Evaluating Heterogeneous Treatment Effects: Hyper-contextual advertisement targeting, a Causal Machine Learning approach

Erasmus School of Economics | Erasmus University Rotterdam
Bachelor thesis Econometrics and Operations Research

Student: Abdelmounaim el Yaakoubi
Student Identification Number: 509792ay
Supervisor: Phd.(c) Hong Deng
Second Assessor: Dr. Flavius Frasincar

Date: 4 July 2021

Abstract

Hyper-contextual targeting has been a popular and effective advertisement tool due to global digitalization. To enhance customer responses to marketing campaigns, companies have been collecting data and exploiting not only customer characteristics but also environmental factors. The main contextual factor that is analyzed in this research, is crowdedness in public transport. Previous research has shown that the adaptation of social behaviour to crowdedness in public transport realizes itself in prolonged mobile phone immersion. The purchase likelihood of customers increases as advertisements can relieve the discomfort that arises when personal spaces are invaded. In order to quantify the effect of sudden variations in crowdedness on purchase likelihood, the Causal Forest algorithm is implemented. This Machine Learning technique estimates heterogeneous treatment effects and allows for statistical inference. This research constructs a simulation framework that models customer heterogeneity and the non-linear effects of crowdedness. The advice for hyper-contextual targeting is to implement the Causal Forest algorithm in order to identify endogeneity and exploit customer heterogeneity.

Keywords:
Causal Inference, Machine Learning, Causal Forests, Hyper-contextual advertisement targeting

Contents

1 Introduction

Global digitalization has led to an extraordinary complex interaction between human beings and technological devices. The informational infrastructure in our contemporary society is dependent on the functionality of these technological devices to operate effectively. In particular, smartphones have proven to be an excellent tool for communication and consumption. Currently, daily mobile phone usage has increased exponentially over the last decade, resulting in an average screen time that ascends several hours per day. Marketers aim to capture the "digital presence" of human beings to quantify consumer behaviour and promote their products. Social media platforms are an excellent avenue for self-promotion, campaigns and social interaction with customers as researched by Dwivedi et al. 2021.

The advantage of social media is that it allows for adequate data collection so that advertisements are tailored to the customers' preferences, also known as narrow-casting. Therefore, marketers aim to exploit customer heterogeneity in order to seek innovation-based competitive advantage per Wijekoon, Salunke, and Athaide 2021. Thus, more than ever companies are competing for the attention of potential clients by differentiating their clients based on their unique profile and characteristics.

User characteristics play a crucial role in determining the susceptibility of customers to advertisements. Recently, a more advanced marketing strategy is applied by companies based on contextual factors. These contextual factors can potentially influence customer behaviour. It includes the customers' current location, weather conditions, engagement in specific activities etc. In the research conducted by Andrews et al. 2016, the hyper-contextual factor of physical crowdedness in the context of public transport was correlated with increased sales purchases.

The research has shown that the personal and physical spaces of individuals are invaded in crowded environments such as underground trains. The spatial limitation of underground subways limits outside options and restricts social behaviour. The tendency to cope with this discomfort is prolonged mobile immersion that allows commuters to psychologically withdraw from the physical crowd and their attention focuses increasingly on mobile advertisement per Julsrud and Denstadli 2017. The increased susceptibility to mobile advertisements that arises from this phenomenon is an important hyper-contextual factor that can be exploited by marketers.

The data from the article by Andrews et al. 2016 is publicly available on the website of Informs, which is an Institute for Operations Research and the Management Sciences. The information was gathered by one of the worlds largest telecommunication providers established in Asia. As means to identify the causal effects of public crowding on purchases, the telecommunication provider designed a unique experiment in which the provider sent targeted mobile ads to a selection of their customers for a missed-call-service offer. The mobile service provider used chip-tracking technology to determine physical crowdedness in subway trains. Crowdedness is naturally confounded with the time of day and the day of the week. The potential problem that arises is self-selection, thus contextual factors and consumer characteristics are represented in the data. In the end, purchases were measured and some purchasers and non-purchasers were surveyed regarding their purchase behaviour.

The empirical research Andrews et al. 2016 proved that purchase rates are significantly influenced by crowdedness via a logistic regression model. However, the analytical conclusion proposed are only applicable for moderate crowdedness levels nor the heterogeneity of customers is captured. In theory, certain

non-moderate crowdedness levels that negatively impact purchase likelihood exist due to under-crowding or congestion for example. This research is intended to quantify the non-linearity and heterogeneous treatment effects of sudden variations in crowdedness caused by unanticipated train delays. In the framework of this research, the effects of unanticipated train delays on purchase behaviour are examined by means of a simulation.
The research question now becomes: *"How can the heterogeneous effects of sudden variations in crowdedness on purchase behaviour be examined, taking into account customer heterogeneity and the non-linear effect of crowdedness?"*. The research question is decomposed into the following sub-questions:

- *"How is it possible to simulate customer heterogeneity and the non-linear effects of crowdedness on purchase likelihood?"*
- *"How is it possible to quantify heterogeneous effects of sudden variations in crowdedness and which variables are most significant in explaining customer heterogeneity?"*
- *"How can the thresholds for which the effects of crowdedness are most reliable be obtained and what is the extent to which customer heterogeneity is captured?"*

For the purpose of answering these research questions, this paper is dissected in multiple sections. The *Literature* section provides a broad overview of the academic researches that contributed to this thesis. The *Data* section provides the background of the data collection and a compact overview of the factors that are considered. In the section of the *Experimental Design* criteria for the simulation are formulated and mathematical notation is introduced for the data generating process. The *Methodology* section describes the statistical methods utilized for the replication and extension of the paper by Andrews et al. 2016. Primarily, it is explained how the Causal Forests algorithm by Athey and Wager 2019 is implemented to calculate heterogeneous treatment effects for sudden variations in crowdedness in the form of unanticipated train delays. Additionally, crowdedness thresholds that maximize the adequacy of forest predictions are determined. In the *Result* section the main findings of this research are presented and illustrated in figures and tables. The heterogeneous treatment effects of different crowdedness intervals and the variable importance in capturing heterogeneity are assessed. The *Conclusion* section finalizes the thesis by translating the results into feasible pieces of advice for marketers and companies.

2 Literature

As an extension to the research conducted by Andrews et al. 2016, the effects of sudden variations in crowdedness on consumer response behaviour is assessed. The contemporary tendency towards crowding in public transport is an inward focus that manifests in prolonged mobile phone immersion. The increased susceptibility to mobile advertisements that arises from this, forms an important hyper-contextual factor that can be exploited by marketers. A previous study by Julsrud and Denstadli 2017 has shown that in crowded environments such as underground subways, the personal and physical spaces of individuals are invaded. This spatial constraint poses a behavioural limitation, consequently, leading to experience a reduction in the available outside options. To cope with this discomfort, mobile immersion allows commuters to psychologically withdraw from the physical crowd and their attention focuses increasingly on

mobile advertisement. Currently, individuals turn inwards and pay prolonged attention to their smartphones as an adaptive social behaviour as examined by Maeng, Tanner, and Soman 2013. The immersion in their private mobiles facilitates escapism of unwanted encounters and allows individuals to regain control over their privacy and environment.

Thus, the probability of purchase responses in more crowded trains becomes more likely. However, the effect of crowding on purchase likelihood is rather non-linear with a lower and upper threshold in practice. Initially, commuters may dynamically preserve their personal space as an adaptation to the closed environment of subways. This becomes infeasible after a certain lower threshold level of crowdedness so that commuters resort to turning inwards via mobile phone immersion. In practice, an upper threshold exists because large densities result in congestion that limits mobile phone usage and negatively impacts mobile ad response France-Presse 2014.

In the empirical research conducted by Andrews et al. 2016 potential biases due to self-selection were controlled by including peak versus non-peak traffic hours, weekdays versus weekends, mobile use measurements and by randomly sending mobile ads. In addition, sudden changes in crowdedness induced by train delays were used as an exogenous shock of crowdedness to examine endogenous selection. Lastly, to further evaluate selection biases, Propensity Score Matching is utilized to tests whether purchases are determined by crowdedness rather than customer heterogeneity. This research aims to examine the effect of additional crowdedness by implementing statistical methods that take customer heterogeneity into account.

Causal inference relies on Machine Learning methods that examine treatment effects under the condition of unconfoundedness. To estimate the Average Treatment Effect (ATE), a rich set of covariates is required for a righteous comparison between the test and the control group. Generalized linear regression methods described by Heij et al. 2004 that include confounding variables are widely used. However, the disadvantage of these traditional econometric methods is that the demonstrated estimates cannot fully replicate results from a randomized study according to LaLonde 1986. In the quest for suitable causal inference models, the implementation of propensity scores, that is the probability of being in the treated group, is introduced by Rosenbaum and Rubin 1983. However, according to Hahn 1998 the estimation of the ATE based on the propensity score frameworks potentially causes inefficiency. Robins, Rotnitzky, and Zhao 1995 provide a parametric method that combines propensity scores and response functions. These are reliable methods with "double-robust" and consistent estimators if either the response function or propensity score is specified correctly. However, parametric models fail to capture possible patterns of heterogeneity if subgroups are incorrectly specified. In addition, Chernozhukov, Chetverikov, et al. 2017 introduces Debiased Machine Learning (DML) to estimate the Average Treatment Effects by implementing the score function of Robins, Rotnitzky, and Zhao 1995. Furthermore, Chernozhukov, Demirer, et al. 2018 proposes a strategy that estimates The Conditional Average Treatment Effect (CATE) using non-parametric Machine Learning methods.

The most popular method to estimate heterogeneous treatment effects is the Causal Forest algorithm developed by Athey, Tibshirani, Wager, et al. 2019 that is a non-parametric Machine Learning technique. This algorithm is part of the family of Generalized Random Forest that is an extension of Random Forest of Breiman 2001. In causal inference, this non-parametric estimation framework is widely considered to be one of the most efficient techniques with respect to drawing causal effect inferences of a treatment.

In contrast to Random Forests, the objective of Causal Forests is to maximize heterogeneity between all leaves instead of maximizing predictive accuracy. In order to achieve this, an adaptive weighting mechanism of Causal Trees is implemented alongside a gradient-based approximation to compute parameter estimates. The principle of bootstrapping data samples and "out-of-bag" errors allows for honesty in constructing decision trees. Furthermore, it is proven that under certain conditions the parameter estimates are consistent and approximately Gaussian that allows for likelihood inference. Nie and Wager 2021 developed the R-learner objective function that is a regularization of the objective function that targets Heterogeneous Treatment Effects (HTE). As an additional feature for the improvement of performance, Athey and Wager 2019 implemented an orthogonalization step alongside parameter tuning based on the R-learner.

Wager and Athey 2018 implemented the Causal Forest in an observational study in which they elaborately explained the statistical framework of Causal Forests and proved the applicability of the algorithm. Athey, Tibshirani, Wager, et al. 2019 recently developed an the *grf-package* for the generalized random forest. Because R-software is an open-source program, it allows for the accessibility and the efficient implementation of the Causal Forest algorithm to new data sets.

3 Data

The data is publicly available on the website of Informs which is an Institute for Operations Research and the Management Sciences. The information was gathered by one of the world's largest telecommunication providers established in Asia. The company designed a unique experiment to gather data for the identification of the causal effects of crowding. The telecommunication provider sent targeted mobile ads to a selection of their customers with the following offer (per Andrews et al. 2016, section 3): *"Missed a call and want to know from whom? Subscribe to [Wireless Service Provider's] missed call alert service and be notified by SMS of the calls you missed! Only ¥9 for 3 months! Get ¥3 off if you reply 'Y' to this SMS within the next 20 minutes!"*. The message was sent via SMS, which stands for short message service. The mobile phone service could be purchased by responding to the SMS and costs would be charged. The mobile service company randomly sampled commuters who travelled in the underground subway and targeted this specific subway population by implementing an instant computational and randomization procedure.

Afterwards, purchases were measured and some purchasers and non-purchasers were willing to fill in a survey regarding their purchase behaviour. The mobile service provider was also able to measure physical crowdedness in terms of the number of active mobile users in subway trains. Due to the implementation of chip-tracking technology by the telecommunication company the provider could indicate all active phones. In the end, the company was able to identify all mobile users' phone numbers and their location including those phones that were inactive or turned off. As approximately seventy per cent of the population uses the mobile server, the crowd density was computed by calculating the expected number of commuters within the vehicle and the vehicle surface.

In total, there are 1664 observations from individual customers with information about their profile and the context of the offer. Note that this data-set is just a sample from the original data-set with a total number of 14972 observations from mobile phone users. The dependent variable is the purchase indicator

because the effects of environmental and profile factors on purchase likelihood are examined. The purchase indicator equals one if indeed a purchase was made and it equals zero if a purchase wasn't made. The key dependent variable in this research is crowdedness. However, crowdedness levels are naturally confounded with the time of the day and the day of the week. Self-selection is a potential problem that arises when people self-select into more or less crowded trains depending on these factors. If this is the case, purchases are more likely to be determined by variations in the type of commuter instead of crowdedness levels. In addition, a weekend indicator and a peak-hour indicator are added to the set of explanatory variables which, respectively, indicate one if it is weekend or a peak-hour and otherwise is equals zero. These dummy variables help isolate the systematic differences between weekday and weekend crowdedness. Furthermore, several profile variables are included to account for mobile phone behaviour, which are the mobile users' monthly bills or average revenue per user (ARPU), monthly call minutes of usage (MOU), the messages sent and received (SMS) and data-usage (GPRS). Table 1 provides an overview of all the relevant variables and their descriptions.

Table 1: Descriptions variables in the data-set

Variables	**Description**
purchase	indicates whether indeed the offer resulted in a purchase of the service
crowdedness	indicates the level of crowdedness within the underground train vehicle
weekend	indicates whether the offer was sent during the weekend
peak	indicates whether the offer was sent during a peak-hour
arpu	the average revenue per unit on mobile services
mou	voice time of a certain customer
sms	the amount of monthly text messages sent and received
gprs	a measure of the customers monthly data-usage with the wireless service provider

In order to address endogenous selection and to eliminate alternative explanations for purchase behaviour, the study of Andrews et al. 2016 utilizes an identification strategy. The identification strategy exploits sudden variations in crowdedness that are induced by unanticipated train delays underground and street closures above ground. However, there is no data set available that represented this identification strategy. Therefore, a simulation framework is used based on the provided data to replicate the results of the research.

Table 2 summarizes the statistical properties of the relevant variables in presenting the mean, standard deviation, minimum and maximum value. The dependent variable *purchase* alongside the contextual covariates *weekend* and *peak* are dummy-variables. On average the purchase rate was roughly three per cent as shown in the table. The key explanatory variable *crowdedness* varies between 0.83 and 5.36 commuters per square meter. The purchase rate doubles as the crowdedness levels move from the minimum values to the maximum values. The purchase rates measured 2.1% with fewer than two people per square meter and increased to 4.3% with five people per square meter after controlling for all covariates. The correlation between the profile variables is interestingly high compared to the other variables. This is because mobile phone usage naturally confounds the profile factors.

Table 2: Summary statistics of the relevant variables

Variable	Mean	Standard Deviation	Minimum	Maximum
purchase	0.029	0.1666	0.000	1.000
crowdedness	2.768	1.407	0.830	5.360
weekend	0.328	0.469	0.000	1.000
peak	0.362	0.480	0.000	1.000
log(arpu)	3.927	0.571	2.224	6.403
log(mou)	5.829	1.179	0.000	9.845
log(sms)	5.466	0.966	0.000	7.855
log(gprs)	9.056	2.383	0.000	15.225

4 Experimental design

4.1 Simulation Criteria

The purpose of the experimental design is to construct the data set for the extension of the research. There are certain criteria that have to be fulfilled to simulate the context of the research by Andrews et al. 2016 as accurately as possible. These criteria can be summarized as follows:

- The simulated data has to be an adequate representation of the context discussed in the paper by Andrews et al. 2016. This implies that the relationship between variables and the equations for the data generating process is in correspondence with the information provided in the paper by Andrews et al. 2016. The experimental design will therefore be based on the logistic regression outputs and the data set that are provided. Furthermore, additional features and assumptions are introduced so that general data features are captured and so that applicability is enhanced.

- The simulated data constructs a non-linear relationship between crowdedness and purchase likelihood. In the paper by Andrews et al. 2016, it was explicitly stated the effect of crowdedness is ambiguous. The output for the logistic regression for crowdedness levels below the threshold of two commuters per square meter shows that the effect of crowdedness on purchase likelihood becomes negative. Moreover, it was discussed that in theory there also exists an upper-threshold that potentially negatively influences purchase behaviour as congestion leads to restricted mobility. In a set of Asian metropolises, the maximum capacity is up to eleven commuters per square meter. The assumption for the experimental design is that there exists an upper and lower threshold that negatively affects purchase rates.

- The simulated data constructs individualized heterogeneous effects of the treatment based on their unique profiles. In the paper by Andrews et al. 2016 it was discussed that the effects of an intervention in crowdedness are heterogeneous. For example, individuals with an increased average revenue per unit will be more likely to buy a missed called alert. The assumption for the data generating process is that these individual features do influence the probability of purchase and are more impactful for the treatment group.

To incorporate these features, the data set for different crowdedness intervals are united as the optimal thresholds are not known beforehand. The machine learning algorithm k-means is a robust method to identify clusters based on their characteristics as proposed by Pollard 1981. By clustering each customer into a group based on their profile and contextual factors, the representability of the data can be assessed.

4.2 Data Generating Process

Definition of variables

Dependent variable Y_i (binary): whether a customer made the purchase of the offered mobile service.
Treatment covariate W_i (binary): whether a customer received the offer whilst an intervention in public transport in the form of an unanticipated train delay that induced additional crowdedness.
Explanatory variables X_i (mixed): deterministic contextual and personal factors

Each customer is randomly assigned to be either in the control group or in the treatment group by a Bernoulli distribution:

$$W_i \sim Bernoulli(p = 0.5) \qquad\qquad i = 1, \ldots, n$$

With n being equal to the sample size 6441 and i i the index of each customer. If W_i equals to one, the individual belongs to the treatment group and if W_i equals to zero, an individual belongs to the control group. Furthermore, the crowdedness levels for an individual customer is uniformly sampled:

$$X_i^{crowd} \sim U(a = 0.5, b = 11) \qquad\qquad i = 1, \ldots, n$$

With a being the minimal crowdedness level and b the maximal crowdedness level which is set by 0.5 and 11, respectively. The motivation behind these values is that in the research conducted by Andrews et al. 2016 it was mentioned that levels of crowdedness could be up to eleven commuters per square meter in Asian metropolises. The data has shown that crowdedness levels do not descend below 0.5 and do not ascend above 5.5 commuters per square meter. To model the non-linear effect of crowdedness it is essential that the interval for crowdedness is broadened and that a non-linear function is defined for the parameter of crowdedness. The non-linearity function is a goniometric function with negative values in the tails and positive values in the centre. The function is defined as follows:

$$\theta(x_i^{crowd}) = -0.18cos(0.57x_i^{crowd})$$

Figure 1: R plot of the non-linear function

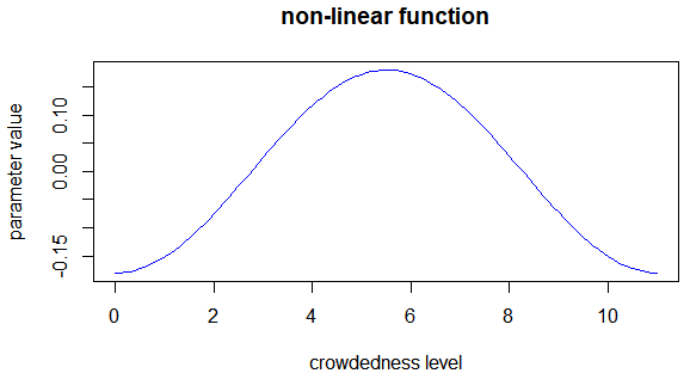


Figure 2 is a visual representation of the non-linear function that is implemented. For levels below approx. two and above approx. seven the effects are negative. Also, the parameter estimate can take up a maximum value of 0.18. To elaborate on this, the treatment effects are calculated based on the customer characteristics. The mathematical notation for the simulated heterogeneous effects is as follows:

$$\mathcal{T}(X_i) = \lambda_1 \theta(x_i^{crowd} + 0.72) + X_i'^{(-crowd)} \lambda^{(-crowd)}$$

The $\mathcal{T}(X_i)$ allows for the distinction between the treatment and the baseline effect. The motivation for adding an additional unit in the non-linear function is that the effect of a train delay is dependent on the level of crowdedness. The average increase in crowdedness due to a train delay was estimated at 0.72 additional commuters per square meter. For instance, an intervention in public transport can lead to an additional negative effect if trains are already highly crowded. In contrast, an intervention in public transport can lead to an additional positive effect if trains are not crowded. Furthermore, it is assumed that all profile and contextual variables influence the treatment effect. The superscript $^{(-crowd)}$ entails that all the explanatory variables are added in the formula except for the crowdedness covariate. The reason behind this is that the effect of crowdedness is represented in the non-linear function. It is assumed that all covariates influence the effect of treatment so they are added in order to distinguish effects between treated individuals. This treatment effect will be multiplied with the treatment indicator making the additional effect of covariates only relevant if the customer belongs to the treatment group. Now that the heterogeneous treatment effects and the non-linear parameter function are set, it is possible to compute purchase utility for each customer. Based on this utility function, the purchase likelihood is calculated as follows:

$$\mathcal{U}(W_i, X_i) = \alpha + \mathcal{T}(X_i) W_i + \theta(x_i^{crowd}) x_i^{crowd} + X_i'^{(-crowd)} \beta$$

$$\rho_i = P(Y_i = 1 | W_i, X_i) = \frac{\exp(\mathcal{U}(W_i, X_i))}{1 + \exp(\mathcal{U}(W_i, X_i))}$$

As a result, the realization of the purchase outcome can be sampled based on the purchase likelihood.

The assignment of a customer to be in the purchase or non-purchase group is sampled as follows:

$$Y_i \sim Bernoulli(p = \rho_i) \qquad\qquad i = 1, \ldots, n$$

This means that the purchase outcome is Bernoulli distributed based on the likelihood of a purchase. The parameter values will be selected in such a way that the purchase rate is approx. 5% and that the treatment group has overall higher purchase rates compared to the control group. Another important feature that the data needs to satisfy, is the variety of purchase rates across different heterogeneous groups. This can be analysed by comparing the purchase rates of the clusters that were made by k-means and see if the explanatory variables indeed influence purchase likelihood.

5 Methodology

5.1 Replication

The empirical analysis that was conducted by Andrews et al. 2016 primarily relied on the logistic regression. Logistic regression is a member of the Generalized Linear Models as described by Heij et al. 2004. In this application, the purchase response Y_i is a binary response variable for which the probability of purchase $P(Y_i = 1)$ depends on explanatory variables including the treatment variable. The logistic regression framework models the logarithmic odds ratio in the following way:

$$log\left(\frac{P(Y_i = 1)}{P(Y_i = 0)}\right) = log\left(\frac{\Lambda(X_i'\beta)}{1 - \Lambda(X_i'\beta)}\right) = X_i'\beta$$

With $\Lambda(t) = e^t/(1 + e^t)$ so that the *logit* model represents the *log-odds* as a linear function of the explanatory variables. The parameter estimates are obtained from iterative procedures that numerically maximize the logarithmic likelihood function. The purchase likelihood of each customer is explicitly related to the treatment, the profile and the contextual variables. In order to address the potential endogenous selection bias, a treatment is introduced that offers the same population of commuters in the subway system. Train delays are an unexpected priori that cause additional crowdedness in subways. By including this variable into the logistic regression one can test the significance of the train delay on determining purchase likelihood. If this effect is insignificant one can argue that an unanticipated train delay in itself should not lead to a higher purchase likelihood, rather the additional crowdedness is what causes it. However, there are limitations to the identification of potential endogeneity. The heterogeneity of customers may lead to different individual treatment effects, whereas in the logistic regression this is considered to be homogeneous. This illustrates the general underlying issue with the *logit* model because explanatory variables may be different in their effect under certain contextual circumstances. Thus, in order to identify potential endogeneity in the context of train delays, a more general framework is proposed to deal with the limitations of the logistic regression.

5.2 Treatment Effects Framework

The estimation of heterogeneous treatment effects relies on the purchase outcome for individuals in the case of them being in the control group $Y_i(control)$ and in the treatment group $Y_i(treatment)$. The discrepancy between these outcomes represents the individual treatment effect (ITE) that is:

$$\tau_i(X) = Y_i(control) - Y_i(treatment)$$

However, the realization of the outcomes can not be obtained simultaneously in practice. Therefore, a potential outcome framework is proposed by G. W. Imbens and Rubin 2015 in which the conditional average treatment effect (CATE) is formulated as follows:

$$\tau_i(X) = E[Y_i(control) - Y_i(treatment)|X_i = x]$$

The identification of the conditional average treatment effect relies on the assumptions of ignorability according to Athey, G. Imbens, et al. 2017. Unbiasedness is ensured by randomness in treatment assignments. The assumption of unconfoundedness states that the potential outcome is independent of the treatment assignment conditioned on the set of covariates. The unconfoundedness assumption is formulated as follows:

$$[Y_i(control), Y_i(treatment)] \quad \perp \quad W_i|X_i$$

In addition, the common support assumption guarantees that there exists a hypothetical counterfactual for each treatment assignment. In other words, this assumption ensures that there is sufficient overlap in the characteristics of treated and untreated individuals to find adequate matches. Consequently, the propensity score which is the probability of an individual being treated, becomes bounded. The following mathematical notation represents the common support assumption:

$$\forall x \qquad 0 < P(W_i = 1|X_i = x) < 1$$

When both assumptions are satisfied, the treatment assignment is said to be strongly ignorable.

5.3 Implementation of Causal Forests

Causal Forests is a Machine Learning technique that fits in the generalized Random Forests framework. The Causal Forests proposed by Athey, Tibshirani, Wager, et al. 2019 is the fundamental building block of this research. Causal Forests are an extension to the Random Forests proposed by Breiman 2001 that are based on decision trees. The Causal Forest algorithm contains four fundamental aspects:

1. Nearest neighbour mechanisms are used to make forest predictions instead of averaging.
2. The honest tree approach is used to reduce bias and allows for valid inference.
3. The split criteria is based on capturing patterns of heterogeneity instead of minimizing predictive error.
4. Gradient-based approximations allow for efficient parameter estimation.

Athey and G. W. Imbens 2015 determined that, under the assumption of unconfoundedness and common support, parameter estimates are consistent and asymptotically distributed which enhances interpretability of the parameter estimates. The Causal Forest (CF) allows for computation of individual and Conditional Average Treatment effects (CATE) and also allows for heterogeneity to be estimated and tested. Furthermore, variable importance can be measured and segmentation can be made based on CATE levels to allow for comparison between groups.
Comparable to Random Forest (RF), a Causal Forest (CF) generates decision trees that are partitioned in a set of leaves. Bootstrapping is used to take sub-samples of the data in order to create different decision trees dependent on prespecified branching criteria. The branching criteria for causal forests are to serve the objective of maximizing heterogeneity between branches. Afterwards, treatment effects are predicted by nearest neighbour mechanisms using the output of the different decision trees.
The advantage of a Causal Forest is that it allows for asymptotically consistent and Gaussian estimates under certain conditions. Furthermore, an advantage of Causal Forests is that in contrast to regression models, exogeneity of covariates is not required. The main function of Causal Forests is to exploit patterns of heterogeneity in customers and it allows us to assess the treatment effect for different customer segments. Moreover, generalized random forests methods capture non-linearity due to the nature of the branching algorithm. Therefore, the non-linear effect of different levels of crowdedness can be captured more pragmatically.

5.3.1 Algorithm

The Causal Forest algorithm that is proposed by Athey, Tibshirani, Wager, et al. 2019 stems from the Generalized Random Forests (GRF). The main concept is to construct multiple decision trees based on bootstrapping and obtaining estimates by averaging all decision trees. Each iteration the decision tree is extended by recursive partitioning of the data points based on the covariates. Each leaf intends to maximize the objective function by classifying the category of subsamples.
The Causal Forests consist in total of B decision trees and $L_b(x)$ denotes the leaf of the $b-th$ tree. Athey and Wager 2019 introduced the following generalized procedure to insure that Causal Forests can effectively capture heterogeneity.

1. Firstly, for each tree $b = 1, \ldots, B$ a subsample $S_b \subseteq \{1, \ldots, n\}$ is drawn based on bootstrapping.
2. Secondly, each subsample S_b is recursively split and the splits are added to form leaves to grow a decision tree for $b = 1, \ldots, B$.
3. Thirdly, predictions are computed using the following formula:

$$\hat{\mu}(x) = E(Y_i|X_i = x) = \sum_{i=1}^{n} \alpha_i(x) Y_i \qquad \alpha_i(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{1}(\{X_i \in L_b(x), i \in S_b\})}{|i : X_i \in L_b(x), i \in S_b|}$$

The utilization of the "out-of-bag" prediction allows for causal inferences. In the notation, this is denoted by the superscript $^{(-i)}$. In practice, the estimation of the treatment effect is exclusive to the tree for which the observation was not incorporated in the subsample of the tree. This entails that all decision trees that did not incorporate a specific observation during the training are identified and these trees are exclusively used to calculate the treatment effect of this specific observation. Furthermore, Athey,

Tibshirani, Wager, et al. 2019 constructed an adaptive kernel method that calculates the conditional average treatment effect. The formula is as follows:

$$\hat{\tau}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} \alpha_i(x)(Y_i - \hat{m}^{(-i)}(X_i))(W_i - \hat{e}^{(-i)}(X_i))}{\frac{1}{n}\sum_{i=1}^{n} \alpha_i(x)(W_i - \hat{e}^{(-i)}(X_i))^2}$$

In the formula $e(x) = P[W_i|X_i = x]$ donates the propensity score and $m(x) = E[Y_i|X_i = x]$ denotes the expected outcome of the response Y_i marginalizing over the treatment. The conditional average treatment effect is based on the "out-of-bag" predictions. Moreover, cross-validation is utilized to optimize the "R-learner" objective function by minimizing "out-of-bag" estimates. Under additional assumption described by Wager and Athey 2018 the conditional average treatment effect is consistent and normally distributed such that:

$$\frac{\hat{\tau}(x) - \tau(x)}{\sqrt{Var(\hat{\tau}(x))}} \rightarrow \ N(0,1)$$

In order to establish unbiasedness in tree predictions, "honest forests" are trained. Honesty in this framework means that two different samples are used for partitioning and estimating the parameter as described by Wager and Athey 2018. In this procedure, treatment effects are estimated utilizing different subsamples to construct the trees. Initially, the tree is trained with the subsample and afterwards, a new subsample runs through the tree until all the observations are contained in a leaf node. Thus, each of the leaf nodes is populated by the new sample.

5.3.2 Heterogeneity: measurements and testing

The objective of causal forests is to capture the potential heterogeneous effects of the treatment. The treatment effect of each explanatory variable is assessed by splitting the data. The data split is determined by the median and in the case of binary dependent variables like the weekend and peak-hour variable, the split is determined by their binary value. Afterwards, the Conditional Average Treatment Effects (CATE) are distinctively assessed. The asymptotic properties of the effects allow for testing the difference in heterogeneous treatment effects is above the median in contrast below the median. The *grf-package* developed by Athey, Tibshirani, Wager, et al. 2019 provides the implementation of Causal Forests in R-software. It enables the relevant coefficients to be calculated and allows for the visualization of trees. The prediction function calculates the conditional average treatment effect based on individual treatment effects of the forest. Additionally, the variance of the treatment effects is given so that the significance of the treatment effect is tested. Chernozhukov, Demirer, et al. 2018 introduced a test for heterogeneity that fits the average treatment effect as a linear function of the "out-of-bag" predictions. The two synthetic predictors are composed as denoted by Athey and Wager 2019 in the following way:

$$C_i = \bar{\tau}(W_i - \hat{e}^{(-i)}(X_i)$$

$$D_i = (\hat{\tau}^{(-i)}(X_i) - \bar{\tau})(W_i - \hat{e}^{(-i)}(X_i)$$

The estimated "out-of-bag" of treatment effects are denoted by $\hat{\tau}^{(-i)}(X_i)$ and $\bar{\tau}$ denotes the average of this effect. Subsequently, $Y_i - \hat{m}^{(-i)}(X_i)$ is regressed on these predictors such that the coefficients have the following interpretation. The coefficient of the C_i represents the absorption of the average treatment

effect. Hence, it suggests that the mean forest prediction is correct if *mean forest prediction* statistic is significantly different from zero and close to one. The coefficient of D_i represents the quality of treatment heterogeneity estimates. Thus, the probability for the presence of heterogeneity is calculated with *differential forest prediction* statistic, which serves as an omnibus test for the presence of heterogeneity. The null of no heterogeneity is rejected if the coefficient is significantly greater than zero and close to one. Furthermore, the *grf-package* also provides a function that measures the importance of the covariates in capturing the presence of heterogeneity.

5.3.3 Cross-validation

The performance of a forest is dependent on several training parameters: the core tree-growing options are minimal node size, sample fraction and number of variables tried for each split. The parameters that control honesty behaviour are the honesty fraction and honesty prune leaves. In addition, the split balance parameter, alpha and imbalance penalty regulate the splitting of the trees.
The Generalized Random Forest package provides a cross-validation procedure to determine optimal parameter values in training. The following cross-validation procedure is implemented as developed by Athey, Tibshirani, Wager, et al. 2019:

- Firstly, as possible parameter values several random points in the parameter space are drawn. By default, there are a hundred distinctive sets of parameter values to be chosen from. The Causal Forest is trained for each set of parameter values so that the "out-of-bag" error is computed.
- Only the trained forests composed of a small number of trees are implemented so that the tuning of parameters is computationally manageable. The "out-of-bag" error will provide biased estimates of the final forest error if such a small number of trees is used. Therefore, the errors are debiased through variance decomposition.
- In the framework of treatment effect estimation, the concept of an error is not as straightforward in contrast to regression forests. The measurement of errors in Causal Forest developed by Nie and Wager 2021 is motivated by residual-on-residual regression proposed by Robinson 1988.
- Finally, a smoothing function determines the optimal parameter values after obtaining the debiased error estimates for each set. The fundamental takeaway is that the optimal parameters are designed to maximize heterogeneity instead of minimizing prediction error.

Tuning can be sensitive. This means that a selection of the tuning parameters, for example, all honest parameters may give better performance than trying to find the optimum of all parameters. Note that for relative smaller data sets, an increasing number of tuned trees may result in more stable results without necessarily increasing training time drastically.

5.4 Determining optimal thresholds

In the paper Andrews et al. 2016 it was stated that the effects of crowdedness are rather non-linear and negative before a lower threshold and after an upper threshold. Yet, in the model that was implemented for the context of underground public transport, only a moderated interval for crowdedness is considered.

Additionally, the effects of a lower threshold were examined by performing a logistic regression for crowd densities under two square meters. However, it is of utmost importance to justify the decision for implementing certain thresholds. In the extension of this research, it prioritizes to detect these thresholds and afterwards, explore the differences in treatment effect for the three intervals that result from these thresholds. The optimal thresholds (c_l, c_h) are computed by splitting the data set into three parts: low, medium and high crowdedness intervals. For each of these distinct data sets, the Causal Forest algorithm (CF) is implemented which provides the Conditional Average Treatment Effect (CATE) for each data set. The goal is to find the threshold levels that maximize the goodness-of-fit of the model. Therefore, the *mean forest prediction* of the calibration test is assessed. If the estimates are approximately one and significantly greater than zero, the predictions of the three Causal Forests are correct and representative of the data. The following algorithm introduces mathematical notation to the threshold detection. It is

Algorithm 1 Threshold detection algorithm

$threshold\_detection(X, \delta)$ ▷ input: data and increment
$(c, \alpha_c, \beta_c) \leftarrow (0, min(x^{crowd}), min(x^{crowd}) + \delta)$
  while $\beta_c \leq max(x^{crowd})$
    $X_c = \bigcup_{i=1}^{n} x_i \mid x_i^{crowd} \in (\alpha_c, \beta_c]$
    $causal\_forest(X_c)$ ▷ implement Causal Forest
    if $\Pr(m.f.p.) > 0.05$ **or** *size error* ▷ mean forest prediction p-value
      $\beta_c \leftarrow \beta_c + \delta$
    else
      $(c, \alpha_c, \beta_c) \leftarrow (c + 1, \beta_c, \beta_c + \delta)$
return α

an iterative process in which the crowdedness interval is incremented with a value δ until the mean forest prediction is significantly greater than zero and the threshold value is set. From the new threshold level on the crowdedness is incremented until the maximum crowdedness level is reached. It is important to note that the Causal Forests only perform if the crowdedness interval contains enough observations to make forest predictions.

5.5 Critical points of the Causal Forest

The disadvantage of most Machine Learning Algorithms is that they require a significant amount of data to provide valid results. However, with a limited sample size of only 6441 observations, it is to be questioned to which extent the Causal Forest provides useful results with an emphasise on not over-fitting the data. As a result, parameters that regulate the tree depth and the splitting criteria need to manipulated in such a way that the tree complexity is reduced. Another drawback of the Causal Forests is that there are no information criteria that allow for efficient comparison between this algorithm and the logistic regression model.

Under certain conditions that are described by Athey and Wager 2019 the estimators of the Causal Forests are consistent and interpretable due to the asymptotic distribution. Consequently, hypothesis testing, the determination of confidence intervals and comparing the output with other statistical methods is straightforward. Nevertheless, some of the assumptions that provide these asymptotic distributions are unquantifiable such as the extent to which the unconfoundedness assumption holds. Furthermore, Wager and Athey 2018 proposes that the conditional mean functions for the treatment effect are required to be

Lipschitz continuous which is not proven to be valid.

6 Results

6.1 Replication Results

Firstly, the logistic regression is performed as the replication part of the research. The provided data was split into two parts, one for which the crowdedness level was below two commuters per square meter and one that was above two commuters per square meter. The logistic regression was performed on both data sets and on the combined data. The following table summarizes the results:

Table 3: Logit regression output for different crowdedness intervals (threshold of two commuters per square meter)

	full crowdedness interval		interval above threshold		interval below threshold	
	estimate	standard error	estimate	standard error	estimate	standard error
intercept	-4.861*	0.640	-5.004*	0.945	-3.648*	1.099
crowdedness	0.150*	0.052	0.105	0.121	-0.209	0.339
weekend	-0.030	0.175	0.062	0.225	-0.300	0.396
peek	-0.043	0.156	-0.147	0.232	0.540	0.470
log(arpu)	0.303	0.168	0.350	0.206	0.238	0.288
log(mou)	-0.056	0.094	-0.032	0.114	-0.110	0.169
log(sms)	0.034	0.105	0.011	0.130	0.087	0.180
log(gprs)	-0.015	0.034	0.003	0.042	-0.065	0.061

* : significantly different from null at 1% percentage point

The results suggest that the effect of crowdedness is significantly greater than zero with an estimate of 0.150 for the full interval. All other variables except for the intercept are insignificant in determining the purchase likelihood according to the t-test.

6.2 Simulation Results

It is important to note that the results shown are dependent on the outcomes of the sampling methods and the realizations of random variables. The purchase rate is 5.5% given the parameter values that are listed in table 4. These are reasonable numbers in the context of our research as higher levels of crowdedness are included and treatments are introduced. In the data generating process, the beta coefficients of the logistic regression output are chosen such that the effects of explanatory variables in the control group are equivalent to that the situation in which no treatment is apparent. The lambdas and alpha are tuned in such a way the purchase rate of circa 5% is respected, whilst providing diversity in the purchase outcomes between control and treatment group. Most importantly, the parameters are chosen such that the logistic regression output gives similar results to what is presented in the paper by Andrews et al. 2016.

Table 4: Parameter values for the Data Generating Process

λ_{crowd}	$\lambda_{weekend}$	λ_{peek}	λ_{arpu}	λ_{mou}	λ_{sms}	λ_{gprs}
0.27	-0.06	-0.09	0.2	-0.1	0.1	-0.05
α	$\beta_{weekend}$	β_{peek}	β_{arpu}	β_{mou}	β_{sms}	β_{gprs}
-4.3	-0.03	-0.045	0.3	-0.055	0.035	-0.015

Table 5 gives a broad overview of purchases per cluster for the treatment and control group. The generated process leads to higher purchase rates for the treatment group that varies across clusters.

Table 5: Purchases of control and treatment group assessed for each cluster

	control				treatment			
cluster	1	2	3	4	1	2	3	4
non-purchase	718	836	835	710	674	777	797	737
purchase	37	40	38	39	49	68	50	36

Table 6 shows the cluster characteristics with the cluster means of each contextual and profile variable. In correspondence with table 5 it shows that the clusters provide sufficient variety in characteristics and purchase outcomes. The k-means algorithm provides groups that are representable in number and show different characteristics for example looking at weekend and peek variables. In cluster four the purchase rate for the control group is higher than the purchase rate of the treatment group. This gives an indication that treatment heterogeneity is apparent.

Table 6: Cluster means of characteristics

cluster	weekend	peek	arpu	mou	sms	gprs
1	0.000	0.140	3.445	5.256	5.423	8.264
2	0.130	0.000	4.395	6.611	5.960	9.596
3	0.940	0.348	3.666	5.043	4.765	9.056
4	0.177	1.000	4.160	6.389	5.739	9.212

In the following table presents the regression output of the simulation for the crowdedness interval of above 2 and below 5.37 commuters per square meter. The motivation behind this is to examine the extent to which the simulation corresponds to the regression output that is presented by Andrews et al. 2016.

Table 7: Logit regression output of the simulated data for the crowdedness interval [2 , 5.36]

	estimate	standard error
intercept	-5.473*	0.839
crowdedness	0.321*	0.095
treatment	0.110	0.178
weekend	0.064	0.207
peek	-0.280	0.194
log(arpu)	0.518	0.207
log(mou)	-0.069	0.116
log(sms)	0.020	0.131
log(gprs)	-0.017	0.043

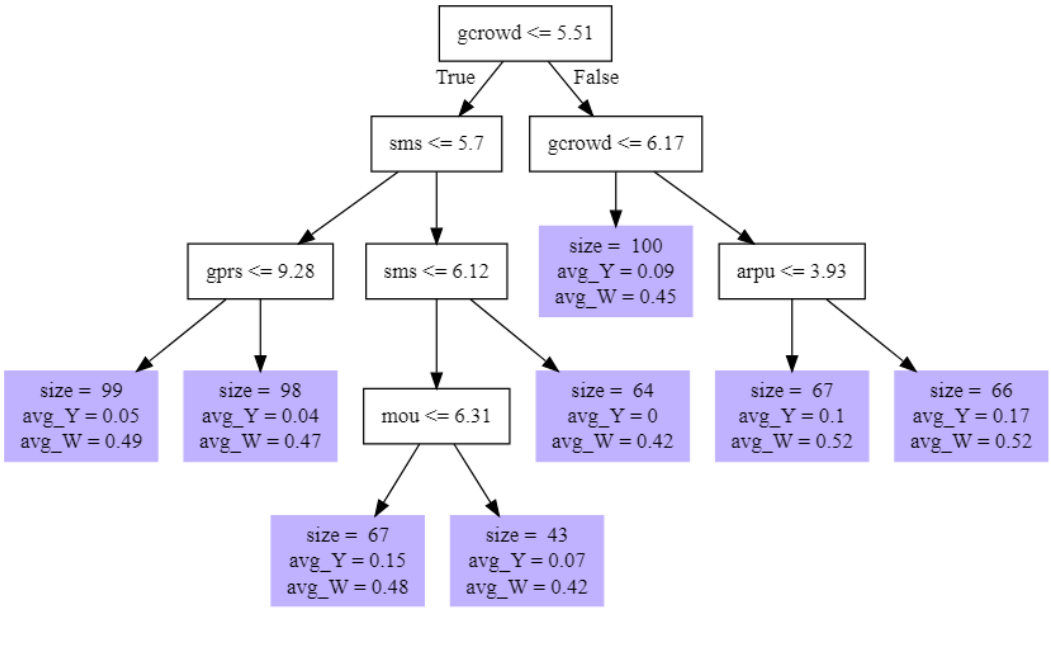
*: significantly different from null at 1% percentage point

The results in Table 7 show representable approximations of the parameter values for the crowdedness and treatment variables with respectively 0.306 and 0.130. Furthermore, only the intercept and the crowdedness variable are significant and all other explanatory variables are not which is in correspondence with the logistic regression output of the article.

6.3 Extension Results

The Causal Forest algorithm is implemented for three distinct levels of crowdedness. Thus, the simulated data set is split into three distinct data sets based on a lower and upper threshold. Figure 2 illustrates a visual example of a Causal Tree for the medium levels of crowdedness. The Causal Forests contains in total 2000 of these Causal Trees. In order to maximize heterogeneity whilst not over-fitting the data, it was required for some parameters to be tuned and others to be fixed. The minimal node size, which is the target for the minimum number of observations in each tree leaf, was set to at least one per cent of the observations in the split data set. This resulted in leaf nodes that were not over-fitted and therefore more representable. The sample fraction of the data used to build each tree was by default set to a half. Other parameters, including honesty parameters, were tuned using cross-validation in order to maximize heterogeneity between the leaf nodes.

Figure 2: Visual representation of a Causal Tree



The optimal thresholds levels (c_l, c_h) were determined based on the quality of the *mean forest prediction* which is the indicator variable for the goodness-of-fit. It is a measure for the "absorption" of the average treatment effect. The estimated coefficient of the mean forest prediction suggests that the predictions in our model are correct if it is approximately one and significantly bigger than zero. Due to the sampling method that is used in the algorithm parameter estimates differ greatly for different intervals. The lower threshold proved to be $c_l = 3.185$ and the optimal upper threshold proved to be $c_h = 7.050$.

- Y_i^l is defined as the dependent variable for the lower interval $crowd \in [0.500, 3.185)$.
- Y_i^m is defined as the dependent variable for the medium interval $crowd \in [3.185, 7.050]$.
- Y_i^h is defined as the dependent variable for the high interval $crowd \in (7.050, 11]$.

The identification of intervals allows for the computation of the Conditional Average Treatment Effects. Table 8 provides the estimates and standard deviations of the CATE for each interval. Additionally, the treatment effects of all observations, alongside the control and treatment groups are given. On a significance level of five per cent non of the treatment effects were found to be significantly greater than zero. Nevertheless, the results implicate that the effect of a sudden variation in crowdedness caused by a train delay has the greatest effect for medium levels of crowdedness. Whereas, the effect of train delays for the lower and higher levels of crowdedness is significantly lower compared to the medium levels of crowdedness. This is in correspondence with the theory that simulated the non-linearity of the crowdedness parameter.

Table 8: Conditional Average Treatment Effects for different crowdedness levels and groups

		Y_i^l		Y_i^m		Y_i^h	
		estimate	st. dev.	estimate	st. dev.	estimate	st. dev.
CATE	*all*	0.0151	0.0091	0.0185	0.0115	0.0135	0.0075
	control	0.0150	0.0092	0.0176	0.0115	0.0125	0.0076
	treated	0.0151	0.0092	0.0196	0.0116	0.0144	0.0076

Additionally, the results of the calibration test are considered in Table 9. It is shown that all the *mean forest prediction* parameters are significant at a five per cent significance level. This suggests that the forest predictions are correctly specified. In other words, the p-value supports the hypothesis that the means of the "out-of-bag" predictions of the Causal Forest are adequate. By construction, the threshold levels were tuned in order to establish coefficients that sufficed this calibration test as best as possible. However, the *differential forest prediction* estimates that indicate the quality of treatment heterogeneity estimates, prove not to be significantly greater than zero. The corresponding p-value suggests that heterogeneity is either insufficiently captured by the model or that the customer heterogeneity wasn't apparent in the data. Consequently, the null hypothesis of no heterogeneity is not rejected and Causal Forest treatment effects are partially relevant predictors.

Table 9: Calibration test of the Causal Forests

	Y_i^l		Y_i^m		Y_i^h	
	estimate	st. dev.	estimate	st. dev.	estimate	st. dev.
mean forest prediction	1.065*	0.638	0.953*	0.665	0.975*	0.567
differential forest prediction	-2.435	1.260	-3.585	-2.951	0.192	0.742

In order to measure which explanatory variables capture the heterogeneity in the model, variable importance is considered. The summary of the output is given in Table 10, which ranks the importance of explanatory variables based on the percentage of heterogeneity that is captured. If the results from the lower interval are considered, it can be shown that the heterogeneity is mainly captured by the profile variables. All of which contribute to a comparable percentage of the explanation. This suggests that it is optimal to the differentiate customers based on their characteristics when crowdedness levels are relatively low. In contrast to the low interval, crowdedness plays a more important role in consumer heterogeneity for medium and high intervals. The profile variables prove to be superior in explaining heterogeneity compared to the contextual factors of the weekend and peak-hour. This indicates that consumer decisions in this context is primarily determined by customer characteristics conditioned on crowdedness levels.

Table 10: Variable Importance in capturing consumer heterogeneity

	Y_i^l		Y_i^m		Y_i^h	
	importance	rank	importance	rank	importance	rank
crowdedness	17,0%	4	16,6%	2	16,2%	2
weekend	5,9%	7	11,1%	7	9,8%	7
peek	8,2%	6	10,6%	6	11,9%	6
log(arpu)	17,6%	3	14,0%	5	15,7%	3
log(mou)	15,2%	5	15,2%	4	15,4%	4
log(sms)	17,9%	2	15,2%	3	16,4%	1
log(gprs)	18,2%	1	17,3%	1	14,5%	5

Lastly, the treatment effect of all the explanatory variables is assessed by splitting the data in two intervals. The data split is determined by the median, then the results are distinctively assessed for variable values below and above the median. In the case of binary dependent variables like the weekend and peak-hour variables, the split is determined by their binary value. For each variable two distinct groups are obtained for which the average treatment effect is calculated. Hence, the difference in heterogeneous treatment effects is computed by subtracting the CATE above the median from the CATE below the median. The difference is indicated by $\Delta CATE$ and significance levels can be obtained due to the asymptotic properties of Conditional Treatment Effects.

Table 11: The heterogenous effects of the explanatory variables

$\Delta CATE$	Y_i^l		Y_i^m		Y_i^h	
	estimate	st. dev.	estimate	st. dev.	estimate	st. dev.
crowdedness	0.0162	0.0254	0.0393	0.0323	-0.0305	0.0200
weekend	-0.0250	0.255	-0.0000	0.0344	-0.0040	0.0221
peek	-0.0314	0.262	-0.0164	0.0327	-0.0399	0.0219
log(arpu)	-0.0216	0.259	-0.000	0.0328	-0.0024	0.0213
log(mou)	0.0058	0.0257	0.0012	0.0326	0.0109	0.1178
log(sms)	-0.0127	0.0261	-0.0050	0.0327	0.0074	0.0214
log(gprs)	-0.0148	0.0258	-0.0060	0.0327	-0.0077	0.2136

Table 11 presents evidence that the non-linearity of crowdedness effects is captured. The results show that the train delays initially have positive effects. Especially for moderated levels of crowdedness, the effect was more than twice as big compared to the lower levels. Note that for higher levels of crowdedness a train delay negatively influences purchases which corresponds to the notion of congestion.

7 Conclusion

Effectively applying marketing strategies such as hyper-contextual targeting is of great importance for companies in the future. This research focuses on contextual factors that explain the purchasing behaviour of individual customers. The effect of additional crowdedness is assessed by logistic regression and the Machine Learning algorithm of the Causal Forest. The logistic regression is limited in capturing customer heterogeneity and to identify possible endogeneity, more nuanced methods are required. The objective of this research is to extend the statistical framework of Andrews et al. 2016 by tackling the limitations of the implemented methods. The following research question is the fundamental building block of the data analysis: *"How can the heterogeneous effects of sudden variations in crowdedness on purchase behaviour be examined, taking into account customer heterogeneity and the non-linear effect of crowdedness?"*.

Starting with the first sub-question: *"How is it possible to simulate customer heterogeneity and the non-linear effects of crowdedness on purchase likelihood?"*. As no data-set was provided on the effects of train delays on purchases, the initial question becomes how to simulate an artificial data-set that mimics reality and incorporates the features of customer heterogeneity and non-linear crowdedness effects. Therefore, the domain of crowdedness was extended to levels that are common in Asian metropolises and a non-linear function for the crowdedness effect was constructed. The representability of the artificial data was checked by performing the logistic regression and comparing the properties of the parameter to the output as presented in the paper by Andrews et al. 2016.

In addition, the resulting sub-question becomes: *"How is it possible to quantify heterogeneous effects of sudden variations in crowdedness and which variables are most significant in explaining customer heterogeneity?"*. The average treatment framework is introduced to quantify the effect of a treatment versus a non-treatment. However, the realization of outcomes can not be obtained simultaneously. Therefore, a potential outcome framework is introduced to be able to calculate treatment effects. An efficient Machine Learning algorithm that is developed by Athey, G. Imbens, et al. 2017 suits the goal of quantifying conditional average treatment effects by maximizing customer heterogeneity. The critique to the methods proposed by Athey, G. Imbens, et al. 2017 is that customer heterogeneity wasn't considered explicitly to identify the potential self-selection problem. The results indeed suggest that the extent to which profile and contextual variables capture customer heterogeneity depends on the crowdedness levels.

The final sub-question is: *"How can the thresholds for which the effects of crowdedness are most reliable be obtained and what is the extent to which customer heterogeneity is captured?"*. Additionally, the algorithm was implemented to formulate a systematic approach of finding crowdedness intervals that optimize model performance. The threshold levels were chosen such that the predictive accuracy of the Causal Forest was significant. Furthermore, Causal Forests allows for the determination of the extent to which heterogeneity is present and which factors are most influential to capture this heterogeneity.

In conclusion, the preference goes to the Causal Forest algorithm in contrast to logistic regression because this method reflects purchase responses more accurately. The emphasis on maximizing customer heterogeneity allows for a more nuanced picture of crowdedness effects and the role of consumer characteristics in their purchaser.

Future research might consist of companies applying market segmentation based on treatment effects. In

this way, marketers can target customers that are most likely to respond to a certain advertisement by linking personal characteristics with environmental factors. The digitalization of purchase processes by incorporating mobile technology as QR-codes or apps, are useful ways for collecting customer data and boosting sales. Marketers should exploit the human tendency towards crowdedness and utilize specific targeting to realize optimal results. There is a promising potential for public transport providers to facilitate companies in their marketing campaigns. As an example, Dutch public service providers integrated infrared sensors in trains that can send mobile messages to boarding passengers. These technological innovations can not only improve commuting efficiency but also be utilized for marketing purposes. If such hyper-contextual indicators indeed influence consumer behaviour, information on crowdedness levels, train delays and other interventions could be valuable for consumer targeting. Consequently, public transport organisations can profit from these advertisements by implementing technological innovations in public services and vehicles. Thus, it is recommended to companies to accumulate data on consumer behaviour by digitalizing the consumer processes of attaining products. Also, it is advisable to public transport provider to digitalize their travelling services.

References

[1] Michelle Andrews, Xueming Luo, Zheng Fang, and Anindya Ghose. "Mobile ad effectiveness: Hyper-contextual targeting with crowdedness". In: *Marketing Science* 35.2 (2016), pp. 218–233. URL: https://pubsonline.informs.org/doi/10.1287/mksc.2015.0905.

[2] Susan Athey, Guido Imbens, Thai Pham, and Stefan Wager. "Estimating average treatment effects: Supplementary analyses and remaining challenges". In: *American Economic Review* 107.5 (2017), pp. 278–81.

[3] Susan Athey and Guido W Imbens. "Machine learning methods for estimating heterogeneous causal effects". In: *stat* 1050.5 (2015), pp. 1–26.

[4] Susan Athey, Julie Tibshirani, Stefan Wager, et al. "Generalized random forests". In: *Annals of Statistics* 47.2 (2019), pp. 1148–1178. URL: https://grf-labs.github.io/grf/REFERENCE.html.

[5] Susan Athey and Stefan Wager. "Estimating treatment effects with causal forests: An application". In: *Observational Studies* 5.2 (2019), pp. 37–51.

[6] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[7] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. "Double/debiased/neyman machine learning of treatment effects". In: *American Economic Review* 107.5 (2017), pp. 261–65.

[8] Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. *Generic machine learning inference on heterogenous treatment effects in randomized experiments*. Tech. rep. National Bureau of Economic Research, 2018.

[9] Yogesh K Dwivedi, Elvira Ismagilova, Nripendra P Rana, and Ramakrishnan Raman. "Social media adoption, usage and impact in business-to-business (B2B) context: A state-of-the-art literature review". In: *Information Systems Frontiers* (2021), pp. 1–23.

[10] Agence France-Presse. "Hong Kong metro seats may be scrapped for smartphone space. Agence France-Presse". In: (2014).

[11] Jinyong Hahn. "On the role of the propensity score in efficient semiparametric estimation of average treatment effects". In: *Econometrica* (1998), pp. 315–331.

[12] Christiaan Heij, Christiaan Heij, Paul de Boer, Philip Hans Franses, Teun Kloek, Herman K van Dijk, et al. "Econometric methods with applications in business and economics". In: (2004).

[13] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[14] Tom Erik Julsrud and Jon Martin Denstadli. "Smartphones, travel time-use, and attitudes to public transport services. Insights from an explorative study of urban dwellers in two Norwegian cities". In: *International Journal of Sustainable Transportation* 11.8 (2017), pp. 602–610.

[15] Robert J LaLonde. "Evaluating the econometric evaluations of training programs with experimental data". In: *The American economic review* (1986), pp. 604–620.

[16] Ahreum Maeng, Robin J Tanner, and Dilip Soman. "Conservative when crowded: Social crowding and consumer choice". In: *Journal of Marketing Research* 50.6 (2013), pp. 739–752.

[17] Xinkun Nie and Stefan Wager. "Quasi-oracle estimation of heterogeneous treatment effects". In: *Biometrika* 108.2 (2021), pp. 299–319.

[18] David Pollard. "Strong consistency of k-means clustering". In: *The Annals of Statistics* (1981), pp. 135–140.

[19] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data". In: *Journal of the american statistical association* 90.429 (1995), pp. 106–121.

[20] Peter M Robinson. "Root-N-consistent semiparametric regression". In: *Econometrica: Journal of the Econometric Society* (1988), pp. 931–954.

[21] Paul R Rosenbaum and Donald B Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1 (1983), pp. 41–55.

[22] Stefan Wager and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1228–1242.

[23] Amali Wijekoon, Sandeep Salunke, and Gerard A Athaide. "Customer heterogeneity and innovation-based competitive strategy: A review, synthesis, and research agenda". In: *Journal of Product Innovation Management* (2021).

A Appendix

R-code: Algorithm for Transaction based segmentation

```
#Bachelor thesis Econometrics
#Quantitative Marketing and Business Analytics

#Abdelmounaim el Yaakoubi [50979ay]


#Install Packages
install.packages("tidyverse")
library(tidyverse)
install.packages("grf")
library("grf")
install.packages("DiagrammeRsvg")
library("DiagrammeRsvg")
install.packages("DiagrammeR")
library("DiagrammeR")
install.packages("NMOF")
library("NMOF")
install.packages("matrixStats")
library("matrixStats")
install.packages("haven")
library("haven")
install.packages("glm")
library("glm")
install.packages("contrib.url")
library("contrib.url")



#Import the datasets 'Dataset1' and 'Data2' from the zip-file povided by INFROMS
Dataset1 = read_sav("C:/Users/Downloads/Dataset1.sav")
Data2 = read_sav("C:/Users/Downloads/Data2.sav")


#Data Analysis

#Cleaning Data
Data2$v3[which(Data2$v3 == 2)] = 0
datas = rbind(Data2, Dataset1)
datas = na.omit(datas)
datas = datas[,c(3:6,11:14)]
colnames(datas)[1] = 'weekend'
colnames(datas)[2] = 'peek'
colnames(datas)[3] = 'crowd'
colnames(datas)[4] = 'purchase'
colnames(datas)[5] = 'arpu'
colnames(datas)[6] = 'mou'
colnames(datas)[7] = 'sms'
```

```
colnames(datas)[8] = 'gprs'
summary(datas)


#Indentification Heterogenious groups
set.seed(1)
clusterdatas = as.matrix(datas[,c(1:2,5:8)])
cluster = kmeans(scale(clusterdatas), 4)
centers = cluster$centers
datas[,9] = cluster$cluster
colnames(datas)[9] = 'group'

#Determining clusters means
rcenters = as.data.frame(matrix(0,4,6))
for(g in 1:4){
  rcenters[g,] = colMeans(as.matrix(clusterdatas[which(datas$group == g),]))
}
colnames(rcenters)[1] = 'weekend'
colnames(rcenters)[2] = 'peek'
colnames(rcenters)[3] = 'arpu'
colnames(rcenters)[4] = 'mou'
colnames(rcenters)[5] = 'sms'
colnames(rcenters)[6] = 'gprs'
rcenters

#Implementation of Logit
lgt = glm(purchase ~ crowd + weekend + peek +arpu + mou + sms + gprs
          ,data = datas[which(datas$crowd>=2),], family = "binomial")
lgt = glm(purchase ~ crowd + weekend + peek +arpu + mou + sms + gprs
          ,data = datas[which(datas$crowd<2),], family = "binomial")
lgt = glm(purchase ~ crowd + weekend + peek +arpu + mou + sms + gprs
          ,data = datas, family = "binomial")
summary(lgt)



#Simulation
set.seed(90)

datas[,10] = runif(6441,0.5,11)
colnames(datas)[10] = 'gcrowd'

datas[,11] = rbinom(6441, 1, 0.5)
colnames(datas)[11] = 'treatment'

nonlin = function(x){
  nl = -.18*cos(0.57*x)
  return(nl)
}

utility = numeric(6441)
indiv = numeric(6441)
```

```
for(i in 1:6441){
  for(t in 0:1){
    indiv[i] =0.27*nonlin(datas$gcrowd[i] + 0.72)
            -0.06*datas$weekend[i] -0.09*datas$peek[i] + 0.2*datas$arpu[i]
                  + -0.1*datas$mou[i] + 0.1*datas$sms[i] - 0.05*datas$gprs

    utility[i] = -4.3 + indiv[i]*as.numeric(datas$treatment[i] == t)
          + nonlin(datas$gcrowd[i])*datas$gcrowd[i] +
              -0.03*datas$weekend[i] -0.045*datas$peek[i] + 0.3*datas$arpu[i]
                  -0.055*datas$mou[i] + 0.035*datas$sms[i] + -0.015*datas$gprs[i]
  }
}

#Expected and realized purchase rate
psm = exp(utility)/(1+exp(utility))
mean(psm)

newpurchase = numeric(6441)
for(i in 1:6441){
  datas$newpurchase[i] = rbinom(1,1,psm[i])
}
colnames(datas)[13] = 'newpurchase'
sum(datas$newpurchase)/6441

#Assesment of simulation quality
table(datas$newpurchase, datas$group, datas$treatment)

lgt = glm(newpurchase ~ gcrowd + treatment + weekend + peek +arpu + mou + sms + gprs
          ,data = datas[which(datas$gcrowd>2 &datas$gcrowd<=5.37 ),], family = "
              binomial")
summary(lgt)



#Causal forests
set.seed(8)

#Causal Forest: low interval
Xl = datas[which(datas$gcrowd<3.185),]
Yl = Xl$newpurchase
Wl = Xl$treatment
Xl = Xl[,c(1:2,5:8,10)]
l = nrow(Xl)
tau.forest_l = causal_forest(Xl, Yl, Wl,min.node.size = floor(0.01*l),
  tune.parameters = c( "mtry", "honesty.fraction", "honesty.prune.leaves", "alpha"))

tree_l = get_tree(tau.forest_l,2000)
plot(tree_l)

average_treatment_effect(tau.forest_l, target.sample = "all")
average_treatment_effect(tau.forest_l, target.sample = "control")
```

```r
average_treatment_effect(tau.forest_l, target.sample = "treated")

variable_importance(tau.forest_l)
test_calibration(tau.forest_l)


#Causal Forest: medium interval
Xm = datas[which(datas$gcrowd>=3.185 & datas$gcrowd<=7.05),]
Ym = Xm$newpurchase
Wm = Xm$treatment
Xm = Xm[,c(1:2,5:8,10)]
m = nrow(Xm)
tau.forest_m = causal_forest(Xm, Ym, Wm, min.node.size = floor(0.01*m),
  tune.parameters = c( "mtry", "honesty.fraction", "honesty.prune.leaves", "alpha"))

tree_m = get_tree(tau.forest_m,2000)
plot(tree_m)

average_treatment_effect(tau.forest_m, target.sample = "all")
average_treatment_effect(tau.forest_m, target.sample = "control")
average_treatment_effect(tau.forest_m, target.sample = "treated")

variable_importance(tau.forest_m)
test_calibration(tau.forest_m)


#Causal Forest: high interval
Xh = datas[which(datas$gcrowd>7.05),]
Yh = Xh$newpurchase
Wh = Xh$treatment
Xh = Xh[,c(1:2,5:8,10)]
h = nrow(Xh)
tau.forest_h = causal_forest(Xh, Yh, Wh, min.node.size = floor(0.01*h),
  tune.parameters = c( "mtry", "honesty.fraction", "honesty.prune.leaves", "alpha"))

tree_h = get_tree(tau.forest_h,2000)
plot(tree_h)

average_treatment_effect(tau.forest_h, target.sample = "all")
average_treatment_effect(tau.forest_h, target.sample = "control")
average_treatment_effect(tau.forest_h, target.sample = "treated")

variable_importance(tau.forest_h)
test_calibration(tau.forest_h)
```