Assessing the Demand Effects of Discounts on Swiss Train Tickets

Martin Huber*
Hannes Wallimann+
Jonas Meier‡

May 2021

Project funded by SBB research fund

Acknowledgment: We are indebted to Pierre Chevalier, Reto Lüscher and Philipp Wegelin for their helpful comments and discussions.

*University of Fribourg, Dept. of Economics and Center for Econometrics and Business Analytics, St. Petersburg State University
+University of Applied Sciences and Arts Lucerne, Competence Center for Mobility
‡University of Bern, Dept. of Economics

Contents

List of Figures

List of Tables

1 Management Summary

This project aims at quantifying the demand effects of discounts on train tickets (the so-called supersaver tickets) issued by the Swiss Federal Railways. To this end, we analyze a survey-based sample of buyers of supersaver tickets based on machine learning methods, a subfield of artificial intelligence.

In a first step, we investigate which customer- or trip-related characteristics (such as the age and gender of the customer, the time or length of the trip, as well as the discount rate itself) importantly predict various aspects of the customers' buying behavior, namely: booking a trip otherwise not realized by train, buying a first- rather than second-class ticket, or rescheduling a trip (e.g. away from rush hours) when being offered a supersaver ticket. In a nutshell, we find that several characteristics like customer's age, the discount level, departure time, and capacity utilization of a trip importantly predict whether supersaver customers book additional trips, reschedule their trips or buy first- rather than second-class tickets. Yet, the quality of the predictions based on the characteristics available in the data differs importantly across the buying behaviors analyzed. While (only) 58% of the decisions to reschedule a trip are correctly predicted, the correct prediction rates amount to 65% for booking an additional trip, and 82% for buying a first- rather than second-class ticket (upselling). Such a predictive machine learning approach can be useful for a customer segmentation, i.e. to classify buyers of supersaver tickets into groups according to their likely demand patterns, e.g. individuals inclined or not inclined to upselling as a function of their characteristics. This might be helpful as a base for designing particular policies or interventions, e.g. for adapting marketing efforts to target specific customer segments.

In our main analysis, we apply so-called causal machine learning to assess the impact of the discount rate on rescheduling a trip, which seems relevant in the light of capacity constraints at rush hours. Typically, it is challenging to assess the impact of discounts on consumer behavior due to two underlying selection problems. First, train connections with low and high demand (e.g. rush vs. off-peak hours) likely differ in terms of the offered discounts such that their effects cannot be easily separated from other factors driving demand. Second, consumers buying a supersaver ticket with a higher discount generally differ from those already buying it at a lower discount in terms of (background characteristics driving) their reservation price or customers paying the regular price, who are not even available in our sample. For this reason, customer groups are not comparable across discount levels. Our study overcomes the first selection problem by controlling for a rich set of trip-related characteristics (e.g. capacity utilization) that capture the varying baseline demand. This implies that only trips that are similar in these characteristics, but vary in terms of discounts are compared to each other to assess the causal effects of interest. To tackle the second selection problem, we only focus at individuals who would have bought the ticket within the same (second or first) class irrespective of the offered discount. As these so-called 'always-buyers' are homogeneous in terms of their buying decision and reservation price, we can evaluate the causal effect of the discounts in this subgroup under the behavioral assumption that anyone purchasing a trip without discount also buys it when being offered a discount. Using statistical parlance, we assume that among the always-buyers (i) the discount rate is quasi-random conditional on our rich set of trip-related characteristics and (ii) the buying decision cannot decrease in the discount rate. We note that the second assumption can even be scrutinized in the data and we do not find evidence for its violation. Nevertheless, our estimated effects should be taken with a grain of salt as they refer to a subgroup of consumers only.

Our results suggest that a one percentage point increase in the discount on average raises the share of rescheduled trips by 0.16 percentage points among the always-buyers. Furthermore, we also find some evidence that the effects might differ depending on specific customer- and trip-related characteristics. For instance, the impact of discounts on the standard ticket price tends to be higher among leisure travelers and during peak hours (when simultaneously controlling for several other characteristics). Therefore, our study provides at least two key insights in the context of Swiss railway transportation. First, it presents the first empirical evidence suggesting that discounts causally induce a subgroup of supersaver customers to reschedule their trip. Second, it points to the likely possibility that the effectiveness of discounts differs across various customer segments. Such a knowledge on which consumers are most (or least) affected by offering a discount in terms of their demand behavior can be helpful for tailoring policies (e.g. marketing campaigns) to specific groups or trips in order to optimize the effectiveness of supersaver tickets.

As a further analysis based on the causal machine learning approach, we assess whether discounts increase customer satisfaction among always buyers according to survey responses. We find positive but very minor effects, suggesting that discounts of 30% or more on average increase customer satisfaction by 0.16 points (on a scale of 1 to 10) among always buyers relative to discounts of less than 30%.

We also consider the discount's effect on upselling, but point out for that this kind of analysis and sample definition, data checks point to a violation of the assumptions required for assessing causal effects. We find that a one percentage point increase in the discount is associated with a rise of 0.589 percentage points in the share of upgraded tickets in our sample. However, given that the assumptions underlying our causal approach are likely not satisfied, this result should be interpreted with much caution.

As a further word of caution concerning our approach, we emphasize that with the current data, it is delicate to address questions about overall revenue (or utilization) as a base for a classical cost-benefit analysis. As our evaluation relies on buyers of supersaver tickets only, it is not representative for the total of railway customers. Therefore, we recommend considering the following options for future evaluations. First, surveying a random sample of customers with single tickets (rather than buyers of discounted tickets) would permit analyzing the demand effects across all customer groups (with single tickets) and account for the likely problem that buyers of supersaver tickets are not comparable to buyers at regular fares. Second, randomizing the discounts (at least for specific departure times and/or destinations) by means of an experiment would help assessing the causal effects without running into the previously mentioned selection problems. Third, increasing the number of surveyed individuals would allow for a more thorough investigation of whether and how the consumer responses differ across subgroups of consumers (e.g. leisure travelers).

Summing up, our study presents novel evidence on how discounts affect consumer behavior in the Swiss railway setting, in particular the willingness to reschedule a trip among customers that would have bought the trip even without discounts. This may provide helpful insights for the future design of discount policies as well as empirical studies to further investigate customer behavior.

2 Introduction

Organizing public transport involves a well-known trade-off between consumer welfare and provider revenue. Typically, consumers value frequency, reliability, space, and low fares (Redman, Friman, Gärling, and Hartig, 2013) while suppliers aim at operating with a minimum number of vehicles to maximize profits. In general, the allocation can be improved as providers do not account for the positive externalities on consumers (Mohring, 1972). In particular, service frequency reduces travelers' access and waiting costs. This so-called 'Mohring-effect' leads to economies of scale, implying the need for subsidies to achieve the first-best solution in terms of welfare. Consequently, it may be socially optimal to subsidize railway companies to reduce fares (Parry and Small, 2009). To assess such a measure's effectiveness on demand, policy-makers would need to know how individuals respond to lower fares. However, it is generally challenging to identify causal effects of discounts on train tickets (or goods and services in general) due confounding or selection. For instance, discounts might typically be provided for dates or hours with low train utilization such that connections with and without discount are not comparable in terms of baseline demand. A naive comparison of sold tickets with and without discount would therefore mix the influence of the discount with that of baseline demand. In this context, we apply machine learning (a subfield of artificial intelligence) to convincingly assess how discounts on train tickets for long-distance connections in Switzerland, the so-called 'supersaver tickets', affect demand, by exploiting a unique data set of the Swiss Federal Railways (SBB) that combines train utilization records with a survey of supersaver buyers. Since customer satisfaction is an important issue for both public transport companies and policy-makers, we also discuss the effects of supservater tickets on customer satisfaction.

More concisely, our study provides four use cases of machine learning for business analytics in the railway industry: (i) Predicting buying behavior among supersaver customers, namely whether customers booked a trip otherwise not realized by train (additional trip), bought a first-class rather than a second-class ticket (upselling), or reschedule their trip, e.g., away from rush hours (demand shift); (ii) analyzing the causal effect of the discount on demand shifts among customers that would have booked the trip even without discount; (iii) analyzing the causal effect of discounts on customer satisfaction among among those booking the trip even without discounts. Use cases (ii) and (iii) are feasible because our unique survey contains information on how supersaver buyers would have decided in the absence of a discount. This is, whether they are so-called 'always buyers' and would have booked the connection even at the regular fare and in the same class. Ignoring the latter condition of bookings in the same class and assuming that individuals upselling their second-class to a first-class ticket are part of the always buyers, we (iv) assess the effect of discounts on upselling. In the case of (iv), however, the results need to interpreted with caution, as the assumptions required for a causal analysis are likely not satisfied due to the more lenient definition of always buyers that might include individuals which are apriori not comparable in terms of their buying behavior.

For all use cases, we use appropriately tailored machine learning techniques, which learn the associations between the demand outcomes of interest, the discount rate, and further customer or trip-related characteristics in a data-driven way in order to avoid model misspecification. Such a targeted combination of predictive and causal machine learning can improve demand forecasting and decision-making in companies and organizations. While predictive machine learning permits optimizing forecasts about demand and customer behavior as a function of observed characteristics, causal machine learning permits evaluating the causal effect of specific interventions like a discount regime for optimizing the offer of such discounts. Concerning the prediction task, we use the so-called random forest, see Breiman (2001), as a machine learner to forecast the supersaver customers' behavior.

Concerning the use cases (ii) to (iv), we rely on two identifying assumptions. First, we impose a selection-on-observables assumption implying that the discount rate is as good as randomly assigned when controlling for our rich set of trip- and demand-related characteristics. Second, we invoke weak monotonicity of any individual's decision to purchase an additional trip (otherwise not realized) in the discount rate, implying that a higher (rather than lower) discount either positively or not affects any customer's buying decision. As a methodological contribution, we formally show how these assumptions permit tackling the selectivity of discount rates and survey response to identify the discount rate's effect on demand shifts (rescheduling away from rush hours) and customer satisfaction for the subgroup 'always buyers'. We define the latter based on the survey information on how customers would have behaved in the absence of a discount. In addition, we discuss testable implications of monotonicity, namely that among all survey respondents, the share of additional trips must increase in the discount rate. Furthermore, the selection on observables assumptions requires that, conditional on trip- and demand-related characteristics, the discount must not be associated with personal characteristics (like age or gender) among always buyers.

We estimate the marginal effect of slightly increasing the (continuously distributed) discount rate based on the causal forest (CF), see Wager and Athey (2018) and Athey, Tibshirani, and Wager (2019), for use cases (ii) to (iv). As a second approach, for these use cases, we apply double machine learning (DML) to assess the effects of a binary definition of discount rates based on splitting the latter into two discount categories of less than 30% (relative to the regular fare) and 30% or more. Further, we investigate the heterogeneity of effects across all of our observed characteristics using the CF. In a second heterogeneity analysis, we investigate whether effects differ systematically across a pre-selected set of characteristics, namely: age, gender, possession of a half fare travel card, travel distance, whether the purpose is business, commute, or leisure, and whether the departure time is during peak hours. To this end, we use the regression approach of Semenova and Chernozhukov (2020).

Our study is related to a growing literature applying statistical and machine learning methods for analyzing transport systems, as well as to methodological studies on causal inference for so-called principal strata, see Frangakis and Rubin (2002), i.e. endogenous subgroups like the always buyers. Typically, it is hard to identify the causal effect of some treatment (or intervention) like a discount on such a non-randomly selected subgroup defined in terms how a post-treatment variable (e.g. buying decision) depends on the treatment (e.g. treatment). One approach is to give up on point identification and instead derive upper and lower bounds on a set of possible effects for groups alike the always buyers based on the aforementioned monotonicity assumption (and possibly further assumptions about the ordering of outcomes of always buyers and other individuals), see for instance Zhang and Rubin (2003), Zhang, Rubin, and Mealli (2008), Imai (2008), Lee (2009), and Blanco, Flores, and Flores-Lagunes (2011). Alternatively, the treatment effect on always buyers is point-identified when invoking a selection-on-observables or instrumental variable assumption for selection into the survey, see for instance Huber (2014), which requires sufficiently rich data on both survey participants and non-participants for modeling survey participation. In contrast to these previous studies, the approach in this paper point-identifies the treatment effect by exploiting the rather unique survey feature that customers were asked about their behavior in the absence of the discount, which under monotonicity permits identifying the principal stratum of always buyers directly in the data.

Furthermore, our work is related to conceptual studies on transport systems, considering for instance the previously mentioned positive externalities of an increased service for customers that are not accounted for by transportation providers. Such externalities typically arise from economies of scale due to fixed costs and a 'Mohring effect', implying that service frequency reduces waiting costs (Mohring, 1972). The study by Parry and Small (2009) suggests that lower fares can boost overall welfare by increasing economies of scale (off-peak) and decreasing pollution and accidents (at peaks). Similarly, De Palma, Lindsey, and Monchambert (2017) argue that time-dependent ticket prices may increase overall welfare as overcrowding during peak hours is suboptimal for both consumers and providers. As public transport is usually highly subsidized, governments may directly manage the trade-off mentioned above. As this involves taxpayer money, it is a question of general interest how the subsidies should be designed. Based on their results, Parry and Small (2009) conclude that even substantial subsidies are justified due to lower fares' positive welfare effect. In contrast, Basso and Silva (2014) find that the contribution of transit subsidies to welfare diminishes once congestion is taxed and alternatives are available, i.e., bus lanes. Irrespective of the specific policy instrument, the consumer's willingness to shift demand drives these policies' effectiveness. While many factors affect this willingness, most studies conclude that consumers are price sensitive (Paulley, Balcombe, Mackett, Titheridge, Preston, Wardman, Shires, and White, 2006). In this context, we aim at contributing to a better understanding of how time-dependent pricing translates to consumer decisions and well-beings.

More broadly, our paper relates to the literature on policies targeting demand shifts. Among these, the setting of car parking costs, fiscal regulations, or even free public transport has been analyzed (e.g. Batty, Palacin, and González-Gil, 2015, Rotaris and Danielis, 2014, Zhang, Lindsey, and Yang, 2018, De Witte, Macharis, Lannoy, Polain, Steenberghen, and Van de Walle, 2006). Another stream of literature applies machine learning algorithms in the context of public transport. Examples are short-term traffic flow forecasts for bus rapid transit (Liu and Chen, 2017) or metro (Liu, Liu, and Jia, 2019) services. Further, Hagenauer and Helbich (2017) and Omrani (2015) implement machine learning algorithms to predict travel mode choices. Yap and Cats (2020) predict disruptions and their passenger delay impacts for public transport stops. In other research fields, also applications of causal (rather than predictive) machine learning are on the rise (see for instance Yang, Chuang, and Kuan, 2020, Knaus, 2021). This is, to the best of our knowledge, the first study using causal machine learning in the context of public transport. Finally, a growing literature discusses the opportunities of data-driven business decision-making (Brynjolfsson and McElheran, 2016) by assessing the relevance of predictive and causal machine learning. Ascarza (2018) and Hünermund, Kaminski, and Schmitt (2021) show that companies may gain by designing their policies based on causal machine learning. For instance, firms can target the relevant consumers much more effectively when accounting for their heterogeneity in terms of reaction to a treatment. Our study provides a use case of how the machine learning-based assessment of discounts could be implemented also in other businesses and industries facing capacity constraints.

This study proceeds as follows.[1] Section 3 presents the institutional setting of passenger railway transport in Switzerland. Section 4 describes our data, coming from a unique combination of a customer survey and transport utilization data. Section 5 discusses the identifying assumptions underlying the causal machine learning approach as well as testable implications. Section 6 outlines the predictive and causal machine learning methods. Section 7 presents the empirical results of all use cases. Section 8 concludes. Appendices A to E provide supplementary material, namely further tables and graphs on the analyses conducted.

[1] Part of this study has already been published as a working paper, see Huber, Meier, and Wallimann (2021). This report includes further analysis and materials compared to the working paper.

3 Institutional Background

The railway system in Switzerland is known for its high quality of service. Examples include the high level of system integration with frequent services, synchronized timetables, and comprehensive fare integration, see Desmaris (2014). In Switzerland, a country of railway tradition, the state owned incumbent Swiss Federal Railways (SBB) operates the long distance passenger rail market as monopolist (Thao, von Arx, and Frölicher, 2020). Furthermore, nationally operating long-distance coaches may only be approved if they do not 'substantially' compete with existing services. Thus, the SBB competes exclusively with motorized private transport in Swiss long-distance traffic. The company also owns most of the rail infrastructure, which is funded by the Federal Government. However, since the end of 2020 the companies Berne-Lötschberg-Simplon Railways (BLS) and Southeast Railways (SOB) operate a few links on behalf of the SBB. Different to regional public transport that Swiss taxpayers subsidize with approximately CHF 1.9 bn per year, the operation of the long distance public transport itself has to be self-sustaining (Wegelin, 2018).

Because of the monopoly position of the SBB in long distance passenger transport, the prices are screened by the Swiss 'price watchdog' (or price monitoring agency) to prevent abuse. Based on the price monitoring act, the watchdog keeps a permanent eye on how prices and profits develop. By the end of 2014, the watchdog concluded that the SBB charged too high prices. As a consequence and through a mutual agreement, the SBB and the Swiss price watchdog agreed on a significantly higher supply of supersaver tickets, which were first offered in 2009. Using a supersaver ticket, customers can travel on long distance public transport routes with a discount of up to 70%. Thereafter, additional agreements were regularly reached regarding number and scope of the supersaver tickets. While only a few thousand supersaver tickets were sold in 2014, sales increased to about 8.8 million in 2019, see Lüscher (2020).

From the SBB's perspective, these tickets can serve two purposes. First, the tickets might be used as means to balance out the utilization of transport services. For instance, supersaver tickets could reduce the high demand during peak hours which is a key challenge for public transport. Thus, balancing the demand may reduce delays and increase the number of free seats which is valued by the consumers. The average load of SBBs' seats amounts to 30% in the long distance passenger transport.[2] For this reason, there is in the literal sense room for improving the allocation. Second, price sensitive customers

[2]See https://reporting.sbb.ch/verkehr.

can be acquired during off-peak hours at rather negligible marginal costs.

Despite the increasing interest in the supersaver tickets in recent years, many users of the Switzerland public transport network purchase a so-called 'general abonnement' travel ticket (GA). This (annually renewed) subscription provides free and unlimited access to the public transport network in Switzerland. In 2019, about 0.5 million individuals owned a GA in Switzerland, roughly 6% of the Swiss population. The GA's cost amounts to 3,860 and 6,300 Swiss francs for the second and first class, respectively. In the same year, about 2.7 million individuals held a relatively cheap half fare travel ticket amounting to 185 Swiss francs. The latter implies a price reduction of 50% for public transport tickets in Switzerland. Overall, discounts provided through supersaver tickets are slightly lower for owners of half fare tickets, as the SBB aims to attract non-regular public transport users. In our causal analysis, we therefore also control for the possession of a half fare ticket.

4 Data

To investigate supersaver tickets' effect, we use a unique cross-sectional data set provided by the SBB. Our sample consists of randomly surveyed buyers of supersaver tickets that purchased their tickets between January 2018 and December 2019. These survey data are matched with data on distances between any two railway stops as well as utilization-related information relevant for the supply and calculation of discounts. In section 7, we provide descriptive statistics for these data.

4.1 Survey Data

The customer survey is our primary data source. It for instance includes the outcome variables 'demand shift' (use case (ii)), a binary indicator of whether an interviewee rescheduled her or his trip due to buying a supersaver ticket. 'Yes' means that the departure time has been advanced or postponed because of the discount. A second outcome variable (use case (iii)) characterizing customer behavior is an indicator for upselling, i.e. whether someone purchased a first rather than a second class ticket as a reaction to the discount. Another question asks whether an interviewee would have bought the train trip in the same or a higher class even without being offered a discount, which permits judging whether an additional trip has been sold through offering the discount and allows identifying the subgroup of always buyers under the assumptions outlined further below. Our continuously distributed treatment variable is the discount rate of a supersaver ticket relative to the standard fare, which may take positive values of up to 70%.

Furthermore, we observe two kinds of covariates, namely trip- or demand-related factors and personal characteristics of the interviewee. The former are important control variables for our causal identification strategy outlined below and include the difference between the days of purchase and travel, the weekday, month, and year, an indicator for buying a half fare ticket, departure time, peak hour,[3] number of tickets purchased per person, class (first or second), indicators for leisure trips, commutes, or business trips, the number of companions (by children and adults if any) and a judgment of how complicated the ticket purchase was on a scale from 1 (complicated) to 10 (easy). Furthermore, it consists of indicators for the point of departure, destination, and public holidays. The personal characteristics include age, gender, migrant status, language (German, French, Italian), and indicators for owning a half fare travel ticket or other subscriptions like those of regional tariff associations, specific connections, and Gleis 7 ('rail 7'). The latter is a travelcard for young adults not older than 25, providing free access to public transport after 7pm.

4.2 Factors Driving the Supply of Supersaver Tickets

In addition to the survey, we have access to factors determining the supply of supersaver tickets with various discounts. This is crucial for our causal analysis that hinges on on controlling for all characteristics jointly affecting the the discount rate and the outcome. While information on the distances between railway stops in Switzerland is publicly available,[4] the SBB provides us for the various connections with information on utilization data, the number of offered seats, and contingency schemes, which define the quantity of offered discounts. This allows us to account for travel distance, offered seats, capacity utilization, and quantities of offered supersaver tickets for various discount levels as well as quantities of supersaver tickets already sold (both quantities at the time of purchase). Furthermore, we create binary indicators for the 27 different contingency schemes of the SBB present in our data, which change approximately every month.

The variables listed in the previous paragraph are important, as the SBB

[3]Peak hour is defined as a departure time between 6am and 8:59am or between 4pm and 18:59pm, from Monday to Friday. This time windows is chosen on the base of the SBB's train-path prices. For further details, see https://company.sbb.ch/en/sbb-as-business-partner/services-rus/onestopshop/services-and-prices/the-train-path-price.html (assessed on March 24 2021).

[4]See the Open Data Platform of the SBB: https://data.sbb.ch/explore/dataset/linie-mitbetriebspunkte (accessed on March 24 2021).

calculates the supply of supersaver tickets based on an algorithm considering four type of inputs: Demand forecasts, advance booking deadlines, number of supersaver tickets already sold, and contingency schemes defining the amount and the size of offered discounts based on the three previous inputs. The schemes are set as a function of the SBB's self-imposed goals such as customer satisfaction but also depend on the requirements imposed by the price watchdog. The algorithm calculates a journey's final discount as a weighted average of all discounts between any two adjacent railway stops along a journey. The weights depend on the distances of the respective subsections of the trip. To approximate the (not directly available) demand forecasts of the SBB, we consider the quarterly average of capacity utilization and the number of offered seats for any two stops, which are available by (exact) departure time, workday, class, and weekend. In addition, we make use of indicators for place of departure, destination, month, year, weekday and public holidays. We use this information to reconstruct the amount and size of offered discounts by taking values from the contingency schemes that correspond to our demand forecast approximation combined with the difference between buying and travel days. Comparing this amount and size of offered discounts with a buyer's discount, we estimate the number of supersaver tickets already sold for the exact date of purchase.

4.3 Sample Construction

Our initial sample contains 12,966 long-distance train trips that cover 61,469 sections between two adjacent stops. For 12.2% of these sections, there is no information on the capacity utilization available, which can be due to various reasons. First, for some cases, capacity utilization data is missing. Second, passengers traveling long-distance may switch to regional transport in exceptional cases causing problems for determining utilization. A further reason could be issues in data processing. Altogether, missing information occurs in 3,967 trips of our initial sample. We tackle this problem by dropping all journeys with more than 50% of missing information, which is the case for 320 trips or 2.5% of our initial sample. After this step, our evaluation sample consists of 12,646 trips. For the remaining 3,647 trips with missing information (which now account for a maximum of 50% of all sections of a journey), we impute capacity utilization as the average of the remaining sections of a trip. In our empirical analysis, we include an indicator for whether some trip information has been imputed as well as the share of imputed values for a specific trip as control variables. Finally, we note that our causal analysis makes (in contrast to the predictive analysis) only use of a subsample, namely observations identified as always buyers who would have traveled even without a discount, all in all 6,112 observations.

5 Identification

We subsequently formally discuss the identification strategy and assumptions underlying our causal analysis of the discounts among always buyers.

For simplification we discuss the identification strategy of the use case (ii), that is analyzing the effect of discounts on demand shift. The identification strategy remains the same for customer satisfaction in use case (iii). However, the definition of the subgroup always buyers changes in use case (iv). We discuss this modification when presenting the empirical results in Section 7.

5.1 Definition of Causal Effects

Let D denote the continuously distributed treatment 'discount rate' and Y the outcome 'demand shift', a binary indicator for rescheduling a trip due to being offered a discount. More generally, capital letters represent random variables in our framework, while lower case letters represent specific values of these variables. To define the treatment effects of interest, we make use of the potential outcome framework, see for instance Rubin (1974). To this end, $Y(d)$ denotes the potential outcome hypothetically realized when the treatment is set to a specific value d in the interval $[0, Q]$, with 0 indicating no discount and Q indicating the maximum possible discount. For instance, $Q = 0.7$ would imply the maximum discount of 70% of a regular ticket fare. The realized outcome corresponds to the potential outcome under the treatment actually received, i.e. $Y = Y(D)$, while the potential outcomes under discounts different to one received remain unknown without further statistical assumptions.

A further complication for causal inference is that our survey data only consists of individuals that purchased a supersaver ticket, a decision that is itself an outcome of the treatment, i.e. the size of the discount. Denoting by S a binary indicator for purchasing a supersaver ticket and by $S(d)$ the potential buying decision under discount rate d, this implies that we only observe outcomes Y for individuals with $S = 1$. In general, making the survey conditional on buying introduces Heckman-type sample selection (or collider) bias, see Heckman (1976) and Heckman (1979), if unobserved characteristics affecting the buying decision S also likely affect the inclination of shifting the timing of the train journey Y. Furthermore, it is worth noting that $S = S(D)$ implies that buying a supersaver ticket is conditional on receiving a non-zero discount. For this reason, non-treated subjects paying regular fares (with $D = 0$) are not observed in our data. Yet, the outcome in our sample is defined relative to the behavior without treatment, as Y indicates whether a has passenger has changed the timing of the trip because of a discount. This implies that $Y(0) = 0$ by definition, such that the causal effect of some positive discount d vs. no discount is $Y(d) - Y(0) = Y(d)$ is directly observable among observations that actually received d. However, it also appears interesting to investigate whether the demand shift effect varies across different (non-zero) discount rates $d \in (0, Q]$ to see whether the size matters. This is complicated by the fact that supersaver customers with different discount rates that are observed in our data might in general differ importantly in terms of background characteristics also affecting the outcome, exactly because they bought their trip and were selected into the survey under non-comparable discount regimes. Our causal approach aims at tackling exactly this issue to establish customer groups that are comparable across discount rates in order to identify the effect of the latter.

Based on the potential notation, we can define different causal parameters of interest. For instance, the average treatment effect (ATE) of providing discount levels d vs. d' (for $d \neq d'$) on outcome Y, denoted by $\Delta(d, d')$, corresponds to

$$\Delta(d, d') = E[Y(d) - Y(d')]. \tag{1}$$

Furthermore, the average partial effect (APE) of marginally increasing the discount level at $D = d$, denoted by $\theta(d)$, is defined as

$$\theta(d) = E\left[\frac{\partial Y(D)}{\partial D}\right]. \tag{2}$$

Accordingly, $\theta(D)$ corresponds to the APE when marginally increasing the actually received discount of any individual (rather than imposing some hypothetical value d for everyone).

The identification of these causal parameters based on observable information requires rather strong assumptions. First, it implies that confounders jointly affecting D and Y can be controlled for by conditioning on observed characteristics. In our context, this appears plausible, as treatment assignment is based on variables related to demand (like weekdays or month), contingency schemes, capacity utilization, and supersaver tickets already sold - all of which is available in our data, as described in section (4). Second, identification requires that selection S is as good as random (i.e., not associated with outcome Y) given the observed characteristics and the treatment, an assumption known as missing at random (MAR), see for instance Rubin (1976) and Little and Rubin (1987). However, the latter condition appears unrealistic in our framework, as our data lack important socio-economic characteristics likely affecting preferences and reservation prices for public transport, namely education, wealth, or income. For this reason, we argue that the ATE and APE among the individuals selected for the survey ($S = 1$), i.e. conditional on buying a supersaver ticket, which are defined as

$$\Delta_{S=1}(d, d') = E[Y(d) - Y(d')|S = 1], \theta_{S=1}(D) = E\left[\frac{\partial Y(D)}{\partial D}\right], \tag{3}$$

cannot be plausibly identified either. The reason is that if an increase in the discount rate induces some customers to buy a supersaver ticket, then buyers with lower and higher discounts will generally differ in terms of their average reservation prices and related characteristics (as education or income), which likely also affect the demand-shift outcome Y.

To tackle this sample selection issue, we exploit the fact that our data provide information on whether the supersaver customers would have purchased a ticket for this specific train trip also in the absence of any discount. Provided that the interviewees give accurate responses, we thus have information on $S(0)$, the hypothetical buying decision without treatment. Under the assumption that each customer's buying decision is weakly monotonic in the treatment in the sense that anyone purchasing a trip in a specific travel class (e.g., second class) without discount would also buy it for that class in the case of any positive discount, this permits identifying the group of always buyers. Importantly, we therefore define always buyers as those that would buy the trip not in a lower travel class (namely second rather than first class) without discount. For alway buyers, $S(0) = S(d) = 1$ for any $d > 0$, such that their buying decision is always one and thus not affected by the treatment, implying the absence of the selection problem. In the denomination of Frangakis and Rubin (2002), the always buyers constitute a so-called principal stratum, i.e., a subpopulation defined in terms of how the selection reacts to different treatment intensities. Therefore, sample selection bias does not occur within such a stratum, in which selection behavior is by definition homogeneous. For this reason, we aim at identifying the ATE and APE on the always buyers:

$$\begin{aligned} \Delta_{S(0)=1}(d, d') &= E[Y(d) - Y(d')|S(0) = 1] \\ &= E[Y(d) - Y(d')|S(0) = S(d'') = 1] \text{ for } d'' \in (0, Q], \\ \theta_{S(0)=1}(D) &= E\left[\frac{\partial Y(D)}{\partial D}\Big|S(0) = 1\right] \\ &= E\left[\frac{\partial Y(D)}{\partial D}\Big||S(0) = S(d'') = 1\right] \end{aligned} \tag{4}$$

where the second equality follows from the monotonicity of S in D that is formalized further below.

Figure 1: Causal framework

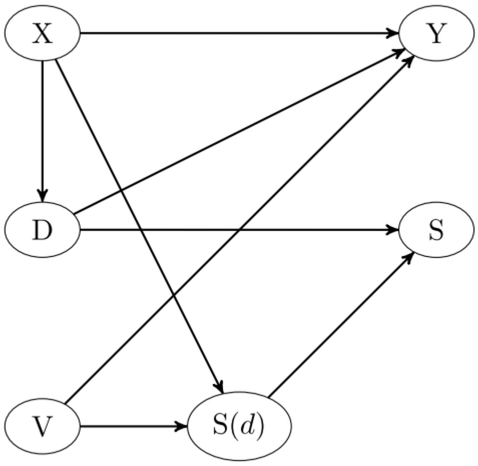


Figure 1 provides a graphical illustration of our causal framework based on a directed acyclic graph, with arrows representing causal effects. Observed covariates X that are related to demand are allowed to jointly affect the discount rate D and the demand-shift outcome Y. X may influence the potential purchasing decision under a hypothetical treatment $S(d)$, implying that buying a ticket given a specific discount depends on observed demand drivers like weekday, month, etc. Furthermore, unobserved socio-economic characteristics V (like the reservation price) likely affect both $S(d)$ and Y. This introduces sample selection when conditioning on S, e.g. by only considering survey respondents ($S = 1$). We also note that S is deterministic in D and $S(d)$ (as $S = S(D)$), even when controlling for X. This is the case because conditional on $S = 1$, D is associated with V, which also affects Y, thus entailing confounding of the treatment-outcome relation. A reason for this is for instance that buyers under higher and lower discounts are generally not comparable in terms of their reservation prices. In the terminology of Pearl (2000), S is a collider that opens up a backdoor path between D and Y through V. Theoretically, this could be tackled by jointly conditioning on the potential selection states under treatment values d vs. d' considered in the causal analysis, namely $S(d), S(d')$, as controls for the selection behavior. This is typically not feasible in empirical applications when only the potential selection corresponding to the actual treatment assignment is observed, $S = S(D)$. In our application, however, we do have information on $S(0)$ and can thus condition on being an always buyer under the mentioned monotonicity assumption.

5.2 Identifying Assumptions

We now formally introduce the identification assumptions underlying our causal analysis.

Assumption 1 (identifiability of selection under non-treatment):
$S(0)$, is known for all subjects with $S = 1$.

Assumption 1 is satisfied in our data in the absence of misreporting, as subjects have been asked whether they would have bought the train trip even in the absence of discount.

Assumption 2 (conditional independence of the treatment):
$Y(d), S(d) \perp D|X$ for all $d \in (0, Q]$.

By Assumption 2, there are no unobservables jointly affecting the treatment assignment on the one hand and the potential outcomes or selection states under any positive treatment value on the other hand conditional on covariates X. This assumption is satisfied if the treatment is quasi-random conditional on our demand-related factors X. Note that the assumption also implies that $Y(d) \perp D|X, S(0) = 1$ for all $d \in (1, Q]$.

Assumption 3 (weak monotonicity of selection in the treatment):
$\Pr(S(d) \geq S(d')|X) = 1$ for all $d > d'$ and $d, d' \in (1, Q]$.

By Assumption 3, selection is weakly monotonic in the treatment, implying that a higher treatment state can never decrease selection for any individual. In our context, this means that a higher discount cannot induce a customer to not buy a ticket that would have been purchased under a lower discount. An analogous assumption has been made in the context of nonparametric instrumental variable models, see Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996), where, however, it is the treatment that is assumed to be monotonic in its instrument. Note that monotonicity implies the testable implication that $E[S - S(0)|X, S = 1, D = d] = E[(1 - S(0)|X, S = 1, D = d]$ weakly increases in treatment value d. In words, the share of customers that bought the ticket because of the discount must increase in the discount rate in our survey population when controlling for X.

Assumption 4 (common support):
$f(d|X, S(0) = 1) > 0$ for all $d \in (1, Q]$.

Assumption 4 is a common support restriction requiring that $f(d|X, S(0) = 1)$, the conditional density of receiving a specific treatment intensity d given X and $S(0) = 1$ (or conditional probability if the treatment takes discrete values), henceforth referred to as treatment propensity score, is larger than zero among always buyers for the treatment doses to be evaluated. This implies that the demand-related covariates X do not deterministically affect the discount rate received such that there exists variation in the rates conditional on X.

Our assumptions permit identifying the conditional ATE given X (CATE), denoted by $\Delta_{X,S(0)=1}(d,d') = E[Y(d) - Y(d')|X, S(0) = 1]$ for $d \neq d'$ and $d, d' \in (1, Q]$. To see this, note that

$$\begin{aligned}\Delta_{X,S(0)=1}(d,d') &= E[Y|D=d, X, S(0)=1] - E[Y|D=d', X, S(0)=1],\\ &= E[Y|D=d, X, S(0)=1, S=1]\\ &- E[Y|D=d', X, S(0)=1, S=1],\end{aligned} \tag{5}$$

where the first equality follows from Assumption 2 and the second from Assumption 3, as monotonicity implies that asymptotically, $S = 1$ if $S(0) = 1$. Together with Assumption 1, which postulates the identifiability of $S(0)$, it follows that the causal effect on always buyers is nonparametrically identified, given that common support (Assumption 4) holds. If follows that the ATE among always buyers is identified by averaging over the distribution of X given $S(0) = 1, S = 1$:

$$\begin{aligned}\Delta_{S(0)=1}(d,d') &= E[E[Y|D=d, X, S(0)=1, S=1]\\ &-E[Y|D=d', X, S(0)=1, S=1]|S(0)=1, S=1].\end{aligned} \tag{6}$$

Furthermore, considering (5) and letting $d - d' \to 0$ identifies the conditional average partial effect (CAPE) of marginally increasing the treatment at $D = d$ given $X, S(0) = 1$, denoted by $\theta_{X,S(0)=1}(D) = E\left[\frac{\partial Y(D)}{\partial D}\middle| |X, S(0) = 1\right]$:

$$\theta_{X,S(0)=1}(d) = \frac{\partial E[Y|D=d, X, S(0)=1, S=1]}{\partial D}. \tag{7}$$

Accordingly, the APE among always buyers that averages over the distributions of X and D is identified by

$$\theta_{S(0)=1}(D) = E\left[\frac{\partial E[Y|D, X, S(0)=1, S=1]}{\partial D}\right]. \tag{8}$$

Our identifying assumptions yield a testable implication if some personal characteristics (like customer's age) that affect $S(d)$ are observed, which we henceforth denote by W. In fact, D must be statistically independent of W conditional on $X, S(0) = 1, S = 1$ if X is sufficient for avoiding any cofounding of the treatment-outcome relation. To see this, note that personal characteristics must by Assumption 2 not influence the treatment decision conditional on X. This statistical independence must also hold within subgroups (or principal strata) in which sample selection behavior (and thus sample selection/collider bias) is controlled for like the always buyers, i.e. conditional on $S(d), S = 1$.

6 Estimation based on machine learning

In this section, we outline the predictive and causal machine learning approaches used in our empirical analysis of the evaluation sample. For simplification, we again present the machine learning approaches used to analyze the outcome 'demand shift' for both, predictive and causal (use case (ii)) machine learning.

6.1 Predictive Machine Learning

Let $i \in \{1, ..., n\}$ be an index for the different interviewees in our sample of size n and $\{Y_i, D_i, X_i, W_i, S_i(0)\}$ denote the outcome, treatment, the covariates related to the treatment and the outcome, the observed personal characteristics, and the buying decision without discount of these interviewees that by the sampling design all satisfy $S_i = 1$ (because they are part of the survey). Therefore, Y_i represents customer i's demand shift (rescheduling behavior) under customer i's received discount rate D_i relative to no discount. We in a first step investigate which observed predictors among the covariates X, W as well as the size of the discount D are most powerful for predicting demand shifts by machine learning algorithms. We point out that this analysis is of descriptive nature as it does not yield the causal effects of the various predictors, but merely their capability of forecasting Y. In particular, our approach averages the predictions of Y over different levels of treatment intensity D and thus different customer types in terms of reservation price (related to $S(0)$) and unobserved background characteristics that likely vary with the treatment level.

Therefore, we also perform the prediction analysis within subgroups defined upon the treatment level to see whether the set of important predictors is affected by the treatment intensity. To this end, we binarize the treatment such that it consists of two categories, namely (non-zero) discounts below 30%, i.e. covering the treatment range $d \in (0, 0.3)$, and more substantial discounts of 30% and more, $d \in [0.3, 0.7]$, as 70% is the highest discount observed in our data. In the same manner, we also assess the predictive power when considering the decision to buy a trip that would not have been realized without discount (additional trip), i.e. $S_i - S_i(0)$, as outcome. As $S_i = 1$ is equal to one for everyone, the outcome corresponds to $1 - S_i(0)$ and indicates whether someone has been induced purchase the ticket because of the discount, i.e. is not an always buyer. As a further consumer behavior-related outcome to be predicted, we also consider buying a first class rather than second class ticket because of the discount (upselling).

Prediction is based on the random forest, a nonparametric machine learner suggested by Breiman (2001) for predicting outcomes as a function of covariates. Random forests rely on repeatedly drawing subsamples from the original data and averaging over the predictions in each subsample obtained by a decision tree, see Breiman, Friedman, Olshen, and Stone (1984). The idea of decision trees is to recursively split the covariate space, i.e. the set of possible values of X, W, into a number of non-overlapping subsets (or nodes). Recursive splitting is performed such that after each split, a statistical goodness-of-fit criterion like the sum of squared residuals, i.e. the difference between the outcome and the subset-specific average outcome, is minimized across the newly created subsets. Intuitively, this can be thought of as a regression of the outcome on a data-driven choice of indicator functions for specific (brackets of) covariate values. At each split of a specific tree, only a random subset of covariates is chosen as potential variables for splitting in order to reduce the correlation of tree structures across subsamples, which together with averaging predictions overall subsamples reduces the estimation variance of the random forest when compared to running a single tree in the original data. Even when using an excessive number of splits (or indicator functions for covariate values) such that some of them do not importantly predict the outcome, averaging over many samples will cancel out those non-predictive splits that are only due to sampling noise. Forest-based predictions can be represented by smooth weighting functions that bear some resemblance with kernel regression, with the important difference that random forests detect predictive covariates in a data-driven way. We use the *randomforest* package by Breiman (2018) for the statistical software R to implement the random forest based on growing 1,000 decision trees.

6.2 Causal Machine Learning

Our second part of the analysis assesses the causal effect of increasing discount rates on demand shifts among always buyers while controlling for the selection into the survey and the non-random assignment of the treatment based on Assumptions 1 to 4 of section 5.[5] We apply the causal forest (CF) approach of Wager and Athey (2018), and Athey, Tibshirani, and Wager (2019) to estimate the CAPE and APE of the continuous treatment, as well as the double machine learning (DML) approach of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) to estimate the ATE of a binary treatment of a discount $\geq 30\%$ vs. $< 30\%$ in the sample of always buyers.

[5] For simplification we outline the causal machine learning approaches used in our empirical analysis of the use case (ii), that is analyzing the effect of discounts on demand shift.

The CF adapts the random forest to the purpose of causal inference. It is based on first running separate random forests for predicting the outcome Y and the treatment D as a function of the covariates X using leave-one-out cross-fitting. The latter implies that the outcome or treatment of each observation is predicted based on all observations in the data but its own, in order to safeguard against overfitting bias. Second, the predictions are used for computing residuals of the outcomes and treatments, in which the influence of X has been partialled out. Finally, a further random forest is applied to average over so-called causal trees, see Athey and Imbens (2016), in order to estimate the CAPE. The causal tree approach contains two key modifications when compared to standard decision trees. First, instead of an outcome variable, it is the coefficient of regressing the residual of Y on the residual of D, i.e. the causal effect estimate of the treatment, that is to be predicted. Recursive splitting aims to find the largest effect heterogeneities across subsets defined in terms of X to estimate the CAPE accurately. Secondly, within each subset, different parts of the data are used for estimating (a) the tree's splitting structure (i.e., the definition of covariate indicator functions) and (b) the causal effect of the treatment to prevent spuriously large effect heterogeneities due to overfitting.

The CAPE estimate obtained by CF can be thought of as a weighted regression of the outcome residual on the treatment residual. The random forest-determined weight reflects the importance of a sample observation for assessing the causal effect at specific values of the covariates. After estimating the CAPE given X, the APE is obtained by appropriately averaging over the distribution of X among the always buyers. For implementing CAPE and APE estimation, we use the *grf* package by Tibshirani, Athey, Friedberg, Hadad, Hirshberg, Miner, Sverdrup, Wager, and Wright (2020) for the statistical software R. We set the number of trees to be used in a forest to 1000. We select any other tuning parameters like the number of randomly chosen covariates considered for splitting or the minimum number of observations per subset (or node) by the built-in cross-validation procedure.

We also estimate the ATE among always buyers in our sample based on DML for a binary treatment defined as $\tilde{D} = I\{D \geq 0.3\}$, with $I\{\cdot\}$ denoting the indicator function that is equal to one if its argument is satisfied and zero otherwise. Furthermore, let $\mu_d(X) = E[Y|\tilde{D} = d, X, S(0) = 1, S = 1]$ denote the conditional mean outcome and $p_d(X) = \Pr(\tilde{D} = d|X, S(0) = 1, S = 1)$ the propensity score of receiving treatment category d (with $d = 1$ for a discount $\geq$ 30% and $d = 0$ otherwise) in that population. Estimation is based on the sample analog of the doubly robust identification expression for the ATE, see Robins, Rotnitzky, and Zhao (1994) and Robins and Rotnitzky

(1995):

$$\Delta_{S(0)=1}(1,0) = E\Big[\mu_1(X) - \mu_0(X) + \frac{(Y-\mu_1(X))\cdot\tilde{D}}{p_1(X)} - \frac{(Y-\mu_0(X))\cdot(1-\tilde{D})}{p_0(X)}\Big|S(0)=1, S=1\Big]. \quad (9)$$

We estimate (9) using the *causalweight* package for the statistical software R by Bodory and Huber (2018). As machine learners for the conditional mean outcomes $\mu_D(X)$ and the propensity scores $p_D(X)$ we use the random forest with the default options of the *SuperLearner* package of van der Laan, Polley, and Hubbard (2007), which itself imports the *ranger* package by Wright and Ziegler (2017) for random forests. To impose common support in the data used for ATE estimation, we apply trimming threshold of 0.01, implying that we drop observations with estimated propensity scores smaller than 0.01 (or 1%) and larger than 0.99 (or 99%) from our sample.

7 Empirical results

In this section, we start with presenting descriptive statistics of our data used in this study. Then, this section shows the results of our four use cases of machine learning for business analytics in the railway industry. For use case (i), we present the results of the predicting buying behavior among supersaver tickets. This is followed with the use cases (ii), (iii) and (iv), discussing the effect of discounts on demand shift, customer satisfaction and upselling respectively. For the latter three use cases we always present first the (causal) effects of discounts and second the effect heterogeneity.

7.1 Descriptive Statistics

Before discussing the results of our machine learning approaches, we present some descriptive statistics for our data in Table 1, namely the mean and the standard deviation of selected variables by always buyer status and binary discount category ($\geq 30\%$ and $< 30\%$). We see that discounts and regular ticket fares of always buyers are on average lower than those of other customers. Another interesting observation is that in either discount category, we observe less leisure travelers among the always buyers than among other customers, which can be rationalized by business travelers responding less to price incentives by discounts. This is also in line with the finding that always buyers tend to purchase more second class tickets. More generally, we see non-negligible variation in demand-related covariates across the four subsamples defined in terms of buying behavior and discount rates. For instance, among always buyers, the total amount of supersaver tickets offered is on average larger in the higher discount category, while it is lower among the remaining clients.[6] This suggests that neither the treatment nor being an always buyer is quasi-random, a problem we aim to tackle based on our identification strategy outlined in section 5. Concerning the demand-shift outcome, we see that always buyers change the departure time less frequently than others. With regard to upselling, we observe that the relative amount of individuals upgrading their 2nd class to a 1st class ticket is rather stable across both discount categories, i.e. $\geq 30\%$ and $< 30\%$.

7.2 Predicting buying decisions

We subsequently present our predictive analysis based on the random forest and investigate which covariates importantly predict three outcomes, namely whether customers booked a trip otherwise not realized by train (additional trip), bought a first-class rather than a second-class ticket (upselling), or rescheduled their trip e.g. away from rush hours (demand shift). For this purpose, we create three distinct datasets in which the values of the respective binary outcome are balanced, i.e. 1 (for instance, upselling) for 50% and 0 (no upselling) for 50% of the observations. We balance our data because we aim to train a model that predicts both outcome values equally well. Taking the demand shift outcome as an example, our data with non-missing covariate or outcome information contain 3481 observations with $Y = 1$ and 9576 observations with $Y = 0$. We retain all observations with $Y = 1$ and randomly draw 3481 observations with $Y = 0$ to obtain such a balanced data set. In the next step, we randomly split these 6962 observations into a training set consisting of 75% of the data and a test set (25%). In the training set, we train the random forest using the treatment D and all covariates X, W as predictors. In the test set, we predict the outcomes based on the trained forest, classifying e.g. observations with a demand shift probability ≥ 0.5 as 1. We then compare the predictions to the actually observed outcomes to assess model performance based on the correct classification rate (also known as accuracy), i.e. the share of observations in the test data for which the predicted outcome corresponds to the actual one.

For each of the outcomes, Table 2 presents the 30 most predictive covariates in the training set ordered in decreasing order according to a variable

[6] In Appendix A we present two tables showing the total and relative amount of supersaver tickets offered and sold by the SBB in more detail.

Table 1: Mean and standard deviation by discount and type

discount	< 30%		≥ 30%	
always buyers	No	Yes	No	Yes
discount	0.21	0.19	0.57	0.53
	(0.07)	(0.08)	(0.12)	(0.13)
regular ticket fare	44.36	36.14	47.19	32.91
	(29.38)	(25.47)	(30.14)	(23.78)
age	47.22	47.68	45.59	48.77
	(15.36)	(16.14)	(15.80)	(16.49)
gender	0.51	0.55	0.53	0.59
	(0.50)	(0.50)	(0.50)	(0.49)
diff. purchase travel	3.42	3.23	7.72	7.19
	(6.96)	(6.76)	(11.23)	(10.30)
distance	136.49	127.86	126.15	116.76
	(77.38)	(71.49)	(69.98)	(66.04)
capacity utilization	35.51	39.19	26.46	33.15
	(14.16)	(14.31)	(13.24)	(13.75)
seat capacity	328.28	429.57	303.83	445.14
	(196.19)	(196.10)	(185.42)	(188.54)
offer total	33.95	44.10	70.97	98.34
	(42.57)	(50.68)	(69.57)	(84.45)
sold total	28.04	37.29	13.70	25.75
	(41.92)	(50.31)	(36.37)	(53.67)
half fare travel ticket	0.74	0.79	0.62	0.74
	(0.44)	(0.40)	(0.49)	(0.44)
leisure	0.77	0.69	0.82	0.76
	(0.42)	(0.46)	(0.39)	(0.43)
class	1.38	1.65	1.33	1.73
	(0.48)	(0.48)	(0.47)	(0.44)
Swiss	0.89	0.92	0.88	0.88
	(0.31)	(0.28)	(0.33)	(0.32)
demand shift	0.31	0.19	0.31	0.23
	(0.46)	(0.40)	(0.46)	(0.42)
upselling	0.49	0.00	0.49	0.00
	(0.50)	(0.00)	(0.50)	(0.00)
obs.	1151	2221	5529	3745

Notes: Regular ticket fare is in Swiss francs. 'diff. purchase travel' denotes the difference between purchase and travel day. 'Offer total' and 'sold total' denote the total amount of supersaver tickets offered and the total amount of supersaver tickets sold respectively.

importance measure. The latter is defined as the total decrease in the Gini index (as a measure of node impurity in terms of outcome values) in a tree when including the respective covariate for splitting, averaged over all trees in the forest. The results suggest that trip- and demand-related characteristics like seat capacity, utilization, departure time, and distance are important predictors. Concerning personal characteristics, also customer's age appears to be relevant. Furthermore, also the treatment intensity D has considerable predictive power. Interestingly, specific connections (defined by indicators for points of departure and destination) turn out to be less important characteristics conditional on the other covariates already mentioned.

At the bottom of Table 2 we also report the correct classification rates for the three outcomes. While the accuracy in predicting a demand shift amounts to 58%, which is somewhat better than random guessing but not particularly impressive, the performance is more satisfactory for predicting decisions about additional trips with an accuracy of 65% and quite decent for upselling (82%). We note that when predicting upselling, we drop the variables 'class', which indicates whether someone travels in the first or second class, and 'seat capacity', which refers to the capacity in the chosen class, from the predictors. The reason is that upselling is defined as switching from second to first class, and therefore, the chosen class and the related seat capacities are actually part of the outcome to be predicted. Tables 18 and 17 in the Appendix E present the predictive outcome analysis separately for subsamples with discounts $\geq 30\%$ and $< 30\%$, respectively. In terms of which classes of variables are most predictive (trip- and demand-related characteristics, age, discount rate) and also in terms of accuracy, the findings are rather similar to those in Table 2. In general, machine learning appears useful for forecasting customer behavior in the context of demand for train trips, albeit not equally well for all aspects of interest. Such forecasts may for instance serve as a base for customer segmentation, e.g. into customer groups more and less inclined to book an additional trip or switch classes or departure times, in order to specifically target them by particular interventions like marketing campaigns.

Table 2: Predictive outcome analysis

demand shift		upselling		additional trip	
variable	importance	variable	importance	variable	importance
departure time	142.694	capacity utilization	295.924	seat capacity	147.037
seat capacity	121.42	offer level B	188.861	D	128.086
age	119.846	offer level C	149.911	age	123.948
capacity utilization	119.606	D	132.095	departure time	123.516
D	112.474	age	100.258	capacity utilization	113.160
distance	112.143	departure time	98.909	distance	101.730
offer level B	84.142	offer level A	93.303	offer level B	84.989
diff. purchase travel	81.167	distance	87.319	diff. purchase travel	80.236
offer level C	76.238	offer level D	85.408	offer level A	78.507
offer level A	75.971	diff. purchase travel	62.841	offer level C	77.097
number of sub-journeys	73.096	number of sub-journeys	55.978	number of sub-journeys	69.443
offer level D	61.763	rel. sold level A	44.505	ticket purchase complexity	64.498
ticket purchase complexity	57.071	ticket purchase complexity	41.819	offer level D	56.888
rel. sold level A	51.377	offer level E	37.159	class	51.456
rel. amount imputed values	42.222	rel. sold level B	34.462	rel. sold level A	46.969
rel. sold level B	38.144	rel. amount imputed values	30.747	rel. amount imputed values	38.869
adult companions	34.176	rel. sold level C	28.635	rel. sold level B	36.484
rel. sold level C	28.201	adult companions	25.115	half fare	35.785
offer level E	25.714	2019	18.88	adult companions	34.446
gender	23.707	gender	18.47	halfe fare travel ticket	28.465
amount purchased tickets	19.575	rush hour	17.448	gender	25.419
German	18.659	Saturday	16.173	rel. sold level C	24.679
travel alone	18.605	German	15.457	offer level E	22.556
2019	18.082	leisure	15.304	leisure	20.438
French	17.906	amount purchased tickets	15.112	no subscriptions	19.793
saturday	17.487	travel alone	14.792	amount purchased tickets	19.283
Friday	17.272	half fare	14.306	German	19.119
peak hour	17.064	French	14.161	travel alone	18.139
class	16.973	Thursday	13.413	2019	17.192
leisure	16.892	scheme 20	13.411	French	17.026
correct prediction rate	0.581		0.817		0.653
balanced sample size	6962		6738		7000

Notes: 'Offer level A', 'offer level B', 'offer level C', 'offer level D' and 'offer level E' denote the amount of supersaver tickets with discount A, B, C, D and E respectively. 'Diff. purchase travel' denotes the difference between purchase and travel day. 'Rel. sold level A', 'rel. sold level B', 'rel. sold level C' and 'rel. sold level D' denote the relative amount of supersaver tickets offered with discount A, B, C and D respectively. The relative amounts are in relation to the seats offered. 'No subscriptions' indicates not possessing any subscription. For predicting upselling, the covariates 'class' and 'seat capacity' are dropped.

7.3 Testing the identification strategy

Before presenting the results for the causal analyses, we consider two different methods to partially test the assumptions underlying our identification strategy. First, we test Assumption 3 (weak monotonicity) by running the CF and DML procedures as well as a conventional OLS regression in which we use buying an additional trip $(1 - S(0))$, i.e. not being an always buyer, as outcome variable and X as control variables in our sample of supersaver customers. The CF permits estimating the conditional change in the share of surveyed customers induced to buy an additional trip by modifying the discount rate D given X, i.e. $\frac{\partial E[(1-S(0))|D,X,S=1]}{\partial D}$, as well as the average thereof across X conditional on sample selection, $E\left[\frac{\partial E[(1-S(0))|D,X,S=1]}{\partial D}\middle|S=1\right]$. DML, on the other hand, yields an estimate of the average difference in the share of additional trips across the high and low treatment categories conditional on sample selection, $E[E[(1-S(0))|D<0.3, X, S=1] - E[(1-S(0))|D \geq 0.3, X, S=1]|S=1]$. Finally, the OLS regression of $(1 - S(0))$ on D and all X in our sample tests monotonicity when assuming a linear model.

Table 3 reports the results that do not provide any evidence against the monotonicity assumption. When considering the continuous treatment D, the CF-based estimate of $E\left[\frac{\partial E[(1-S(0))|D,X,S=1]}{\partial D}\middle|S=1\right]$ is highly statistically significant and suggests that augmenting the discount by one percentage point increases the share of customers otherwise not buying the ticket by 0.56 percentage points on average. Furthermore, any estimates of the conditional change $\frac{\partial E[(1-S(0))|D,X,S=1]}{\partial D}$ are positive, as displayed in the histogram of Figure 2, and 82.2% of them are statistically significant at the 10% level, 69.1% at the 5% level. Furthermore, the OLS coefficient of 0.544 is highly significant. Likewise, the statistically significant DML estimate points to an increase in the share of additional trips by 18.4 percentage points when switching the binary treatment indicator from $D < 0.3$ to $D \geq 0.3$. Taken at face value, our main estimate of 0.56 implies that if the Swiss Federal Railways offered a discount of 20% (instead of no discount), 11.2% additional trips were booked. However, it is important to mention that these numbers must be interpreted with great caution as they are based on a selective sample of buyers of supersaver tickets and cannot be easily generalized to all (i.e. regular fare and supersaver) single tickets purchased by the total of railway customers. For this reason, our limited sample does not appear suitable for conducting a general cost-benefit analysis for the supersaver tickets.

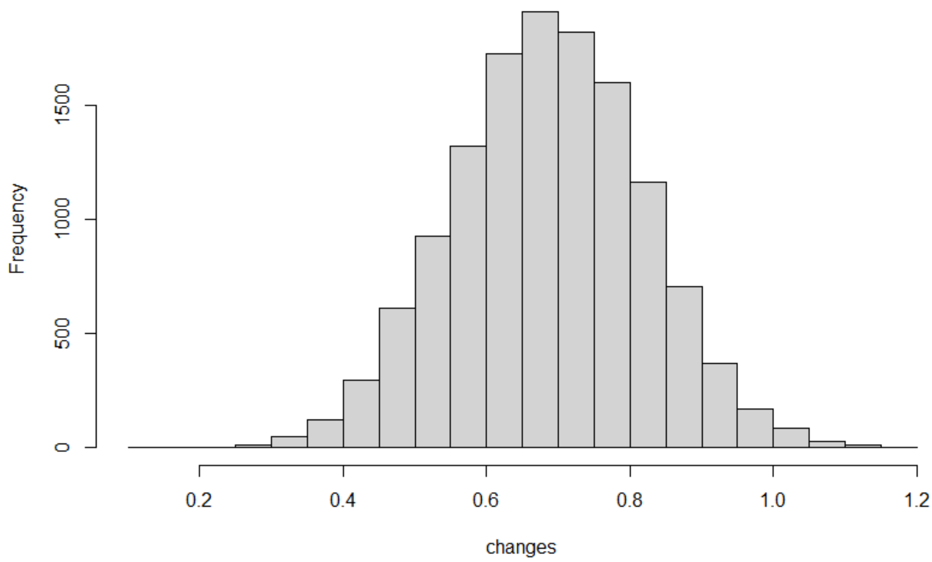
We also test the statistical independence of D and W conditional on X in our sample of always buyers, as implied by our identifying assumptions, see the discussion at the end of section 5. To this end, we randomly split the

Table 3: Monotonicity tests

	CF: av. change	OLS: coef.	DML: $D \geq 0.3$ vs $D < 0.3$
change in $(1 - S(0))$	0.564	0.544	0.184
standard error	0.060	0.031	0.009
p-value	0.000	0.000	0.000
trimmed observations			1760
number of observations	12924		

Notes: 'CF', 'OLS', and 'DML' stands for estimates based on causal forests, linear regression, and double machine learning, respectively. 'trimmed observations' is the number of trimmed observations in DML when setting the propensity score-based trimming threshold to 0.01*. Control variables consist of* X*.*

Figure 2: Monotonicity given X



evaluation data into a training set (25% of observations) and a test set (75% of all observations). In the training data set, we run a linear lasso regression (Tibshirani, 1996) of D on X in order to identify important predictors by means of 10-fold cross-validation. In the next step, we select all covariates in X with non-zero lasso coefficients and run an OLS regression of D on the selected covariates in the test data. Finally, we add W to that regression in the test data and run a Wald test to compare the predictive power of the models with and without W. We repeat the procedure of splitting the data, performing the lasso regression in the training set, and running the OLS regressions and the Wald test in the test set 100 times. This yields an average p-value of 0.226, with 15 out of 100 p-values being smaller than 5%. These results do not provide compelling statistical evidence that W is associated with D conditional on X, even though the training sample is relatively small and thus favors selecting too few predictors in X (due to the cross-validation that trades off bias due to including fewer predictors and variance due to including more predictors).

We note that performing lasso-based variable selection and OLS-based testing in different (training and test) data avoids correlations of these steps that could entail an overestimation of the goodness of fit. Nonetheless, our findings remain qualitatively unchanged when performing both steps in all of the evaluation data. Repeating the cross-validation step for the lasso-based covariate selection 100 times and testing in the total sample yields an even higher average p-value of 0.360. Finally, we run a standard OLS regression of D on all elements of X (rather than selecting the important ones by lasso) in the total sample and compare its predictive power to a model additionally including W. Also in this case, the Wald test entails a rather high p-value of 0.343. In summary, we conclude that our tests do not point to the violation of our identifying assumptions.

7.4 Causal effect of discounts on demand shift

Table 4 presents the main results of our causal analysis, namely the estimates of the discount rate's effect on the demand shift outcome, which is equal to one if the discount induced rescheduling the departure time and zero otherwise. We note that all covariates, i.e. both the trip- or demand-related factors X and the personal characteristics W, are used as control variables, even though we have previously claimed that X is sufficient for identification. There are, however, good reasons for including W as well in the estimations. First, conditioning on the personal characteristics available in the data may reduce estimation bias if X is - contrarily to our assumptions and to what our tests suggest - not fully sufficient to account for confounding. Second, it can also reduce the variance of the estimator, e.g. if some factors like age are strong predictors of the outcome. Third, having W in the CF allows for a more fine-grained analysis of effect heterogeneity based on computing more 'individualized' partial effects that (also) vary across personal characteristics.

Considering the estimates of the CF, we obtain an average partial effect (APE) of 0.161, suggesting that increasing the current discount rate among always buyers by one percentage point increases the share of rescheduled

Table 4: Effects on demand shift

	CF: APE	DML: ATE $D \geq 0.3$ vs $D < 0.3$
effect	0.161	0.038
standard error	0.072	0.010
p-value	0.025	0.000
trimmed observations		151
number of observations	5903	

Notes: 'CF' and 'DML' stands for estimates based on causal forests, linear regression, and double machine learning, respectively. 'trimmed observations' is the number of trimmed observations in DML when setting the propensity score-based trimming threshold to 0.01. *Control variables consist of both X and W.*

trips by 0.16 percentage points. This effect is statistically significant at the 5% level. As a word of caution, however, we point out that the standard error is non-negligible such that the magnitude of the impact is not very precisely estimated. When applying DML, we obtain an average treatment effect (ATE) of 0.038 that is significant at the 1% level, suggesting that discounts of 30% and more on average increase the number of demand shifts by 3.8 percentage points compared to lower discounts, which is qualitatively in line with the CF. Furthermore, we find a decent overlap or common support in most of our sample in terms of the estimated propensity scores across lower and higher discount categories considered in DML, see the propensity score histograms in Appendix B. This is important as ATE evaluation hinges on the availability of observations with comparable propensity scores across treatment groups. Only 151 out of our 5903 observations are dropped due to too extreme propensity scores below 0.01 or above 0.99 (pointing to a violation of common support).[7] In summary, our results clearly point to a positive average effect of the discount rate on trip rescheduling among always buyers, which is, however, not overwhelmingly large.

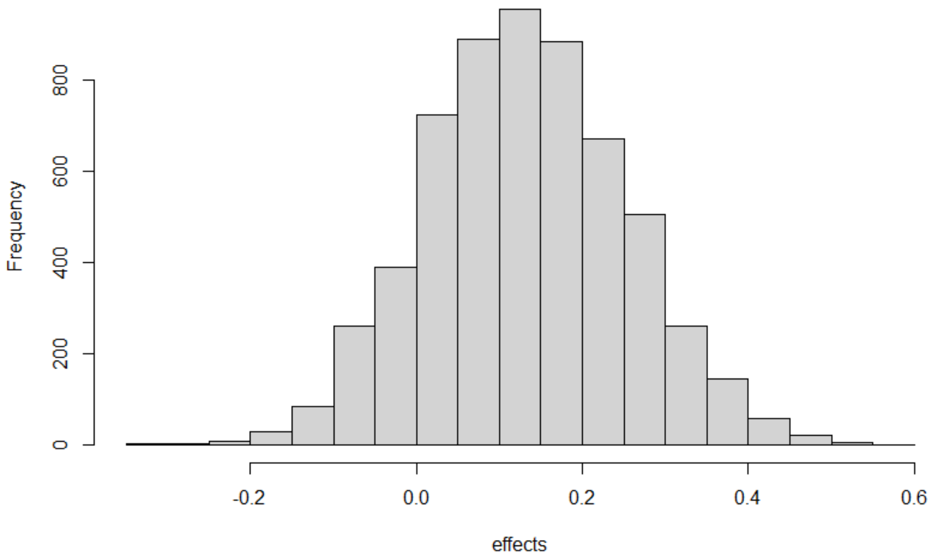
7.5 Demand shift: Effect heterogeneity

In this section, we assess the heterogeneity of the effects of D on Y across interviewees and observed characteristics. Figure 3 shows the distribution the CF-based conditional average effects (CAPE) of marginally increasing the discount rate given the covariates values of the always buyers in our sample (which are also the base for the estimation of the APE). While the CAPEs

[7] Our findings of a positive ATE remain robust when setting the propensity score-based trimming threshold to 0.02 (ATE: 0.042) or 0.05 (ATE: 0.045).

are predominantly positive, they are quite imprecisely estimated. Only 2.9% and 0.8% of the positive ones are statistically significant at the 10% and 5% levels, respectively. Further, only 0.1% of the negative ones are statistically significant at the 10% level. Yet, the distribution points to a positive marginal effect for most always buyers and also suggests that the magnitude of the effects varies non-negligibly across individuals.

Figure 3: CAPEs on demand shift



Next, we assess the effect heterogeneity across observed characteristics based on the CF results. First, we run a conventional random forest with the estimated CAPEs as the outcome and the covariates as predictors to assess the covariates' relative importance for predicting the CAPE, using the decrease in the Gini index as importance measure as also considered in section 7.2. Table 5 reports the 20 most predictive covariates ordered in decreasing order according to the importance measure. Demand-related characteristics (like seat capacity, utilization, departure time, and distance) turn out to be the most important predictors for the size of the effects, also customer's age has some predictive power. Similarly as for outcome prediction in section 7.2, specific connections (characterized by points of departure or destination) are less important predictors of the CAPEs given the other information available in the data.

While Table 5 provides information on the best predictors of effect hetero-

Table 5: Most important covariates for predicting CAPEs on demand shift

covariate	importance
seat capacity	11.844
offer level C	11.164
capacity utilization	5.144
departure time	5.122
distance	4.287
offer level D	4.015
class	3.434
saturday	2.933
age	2.429
number of sections	2.373
diff. purchase travel	2.110
offer level A	1.634
offer level B	1.610
half fare	1.524
scheme 17	1.496
half fare travel ticket	1.373
rel. sold level B	0.901
ticket purchase complexity	0.847
leisure	0.773
rel. sold level A	0.770

Notes: 'Offer level A', 'offer level B', 'offer level C' and 'offer level D' denote the amount of supersaver tickets with discount A, B, C and D respectively. 'Rel. offer level A', 'rel. offer level B' and 'rel. offer level C' denote the relative amount of supersaver tickets offered with discount A, B and C. The relative amounts are in relation to the seats offered.

geneity, it does not give insights on whether effects differ importantly and statistically significantly across specific observed characteristics of interest. For instance, one question relevant for designing discount schemes is whether (marginally) increasing the discounts is more effective among always buyers so far exposed to rather small or rather large discounts. Therefore, we investigate whether the CAPEs are different across our binary treatment categories defined by $\tilde{D}$ (30% or more and less than 30%). To this end, we apply the approach of Semenova and Chernozhukov (2020) based on (i) plugging the CF-based predictions into a modified version of the doubly robust functions provided within the expectation operator of (9) that is suitable for a continuous D and (ii) linearly regressing the doubly robust functions on the treatment indicator $\tilde{D}$. The results are reported in the upper panel of Table 6. While the point estimate of -0.104 suggests that the demand shifting effect of increasing the discount is on average smaller when discounts are already quite substantial (above 30%), the difference is far from being statistically significant at any conventional level.

Using again the method of Semenova and Chernozhukov (2020), we also investigate the heterogeneity among a limited and pre-selected set of covariates that appears interesting for characterizing customers and their travel purpose,

Table 6: Effect heterogeneity analysis for demand shift

	effect	st. err.	p-value
Discounts categories ($D \geq 0.3$ vs $D < 0.3$)			
APE for $D < 0.3$ (constant)	0.209	0.089	0.019
Difference APE $D \geq 0.3$ vs $D < 0.3$ (slope coef.)	-0.104	0.122	0.395
Customer and travel characteristics			
constant	-0.154	0.295	0.602
age	-0.002	0.004	0.556
gender	-0.022	0.129	0.866
distance	-0.000	0.001	0.697
leisure trip	0.297	0.165	0.072
commute	0.241	0.241	0.316
half fare travel ticket	0.228	0.142	0.109
peak hours	0.222	0.133	0.094

Notes: Business trip is the reference category for the indicators 'leisure trip' and 'commute'.

namely age, gender, and travel distance, as well as indicators for leisure trip and commute (with business trip being the reference category), traveling during peak hours, and possession of a half fare travel tickets. As displayed in the lower panel of Table 6, we find no important effect heterogeneities across the age or gender of always buyers or as a function of travel distance conditional on the other information included in the regression, as the coefficients on these variables are close to zero. In contrast, the effect of demand shift is (given the other characteristics) substantially larger among always buyers with a half fare travel tickets and among commuters, however, neither coefficient is statistically significant at the 10% level (even though the half fare coefficient is close).

For leisure trips, the coefficient is even larger (0.297), suggesting that all other included variables equal, a one percentage point increase in the discount rate increases the share of rescheduled trips by 0.29 percentage points more among leisure travelers than among always buyers traveling for business. The coefficient is statistically significant at the 10% level, even though we point out that the p-value does not account for multiple hypothesis testing of several covariates. This finding can be rationalized by leisure travelers being likely more flexible in terms of timing than business travelers. Also the coefficient on peak hours is substantially positive (0.222) and statistically significant at the 10% level (again, without controlling for multiple hypothesis testing). This could be due to peak hours being the most attractive travel time, implying that custumers are more willing to reschedule their trips when being offered a discount within peak hours. We conclude that even though several coefficients appear non-negligible, statistical significance in our heterogeneity analysis is overall limited, which could be due to the (for the purpose of investigating effect heterogeneity) limited sample of several thousand observations.

7.6 Causal effect of discounts on customer satisfaction

To assess the effects of supersaver tickets on customer satisfaction, we use the same causal approach as in the previous sections when evaluating the effects on demand shift. Customer satisfaction is measured by the question *What is your overall impression of SBB?*, which takes the values 1 (very bad) to 10 (very good), see Appendix C for more details. Table 7 presents the estimates of the discount rate's effect on the customer satisfaction based on CF and DML. All covariates, i.e. both the trip- or demand-related factors X and the personal characteristics W, are used as control variables.

Table 7: Effects on customer satisfaction

	CF: APE	DML: ATE $D \geq 0.3$ vs $D < 0.3$
effect	0.248	0.159
standard error	0.249	0.037
p-value	0.320	0.000
trimmed observations		163
number of observations	5963	

Notes: 'CF' and 'DML' stands for estimates based on causal forests, linear regression, and double machine learning, respectively. 'trimmed observations' is the number of trimmed observations in DML when setting the propensity score-based trimming threshold to 0.01*. Control variables consist of both X and W.*

Considering the estimates of the CF, we obtain an average partial effect (APE) of 0.248, suggesting that increasing the current discount rate among always buyers by one percentage point increases the customer satisfaction by 0.0025 points (on a 10-points scale).[8] However, this small positive effect is far from being statistically significant. When applying DML, we obtain an average treatment effect (ATE) of 0.159 that is significant at the 1% level, suggesting that discounts of 30% and more on average moderately increase the customer satisfaction by 0.16 points compared to lower discounts. Note that 163 out of our 5963 observations are dropped due to too extreme propensity scores below 0.01 or above 0.99 (pointing to a violation of common

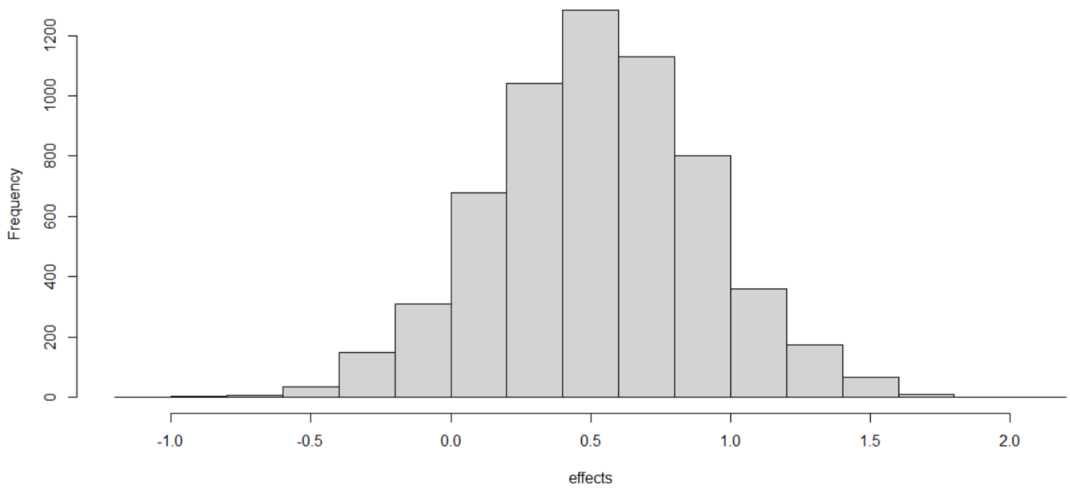
[8] The interpretation of the APE differs from those in the other analyses in that it is expressed in points on a 10-points scale rather than percentage points.

support). In summary, our results point to a positive, but rather minor effect of discounts on customer satisfaction.

7.7 Satisfaction: Effect heterogeneity

Despite the non-significant APE in the previous section, we now turn to assessing the heterogeneity of the effects of D on Y across always buyers and observed characteristics. Figure 4 shows the distribution the CF-based conditional average effects (CAPE) of marginally increasing the discount rate given the covariates values of the always buyers in our sample (which are also the basis for the estimation of the APE). While the CAPEs are predominantly positive, they are quite imprecisely estimated. Only 3.1% and 1.0% of the positive ones are statistically significant at the 10% and 5% levels, respectively. Yet, the distribution points to a positive marginal effect for most always buyers.

Figure 4: CAPEs on satisfaction



Next, we assess the effect heterogeneity across observed characteristics based on the CF results. First, we run a conventional random forest with the estimated CAPEs as the outcome and the covariates as predictors to assess the covariates' relative importance for predicting the CAPE. Table 8 reports the 20 most predictive covariates ordered decreasingly according to the importance measure. Similarly as for the demand shift outcome, demand-related characteristics (like seat capacity, utilization, departure time, and distance) turn out to be the most important predictors for the size of the effects, also customer's age has some predictive power. These variables are therefore relatively important for identifying different customer groups in terms of a discount's impact on customer satisfaction. Again, specific connections (characterized by points of departure or destination) are less important predictors of the CAPEs given the other information available in the data, with the exception of the destination 'Geneva Airport'.

Table 8: Most important covariates for predicting CAPEs on satisfaction

covariate	importance
capacity utilization	77.322
age	64.664
distance	52.699
departure time	48.849
distance	45.446
seat capacity	40.177
rel. sold level A	32.103
destination Geneva Airport	31.707
offer level B	28.395
offer level C	27.238
rel. amount imputed values	25.967
ticket purchase complexity	25.228
rel. sold level B	23.703
half fare travel ticket	23.535
offer level D	22.967
diff. purchase travel	21.231
offer level A	19.531
half fare	17.711
peak hour	14.856
imputed values	10.018

Notes: 'offer level A', 'offer level B', 'offer level C' and 'offer level D' denote the amount of supersaver tickets with discount A, B, C and D respectively. 'Rel. offer level A', and 'rel. offer level B' denote the relative amout of supersaver tickets offered with discount A, B and C. The relative amounts are in relation to the seats offered.

While Table 8 provides information on the best predictors of effect heterogeneity, it does not give insights on whether effects differ importantly and statistically significantly across specific observed characteristics of interest. For instance, one relevant question for designing discount schemes is whether (marginally) increasing the discounts is more effective among always buyers so far exposed to rather small or rather large discounts. Therefore, we investigate whether the CAPEs are different across our binary treatment categories defined by $\tilde{D}$ (30% or more and less than 30%). To this end, we again apply the approach of Semenova and Chernozhukov (2020). The results are reported in the upper panel of Table 9. While the point estimate of -0.104 suggests that the effect on customer satisfaction of increasing the discount is on average smaller when discounts are already quite substantial (above 30%), the difference is far from being statistically significant at any conventional level.

Table 9: Effect heterogeneity analysis for satisfaction

	effect	st. err.	p-value
Discounts categories ($D \geq 0.3$ vs $D < 0.3$)			
APE for $D < 0.3$ (constant)	0.209	0.089	0.019
Difference APE $D \geq 0.3$ vs $D < 0.3$ (slope coef.)	-0.104	0.122	0.395
Customer and travel characteristics			
constant	0.937	1.066	0.380
age	-0.001	0.015	0.963
gender	-0.628	0.452	0.165
distance	0.001	0.004	0.885
leisure trip	0.485	0.601	0.420
commute	0.746	0.859	0.385
half fare travel ticket	-0.919	0.525	0.080
peak hours	0.479	0.487	0.325

Notes: Business trip is the reference category for the indicators 'leisure trip' and 'commute'.

Using again the method of Semenova and Chernozhukov (2020), we also investigate the heterogeneity among a limited and pre-selected set of covariates that appears interesting for characterizing customers and their travel purpose, namely age, gender, and travel distance, as well as indicators for leisure trip and commute (with business trip being the reference category), traveling during peak hours, and possession of a half fare travel tickets. As displayed in the lower panel of Table 9, we find no important effect heterogeneities across the age or the travel distance among always buyers conditional on the other information included in the regression, as the coefficients on these variables are close to zero. Different, the effect is positive for commuter, leisure travelers and peak hour. However, neither coefficient is statistically significant at the 10% level. In contrast, the effect on customer satisfaction is (given the other characteristics) lower among always buyers with a half fare travel tickets and for men. However, only the effect of half fare travel tickets is statistically significant at the 10% level.

7.8 Effect of discounts on upselling

In this and the following section, we discuss the effects of discounts on upselling behavior. The latter are individuals upgrading their second-class to a first-class ticket. In the previous use cases, we defined always buyers as individuals that would even have traveled at the regular fare and in the same class. We now drop the latter constraint from our sample definition and permit individuals switching from the second to the first class to be part of the subgroup of always buyers too (as long as they would have bought the trip without discount). This, however, reduces the credibility of the assumptions required for causal analysis, as our evaluation is now based on individuals that are likely more dissimilar in terms of buying behavior than those not switching the travel class. Indeed, we find that our previously discussed test of the selection on observables assumption, namely that W must not be associated with D conditional on X, fails for this sample definition, see Appendix D (which also provides the results for the monotonicity test that among all survey respondents the share of additional trips must increase in the discount rate). Such a selection bias could be circumvented in future studies by surveying a random sample of individuals (not only buyers of supersaver tickets) and randomizing the discounts by means of an experiment.

Table 10: Effects on upselling

	CF: APE	DML: ATE $D \geq 0.3$ vs $D < 0.3$
effect	0.589	0.163
standard error	0.031	0.008
p-value	0.000	0.000
trimmed observations		774
number of observations	9422	

Notes: 'CF' and 'DML' stands for estimates based on causal forests, linear regression, and double machine learning, respectively. 'trimmed observations' is the number of trimmed observations in DML when setting the propensity score-based trimming threshold to 0.01*. Control variables consist of both X and W.*

Table 10 presents the effect estimates based on CF and DML when using both trip- or demand-related factors X and the personal characteristics W as control variables. However, note that the covariates 'class' and 'seat capacity' are not included. The reason is that these variables are directly related to the outcome of interest, because every individual upgrading her ticket rides first class, while seat capacity refers to a specific class (1st or 2nd). Considering the estimates of the CF, we obtain an average partial effect (APE) of 0.589, suggesting that increasing the current discount rate among always buyers (new definition) by one percentage point increases the share of tickets being upgraded by 0.59 percentage points. This effect is substantially higher than that for demand shift (0.161) and statistically significant at the 1% level. When applying DML, we obtain an average treatment effect (ATE) of 0.163, which is also significant at the 1% level. It suggests that discounts of 30% and more on average increase the number of demand shifts by 16.3 percentage points compared to lower discounts, which is qualitatively in line with the

CF. Again, it is substantially higher than the effect on demand shift (0.038). Common support appears acceptable for most of our sample in terms of the estimated propensity scores across lower and higher discount categories considered in DML: 774 out of our 5903 observations are dropped due to too extreme propensity scores below 0.01 or above 0.99 (pointing to a violation of common support). In summary, our results clearly point to a positive average effect of the discount rate on trip rescheduling among always buyers if taken at face value. However, we bear in mind that these effects might be substantially biased as the assumptions underlying our causal analysis are likely to be violated.

7.9 Upselling: Effect heterogeneity

In this section, we assess the heterogeneity of the effects of D on Y across interviewees and observed characteristics. Figure 5 shows the distribution the CF-based conditional average effects (CAPE) of marginally increasing the discount rate given the covariates values of the newly defined always buyers in our sample. We see that the CAPEs are predominantly positive and that the magnitude of the effects varies non-negligibly across individuals.

Next, we assess the effect heterogeneity across observed characteristics based on the CF results. First, we run a conventional random forest with the estimated CAPEs as the outcome and the covariates as predictors to assess the covariates' relative importance for predicting the CAPE, using again the decrease in the Gini index as importance measure. Table 11 reports the 20 most predictive covariates in decreasing order according to the importance measure. Demand-related characteristics (like utilization, departure time, and distance) turn out to be the most important predictors for the size of the effects, also customer's age has some predictive power. As in our previous heterogeneity analyses, particular connections are less important predictors of the CAPEs given the other information available in the data.

Figure 5: CAPEs for upselling

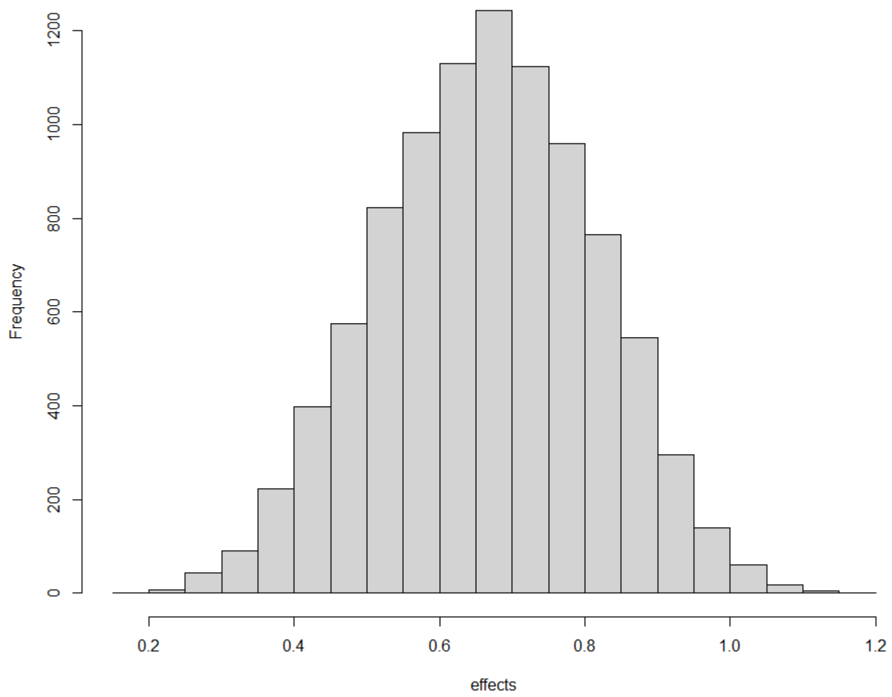


Table 11: Most important covariates for predicting CAPEs on upselling

covariate	importance
capacity utilization	83.686
age	15.596
offer level C	13.407
departure time	10.518
distance	7.734
offer level B	7.454
Saturday	4.185
half fare	4.156
offer level A	3.842
scheme 20	3.654
offer level D	2.988
diff. purchase travel	2.907
half fare travel ticket	2.535
number of sub-journeys	2.245
rel. sold level A	2.093
rel. sold level B	1.902
peak hours	1.744
ticket purchase complexity	1.555
scheme 6	1.458
rel. amount imputed values	1.231

Notes: 'offer level A', 'offer level B', 'offer level C' and 'offer level D' denote the amount of supersaver tickets with discount A, B, C and D respectively. 'Rel. offer level A' and 'rel. offer level B' denote the relative amount of supersaver tickets offered with discount A and B. The relative amounts are in relation to the seats offered. 'Diff. purchase travel' denotes the difference between purchase and travel day.

While Table 11 provides information on the best predictors of effect heterogeneity, it does not give insights on whether effects differ importantly and statistically significantly across specific observed characteristics of interest. Similarly to our previous analyses on effect heterogeneity, we investigate whether the CAPEs are different across our binary treatment categories defined by $\tilde{D}$ (30% or more and less than 30%). To this end, we again apply the approach of Semenova and Chernozhukov (2020) based on (i) plugging the CF-based predictions into a modified version of the doubly robust functions provided within the expectation operator that is suitable for a continuous D and (ii) linearly regressing the doubly robust functions on the treatment indicator $\tilde{D}$. The results are reported in the upper panel of Table 12. The point estimate of -0.064 suggests that the upselling effect of increasing the discount is on average smaller when discounts are already quite substantial (above 30%). However, the difference is not statistically significant at any conventional level.

Using again the method of Semenova and Chernozhukov (2020), we also investigate the heterogeneity among a limited and pre-selected set of covariates that appears interesting for characterizing customers and their travel purpose,

Table 12: Effect heterogeneity analysis for upselling

	effect	st. err	p-value
Discounts categories ($D \geq 0.3$ vs $D < 0.3$)			
APE for $D < 0.3$ (constant)	0.742	0.071	0.000
Difference APE $D \geq 0.3$ vs $D < 0.3$ (slope coef.)	-0.064	0.094	0.493
Customer and travel characteristics			
constant	1.320	0.230	0.000
age	-0.009	0.003	0.002
gender	-0.062	0.099	0.528
distance	0.001	0.001	0.350
leisure trip	-0.034	0.154	0.827
commute	-0.170	0.196	0.386
half fare travel ticket	-0.220	0.107	0.040
peak hours	-0.064	0.107	0.548

Notes: Business trip is the reference category for the indicators 'leisure trip' and 'commute'.

namely age, gender, and travel distance, as well as indicators for leisure trip and commute (with business trip being the reference category), traveling during peak hours, and possession of a half fare travel tickets. As displayed in the lower panel of Table 12, we find a negative and significant effect of age, conditional on the other information included in the regression. Further, we see that the possession of a half fare travel ticket has a negative effect on upselling. This coefficient is statistically significant at the 5% level (again, without controlling for multiple hypothesis testing). We find no other effect heterogeneities being significant at the 10% level.

8 Conclusion

In this study, we applied causal and predictive machine learning to assess the effects of discounts on train tickets issued by the Swiss Federal Railways (SBB), the so-called 'supersaver tickets', in four use cases. Our study is based on unique data that combines a survey of supersaver customers with rail trip- and demand-related information provided by the SBB.

In use case (i), we analyzed which customer- or trip-related characteristics (including the discount rate) are predictive for three outcomes characterizing buying behavior, namely: booking a trip otherwise not realized by train (additional trip), buying a first- rather than second-class ticket (upselling), or rescheduling a trip (e.g. a demand shift away from rush hours) when being offered a supersaver ticket. The random forest-based results suggested that customer's age, demand-related information for a specific connection (like seat capacity, departure time, and utilization), and the discount level permit forecasting buying behavior to a certain extent, with correct classification rates amounting to 58% (demand shift), 65% (additional trip), and 82% (upselling), respectively.

As predictive machine learning cannot provide the causal effects of the predictors involved, we applied causal machine learning to assess the causal impact of the discount rate on buying behavior. In particular, in use case (ii), we investigated the effect of discounts on demand shift among always buyers who would have traveled even without a discount in the same travel class. This appears interesting in the light of capacity constraints at rush hours. To this end, we invoked the identifying assumptions that (a) the discount rate is quasi-random conditional on our covariates and (b) the buying decision increases weakly monotonically in the discount rate and exploited survey information about customer behavior in the absence of discounts. We also considered two approaches for partially testing our assumptions, which did not point to a violation of the latter. Our main results based on the causal forest suggested that increasing the discount rate by one percentage point entails an average increase of 0.16 percentage points in the share of rescheduled trips among always buyers. This finding was corroborated by double machine learning with just two discount categories, suggesting that discount rates of 30% and more on average increase the share of rescheduled trips by 3.6 percentage points compared to lower discounts. Furthermore, when investigating effect heterogeneity across a pre-selected set of characteristics, we found the causal forest-based effects to be higher (with marginal statistical significance when not controlling for multiple hypothesis testing) for leisure travelers and during peak hours when also controlling for customer's age, gender, possession of a half fare travel card, and travel distance. Finally, our effect heterogeneity analysis also revealed that demand-related information is most predictive for the effect of the discount rate.

In use case (iii), we assessed the causal effect of discounts on customer satisfaction, again among always buyers and based on the same identifying assumptions as in the second use case. We found that increasing the discount rate has a relatively small positive effect on customer satisfaction. However, the effect was only statistically significant when applying double machine learning with a binary treatment (and not when considering the causal forest). Investigating effect heterogeneity across observables suggested that the effects are lower among always buyers with a half-fare travel ticket.

Finally, use case (iv) concerned the effect of supersaver tickets on upselling, i.e., purchasing a first-class rather than a second-class ticket, using a more lenient definition always buyers that permitted for changes in the travel class. Applying the causal forest suggested that increasing the discount by 1 percentage point increases the share of tickets being upgraded by 0.59 percentage points. Investigating effect heterogeneity across a pre-selected characteristics pointed to lower effects among elderly and those customers possessing a half-fare ticket. However, due to the more lenient definition of always buyers, the assumptions required for the identification of causal effects are likely violated. Therefore, our findings in use case (iv) need to be interpreted with caution, as the estimates might be substantially biased.

Using state-of-the art machine learning tools, our study appears to be the first (at least for Switzerland) to provide empirical evidence on how discounts on train tickets affect customers' willingness to reschedule trips. These insights may be helpful for designing discount schemes aiming at balancing out train utilization across daytime and reducing overload during peak hours. Even though the overall impact on the demand shifts on always buyers might not be as large as one could hope for, the causal forest pointed to the existence of customer segments that are likely more responsive and could be scrutinized when collecting a larger amount of data than available for our analysis. In addition, our study is the first to provide empirical results on how discounts affect customer satisfaction and upselling related to train trips in Switzerland, as well as on the prediction of various aspects of customer behavior. More generally, our study can be regarded as use cases for how predictive and causal machine learning can also be fruitfully applied for business analytics and as decision support for optimizing specific interventions like discount schemes based on impact evaluation. Future studies would ideally aim not only for a larger, but also a more representative survey covering also buyers of regular (rather than supersaver) tickets and could in addition rely on a random assignment of discounts by means of an experiment in order to avoid the previously mentioned selection issues.

Appendix

A Statistics on offered and sold supersaver tickets

This section provides two tables on the total and relative amount of supersaver tickets offered and sold by the SBB. While these descriptive statistics give an overview of the supply of discounted tickets, they must be interpreted with caution. In particular, they represent averages at the time of purchase of the individuals in our survey-based sample and not the average of the entire supply. The offers of supersaver tickets, which change over time (become smaller), always refer to the time of purchase. The number of tickets sold is calculated based on information about a customers' discount (according to the survey data) and the offered discounts within specific discount levels (according to factors determining the supply of supersaver tickets). The latter is indexed by A to E depending on the discount level, differing (slightly) between half-fare and full-fare travelers. The index is in descending order. For instance, A denotes the highest possible discount of 70% on the standard ticktet price. On the other hand, E denotes the lowest possible discount being 10% and 20% for half-fare and full-fare travelers, respectively. Table 13 gives the absolute numbers of discounts and Table 14 the relative shares by discount level and customer type (i.e. always buyers: yes or no). For instance, we see in Table 13 that the SBB offers on average a total of 98.34 supersaver tickets when an individual's discount is higher than or equal to 30% and the individual is an always buyer. This offer represents 22% of all seats (Table 14).

Table 13: Mean and standard deviation of numbers of supersaver tickets offered and sold by discount level and customer type

discount	$< 30\%$		$\geq 30\%$	
always buyers	No	Yes	No	Yes
offer total	33.95	44.10	70.97	98.34
	(42.57)	(50.68)	(69.57)	(84.45)
offer sold total	28.04	37.29	13.70	25.75
	(41.92)	(50.31)	(36.37)	(53.67)
offer level A	10.80	14.02	31.76	40.76
	(29.36)	(37.89)	(54.90)	(70.28)
offer level B	10.26	13.53	23.59	34.61
	(22.30)	(25.81)	(35.33)	(45.75)
offer level C	8.14	10.10	11.42	16.99
	(14.05)	(15.95)	(17.27)	(21.90)
offer level D	3.88	5.40	3.51	5.14
	(6.66)	(9.15)	(7.13)	(9.39)
offer level E	0.87	1.05	0.69	0.85
	(2.61)	(3.02)	(2.37)	(2.80)
sold level A	10.80	14.02	10.13	18.74
	(29.36)	(37.89)	(31.92)	(46.84)
sold level B	10.26	13.53	2.55	5.77
	(22.30)	(25.81)	(12.66)	(21.06)
sold level C	5.35	7.24	1.01	1.23
	(11.96)	(14.22)	(6.13)	(6.69)
sold level D	1.57	2.28	0.00	0.00
	(4.56)	(6.37)	(0.00)	(0.00)
sold level E	0.05	0.22	0.00	0.00
	(0.53)	(1.38)	(0.00)	(0.00)
Obs.	1151	2221	5529	3745

Notes: 'Offer level A' denotes the number of offered supersaver tickets with discount level A. 'Sold level A' denotes the number of supersaver tickets sold with discount level A.

Table 14: Mean and standard deviation of shares of supersaver tickets offered and sold by discount level and customer type

discount	< 30%		≥ 30%	
always buyers	No	Yes	No	Yes
rel. offer total	0.11	0.10	0.24	0.22
	(0.11)	(0.11)	(0.17)	(0.16)
rel. sold total	0.09	0.09	0.04	0.06
	(0.11)	(0.11)	(0.10)	(0.12)
rel. sold level A	0.04	0.03	0.11	0.10
	(0.10)	(0.09)	(0.18)	(0.15)
rel. offer level B	0.03	0.03	0.08	0.08
	(0.05)	(0.05)	(0.08)	(0.09)
rel. offer level C	0.02	0.02	0.04	0.04
	(0.04)	(0.03)	(0.05)	(0.04)
rel. offer level D	0.01	0.01	0.01	0.01
	(0.02)	(0.02)	(0.02)	(0.02)
rel. offer level E	0.00	0.00	0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.01)
rel. sold level A	0.04	0.03	0.03	0.04
	(0.10)	(0.09)	(0.09)	(0.10)
rel. sold level B	0.03	0.03	0.01	0.01
	(0.05)	(0.05)	(0.03)	(0.04)
rel. sold level C	0.02	0.02	0.00	0.00
	(0.03)	(0.03)	(0.01)	(0.01)
rel. sold level D	0.00	0.01	0.00	0.00
	(0.01)	(0.01)	(0.00)	(0.00)
rel. sold level E	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Obs.	1151	2221	5529	3745

Notes: 'Rel. offer level A' denotes the share of supersaver tickets offered with discount level A. 'Rel. sold level A' denotes the share of supersaver tickets sold with discount level A. The shares are relative to the seats offered.

B Propensity score plots

Figure B.1: Propensity score estimates in the higher discount category ($D \geq 0.3$)

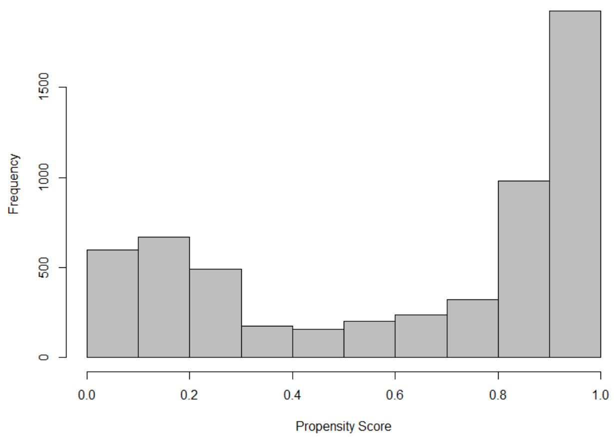
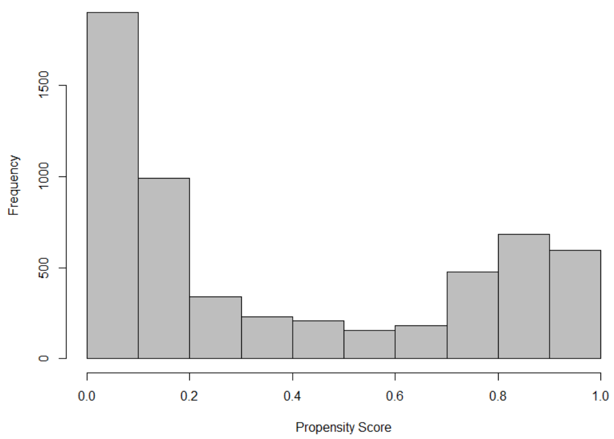


Figure B.2: Propensity score estimates in the lower discount category ($D < 0.3$)



C Descriptive statistics for customer satisfaction

Thsi section presents descriptive statistics for the survey question *What is your overall impression of SBB?* The variable takes values from 1 (very bad) to 10 (very good). Table 15 shows the mean and standard deviation by discount categories and customer type (always buyers: yes or no). In general, the survey participants are quite satisfied, with the highest average satisfaction occurring among discounts $\geq$ 30% and for always buyers (albeit differences across groups are generally small).

Table 15: Mean and standard deviation of satisfaction with SBB by discount level and type

discount	$< 30\%$		$\geq 30\%$	
always buyers	No	Yes	No	Yes
customer satisfaction	7.40	7.62	7.74	7.86
	(1.73)	(1.54)	(1.62)	(1.51)
Obs.	1151	2221	5529	3745

Note: Customer satisfaction is based on a scale from 1 (very bad) to 10 (very good).

D The effect on upselling: testing the identification strategy

As for the use cases (ii) and (iii), we consider two different methods to partially test the assumptions underlying our identification strategy when assessing the discounts' effects on upselling based on the modified definition of always buyers. First, we test weak monotonicity (Assumption 3) by running the causal forest (CF) and double machine learning (DML) procedures as well as a conventional OLS regression in which we use buying an additional trip $(1 - S(0))$, i.e. not being an always buyer, as outcome variable and X as control variables in our sample of supersaver customers. The CF permits estimating the conditional change in the share of surveyed customers induced to buy an additional trip by modifying the discount rate D given X, i.e. $\frac{\partial E[(1-S(0))|D,X,S=1]}{\partial D}$, as well as the average thereof across X conditional on sample selection, $E\left[\frac{\partial E[(1-S(0))|D,X,S=1]}{\partial D}\middle|S=1\right]$. DML, on the other hand, yields an estimate of the average difference in the share of additional trips across the high and low treatment categories conditional on sample selection, $E[E[(1-S(0))|D<0.3,X,S=1] - E[(1-S(0))|D \geq 0.3,X,S=1]|S=1]$. Finally, the OLS regression of $(1 - S(0))$ on D and all X in our sample tests monotonicity when assuming a linear model.

Table 16 reports the results that do not provide evidence against the monotonicity assumption. When considering the continuous treatment D, the CF-based estimate of $E\left[\frac{\partial E[(1-S(0))|D,X,S=1]}{\partial D}\middle|S=1\right]$ is highly statistically significant and suggests that augmenting the discount by one percentage point increases the share of customers otherwise not buying the ticket by 0.207 percentage points on average. However, this is only about half of the magnitude when compared to the original definition of always buyers (0.56). Any (almost all) estimates of the conditional change $\frac{\partial E[(1-S(0))|D,X,S=1]}{\partial D}$ are positive, as displayed in the histogram of Figure D.1, and 21.1% of them are statistically significant at the 10% level, 11.2% at the 5% level. Also the OLS coefficient of 0.291 is highly significant, but also lower than for the original definition of always buyers (0.544). Likewise, the statistically significant DML estimate points to an increase in the share of additional trips by 8.4 percentage points when switching the binary treatment indicator from $D < 0.3$ to $D \geq 0.3$.

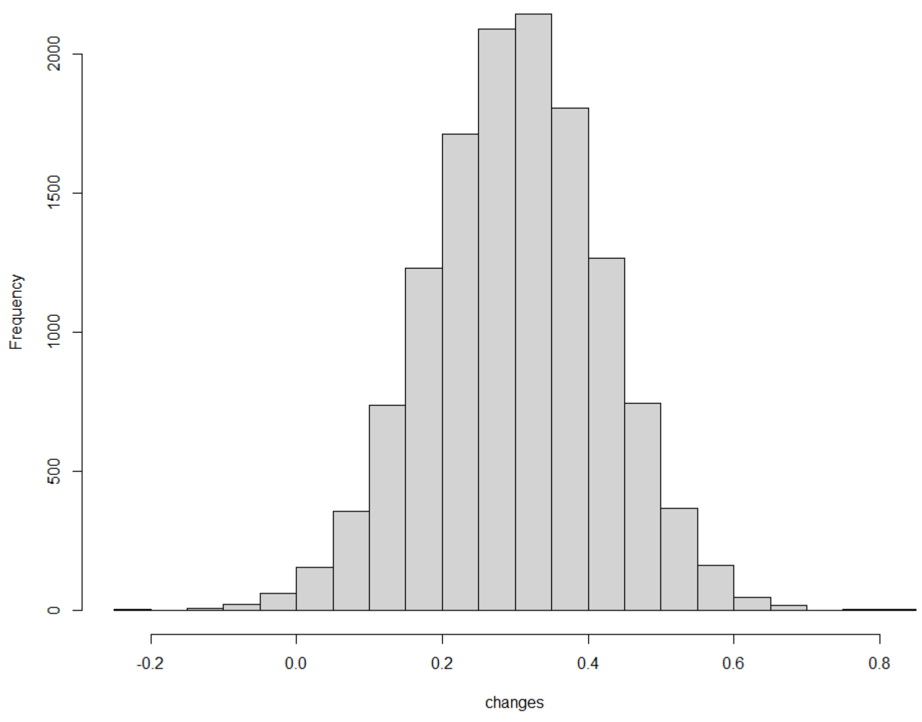
Table 16: Monotonicity tests

	CF: av. change	OLS: coef.	DML: $D \geq 0.3$ vs $D < 0.3$
change in (1-S(0))	0.207	0.236	0.084
standard error	0.061	0.028	0.008
p-value	0.001	0.000	0.000
trimmed observations			1826
number of observations	12924		

Notes: 'CF', 'OLS', and 'DML' stands for estimates based on causal forests, linear regression, and double machine learning, respectively. 'trimmed observations' is the number of trimmed observations in DML when setting the propensity score-based trimming threshold to 0.01. *Control variables consist of* X.

We also test the statistical independence of D and W conditional on X in our newly defined sample of always buyers, as implied by our identifying assumptions. To this end, we randomly split the evaluation data into a training set (25% of observations) and a test set (75% of all observations). In the training data set, we run a linear lasso regression (Tibshirani, 1996) of D on X in order to identify important predictors by means of 10-fold cross-validation. In the next step, we select all covariates in X with non-zero lasso coefficients and run an OLS regression of D on the selected covariates in the test data. Finally, we add W to that regression in the test data and run a Wald test to compare the predictive power of the models with and without W. We repeat the procedure of splitting the data, performing the lasso regression in the training set, and running the OLS regressions and the Wald test in the test set 100 times. This yields an average p-value of 0.091, with 48 out of 100 p-values being smaller than 5%. These results provide statistical evidence that W might be associated with D conditional on X. Therefore, the causal effects of the discounts on upselling should be interpreted with much caution.

Figure D.1: Monotonicity given X



E Predictive outcome analysis separately for subsamples

The Tables 18 and 17 present the predictive outcome analysis separately for subsamples with discounts $\geq 30\%$ and $< 30\%$, respectively.

Table 17: Predictive outcome analysis, $D < 0.3$

demand shift		upselling		additional trip	
variable	importance	variable	importance	variable	importance
departure time	37.33	capacity utilization	41.387	seat capacity	25.342
seat capacity	27.871	offer level D	27.669	age	21.639
capacity utilization	26.508	age	22.145	capacity utilization	20.168
distance	26.31	D	19.077	distance	18.970
age	26.223	offer level C	17.324	departure time	18.527
D	25.08	departure time	16.538	D	18.076
number of sub-journeys	17.403	distance	15.897	ticket purchase complexity	16.637
offer level C	15.643	offer level B	15.696	offer level C	12.085
diff. purchase travel	15.299	rel. sold level B	10.992	rel. sold level B	11.709
ticket purchase complexity	15.116	number of sub-journeys	10.019	offer level D	11.641
rel. sold level B	15.03	diff. purchase travel	9.573	number of sub-journeys	11.347
offer level D	15.012	rel. sold level C	9.367	diff. purchase travel	10.328
rel. sold level C	14.625	offer level A	7.857	offer level B	10.185
offer level B	14.413	rel. sold level D	7.33	rel. sold level C	8.993
rel. sold level A	11.856	ticket purchase complexity	7.319	rel. sold level A	8.162
offer level A	11.329	offer level E	7.183	offer level A	7.643
rel. amount imputed values	10.079	rel. sold level A	6.769	class	7.381
rel. sold level D	9.625	rel. amount imputed values	5.422	rel. sold level D	6.964
adult companions	8.503	adult companions	4.881	rel. amount imputed values	6.189
offer level E	6.511	rush hour	4.143	adult companions	5.785
gender	5.214	leisure	3.602	leisure	5.398
leisure	5.154	gender	3.599	offer level E	4.866
destination Geneva Airport	4.83	2019	3.597	gender	4.692
departure Zuerich	4.736	travel alone	3.047	German	3.801
class	4.598	Friday	2.92	halfe fare travel ticket	3.686
travel alone	4.59	German	2.825	travel alone	3.540
peak hour	4.545	French	2.479	French	3.493
Friday	4.524	departure Zuerich	2.429	Friday	3.419
German	4.522	destination Zuerich Airport	2.427	half fare	3.163
amount purchased tickets	4.349	scheme 20	2.427	2019	3.136
correct prediction rate	0.555		0.772		0.605
balanced sample size	1642		1140		1202

Notes: 'Diff. purchase travel' denotes the difference between purchase and travel day. 'Rel. offer level A', 'rel. offer level B', 'rel. offer level C' and 'rel. offer level D' denote the relative amount of supersaver tickets offered with discount A, B, C and D respectively. The relative amounts are in relation to the seats offered. 'Offer level A', 'Offer level B', 'Offer level C', 'Offer level D' and 'Offer level E' denotes the amount of supersaver tickets with discount A, B, C, D and E respectively. 'No subscription' indicates not possessing any subscription. For predicting upselling, the covariates 'class' and 'seat capacity' are dropped.

Table 18: Predictive outcome analysis, $D \geq 0.3$

demand shift		upselling		additional trip	
variable	importance	variable	importance	variable	importance
departure time	114	seat capacity	246.396	capacity utilization	133.936
seat capacity	95.799	offer level B	178.212	age	105.107
age	95.209	offer level C	127.327	departure time	100.091
capacity utilization	89.422	D	100.618	capacity utilization	97.889
distance	85.503	offer level A	88.947	distance	85.647
D	80.447	Tageszeitinmin	82.658	D	83.671
diff. purchase travel	69.276	age	78.886	offer level B	73.399
offer level B	68.75	distance	73.885	offer level A	69.936
offer level A	65.766	offer level D	72.452	diff. purchase travel	67.823
offer level C	60.513	diff. purchase travel	55.321	offer level C	64.689
number of sub-journeys	57.767	number of sub-journeys	48.622	number of sub-journeys	58.100
offer level D	44.626	rel. sold level A	38.997	class	54.857
ticket purchase complexity	44.434	ticket purchase complexity	31.991	ticket purchase complexity	49.348
rel. sold level A	39.796	offer level E	27.041	offer level D	46.685
rel. amount imputed values	32.144	rel. amount imputed values	25.586	rel. sold level A	42.589
adult companions	25.925	adult companions	24.73	rel. amount imputed values	35.411
rel. sold level B	20.536	rel. sold level B	17.099	half fare	31.308
gender	18.629	gender	15.835	adult companions	28.732
offer level E	17.521	2019	15.416	half fare travel ticket	23.900
travel alone	15.15	Saturday	14.256	gender	20.562
amount purchased tickets	14.939	amount purchased tickets	13.396	rel. sold level B	20.036
German	14.855	rush hour	12.947	offer level E	18.595
French	14.415	German	12.878	German	16.257
2019	14.387	leisure	12.344	amount purchased tickets	15.847
Sunday	13.821	travel alone	12.28	leisure	15.500
destination Zuerich Airport	13.387	half fare	11.882	no subscription	15.434
Saturday	13.378	Friday	11.646	travel alone	15.132
class	13.27	scheme 20	11.559	Swiss	14.951
half fare	13.258	French	11.477	Saturday	14.613
rel. amount imputed values	13.048	Sunday	11.39	2019	14.279
correct prediction rate	0.589		0.809		0.629
balanced sample size	5320		5598		5798

Notes: 'Diff. purchase travel' denotes the difference between purchase and travel day. 'Rel. offer level A', 'rel. offer level B', 'rel. offer level C' and 'rel. offer level D' denote the relative amount of supersaver tickets offered with discount A, B, C and D respectively. The relative amounts are in relation to the seats offered. 'Offer level A', 'Offer level B', 'Offer level C', 'Offer level D' and 'Offer level E' denotes the amount of supersaver tickets with discount A, B, C, D and E respectively. 'No subscription' indicates not possessing any subscription. For predicting upselling, the covariates 'class' and 'seat capacity' are dropped.

Bibliography

ANGRIST, J., G. IMBENS, AND D. RUBIN (1996): "Identification of Causal Effects using Instrumental Variables," *Journal of American Statistical Association*, 91, 444–472.

ASCARZA, E. (2018): "Retention futility: Targeting high-risk customers might be ineffective," *Journal of Marketing Research*, 55(1), 80–98.

ATHEY, S., AND G. IMBENS (2016): "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360.

ATHEY, S., J. TIBSHIRANI, AND S. WAGER (2019): "Generalized random forests," *The Annals of Statistics*, 47, 1148–1178.

BASSO, L. J., AND H. E. SILVA (2014): "Efficiency and substitutability of transit subsidies and other urban transport policies," *American Economic Journal: Economic Policy*, 6(4), 1–33.

BATTY, P., R. PALACIN, AND A. GONZÁLEZ-GIL (2015): "Challenges and opportunities in developing urban modal shift," *Travel Behaviour and Society*, 2(2), 109–123.

BLANCO, G., C. A. FLORES, AND A. FLORES-LAGUNES (2011): "Bounds on Quantile Treatment Effects of Job Corps on Participants' Wages," Discussion paper.

BODORY, H., AND M. HUBER (2018): "The causalweight package for causal inference in R," *SES Working Paper 493, University of Fribourg.*

BREIMAN, L. (2001): "Random forests," *Machine Learning*, 45, 5–32.

BREIMAN, L. (2018): "randomForest: Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.6-12," *Software available at URL: https://cran. r-project. org/package= randomForest.*

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984): *Classification and Regression Trees*. Wadsworth, Belmont, California.

Brynjolfsson, E., and K. McElheran (2016): "The rapid adoption of data-driven decision-making," *American Economic Review*, 106(5), 133–39.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

De Palma, A., R. Lindsey, and G. Monchambert (2017): "The economics of crowding in rail transit," *Journal of Urban Economics*, 101, 106–122.

De Witte, A., C. Macharis, P. Lannoy, C. Polain, T. Steenberghen, and S. Van de Walle (2006): "The impact of "free" public transport: The case of Brussels," *Transportation Research Part A: Policy and Practice*, 40(8), 671–689.

Desmaris, C. (2014): "The reform of passenger rail in Switzerland: More performance without competition," *Research in Transportation Economics*, 48, 290–297.

Frangakis, C., and D. Rubin (2002): "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29.

Hagenauer, J., and M. Helbich (2017): "A comparative study of machine learning classifiers for modeling travel mode choice," *Expert Systems with Applications*, 78, 273–282.

Heckman, J. J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Simple Estimator for such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

——— (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153–161.

Huber, M. (2014): "Treatment evaluation in the presence of sample selection," *Econometric Reviews*, 33, 869–905.

Huber, M., J. Meier, and H. Wallimann (2021): "Business analytics meets artificial intelligence: Assessing the demand effects of discounts on Swiss train tickets," *arXiv working paper*.

Hünermund, P., J. Kaminski, and C. Schmitt (2021): "Causal Machine Learning and Business Decision Making," .

Imai, K. (2008): "Sharp bounds on the causal effects in randomized experiments with 'truncation-by-death'," *Statistics & Probability Letters*, 78, 144–149.

Imbens, G. W., and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Knaus, M. C. (2021): "A double machine learning approach to estimate the effects of musical practice on student's skills," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(1), 282–300.

Lee, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76, 1071–1102.

Little, R., and D. Rubin (1987): *Statistical Analysis with Missing Data*. Wiley, New York.

Liu, L., and R.-C. Chen (2017): "A novel passenger flow prediction model using deep learning methods," *Transportation Research Part C: Emerging Technologies*, 84, 74–91.

Liu, Y., Z. Liu, and R. Jia (2019): "DeepPF: A deep learning based architecture for metro passenger flow prediction," *Transportation Research Part C: Emerging Technologies*, 101, 18–34.

Lüscher, R. (2020): "10 Jahre Sparbillette – Attraktive Preise und höhere Nachfrage für den öV.," Discussion paper.

Mohring, H. (1972): "Optimization and scale economies in urban bus transportation," *The American Economic Review*, 62(4), 591–604.

Omrani, H. (2015): "Predicting travel mode of individuals by machine learning," *Transportation Research Procedia*, 10, 840–849.

Parry, I. W., and K. A. Small (2009): "Should urban transit subsidies be reduced?," *American Economic Review*, 99(3), 700–724.

Paulley, N., R. Balcombe, R. Mackett, H. Titheridge, J. Preston, M. Wardman, J. Shires, and P. White (2006): "The demand for public transport: The effects of fares, quality of service, income and car ownership," *Transport policy*, 13(4), 295–306.

Pearl, J. (2000): *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge.

Redman, L., M. Friman, T. Gärling, and T. Hartig (2013): "Quality attributes of public transport that attract car users: A research review," *Transport policy*, 25, 119–127.

Robins, J. M., and A. Rotnitzky (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90, 122–129.

Robins, J. M., A. Rotnitzky, and L. Zhao (1994): "Estimation of Regression Coefficients When Some Regressors Are not Always Observed," *Journal of the American Statistical Association*, 90, 846–866.

Rotaris, L., and R. Danielis (2014): "The impact of transportation demand management policies on commuting to college facilities: A case study at the University of Trieste, Italy," *Transportation Research Part A: Policy and Practice*, 67, 127–140.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (1976): "Inference and Missing Data," *Biometrika*, 63, 581–592.

Semenova, V., and V. Chernozhukov (2020): "Debiased machine learning of conditional average treatment effects and other causal functions," *forthcoming in the Econometrics Journal.*

Thao, V. T., W. von Arx, and J. Frölicher (2020): "Swiss cooperation in the travel and tourism sector: long-term relationships and superior performance," *Journal of Travel Research*, 59(6), 1044–1060.

Tibshirani, J., S. Athey, R. Friedberg, V. Hadad, D. Hirshberg, L. Miner, E. Sverdrup, S. Wager, and M. Wright (2020): "grf: Generalized Random Forests," *R package version 1.2.0.*

Tibshirani, R. (1996): "Regresson shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society*, 58, 267–288.

van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007): "Super Learner," *Statistical Applications in Genetics and Molecular Biology*, 6.

WAGER, S., AND S. ATHEY (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242.

WEGELIN, P. (2018): "Is the mere threat enough? An empirical analysis about competitive tendering as a threat and cost efficiency in public bus transportation," *Research in Transportation Economics*, 69, 245–253.

WRIGHT, M. N., AND A. ZIEGLER (2017): "ranger: A fast implementation of random forests for high dimensional data in C++ and R," *Journal of Statistical Software*, 77, 1–17.

YANG, J.-C., H.-C. CHUANG, AND C.-M. KUAN (2020): "Double machine learning with gradient boosting and its application to the Big N audit quality effect," *Journal of Econometrics*, 216(1), 268–283.

YAP, M., AND O. CATS (2020): "Predicting disruptions and their passenger delay impacts for public transport stops," *Transportation*, pp. 1–29.

ZHANG, J., R. LINDSEY, AND H. YANG (2018): "Public transit service frequency and fares with heterogeneous users under monopoly and alternative regulatory policies," *Transportation Research Part B: Methodological*, 117, 190–208.

ZHANG, J., AND D. B. RUBIN (2003): "Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by 'Death'," *Journal of Educational and Behavioral Statistics*, 4, 353–368.

ZHANG, J., D. B. RUBIN, AND F. MEALLI (2008): "Evaluating The Effects of Job Training Programs on Wages through Principal Stratification," in *Advances in Econometrics: Modelling and Evaluating Treatment Effects in Econometrics*, ed. by D. Millimet, J. Smith, and E. Vytlacil, vol. 21, pp. 117–145. Elsevier Science Ltd.