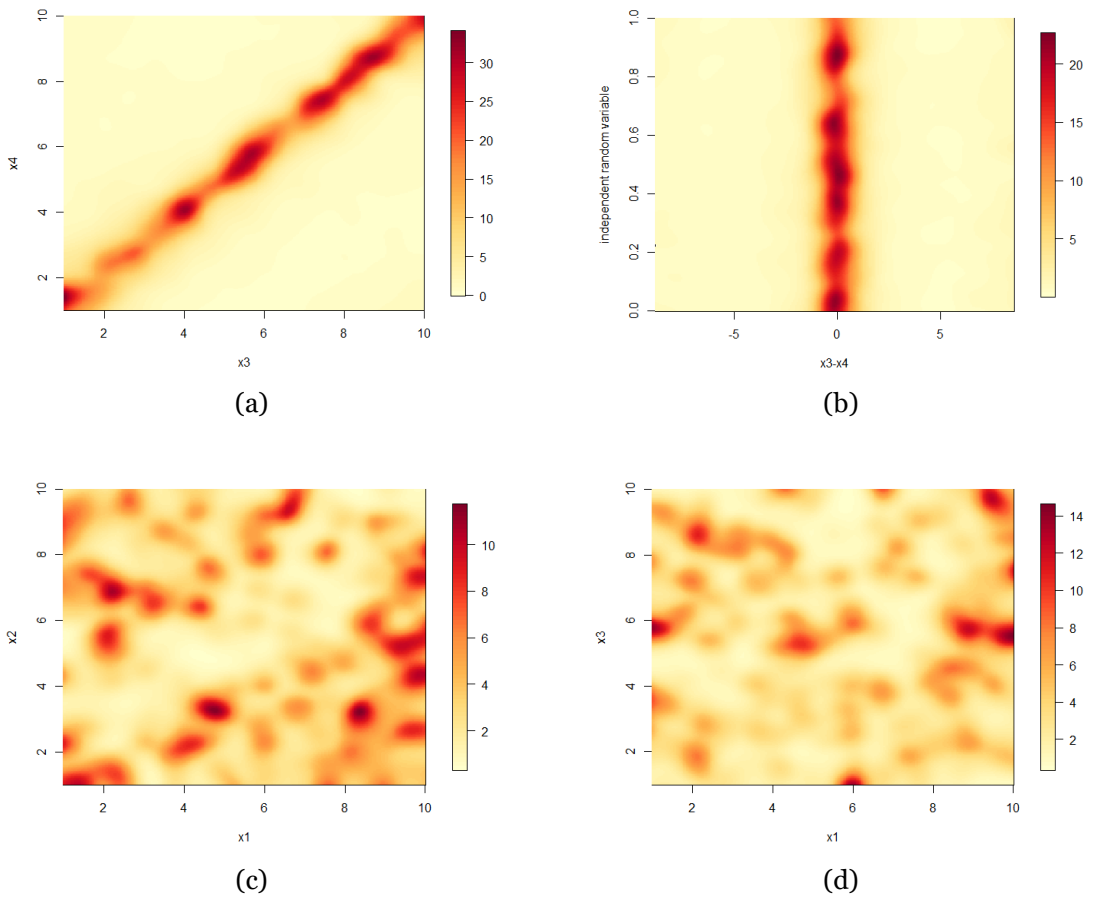


Figure 5 a-d. ANN model evaluation using residual heat maps

Industry Example

Problem Statement

This study is motivated by an industry application in cooperation with a large car manufacturer which operates a leasing business. The data analysis problem in this context concerns the question of how the achievable resale price of leasing cars can be estimated based on car characteristics and the leasing contract. A forecast of resale prices is especially important for specifying a proper leasing rate at the beginning of the leasing period. The leasing rate should compensate the expected loss in value of the car over the period of use. While the leasing rate is fixed in the leasing contract, the resale price after the leasing period remains uncertain and depends on different features such as the age of the car and the mileage. An accurate forecast is important because either a systematic overestimation or underestimation can have negative consequences. If the expected resale price is higher than the realized value on the used car market, the leasing rate is not sufficient to compensate the loss in value of the car. If in turn the expected resale price is lower than the realized value, the leasing rate is unnecessarily high, which induces avoidable competitive disadvantages. In our research we therefore address the resale price risk (Lessmann et al. 2010; Prado and Ananth 2012; Gleue et al. 2017) by implementing an operative decision support system within the company which automatically collects, stores and analyzes data on the leasing contracts and the respective cars. The data sets are labeled with the realized resale price of a specific car on the used car market. The application is thus a typical regression problem in which certain features such as age, mileage and several other car and leasing contract specific variables determine the realized resale price. Our analysis is currently based on

more than 250,000 completed leasing contracts over a time period from 2011/01 to 2016/12. Table 1 summarizes the available features.

Table 1. Initial Features

Feature	Type	Description
Accident free	categorical	Accident free (yes/no)
Color	continuous	Extra charge for special paintwork (in % of list price)
Customer	categorical	Customer type (end customer or reseller)
Distribution center	categorical	Location of the dealership (sales region/state)
Equipment	continuous	Extra charge for optional equipment (in % of list price)
Financing type	categorical	How the vehicle was financed (leased, financed, purchased)
Initial list price	continuous	Original, "historical" list price of the vehicle
Mileage	continuous	Vehicle mileage in km
Vehicle age	continuous	Vehicle age (registration date to resale date in days)
Engine capacity	continuous	Engine capacity in cubic centimeters
Engine power	continuous	Engine power in horsepower
Four wheel	categorical	Four-wheel drive indicator (yes/no)
Fuel type	categorical	Gasoline, Diesel, CNG, LPG
Gear number	categorical	Number of gears
Gear type	categorical	Transmission type: automatic, double-clutch, manual
Vehicle type age	continuous	Age of the vehicle type since market launch in days
Vehicle specifics	categorical	Other specifications of the vehicle type appearance
Resale price	continuous	Remaining value of the car at the end of the leasing period

In such a practical application there are many domain experts such as managers and decision-makers who have worked in the car leasing market for many years and understand its characteristics. We therefore use this case to demonstrate the feature engineering capabilities based on heat map visualizations. As the machine learning technique, we implement feed-forward ANN ensembles to learn the dependencies in the data. The next subsection describes the basic data pre-processing and analysis procedure.

Analysis Approach

Rather than algorithm optimization and the benchmarking of different machine-learning types, we document our solution for the given analysis problem by focusing on the iteratively identified difficulties and improvements of the models. A certain amount of basic pre-processing is performed in a first step. The categorical values from Table 1 are transformed into dummy variables while the continuous values are standardized according to the z-score

$$z = \frac{x - \mu}{\sigma} \tag{9}$$

in which μ denotes the mean and σ the standard deviation. In a second step a simple linear regression analysis is performed and the residuals are analyzed. Residual and QQ-plots show strong patterns for non-linear dependencies. By sorting the residuals with respect to time and calculating mean residuals for each month, we generated a time series of model errors and found evidence of seasonal patterns in the data. After controlling for all vehicle specific values, it was found that the seasonal pattern revealed higher prices on the used car market during spring and the lowest prices at the end of the year. Based on these observations, we introduced monthly dummy variables into the analysis. Although general macroeconomic factors may influence resale prices, they are not incorporated in the models because the forecasting

application requires the use of only those variables available at the time the forecast is performed. This means that the general market conditions or economic situation at the time the cars are actually resold are unknown in advance or need to be forecasted. In order to avoid a look-ahead bias we declined the use of these factors. After this basic pre-processing, we split the data into training and testing intervals. In this study the first three years from 2011/01 to 2013/12 are used for model-building and the subsequent three years from 2014/01 to 2016/12 are used for out-of-sample evaluation. Forecast horizons of up to three years are typical in this application for managing the risk of the current leasing portfolio. For the training data we perform an outlier analysis. Firstly, unrealistic values such as negative resale prices and resale prices which are higher than the initial price are removed. A pre-processing ANN is then trained on the entire training data and the 5% data sets with the highest in-sample error (absolute values) are removed. A manual analysis of the detected outliers reveals problems with the data quality regarding the overall condition of the car (e.g. falsely indicated accidents). The cleaned and processed training data set is now used for training ANN ensembles of 30 feed-forward ANNs (two hidden layers containing 50 and 25 neurons, respectively) with random weight initialization. The actual forecast is calculated as the average of the 30 results. In order to prevent the problem of overfitting, an early stopping approach is adopted. We thus split the training data into portions of 80% and 20%. The 80% portion is used for the iterative weight update process while the 20% portion serves as validation data for monitoring the out-of-sample error. The training process is terminated if the error on the validation set no longer decreases. After training the ANNs using past data, the models can be applied to previously unseen data in order to forecast future resale prices. There is much leeway for tuning the models further, e.g. by using ensembles of more and different model types, extending hyperparameter optimization and applying different regularization methods like dropout. In this study, however, we explicitly focus on the features and the data representation. This standard setup is thus used to illustrate the evaluation and adjustment process. In the following section the use cases of heat map visualizations are illustrated by way of three examples.

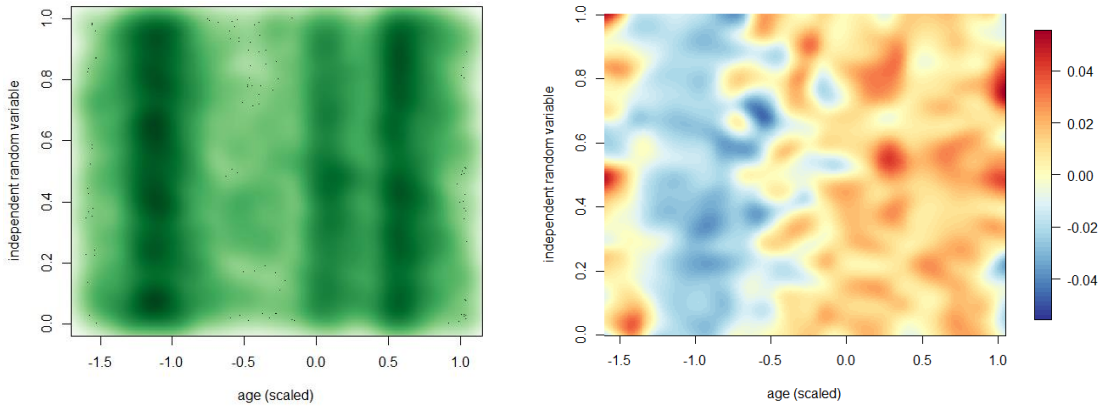
Examples of Feature Engineering

The analysis of resale prices in the presented form would already be of economically-relevant value for supporting decisions in negotiations on leasing rates and internal risk management applications. In this section we describe the analysis of the results on the out-of-sample data set and how possible problems and solution approaches for further improvements can be identified. We hence present three use cases for applying residual heat maps and explain the generated insights.

In a first step the resulting model is used for feature ranking in order to assess which of the inputs have the greatest influence on forecasting performance. For this purpose we use a perturbation ranking procedure (Breiman 2001). Using the out-of-sample data, the output of the ANN model is calculated n times, where n is the number of available features in the current model specification. In each calculation, one of these features is "destroyed" by randomly shuffling the values. By this means, the statistical properties of the feature such as the mean and standard deviation are retained while possible dependencies with the output value are eliminated. The importance of a feature is measured by the increase of an error measure (in this case the RMSE) when using the destroyed feature. This procedure shows that vehicle age is the most important feature in the current model followed by the mileage. In a next step the heat map visualization technique is applied to the most important continuous features by plotting them against an artificial variable, which is completely independent of all input and output values. This permits an observation of the effects on the error distribution of each feature individually. The residuals r_i are calculated as

$$r_i = y_i - \frac{1}{30}\sum_{j=1}^{30} ANN_j(X_i) \tag{10}$$

where y_i is the realized resale price of car i on the used car market and X_i is the feature vector for car i used to calculate the average of the 30 ANN models representing the prediction. In order to prevent outliers in the out-of-sample data from distorting the visual representation of possible systematic patterns, we remove the 5% highest errors before calculating the heat map. Figure 6 shows the result for the feature "age" (although the axes in this paper represent scaled values for anonymization, actual values should be used in practice for better comprehensibility). Figure 6a gives the distribution of data in the constructed feature space while Figure 6b shows the corresponding residual heat map.

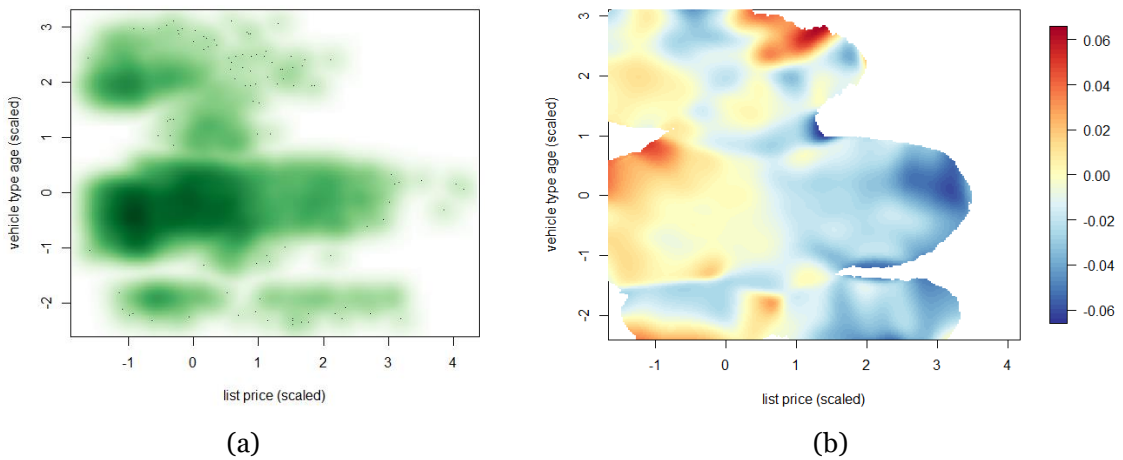


(a) (b)

Figure 6 a-b. Residual heat map of the feature "age"

The smoothed scatter plot (Figure 6a) indicates that there are three major clusters in the distribution of the feature "age". The corresponding residual heat map (Figure 6b) reveals a clear overestimation (blue region) of the resale price for the cluster of younger cars (scaled age values below -0.5). Internal discussions with domain experts reveal that the age of the cars in the identified region can be a proxy for a different market segment of short-leasing vehicles with different pricing policies. Incorporating previously unused information on short-leasing policies with additional dummy variables improves the results of the forecasting model significantly and in economically-relevant terms. The resulting heat map following model adjustment shows no relevant abnormalities and is comparable to the white noise simulation shown in Figure 2. All other analyzed continuous features also reveal no obvious deviations from the white noise example.

In a subsequent step, two features are always plotted against each other. The example presented in Figure 7 shows the bivariate distribution (Figure 7a) and the residual heat map (Figure 7b) of the features "list price" and "vehicle type age". Both are identified as belonging to the five most important features based on perturbation ranking.

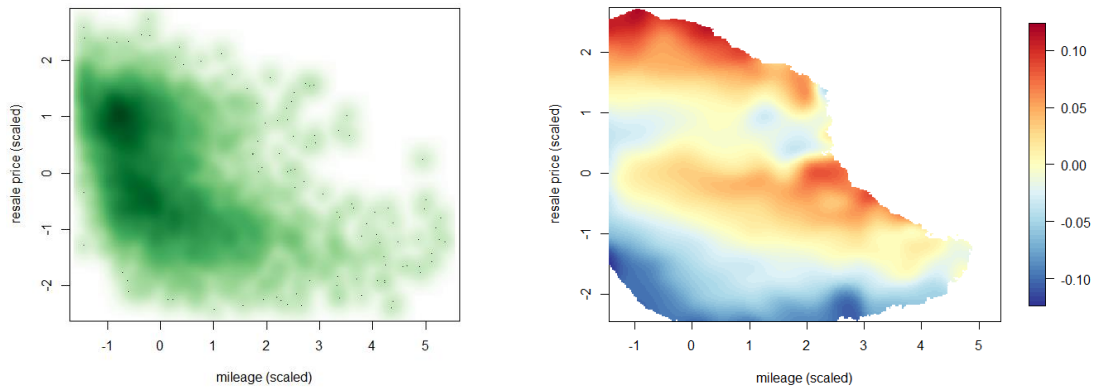


(a) (b)

Figure 7 a-b. Residual heat map of the features "vehicle type age" and "list price"

Again, a systematic pattern in the residuals can be observed. As indicated by the blue regions in Figure 7b, the ANN model systematically overestimates the resale price of cars with higher initial list prices. Based on this visual representation, a problem regarding data quality is identified. The ANN model uses features for special equipment prices. The initial list price, however, is a combination of the car price and the special equipment price. In discussions with domain experts it became clear that the special equipment price is not always reported correctly in the data base because dealers sometimes report only the combined price without additional information. Nevertheless, the feature for special equipment is reported to be zero if the car in fact has no special equipment, or simply if no information is available. Based on their practical experience, market dealers state that more special equipment leads to lower resale prices (in percentage terms of the initial price), which explains the overestimation of the ANN model in this region. Subsequent processing of the stored data so as to correctly represent the available information was found to improve the results of the model in economically-relevant terms. But even after accounting for this adjustment, a similar pattern in attenuated form was observed. A possible reason for this could be the manual input of this information. As this may be due to other reasons, however, this phenomenon is still under investigation. This example shows that the heat map can serve as a monitoring tool for continuously overseeing the data and model quality in conjunction with new incoming data.

A third application of the heat map visualization is to obtain an intuitive view of the dependencies between an input feature and the dependent variable. The most important continuous features are hence plotted against the resulting car resale prices. Figure 8 shows a smoothed scatter plot (Figure 8a) and the residual heat map (Figure 8b) with the dependent resale price plotted against car mileage.



(a) (b)

Figure 8 a-b. Residual heat map for the feature "mileage" plotted against the actual resale price

As expected, Figure 8b confirms that the upper (red) and lower (blue) regions in the data space are interspersed with higher errors. This can be due to simple outliers in the data or data quality problems. In the case of the blue region, a possible explanation is the incorrect documentation of (smaller) car accidents, which tends to decrease the resale price. Since accidents are labeled manually, there is a lot of room for errors in the data. Discussions on the upper red region in the data space also provide a possible explanation of unexpected high resale prices. Experience shows that in this particular vehicle segment it is common to install additional extras in cars after the leasing contract has been specified. Although additional extras are currently not documented in the analyzed data set, they tend to raise the total value of the car. This results in an unexpectedly high resale price. For further analyses, these additional extras can now be explicitly monitored by dealers to include them as an additional feature for future forecasts. The observation of the red region in the proximity of zero of the scaled resale price values (representing an underestimation, i.e. the resale prices are higher than expected from the models) is puzzling. This pattern is first observed after a certain number of kilometers driven and then tends to increase in magnitude. A hypothesis generation reveals that in rural areas the competition among vehicle dealers is much lower and experience shows that especially in these regions, cars with a higher mileage can achieve higher resale prices. Another possible

reason for this pattern could be maintenance measures, which are performed after a certain mileage and can increase the value of a car. Although these generated plausible hypotheses form a basis for further investigations along the described paths, they cannot be confirmed at the present time. Nevertheless, they can provide an indication of the direction to follow in future research. The heat map technique does not claim to visualize the truth, but rather to trigger new ideas among experts for targeted investigations/solutions.

Discussion and Outlook

The presented examples in this study show the benefits of visual model evaluation for incorporating domain knowledge in the iterative data science process. The proposed heat map approach provides an intuitive view on the performance of machine learning models. The model error can be visualized as a color scale in the same two-dimensional data space as simple scatter plots. Systematic errors or location-dependent error clusters can be easily revealed and possible reasons can be discussed with data scientists and domain experts on the same level of complexity. While feature engineering is mostly a very complex and time-consuming task without the necessary domain knowledge about the underlying data, discussions and exchanges of ideas between data scientists and domain experts can facilitate the process of finding the right features to explain certain observations of model behavior. Future data collection can also benefit from the generated insights. The industry example shows how the method can be used in a practical application to intuitively understand the results of a model. The method enables data scientists to pose easily comprehensible questions about complex models which can be answered on the basis of correct domain knowledge. In our view, an understanding of the results, an awareness of the model performance in specific regions and participation in the development process are crucial for winning the trust of decision-makers and their acceptance of the resulting data products/decision support systems. While this study focuses on the visualization method and its application in practice, further research is necessary to underpin this statement and measure the actual benefits regarding trust, acceptance and the overall outcome of model performance. Qualitative and quantitative studies in different industry applications should be conducted for this purpose. Further investigations are necessary to understand how the visual appearance of the proposed heat map visualization influences the discussions and hypothesis generation. While the literature highlights the disadvantages of a rainbow color scale (Borland and Taylor 2007), our experience show that the resulting heat maps are perceived as far more intuitive and appealing among domain experts. Despite the disadvantages of rainbow colors, it is worth a comparative study. It would also be interesting to carry out studies on classification problems in order to assess the comprehensibility and interpretability of the results in this context. In our examples we limited our attention to regression problems. For classification purposes, however, the binary model error cannot be directly used on a color scale. It would be interesting to know whether the results (for example, weighted by the loss function) are also intuitive for the user in the same way as the concrete errors from a regression problem. Besides the communication aspect between data scientists and domain experts, the method is also suitable for big data model diagnostics in general. Compared to residual plots, the appearance of the heat map is independent of the sample size. It is possible to identify regions in which model errors deviate systematically from a white noise depending on the location in a two-dimensional data space. Performance indicators such as the RMSE reduce the information about model quality to a single number and therefore fail to account for this important component of information. The described approach can also be used to carry out continuous quality control, in particular regarding changes in the data, e.g. in order to recognize when a new training of the models becomes necessary. The time factor is essential for this purpose and represents an interesting application for residual heat maps due to the fact that the technique can easily incorporate time as a potential feature on one of the axes.

The proposed technique also suffers from various difficulties and drawbacks. Visualization in general is limited to perceptible dimensions. The heat map is limited to the representation of two variables on the coordinate axes and the corresponding errors on a color scale. On the other hand, most machine learning applications implement a high-dimensional feature space. We try to limit the problem by first calculating feature importance and then plotting the most valuable features against each other. The representation as a scatter plot matrix with corresponding residual heat maps can also help to provide a better overview. Another serious limitation is the inability to correctly visualize categorical features. Moreover, continuous values must have a reasonably broad distribution in the data space in order to permit a meaningful error representation by means of a heat map approach. Categorical features can be incorporated indirectly by

splitting the data based on different categories and comparing the heat maps on the different sub-samples. Heat map comparisons face several drawbacks, however, because the color scale is relative and therefore only possible visual patterns can be compared rather than model performance or different data in general. Previous research by Cleveland and McGill (1984) also shows that the elementary perceptual task of color saturation perception is one of the least accurate when people extract quantitative information from graphs. This underlines the problem when using heat maps for quantitative model comparisons. In addition, the current approach does not visualize the density of available data points directly in the heat map. Hence, the results can be misleading or may lead to wrong decisions and assessments regarding overall model quality. On the other hand, it is also possible to overlook systematic errors, for example, if the errors fluctuate with a high frequency but kernel smoothing covers these patterns. These limitations can thus result in a general criticism of the approach itself. The aim of the visual representation of model errors is not a statistically valid assessment of models per se but rather a stimulus to trigger human creativity when assessing the model results. Quantitative model comparisons should further be carried out using standard approaches such as the RMSE, etc., which cannot be replaced by the proposed method. Used in a responsible way, it may be argued that the approach is of some value in the creative aspect of data science for explaining error phenomena in the data and identifying features that explain these patterns. For this reason, the parameters must be set correctly by an experienced user because the technique in its current version is not a plug and play solution and incorrect parameters can significantly influence the visual appearance of the heat map. Future work should concentrate on the robustness of the method and further develop the approach to make it easier to use in practice by automatically setting correct parameters based on the available data.

In addition to the provided R code, we are currently developing a web-based application to make the functionality accessible to a broad user group. This tool will be available on GitHub (https://github.com/eilersde) for further developments such as the incorporation of other visualization techniques, especially for residual analysis, evaluations, applicability checks in different domains and as a possible basis for future studies focused on trust and acceptance. In future research we will concentrate on a design-oriented development of an extended version of the tool towards business needs such as the live updating of error dashboards for quality monitoring with new incoming data. This might include e.g. an early-warning system for new patterns or market developments and more interactive graphics offering zoom and data selection options. In order to ensure a real-time interaction with the heat map, the performance of the calculations needs to be improved in future work. It is also planned to further test the utility of the presented technique by means of controlled experiments with different user groups. Generally speaking, however, easily comprehensible error visualizations and tools are unable to explain the complex black-box model underlying the analyses and hence general reservations and mistrust may still remain.

Conclusion

> "We can tell stories with graphics, but we can also let people build their own stories with them."
> – Alberto Cairo (Cairo 2016)

In this study we address the question of how to make better use of knowledge from domain experts for improving the quality of data analysis models. In typical industry applications we identify a gap between data scientists, who are skilled experts in conducting analytics tasks, and domain experts, who understand business problems and have valuable experience in their field. As a step towards improving this situation, we introduce a visualization technique for model evaluation based on heat maps. The familiar and easily comprehensible concept of heat maps facilitates improved communication between data scientists and domain experts regarding the quality of the analysis models on the same level of complexity. Insights from a real-world industry example show how the visualization technique can help to incorporate domain knowledge during the iterative process of model construction, evaluation and adjustment. The improvements are mainly based on an identification of systematic problems in the data space and possible explanatory features to address the observed phenomena. The heat map approach provides an intuitive context for model performance/errors and serves as a basis for discussion among different expert groups. There are many challenges facing further research regarding the proposed heat map visualization and the general question of how to better integrate different expert groups for successful data science in practice. By providing the code package, explanatory examples and a first easy-to-use tool, we hope to contribute to the scientific community and encourage research in the further development of solutions or the evaluation of ideas in different contexts.

Acknowledgements

We thank Julian Hamann, Nikolas Stege, Ian Westwood and the anonymous reviewers for their valuable comments on our manuscript. Jean-Henrick Schünemann provided excellent research assistance. We also thank the car manufacturer for providing the data and all its employees involved for the cooperation.

References

Agarwal, R., and Dhar, V. 2014. "Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research," Information Systems Research (25:3), pp. 443–448.

Al-Kassab, J., Ouertani, Z. M., Schiuma, G., and Neely, A. 2014. "Information visualization to support management decisions," International Journal of Information Technology & Decision Making (13:02), pp. 407–428.

Alpar, P., and Schulz, M. 2016. "Self-Service Business Intelligence," Business & Information Systems Engineering (58:2), pp. 151–155.

Bengio, Y., Courville, A., and Vincent, P. 2013. "Representation Learning: A Review and New Perspectives," IEEE Transactions on Pattern Analysis and Machine Intelligence (35:8), pp. 1798–1828.

Bishop, C. M. 1995. *Neural networks for pattern recognition*, Oxford university press.

Borland, D., and Taylor, R. M. 2007. "Rainbow Color Map (Still) Considered Harmful," IEEE Computer Graphics and Applications (27:2), pp. 14–17.

Breiman, L. 2001. "Random Forests," Machine Learning (45:1), pp. 5–32.

Bresciani, S., and Eppler, M. J. 2009. "The Benefits of Synchronous Collaborative Information Visualization: Evidence from an Experimental Evaluation," IEEE Transactions on Visualization and Computer Graphics (15:6), pp. 1073–1080.

Brooks, M., Amershi, S., Lee, B., Drucker, S. M., Kapoor, A., and Simard, P. 2015. "FeatureInsight: Visual support for error-driven feature ideation in text classification," IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 105–112.

Buehler, K. S., and Pritsch, G. 2003. "Running with risk," McKinsey Quarterly (4), pp. 40–49.

Cairo, A. 2016. *The Truthful Art: Data, Charts, and Maps for Communication*, New Riders.

Carr, D. B., Littlefield, R. J., Nicholson, W. L., and Littlefield, J. S. 1987. "Scatterplot Matrix Techniques for Large N," Journal of the American Statistical Association (82:398), pp. 424–436.

Cleveland, W. S., and McGill, R. 1984. "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," Journal of the American Statistical Association (79:387), pp. 531–554.

Conway, D. 2010. "The Data Science Venn Diagram," (available at https://goo.gl/KvYVx5; retrieved April 20, 2017).

Domingos, P. 2012. "A Few Useful Things to Know About Machine Learning," Commun. ACM (55:10), pp. 78–87.

Dumbill, E., Liddy, E. D., Stanton, J., Mueller, K., and Farnham, S. 2013. "Educating the Next Generation of Data Scientists," Big Data (1:1), pp. 21–27.

Eilers, D., and Breitner, M. H. 2017. "A Picture is Worth a Thousand Words: Visual Model Evaluation in Data Science Applications," Wirtschaftsinformatik 2017 Proceedings.

Endert, A., Hossain, M. S., Ramakrishnan, N., North, C., Fiaux, P., and Andrews, C. 2014. "The human is the loop: new directions for visual analytics," Journal of Intelligent Information Systems (43:3), pp. 411–435.

Fransecky, R. B., and Debes, J. L. 1972. *Visual Literacy: A Way to Learn–A Way to Teach*, Washington, D.C.: Association for Educational Communications and Technology.

Franz, M., Scholz, M., and Hinz, O. 2015. "2D versus 3D Visualizations in Decision Support – The Impact of Decision Makers' Perceptions," ICIS 2015 Proceedings.

Ghiassi, M., Zimbra, D., and Lee, S. 2016. "Targeted Twitter Sentiment Analysis for Brands Using Supervised Feature Engineering and the Dynamic Architecture for Artificial Neural Networks," Journal of Management Information Systems (33:4), pp. 1034–1058.

Gleue, C., Eilers, D., von Mettenheim, H.-J., and Breitner, M. H. 2017. "Decision Support for the Automotive Industry: Forecasting Residual Values using Artificial Neural Networks," Wirtschaftsinformatik 2017 Proceedings.

Harvey, D. Y., and Todd, M. D. 2015. "Automated Feature Design for Numeric Sequence Classification by Genetic Programming," IEEE Transactions on Evolutionary Computation (19:4), pp. 474–489.
Heaton, J. 2016. "An empirical analysis of feature engineering for predictive modeling," in SoutheastCon 2016, pp. 1–6.
Keim, D. A. 2002. "Information visualization and visual data mining," IEEE Transactions on Visualization and Computer Graphics (8:1), pp. 1–8.
Keim, D. A., Panse, C., Sips, M., and North, S. C. 2004. "Visual data mining in large geospatial point sets," IEEE Computer Graphics and Applications (24:5), pp. 36–44.
Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., and Melançon, G. 2008. "Visual Analytics: Definition, Process, and Challenges," in Information Visualization, Lecture Notes in Computer Science, A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North (eds.), Springer Berlin Heidelberg, pp. 154–175.
Kelleher, C., and Wagener, T. 2011. "Ten guidelines for effective data visualization in scientific publications," Environmental Modelling & Software (26:6), pp. 822–827.
Klemm, P., Lawonn, K., Glaßer, S., Niemann, U., Hegenscheid, K., Völzke, H., and Preim, B. 2016. "3D Regression Heat Map Analysis of Population Study Data," IEEE Transactions on Visualization and Computer Graphics (22:1), pp. 81–90.
Köpp, C., Mettenheim, H.-J. von, and Breitner, M. H. 2014. "Decision Analytics with Heatmap Visualization for Multi-step Ensemble Data," Business & Information Systems Engineering (6:3), pp. 131–140.
Krause, J., Perer, A., and Bertini, E. 2016. "Using Visual Analytics to Interpret Predictive Machine Learning Models," arXiv:1606.05685 [cs, stat] (available at http://arxiv.org/abs/1606.05685).
Kuhn, M., and Johnson, K. 2013. *Applied Predictive Modeling*, Springer Science & Business Media.
Lash, M. T., and Zhao, K. 2016. "Early Predictions of Movie Success: The Who, What, and When of Profitability," Journal of Management Information Systems (33:3), pp. 874–903.
Lessmann, S., Listiani, M., and Voß, S. 2010. "Decision Support in Car Leasing: A Forecasting Model for Residual Value Estimation," ICIS 2010 Proceedings.
Markovitch, S., and Rosenstein, D. 2002. "Feature Generation Using General Constructor Functions," Machine Learning (49:1), pp. 59–98.
Power, D. J. 2008. "Decision Support Systems: A Historical Overview," in Handbook on Decision Support Systems 1, International Handbooks Information System, Springer Berlin Heidelberg, pp. 121–140.
Prado, S., and Ananth, R. 2012. "Breaking Through Risk Management, a Derivative for the Leasing Industry," Journal of Financial Transformation (34), pp. 211–218.
Schocken, S., and Ariav, G. 1994. "Neural networks for decision support:: Problems and opportunities," Decision Support Systems (11:5), pp. 393–414.
Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., and Carlsson, C. 2002. "Past, present, and future of decision support technology," Decision Support Systems (33:2), pp. 111–126.
Sun, G.-D., Wu, Y.-C., Liang, R.-H., and Liu, S.-X. 2013. "A Survey of Visual Analytics Techniques and Applications: State-of-the-Art Research and Future Challenges," Journal of Computer Science and Technology (28:5), pp. 852–867.
Tam, G. K. L., Fang, H., Aubrey, A. J., Grant, P. W., Rosin, P. L., Marshall, D., and Chen, M. 2011. "Visualization of Time-Series Data in Parameter Space for Understanding Facial Dynamics," Computer Graphics Forum (30:3), pp. 901–910.
Tam, G. K. L., Kothari, V., and Chen, M. 2017. "An Analysis of Machine- and Human-Analytics in Classification," IEEE Transactions on Visualization and Computer Graphics (23:1), pp. 71–80.
Thomas, J. J., and Cook, K. A. 2006. "A visual analytics agenda," IEEE Computer Graphics and Applications (26:1), pp. 10–13.
Wong, B. 2011. "Points of view: Color blindness," Nature Methods (8:6), pp. 441–441.
Yu, H.-F., Lo, H.-Y., Hsieh, H.-P., Lou, J.-K., McKenzie, T. G., Chou, J.-W., Chung, P.-H., Ho, C.-H., Chang, C.-F., Wei, Y.-H., Weng, J.-Y., Yan, E.-S., Chang, C.-W., Kuo, T.-T., Lo, Y.-C., Chang, P. T., Po, C., Wang, C.-Y., Huang, Y.-H., Hung, C.-W., Ruan, Y.-X., Lin, Y.-S., Lin, S., Lin, H.-T., and Lin, C.-J. 2010. "Feature Engineering and Classifier Ensemble for KDD Cup 2010," KDD Cup and Workshop 2010 Proceedings.