Hybrid Intelligence with Commonality Plots: A First Aid Kit for Domain Experts and a Translation Device for Data Scientists

Nikolas Stege[1,2], Michael H. Breitner[1]

[1] Leibniz Universität Hannover, Information Systems Institute, Hannover, Germany
{stege,breitner}@iwi.uni-hannover.de
[2] Ernst & Young GmbH, Hannover, Germany
{nikolas.stege}@de.ey.com

Abstract. There is a large gap between domain experts capable to identify business needs and data scientists who use insight producing algorithms, but often fail to connect these to the bigger picture. A major challenge for companies and organizations is to integrate practical data science into existing teams and workflows. We are driven by the assumption that efficient data science requires cross-disciplinary teams able to communicate. We present a methodology that enables domain experts and data scientists to analyze and discuss findings and implications together. Motivated by a typical problem from auditing we introduce a visualization method that helps to detect unusual data in a subset and highlights potential areas for investigation. The method is a first aid kit applicable regardless whether unusual samples were detected by manual selection of domain experts or by algorithms applied by data scientists. An applicability check shows how the visualizations facilitate collaboration of both parties.

Keywords: Commonality Plots, Domain Knowledge, Hybrid Intelligence, Visualization, Data Science

1 Introduction and Motivation

Referring to the term data science the spotlight is usually put on the application of state of the art models, machine learning algorithms, on how to tune the algorithms' hyperparameters and on how to optimize scalability and overall performance. Those are all very important aspects in the world of data analytics. However, especially in the context of business intelligence and the overall creation of value there is large potential for information systems researchers and practitioners in the field of data science besides the optimization of algorithms, see [1-2] for an overview. Efficient data science requires a skillset that covers the range from the engineering side (data capturing and processing) to the business side (domain expertise and storytelling) and it comes to no surprise that individuals with such cross-disciplinary skillsets are rare [3]. Because of that shortage, the biggest challenge is to figure out how to efficiently

15th International Conference on Wirtschaftsinformatik,
March 08-11, 2020, Potsdam, Germany

https://doi.org/10.30844/wi_2020_c7-stege

integrate practical data science into existing teams, workflows and processes. We are guided by the research question of how to improve the communication and collaboration between data scientists (engineering side) and domain experts (business side). As this question is broad, we approach it bottom-up with a specific problem motivated from the audit domain, namely the identification and assessment of unusual items. We introduce a visualization method (commonality plots) that supports the interpretation of unusual items, regardless of how these items have been identified in the first place. For the applicability check we use a structured dataset provided for auditing purposes to show how commonality plots can facilitate the dialogue of data scientists and auditors. However, our objective is to provide a method that is not restricted to auditing, but can be applied to many domains where structured data is available for investigation. In the end, every result of a pattern recognition algorithm will be some kind of subset of an overall population. We therefore argue that a visualization method capable of examining samples from structured datasets is suitable for adaption to other domains without undue effort.

The next section contains a detailed description of the problem statement as well as its localization in a stylized process. Section 3 contains the theoretical background with definitions for commonality and the likelihood of occurrence of commonality. Section 4 contains the applicability check that demonstrates how to visualize and interpret the commonality measures. In Section 5 we discuss results of our approach and outline potential areas for further research; Section 6 concludes.

2 Problem Definition

There is a large body of literature on how the quality of an audit is critically dependent on the auditors' judgment and the related derivation of conclusions regarding the financial statements of companies and organizations [4]. Accordingly, audit efficiency highly depends on the auditors' competency "in recognizing patterns in financial data and in hypothesizing likely causes of those patterns to serve as a guide for further testing" and investigation [5]. In the light of unprecedented computational power and the transformational nature of advanced technologies and analytics in general the auditing sector is facing a change in paradigms [6]. It has become ever more challenging to gain insight from the vast volumes of structured and unstructured data available in order to assure a high quality audit not only based on samples but on entire company transactions. Thus, the accounting companies have "invested heavily in technological innovation" and personnel [7]. In order to unlock the investments' potential, a common level of communication needs to be established for the experienced auditors and the new wave of tech-savvy employees. This is a major challenge not only for the auditing and accounting sector. It can be generalized in terms of how to successfully integrate practical data science and methodologies into the common working world. This makes the overall facilitation of communication and collaboration between domain experts and data scientists a highly relevant field for information systems research [8].

https://doi.org/10.30844/wi_2020_c7-stege

Figure 1 lists typical tasks to be performed by an auditor from a data perspective. The idea is to lay out the general process without any links to specific audit procedures that obscure the broader picture with too much detail.

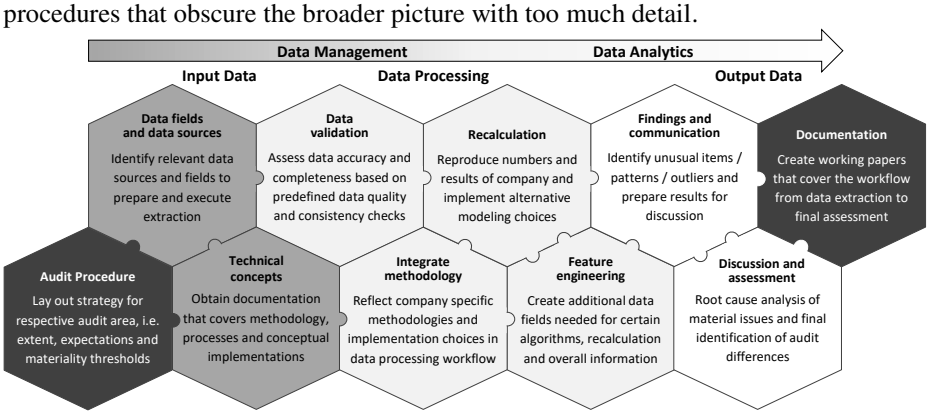

Figure 1. Stylized Data-Driven Audit Process

Following [9-10] the general process of extracting insights from data can be broken down into two main sub-processes: data management (acquisition, validation and enrichment) and data analytics (analyses and interpretation). We suggest to split the data management part into input data and data processing as both can account for months of work on their own. Plus, for auditing purposes, the process needs to be extended to account for documentation, often the most important part as a proper audit is completely relying on audit evidence. In general, the skillset of a data scientist "should be interdisciplinary and cover critical analytical and IT skills, business and domain knowledge, and communication skills required in a complex data-centric business environment" [2]. While such a skillset is certainly beneficial for each step of the process outlined in Figure 1, our method focuses on its rear part. We assume that data acquisition, validation, feature engineering and recalculation have all been finished so that our starting point is a perfectly preprocessed dataset. Additionally, let's assume that we already have identified a number of samples that potentially contain unusual items. At this point, we are left with questions like the following:

- Do the items in the sample have something in common that is not directly apparent?
- If there is a common pattern, how likely is its occurrence?
- Is the pattern explainable, and if not, is it worth spending time on further investigation or on the extension of the sample?

While these are important questions that arise during an audit, they will most likely also be relevant to other domains. Regardless whether the identification of items was conducted manually by a domain expert based on experience (educated guess) or by a data scientist with the help of pattern recognition algorithms (algorithmic guess), these questions are worth considering for both parties alike.

https://doi.org/10.30844/wi_2020_c7-stege

In order to answer these questions we introduce a method that visualizes commonalities in subsets and highlights items that have a low likelihood of occurrence. Visualization in general helps to facilitate all phases from the description, collection and processing of information to comprehension and knowledge discovery. The use of interactive visualization has proven to facilitate communication and significantly increase individual learning as well as overall team performance [11]. Following [12] visualization is generally defined as the visual representation of information in order to enable communication and exploration. Moreover, when visualizing data the intention must not be to merely draw a pretty picture and leave it as is, but instead create something that enables people to extract their own insights and conclusions from what the data is telling [13]. When configured properly, visualization can be an efficient tool for making complex interrelationships intuitively understandable for users. Our goal is hence to provide plots that put domain experts and data scientists on the same level regarding the interpretation and discussion of results, insights and implications. This facilitates the dialogue and collaboration between both parties (hybrid intelligence). Last but not least, to cover the whole process in Figure 1, visualizations are very useful for documentation.

3 Towards the Visualization of Commonality

Our starting point is a dataset that has already gone through all the necessary preprocessing steps. Compare Figure 1, the dataset has been extracted, validated and enriched. Let D denote such a preprocessed dataset with an arbitrary number of columns c and rows r. Additionally, from D we have obtained a subset d, that is $d \subset D$ with $c(d) = c(D)$ and $r(d) < r(D)$. We now want to define a measure that tells us if there is something that the items in d have in common, and if so, how unusual the commonality is. Figure 2 illustrates our framework of the approach. The measures and notations are described in the following sections. On the upside, unless the entire dataset is altered, the measures only have to be calculated once to separately store them. Then, regardless the subset, the measures can always be loaded and reapplied. Subject to the size of the subsets, this makes the approach highly scalable. On the downside, our approach is not a one-size-fits-all measure, but differs depending on the data type of the data fields. In statistics levels of measurement are typically divided into four types: nominal, ordinal, interval, and ratio data [14]. Here, we start with the distinction between categorical and continuous data and will later on discuss how to adjust for a more granular differentiation.

https://doi.org/10.30844/wi_2020_c7-stege

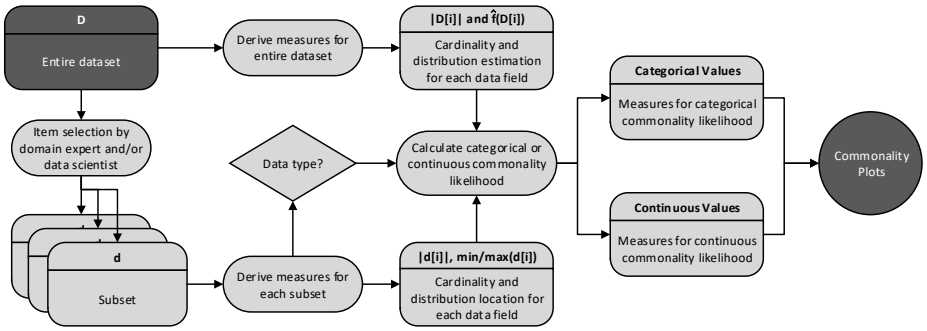

Figure 2. Commonality Framework

3.1 Categorical Values

In general, categorical data can only take specific values that are clearly separable [14]. More precisely, the following definitions and arguments apply to Booleans, integers, characters and strings.

Definition of Categorical Commonality. Within our framework (Figure 2), we use the concept of cardinality as the basis for the definition of categorical commonality. Cardinality describes the number of different elements in a dataset. Let $|d[i]|$ denote the cardinality of a certain data field i in subset d, where $i = 1, \dots, c(D)$. Then, the minimum cardinality equals 1 if all the values in $d[i]$ are identical and the maximum is equal to the number of rows of the subset $r(d)$ if all values in $d[i]$ are different:

$$|d[i]| \in \{x \in \mathbb{N} |\ 1 \leq x \leq r(d[i])\}, \quad i = 1, \dots, c(D). \tag{1}$$

From (1), we say a data field of the subset has commonality if the cardinality of the data field in the subset is equal to 1 while its cardinality with regards to the original dataset exceeds 1. That is to say, commonality exists if the following holds:

$$\text{Commonality} \coloneqq |d[i]| = 1 \ \wedge |D[i]| > 1, \quad i = 1, \dots, c(D). \tag{2}$$

To summarize, we take the whole dataset, count the unique elements within a categorical data field and repeat the procedure for the subset. We then compare the resulting numbers: if the subset only holds a single unique element while the whole dataset holds more than one unique element, we say there is commonality in the subset. We include the constraint in (2) due to the fact that a cardinality of 1 in the whole dataset will undoubtedly result in a cardinality of 1 for any subset, regardless of the composition. Such causality is obvious and there is no need to visualize the obvious.

Definition of Categorical Commonality Likelihood. The intention is to find a likelihood measure for the occurrence of commonality. The measure must be close to zero in case the observed commonality is highly improbable and close to one in case commonality is most likely. Such a measure can be derived from the multivariate

https://doi.org/10.30844/wi_2020_c7-stege

hypergeometric distribution that is used in combinatorics. The distribution can be described with the typical example of drawing balls of different color from an urn without replacement: let there be K_i balls of color i in the urn and there are c different colors in total, so that the urn contains a total of $N = \sum_{i=1}^{c} K_i$ balls. If we take out n balls at random without replacement, it follows that the number of balls k_i of each color $i = 1, \dots, c$ in the resulting sample has the multivariate hypergeometric distribution. Consequently, the probability P_c for obtaining a specific composition of colored balls in the sample can be derived as follows [15-16]:

$$P_c(X_1 = k_1 \wedge X_2 = k_2 \wedge \cdots \wedge X_c = k_c) = \frac{\binom{K_1}{k_1} \cdot \binom{K_2}{k_2} \cdot \ldots \cdot \binom{K_c}{k_c}}{\binom{N}{n}}. \quad (3)$$

Let us now change the urn and ball terminology by reinterpreting the sample as our subset and the colors as the cardinality of a data field in the whole dataset. Then, N is its number of rows and n is the number of rows of the subset. Furthermore, let k_j denote the value count in a subset's data field that shows commonality as defined in (2). Then, the probability in (3) reduces to the likelihood measure for the occurrence of commonality we were looking for:

$$P_1\big(X_j = k_j \mid \{X_i = 0, \forall\, i \neq j\}\big) = \frac{\binom{K_j}{k_j}}{\binom{N}{n}} \Rightarrow \frac{\binom{K_j}{n}}{\binom{N}{n}}. \quad (4)$$

To summarize, if a categorical data field shows commonality as in (2), all values in the data field are identical. We count how often the value appears in the whole data set (K_j) and how often the value appears in the subset (k_j). As the subset's data field only contains identical values, it follows that k_j is equal to the sample size n. In combination with the size of the whole dataset N we set up the binomial coefficients for numerator and denominator in (4). For the properties of binomial coefficients see for example [17]. The resulting measure yields values close to zero in case it is highly improbable to end up with commonality (the larger the difference between N and K) and values close to one in case commonality is most likely (the smaller the difference between N and K).

3.2 Continuous Values

Unlike categorical data types, continuous data can take any value within a specific range for which the intention regarding measurement is usually not based on counting [14]. In this context, the following definitions and arguments apply to floating point numbers.

Definition of Continuous Commonality. Unfortunately, the concept of cardinality in (2) is not an applicable measure for floating point numbers. This is due to the fact that cardinality describes the number of different elements in a dataset whereas floats have no clear boundaries within their value range, and as stated, are not for counting. Thus, for floats we turn towards the distribution of the values within a data field that holds

https://doi.org/10.30844/wi_2020_c7-stege

continuous data. Let f denote the probability density function of the distribution of such a data field in dataset $D[i]$, so that

$$\int_{-\infty}^{\infty} f(D[i])dD[i] = 1. \quad (5)$$

Following [18-19], we can use kernel density estimation to find an estimate for the density function in (5) from the observed data $D[i]$. The kernel estimator is defined by

$$\hat{f}(D[i]) = \frac{1}{nh} \cdot \sum_{j=1}^{n} K\left(\frac{D[i]-X_j}{h}\right), \quad (6)$$

where parameter h is the bandwidth that describes the width of kernels K that are created for each observed value within $D[i]$ and are then added up to form the estimated density curve. Now, we simply define continuous commonality as a state in which the value range of the subset is much narrower than the value range regarding the entire dataset. That is, in reference to (6) we say that continuous commonality exists, if all values in $d[i]$ lie close to each other when projected on $\hat{f}(D[i])$. As this is a heuristic approach we do not provide a more precise definition and leave it to the user to decide what is close enough for commonality and what is not.

Definition of Continuous Commonality Likelihood. As with categorical data, the intention is to find a likelihood measure for the occurrence of commonality. Since $\hat{f}$ in (6) is an estimator for the density function, we can use it to derive probabilities. More precisely, we are looking for the probability of the value range of the subset, that is described by the interval $[min(d[i]), max(d[i])]$. The probability measure for continuous commonality then results from inserting these interval limits into (5):

$$P(min(d[i]) \leq X \leq max(d[i])) = \int_{\min(d[i])}^{\max(d[i])} \hat{f}(D[i])dD[i]. \quad (7)$$

As desired, the expression describes a measure that takes values between 0 for single or identical values and 1 in case the value range of the subset equals the range of the entire dataset.

In summary, the overall intention of the defined measures is not to be exact, but to allow for comparability of the various data fields in a subset with respect to the likelihood of the data fields' value occurrence. That is, the measures in this section must be seen as guidance for focusing discussions and interpretations. Please note that we will use the complementary probability for the visualizations to the described measure in (4) as we are interested in the unusualness of the occurrence of certain patterns.

3.3 Challenges and Minor Adjustments

Our following listing is not exhaustive, but highlights key issues that might complicate visualization and interpretation.

https://doi.org/10.30844/wi_2020_c7-stege

Missing Values. Following [20], missing data is defined as unobserved values that could have been meaningful for further analyses, if observed. When dealing with missing data, it is particularly important to find out why values are missing in order to rely on potential conclusions drawn from the data. Missing data can derive from, for example, non-response, sensor failure, high level of noise, or simply unknown information [21]. However, often the reason for the occurrence of missing values might be unclear. In that case, we suggest to impute all missing values as follows. For categorical data fields missing values can be treated as a separate category by assigning them a specific string, character or integer. Missing continuous values can be imputed with straightforward measures of center like mean, median or mode. In addition it is necessary to create a column that flags each record that contains imputed continuous values in order to be able to visualize them on the density curve. Please note that the main intention of such an approach is to assess whether missing values of certain records in the data set are unusual, not to find the best possible imputation. More precisely, in reference to the framework of missing data types [22], the intention is to facilitate the differentiation between structural deficiencies in the data and random occurrences as a prerequisite for an adequate imputation of missing values. For information on in-depth imputation methods see [23].

Hybrid Data Types. When considering data types in the context of commonality there are two conditions that we refer to as hybrid, namely categorical data that resembles continuous data (i) and vice versa (ii).

(i) Ordinal data refers to values that have meaningful order and can be ranked so that higher values represent more of a certain characteristic than lower values [14]. If a categorical data field contains ordinal data and there is high cardinality, then measuring categorical commonality will most likely skew the picture as information about the value order is lost. For such data fields it makes more sense to measure continuous commonality rather than categorical commonality.

(ii) In case a continuous data field has low cardinality, floating point numbers in that data field might represent categories rather than a distribution. For such data fields it makes more sense to measure categorical commonality, as the value order information is of lesser (or zero) importance compared to the categorical allocation of values in the data field.

Again, the intention is not to measure exact probabilities, but to provide guidance. With this in mind, for interpretation it can even help to measure both categorical and continuous commonality when dealing with hybrid data types.

Date and Time Values. Given that date and time values come in various forms, working with such data types can be quite complex. In order to measure commonality, we suggest to convert data fields that contain date and time values to integers regardless of whether they have been made available as string or any other kind of date format. In this way, the information about the value order is preserved and, if desired, both categorical and continuous commonality can be measured.

https://doi.org/10.30844/wi_2020_c7-stege

Heavy Tailed Distributions. Statisticians use the term heavy tails to describe distributions that contain extreme events. With respect to continuous data types, one often encounters a skewness towards large values in a dataset, meaning that one or few records are much larger than the bulk of the data. Kernel density estimation which we apply in the process of measuring continuous commonality comes with some drawbacks, especially with respect to heavy tailed distributions [15]. Thus, we use log transformation on the data, if heavy tailed, prior to estimating kernel density as in (6).

4 Applicability Check

Our approach described in the previous section is now applied to a real world problem from the audit. As starting point, we use an anonymized dataset extracted from the risk data warehouse of a European bank and describe the visualization setup with the help of an arbitrary sample (Section 4.1). We then perform an applicability check: we take the roles of an auditor and a data scientist and have each of them draw unusual items from the bank's dataset. Upon that we present how to examine those subsets with the help of commonality plots (Section 4.2). From the results we derive limitations and recommendations (Section 4.3).

4.1 Data Description and Visualization Setup

As illustrative example of our approach we use a dataset that contains typical information that a bank is required to provide during an audit for a specific reporting date. As such, the resulting information represents an unalterable snapshot of the bank's portfolio as of that reporting date. The data is an anonymized random sample with 15 columns and 1,000 rows based on real transactions extracted from the risk data warehouse of a European bank. Table 1 shows a summary of the dataset.

Table 1. Dataset Summary

Data field	*Cardinality*	*Data type*	*Description*
rep_date	*1*	*hybrid*	Dataset extraction date / reporting date
legal_entity	*2*	*categorical*	Name of legal entity / subsidiary
portfolio	*2*	*categorical*	Name of portfolio
stage	*3*	*categorical*	Risk class of the expected credit loss model
country	*10*	*categorical*	Contract country of origin
product_type	*12*	*categorical*	Contract product type
dayspastdue	*14*	*hybrid*	No. of days that a payment is delayed
init_rating	*26*	*hybrid*	Rating at origination
rating	*26*	*hybrid*	Current rating
contract_age	*32*	*hybrid*	Age of contract [in years]
loss_provision	*296*	*continuous*	Loan loss provision
start_date	*585*	*hybrid*	Date of origination of the contract
exposure	*611*	*continuous*	Gross carrying amount

https://doi.org/10.30844/wi_2020_c7-stege

prob_dflt	*757*	*continuous*	Probability of default
contract_id	*1000*	*categorical*	Unique contract identifier

The table also marks those data fields we claim to be hybrid. For further analysis we decide to exclude all data fields that have identical values throughout the entire dataset. This applies to the data field that holds information about the reporting date ($|D[rep\_date]| = 1$). All calculations were carried out using a notebook with a four-core i7 2.90 GHz processor and 32 GB of RAM. Performance tests reveal a reasonably linear correlation between computational effort and the size of the dataset: 59.4 ms $\pm$ 6.2 ms per run for one thousand rows, 6.8 s $\pm$ 47.2 ms per run for one million rows, 76 s $\pm$ 2.4 s per run for ten million rows.

In order to describe the visualization setup let us assume that we have obtained an arbitrary subset from the data for which we have calculated the commonality measures defined in Section 3. Then, for the visualization of categorical commonality we use bar charts to present the commonality likelihood measures from (4) as this type of chart is easy to interpret and very useful for displaying comparative data. The bars are sorted by the level of likelihood and are displayed horizontally for clearer labeling. In order to facilitate interpretation of the likelihood measures we additionally provide the ratio of value appearance in the subset versus appearance in the entire dataset. For this we use pie charts as they are among the most popular options for displaying compositions, especially when there are only a few categories. More on that in the following section. For the visualization of continuous commonality, we introduce a projection method that facilitates the visual investigation of numerous distributions. Let us have a look at the distributions of the data fields $contract\_age$ and $exposure$ as an example (see Figure 3). Along the distribution we position the values from the arbitrary subset (red marks). The size of the marks is an indication for the density of the values: the higher the density, the lower the size of the mark. Additionally, we integrate a rug plot, which is a one-dimensional display of the distribution that uses short lines for each value occurrence [24]. As a result, it is now possible to compress the plot in favor of a minimalistic design without losing much information about the distribution. While usually two-dimensional scatter plots or heat maps are used to visualize distributions and patterns in data, we end up with one-dimensional plots that can be stacked to examine all continuous data fields in the dataset at a glance.

https://doi.org/10.30844/wi_2020_c7-stege

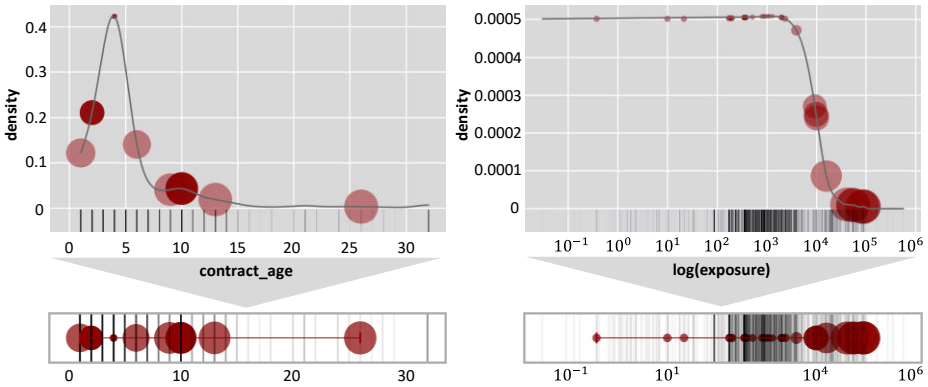

Figure 3. Projection Method for the Continuous Commonality Plot

4.2 Subset Selection and Interpretation

In this section we demonstrate how commonality contributes to the examination of data in general (example 1) and how visualizing commonality can facilitate the dialogue between a data scientist and an auditor (example 2). For this purpose we separately select two subsets from the dataset – one in the role of an auditor and one as a data scientist. Both subsets are examined with the help of commonality plots.

Example 1: Auditor Subset Selection. The auditor knows from the bank's lending policy that loans are only granted to borrowers with a certain level of credit-worthiness. For credit quality assessment the bank uses a rating scale that ranges from 1 (highest quality, lowest level of credit risk) to 30 (lowest quality, default without prospect for recovery). The bank will not grant loans with a rating grade worse than 20. Thus, the auditor considers deviations from this policy to be unusual and selects all items of $D[init\_rating] > 20$ for further investigation. The auditor uses commonality plots to further examine the resulting subset (see Figure 4, left-hand side, for overview purposes we only show continuous commonality plots where $1 - P > 5\%$). The categorical plot

https://doi.org/10.30844/wi_2020_c7-stege

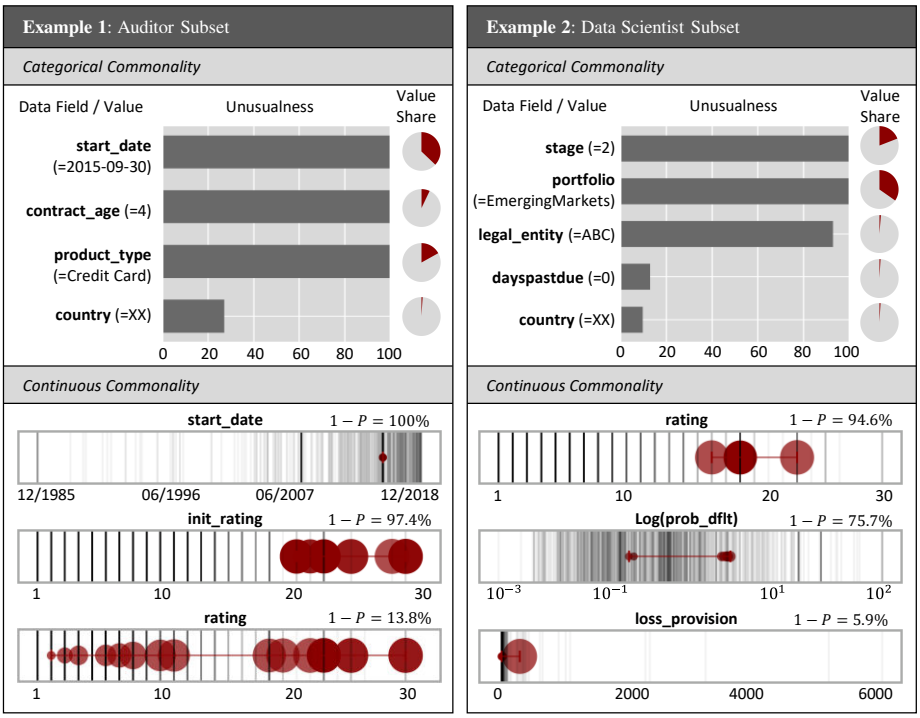

Figure 4. Commonality Plots for Auditor and Data Scientist Subset

reveals that all items of the auditor's subset have the same product type, age, date of issue and country of origin. While the fact that all items are from the same country is not unusual according to the bar size, the categorical commonality for the remaining data fields appears to be highly unusual. The pie charts help the auditor to better classify the unusualness: a contract age of four years seems to be an age which occurs very often in the entire dataset, so the resulting share in the subset is very small. This needs to be taken into account when comparing the likelihood of occurrence of the contract age (lower share, higher likelihood of occurrence) to that of start date and product type (higher share, lower likelihood of occurrence). Regarding the continuous plot for the start date, the unusualness of occurrence of identical values is emphasized ($1 - P = 100\%$). The continuous plot for the initial rating distribution reflects the auditor's selection (all initial ratings worse than 20). The auditor uses the insights from the commonality plots as basis to question the bank. After consultation with the bank it turns out there was a merger with a retail bank in September 2015. Following the merger, the date of issue of the newly added contracts was set to the date of the merger and not to the original date of issue. Since the credit quality at origination is directly linked to the date of issue, this lead to incorrectly recorded initial ratings, which in turn may lead to misstatements in the financial statement of the bank.

Example 2: Data Scientist Subset Selection. The data scientist applies pattern recognition algorithms to the distributions of all continuous data fields and finds an

https://doi.org/10.30844/wi_2020_c7-stege

unusual value cluster in $D[prob\_dflt]$. Thus, the data scientist selects all items belonging to the cluster for further investigation. Now, without going into technical details of outlier detection, the data scientist can use commonality plots to examine the subset in close collaboration with the auditor. The commonality plot for the data scientist's subset is shown on the right-hand side in Figure 4. The categorical plot reveals that all items in the subset come from a certain portfolio and country and are allocated to the second risk class that holds contracts for which the credit quality has significantly deteriorated since origination. The unusualness of risk class and portfolio is emphasized by the pie charts. None of the contracts in the sample have delayed payments, which is questionable as an allocation to the second risk class is often accompanied by arrears. Despite the credit quality deterioration, the loan loss provision recognized for these items is very low with reference to the continuous plot of the distribution. In addition, the low default probability does not relate to the credit rating that is recorded for the contracts (high rating is equal to poor credit quality). After consultation, it turns out that subsidiary "ABC" of the bank provides percentages in a different value range ($[0,1]$) than the way in which percentages are displayed and shared by the affiliate ($[0,100]$). If undetected, the inconsistent measures lead to a significant undervaluation of the credit risk of the subsidiary in the financial statements.

To summarize, in example 1 an auditor uses domain knowledge to select a sample of unusual items. We then demonstrate how commonality plots generally support examination and the discussion of findings and implications. Example 2 goes one step further and shows how a data scientist can be integrated into the workflow to make it hybrid. The data scientist is capable to reveal patterns that are not apparent with pure domain knowledge and commonality plots help to facilitate a dialogue between data scientist and auditor. Without a sound collaboration there is high risk that the potential misstatements we describe remained undetected.

4.3 Limitations and Recommendations

Besides illustrating the benefits for examination and discussion, Figure 4 also helps to reveal drawbacks of the visualization method. A major difficulty is the overall calibration of the likelihood measures. For categorical commonality a combination of likelihood measures (bar chart) and total share (pie chart) to a joint measure will most likely yield more granular results and improve interpretability. In terms of continuous commonality, the calibration is more complex. If the bulk of the values of a subset is located on one end of the distribution and there is a single outlier on the other end, the likelihood measure will be close to 1 as we calculate the integral of the entire density curve. Additionally, if the bulk happens to occur with high density, its values will be displayed with a small marker size. For enhancement we suggest to integrate a function that adds a certain weight to the number of occurrences in clusters to separate bulks from outliers when measuring the likelihood of continuous commonality. Another drawback is that categorical commonality, as we define it, only reveals "pure" or "first-hand" commonality, i.e., a cardinality of one. It follows

https://doi.org/10.30844/wi_2020_c7-stege

that a single outlier in an otherwise identical subset is enough to exclude a data field from further investigation. Thus, it seems worthwhile to expand the categorical commonality to also account for cases of "second-hand" commonality. As mentioned in Section 3, we use log transformation to avoid the flaws of kernel density estimation regarding heavy-tailed distributions. This highly influences the appearance of the plots and can also affect the likelihood measures. The same applies to kernel density estimator parameter tuning or a completely different estimation approach. It is important to keep these limitations in mind when drawing conclusions from the resulting measures and visualizations.

5 Discussion and Outlook

The applicability check shows how an auditor can use commonality plots as indication of where to focus further investigation. More generally, visualizing commonality can be seen as a first aid kit for domain experts to gain insights and formulate better questions for discussion. From a different perspective, commonality plots facilitate the interpretation of the work of data scientists, who may be the only ones capable of revealing certain phenomena, but in many cases do not have sufficient domain knowledge to explain these phenomena. Following, commonality plots can be viewed as a translation device that helps bridging the gap between data science and domain expertise, and thus provides decision support. The proposed method performs best for structured datasets that remain unaltered throughout the analysis, such as data at a specific cut-off date. On the other hand, our method is least efficient for the analysis of time series data that is constantly updated, such as real-time data, as this also requires a constant update of the commonality measures that need to be calculated for the entire dataset. Against this background, the applicability check can be adapted to other domains, where specific subsets of structured datasets are examined at a certain cut-off date or for a fixed period. For instance in health domains (What does a certain group of patients or disease patterns have in common?), marketing (What do peak sales or certain customers have in common?) or machine maintenance (Is there something common about specific machine failures?). The first example of the applicability check shows that such questions can be answered with the help of commonality plots by a respective domain expert him or herself. A great potential of our method lies in the interpretability of resulting subsets obtained from sophisticated pattern recognition algorithms. In other words, our method becomes most beneficial when applied in a hybrid framework where domain knowledge and data science are combined. On the downside, the limitations outlined in the previous section show that the biggest strength of our method is also its biggest weakness: the visualization method is capable to highlight potential areas for further investigation, but, in its current state, may also obscure certain information that is actually relevant. Conceptual enhancements such as the proposed second-hand commonality and weighting functions for the likelihood measures are helpful to limit the drawbacks. Besides these technical limitations, the overall challenge is to integrate our commonality and visualization method into the regular workflow of (structural) data

https://doi.org/10.30844/wi_2020_c7-stege

analytics. We describe most building blocks necessary for this purpose, but these still need to be properly calibrated and automated in order to increase overall interpretability, reliance, and acceptance. For this purpose, the measures for the entire dataset can be automatically calculated and stored in the background. Similarly, likelihood measures can be calibrated automatically, so that whenever subsets undergo investigation, commonality plots are instantly and seamlessly made available, regardless the domain or size of the datasets. Until now, our overall assumption was that the process starts with a dataset that has already undergone preprocessing. However, there is great potential in setting an earlier starting point to properly align commonality measures and data preprocessing. In this context, e.g., aligned and collaborative feature engineering can reduce the cardinality of certain data fields to enable a more precise and distinctive analysis of commonality. This also enables a critical assessment of the impact of missing or erroneous data at an earlier stage. Ultimately, the visualization method must be applied with high caution as minor changes in calibration can have significant effects on the visual appearance of the commonality plots. As this is not a one-size-fits-all visualization method, the derivation of findings and conclusions critically depend on the users' ability to carefully weigh and classify the results. Otherwise there is a high risk of misjudgment that can have serious consequences in the respective domain. Future research can focus on a proper calibration of the measures to make results more transparent and to increase overall interpretability of the visualizations. Equally important is a provision of commonality plots without much technical effort and for various kinds of datasets due to the fact that seamless availability is essential for increased acceptance. We will also concentrate on applying our method to open datasets and plan to publish the code as well as the results for reproducibility and benchmarking.

6 Conclusions

We introduce a method that bridges the gap between domain experts capable of identifying business needs and data scientists with toolboxes full of insight producing algorithms. Our research is driven by a typical challenge from the audit domain, which is answering the question of whether there is something unusual about items in a sample. We introduce commonality plots that visualize the likelihood of occurrence of values in a given subset. In our applicability check we take the roles of an auditor and a data scientist and demonstrate how commonality plots can be applied to support investigation and the discussion of findings and implications. The overall intention of our applicability check is not only to explain functionalities, but also to show that commonality plots are a translation device that facilitates the integration of practical data science into existing workflows by improving communication and collaboration (hybrid intelligence). In its current state, our commonality measures are far from being capable to reveal the whole truth about a subset of data. Most likely, they never will. On the upside, despite the limitations, if applied with caution, commonality plots have the potential to create value by enabling cross-disciplinary teams to reveal, interpret and discuss findings and implications together. As our method is not

https://doi.org/10.30844/wi_2020_c7-stege

restricted to auditing but designed to be universally applicable for structured datasets, we hope this encourages researchers and practitioners to apply and further develop it in their respective domains.

References

1. Agarwal, R., Dhar, V.: Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. Information Systems Research (ISR). 25(3), 443–448 (2014)
2. Chen, H., Chiang, R., Storey, V.: Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly (MISQ), 36(4), 1165–1188 (2012)
3. Reif, R.: The Data Scientist Shortage is Huge. Here's How to Beat It. https://insidebigdata.com/2018/12/27/data-scientist-shortage-huge-heres-beat (Accessed: July 25, 2019)
4. Knechel, W.R.: Audit Quality and Regulation. International Journal of Auditing. 20(3), 215–223 (2016)
5. Bedard, J.C., Biggs, S.F.: Pattern Recognition, Hypotheses Generation, and Auditor Performance in an Analytical Task. Accounting Review. 66(3), 622–642 (1991)
6. Raphael, J.: Rethinking the Audit: Innovation is Transforming How Audits are Conducted – and Even What it Means to Be an Auditor. Journal of Accountancy. 223(4), 28 (2017)
7. Kokina, J., Davenport, T.H.: The Emergence of Artificial Intelligence: How Automation is Changing Auditing. Journal of Emerging Technologies in Accounting. 14(1), 115–122 (2017)
8. Eilers, D., Köpp, C., Gleue, C., Breitner, M.H.: It's Not a Bug, It's a Feature: How Visual Model Evaluation Can Help to Incorporate Human Domain Knowledge in Data Science. Proceedings of the International Conference on Information Systems (ICIS) (2017)
9. Gandomi, A., Haider, M.: Beyond the Hype: Big Data Concepts, Methods, and Analytics. International Journal of Information Management. 35, 137–144 (2015)
10. Labrinidis, A., Jagadish, H.V.: Challenges and Opportunities with Big Data. Proceedings of the VLDB Endowment. 5(12), 2032–2033 (2012)
11. 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(6), 1073–1080 (2009)
12. Alpar, P., Schulz, M.: Self-Service Business Intelligence. Business & Information Systems Engineering. 58(2), 151–155 (2016)
13. Cairo, A.: The Truthful Art: Data. Charts, and Maps for Communication. New Riders (2015)
14. Boslaugh, S.: Statistics in a Nutshell: A Desktop Quick Reference. O'Reilly Media (2012)
15. Hoadley, B.: The Compound Multinomial Distribution and Bayesian Analysis of Categorical Data from Finite Populations. Journal of the American Statistical Association. 64(325), 216–229 (1969)
16. Janardan, K.: Chance Mechanisms for Multivariate Hypergeometric Models. Sankhyā: The Indian Journal of Statistics. 35(4), 465–478 (1973)
17. Wasserman, L.: All of Statistics: A Concise Course in Statistical Inference. Springer (2013)
18. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Routledge (1998)
19. Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman and Hall/CRC (1994)
20. Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data. John Wiley & Sons (2019)

https://doi.org/10.30844/wi_2020_c7-stege

21. García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R.: Pattern Classification with Missing Data: A Review. Neural Computing and Applications. 19(2), 263–282 (2010)
22. Kuhn, M., Johnson, K.: Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press (2019)
23. Van Buuren, S.: Flexible Imputation of Missing Data. Chapman and Hall/CRC (2018)
24. Hilfiger, J.J.: Graphing Data with R: An Introduction. O'Reilly Media (2015)

https://doi.org/10.30844/wi_2020_c7-stege