6-18-2022

Sharing is Caring: Using Open Data To Improve Targeting Policies

Jannik Rößler
Cologne Institute for Information Systems, roessler@wim.uni-koeln.de

Roman Tilly
Cologne Institute for Information Systems, tilly@wim.uni-koeln.de

Matthias Faska
Cologne Institute for Information Systems, matthias.faska@gmail.com

Detlef Schoder
Cologne Institute for Information Systems, schoder@wim.uni-koeln.de

SHARING IS CARING: USING OPEN DATA TO IMPROVE TARGETING POLICIES

Research Paper

Jannik Rößler, University of Cologne, Cologne, Germany, roessler@wim.uni-koeln.de

Roman Tilly, University of Cologne, Cologne, Germany, tilly@wim.uni-koeln.de

Matthias Faska, University of Cologne, Cologne, Germany, matthias.faska@gmail.com

Detlef Schoder, University of Cologne, Cologne, Germany, schoder@wim.uni-koeln.de

Abstract

When it comes to predictive power, companies in a variety of sectors depend on having sufficient data to develop and deploy business analytics applications, for example, to acquire new customers. While there is a vast literature on enriching internal data sets with external data sources, it is still largely unclear whether and how open data can be used to enrich internal data sets to improve business analytics. We choose a particular business analytics problem – designing targeting policies to acquire new customers – to investigate how an internal data set of a German grocery supplier can be enriched with open data to improve targeting policies. Using the enriched data set, we can improve the response rate of several well-established targeting policies by more than 30% in back-testing. Based on these results, we encourage firms and researchers to use, leverage, and share open data to enhance business analytics.

Keywords: open data, targeting policies, use case, marketing, business analytics.

1 Introduction

Business analytics and big data have gained a lot of attention in recent years because they support many opportunities for researchers and practitioners to create value, gain insights, and make data-driven decisions (Sharma et al., 2014). For instance, Sharma et al. (2014) argue that "improvements in organizational performance are likely to be an outcome of superior decision-making processes enabled by business analytics" (Sharma et al., 2014, p. 433).

The design of targeting policies is one type of decision problem that has benefited greatly from business analytics (Simester et al., 2019). Targeting policies are used across a broad spectrum of domains from marketing to politics and to medicine to match different treatments (product ads, political messages, drugs) to different individuals (customers, voters, patients) (Simester et al., 2020). For example, electricity suppliers want to send different promotions to different households to prevent them from churning; politicians want to use ads, direct mail, and phone calls to push residents to vote for their parties; and doctors want to treat sick patients with the correct treatment.

The literature on using heterogeneity in individuals' responses as a way to optimize decision making has gained considerable attention in recent years, providing researchers and practitioners with methods and insights that have improved our understanding of targeting policies (Ascarza, 2018; Athey et al., 2019; Devriendt et al., 2018). For example, Ascarza (2018) conducted two field experiments to eliminate a common misconception: that it may be futile to target customers at the highest risk of churning. Ascarza proposes instead to use the observed heterogeneity in response to the intervention and target only those customers who are likely to remain because of the intervention. The effectiveness of such

methods has been validated in both theoretical and experimental settings (Ascarza, 2018; Athey et al., 2019; Gubela et al., 2019). Nevertheless, sophisticated algorithms are not enough for optimizing targeting policies, as the outcome of a data mining problem is also determined by the underlying data set (Baesens et al., 2009; Simester et al., 2019).

To optimize targeting policies to acquire new customers, extend cross- or upselling offerings, and take preventive actions against customer churn, companies need data on both existing and prospective individual customers. Those data can come from various sources. With respect to existing customers, companies typically have data on product use, interactions with customer service, and past purchases (Simester et al., 2019). Companies can buy data on prospective customers from data vendors (D'Haen et al., 2013; Simester et al., 2019). Acquiring additional data from external sources can be an opportunity to improve business analytics (Afonso et al., 2019; Baecke & Van den Poel, 2011; D'Haen et al., 2013; Zheng et al., 2014). For example, D'Haen et al. (2013) demonstrated that combining web data and commercial data was most effective when acquiring new customers for a German B2B mail order company. Acquiring external data, however, is often expensive, time-consuming, and the data may be inaccurate (D'Haen et al., 2013; Simester et al., 2019).

One opportunity to overcome these issues could be to use open data sources to enrich internal data to give the data set greater predictive power. We refer to open data as data that are machine readable, accessible (e.g. through APIs), shareable, and usable by anyone, with no or only very limited restrictions (Janssen et al., 2012; Vetrò et al., 2016). This excludes data that are personal, confidential, for which fees are required, or which are protected by very restrictive copyrights or license terms (Janssen et al., 2012). Open data fall into many different application areas, such as geographic, demographic, mobility, weather, financial, legal, science, and many more (Hossain et al., 2016). Using open data sources seems especially promising given that the volume and breadth of open data held by research institutions (Link et al., 2017; Zilioli et al., 2019) and government organizations (Janssen et al., 2012), for example, has increased tremendously (Hossain et al., 2016). An increasing number of governments have enacted laws to increase the availability of raw data from public authorities and administrations (e.g., regulations regarding re-use and publication of public sector information according to Directive (EU) 2019/1024); and the Information Systems community in particular has pushed for further research into open data (Link et al., 2017). However, there is still a lack of research regarding whether and how open data can be used to enrich internal data sets and improve business analytics.

Thus, we ask the following research question: *Can open data on (aggregated) poll results and socio-demographic data improve the targeting policy of a grocery supplier's marketing campaign?* We collect voting results from the German parliamentary election of 2017, create socio-demographic features using two general election population surveys, and use geospatial information about German constituencies. This information is then used to enrich a supplier's marketing data set. Finally, we compare the performance of various targeting policies with and without the enrichment of open data sources.

We find that the enriched data set improved the response rate of the targeting policies by more than 30%. Further, the performance of all targeting policies improved when trained on the enriched data set compared to when trained on the internal data set only. This leads us to recommend both the provision and distribution of more open data sources. Our findings are generalizable to a large variety of business settings, beyond those investigated in this work (e.g., data mining in general, predictive modeling, prescriptive analytics), where open data sources can be used to enrich an internal data set and improve overall outcomes.

The paper continues in Section 2 with a literature review, first concerning targeting policies and then open data. We illustrate the research method in Section 3 with a description of the open data set used, the mining process for the socio-demographic features, the geospatial information, the data set that was provided by the German supplier, and the evaluation procedure of the targeting policies. Section 4 presents our results. Section 5 includes the discussion, our contribution to research and practice, and the limitations of our study and suggestions for future research.

2 Literature Review

2.1 Targeting Policies

There has been a growth of interest in marketing research in evaluating targeting policies using randomized, controlled experiments with a randomized-by-action (RBA) design in which customers (including a control group) are randomly assigned to marketing actions (Ascarza, 2018; Simester et al., 2020). Leveraging the data from such experiments, researchers and practitioners estimate individual treatment effects (ITEs), that is, the change in probability for an individual to exhibit a specific behavior that is caused by the treatment (Devriendt et al., 2018).

More formally, following the Rubin causal model (Rubin, 1974), we can compare the outcomes we observe and the counterfactual outcomes we would have observed under a different treatment. Consider a setup with *N* individuals, *i=1,...,N*. Let $T_i \in \{0,1\}$ be a binary treatment indicator, with $T_i = 1$ if that individual i has been subject to a treatment, and $T_i = 0$ otherwise. For each individual, we denote $Y_i(T_i = 1) = Y_i(1)$ as the outcome of individual *i* being subject to a treatment, and $Y_i(T_i = 0) = Y_i(0)$ as the outcome of individual *i* not being subject to a treatment. Thus, the ITE can be approximated by using the conditional average treatment effect (CATE), also called uplift, as

$$\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x] \quad (1)$$

where X_i is a vector of features or covariates.

A variety of methods have been proposed in different streams of work to estimate ITEs. In the heterogeneous treatment effects (HTEs) literature, statistics and economics researchers mainly propose using nonparametric statistical estimation to calculate the CATE. For example, Athey and Imbens (2016) propose the causal tree/honest tree, which was later extended to the causal forest/double-sample causal tree (Wager & Athey, 2018). Similarly, Su et al. (2012) approach CATE estimation by using the causal inference tree – a regression tree ensemble. Hahn et al. (2020) found that their Bayesian causal forest model, which models the degree of shrinkage on the treatment effect directly and separately of the prognostic effect, performs especially well, with strong confounding, targeted selection and weak treatment effects. In contrast to these approaches, Künzel et al. (2019) and Nie and Wager (2021) use meta-learning methods built on base algorithms such as random forest or neural networks to estimate the CATE. Finally, a number of authors propose neural networks to predict a CATE function (Alaa et al., 2017; Farrell et al., 2021; Schwab et al., 2019; Shalit et al., 2017; Shi et al., 2019; Yao et al., 2018). For example, Shalit et al. (2017) used a neural network with two separate heads, one for the response prediction with treatment and the other for the response prediction without treatment. Subsequently, samples were used to update only one of the heads depending on the observed treatment. More recently, Farrell et al. (2021) developed a framework based on deep learning – that is, neural networks – which retains the interpretability and economic meaning of classical economic models combined with the predictive power of machine learning algorithms. The authors show that with this framework, estimates and inferences become economically meaningful.

In another stream of work, marketing, information systems, and machine learning researchers tackle the same set of problems using so-called *uplift modeling*. Similar to the HTE literature, researchers in uplift modeling tend to favor nonparametric models – that is, tree-based algorithms – over (semi)-parametric models. For example, Rzepakowski and Jaroszewicz (2012) modify the conventional CART algorithm by using one of three different distribution divergence measures: Kullback-Leibler divergence; squared Euclidean distance; and chi-squared divergence. This approach was later extended to a Random Forest (Sołtys et al., 2015). Radcliffe and Surry (2011) propose Significance-Based Uplift Trees that fit and measure the quality of each candidate split using a linear model. More recently, Zhao et al. (2017) suggested the Unbiased Contextual Treatment Selection (UCTS) algorithm, which recursively maximizes the expected response at each split. Similar to Athey and Imbens (2016), they split a given data set into two samples, one for building the trees and another one for predicting the estimates in the leaves.

Finally, a number of researchers compare and benchmark various methods on real and synthetic data: Devriendt et al. (2018), Gubela et al. (2019), Dorie et al. (2019), Gutierrez and Gerárdy (2017), and Olaya et al. (2020).

2.2 Open Data for Data Enrichment

Open data are one form of external data that can be used to enrich a given data set in order to improve business analytics, such as targeting models (Yan & Weber, 2018). The increasing volume of data being published as "open" creates growing potential to put open data into use and thereby improve business analytics applications and address new research problems (Hossain et al., 2016; Yadav et al., 2017; Zilioli et al., 2019). For example, Yadav et al. (2017) analyzed various open mobility data sets in nine smart cities (e.g., New York, Dublin, and Barcelona) and found that the data sets to be fostering organic innovation, improving the efficiency and effectiveness of services, and having a positive impact in various smart city-related domains by improving parking and traffic management, increasing environmental awareness, and heightening people's awareness of active options such as walking and cycling. They conclude that open data initiatives have a positive impact on society, and encourage other researchers and practitioners to maximize their utilization of open data resources.

Combining data from different sources can create new opportunities for analysis and ease further development initiatives (Janssen et al., 2012; Wu et al., 2018), but published studies demonstrating such advantages are sparse (Hopf et al., 2017). For example, Wu et al. (2018) combined various open data sources such as government data and climate data to predict the next dengue fever epidemic. They conclude that mining open data reveals the main drivers of the dengue fever epidemic (location and time) in ways that can be used to develop a dengue control strategy. Zheng et al. (2014) enriched data on complaints about noise with social media data, road network data, and points of interests to model the noise situation in New York City. Finally, Hopf et al. (2017) enriched an internal, company-owned data set with open government data to improve the prediction of household characteristics (i.e., size and type of dwelling, number of residents, and type of heating). They combined data on electricity consumption and household location with features extracted from OpenStreetMap on topology (e.g., frequency of objects, distance to city center), points of interest (e.g., frequency and distance to shops, cafes, etc.), buildings types, and types of land use. Overall, the authors found that "a large portion of the statistical data is hardly usable" because of the "low geographical granularity … and the low number of statistical data sets that are available for different countries" (Hopf et al., 2017, pp. 15–16).

While there has been some research on the potential of open data for data enrichment, the broader landscape of open data seems to be awaiting further exploration. Three systematically different areas of open data offer distinct opportunities for research: open government data; open research data; and collaborative open data projects/communities; First, regarding open data provided by governments, some large-scale initiatives were launched in the European Union (EU) in the past. The Infrastructure for Spatial Information in Europe (INSPIRE) initiative, regulated by Directive 2007/2/EC (Directive 2007/2/EC, 2007) defines a technical and organizational framework according to which EU member states are required to publish geospatial open data on 34 different themes, such as administrative units, hydrography, transport networks, geology, land use, and buildings. Although the number of available data sets and services still vary among member states (see https://inspire-geoportal.ec.europa.eu/overview.html for an overview), the technical standards (e.g., data models and XML schemas) and distribution infrastructure established by INSPIRE provide a harmonized high-quality source of open data.

Second, open research data is another growing opportunity for external data sources. For example, the "Principles and Guidelines for Access to Research Data from Public Funding" of the Organisation for Economic Co-operation and Development (OECD) (Pilat & Fukasaku, 2007) provide recommendations regarding, among others, openness, flexibility, transparency, interoperability, and quality of open research data. Similarly, the EU's Horizon 2020 research funding program implemented a research data sharing policy of "open by default, closed by exception" (Comission, 2017). Consequently, there are numerous research data sets already publicly available. The meta-registry re3data

(https://www.re3data.org) lists more than 2,600 research data registries worldwide that cover various research disciplines and data types. In addition, several data journals (e.g., *Scientific Data* and *Data in Brief)*, along with other types of publications in categories such as and "data paper", "data report", and "data descriptor" emerged in the publication landscape rather recently (Walters, 2020).

Collaborative projects are a third source of open data; examples include OpenStreetMap (https://www.openstreetmap.org), Wikidata (https://www.wikidata.org), and openSenseMap (https://opensensemap.org). In many of these projects, a group of volunteers form virtual communities that create and curate data within a common technical and organizational infrastructure for a shared goal or purpose, and mostly without any direct compensation (Preece, 2001; Rheingold, 2000; Sproull & Arriaga, 2012). Sometimes these virtual communities working on collaborative projects collect data for research projects, such as to monitor wildlife species (Locke et al., 2019), or to amass general geographic information (Haklay, 2013).

While structurally different, these three areas have in common that data sets are typically not readily available at the low level of granularity required to combine them with internal data, such as data on individual customers. They may, however, be transformed to fit internal data structurally and semantically, and thus add valuable information for business analytics.

3 Research Method

We present our research method in this section, including descriptions of the open data set used, our mining process for the socio-demographic features, the geospatial information, the data set provided by a German supplier, and the evaluation procedure of the targeting policies. Figure 1 illustrates the research methodology.

3.1 Open Data and Mining Socio-Demographic Features

To construct the open data set and mine additional socio-demographic features, we combined the voting results of the German parliamentary election of 2017 and the results of two general election population surveys. The idea was that respective parties have specific socio-demographic characteristic in common. For example, retired citizens tend to vote for Germany's Christian Democratic Party (CDU, a conservative party), while voters for Bündnis 90/Die Grünen (Germany's Green Party) are more likely to be university students (Bukow, 2017; Jung, 2019).

The German federal territory is divided into 299 parliamentary electoral districts, each with about 250,000 citizens (see https://www.bundeswahlleiter.de/en/service/glossar/w/wahlkreise.html). The electoral districts are again divided into roughly 70,000 constituencies (roughly the equivalent of precincts in U.S. elections) nationwide, each with a maximum of 2,500 citizens; voting results from each are downloadable from a government website (https://www.bundeswahlleiter.de/en/index.html). We only considered voting data from cities with more than 100,000 voters that thus share relatively high population densities. The members of the German parliament are elected in a two-vote system, with the first vote for direct candidates and the second vote to elect a party list in each state. We only used these second votes because they more specifically indicate voters' political attitudes and party preferences. We considered only votes for the following parties: Christliche Demokratische Union *(CDU)*, Christliche-Soziale Union *(CSU),* Sozial-demokratische Partei Deutschlands *(SPD),* Die Linke *(LINKE),* Bündnis 90/Die Grünen *(GRUENE),* Freie Demokratische Partei *(FDP),* and Alternative für Deutschland (*AFD*). The remaining parties were removed from the data set due to their low vote totals. For each constituency, we created the following eight voting features: the absolute number of votes for each party, the number of persons entitled to vote, and the election turnout.

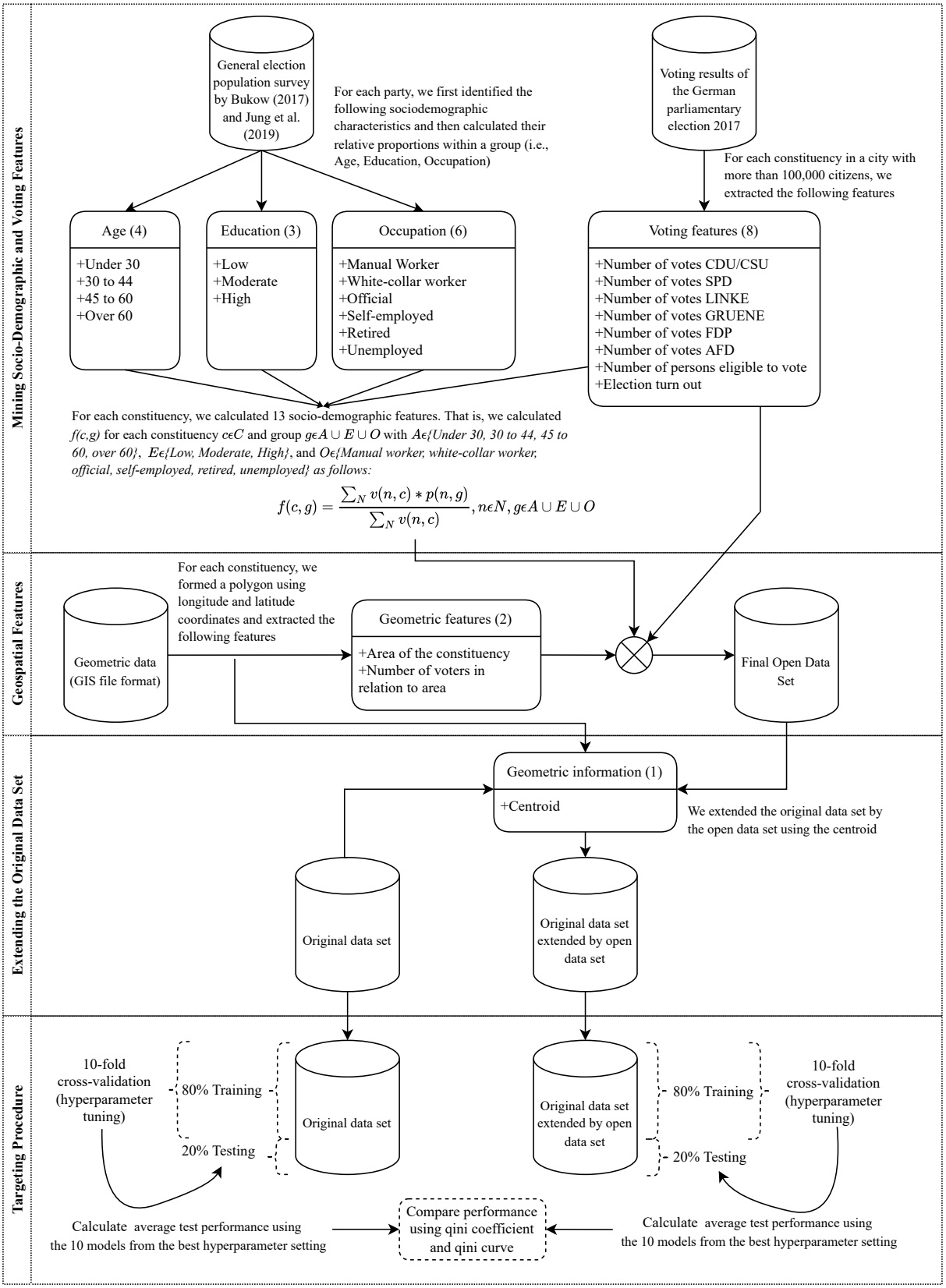

Figure 1. Overview of research methodology.

In addition to the results of the German parliamentary election, we took into account socio-demographic structures of different population groups by using the surveys of Jung et al. (2019) and Bukow (2017). More specifically, we focused on *age, education,* and *occupation* as socio-demographic characteristics and how these are distributed across the different parties. To unify the two surveys with respect to age, we reduced the five groups proposed in Bukow (2017) to four, distributing the values of the eliminated group equally among the adjacent age groups and leaving the following age groups: *under 30*, *30 to 44*, *45 to 60*, and *over 60*. We used the categories proposed for education by Bukow (2017), labeling the categories of Jung et al. (2019) as follows: *low* for completing basic school with a "leaving certificate"; *medium* for completing intermediate school with a "leaving certificate"; and *high* for completing a-level exams and earning a university degree. (A "leaving certificate" qualifies a student for the next level in the German education system.) In both surveys, voters are classified into the following six occupational categories: *manual worker*, *white-collar worker*, *official*, *self-employed*, *retired*, and *unemployed*. Finally, we calculated the relative proportions across the respective groups of a socio-demographic aspect (i.e., age, education, and occupation) individually for all political parties (i.e., CDU/CSU, SPD; LINKE, GRUENE, FDP, and AFD).

Tables 1, 2, and 3 summarize the relative ratio of voters across the socio-demographic aspects and their groups. For example, consider 100 voters who voted for the FDP. According to the relative proportions listed in Table 1, the following distribution of voters can be derived in the age groups: 28 voters are under age 30; 25 voters are between 30 and 44 years old; 24 voters are between 45 and 60 years old; and 23 voters are over age 60.

Age	CDU/CSU	SPD	LINKE	GRUENE	FDP	AFD
Under 30	0.19	0.23	0.29	0.30	0.28	0.23
30 to 44	0.24	0.21	0.25	0.28	0.25	0.29
45 to 60	0.25	0.26	0.24	0.28	0.24	0.28
Over 60	0.32	0.30	0.22	0.14	0.23	0.20

Table 1. Relative proportions across age groups by political party (calculated based on Bukow (2017) and Jung et al. (2019)).

Education	CDU/CSU	SPD	LINKE	GRUENE	FDP	AFD
Low	0.37	0.41	0.22	0.17	0.25	0.35
Moderate	0.33	0.31	0.37	0.25	0.31	0.43
High	0.30	0.28	0.41	0.58	0.44	0.22

Table 2. Relative proportions across education groups by political party (calculated based on Bukow (2017) and Jung et al. (2019)).

Occupation	CDU/CSU	SPD	LINKE	GRUENE	FDP	AFD
Manual worker	0.14	0.19	0.18	0.09	0.12	0.24
White-collar worker	0.17	0.16	0.16	0.19	0.16	0.14
Official	0.19	0.16	0.11	0.25	0.18	0.12
Self-employed	0.18	0.10	0.16	0.19	0.28	0.14
Retired	0.22	0.20	0.16	0.11	0.15	0.13
Unemployed	0.10	0.19	0.23	0.18	0.10	0.23

Table 3. Relative proportions across occupation groups by political party (calculated based on Bukow (2017) and Jung et al. (2019)).

We created 13 socio-demographic features for each constituency: four with respect to age (i.e., under 30, 30 to 44, 44 to 60, over 60); three concerning education (i.e., low, moderate, high), and six with respect to occupation (i.e., manual worker, white-collar worker, official, self-employed, retired, unemployed). The following formula was used to generate the feature $f(c, g)$ for each constituency $c \in C$ and group $g \in A \cup E \cup O$ with $A \in \{under\ 30,\ 30\ to\ 44,\ 45\ to\ 60,\ Over\ 60\}$, $E \in \{low,\ moderate,\ high\}$, and $O \in \{manual\ worker,\ white\text{-}collar\ worker,\ official,\ self\text{-}employed,\ retired,\ unemployed\}$:

$$f(c, g) = \frac{\sum_N \big(v(n, c) * p(n, g)\big)}{\sum_N v(n, c)}, n \in N, g \in A \cup E \cup O \tag{2}$$

where $N \in \{CDU/CSU,\ SPD,\ LINKE,\ GRUENE,\ FDP,\ AFD\}$, v is the total number of votes a political party n received in a given constituency $c \in C$, and p is the relative socio-demographic group g of party $n \in N$ we calculated in the previous step (i.e., Tables 1 to 3). For example, let $\tilde{g}$ be the group *under 30.* In addition, let the absolute votes received by each political party in a random selected constituency $\tilde{c}$ be as follows: $v(CDU/CSU, \tilde{g}) = 145$, $v(SPD, \tilde{g}) = 125$, $v(LINKE, \tilde{g}) = 76$, $v(GRUENE, \tilde{g}) = 212$, $v(FDP, \tilde{g}) = 89$, $v(AFD, \tilde{g}) = 25$. The outcome for $f(c = \tilde{c}, g = \tilde{g})$ according to Table 1 is 0.26. The feature provides information regarding the relative proportion of voters under 30 for a given constituency. More specifically, it states that the probability of belonging to this age group is 26% for a randomly selected voter in this specific constituency.

3.2 Geospatial Data

Geospatial information was used to join the socio-demographic features with the given data set. More specifically, we utilized several latitude-longitude coordinates to define a constituency in the form of a polygon. As we describe in the next section, the internal data set contains various households in Germany, including their latitude-longitude coordinates. Thus, we can easily enrich the internal data set with the voting and socio-demographic features we just created by assigning the household a constituency according to its latitude-longitude coordinates.

In Germany, government departments are responsible for building, hosting, and maintaining the spatial arrangement of constituencies. The majority of geospatial data is published under *Creative Commons* or *Data license Germany* and was made available on an open data platform of the authority. Where the required geospatial data was not yet available via open data, we requested the information directly from the authorities responsible for open government, elections, and statistics.

The geometry data sets of the constituencies are available in machine-readable geographic information system (GIS) file formats, for example, as *ESRI shapefile* or *GeoJSON*. The files contain the geometry data, that is, points, polygons, or multi-polygons, as well as relevant information regarding their coordinate reference system (CRS). At a geometric level, each constituency is defined by latitude-longitude coordinates of several points. All points are assembled into a polygon and represent the shape of a constituency. Further, we calculated the corresponding shape area, the number of voters in relation to the area, and the centroid for all constituencies. While we used the area and the number of voters in relation to the area as additional features, the centroid for all constituencies was used to enrich the internal data set with the voting and socio-demographic features of the constituency with the shortest distance between its centroid and the household-level latitude-longitude coordinates.

3.3 Marketing Data Set

For the application, we utilized an internal data set (referred to in the following as the *original data set*) from a grocery supplier in Germany. Data were collected during a print marketing campaign to investigate how a discount offer could help acquire new customers. Individuals were randomly divided into control and treatment groups. All customers in the treatment group were subject to the same treatment: a discount on their next purchase. The costs for each treatment, including expenses for sending the offer via mail and the discount itself, were 20€. Customers in the control group received no purchasing incentive. After three weeks, the company tracked who did or did not purchase a product.

The data set comprises 47,659 households: 17,069 in the treatment group and 30,590 in the control group. The response rates in both groups are low, with 1% (treatment group) and 0.5% (control group), respectively. The overall uplift – that is, the difference between treatment response rate and control response rate – is 0.5%. Overall, the data set is imbalanced, with more non-responders than responders in both groups: 169 treatment responder (0.35%), 16,900 treatment non-responder (35.46%), 161 control responder (0.34%), and 30,429 control non-responder (63.85%).

The internal data set has 23 features including binary and categorical features covering general consumer behavior (e.g., level of consumption or interest in fashion), information regarding the household (e.g., duration of residence) and house (e.g., size of the backyard), as well as affinities (e.g., affinity for print media). In contrast, the *enriched data set* contains 46 features, as we enriched the original data set by eight voting features, 13 socio-demographic features, and two geospatial features. Thus, the enriched data set contains additional information about the constituency in which the household is located and, more specifically, the distributions of age, education, and occupation within the constituency. For example, one can assess whether a constituency has a relatively high or low level of education, whether residents are more likely to be unemployed or retired, and whether residents tend to be younger or older. Further, the geospatial features cover the size of the constituency, both in terms of square meters and in terms of the population eligible to vote.

3.4 Targeting Procedure

In order to ensure a comprehensive comparison between the original and enriched data sets, we decided to evaluate various nonparametric, well-established targeting policies. More specifically, we implemented the approach by Sołtys et al. with Euclidean distance as distribution divergence measure (D-ED) (Sołtys et al., 2015), the X-Learner (TM-X) (Künzel et al., 2019), the R-Learner (MC-R) (Nie & Wager, 2021), and the Bayesian Causal Forest approach (D-BCF) (Hahn et al., 2020). Note that TM-X- and MC-R- were based on Random Forest as an underlying algorithm. Hyperparameter tuning has been applied for two parameters and for each approach: *n_estimators,* and *max_depth.* Other hyperparameters were set to default. D-ED, TM-X, and MC-R were implemented using causalml (Chen et al., 2020) while D_BCF was implemented using the XBCF package (see https://github.com/socket778/XBCF).

Typically, the performance of predictive models is measured by comparing actual values with predicted values. However, when evaluating the effect of targeting policies on real-world data sets, the ground truth is missing because an individual cannot be in both the treatment and control groups simultaneously. This is called the fundamental problem of causal inference (Holland, 1986).

To make hyperparameter tuning and final evaluation reliable, we used the following procedure. We split both data sets – original and enriched – into 80% training and 20% testing while stratifying on the treatment and response variable. Next, we used the training samples and 10-fold cross-validation to select the best hyperparameters for each approach, again, using stratification with the treatment and response variable. Predictions and qini-related metrics were computed for each approach, data set, and validation fold. Subsequently, the metrics were averaged such that we know which hyperparameter setting yields the highest average unscaled qini coefficient given the approach and data set. The metrics – that is, the deciles of the qini curve and the unscaled qini coefficient (UQC) – are commonly used in uplift modeling (Gubela et al., 2019). The qini curve is a decile-based evaluation metric that plots the cumulative number of incremental response, also called uplift, as a function of the number of targeted individuals ranked by the model from high to low (Radcliffe & Surry, 2011). In each decile, we compare the k percent highest scores of the treatment and control groups and estimate the uplift as the difference in response rates between these two groups. Thus, the higher the response rate in the treatment group compared to the response rate in the control group, the higher the uplift value for a decile. The unscaled qini coefficient serves as a single number metric; it is defined as the ratio of the area under the actual qini curve to the area under the diagonal, random curve (Radcliffe & Surry, 2011). If the value is greater than 1, the actual model is better than random targeting. In general, the higher the value, the better the uplift modeling approach. Finally, we compared and evaluated the models based on the average test

performance. That is, we used the 10 models from cross-validation with the best hyperparameter settings, estimated their scores on the independent test sample, and calculated the average values across the ten models.

The final results on the test set are evaluated with a statistical test to detect whether the differences in performance are significantly different. Following the approach suggested by Demsar (2006), we performed a nonparametric Friedman test (Friedman, 1940) followed by a post hoc Nemenyi test (Nemenyi, 1963). The null hypothesis of the Friedman test states that there are no significant differences in performance across all models. We attempted to reject this hypothesis with an alpha value of $\alpha = 0.05$. After rejecting the Friedman test, we performed a post hoc Nemenyi test to compare the models trained on the original data set with the models trained on the enriched data set.

4 Results

Table 4 summarizes the average UQC and their significances for all approaches on both data sets. In the following section, we elaborate on these findings.

We found that each method performed better when trained on the enriched data set versus when trained on the original data set. D-ED achieved an average UQC of 1.3936 on the original data set and an average UQC of 1.7511 (*difference to original:* 0.3575) on the enriched data set. TM-X performed worse on the original data set, with an average UQC of 1.5468, than on the enriched data set, with an average UQC of 1.6915 (0.1447). MC-R obtained an average UQC of 1.3979 on the original data set and an average UQC of 1.4574 (0.0595) on the enriched data set. Finally, D-BCF achieved an average UQC of 1.6277 on the original data set and an average UQC of 1.6574 (0.0297) on the enriched data set. Note that the D-ED approach trained on the enriched data set was the best performing model, followed by TM-X, D-BCF, and MC-R.

Approach	**Unscaled Qini Coefficient**	
	Original Data Set	**Enriched Data Set**
Random forest with Euclidean distance distribution divergence (D-ED) (Sołtys et al., 2015)	1.3936	1.7511***
X-Learner (TM-X) (Künzel et al., 2019)	1.5468	1.6915
R-Learner (MC-R) (Nie & Wager, 2021)	1.3979	1.4574
Bayesian Causal Forest (D-BCF) (Hahn et al., 2020)	1.6277	1.6574

*Table 4. Average unscaled qini coefficient on the original and enriched test data sets. The higher the value, the better. Note: ***p<0.01*

The Friedman test was applied to the results shown in Table 4 to check for the existence of statistical differences among the performances of the different methods. The estimated p-value for this test was smaller than 0.01; thus, we conclude that there is at least one statistically significant difference in performance among the algorithms. Subsequently, we applied the post hoc Nemenyi test to compare all of the models trained on the original data set with all of the models trained on the enriched data set. We found that D-ED is the only approach for which the difference in performance between original and enriched data set is statistically significant with a p-value of 0.0032. The other methods achieved p-values greater than 0.1.

Figure 2 – showing the performance of all algorithms on original (dark green line) and enriched (purple line) data set – illustrates these findings using average qini curves. Recall that the qini curve plots the cumulative uplift (y-axis), here in relative numbers, as a function of the fraction of people targeted from the campaign's total population (x-axis). The incremental number of responses is a helpful indicator that informs practitioners and analysts in charge regarding the relationship between the response rates in the treatment and control groups. The higher the uplift, the greater the response rate in the treatment group compared to that in the control group.

We can clearly see that the areas under the average qini curves in the enriched plot are larger than the areas under the average qini curves in the original plot for all approaches. Further, as a result of enriching the data set with open data sources, we can exceed the uplift value of 0.0047 prior to the last decile – namely in the fourth decile using the D-ED approach, with an uplift value of 0.005. The other approaches also achieve higher uplift values on the same deciles when trained on the enriched data set versus when trained on the original data set. For example, in the third decile, TM-X achieved an average uplift value of 0.0038 when trained on the original data set and an average uplift value of 0.0046 when trained on the enriched data set.

Multiple targeting policies can be deployed that increase the revenue from the marketing campaign by reducing the number of individuals targeted and increasing the number of customers buying products. Note that targeting all individuals results in expenses of 341,380€ (17,069 individuals treated at a cost of 20€ each). The best strategy is to target only 40% of the individuals, thus reducing the number of contacts by 60%, from 17,069 to 6,828. Compared to targeting all individuals, the company can save 204,820€ while obtaining 6% more response (average uplift value targeting all individuals: 0.0047). Compared to the same approach trained on the original data set, response rates have increase by more than 30% (average uplift value targeting 40% using D-ED trained on original data: 0.0033). In another targeting policy, we can use the TM-X approach trained on the enriched data set to target only 30% of the customers, thus, reducing the number of contacts from 17,069 to 5,121 while achieving an uplift of 0.0046. Compared to targeting all customers we can save 238,960€ while obtaining marginally less uplift (difference in uplift: -0.0001). However, compared to the TM-X approached trained on the original data set that is an increase in uplift by almost 18% (average uplift value targeting 30% using TM-X trained on original data: 0.0038).

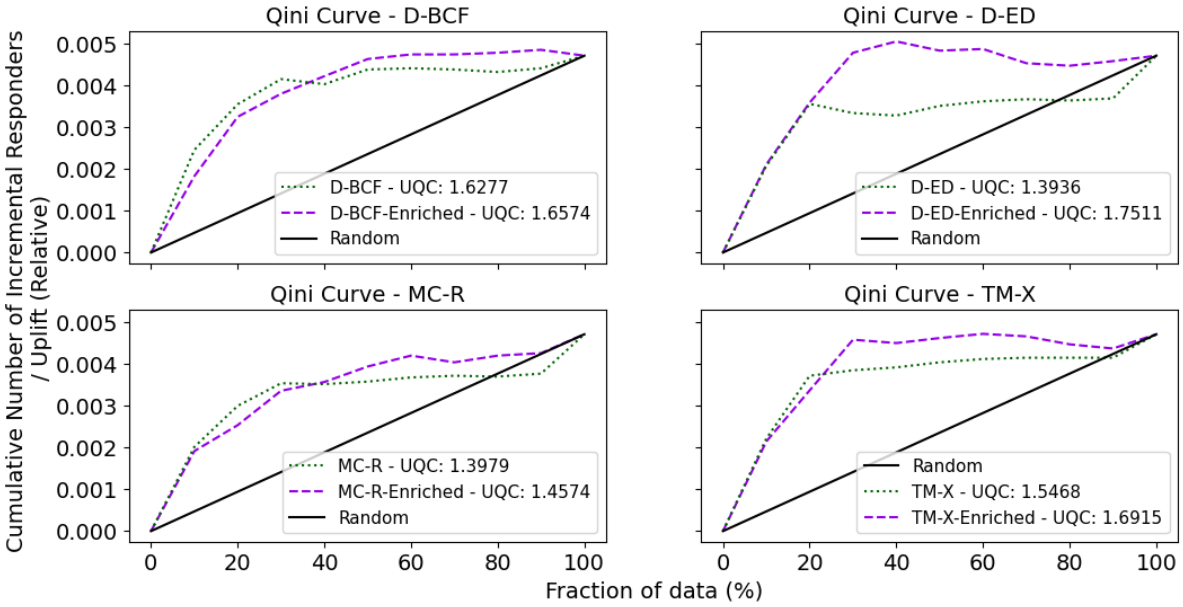

Figure 2. *Average qini curves for each approach trained on enriched (dark green) and original (purple) data sets.*

5 Discussion and Conclusion

With the open data movement picking up speed, academia and practice have the possibility to put open data into use and thereby create new analysis opportunities and address new research problems. However, whether and how open data can be used to enrich internal data sets to improve business analytics remain open questions. The findings of this study illustrate that open data can help advance analysis. We enriched a data set from a marketing campaign of a German grocery supplier using open data sources to improve targeting policies. We found that the enriched data set improved the response rate of the targeting policies by more than 30%.

5.1 Contribution to Research

Prior studies concluded that the outcome of a data mining problem is also determined by the underlying data set (Baesens et al., 2009; Simester et al., 2019). Further, researchers have examined how enriching a given data set by external data sources can improve applications and services (Afonso et al., 2019; Baecke & Van den Poel, 2011; D'Haen et al., 2013; Zheng et al., 2014). They also acknowledge the use of open data (Wu et al., 2018). However, use cases demonstrating that enriching a given data set using open data sources to improve the overall outcome (e.g., response rate) are rare – especially in the targeting policy literature (Hopf et al., 2017). This study addresses this gap by showing that targeting policies can be improved by incorporating open data sources. Further, other IS and marketing researchers are informed through our study about the potential impact of open data on business analytics. While we demonstrate the benefits of open data when acquiring new customers, we believe that open data can also have a huge impact on other use cases, not only in marketing (e.g., churn prevention or cross- or up-selling) but also in political science (e.g., political campaigns) and health care (e.g., estimating the effect of a specific drug). Finally, our study contributes to the sparse literature that demonstrates the usefulness of open government data at a low level of granularity – that is, aggregated data – motivating the analysis of other aggregated data sources such as research and collaborative open data in a business analytics context.

5.2 Contribution to Practice

In practice, many companies lack sufficient (individual-level) data to apply proper data mining tools (Simester et al., 2019). This is especially true with respect to firms trying to acquire new customers, because there is no purchase history for individuals who have not previously bought a product or a service. Further, in churn management, the risk of targeting customers who churn because of the treatment, referred to as *sleeping dogs* or *do-not-disturbs* (Devriendt et al., 2018), makes it difficult to target the proper individuals. In both cases, customers who should be targeted need to be selected carefully, as the costs of the targeting can easily exceed the benefits because of various costs, including the cost of the effort to select potential candidates, contact costs, and costs when individuals accept the treatment. Thus, firms need data sets that are sufficient in terms of predictive power to enable and improve targeting policies.

We recommend leveraging open data sources to enrich a given data set. With our study, we demonstrated that an individual-level, internal data set enriched by open data sources leads to better targeting policies. Thus, our study informs analysts in charge of targeting policies and other practitioners about the important role of open data. As more and more open data becomes available, using open data is an especially important consideration.

5.3 Limitations and Future Research

We are aware that our research has some limitations that also serve as excellent avenues for future research. First, we were constrained to evaluate our idea using a single data set. Thus, we invite other researchers to demonstrate the effectiveness of enriching their data sets with open data sources not only using government data but also open research data (e.g., vegetation plot database of the Ecological Society of America's Panel on Vegetation Classification) or data from collaborative projects (e.g., points of interesting using OpenStreetMap).

Second, adding more (open) data does not necessarily lead to better results (Moro et al., 2017). Although our results look promising, practitioners and researchers must not take it for granted that enriching a given data set will always improve the outcome. Rather, we encourage other scholars to search for open data sources that improve the understanding of the problem at hand.

Third, while we demonstrate how open data can be used to improve a targeting policy, companies and researchers alike need to consider who to target carefully, using the results of the analysis. Methods such as partial dependencies, feature importance (of tree-based algorithms), or simply covariates coefficients could be used to shed light on customer characteristics that are related to a change in behavior because

of a treatment. For example, it might be that customers who live in a particular constituency with a high share of voters for a particular party respond more favorably to a treatment.

Fourth, while we demonstrated the potential impact of government and research open data in a targeting policies context, we invite other IS researchers to analyze systematically the impact of (aggregated) open data on three levels: open data sources (i.e., governmental, research, and collaborative); domain of interest (e.g., marketing, political science, health care); and business analytics method (e.g., supervised vs. unsupervised learning, classification, regression).

Finally, it was a labor-intensive task to collect and combine the data sources, for three reasons: the various ways in which open data is shared (e.g., open data platform, email attachments, or other portals); the many different formats of constituencies (e.g., ESRI shapefile or GeoJSON) and CRSs (e.g., Universal Transverse Mercator, Gauss Krueger, or World Geodetic System); and different licenses (e.g., dl-de/by-2-0, CC BY 3.0 DE, or CC BY 4.0 DE). In asking academia and practice to provide more open data sources, we urge consistency, easy access, and shareability in the open data movement. Without these conditions, open data use cases will remain rare and, thus, open data awareness will continue to be limited.

References

Afonso, B., Melo, L., Oliveira, W., Sousa, S., & Berton, L. (2019). Housing Prices Prediction with a Deep Learning and Random Forest Ensemble. *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 389–400. https://doi.org/10.5753/eniac.2019.9300

Alaa, A. M., Weisz, M., & van der Schaar, M. (2017). Deep Counterfactual Networks with Propensity-Dropout. *ArXiv:1706.05966 [Cs, Stat]*. http://arxiv.org/abs/1706.05966

Ascarza, E. (2018). Retention Futility: Targeting High-Risk Customers Might be Ineffective. *Journal of Marketing Research*, *55*(1), 80–98. https://doi.org/10.1509/jmr.16.0163

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178. https://doi.org/10.1214/18-AOS1709

Baecke, P., & Van den Poel, D. (2011). Data augmentation by predicting spending pleasure using commercially available external data. *Journal of Intelligent Information Systems*, *36*(3), 367–383. https://doi.org/10.1007/s10844-009-0111-x

Baesens, B., Mues, C., Martens, D., & Vanthienen, J. (2009). 50 years of data mining and OR: Upcoming trends and challenges. *Journal of the Operational Research Society*, *60*(sup1), S16–S23. https://doi.org/10.1057/jors.2008.171

Bukow, S. (2017). *Bundestagswahl 2017: Ergebnisse und Analysen*. Heinrich-Böll-Stiftung. https://www.boell.de/de/2017/09/28/bundestagswahl-2017-ergebnisse-und-analysen

Chen, H., Harinen, T., Lee, J.-Y., Yung, M., & Zhao, Z. (2020). CausalML: Python Package for Causal Machine Learning. *ArXiv:2002.11631*. http://arxiv.org/abs/2002.11631

Comission, E. (2017). *Guidelines to the rules on open access to scientific publications and open access to research data in horizon 2020*.

Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *The Journal of Machine Learning Research*, *7*, 1–30.

Devriendt, F., Moldovan, D., & Verbeke, W. (2018). A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone Toward the Development of Prescriptive Analytics. *Big Data*, *6*(1), 13–41. https://doi.org/10.1089/big.2017.0104

D'Haen, J., Van den Poel, D., & Thorleuchter, D. (2013). Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique. *Expert Systems with Applications*, *40*(6), 2007–2012. https://doi.org/10.1016/j.eswa.2012.10.023

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition. *Statistical Science*, *34*(1), 43–68. https://doi.org/10.1214/18-STS667

Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), (2007). https://eur-lex.europa.eu/eli/dir/2007/2/oj

Farrell, M. H., Liang, T., & Misra, S. (2021). Deep Learning for Individual Heterogeneity: An Automatic Inference Framework. *ArXiv:2010.14694 [Cs, Econ, Math, Stat]*. http://arxiv.org/abs/2010.14694

Friedman, M. (1940). A Comparison of Alternative Tests of Significance for the Problem of m Rankings. *The Annals of Mathematical Statistics*, *11*(1), 86–92.

Gubela, R., Bequé, A., Lessmann, S., & Gebert, F. (2019). Conversion Uplift in E-Commerce: A Systematic Benchmark of Modeling Strategies. *International Journal of Information Technology & Decision Making*, *18*(03), 747–791. https://doi.org/10.1142/S0219622019500172

Gutierrez, P., & Gérardy, J.-Y. (2017). Causal Inference and Uplift Modelling: A Review of the Literature. *Proceedings of The 3rd International Conference on Predictive Applications and APIs*, *67*, 1–13. https://proceedings.mlr.press/v67/gutierrez17a.html

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, *15*(3), 965–1056. https://doi.org/10.1214/19-BA1195

Haklay, M. (2013). Citizen Science and Volunteered Geographic Information: Overview and Typology of Participation. In D. Sui, S. Elwood, & M. Goodchild (Eds.), *Crowdsourcing Geographic Knowledge* (pp. 105–122). Springer Netherlands. https://doi.org/10.1007/978-94-007-4587-2_7

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Hopf, K., Riechel, S., Sodenkamp, M., & Staake, T. (2017). Predictive Customer Data Analytics – The Value of Public Statistical Data and the Geographic Model Transferability. *ICIS 2017 Proceedings*. https://aisel.aisnet.org/icis2017/DataScience/Presentations/9

Hossain, M. A., Dwivedi, Y. K., & Rana, N. P. (2016). State-of-the-art in open data research: Insights from existing literature and a research agenda. *Journal of Organizational Computing and Electronic Commerce*, *26*(1–2), 14–40. https://doi.org/10.1080/10919392.2015.1124007

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, *29*(4), 258–268. https://doi.org/10.1080/10580530.2012.716740

Jung, M. (2019). Bedingt regierungsbereit – Eine Analyse der Bundestagwahl 2017. In K.-R. Korte & J. Schoofs (Eds.), *Die Bundestagswahl 2017: Analysen der Wahl-, Parteien-, Kommunikations- und Regierungsforschung* (pp. 23–45). Springer Fachmedien. https://doi.org/10.1007/978-3-658-25050-8_2

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116

Link, G., Lumbard, K., Germonprez, M., Conboy, K., & Feller, J. (2017). Contemporary Issues of Open Data in Information Systems Research: Considerations and Recommendations. *Communications of the Association for Information Systems*, *41*, 587–610. https://doi.org/10.17705/1CAIS.04125

Locke, C. M., Anhalt-Depies, C. M., Frett, S., Stenglein, J. L., Cameron, S., Malleshappa, V., Peltier, T., Zuckerberg, B., & Townsend, P. A. (2019). Managing a large citizen science project to monitor wildlife. *Wildlife Society Bulletin*, *43*(1), 4–10. https://doi.org/10.1002/wsb.943

Moro, S., Cortez, P., & Rita, P. (2017). A framework for increasing the value of predictive data-driven models by enriching problem domain characterization with novel features. *Neural Computing and Applications*, *28*(6), 1515–1523. https://doi.org/10.1007/s00521-015-2157-8

Nemenyi, P. B. (1963). *Distribution-free multiple comparisons.* Princeton University.

Nie, X., & Wager, S. (2021). Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *Biometrika*, *108*(2), 299–319. https://doi.org/10.1093/biomet/asaa076

Olaya, D., Coussement, K., & Verbeke, W. (2020). A survey and benchmarking study of multitreatment uplift modeling. *Data Mining and Knowledge Discovery*, *34*(2), 273–308. https://doi.org/10.1007/s10618-019-00670-y

Pilat, D., & Fukasaku, Y. (2007). OECD principles and guidelines for access to research data from public funding. *Data Science Journal*, *6*, OD4–OD11.

Preece, J. (2001). Sociability and usability in online communities: Determining and measuring success. *Behaviour & Information Technology*, *20*(5), 347–356. https://doi.org/10.1080/01449290110084683

Radcliffe, N. J., & Surry, P. D. (2011). Real-World Uplift Modelling with Significance-Based Uplift Trees. *White Paper TR-2011-1, Stochastic Solutions*, 1–33.

Rheingold, H. (2000). *The Virtual Community, revised edition: Homesteading on the Electronic Frontier*. MIT press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rzepakowski, P., & Jaroszewicz, S. (2012). Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, *32*(2), 303–327. https://doi.org/10.1007/s10115-011-0434-0

Schwab, P., Linhardt, L., & Karlen, W. (2019). Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks. *ArXiv:1810.00656 [Cs, Stat]*. http://arxiv.org/abs/1810.00656

Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. *International Conference on Machine Learning*, 3076–3085. http://proceedings.mlr.press/v70/shalit17a.html

Sharma, R., Mithas, S., & Kankanhalli, A. (2014). Transforming decision-making processes: A research agenda for understanding the impact of business analytics on organisations. *European Journal of Information Systems*, *23*(4), 433–441. https://doi.org/10.1057/ejis.2014.17

Shi, C., Blei, D. M., & Veitch, V. (2019). Adapting Neural Networks for the Estimation of Treatment Effects. *ArXiv:1906.02120 [Cs, Stat]*. http://arxiv.org/abs/1906.02120

Simester, D., Timoshenko, A., & Zoumpoulis, S. I. (2019). Targeting Prospective Customers: Robustness of Machine-Learning Methods to Typical Data Challenges. *Management Science*, *66*(6), 2495–2522. https://doi.org/10.1287/mnsc.2019.3308

Simester, D., Timoshenko, A., & Zoumpoulis, S. I. (2020). Efficiently Evaluating Targeting Policies: Improving on Champion vs. Challenger Experiments. *Management Science*, *66*(8), 3412–3424. https://doi.org/10.1287/mnsc.2019.3379

Sołtys, M., Jaroszewicz, S., & Rzepakowski, P. (2015). Ensemble methods for uplift modeling. *Data Mining and Knowledge Discovery*, *29*(6), 1531–1559. https://doi.org/10.1007/s10618-014-0383-9

Sproull, L., & Arriaga, M. (2012). Online Communities. In H. Bidgoli (Ed.), *Handbook of Computer Networks* (pp. 898–914). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118256107.ch58

Su, X., Kang, J., Fan, J., Levine, R., & Yan, X. (2012). Facilitating Score and Causal Inference Trees for Large Observational Studies. *Journal of Machine Learning Research*, *13*, 2955–2994.

Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, *33*(2), 325–337. https://doi.org/10.1016/j.giq.2016.02.001

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Walters, W. H. (2020). Data journals: Incentivizing data access and documentation within the scholarly communication system. *Insights the UKSG Journal*, *33*, 18. https://doi.org/10.1629/uksg.510

Wu, C., Kao, S.-C., Shih, C.-H., & Kan, M.-H. (2018). Open data mining for Taiwan's dengue epidemic. *Acta Tropica*, *183*, 1–7. https://doi.org/10.1016/j.actatropica.2018.03.017

Yadav, P., Hasan, S., Ojo, A., & Curry, E. (2017). The Role of Open Data in Driving Sustainable Mobility in Nine Smart Cities. *Research Papers*, 1248–1263.

Yan, A., & Weber, N. (2018). Mining Open Government Data Used in Scientific Research. In G. Chowdhury, J. McLeod, V. Gillet, & P. Willett (Eds.), *Transforming Digital Worlds* (pp. 303–313). Springer International Publishing. https://doi.org/10.1007/978-3-319-78105-1_34

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., & Zhang, A. (2018, January 1). *Representation Learning for Treatment Effect Estimation from Observational Data*. NeurIPS. https://openreview.net/forum?id=By-yGDZ_bS

Zhao, Y., Fang, X., & Simchi-Levi, D. (2017). A Practically Competitive and Provably Consistent Algorithm for Uplift Modeling. *2017 IEEE International Conference on Data Mining (ICDM)*, 1171–1176. https://doi.org/10.1109/ICDM.2017.157

Zheng, Y., Liu, T., Wang, Y., Zhu, Y., Liu, Y., & Chang, E. (2014). Diagnosing New York city's noises with ubiquitous data. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 715–725. https://doi.org/10.1145/2632048.2632102

Zilioli, M., Lanucara, S., Oggioni, A., Fugazza, C., & Carrara, P. (2019). Fostering Data Sharing in Multidisciplinary Research Communities: A Case Study in the Geospatial Domain. *Data Science Journal*, *18*, 15. https://doi.org/10.5334/dsj-2019-015