LUND UNIVERSITY
School of Economics and Management

Master's Programme in Data Analytics and Business Economics

Estimating the Impact of Website Changes on Conversion Rates

A Double Machine Learning Approach

by

Jan Jakub Jarco, ja1710ja-s@lu.se

Abstract This study sought to evaluate the historical impact of changes to an ordering page of an online travel agency on its conversion rates. Data gathered from the website over a year, detailing aspects such as travel dates, prices, itineraries, number of passengers, travel time, and carriers, was analyzed. External data sources were also included, with the dataset covering 12 changes to the website's layout and payment process. The changes' effectiveness was assessed using three methods: comparing conversion rates before and after the changes, a modified linear regression model, and the Double Machine Learning (DML) method with Random Forests as the base learners. The analysis revealed that the only modification with a statistically significant positive impact on conversion rates was related bug fixing. Most changes did not significantly affect conversion rates, and some even demonstrated a non-significant negative impact. The DML method proved a useful tool in this context, outperforming simpler comparison methods with better control for confounding variables and reducing potential bias in Average Treatment Effect (ATE) estimation. However, estimates from the DML model were sensitive to the analysis time window. This study suggests future website design should focus on user-friendly and intuitive design, clear and detailed information provision, and careful evaluation of changes' potential impact on user experience.

Keywords: website design, causal inference, observational study, average treatment effects, double machine learning

Master's Thesis (15 credits ECTS)
May 2023
Supervisor: Simon Reese

Acknowledgements

I want to thank my supervisor, Simon Reese, for introducing me to the world of causal inference through his lectures during this program. This helped me see statistics in a new way, beyond simple correlations. I am very thankful for a huge support while I was writing my thesis. I'm also grateful to Robert from AOB Travel for sharing data for my study and helping me with difficult questions, without this input this study wouldn't ever be made. I want to thank my former math teacher, Grażyna Zielińska, for showing me the amazing world of mathematics and teaching me the right way to think in science. Lastly, I'm thankful to my mother who showed me how exciting and worth exploring the world is and to my dad who taught me the way to be consistent and clever in my journey.

List of Figures

1 Screenshot of flysmarter.dk home page layout 5

2 High level flow for A/B testing procedure, adapted from Kohavi, Deng, et al. 2012 7

3 Directed Acyclic Graph of simplified Structural Causal Model 9

4 Correlation map of features used as covariates for each treatment generated on full dataset 15

5 Plot of frequencies from the trained Random Forest model. The features along the y-axis represent the different predictors used in the model. The frequency metric on the x-axis represents the importance of each feature. Features with higher scores have more impact on the model's decisions. The features we consider confounders are clearly labeled. 16

6 Diagram showing how assignment to control and treatment groups was conducted. For each treatment evaluated in the study appropriate time window in days was calculated. From two time windows' lengths $T_1 - T_0$ and $T_2 - T_1$ the lower value constitutes as a symmetrical time window's length. For a given time window control period and treatment period were assigned ensuring the lack of overlap between periods analyzed for each treatment. 17

7 Aggregated weekly conversion rates index for the Danish market with dashed lines analyzed changes to the website were shown together with analyzed placebo effects evaluated 18

8 Directed Acyclic Graph of Structural Causal Model 20

9 Scatter plot showing distribution of ATE coefficients with standard, not adjusted p-values of these coefficients 26

10 Scatter plot showing the distribution of ATEs from Differences in Means and DML PLR 27

List of Tables

1 List of changes made to the website in analyzed period 19

2 List of placebo changes evaluated in the modelling process 20

3 Estimated ATEs (in percentage points) and Holm's adjusted p-values for each treatment, where . if p<0.1, * if p<0.05, ** if p<0.01, *** if p<0.001. 24

4 Estimated ATEs (in percentage points) and Holm's adjusted p-values for each treatment, where . if p<0.1, * if p<0.05, ** if p<0.01, *** if p<0.001. 25

5 Estimated ATEs (in p.p.) and Holm's adjusted p-values DML PLR with different time windows 27

6 Summary of variables used as covariates in the modelling 34

Contents

1 Introduction 5

2 Literature Review 7
2.1 Online Controlled Experiments . . . 7
2.1.1 Randomized Control Trials . . . 7
2.1.2 A/B Testing . . . 7
2.1.3 Limitations of A/B Testing . . . 8
2.1.4 Scenarios Restricting the Feasibility of Online Controlled Experiments . . . 8
2.2 Average Treatment Effects estimation . . . 8
2.2.1 Methods for Average Treatment Effects estimation . . . 9
2.3 Research Gap . . . 10

3 Data 11
3.1 Data Collection . . . 11
3.1.1 Company data . . . 11
3.1.2 External Data Sources . . . 12
3.2 Data Preprocessing . . . 12
3.3 Feature engineering . . . 14
3.4 Feature selection . . . 15
3.5 Data Splitting and Transformation . . . 16

4 Methodology 18
4.1 ATE estimation using conversion rates . . . 18
4.2 Treatments description . . . 19
4.3 Placebo changes . . . 19
4.4 Structural Causal Model . . . 20
4.5 Models for ATE Estimation . . . 21
4.5.1 Comparing differences . . . 21
4.5.2 Linear Regression . . . 21
4.5.3 Double Machine Learning . . . 22
4.6 Ensemble methods as base learners . . . 23

5 Results and Discussion 24
5.1 Comparison of Models' Results . . . 24
5.1.1 Including placebo treatments . . . 25
5.2 Double Machine Learning . . . 26
5.3 Insights and Analysis of Results . . . 28
5.3.1 Implications and Recommendations for Future Development . . . 28
5.4 Feasibility of Double Machine Learning . . . 29
5.5 Limitations of the Study . . . 29
5.6 Future Work . . . 30

6 Conclusions 31

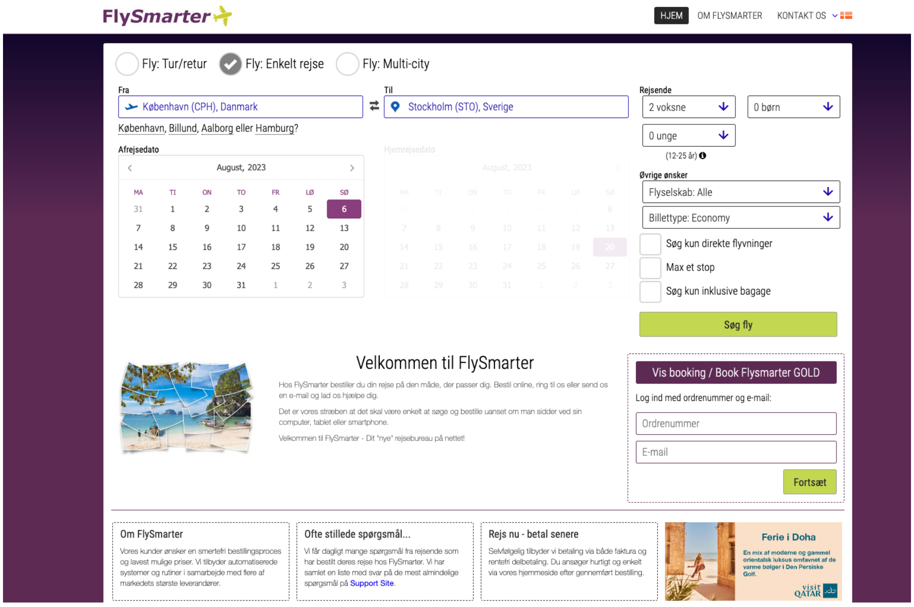
A Appendix 34
A.1 Data Dictionary . . . 34

1 Introduction

The travel industry has seen significant growth in recent years, with online travel agencies (OTAs) playing a crucial role in facilitating travel bookings for consumers. One such OTA, AOB Travel, is a Swedish agency operating in Sweden, Denmark, Norway, Finland, the Netherlands, and Germany. As a meta-search engine, AOB Travel aggregates and displays results from various flight search engines on their own servers and presents these results to travel search engines such as Momondo, Skyscanner, Flygresor, and Travelmarket. Customers seeking travel options can compare prices from multiple travel agencies through the travel search engine and select from which OTA they will order flight tickets. They are then redirected from these travel search engines to the ordering website maintained by AOB Travel, where they can finalize an order for flight tickets.

A well-designed website is the key to achieving high returns in e-commerce. Particularly, within the online travel agency sector, the effectiveness of the website is more important than anything else due to its direct impact on customer behavior and conversion rates. About 80% of consumers looking for flight tickets visit more than one provider before ordering a ticket, suggesting a user-friendly website design alongside an attractive price is crucial for the effectiveness of online travel agencies (Statista 2017). Within the tourism industry, the website serves not only as a platform for business transactions but also as an essential part of the marketing strategy (Chan et al. 2021). This is the reason why web developers need to constantly make changes to the front end and back end of the website to maintain high performance.

Role of AOB Travel in the study As an online travel agency, AOB Travel offers a broad range of travel options from various providers, making it a suitable candidate for examining the effects of changes to the website on a diverse set of customers. Particularly, this study focuses on the Danish market where the flysmarter.dk webpage is available to customers. This AOB Travel's website is the one with the most significant number of changes, providing valuable insights into the impact of changes in the layout and payment process on the conversion rate (AOB Travel 2023).



○ Fly: Tur/retur ● Fly: Enkelt rejse ○ Fly: Multi-city

Fra: København (CPH), Danmark

Til: Stockholm (STO), Sverige

København, Billund, Aalborg eller Hamburg?

Rejsende: 2 voksne; 0 børn; 0 unge (12-25 år)

Øvrige ønsker: Flyselskab: Alle; Billettype: Economy

□ Søg kun direkte flyvninger

□ Max et stop

□ Søg kun inklusive bagage

Velkommen til FlySmarter

Hos FlySmarter bestiller du din rejse på den måde, der passer dig. Bestil online, ring til os eller send os en e-mail og lad os hjælpe dig.

Det er vores stræben at det skal være enkelt at søge og bestille uanset om man sidder ved sin computer, tablet eller smartphone.

Velkommen til FlySmarter - Dit "nye" rejsebureau på nettet!

Log ind med ordrenummer og e-mail:

Ordrenummer

E-mail

Om FlySmarter

Vores kunder ønsker en smertefri bestillingsproces og lavest mulige priser. Vi tilbyder automatiserede systemer og rutiner i samarbejde med flere af markedets største leverandører.

Ofte stillede spørgsmål...

Vi får dagligt mange spørgsmål fra rejsende som har bestilt deres rejse hos FlySmarter. Vi har samlet en liste med svar på de mest almindelige spørgsmål på Support Site.

Rejs nu - betal senere

Selvfølgelig tilbyder vi betaling via både faktura og rentefri delbetaling. Du ansøger hurtigt og enkelt via vores hjemmeside efter gennemført bestilling.

Figure 1: Screenshot of flysmarter.dk home page layout

Aim of the study This research paper aims to estimate the average treatment effects of changes in the ordering page and how they contribute to the ratio of orders and visits on their website from travel search engines to provide actionable insights for future website optimization processes.

Usually, web developers would use A/B testing for the evaluation of changes to the website. This method has been the gold standard for evaluation in online controlled experiments (Kohavi, Deng, et al. 2012; Siroker and Koomen 2013; Luca and Bazerman 2021). This method allows for assessing the causal impact of the proposed changes in controlled randomized experiments, offering robust insights for decision-making.

However, there are circumstances where web developers seek to provide insights on historical changes where online controlled experiments were not conducted. Also, conducting A/B tests may be unfeasible due to a lack of control over the causal action, high opportunity costs, expensive implementation, impracticality of proper randomization, or ethical and legal limitations (Kohavi, Tang, and Y. Xu 2020). In such scenarios, methods that leverage machine learning algorithms offer alternative ways of identifying Average Treatment Effects (ATEs) in the presence of confounding factors and have gained increased attention due to their capabilities to capture complex relations between variables (Athey and Wager 2019; Karmakar, Majumder, and Gangaraju 2023). In particular, this thesis will focus on Double Machine Learning (DML), an approach that has shown promising results for bias reduction in observational studies when compared to standard methods (Chernozhukov et al. 2018).

Data about website visits and resulting orders were gathered to evaluate the impact of changes and provide insights for future website design. This data, including details about travel dates, prices, itineraries, number of passengers, travel time, and carriers, helped identify potential confounding variables. Data from external sources was incorporated to expand the dataset. Over the past year, a total of 12 changes were made to the website's layout and payment process.

To test each change, the data was divided into two sub-groups - the period before the change (the control group) and the period after the change (the treatment group). For each treatment, the set of confounding variables and an output variable were evaluated, with the output variable indicating whether the visit resulted in order or not.

The effectiveness of the changes was assessed using three different methods:

1. The conversion rates before and after the changes were compared
2. A modified linear regression model was used, which included the ordinary treatment variable along with the set of confounders
3. The Double Machine Learning method was employed, utilizing Partially Linear Regression with Random Forests as the base learners

In the context of evaluating the impact of changes on the conversion rate of an ordering website, the Double Machine Learning (DML) framework appears to be a suitable choice, as will be further discussed in the following sections.

The remainder of this paper is structured as follows: The literature review in Section 2 provides an overview of causal inference in website performance, with an exploration of Randomized Control Trials (RCTs), A/B testing as an analogue of RCTs, its limitations, scenarios restricting the feasibility of randomized controlled experiments, and implications for the study. Also covered is the estimation of Average Treatment Effects (ATE), various models for ATE estimation and ensemble methods used as base learners.

Section 3 presents a discussion of the data used in this research. An overview of AOB Travel is provided, the data collection process is described, and the steps for data preprocessing, feature engineering, selection, and data splitting are outlined.

The methodology is detailed in Section 4, starting with ATE estimation using conversion rates, followed by the description of treatments, placebo changes, and the structural causal model utilized in this research. Lastly, the models and their evaluation are explained.

In Section 5, the results of the study are presented, including model training and evaluation, differences in conversion rates, and a comparison of different models' results.

Finally, Section 6 concludes the research, offering potential areas for improvement, and discussing the implications for future booking system design. Additional details such as a data dictionary and detailed model results are provided in the Appendices.

2 Literature Review

2.1 Online Controlled Experiments

2.1.1 Randomized Control Trials

Randomized Control Trials (RCTs) are a gold standard of experimental research methodology. The main principle behind an RCT is the random assignment of participants to treatment and control groups to eliminate confounding factors. As discussed in (Imbens and Rubin 2015), there are four primary assumptions for RCTs: (i) the assignment is individualistic - with the dependence on values of covariates and potential outcomes for other units limited; (ii) the assignment is probabilistic, where each unit has a non-zero probability of receiving either treatment; (iii) the assignment is unconfounded, meaning it does not depend on potential outcomes given covariates; and (iv) the assignment mechanism has a known, controlled functional form. Various types of RCTs exist, each defining the treatment assignment of the population differently, including Bernoulli trials, completely randomized experiments, stratified randomized experiments, and paired randomized experiments.

2.1.2 A/B Testing

A/B testing, also referred to as online controlled experiments, is a fundamental practice in web analytics, resembling the principles of Randomized Controlled Trials (RCTs). This methodology is employed to measure the impact of various changes in a website's design or feature on user behavior. In an A/B test, the 'treatment', which may be a modification to the website's layout or features, is randomly assigned to a portion (typically 50%) of users (Group B). The remainder of the users (Group A) are exposed to the existing or original version of the website, thus acting as the control group.

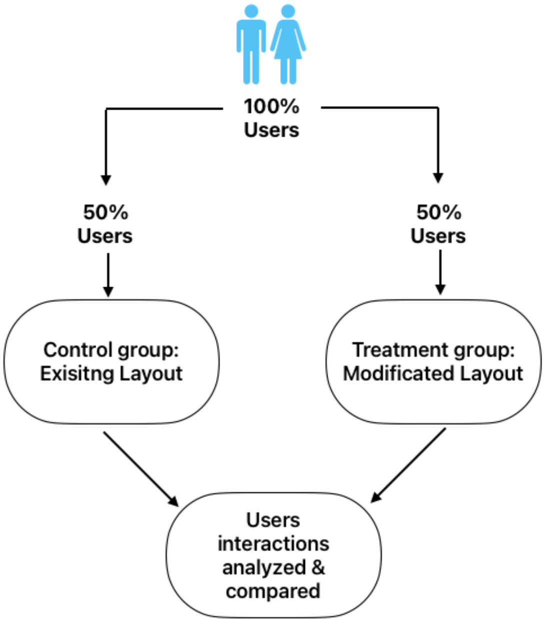


Figure 2: High level flow for A/B testing procedure, adapted from Kohavi, Deng, et al. 2012

This principle of random assignment substantially reduces the influence of confounding variables on the observed outcomes. Consequently, it allows for causal inferences to be drawn between the change implemented on the website and subsequent user behavior. In terms of user interactions, both groups' behaviors are analyzed and compared to discern any statistically significant differences that can be attributed to the implemented change (Kohavi, Deng, et al. 2012).

In mathematical terms, the experimental setup can be represented as $T \in \{0, 1\}^N$, where T_i denotes the status of the i^{th} user - if they belong to the control group or the treatment group. If $T_i = 0$, the i^{th} user is part of the control group and experiences the existing layout, while if $T_i = 1$, the user is part of the treatment group and experiences the modified layout (Imbens and Rubin 2015).

2.1.3 Limitations of A/B Testing

Despite the conceptual similarities between A/B testing and RCTs, there are crucial limitations in A/B testing (Kohavi, Deng, et al. 2012; Saxena 2020). It requires a large number of users to achieve statistical significance and, unlike in RCTs, the treatment cannot always be randomly assigned due to technical constraints or significant design changes, for instance introducing a new payment system applicable to all users at once. Furthermore, concurrent changes on the website can confound the effects, making it difficult to ascribe observed changes to a specific treatment.

2.1.4 Scenarios Restricting the Feasibility of Online Controlled Experiments

While online controlled experiments or A/B testing are powerful tools, certain scenarios can restrict their feasibility. In this study, the main reason why drawing conclusions from an online controlled experiment could not be applied is the issue with data collection. The web analytics team was pursuing to analyze historical changes to the website while suitable data was not available. The team was seeking new insights on changes already made to adjust their future decisions in website development, for the cases where A/B testing was not applied.

Analyzing other potential scenarios when A/B testing is not possible to implement we can distinguish:

A lack of control over the causal action could prevent conducting these experiments. For instance, if the action tested depends on users' personal choices, such as switching from one phone brand to another, the organization cannot control it.

The number of available units can also be a limiting factor. If too few units exist for analysis, running a controlled experiment can be problematic. A single-event occurrence, such as a merger or acquisition, exemplifies this as it complicates estimating the counterfactual.

High opportunity costs can also impede setting up a control group. This might occur when dealing with rare events or when the desired outcome requires a long observation period. Consider, for example, evaluating the impact of Super Bowl advertisements or measuring website revisits for a car purchase happening years after the current one.

Expensive changes relative to their perceived value may render A/B testing infeasible. An example would be forcibly signing users out after a certain time period or abstaining from displaying ads on a search engine.

In certain cases, the desired unit of randomization cannot be appropriately randomized. For instance, it's practically impossible to randomize TV ad viewership, resulting in fewer units for analysis, and consequently, lower statistical power.

Finally, ethical or legal limitations may restrict the scope of online controlled experiments. Withholding beneficial medical treatments for the sake of testing, for instance, would be unethical.

In such scenarios, alternatives like small-scale user experience studies, surveys, and observational studies often prove more practical for estimating effects, although these methods fall lower in the evidence hierarchy (Kohavi, Tang, and Y. Xu 2020).

2.2 Average Treatment Effects estimation

The Average Treatment Effect (ATE) estimates the impact of an intervention, averaged across a whole population. The ATE is used to compare the average outcome if everyone in the population was exposed to the treatment, versus the average outcome if no one were exposed. Its formula is:

$$ATE = E[Y(T = 1) - Y(T = 0)] \tag{1}$$

Here, $Y(t = 1)$ represents the potential outcome with the treatment (from the treatment group) and $Y(t = 0)$ the potential outcome without the treatment (from the control group). In the context of this study, the treatment refers to the changes in the website layout and payment system, and the outcome would be the orders made on the website. Estimating the ATE in an A/B testing setting under the condition of meeting all the requirements of the gold standard of Randomized Control Trial would provide a reliable estimation of ATEs. However, as described in Section 2.1.3, adhering to the RCT framework is not always possible and often requires a lot of resources. The aim of this study is to research the potential usage of causal models in observational studies where RCT principles cannot be applied.

Confoundedness In observational studies, drawing conclusions about the impact of treatments on output variables is a challenging task due to the problem of confounding variables that are related to both the treatment and output variables. Rosenbaum's study (Rosenbaum and Rubin 1983) suggests that bias in estimating the ATE can be minimized by conditioning on a well-adjusted set of covariates, presented in a vector x. This approach can be represented in a simplified way with a Directed Acyclic Graph (DAG) showing a Structured Causal Model (SCM) that adjusts ATEs estimation with the usage of a set of covariates denoted with X. Figure 3, the impact of covariates X on both treatment variable T and output variable Y is a confounding factor. It needs to be controlled if the effect of T on Y is to be identified. On the other hand, we see that at the same time, covariates and treatment variables have colliding relation towards output variable Y (Pearl, Glymour, and Jewell 2016).

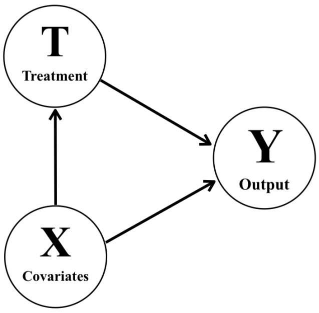


Figure 3: Directed Acyclic Graph of simplified Structural Causal Model

The formula for the ATE adjusted by these covariates is:

$$ATE(x) = E[Y(T = 1) - Y(T = 0)] \tag{2}$$

This formula estimates the ATE, given the covariates in x. The treatment effect is considered unconfounded with the outcome, conditional on these covariates. This adjustment can potentially remove bias from estimating average treatment effects.

It's crucial to highlight that our methodology fundamentally assumes the absence of unobserved confounders, meaning that, given the covariates, potential outcomes are expected to be independent of treatment assignment.

2.2.1 Methods for Average Treatment Effects estimation

Propensity Score Matching Propensity score matching is a widely used method for estimating ATE in observational studies. Introduced by Rosenbaum and Rubin 1983 this involves estimating the probability of receiving treatment based on observed covariates and then matching treated and control units with similar propensity scores. This method helps to balance the distribution of covariates between treated and control groups, reducing confounding bias and allowing for a more accurate estimation of ATE (Imbens and Rubin 2015). Different learning algorithms can be used for estimation of propensity scores, starting from logistic regression, and ensemble methods to advanced machine learning frameworks like artificial neural networks (J. Xu et al. 2022; Shi, Blei, and Veitch 2019).

Instrumental Variables Instrumental variables (IV) methods can be used to estimate treatment effects in the presence of unmeasured confounding, provided that the sample size is sufficiently large and the assumptions of the IV analysis are met (Angrist and Pischke 2009; Angrist and Pischke 2015). This involves approximation of random assignment with well adjusted set of instrumental variables that have a large impact on the treatment variable and output variable. Adjusting for measured instruments in the analysis can reduce the variability in the IV estimates. However, violating the IV assumptions that are very restrictive can lead to biased estimates, sometimes even more biased than those from naïve linear regression models (John et al. 2019). Most common method for the estimation of ATE with inclusion of instrumentals variables is two-stage least-squares regression that is not capable of handling complex relations.

Causal ML Techniques Causal ML techniques, such as meta-learners, direct uplift estimation, and tree-based algorithms, can be employed to estimate conditional average treatment effects (CATE) (Karmakar, Majumder, and Gangaraju 2023). These methods can capture heterogeneous treatment effects across different subgroups of users, allowing for better prescriptive recommendations and implementation strategies. One of the most famous examples for such estimation is Causal Forests (Imbens and Rubin 2015) whose application for observational data with clustered errors was described by Athey and Wager (2019).

In this context, Double Machine Learning (DML) has emerged as a promising approach for estimating average treatment effects in observational studies where traditional experimental methods are not feasible (Chernozhukov et al. 2018). DML combines machine learning techniques with causal inference to address the issue of confounding bias and provide consistent estimates of treatment effects (Karmakar, Majumder, and Gangaraju 2023). DML is particularly well-suited for high-dimensional settings, where the number of covariates is large, and the relationships between them are complex and non-linear. Moreover, DML is robust to model misspecification, which is a common issue in observational studies (Chernozhukov et al. 2018). This robustness is achieved by using cross-fitting, reducing the risk of overfitting and improving the accuracy of treatment effect estimates (ibid.) which is described in detail in section 4.5.3.

2.3 Research Gap

Over the last decade, a myriad of methods has been developed to effectively estimate average treatment effects in observational studies, addressing a high-dimensional set of confounding variables. Numerous research papers have focused on the use of these methods in simulated observational data studies (Chen and Au 2019; John et al. 2019; J. Xu et al. 2022). However, there seems to be a gap when it comes to implementing these methods on real-world datasets, specifically for website changes aiming to provide valuable insights for website optimization.

In this context, Double Machine Learning (DML) emerges as a valuable technique for estimating average treatment effects, especially for website performance optimization where traditional experimental methods might not be applicable. DML leverages machine learning techniques and addresses confounding bias, offering more accurate and consistent estimates of treatment effects. This makes DML a promising approach for this research area (Chernozhukov et al. 2018; Karmakar, Majumder, and Gangaraju 2023).

3 Data

3.1 Data Collection

3.1.1 Company data

The data collection for this study was conducted using three primary sources from Elasticsearch, a distributed, free, and open search and analytics engine. These sources provided comprehensive information on customer visits to the flysmarter.dk webpage through travel search engines, as well as the details of orders placed on the platform. The data sets were merged to create a holistic view of customer interaction with the flysmarter.dk website, from initial visitation to flight booking.

Clicks data The first data source is the "Clicks" table, which contains information on customer visits to the flysmarter.dk webpage through various travel search engines. This data set includes variables related to the following:

1. Visit details, such as timestamps, IP addresses, and mobile usage
2. Itinerary characteristics, such as marketing and operating carriers, currency, sales price, and segment count
3. Flight attributes, including travel time, baggage availability, and direct flight status
4. Results set positions and prices, including the cheapest, fastest, and best prices
5. Customer behavior metrics, such as weekday/weekend visits and passenger count

Orders Data The second data source is the "Orders" table, which captures information on orders placed on the flysmarter.dk webpage. This data set contains variables related to the following:

1. Order details, including order IDs, order types, and timestamps
2. Payment information, such as payment status, transaction details, and cancellation information
3. Customer behavior metrics, including IP addresses, mobile usage, and search parameters
4. Experiment-related variables, capturing information on various experiments conducted by the platform
5. Total price and addon information for the orders

The two tables were merged to provide a comprehensive view of customer interaction with the flysmarter.dk website, from initial visitation to flight booking. This integrated data set included multiple details about desired destinations, travel distances, travel times, customer characteristics, timestamps, and passenger information, which were used as confounding variables in the study of average treatment effects of changes on the website.

To distinguish data about the Danish market merged table was filtered using currency displayed to the user after clicking from the travel search engine taken from "Clicks" table.

Changes to order website The last source combined for the analysis was the list of changes to the flysmarter.dk website provided by AOB Travel. This list constitutes as set of changes evaluated in this study in the estimation of average treatment effects (ATEs). More detailed description of changes to the order website evaluated in this study can be found in Section 4.2.

The web developers at AOB Travel are continuously updating the website, specifically making modifications to the pricing rules and enhancing their proprietary search engine. This engine aggregates relevant flights for user queries from various travel search engines, including Momondo and Skyscanner, and then presents these options to the end user.

These changes directly influence the number of visits to the flysmarter.dk website. However, due to the high frequency of these adjustments, they aren't systematically tracked by the team.

While these changes do increase the number of visitors to the website, they don't necessarily have a direct influence on the conversion rate once the user has been redirected to the flysmarter.dk webpage. As the primary goal of this study is to analyze factors affecting the conversion rate, these particular modifications were not included in the analysis.

3.1.2 External Data Sources

Google Trends Data Google Trends data provides valuable insights into the relative search popularity of searched keywords. The data measures the search volume for each keyword as an index, with values relative to the maximum point of interest for a particular search term within the Google search engine during the analyzed period. This information allows us to track fluctuations in demand and identify patterns or trends that may impact the effectiveness of changes made to the website of online travel agency (Trends 2023).

To analyze the interest and demand for flight tickets over time in the Danish market, weekly data on the popularity of search queries was extracted from Google Trends for the entire modelling period. The following Danish keywords were selected as the most relevant and popular search terms for flight tickets: ”flybilletter” (flight tickets), ”momondo” (most popular travel search engine in Denmark), ”direkte fly” (direct flights), ”skyscanner” (travel search engine), and ”flybilletter billige” (cheap flight tickets). These keywords represent common search queries that potential customers use when looking for flight tickets and related services in the Danish market.

Skytrax Airlines Rating The ”Clicks” table contained information on IATA airline codes for each airline associated with every flight in an itinerary. These codes were matched with their respective airline names using the list of airline codes available on Wikipedia (Wikipedia 2023).

To assess the quality of carriers in the itineraries, airline ratings were obtained from Skytrax. Skytrax is a UK-based consultancy which conducts research for commercial airlines and audits airline standards. It is known for its annual World Airline Awards and Airline Star Ratings, which provide comprehensive ratings and rankings of airlines worldwide (Skytrax 2023).

Airports Coordinates For each observation in the ”Clicks” table, there was information about the origin and destination airports for each flight within the itinerary. To calculate the distances of the itineraries, it was necessary to use the geographical coordinates of the airports, which were collected from the IP2Location geolocation database. This database covers both IATA and ICAO airport codes and includes the airport name, latitude, and longitude (IP2Location 2018). Using the GeoPy package, distances were calculated for each flight to obtain the travel distance metric (Esmukov 2022).

Exchange Rates In order to standardize the sales prices in the ”Clicks” table, all currencies were converted to Swedish Crowns (SEK). This conversion was necessary because AOB is a Swedish travel agency and having all the prices in the local currency facilitates the analysis. To perform this conversion, exchange rates were obtained from the Free Currency API. This API provides real-time exchange rates for various currencies, allowing for accurate currency conversions (FreecurrencyAPI 2023).

3.2 Data Preprocessing

This subsection provides a comprehensive overview of the steps undertaken in data cleaning, preprocessing, and feature engineering. These are fundamental procedures that ensure the reliability, relevance, and validity of the data, addressing potential confounding factors, and improving the interpretability of the study’s findings.

Data Merging: The **Clicks** and **Orders** tables were filtered using a relevant currency column, after which they were merged through a left join operation. The Orders table was merged into the Clicks table using the *clicks_id* column.

Travel Distance Calculation: For every observation in the merged dataset, distances for each flight were calculated using data from the Airports coordinates table. Total travel distance is a valuable feature for customer segmentation, differentiating between types of travel based on the travelling distance.

Currency Conversion: All sales price values were converted to Swedish crowns (SEK) using data from the FreeCurrency API. As AOB is a Swedish travel agency, normalizing all prices to the local currency simplifies the analysis.

Airlines Ratings: Using the Skytrax Airlines Rating, the minimum, average, and maximum ratings of the airlines involved in the flights of each itinerary were computed. This approach offers a comprehensive evaluation of the airline quality within each itinerary, enabling a more profound understanding of customer choice and the impact of payment process changes.

Google Trends Data: Google Trends search popularity index data were added to the merged table based on the date of each click. Given the weekly granularity of Google Trends data, values from the previous week were also included.

3.3 Feature engineering

In the subsequent feature engineering stage, additional features were created to better characterize customers and their behaviors. Such features enhance the base learners' ability to estimate Average Treatment Effects (ATEs) and provide meaningful business insights.

Ratio-based features These include the ratio-based features calculated using relevant combinations of the following variables: sales price, number of passengers in order, travel time, travel distance, and carrier's quality rating. All of the ratio-based features were described in Table **??**.

Counting variables To better capture the volatility of website traffic on the ordering page *clicks_count* measure was introduced, it measures the number of visits on the ordering website with the number of clicks directed from the travel search engine to the ordering website. To better measure the complexity of travel that a customer is viewing *clicks_segment_count* number of segments (unique number of combinations of origin and destination airports in the travel) and *carriers_marketing_ratings_count* representing the number of unique airlines within travel.

Price option variables Furthermore, based on the three price options typically displayed on a travel search engine (cheapest, best, and fastest), binary and continuous variables were created to identify the option chosen by the customer and the price difference between the chosen option and the other available options.

Timestamp variables New features were extracted based on the precise timestamp of the click event on the AOB Travel website, including the date, weekday, and hour of the day. To use these variables in the model One Hot Encoding was used for *clicks_created_at_datetime_weekday* to obtain binary variables for each day of the week. For *clicks_created_at_datetime_hour* variables were grouped by time intervals by hours 0:00-6:00, 7:00-12:00, 13:00-18:00, 19:00-24:00 identifying hours when the customer was looking for travel options.

3.4 Feature selection

The final stage involved the selection of variables, with the final set guided by their relevance to the study, their ability to address confounding issues, and their potential to enhance our understanding of customer behaviors and the effects of different treatments.

To make sure that any of the variables do not repeat any important information variable selection process was guided using correlation matrix and basic Random forest regression model regressing potential covariates X_i on output variable Y.

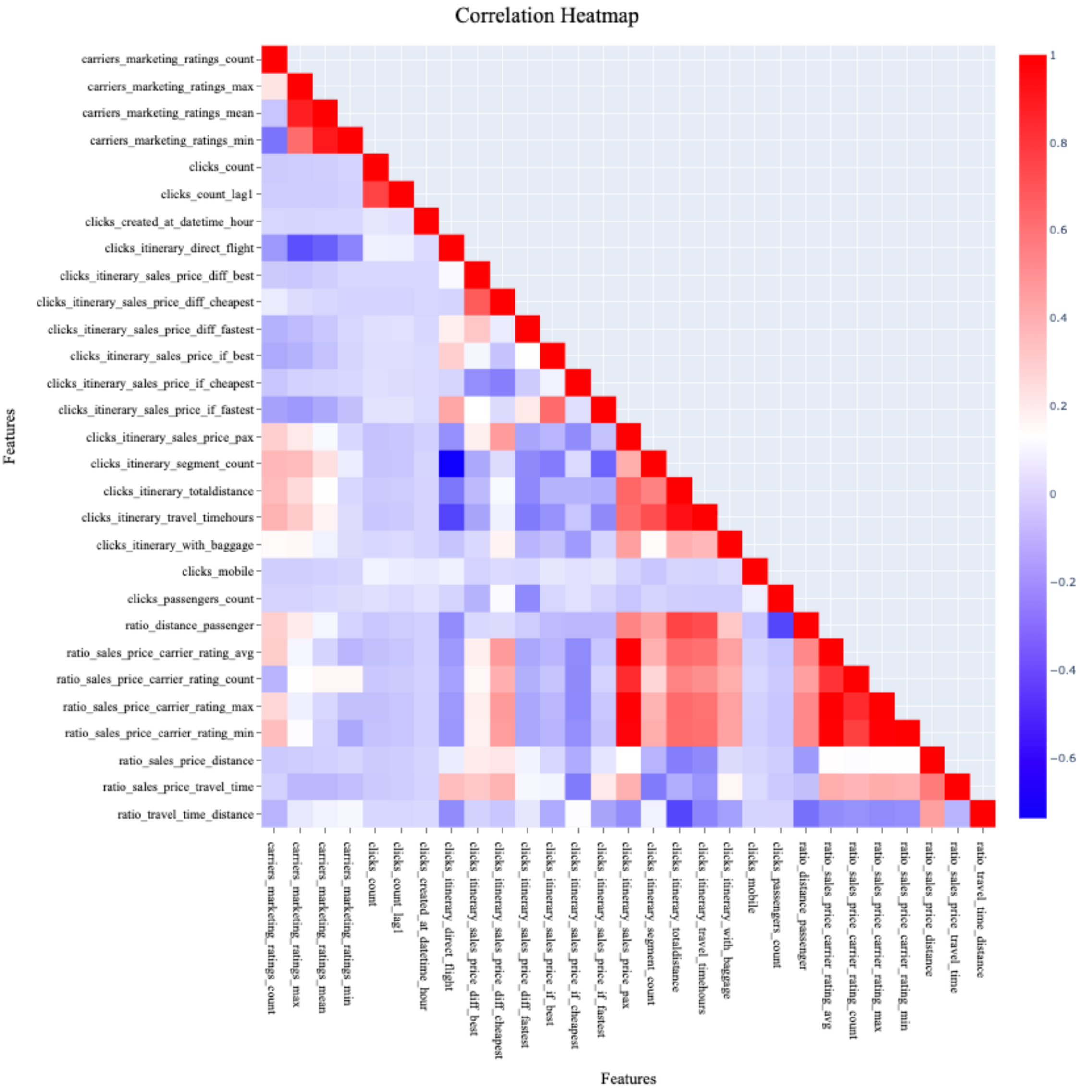


Figure 4: Correlation map of features used as covariates for each treatment generated on full dataset

All variables measuring seasonality with Google Trends data were excluded from the analysis due to the weekly granularity of the data. These variables were not capturing changes with good enough dynamics, instead a proxy variable of ’clicks_count’ was used as a number of clicks in the day reflecting the volume of daily traffic on the website. Also ’clicks_created_at_datetime_weekday’ and resulting binary variables as some days of the week were not present in both control and treatment groups affecting estimation of average treatment effects.

In Figure 4 final set of confounding variables was shown with standard Pearson’s correlation coefficients. Figure 5 shows that Random forest was the model for base learners, a relevancy of this method for this setting will be extensively described in 4.6

From the basic Random forest model relevant frequency metrics were extracted and they are shown in Figure 5. The frequency metric refers to the number of times a feature is used to split the data across all trees. We can see that variables related to price, travel characteristics and the ratio between them play a key role in mitigating the problem of confoundedness

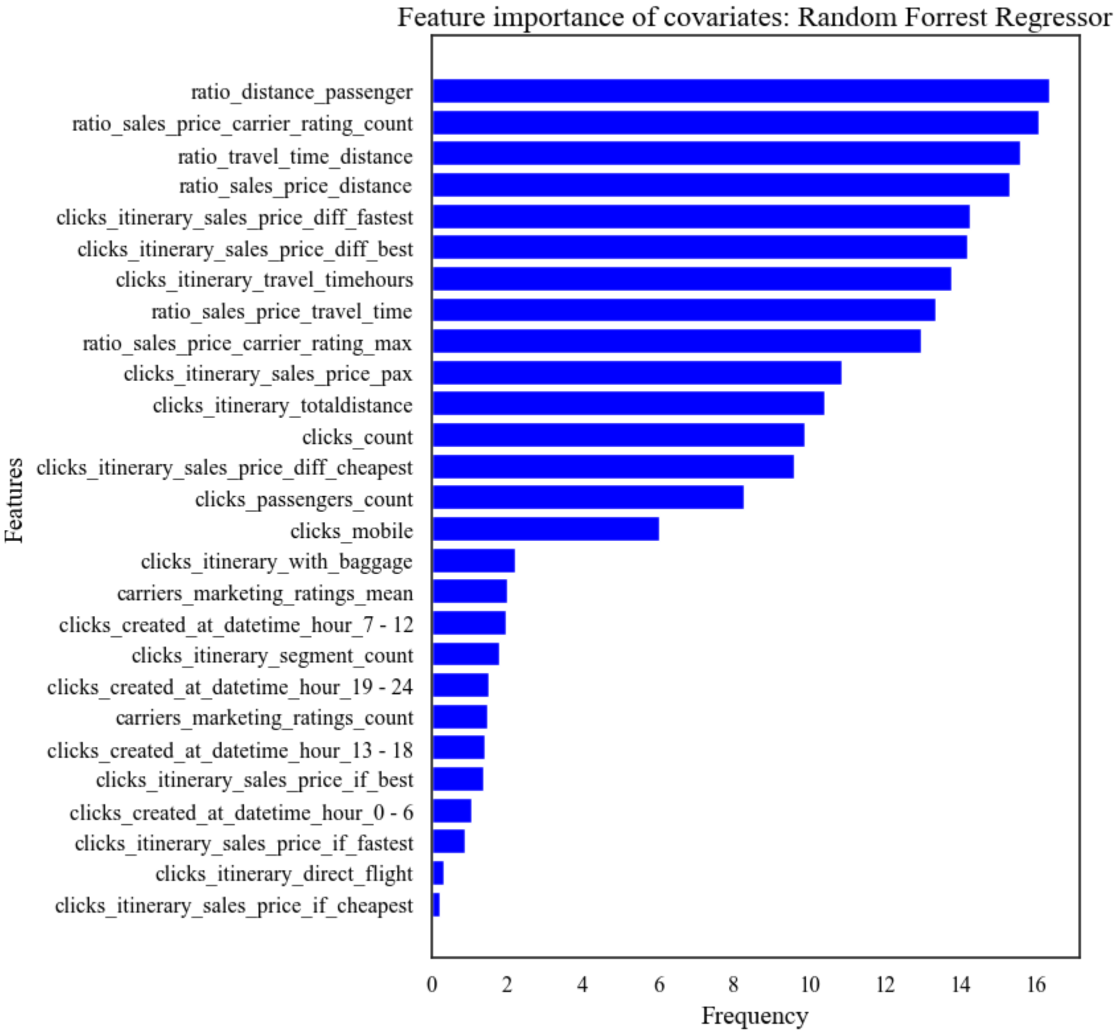


Figure 5: Plot of frequencies from the trained Random Forest model. The features along the y-axis represent the different predictors used in the model. The frequency metric on the x-axis represents the importance of each feature. Features with higher scores have more impact on the model's decisions. The features we consider confounders are clearly labeled.

Final dataset description The final dataset consisted of 39 variables in total. All the variables used as confounding variables were listed in the Appendix in Figure A.1

3.5 Data Splitting and Transformation

In the course of this study, a unique data partitioning approach was adopted to accommodate each treatment variable, including placebo effects. Separate datasets were generated for each treatment variable, adhering to the symmetrical time window as indicated in Figure 7. For each treatment variable, this methodology produced a unique dataset encapsulating observations from a specific time interval before and after the associated change.

The observations collected before the changes to the website were designated as the control group, whereas observations from a symmetrical time window after each change formed the treatment group. This procedure was replicated for each change, yielding distinct datasets for each.

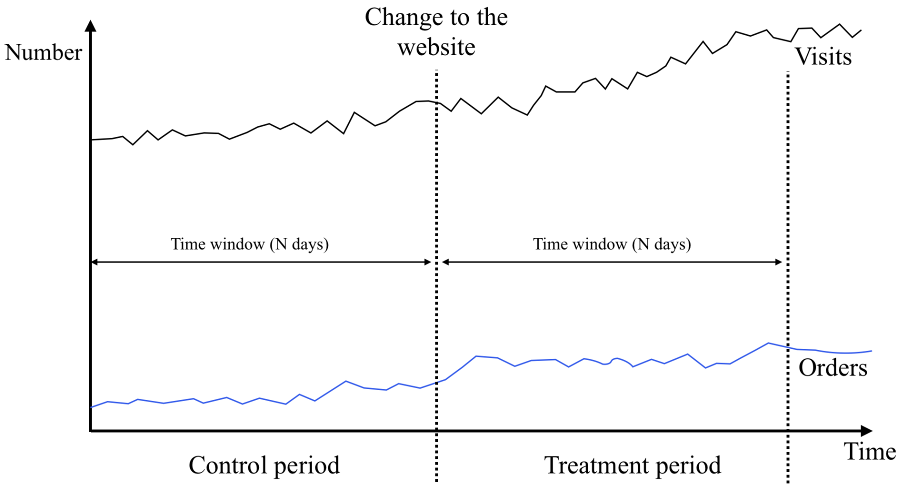


Figure 6: Diagram showing how assignment to control and treatment groups was conducted. For each treatment evaluated in the study appropriate time window in days was calculated. From two time windows' lengths $T_1 - T_0$ and $T_2 - T_1$ the lower value constitutes as a symmetrical time window's length. For a given time window control period and treatment period were assigned ensuring the lack of overlap between periods analyzed for each treatment.

The Quantile transformation with a number of quantiles set to 1000 was employed on all numerical continuous variables in the dataset to ensure that all variables have comparable mean and variance. This transformation is crucial as it adjusts the data to follow a normal distribution, enhancing the predictive power of many machine learning algorithms. The transformation was fitted only on the training data to prevent any information leakage from the test data, thereby preserving the integrity of the evaluation. The parameters derived from the quantile transformation were subsequently applied to adjust the corresponding test datasets. This process ensures consistency between the training and test datasets, enabling a more accurate assessment of the model's performance.

An identical methodology was adopted for placebo effects. During time windows without any significant changes, hypothetical placebo changes were imputed. This provided a basis for comparison and helped evaluate the models' efficacy in discerning the actual effects of changes. Thus, this process of data splitting and transformation not only enriches the dataset but also facilitates a robust evaluation of the model's performance in interpreting and predicting treatment effects.

4 Methodology

4.1 ATE estimation using conversion rates

In order to measure the performance of the order website there is a necessity to provide a relevant metric that is possible to calculate in the short term, possible to extract for both control and treatment groups and sensitive to changes in time (Kohavi, Tang, and Y. Xu 2020). The metric that is used in this study and meets all the criteria listed before is conversion rate. The conversion rate (CR) can be calculated using the formula:

$$CR = \frac{\text{Number of orders}}{\text{Number of web visits}} \tag{3}$$

Let Y_i be a random variable indicating whether the i-th web visit results in an order (with $Y_i = 1$ for an order and $Y_i = 0$ otherwise) and let n denote the total number of web visits in your sample. The sample average of Y, denoted by $\hat{E}(Y)$, estimates the expectation of Y in the population and is calculated as:

$$\hat{E}(Y) = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{4}$$

The empirical estimate of E(Y), denoted as $\hat{E}(Y)$, is the sample average of Y, which equals the conversion rate. This $\hat{E}(Y)$ equals the conversion rate (CR), which is the ratio of the number of orders to the number of web visits in your sample:

$$CR = \frac{\text{Number of orders}}{\text{Number of web visits}} = \hat{E}(Y) \tag{5}$$

Let's underline that $\hat{E}(Y)$ is an estimate of the true conversion rate in the population, which is unknown and is represented by $E(Y)$.

Conversion rates usually take values in the range not exceeding 4.2% (Alexander 2023). However, due to special position of flysmarter.dk platform in the process, where customers already compared prices for the particular travel between other OTAs on a travel search engine and clicked the direct link to flysmarter.dk website these conversion rates are exceeding this value substantially. conversion coefficients represent crucial information for business operations, and sharing this data with competitors may pose a risk to AOB Travel. For this reason, the conversion coefficients demonstrated in this study have been indexed.

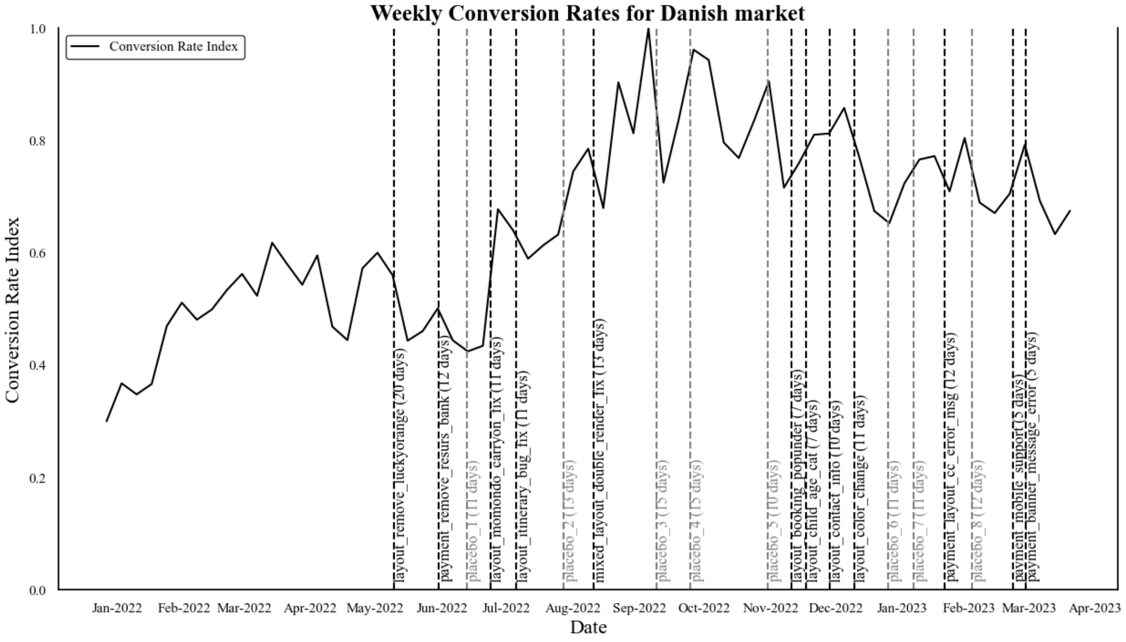


Figure 7: Aggregated weekly conversion rates index for the Danish market with dashed lines analyzed changes to the website were shown together with analyzed placebo effects evaluated

The direct connection between conversion rate and output predicted output variable was demonstrated in the equations below. Additional control for confounding variables allows us to draw more reliable conclusions about ATE estimates.

$$A\hat{T}E = \hat{E}[CR_{treatment} - CR_{control}] \tag{6}$$

$$\tag{7}$$

4.2 Treatments description

Table 2 outlines a series of changes implemented in the ordering web page of flysmarter.dk, evaluated in this study. Starting from May these modifications span across different categories, primarily focused on layout and payment features. The layout alterations range from subtle changes such as color adjustments, modifications to child age category search parameters, and fixes to existing bugs, to more noticeable ones such as the removal of third-party components (Luckyorange and Google maps) for faster page loading, and extension of company contact information on the booking step. Similarly, in the payment category, changes revolved around enhancing user experience and extending payment options. Key updates include the removal of specific payment options, more explicit error messaging for failed credit card transactions, the introduction of mobile payments support, and displaying supported payment methods on the Danish site. Overall, these changes were strategically implemented to boost site performance and user experience, in turn influencing conversion rates.

Abbreviated name	Category	Date	Description
layout_remove_luckyorange	layout	09/05/2022 13:18	Remove Luckyorange visualizations and Google Maps to make page faster
payment_remove_resurs_bank	payment	30/05/2022 10:35	Remove Resurs Bank payment option
layout_momondo_carryon_fix	layout	23/06/2022 16:27	Fix Momondo carryon count
layout_itinerary_bug_fix	layout	05/07/2022 15:25	Fix bug hiding itinerary on step 4
mixed_layout_double_render_fix	mixed	10/08/2022 07:23	Fix double render issue
layout_booking_popunder	layout	10/11/2022 10:16	Booking.com popunder window/tab
layout_child_age_cat	layout	17/11/2022 10:48	Child search changed from using age number to age categories (infant / child)
layout_contact_info	layout	28/11/2022 08:11	Extended company contact information on booking step
layout_color_change	layout	09/12/2022 15:10	Small color change
payment_layout_cc_error_msg	payment	20/01/2023 14:23	More specific error messages for failing credit card payments (better explanation why payment failed)
payment_mobile_support	payment	21/02/2023 11:31	Introduction of mobile payments support
payment_banner_message_error	payment	27/02/2023 10:48	Banner on Danish site displaying supported payment methods

Table 1: List of changes made to the website in analyzed period

4.3 Placebo changes

To better understand how well the models for Average Treatment Effect (ATE) work, a new method was used: null or "placebo" changes were added. These changes, based on an idea from Rosenbaum's work on hypothesis testing (Rosenbaum 2020), are fictitious changes that are treated as real ones. These changes were added to data in between the actual changes ensuring that time periods do not affect actual changes to the website evaluated in this study.

The idea behind this study is straightforward: if it is shown in this paper that these fake changes have no effect, it means that the models used in this study are effective in identifying real ATEs. These placebo changes were carefully introduced during instances when significant and minor disparities were observed in the binary output of the control and treatment groups.

By employing this new method, the process of uncovering cause-and-effect relationships is strengthened, providing a more comprehensive evaluation of the model's performance. In future analyses, a comparison will be made between the model's outcomes and the ATEs resulting from actual changes, taking into consideration the magnitude of differences between the control and treatment groups.

Table 2: List of placebo changes evaluated in the modelling process

Abbreviated name	Date	Description	Time window
placebo_1	12/06/2022 09:00	Not applicable	11
placebo_2	27/07/2022 13:00	Not applicable	13
placebo_3	08/09/2022 16:00	Not applicable	15
placebo_4	24/09/2022 10:00	Not applicable	15
placebo_5	30/10/2022 15:00	Not applicable	10
placebo_6	25/12/2022 14:00	Not applicable	11
placebo_7	06/01/2023 09:00	Not applicable	11
placebo_8	02/02/2023 14:00	Not applicable	12

4.4 Structural Causal Model

To enhance the comprehension of the study's setup and the types of relationships that can be discovered in the data used, the researchers introduced a Structural Causal Model (SCM) (Pearl, Glymour, and Jewell 2016). Within the SCM, several factors influencing the likelihood of customers making positive outputs (i.e., flight ticket orders) can be distinguished. The variables described as S, D, C, P, and F, located in the lower left part of the graph, represent groups of confounding variables. Detailed descriptions of these variables, along with their relevant categories, can be found in Table A.1 in the appendix.

The primary objective of this study is to estimate the average treatment effects, as indicated by the dark blue arrow. The main goal is to identify the presence and magnitude of these effects, which can influence a higher probability of flight ticket orders on the website and, in turn, lead to an increased conversion rate (i.e., the ratio of orders to the number of visits on the ordering page).

Furthermore, the relationships between the set of covariates represented by the symbols S, D, C, P, and F directly relate to the output variable Y, as depicted by the feature importances in Figure 5. This observation is supported by the strong predictive power of the final model used for the feature selection procedure, as described in Section 3.4.

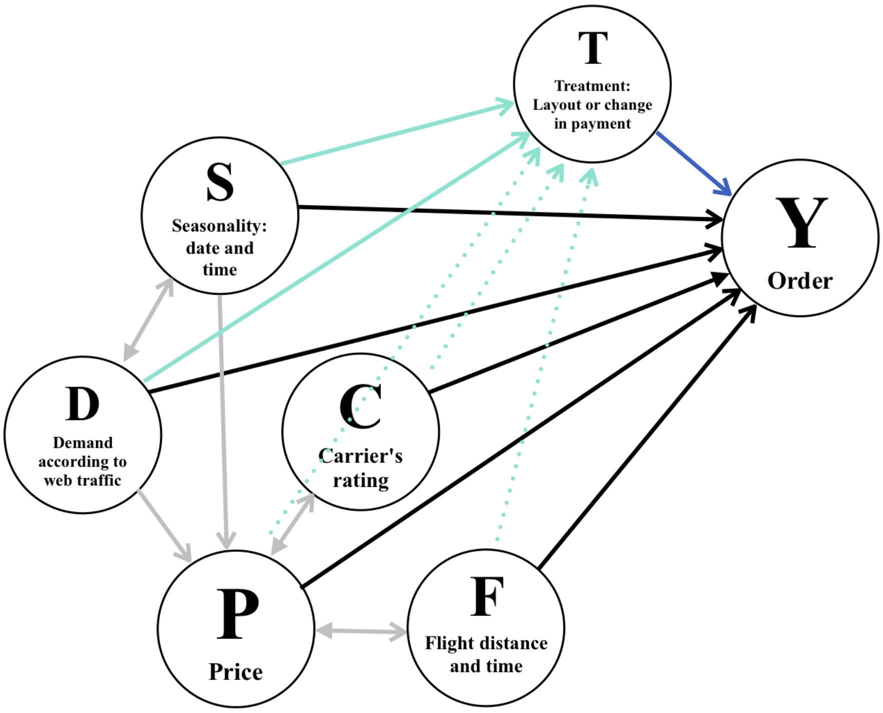

T
Treatment: Layout or change in payment
S
Seasonality: date and time
Y
Order
D
Demand according to web traffic
C
Carrier's rating
P
Price
F
Flight distance and time

Figure 8: Directed Acyclic Graph of Structural Causal Model

The treatment variable is closely associated with time intervals corresponding to changes made on the website (where the control group consists of visitors before the change and the treatment group includes those after the change). This impacts the set of confounding variables used in this study. Changes in web traffic volume on the website directly impact exposure to treatment. At times with higher demand for flight tickets, customers can behave differently; they can possibly be less vulnerable to changes in the page layout or, conversely, they can be more likely to change travel agencies to buy flight tickets for certain travel. On some days of the week or times of the day, traffic differs considerably, and also the characteristics of customers that are looking for tickets and visiting the website can be different, so the response to changes to the website may also differ.

Because possible time windows differ considerably between different treatments tested - from 5 to 20 days - there was limited scope for variable selection. Variables associated with time, such as days of the week, binary variables for weekend days, and Google Trends data describing demand (only available in weekly granularity), included values present only in one of the treatment and control groups. This implies that the volume of subjects between the control and treatment groups was very small. To resolve this issue for seasonality (S), only binary variables for the time of the day were used, and for the demand metric (D), the number of visits in the day was used as a proxy variable for demand and patterns in days of the week as well.

The remaining confounding variable categories describe travel characteristics in detail. Multiple metrics describing price (and chosen option on the meta site), travel time and distance, carrier's rating, a device used, and travel with luggage or not, were used as they were found to be relevant for possible confoundedness and potentially different responses to the changes on the website.

The objective of this study is to separate the effects of variables S, D, C, P, and F on the output variable Y from the potential effects of the treatment. This setup aims to minimize the probability of omitting relevant variables given the available data. However, it should be noted that certain additional factors were not taken into account due to problems in data collection.

4.5 Models for ATE Estimation

In the pursuit of unbiased Average Treatment Effects estimation, several advanced methodologies have emerged that attempt to mitigate the confounding factors often present in observational studies. Among these, simple comparing differences in conversion rate, Linear Regression, and Double Machine Learning (DML) with Partially Linear Regression have shown considerable promise. Random Forest, an ensemble learning method, is utilized as the primary learning algorithm in DML.

4.5.1 Comparing differences

A standard business scenario that could be applied is to compare the average conversion rates before and after a change within a symmetrical time window. This approach is the most accessible because simply comparing differences does not require any additional work other than computing the ratio between the number of visits to the website and the number of orders within the time period. In this method, the web analyst does not account for covariates that could have an impact on the results and the potential effect of the change. However, this method may not be ideal for estimating the Average Treatment Effect (ATE). While it may provide a straightforward comparison of pre- and post-change conversion rates, it does not account for potential confounding variables. The lack of control for these covariates can lead to biased estimations of the treatment effect.

4.5.2 Linear Regression

Linear regression can be utilized to estimate Average Treatment Effects (ATE) when there's an additional treatment variable alongside a set of covariates. This approach is often utilized as a baseline method to compare their effectiveness versus other more robust modelling methods (for example (John et al. 2019)). Consider a setting where we have a treatment variable $T_i \in \{0, 1\}$, a set of covariates $X_i, i \in \{1, 2, ..., n\}$ and an outcome variable Y for each individual i in the dataset.

The potential outcomes model can be represented as follows:

$$Y(T) = \alpha + \theta_0 T + \beta_1 X_1 + ... + \beta_n X_n + \epsilon_i, \tag{8}$$

where $Y(T)$ denotes the potential outcome of individual i when the treatment variable is set to T, α is the intercept, θ_0 is the coefficient of the treatment variable, β_2 represents the vector of coefficients associated with the covariates, and ϵ_i is the error term.

In this model, the ATE is given by the parameter θ_0, which measures the expected change in the outcome Y when the treatment T is changed from 0 to 1, holding the covariates X constant.

This method gives us the estimated ATE under the assumption of unconfoundedness: given the covariates, the potential outcomes are independent of treatment assignment. This strong assumption that was made for this method is not met as relevant reasoning will be provided in Section 4.4.

4.5.3 Double Machine Learning

Double Machine Learning (DML) is an approach that combines machine learning techniques with statistical theory to estimate structural parameters such as average treatment effects (ATE), while handling high-dimensional nuisance parameters. It uses machine learning methods to estimate these nuisance parameters and applies a debiasing step to mitigate regularization bias and overfitting.

Double Machine Learning with Partially Linear Regression The power of DML is truly realized when it is applied to a Partially Linear Regression (PLR) model. Consider the following PLR model:

$$Y = T\theta_0 + g_0(X) + U, \qquad E[U \mid X, T] = 0, \tag{9}$$

$$T = m_0(X) + V, \qquad E[V \mid X] = 0. \tag{10}$$

The PLR model represents the relationship between the outcome variable Y, the treatment variable T, and the set of covariates X. θ_0 is the parameter of interest, the causal effect or treatment effect, and $g_0(X)$ and $m_0(X)$ are nuisance functions. A naive approach to estimate θ_0 would be to construct a machine learning estimator $\hat{\theta_0}$ using the regression function $T\theta_0 + g_0(X)$.

Procedure Suppose we split the sample into two parts, a main part I_1 and an auxiliary part I_2. We obtain an estimator for $\hat{\theta_0}$ with following algorithm adapted from Reese 2022:

1. Train arbitrary machine learning algorithm using I_2 to derive $Y_i = \hat{g_0}(X_i)$ - Y task
2. Train arbitrary machine learning algorithm using I_2 to derive $T_i = \hat{m_0}(X_i)$ - β task
3. Obtain prediciton errors $\hat{u_i}$ and $\hat{v_i}$ with following equations:

$$\hat{u_i} = Y_i - \hat{g_0}(X_i), \hat{v_i} = T_i - \hat{m_0}(X_i) \tag{11}$$

4. Obtain estimate of θ_0 on main part of the data I_1 with:

$$\hat{\theta_0} = \left(\sum_{i \in I_1} \hat{v_i}^2\right)^{-1} \left(\sum_{i \in I_1} \hat{v_i}\hat{u_i}\right) \tag{12}$$

In this study described procedure was conducted using 5 splits of the data for each treatments separately. This enabled the derivation of more robust estimates of θ_0. By employing multiple splits, we tackle the advantage of cross-validation, enhancing the generalizability of our estimates and reducing the risk of over-fitting. This approach ultimately strengthens the reliability and validity of the DML modelling outcomes.

This estimator, $\hat{\theta_0}$, will generally have a slower than $1/\sqrt{n}$ rate of convergence due to bias in learning g_0. Chernozhukov et al. overcome regularization biases using orthogonalization. Specifically, they partial out the effect of X from T to obtain the orthogonalized regressor $v = T - m_0(X_i)$, where $\hat{V} = T - \hat{m_0}(X)$, with $\hat{m_0}$ being a machine learning estimator of m_0 obtained using the auxiliary sample. This step removes the dependence of the treatment variable on controls, thus mitigating regularization bias.

Advantages Through these steps, DML in a PLR model allows the estimation of the ATE parameter θ_0 even in the presence of high-dimensional nuisance parameters, making it a robust method for causal inference in complex settings (Chernozhukov et al. 2018). Another important advantage of DML implementation is that it does not require specification of how X_i affects T_i and Y_i respectively (Reese 2022). This allows a much simpler modelling process and allows researchers to take advantage of machine learning and capture complex relations between variables that are difficult to observe. A researcher's limited abilities to digest complex causal inference tasks can be mitigated with machine learning. These limited abilities are very often referred to using Simpson's Paradox (Pearl, Glymour, and Jewell 2016), such a problem can be resolved with the usage of the complex dataset along with a large number of samples and machine learning methods to capture sophisticated relations within the data.

The behavior of $\hat{\theta}_0$ as a coefficient in the Ordinary Least Squares (LS) regression, despite the use of machine learning (ML) methods, provides several additional advantages. This characteristic ensures that estimates of $\hat{\theta}_0$ are characterized by reduced bias and consistency, providing more confidence in the estimates obtained. Moreover, it also means that the interpretation of $\hat{\theta}_0$ remains straightforward, similar to interpreting coefficients in a linear regression model. This simplicity in interpretation is often lost in complex ML models. Furthermore, traditional statistical inference techniques such as hypothesis tests and confidence intervals can be applied directly to $\hat{\theta}_0$, making the double machine learning framework more flexible and easier to integrate into classic statistical analysis.

4.6 Ensemble methods as base learners

Random Forest, a machine learning algorithm introduced by Breiman (Breiman 2001), is employed as the base learner in our study. The Random Forest algorithm is employed in both its regressor and classifier forms, depending on the setting of the modelling technique.

Random Forest is an ensemble method that constructs multiple decision trees and combines their outputs. For classification, it uses majority voting, while for regression, it averages the predictions of individual trees. Both were used as base learners in the Double Machine Learning approach. Random forest classifier was implemented for β tasks and Random Forest Regressor for Y tasks. In essence, a Random Forest model can be represented by:

$$RF(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{13}$$

where $RF(x)$ is the Random Forest prediction for an input x, $T_b(x)$ represents the prediction of the b-th tree, and B is the total number of trees in the forest.

Each tree $T_b(x)$ is independently grown using a bootstrapped sample of the data and uses random feature selection for each node split to ensure diversity among the trees. This randomness helps in robust predictive performance and mitigates overfitting.

For a more detailed mathematical explanation and intuition behind Random Forest, it is recommended to refer to the original study by Breiman.

5 Results and Discussion

5.1 Comparison of Models' Results

In the following analysis, the results from three models: Difference in Means, Ordinary Least Squares (OLS) Linear Regression, and Double Machine Learning Partially Linear Regression (DML PLR) are compared for actual and placebo treatments together.

	Differences		OLS		DML PLR			
Treatment name	ATE (p.p)	Holm's p-value	ATE (p.p)	Holm's p-value	ATE (p.p)	Holm's p-value	N	Time window
layout_remove_luckyorange	-0.670	0.244	-0.630	0.366	-0.460	1.000	23 990	20
payment_remove_resurs_bank	0.180	1.000	0.220	1.000	0.110	1.000	14 022	12
layout_momondo_carryon_fix	2.050	0.000 ***	2.200	0.000 ***	2.100	0.001 ***	10 198	11
layout_itinerary_bug_fix	0.540	1.000	0.440	1.000	0.540	1.000	9 750	11
mixed_layout_double_render_fix	0.160	1.000	0.300	1.000	0.500	1.000	22 776	13
layout_booking_popunder	-1.120	0.387	-0.860	0.942	-1.650	0.513	9 994	7
layout_child_age_cat	0.740	1.000	0.650	1.000	0.690	1.000	9 511	7
layout_contact_info	0.730	1.000	0.600	1.000	0.600	1.000	13 239	10
layout_color_change	-1.140	0.188	-1.320	0.107	-0.880	0.938	13 860	11
payment_layout_cc_error_msg	0.200	1.000	0.670	0.366	0.860	0.125	28 359	12
payment_mobile_support	0.400	1.000	0.260	1.000	0.210	1.000	8 045	5
payment_banner_message_error	0.470	1.000	0.960	1.000	1.610	0.961	7 574	5

Table 3: Estimated ATEs (in percentage points) and Holm's adjusted p-values for each treatment, where . if p<0.1, * if p<0.05, ** if p<0.01, *** if p<0.001.

Table 4 demonstrates a comparison of all changes made to the flysmarter.dk website during last year. Estimated average treatment effects were shown in percentage points to enhance the readability of results. An exemplary interpretation of ATE for 'layout_momondo_carryon_fix' would be that on average this change lifted the conversion rate by 2.1 p.p. (e.g. from 7.0% to 9.1%) if all other variables were constant. This implies an overall high increase on average in the probability of conversion during the visit on the website after the implementation of this change.

Holm's procedure was used to correct the problem of multiple comparisons. When multiple comparisons are made, the probability of making at least one Type I error (incorrectly rejecting a true null hypothesis) increases. Holm's procedure is a method to control the Family-Wise Error Rate (FWER), which is the probability of making at least one Type I error (Holm 1979).

From the table, we can make the following specific observations:

1. The 'layout_momondo_carryon_fix' treatment shows a significant positive Average Treatment Effect (ATE) across all three models. This suggests that the treatment has had a significant positive impact on the conversion rate at the 0.001 level.

2. Conversely, the 'layout_remove_luckyorange' treatment shows a negative ATE in the Difference in Means and OLS models, though it is not significant. The DML PLR model doesn't indicate a significant ATE. This could suggest that this treatment may have had a negative impact on the conversion rate, although the evidence is inconclusive when considering the DML PLR model.

3. Treatments such as 'payment_remove_resurs_bank', 'layout_itinerary_bug_fix','mixed_layout_double_render_fix', 'layout_booking_popunder', 'layout_child_age_cat','layout_contact_info', 'layout_color_change', 'payment_mobile_support', and 'payment_banner_message_error' do not show a significant ATE in any of the models.This suggests that these treatments did not have a noticeable impact on the conversion rate.

4. The 'payment_layout_cc_error_msg' treatment indicates a positive ATE in the DML PLR model but the coefficient is not significant on any typical level, we can draw similar conclusions for Difference in Means and OLS models do not show significant ATE. This points out that the treatment shows a higher estimate of ATE when using the DML PLR model, but it is not significant.

This table underlines the importance of accounting for confounding factors when estimating average treatment effects, as evidenced by the varying significance levels across the three models.

To sum up, it is evident from the analysis that the estimated ATE and its significance vary between these three techniques for the same treatments. However, the overall conclusion about a set of changes to the website that had a significant impact on conversion is identical for all the models compared.

5.1.1 Including placebo treatments

Models that consider covariates (OLS and Double ML with PLR) provide a more robust approach for distinguishing the impact of actual changes versus placebo changes, which are expected to show no impact on the outcome variable. Interestingly, the DML PLR model indicates that none of the placebo treatments significantly impact the conversion rate at the 0.05 level of significance. This suggests that the Double Machine Learning procedure, which includes additional control over confounding variables, offers a more reliable approach for estimating average treatment effects.

Treatment name	Differences		OLS		DML PLR		N	Time window
	ATE (p.p)	Holm's p-value	ATE (p.p)	Holm's p-value	ATE (p.p)	Holm's p-value		
layout_remove_luckyorange	-0.670	0.366	-0.630	0.659	-0.460	1.000	23 990	20
payment_remove_resurs_bank	0.180	1.000	0.220	1.000	0.110	1.000	14 022	12
layout_momondo_carryon_fix	2.050	0.000 ***	2.200	0.000 ***	2.100	0.001 **	10 198	11
layout_itinerary_bug_fix	0.540	1.000	0.440	1.000	0.540	1.000	9 750	11
mixed_layout_double_render_fix	0.160	1.000	0.300	1.000	0.500	1.000	22 776	13
layout_booking_popunder	-1.120	0.602	-0.860	1.000	-1.650	0.924	9 994	7
layout_child_age_cat	0.740	1.000	0.650	1.000	0.690	1.000	9 511	7
layout_contact_info	0.730	1.000	0.600	1.000	0.600	1.000	13 239	10
layout_color_change	-1.140	0.273	-1.320	0.185	-0.880	1.000	13 860	11
payment_layout_cc_error_msg	0.200	1.000	0.670	0.667	0.860	0.216	28 359	12
payment_mobile_support	0.400	1.000	0.260	1.000	0.210	1.000	8 045	5
payment_banner_message_error	0.470	1.000	0.960	1.000	1.610	1.000	7 574	5
placebo_1	-0.590	1.000	-0.420	1.000	-0.520	1.000	12 300	11
placebo_2	1.240	0.035 *	0.780	1.000	1.360	1.000	16 761	13
placebo_3	-0.800	0.606	-0.430	1.000	-0.160	1.000	20 349	15
placebo_4	1.680	0.000 ***	0.830	0.751	0.100	1.000	21 688	15
placebo_5	0.510	1.000	-0.480	1.000	0.030	1.000	11 857	10
placebo_6	-0.700	1.000	-0.330	1.000	-0.200	1.000	17 373	11
placebo_7	0.470	1.000	0.200	1.000	-0.010	1.000	25 833	11
placebo_8	-0.900	0.132	-0.470	1.000	-0.210	1.000	25 523	12

Table 4: Estimated ATEs (in percentage points) and Holm's adjusted p-values for each treatment, where . if p<0.1, * if p<0.05, ** if p<0.01, *** if p<0.001.

Based on the updated table which in this case includes the effects of actual and placebo treatments, we can make some important conclusions and observations:

In terms of robustness, the OLS model and DML PLR model perform better than the Difference in Means model. This is likely because these models take into account covariates, leading to a more accurate estimation of the average treatment effect (ATE). In particular, the DML PLR model, which is designed to deal with potential confounding variables, stands out as a particularly reliable tool for estimating ATE.

The DML PLR model shows that none of the placebo treatments has a significant impact on the conversion rate at the 0.05 level of significance. This indicates the model's strength in effectively separating actual effects from placebo effects, further highlighting the importance of controlling for confounding variables when estimating ATE.

Looking at the 'layout_momondo_carryon_fix' treatment, the positive ATE is significant in all three models. This suggests that this particular treatment had a notable positive impact on the conversion rate. However, for 'layout_remove_luckyorange', while the negative ATE is significant in the Difference in Means and OLS models, the DML PLR model does not find it significant. This discrepancy indicates that the latter treatment might have a potentially negative effect on the conversion rate, although it does not hold when controlling for confounding variables.

In terms of placebo treatments, it's notable that 'placebo_4' shows a significant positive ATE in the Difference in Means method, while it is not significant in the OLS and DML PLR models. This highlights the potential for unobserved factors affecting the outcome in the Difference in Means and

OLS models, which are better accounted for in the DML PLR model due to its higher capabilities owing to implementing machine learning models base learners.

Overall, these results underline the importance of choosing the appropriate model and taking into account its assumptions for causal inference. In particular, the DML PLR model's ability to control for confounding variables appears to provide a more accurate and reliable estimation of average treatment effects, making it a robust tool for these kinds of analyses, effectively distinguishing placebo effects..

5.2 Double Machine Learning

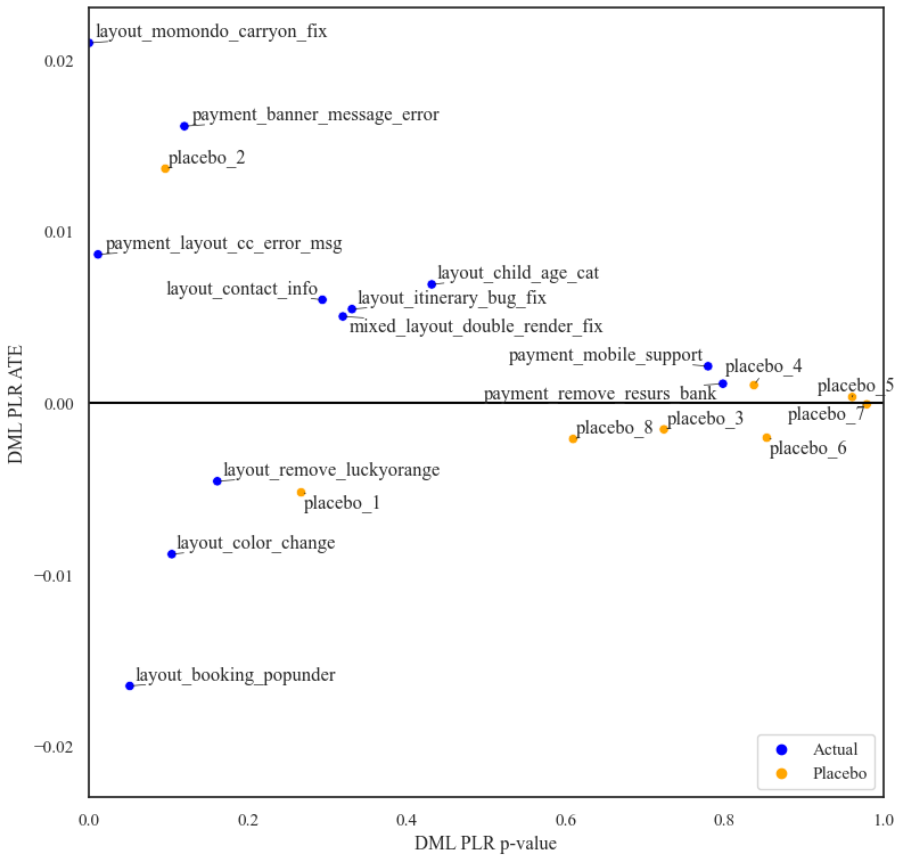


Figure 9: Scatter plot showing distribution of ATE coefficients with standard, not adjusted p-values of these coefficients

This section is dedicated to the Double Machine Learning model and provides a more intricate analysis of its results. Figure 9 presents a visualization of the Average Treatment Effects (ATEs) coefficients along with their corresponding p-values. As an exception here standard p-values (not adjusted with Holm's procedure) were shown for a broader perspective on the outcomes from the DML model. The diagram effectively demonstrates the model's proficiency in distinguishing real changes from placebo changes. The vast majority of placebo effects are characterized by ATE estimates near zero and high p-values, signifying no substantial improvement in conversion rates.

An exception is observed with 'placebo_2', which is positioned among the actual changes to the website. The reason for this occurrence lies in a substantial increase in the conversion rate during the period when this placebo change was implemented. This exposes a weakness in the model: it tends to perform poorly under conditions of high volatility in the visit-to-order ratio.

Figure 10 illustrates the relationship between ATEs estimated with DML and those derived from simple conversion rate difference comparisons. A strong positive correlation is noticeable between these two sets of ATE estimates. Any deviations from a linear relationship between these two groups highlight the influence of confounding variables incorporated in the DML modelling. Most placebo effects were "shifted" towards zero with DML, suggesting the model's overall effective performance and its ability to reduce bias in ATE estimation.

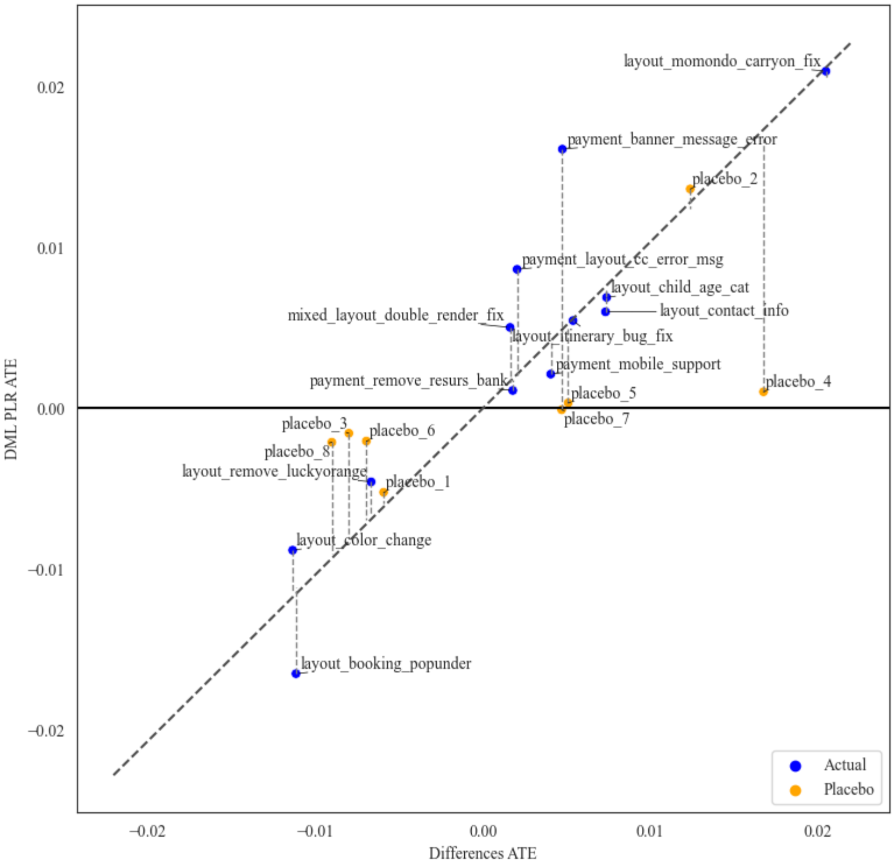


Figure 10: Scatter plot showing the distribution of ATEs from Differences in Means and DML PLR

Treatment name	DML PLR (1 day) ATE	Holm's p-value	DML PLR (3 days) ATE	Holm's p-value	DML PLR (5 days) ATE	Holm's p-value	DML PLR ATE	Holm's p-value	Time window
layout_remove_luckyorange	-6.590	1.000	0.930	1.000	-0.390	1.000	-0.460	1.000	20
payment_remove_resurs_bank	-3.520	0.034 *	-1.310	1.000	0.010	1.000	0.110	1.000	12
layout_momondo_carryon_fix	2.430	1.000	0.940	1.000	0.790	1.000	2.100	0.001 ***	11
layout_itinerary_bug_fix	0.690	1.000	0.260	1.000	-0.920	1.000	0.540	1.000	11
mixed_layout_double_render_fix	-2.100	1.000	0.770	1.000	0.580	1.000	0.500	1.000	13
layout_booking_popunder	-1.160	1.000	-0.720	1.000	-0.990	1.000	-1.650	0.513	7
layout_child_age_cat	2.420	1.000	2.960	1.000	0.360	1.000	0.690	1.000	7
layout_contact_info	1.240	1.000	1.160	1.000	1.500	1.000	0.600	1.000	10
layout_color_change	-3.640	1.000	0.700	1.000	0.060	1.000	-0.880	0.938	11
payment_layout_cc_error_msg	1.120	1.000	-0.310	1.000	1.980	0.061 .	0.860	0.125	12
payment_mobile_support	-1.000	1.000	0.380	1.000	0.050	1.000	0.210	1.000	5
payment_banner_message_error	2.620	1.000	0.980	1.000	1.610	1.000	1.610	0.961	5

Table 5: Estimated ATEs (in p.p.) and Holm's adjusted p-values DML PLR with different time windows

The application of Holm's adjustment to the results of DML PLR for varying time windows is depicted in the table above. As observed in the previously mentioned results (Table 4), the application of Holm's correction for multiple comparisons typically results in fewer significant results, reducing the risk of Type I errors.

Considering the 1-day, 3-day, and 5-day time windows the additional model analysis was conducted to assess the model's behavior for different time windows. Such analysis was prepared in order to draw conclusions about the recommended length of the time window to ensure the reliability of modelling results. As shown in Table 5 for 1-day there is only one treatment ('payment_remove_resurs_bank') that indicates a significant effect on conversion rate with Holm's adjusted p-value at a level of 0.034. For 3-day, and 5-day time windows no significant results can be observed after Holm's adjustment.

Analyzing estimates of ATEs for different time windows we can observe that ATEs vary considerably when changing time intervals. This means that estimation is not stable across different time windows, indicating that the estimated average treatment effect is sensitive to the length of the time window considered.

For instance, the treatment 'payment_remove_resurs_bank' has an estimated ATE of -3.520 in the 1-day time window, which decreases to -1.310 in the 3-day window and is nearly zero in the 5-day window. This may suggest that the effect of this treatment is more immediate and diminishes over time.

Conversely, the 'layout_momondo_carryon_fix' treatment has an estimated ATE of 2.430 on the 1-day window, which decreases to 0.940 on the 3-day window and further to 0.790 on the 5-day window. However, when considering the larger time window used in the DML PLR model, it shows a significant effect with an ATE estimate of 2.100.

These observations underline the importance of considering the appropriate time window for assessing treatment effects. The selection of a time window can have a significant impact on the magnitude and even the direction of the estimated ATEs, potentially leading to different conclusions regarding the effectiveness of the treatments.

5.3 Insights and Analysis of Results

As outlined in Table 4, the only modification to the website that showed a statistically significant impact on improving the conversion rate, following Holm's adjustment, was 'layout_momondo_carryon_fix.' This specific change resolved an issue with adding extra carry-on luggage after redirection from momondo.dk, which is the leading travel search engine in the Danish market. Given that a considerable number of customers potentially faced this issue while trying to add carry-on luggage to their flight tickets.

It's worth emphasizing that the enhancement 'payment_layout_cc_error_msg,' which provided more detailed information about the reasons for unsuccessful payments, achieved a significance level of 0.125 after Holm's adjustment. While not statistically significant at typical levels, this result suggests a meaningful impact on user experience during the payment process.

However, the remainder of the changes assessed in this study did not exhibit any significant effects in terms of conversion rate improvement. Specifically, none of the modifications to the webpage layout and payment process had a noticeable impact on the website's conversion rates.

On the other hand, some changes resulted in a noticeable, albeit statistically non-significant, negative impact on the website's performance. For instance, 'layout_booking_popunder' (an additional window to view an apartment on booking.com) could potentially confuse customers, thereby reducing the clarity of the ordering process and discouraging customers from purchasing flight tickets alone. Another modification, 'layout_color_change' (a change in the color layout), had a negative effect according to the OLS Holm's adjusted p-value at 0.11, although it was not statistically significant in the DML PLR model that can be proof for rejecting this hypothesis.

5.3.1 Implications and Recommendations for Future Development

Based on the analyses and insights gained from this study, several implications and recommendations can be derived for future improvements to the website design and functionality.

Firstly, the value of user-friendly and intuitive design cannot be overemphasized. The 'layout_momondo_carryon_fix' enhancement, which resolved an issue that a considerable number of users might have encountered, had a significant positive impact on the conversion rate. This reinforces the importance of ensuring a seamless user experience at every step of the customer journey. Such issues should be identified and resolved shortly to minimize a potential loss in the number of resulting orders.

Secondly, while the improvements in providing more detailed information about unsuccessful payments did not reach a typical significance level, the trend towards a positive impact is encouraging. This suggests that providing clearer, more detailed information to the user, especially during critical steps such as payments, can potentially improve conversion rates. Future design improvements should thus prioritize transparency and clarity of information, especially in areas of the website that directly impact the completion of transactions.

However, caution should be exercised in implementing enhancements that may potentially confuse or distract users. Changes such as 'layout_booking_popunder' might have led to a decrease in conversion rate due to the addition of a potentially confusing element in the user experience. Therefore, changes should be carefully evaluated not only for their functionality but also for their potential impact on the overall user experience. If a change has the potential to confuse or distract users, it might be better to reconsider or modify the approach. Moreover 'layout_remove_luckyorange' change that had an impact on user experience showed a negative (but not statistically significant) effect on conversion rate. This change implemented to enhance website loading times could potentially affect users in terms of reduced informativeness and a less attractive layout visible to the customers.

In conclusion, the future development of the website should continue to prioritize a user-friendly and intuitive design, provide clear and detailed information to users, and carefully evaluate the potential impact of changes on user experience.

5.4 Feasibility of Double Machine Learning

The Double Machine Learning approach, as implemented in this study, has demonstrated a robust and efficient methodology for estimating Average Treatment Effects (ATE) in a website change evaluation context. It not only provides a nuanced understanding of the impact of implemented changes on the conversion rate but also illustrates the importance of accounting for confounding factors, an aspect that may not be addressed when using simpler comparison methods including simple comparing differences in conversion rates or OLS Linear Regression.

DML has shown its strength in comparing differences in conversion rates and OLS by better controlling for confounding variables, ensuring that the estimated treatment effects are more reliable. The model was able to effectively distinguish placebo treatments from real changes, highlighting its capability in reducing potential bias in ATE estimation.

Nevertheless, it's important to note that estimates from the DML model are highly sensitive to the time window used for analysis. As demonstrated in this study, changing the length of the time window can lead to varying ATEs, highlighting that the choice of an appropriate time window is critical for obtaining reliable estimates.

In essence, the use of Double Machine Learning in the evaluation of website changes offers a more refined analysis and a deeper understanding of the effects of different treatments on conversion rates. However, it is essential to carefully select the appropriate time window for the analysis to ensure the reliability of the results. Therefore, future work may focus on developing guidelines or strategies for selecting an optimal time window for different types of treatments in website change evaluation.

5.5 Limitations of the Study

This study, despite providing valuable insights into the impact of specific changes on the flysmarter.dk website's conversion rate has a few notable limitations:

- **Overlapping changes:** If more than one change was made to the site at the same time, it might be difficult to separate the individual impacts of those changes on the conversion rate. We can distinguish two types of these changes: changes to AOB Travel's search engine and pricing rules may indirectly influence consumer behavior and consequently the conversion rate. Also, The study might not fully account for the cumulative effects or interactions between simultaneous changes.

- **Seasonality:** The analysis might not fully account for seasonal variations in the number of flight bookings. Travel and booking patterns can vary a lot throughout the year due to factors like holidays, school vacations, and travel seasons. Without well-adjusted variables, these patterns could confound the impact of the site changes on conversion rates. This limitation was met due to the granularity of Google Trends data. While it provides trends over time, it is not too specific about the interest of customers on a daily (or even more granular) basis.

- **Potential omitted covariates:** Even though the data set was rich in terms of the number and quality of covariates, the addition of more variables could enhance the model. For instance, the time window between the order date and the flight date could be a potentially beneficial variable to include. Additional covariates would allow for a more complex and thorough model, potentially highlighting further associations or trends.

- **Number of tested changes:** Finally, the relatively small number of changes tested in this study could limit the breadth of insight provided. Having a larger set of changes to evaluate would likely yield more varied and substantial findings, thereby providing a richer pool of data from which the web development team can draw insights for future site enhancements.

- **External factors:** E-commerce environment is very volatile and full of potential external factors that can potentially impact conversion rates on flysmarter.dk. For example, although AOB Travel does not lead any advertising campaigns, these can be run by momonodo and skyscanner offering special offers for certain destinations or occasions and impacting the traffic to AOB Travel websites.

5.6 Future Work

The study shows several potential directions for further research and improvements. Firstly, refining the model to better handle periods of high volatility in the visit-to-order ratio could lead to more accurate and robust ATE estimates. To mention

Secondly, more focus can be put on tuning the base learners and testing other algorithms. In the process of modelling other methods were explored (XGBoost, LightGBM) however they did not provide promising results in terms of robustness versus separating actual changes from placebo effects. Because datasets used for the analysis were relatively big, Double ML or other potential methods could benefit from the advantages of artificial neural networks that are very promising for causal inference settings as they handle very complex relations within data (Shi, Blei, and Veitch 2019).

Thirdly, while the study has demonstrated the potential of DML PLR models for reducing bias in ATE estimation, alternative techniques could be explored and compared for a more comprehensive understanding of their performance. These can be Propensity Score Matching, Instrumental Variables and many causal machine learning techniques.

Lastly, the potential for implementation of transfer learning techniques can be distinguished as the data structure is consistent throughout the analyzed period. Transfer learning could be especially valuable for causal inference in this context as it can leverage insights gained from previous changes and their impacts to better predict the effect of future changes. This approach has the potential to strengthen the model by utilizing patterns identified in past data to improve the accuracy of causal inference, particularly when predicting outcomes for new, unseen changes in pricing rules or website enhancements.

6 Conclusions

This study aimed to estimate the Average Treatment Effects (ATEs) of changes made to the ordering page on flysmarter.dk's website. The main focus was to understand how these modifications influenced the ratio of conversion rates and provide actionable insights for future optimization. The primary approach used to measure these effects was Double Machine Learning (DML), a methodology that has shown considerable potential in reducing bias in observational studies compared to other used methods, including comparing differences in conversion rates and OLS Linear Regression.

The results of this study revealed that only one change, 'layout_momondo_carryon_fix', significantly improved the conversion rate. Some other modifications, although not statistically significant, indicated trends towards both positive and negative directions, suggesting further areas of exploration for future development.

As in addition to our conclusions, this study also emphasizes the importance of the user experience in website design. The significant impact of the 'layout_momondo_carryon_fix' change reconfirms this; user-facing issues need a fast bug-fixing procedure to prevent potential conversion losses.

Providing clear, detailed information, especially during critical steps like the payment process, can improve user experience and potentially boost conversion rates. Future enhancements should prioritize user transparency and information clarity, especially in transaction-critical areas.

Additionally, changes should be carefully evaluated for their overall impact on user experience, not just their functionality. Any modifications, like the 'layout_booking_popunder' change, that may introduce user confusion could potentially decrease conversion rates.

Finally, aesthetic changes, such as 'layout_color_change,' and changes intended to improve load times, like 'layout_remove_luckyorange,' should be cautiously considered as they can significantly impact user perception and interaction.

In summary, future website design should focus on enhancing the user interface, providing clear information, and carefully evaluating the potential impact of any changes on the user experience to improve overall website performance.

From a methodological perspective, in the context of this study, Double Machine Learning (DML) proved to be a robust and efficient method. The approach utilized a comprehensive understanding of how changes on the website impacted the conversion rate. What is important it showed the importance of accounting for confounding factors, an aspect often omitted with simpler comparison methods.

However, the results produced by the DML model were observed to be highly sensitive to the time window chosen for analysis. This study underscored the importance of selecting an appropriate time window for the analysis to ensure the reliability of results. Therefore, one possible direction for future work could involve developing guidelines for choosing the optimal time window for different treatments in website change evaluation.

While this study provided valuable insights, some limitations were identified. These include overlapping changes on the website, potential omission of important covariates, seasonal fluctuations in flight bookings, and the limited number of changes tested. Future studies could seek to address these limitations for a more robust and insightful analysis.

Moreover, considering external factors that could influence the e-commerce environment and subsequently affect conversion rates is crucial. Such externalities could range from promotional campaigns run by travel search engines to global events or economic shifts that may impact consumer behavior.

Going forward, several potential directions for research and improvements have been identified. These involve refining the model to better handle periods of high volatility in the visit-to-order ratio, focusing more on tuning the base learners and testing other algorithms, exploring alternative techniques for a more comprehensive understanding of their performance, and considering the implementation of transfer learning techniques.

In conclusion, the results from this study offer valuable insights that could guide future website optimization processes. By integrating these findings with continued innovation and testing, there are substantial opportunities to further enhance the user experience and, in turn, improve conversion rates on the website.

References

Alexander, Lucy (Mar. 14, 2023). "Ecommerce Conversion Rates Across Industries (And How to Raise Yours)". In: URL: https://blog.hubspot.com/marketing/ecommerce-conversion-rates-across-industries.

Angrist, Joshua David and Jörn-Steffen Pischke (2009). *Mostly harmless econometrics: an empiricist's companion*. OCLC: ocn231586808. Princeton: Princeton University Press. 373 pp. ISBN: 978-0-691-12034-8 978-0-691-12035-5.

— (2015). *Mastering 'metrics: the path from cause to effect*. Princeton ; Oxford: Princeton University Press. 282 pp. ISBN: 978-0-691-15283-7 978-0-691-15284-4.

AOB Travel (Feb. 6, 2023). "AOB Travel: Online Travel Agency (OTA)".

Athey, Susan and Stefan Wager (2019). "Estimating Treatment Effects with Causal Forests: An Application". In: Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.1902.07409. URL: https://arxiv.org/abs/1902.07409 (visited on 05/13/2023).

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. ISSN: 08856125. DOI: 10.1023/A:1010933404324. URL: http://link.springer.com/10.1023/A:1010933404324 (visited on 05/17/2023).

Chan, Irene Cheng Chu et al. (Sept. 2021). "Website design in tourism and hospitality: A multilevel review". In: *International Journal of Tourism Research* 23.5, pp. 805–815. ISSN: 1099-2340, 1522-1970. DOI: 10.1002/jtr.2443. URL: https://onlinelibrary.wiley.com/doi/10.1002/jtr.2443 (visited on 05/21/2023).

Chen, Aiyou and Timothy C. Au (Aug. 8, 2019). *Robust Causal Inference for Incremental Return on Ad Spend with Randomized Paired Geo Experiments*. arXiv.org. URL: https://arxiv.org/abs/1908.02922v3 (visited on 05/23/2023).

Chernozhukov, Victor et al. (Feb. 1, 2018). "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21.1, pp. C1–C68. ISSN: 1368-4221, 1368-423X. DOI: 10.1111/ectj.12097. URL: https://academic.oup.com/ectj/article/21/1/C1/5056401 (visited on 03/05/2023).

Esmukov, Kostya (Nov. 13, 2022). *GeoPy*. Version 2.3.0.

FreecurrencyAPI (Mar. 27, 2023). *Currency Exchange Rates*. URL: https://freecurrencyapi.com/.

Holm, Sture (1979). "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian Journal of Statistics* 6.2. Publisher: [Board of the Foundation of the Scandinavian Journal of Statistics, Wiley], pp. 65–70. ISSN: 03036898, 14679469. URL: http://www.jstor.org/stable/4615733 (visited on 05/23/2023).

Imbens, Guido W. and Donald B. Rubin (Apr. 6, 2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. 1st ed. Cambridge University Press. ISBN: 978-0-521-88588-1 978-1-139-02575-1. DOI: 10.1017/CBO9781139025751. URL: https://www.cambridge.org/core/product/identifier/9781139025751/type/book (visited on 05/24/2023).

IP2Location (Aug. 9, 2018). *IATA/ICAO List*. https://raw.githubusercontent.com/ip2location/ip2location-iata-icao/master/iata-icao.csv. URL: https://raw.githubusercontent.com/ip2location/ip2location-iata-icao/master/iata-icao.csv (visited on 05/10/2023).

John, E. R. et al. (Dec. 2019). "Assessing causal treatment effect estimation when using large observational datasets". In: *BMC Medical Research Methodology* 19.1, p. 207. ISSN: 1471-2288. DOI: 10.1186/s12874-019-0858-x. URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0858-x (visited on 05/17/2023).

Karmakar, Somedip, Soumojit Guha Majumder, and Dhiraj Gangaraju (Jan. 4, 2023). "Causal Inference and Causal Machine Learning with Practical Applications: The paper highlights the concepts of Causal Inference and Causal ML along with different implementation techniques". In: *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*. Conference Name: CODS-COMAD 2023: 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD) ISBN: 9781450397971 Place: Mumbai India Publisher: ACM, pp. 324–326. DOI: 10.1145/3570991.3571052. URL: https://dl.acm.org/doi/10.1145/3570991.3571052 (visited on 05/23/2023).

Kohavi, Ron, Alex Deng, et al. (Aug. 12, 2012). "Trustworthy online controlled experiments: five puzzling outcomes explained". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '12: The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing China: ACM, pp. 786–794. ISBN: 978-1-4503-1462-6. DOI: 10.1145/2339530.2339653. URL: https://dl.acm.org/doi/10.1145/2339530.2339653 (visited on 05/14/2023).

Kohavi, Ron, Diane Tang, and Ya Xu (Apr. 2, 2020). *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. 1st ed. Cambridge University Press. ISBN: 978-1-108-65398-5 978-1-108-72426-5. DOI: 10.1017/9781108653985. URL: https://www.cambridge.org/core/product/identifier/9781108653985/type/book (visited on 05/20/2023).

Luca, Michael and Max H. Bazerman (2021). *The power of experiments: decision making in a data-driven world*. First editon. Cambridge, Massachusetts: The MIT Press. 211 pp. ISBN: 978-0-262-04387-8 978-0-262-54227-2.

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell (2016). *Causal inference in statistics: a primer*. Chichester, West Sussex: Wiley. 136 pp. ISBN: 978-1-119-18684-7.

Reese, Simon (2022). "DABN14/STAN52 Lecture 8: ML for causal inference, Double Machine Learning". Lund University.

Rosenbaum, Paul R. (2020). *Design of Observational Studies*. Springer Series in Statistics. Cham: Springer International Publishing. ISBN: 978-3-030-46404-2 978-3-030-46405-9. DOI: 10.1007/978-3-030-46405-9. URL: https://link.springer.com/10.1007/978-3-030-46405-9 (visited on 05/16/2023).

Rosenbaum, Paul R. and Donald B. Rubin (1983). "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1, pp. 41–55. ISSN: 0006-3444, 1464-3510. DOI: 10.1093/biomet/70.1.41. URL: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/70.1.41 (visited on 05/13/2023).

Saxena, Shipra (Oct. 27, 2020). *A/B Testing for Data Science using Python - A Must-Read Guide for Data Scientists*. Analytics Vidhya. URL: https://www.analyticsvidhya.com/blog/2020/10/ab-testing-data-science/ (visited on 05/14/2023).

Shi, Claudia, David Blei, and Victor Veitch (2019). "Adapting Neural Networks for the Estimation of Treatment Effects". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/hash/8fb5f8be2aa9d6c64a04e3ab9f63feee-Abstract.html (visited on 05/24/2023).

Siroker, Dan and Pete Koomen (2013). *A/B testing: the most powerful way to turn clicks into customers*. Hoboken: Wiley. 1 p. ISBN: 978-1-118-65920-5 978-1-118-65917-5 978-1-118-65922-9.

Skytrax (Mar. 15, 2023). *Skytrax World Airline Star Rating 2023*. URL: https://skytraxratings.com/about-airline-rating.

Statista (May 2017). *How many different websites or providers do you usually compare before you book a flight?* Statista Survey. URL: https://www.statista.com/statistics/703536/number-of-websites-or-providers-visited-before-booking-flights/ (visited on 05/21/2023).

Trends, Google (Mar. 20, 2023). *Google Trends: Interest over time for kewyords: flybilletter, momondo, direkte fly, skyscanner, flybilletter billige for Denmaerk region*. URL: https://trends.google.com/trends/explore?date=2022-01-01%202023-03-20&geo=DK&q=flybilletter,momondo,direkte%20fly,skyscanner,flybilletter%20billige&hl=en.

Wikipedia (Apr. 7, 2023). *List of airline codes (A)*. In: *Wikipedia*. Page Version ID: 1148668764. URL: https://en.wikipedia.org/w/index.php?title=List_of_airline_codes_(A)&oldid=1148668764 (visited on 05/10/2023).

Xu, Jiaqin et al. (Dec. 28, 2022). "Estimation of average treatment effect based on a multi-index propensity score". In: *BMC Medical Research Methodology* 22.1, p. 337. ISSN: 1471-2288. DOI: 10.1186/s12874-022-01822-3. URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01822-3 (visited on 05/10/2023).

A Appendix

A.1 Data Dictionary

Variable Name	Category	Description
carriers_marketing_ratings_count	travel characteristics	Number of marketing carrier ratings
carriers_marketing_ratings_mean	travel characteristics	Mean value of marketing carrier ratings
clicks_count	demand	Number of clicks on the itinerary
clicks_created_at_datetime_hour_0-6	time-related	Clicks created between 0:00-6:00
clicks_created_at_datetime_hour_13-18	time-related	Clicks created between 13:00-18
clicks_created_at_datetime_hour_19-24	time-related	Clicks created between 19:00-24:00
clicks_created_at_datetime_hour_7-12	time-related	Clicks created between 7:00-12:00
clicks_created_at_datetime_weekday_1-7	time-related	Clicks created on weekdays from Monday to Sunday
clicks_created_at_datetime_weekend	time-related	Clicks created on weekends
clicks_itinerary_direct_flight	travel characteristics	Direct flights in the clicked itinerary
clicks_itinerary_sales_price_diff_best	price related	Difference between sales price and best price
clicks_itinerary_sales_price_diff_cheapest	price related	Difference between sales price and cheapest price
clicks_itinerary_sales_price_diff_fastest	price related	Difference between sales price and fastest price
clicks_itinerary_sales_price_if_best	price related	Sales price if it's the best price
clicks_itinerary_sales_price_if_cheapest	price related	Sales price if it's the cheapest price
clicks_itinerary_sales_price_if_fastest	price related	Sales price if it's the fastest price
clicks_itinerary_sales_price_pax	price related	Sales price per passenger
clicks_itinerary_segment_count	travel characteristics	Number of segments in the clicked itinerary
clicks_itinerary_totaldistance	travel characteristics	Total distance of the clicked itinerary
click_itinerary_travel_timehours	travel characteristics	Total travel time of the clicked itinerary
clicks_itinerary_with_baggage	travel characteristics	Itineraries with baggage included
clicks_mobile	customer characteristics	Clicks made from mobile devices
clicks_passengers_count	travel characteristics	Number of passengers in the clicked itinerary
diff_clicks_google_trends	demand	Difference between clicks count and google trends data
google_trends_weekly_DK	demand	Weekly Google trends data for Denmark
google_trends_weekly_DK_lag_7	demand	Lagged Google trends data for Denmark by 7 days
ratio_clicks_google_trends	demand	Ratio of clicks count to Google trends data
ratio_distance_passenger	travel characteristics	Ratio of total distance to number of passengers
ratio_sales_price_carrier_rating_count	price related	Ratio of sales price to number of carriers in a travel
ratio_sales_price_carrier_rating_max	price related	Ratio of sales price to maximum carrier rating
ratio_sales_price_distance	price related	Ratio of sales price to total distance
ratio_sales_price_travel_time	price related	Ratio of sales price to total travel time
ratio_travel_time_distance	travel characteristics	Ratio of total travel time to total distance

Table 6: Summary of variables used as covariates in the modelling