WORKSHOP

Quality Aspects of Machine Learning –
Official Statistics between Specific Quality Requirements and Methodological Innovation

September 06 to 08, 2022 in Munich

ABSTRACTS

Qualitätsdimensionen für den Einsatz von ML in der amtlichen Statistik

Johannes Rohde (IT.NRW) und Christian Salwiczek (Statistik Nord)
Johannes.Rohde[at]it.nrw.de, christian.salwiczek[at]statistik-nord.de

Die Statistischen Ämter des Bundes und der Länder haben das Ziel, das Qualitätsniveau der Statistiken dauerhaft sicherzustellen und zu gewährleisten. Dazu werden im Statistischen Verbund eine Vielzahl von Instrumenten zum Management der Datenqualität eingesetzt. Das Qualitätshandbuch beschreibt den Rahmen zur Sicherstellung der Datenqualität in der deutschen amtlichen Statistik. Es liefert Qualitätsrichtlinien, die für alle Phasen der Durchführung amtlicher Statistiken (z. B. Datengewinnung, Aufbereitung, Verbreitung) konkrete Vorgehensweisen beschreiben. Die Umsetzung der Richtlinien sind für alle Fachbereiche verbindlich und dienen damit der Sicherstellung einer hohen Qualität der statistischen Prozesse und Produkte. Maschinelles Lernen, das im Rahmen des statistischen Produktionsprozesses bisherige Arbeiten ersetzend oder ergänzend oder ganz neu eingesetzt wird, muss daher als Mindestanforderung sicherstellen, dass die Qualitätsanforderungen an die statistischen Prozesse und Produkte weiterhin erfüllt werden. Um dies zu gewährleisten muss in einem ersten Schritt geprüft werden, welche Qualitätskriterien der amtlichen Statistik von ML in welchem Umfang, bspw. qualitätserhaltend oder auch qualitätsverbessernd, berührt werden. Zu diesem Zweck wurde vom Statistischen Verbund eine Projektgruppe eingesetzt, deren Ziel es sein soll, Qualitätsdimensionen zu beschreiben, die die Anforderungen an den Einsatz von (ML-)Algorithmen in der amtlichen Statistik definieren. Diese Qualitätsdimensionen könnten bzw. sollten dann in einem weiteren Schritt die Einhaltung der Qualitätskriterien messbar machen, sodass darüber hinaus die Möglichkeit entsteht, den Output der ML-Verfahren mit anderen Algorithmen oder Methoden zu vergleichen.

Quality, ML, Destatis und Internationale Projects

Florian Dumpert (Federal Statistical Office of Germany)
Florian.Dumpert[at]destatis.de

Quality aspects of official statistics in Germany are discussed in the German Official Statistics Network (formed by the Statistical Offices of the Federation and the Länder). Of course, this is without prejudice to complementary considerations in individual institutions (such as the Federal Statistical Office) or at the international level. The lecture presents and discusses considerations of the Federal Statistical Office in the area of the quality of machine learning in official statistics as well as work from an international project.

Equity, inclusion, and fairness in data-driven decision making in the public sector

Christoph Kern und Frauke Kreuter (Uni Mannheim und LMU München)
christoph.kern[at]stat.uni-muenchen.de, frauke.kreuter[at]stat.uni-muenchen.de

This talk discusses quality aspects of machine learning (ML) in official statistics with a focus on equity, inclusion, and fairness. Using case studies involving data-driven decision-making in the public sector, we highlight various forms of biases introduced along the ML modeling pipeline, propagated through downstream tasks, and affecting statistical products. Based on a process model of bias in ML systems, we exemplify how collaborative efforts from statistics, social and survey science can contribute to a fair and safe adoption of ML by uncovering and mitigating misrepresentation in training data, by monitoring implications of data processing and analysis decisions, as well as by studying public perceptions on the algorithmic automation of processes in practice.

Fashions of Artificial Intelligence

Rudolf Seising (Forschungsinstitut für Technik- und Wissenschaftsgeschichte; Deutsches Museum München)
r.seising[at]deutsches-museum.de

When did the term Artificial Intelligence (AI) come into the world? What did it mean and how has it developed since its emergence in the mid-20th century? AI was coined as a term for a field of research in 1955, when the young mathematician John McCarthy at Dartmouth College in Hanover, New Hampshire, planned a “Summer Research Project on Artificial Intelligence”, which he, together with Claude E. Shannon, Marvin Minsky and Nathaniel Rochester, proposed to the Rockefeller Foundation to fund. Having previously failed with his proposal of "Intelligent Machines" as the title of the volume he and Shannon then published as "Automata Studies", McCarthy then pushed through the use of the word "intelligence".

For cybernetics,[1] the "superscience" launched by Norbert Wiener in the last third of the 1940s, the fast-computing digital computers built in the USA at that time became the paradigm of the "rise of the machines": Machines that can control themselves and learn. Here, a thinking of man from the machine emerged, but also a thinking of the machine from man (Thomas Rid).[2] Cybernetics prepared AI in different ways in different contexts. From England, William Ross Ashby tried to bring his view of the new computers as “amplifiers” of human intelligence into cybernetics; Philipp Aumann attested that cybernetics in Germany was characterised by utopias and speculations, which is why he called it a "fashionable science" (“Modewissenschaft”).[3]

Even in earlier times, speculations were made about whether machines exhibit intelligent behaviour, and Alan M. Turing discussed these in his 1950 article "Computing Machinery and Intelligence".[4]

The philosopher John Haugeland described AI as a fashion - "Good Old Fashioned Artificial Intelligence" (GOFAI) in his 1985 book *Artificial Intelligence: The Very Idea*.[5] In this way, 30 years after the Dartmouth event, he distinguished these AI methods of symbolic representations, which have since been called "classical", from newer approaches that do not use explicit high-level symbols, such as mathematical optimisation, statistical classifiers and neural networks. These "new-fashioned AI", "new wave AI" or "new-fangled AI" (NFAI) researches almost totally shape our current concept of AI, mostly without considering its historical developments and meanings. In this lecture, I present a view on the history of AI and these mentioned fashions.

[1] Wiener, N.: *Cybernetics: Or Control and Communication in the Animal and the Machine*. Paris: Hermann & Cie & Camb. Mass.: MIT Press 1948.
[2] Rid, Th.: *Rise of the Machines. A Cybernetic History*. New York: W.W. Norton & Comp. 2016.
[3] Aumann, Ph.: *Mode und Methode. Die Kybernetik in der Bundesrepublik Deutschland*. (Deutsches Museum. Abhandlungen und Berichte - Neue Folge 24), Göttingen: Wallstein Verlag 2009.
[4] Turing, A. M.: Computing Machinery and Intelligence, *Mind*, LIX (236): 433–460.
[5] Haugeland, J.: *Artificial Intelligence: The Very Idea,* Cambridge, Mass.: MIT Press 1985.

Ten propositions on machine learning in official statistics

Arnout van Delden, Joep Burger und Marco Puts (CBS – Statistics Netherlands)
a.vandelden[at]cbs.nl, m.puts[at]cbs.nl

Machine learning (ML) is a relatively new tool in the official statistician's toolbox, already stuffed with design-based, model-assisted and model-based methods. Although both statistical modeling and ML can be used in the production of official statistics, ML offers new opportunities as well as new challenges. We pose ten propositions related to applicability, causality, explainability, hyperparameter tuning, model performance and rare phenomena. We hope they will raise awareness and stimulate discussion among official statisticans considering applying ML.

Challenges and solutions when adopting ML (Machine Learning) and AI (Artificial Intelligence) in large organisations

Joni Karanka und Eleanor Law (ONS – Office for National Statistics)
joni.karanka[at]ons.gov.uk, Eleanor.Law[at]ons.gov.uk

Machine learning and artificial intelligence have many benefits for data and information intensive industries. In the case of official statistics, the use cases include increasing the efficiency of existing statistical operations and processes; creating and making use of new sources of data; and increasing the quality of statistics. Typically, ML and AI projects are championed by individuals and small groups and delivered in a case-by-case basis. When the portfolio of such projects grows, it can lead to issues with support and standards in larger organisations. This leads to the challenge of how we industrialise ML projects to deliver them to agreed quality more efficiently and sustainably.

In this session we want to introduce a series of problems and exchange ideas and best practices on how teams and organisations can best make use of machine learning. Some include:

- Organisation. What are the best ways to organise ML functions within an organisation? Do you centralise or distribute your data ML functions? What are the most common ML roles, ranging from labellers of training data to ML infrastructure engineers?

- Model quality. What are the key quality criteria to evaluate in models? When should a model be revised, updated, or replaced? How do you monitor model drift of different models? What choices can be made in model design to ensure enduring quality, not just good performance at a single point in time?

- Support. How do you support ML services and products that are live in the long term? How can you reduce the support burden of an ML service? How do you build capability and maintain the expertise required to continually quality assure complex models in production?

- Tools and standards. What standards and tools can support organisations in making best use of ML? How do you automate the deployment of models? What tools can save time and effort for developing and maintaining models?

What is Fairness? Implications for FairML

Ludwig Bothmann, Kristina Peters, Bernd Bischl (LMU München)
Ludwig.Bothmann[at]stat.uni-muenchen.de

A growing body of literature in fairness-aware ML (fairML) aspires to mitigate machine learning (ML)-related unfairness in automated decision making (ADM). Typically, this issue is addressed by defining metrics that measure the fairness of an ML model and by proposing methods that ensure that trained ML models achieve low values in those measures. However, the underlying concept of fairness, i.e., the fundamental question of what fairness is, is rarely discussed, leaving a considerable gap between centuries of philosophical discussion and the recent adoption of the concept in the ML community.

In this work, we try to bridge this gap by formalizing a consistent concept of fairness and by translating the philosophical considerations into a formal framework for the evaluation of ML models in ADM systems.

A first important point is that the concept of fairness involves the treatment of individuals, i.e., it cannot be applied to an ML model directly, since the ML model lacks the action aspect. Rather, it can be applied to the final treatment by an ADM system, and since the ML model is part of this ADM system it is possible that the ML model induces fairness problems. A second important point is that the concept of fairness is based on the concept of task-specific equality of individuals; the role of ML is to estimate characteristics that are not directly measurable and which add to the assessment of this task-specific equality.

Usually, the concept of fairness in the ML literature is designed around so-called “protected attributes” (such as gender, ethnicity, religion, etc.) which are more sensitive than others. We derive that fairness problems can already arise without such protected attributes – if the model is not individually well-calibrated. Furthermore, we point out that fairness and predictive performance are not irreconcilable counterparts, but rather that the latter is necessary to achieve the former.

Under the presence of protected attributes, we show how the concept of task-specific equality as a decision basis for fair treatment can be upheld. The crucial point is that people who are not equal in the real world are considered equal in a fictitious, normatively desirable world where the protected attribute has no causal effect on the target variable. We argue why causal considerations are necessary for this context and formulate causal questions which have to be answered in future work.

Eventually, we achieve greater linguistic clarity for the discussion of fairML and clearly assign responsibilities to stakeholders inside and outside ML. While the question of fairness is in general an essential topic in ML, it is especially relevant for official statistics since many decisions about the treatment of individuals are based on those official statistics.

Three ML-assisted strategies for coding diverse data sources

Malte Schierholz (LMU München)
malte.schierholz[at]gmail.com

Official statisticians have access to a wide variety of data, including satellite images or textual data. These "new" data sources are not in a numeric format when they arrive, and it is often required to label them before they become amenable for statistical analysis. Traditionally, humans (called "coders") have been employed to do such labeling tasks, but this process is time-consuming and expensive with large quantities of data. Supervised machine learning (ML) algorithms promise to make the process more efficient, especially when there is a recurrent inflow of new materials that need to be labeled time and again.

Yet, researchers and official statistics require the highest quality of data when they inform the public debate. Supervised machine learning, however, can make erroneous predictions, leaving questions on data quality. This paper shows three strategies how machine learning can still be applied in the data labeling process, discusses the underlying rationales and exemplary use cases of each strategy, and explores their respective advantages and disadvantages.

First, machine learning predictions can be used prescriptively. The labels predicted by the ML model are considered correct, simply because one has agreed to use this algorithm to label the data; alternative labeling approaches are—by definition—not better. This requires high confidence in the selected algorithm, for example because it has been shown to have (near) 100% accuracy or because the pros and cons of the algorithm are well understood and suitable for the analysis.

Second, there is often reason to question the validity of ML results. In many applications the accuracy of ML is below 100%, implying that the possibility of error exists, and ML might not be considered suitable to generate appropriate labels. Even if fully automated coding is beyond reach, the same algorithms (2a) may still help if automatic labeling is possible for parts of the data, reducing the workload of manual coding. Alternatively, (2b) the ML algorithm can be used to generate a set of candidate labels (from a long list of possible outcome categories). They are then subject to human review where the ML predictions get verified or corrected when necessary, ensuring the final label is free of error.

Third, when ML predictions are possibly incorrect, they may still be valuable to estimate aggregate population parameters. The key is to estimate the bias (i.e., the average difference between predicted and ground-truth values) from a representative random sample. The ground-truth values are determined using a (manual) gold-standard coding procedure, but only for the sample, reducing coding costs. Continuous monitoring may be needed to identify shifts over time. This approach guarantees unbiased estimates (unlike a pure ML approach) and promises to reduce standard errors.

Record Linkage of Company Data Sets

Valentin Reich (ifo Institut – Leibniz-Institut für Wirtschaftsforschung)
reich[at]ifo.de

(Preliminary and incomplete. Please do not cite or circulate!)

Linking different micro data sources can open up completely new empirical research opportunities. The process of linking entities from different data sources is called Record Linkage (RL). When the two data sources do not have a common identi_er, records can be linked via probabilistic matching if they are sufficiently similar in attributes such as name or address (Fellegi and Sunter, 1969, Newcombe, 1988). In a common workflow, the researcher uses supervised Machine Learning (ML) to determine whether the units in a pair of records are the same or not given a vector of similarity measures for this pair (Christen, 2012).

However, data from different domains can have greatly varying linkage requirements and often one cannot naively apply the exact same procedure and trained models across projects. Especially when working with records of non-natural persons, such as company level data, the linkage can be particularly challenging and it is important to tailor it to this use case. This paper highlights the specific challenges of company data RL and describes how the LMU-ifo Economics and Business Data Center[1] (EBDC) combines its data sources accordingly.

There are several factors that make supervised ML based RL challenging by itself but with records of non-natural persons there are additional complications. A key reason for this lies in the hierarchical nature of companies, where _rms often belong to a group of _rms, potentially with near identical name and addresses. This can entail, for example, different legal entities for production, trade, and administration. One main differentiator for _rms within a group is the entity name which usually consists of multiple tokens. However, it can vary widely from one database to another which tokens are included and in which order they appear. Relocations, name changes, mergers, restructuring within a corporate group, or sector changes complicate this further. When dealing with panel data, it is even possible that a record should be matched to different entities depending on the period of an observation. A further hurdle for ML supported company RL is that even for humans it may not be obvious how to tell true and false matches apart. This makes the collection of labelled training data for ML difficult and time consuming. The EBDC tackles these challenges with its procedure targeted towards the specific demands of company linkage. We do so, for example by both using specific comparison metrics that work well with tokens of arbitrary order and exploring the use of Natural Language Processing methods which are uniquely applicable when dealing with company records: Because company name tokens have a linguistic meaning that can capture differences within a corporate group, pretrained embedding vectors (Mikolov, Grave, Bojanowski, Puhrsch, and Joulin, 2018) allow to extract that information which can then be used in a variety of preprocessing and comparison steps. Using this procedure, we were able to substantially increase the share of our IDs

[1]The LMU-ifo Economics and Business Data Center is a Munich based research data center at the ifo Institute that provides access at their workstations to company micro data for academic research. Their data includes subjective micro data from the ifo Business Surveys alongside a version of this data enriched with companies' objective balance sheet data from Hoppenstedt and the Bureau van Dijk Databases such as Orbis.

that could be matched with their respective balance sheet data compared to a previous linkage (Hönig, 2010).

Despite these efforts, probabilistic Record Linkage of non-natural persons remains a challenging problem, in particular because of the similarity of entities within a group. Finding ways to solve this problem would greatly facilitate various applications of company level RL.

References

Christen, P. (2012): Data Matching. Springer Berlin Heidelberg, Berlin, Heidelberg.

Fellegi, I. P., and A. B. Sunter (1969): \A theory for record linkage," Journal of the American Statistical Association, 64(328), 1183{1210.

Hönig, A. (2010): Linkage of Ifo Survey and Balance-Sheet Data: The EBDC Business Expectations Panel & the EBDC Business Investment Panel, Schmollers Jahrbuch, 130(4), 635-642.

Mikolov, T., E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin (2018): Advances in Pre-Training Distributed Word Representations, in Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).

Newcombe, H. B. (1988): Handbook of record linkage: methods for health and statistical studies, administration, and business. Oxford University Press, Inc.

The SearchEngine: a Holistic Approach to Matching

Thorsten Doherr (ZEW - Leibniz-Zentrum für Europäische Wirtschaftsforschung)
Thorsten.Doherr[at]zew.de

Matching two databases of unrelated origin is one of the most common tasks in data science expanding the research opportunities way beyond the potential of the isolated databases. For instance, patent data alone provides ample insights of the spatial and temporal distribution of technologies and their markets but linking the patent applicants to firm level databases extends the portfolio from bibliometric assays to economic research of innovation. The different backgrounds motivating the amassing of this data increase the scope of research opportunities as well as the effort required to link their respective entities due to the lack of mutual unique keys like official company-IDs. Other overlaps in the data, i.e. firm names, addresses and so on, have to compensate this omission. In the classical approach the overlapping data would be harmonized in terms of character case, special characters, removal of filler words, truncations and other means of destroying just enough information to achieve a healthy balance between precision and recall. Blocking strategies further improve this balance to facilitate the application of ML methods further eliminating false positives. Arbitrary decisions permeate those approaches throughout all stages to accommodate the idiosyncrasies of the involved databases.

Our approach, although far from being free of arbitrariness, channels the different steps into a holistic procedure where the candidate blocking is based on a heuristic approach that exploits the same Meta information, derived from the databases, driving the ensuing ML refinement. The center stage is taken by a registry containing all words of the overlapping fields of one database. During the initialization of the registry, only basic harmonization is performed on the words, which become entries in the dictionary along with their frequencies in the mutual fields, which define separate chapters. Additional chapters can be established to contain specific tokenization of words like phonetic representations or grams to increase robustness versus misspellings and typos. Every entry can be traced back to all records in the database containing this token or word and vice versa via internal index tables. The frequencies are required to separate the identifying words from the filler words. A high frequency designates a filler word because the potential to narrow the search space is low, while a low frequency word contributes to the relevancy of the candidate list. The underlying engine attaches a weight to every word of a search term based on its frequency and an arbitrary weighting scheme superimposed on the chapters of the registry to implement a search strategy usually putting a higher weight on the context defining fields/chapters. If matching of companies is intended the weight distribution should be skewed towards the firm name while the address fields fulfill auxiliary purposes. The words are processed in descending order of their weights by retrieving the respective list of candidates. A candidate accumulates the weights of the associated lists. Only candidates passing a threshold are selected it into the final list representing the result of the search term. This procedure can be repeated to implement incremental search strategies by merging the result sets, which are not commutative. The base table always provides the data feeding the registry, while the records of other tables provide the search terms. The decision, which table should become the base table, is determined by structural properties of the tables. Size, prevalence of redundant noise and the general focus of the data acquisition are the main factors of this decision. As redundant noise causes more harm in the search

term, noisy tables should have a prior to become base tables because additional noise on the candidates has no impact on the search quality unless it is explicitly stated to rank the results by the relevancy of the candidate noise. Focused data acquisition avoids redundancies by maintaining entity related tables, where an entity, like a firm, is assigned a unique key and variants to its representation in the data are only allowed to capture temporal changes in the context of a panel. As base tables these data sources facilitate much stricter search strategies already eliminating the brunt of false positives by waiving unnecessary tolerance towards ambiguity in the representation of entities. In case of target conflicts between noise and focus, it is reassuring that a decision towards noise and therefore towards recall at the cost of precision can be mitigated with integral Machine Learning approaches filtering false positives.

The algorithm is capable to produce candidate lists for every search term replacing the classical multi-layered blocking of data along arbitrary conditionalities, like same region, similar address, same harmonized name. Blocking is required to reduce the solution space and to provide quality measurements to drive statistical or ML models based on ground truths usually sampled from the current data. As the blocking of our algorithm completely relies on already registered frequencies the quality measurements are readily available. While the search process exploits the frequencies to create a relative order of the words within a search term, the general quality of a candidate is gauged by the absolute frequencies of the matching words, the candidate exclusive words and the omitted words of the search term. To retrieve the latter, it is required to create a separate registry for the search table. Together with the accumulated weight sum, the number of co-candidates, its position within the list of co-candidates, the number of overlapping words between the different fields of the search term and other derived information, every match engenders a multitude of data points to predict false positives among the candidates based on a ML model trained on a manually scrutinized sample. We introduce the standalone tool “SearchEngine” that encompasses all steps from creating the registry, the search, aggregating the Meta data for training and prediction up to sample drawing. The package is complemented by the STATA Neural Net module “brain” and a guideline script.

Climate data for official statistics – 3 machine learning applications

Hendrik Doll (Deutsche Bundesbank)
hendrik.doll[at]bundesbank.de

Climate data is increasingly important for Central Banks as climate risks can affect financial risks and monetary policy. However to date, climate data availability varies. Besides available structured climate data and known data gaps, unstructured sources of climate data exist and remain to be systematically leveraged. Often times, such climate data exists publically but in practice is difficult to use due to a lack of structure or dispersion across many sources.

This presentation outlines an example using web-scraping and clustering to analyse self-proclaimed ESG exchange-traded funds (ETFs). Leveraging on dispersed data we assess the environmental footprints of self-proclaimed ESG ETFs, covering the largest global issuers, using public ETF holdings data and proprietary company-level emission data. Self-proclaimed ESG ETFs seem to have lower average emission intensities than their reference ETFs.

However, investors looking to cover a broad market, while rewarding lowest emitters within a sector, cannot generally do so by investing in self-proclaimed ESG ETFs. Preliminary results emphasize the need for more transparency in sustainable investments and for improving climate-related data availability. Standardization of sustainability criteria, enhanced transparency and data availability is underway on company-level. Standardization on fund-level is yet to come.

The example highlights, how Central Banks can improve access to thus far underused climate data resources. Manifold further applications can be thought of to leverage on unstructured climate data. Potential exists in one project that generates structured emission data from companies' textual non-financial reports using natural language processing. Another application is leveraging on image classification to improve economic variables. These projects are important as Central Banks stand to benefit from high-quality climate data through improved data availability and quality.

Befristungen in der Statistik der gemeldeten Arbeitsstellen: Analysen mit Text Mining

Arsen Çelikel, Joachim Seitz, Jörg Szameitat (Bundesagentur für Arbeit)
Arsen.Celikel[at]arbeitsagentur.de, Joachim.Seitz[at]arbeitsagentur.de,
Joerg.Szameitat2[at]arbeitsagentur.de

Die Statistik der Bundesagentur für Arbeit (BA) erprobt in Kooperation mit dem IT-Systemhaus der BA erstmals Methoden des Text Mining für die Analyse von Stellenanzeigentexten mit dem Ziel, die amtliche Statistik der Gemeldeten Arbeitsstellen zu prüfen und ggf. anzureichern. Ein im August 2021 veröffentlichter Methodenbericht[1] stellt die angewandte Methodik vor, mit der systematisch die Validität des Merkmals „Befristung" untersucht wurde, zeigt die erzielten Ergebnisse auf und gibt einen Ausblick, wie die gewonnenen Erkenntnisse in die statistische Verarbeitung miteinbezogen werden können.

Das Vorhaben erprobt Text Mining erstmals anhand einer breiten Basis amtlicher Daten. Verwendet wird dieselbe Datenquelle, aus der auch die Statistik der gemeldeten Arbeitsstellen erstellt wird, angereichert um die Ausschreibungstexte der Stellenangebote. Diese werden bislang nicht für die statistische Berichterstattung der BA genutzt. Amtliche Statistikdaten als Grundlage für Text Mining heranzuziehen hat mehrere Vorteile gegenüber Web Scraping mit Online-Stellenbörsen. So hat die Statistik der BA bei Text Mining mit gemeldeten Arbeitsstellen zum einen profunde Kenntnis über Struktur und Historizität der Datenquelle und zum anderen sind alle Merkmale verfügbar, die auch in den amtlichen Statistikdaten enthalten sind – und nicht nur Texte.

Mit den um Stellenanzeigentexte erweiterten Statistikdaten wurden unterschiedliche Modelle und Algorithmen getestet. Aufgrund der Ergebnisse, ihrer guten Interpretierbarkeit und der hohen Modellgüte fiel die fachliche Entscheidung auf den sog. XGBoost-Klassifikator. Das Modell könnte die statistische Information, ob ein Stellenangebot befristet oder unbefristet ausgeschrieben ist, gegenüber dem jetzigen Stand verbessern.

Das Befristungsmodell wird derzeit einem einjährigen, manuellen Testbetrieb unterzogen, um die Integration in die regelmäßige Statistikproduktion vorzubereiten. Sobald die Ergebnisse hierzu vorliegen, ist – neben der Entscheidung über eine statistische Veröffentlichung als Ersatz oder Ergänzung für das operativ erfasste Merkmal – ein Verfahren zu entwickeln, das die Arbeitsweise des Modells künftig regelmäßig überwacht und erneut trainiert.

[1] https://statistik.arbeitsagentur.de/DE/Statischer-Content/Grundlagen/Methodik-Qualitaet/Methoden-berichte/gemeldete-Arbeitsstellen/Generische-Publikationen/Methodenbericht-Befristungen-in-der-Sta-tistik-der-gemeldeten-Arbeitsstellen-Analysen-mit-Text-Mining.pdf?__blob=publicationFile&v=3

Data Science Informed by Survey Science: Collecting More Accurate Labels

Jacob Beck, Stephanie Eckman, Frauke Kreuter (LMU München)
jacob.beck[at]stat.uni-muenchen.de

Data scientists are lauded for improving models by using the newest algorithms and tuning hyperparameters, but they are increasingly realizing that the quality of the insights from their models are driven by the quality of the data the models are trained on. However, AI researchers are typically not trained in data collection and can benefit from the expertise of those who are. Frequently, machine learning models rely on high-quality input data, for example, images labeled as dogs or cats or text labeled as positive or negative sentiment. The instruments used to collect these labels are similar to web surveys, except that the questions are about images or text rather than about the labelers themselves. Survey methodologists have studied the features of web surveys that impact data quality and we expect many of these findings will transfer to the collection of training data for AI research. With our labeling experiment, we investigate whether previous findings about the mechanisms of measurement error in web surveys also apply to labeling tasks. Besides structurally skewed labels due to labeling task design, biased training data may also stem from other well-known mechanisms in survey contexts such as representation and coverage issues. Our study tests whether the principles of data quality in web surveys also apply to the collection of labels for machine learning models.

We field two versions of an instrument to code images of urban streets. Labelers are assigned to assess whether a bike lane is blocked by a vehicle and by what kind of vehicle. All images have already been coded previously and thus benchmark labels exist. The instruments can be varied by multiple components. One could be whether respondents have the possibility to reduce the number of items to answer by labeling the images in a specific way. Another modification could potentially be the ordering of questions. By comparing the labels collected in the two versions with each other and with the benchmark, we provide the first evidence that instrument design matters in the collection of labels for data science.

We investigate whether the high dependency of machine learning models on human-labeled data requires additional attention to the process of data collection. Our results will interest data scientists who want to maximize resource efficiency by collecting high quality labels. If labels are collected with higher quality standards in the first place, model performance will also increase since errors that usually transfer from training data into the model are now ruled out

Combining Bayesian estimation and machine learning methods for handling missing values and model selection in complex sampling designs

Christian Aßmann, Christoph Gaasch, Doris Stingl (Otto-Friedrich-Universität Bamberg)
christian.assmann[at]uni-bamberg.de

For large data sets estimation approaches are required that scale with regard to data set dimensions and allow for dealing with issues of missing values and model selection to elicitate interpretable results. We propose a Bayesian estimation approach based on the device of data augmentation that addresses the handling of missing values in multilevel latent regression models arising in complex sample designs. Population heterogeneity is modeled via multiple groups enriched with random intercepts. Bayesian estimation is implemented in terms of a Markov Chain Monte Carlo (MCMC) sampling approach.

Data augmentation (DA) in the Bayesian context offers a conceptually straightforward way to deal with missing values. The vector of unknown quantities can be augmented with the missing values in covariates. Correspondingly, the MCMC sampling scheme incorporates the set of full conditional distributions of the missing values. This approach has the advantage that the modeling of the full conditional distributions can incorporate information available in form of the latent variable serving in the considered model context as a sufficient statistic. In addition, in data contexts with a large number of covariates relative to the number of observations, the Bayesian approach incorporates shrinkage in terms of the involved prior distributions and facilitates updating of information with regard to the modeled relationships. Next, Bayesian estimators of parameters or functions thereof, like context effects and uncertainty measures, are directly accessible without the use of combining rules.

The DA principle has been successfully applied in different contexts ranging from multivariate panel models to social network analysis and educational large-scale assessments by Liu, Taylor, and Belin (2000), Koskinen, Robins, and Pattison (2010), Blackwell, Honaker, and King (2017), and Kaplan and Su (2018). Full conditional distributions of missing values are typically operationalized in terms of a parametric modeling approach as discussed by Grund, Lüdtke, and Robitzsch (2020) and Erler et al. (2016). Goldstein, Carpenter, and Browne (2014), Erler et al. (2016), and Grund, Lüdtke, and Robitzsch (2018) provide a discussion in the context of linear regression models for metrically scaled hierarchical data.

We extend the DA approach towards missing values in covariate data in extended hierarchical structures in LRMs for dependent variables with binary and ordinal scale arising in complex sample designs. We also illustrate that DA allows for direct access to a valid model specification for the missing values incorporating information available in form of sufficient statistics as suggested by the Hammersley-Clifford theorem, see Robert and Casella (2004). Further, specifying the full conditional distributions of missing values in terms of sufficient statistics arising in the hierarchical latent regression context has the potential to reduce the computational burden. In combination with modeling the full conditional distributions of missing values via machine learning methods in terms of non-parametric sequential regression trees as suggested by Burgette and Reiter (2010) and Doove, van Buuren, and Dusseldorp (2014), the DA approach suggested in this paper offers high exibility in empirical application to cope with nonlinear relationships, e.g. interaction terms, within a potentially large set of covariates having different scales. The proposed modeling approach allows hence for

tackling typical research questions arising in social and economic research. It simultaneously addresses the uncertainty associated with the estimation of a latent trait variable and the imputation of missing values in manifest covariate variables. The reciprocal dependence of outcomes and predictors is reflected to the full extent by the Bayesian DA estimation algorithm. The benefits of the suggested fully Bayesian approach arise in terms of methodological stringency and gains in estimation precision. Illustration of the suggested approach is provided by means of a simulation study and an empirical application using a large survey data set having a complex hierarchical sampling design. To highlight the benefits of considering sufficient statistics within the suggested DA approach towards missing values in covariates, we provide a comparison with a classical imputation setup, where the full conditional distributions of missing values are defined on the basis of directly observable quantities only, see e.g. von Hippel (2007). As shown in the simulations, the consideration of sufficient statistics accelerates the computation up to a third and ensure the feasibility of specifying full conditional distributions in multilevel contexts.

Refrences

Blackwell, M., Honaker, J., & King, G. (2017). A unified approach to measurement error and missing data: Details and extensions. Sociological Methods & Research, 46 (3), 342{369. Retrieved from https://doi.org/10.1177/0049124115589052 doi: 10.1177/0049124115589052

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. American Journal of Epidemiology, 172 (9), 1070-1076.

Doove, L. L., van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction e_ects. Computational Statistics & Data Analysis, 72 , 92-104.

Erler, N. S., Rizopoulos, D., van Rosmalen, J., Jaddoe, V. W. V., Franco, O. H., & Lesaffre, E. M. E. H. (2016). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full bayesian approach. Statistics in medicine, 35 (17), 2955{2974. doi: 10.1002/sim.6944

Goldstein, H., Carpenter, J. R., & Browne, W. J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. Journal of the Royal Statistical Society: Series A (Statistics in Society), 177 (2), 553{564. doi: 10.1111/rssa.12022

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data at level 2: A comparison of fully conditional and joint modeling in multilevel designs. Journal of Educational and Behavioral Statistics, 43 (3), 316{353. Retrieved from https://doi.org/10.3102/1076998617738087 doi: 10.3102/1076998617738087

Grund, S., Lüdtke, O., & Robitzsch, A. (2020). On the treatment of missing data in background questionnaires in educational large-scale assessments: An evaluation of different procedures. Journal of Educational and Behavioral Statistics, 107699862095905. doi: 10.3102/1076998620959058

Kaplan, D., & Su, D. (2018). On imputation for planned missing data in context questionnaires using plausible values: A comparison of three designs. Large-scale Assessments in Education, 6 (1), 6. Retrieved from https://doi.org/10.1186/s40536-018-0059-9 doi: 10.1186/s40536-018-0059-9

Koskinen, J. H., Robins, G. L., & Pattison, P. E. (2010). Analysing exponential random graph (p-star) models with missing data using bayesian data augmentation. Statistical Methodology, 7 (3), 366-384.

Liu, M., Taylor, J. M. G., & Belin, T. R. (2000). Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. Biometrics, 56 (4), 1157-1153.

Robert, C. P., & Casella, G. (2004). Monte carlo statistical methods (2nd ed.). New York, NY: Springer Science + Business Media.

von Hippel, P. (2007). Regression with missing ys: An improved strategy for analysing multiply imputed data. Sociological Methodology, 37, 83{117. Retrieved from http://www.jstor.org/stable/20451132