LIU-IDA/STAT-A–19/014–SE

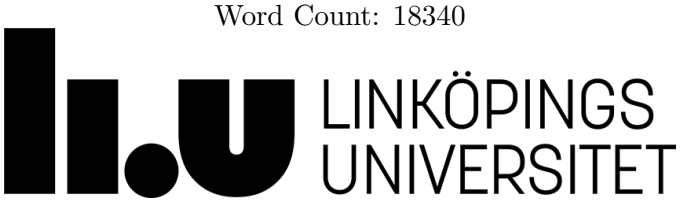
Master Thesis in Statistics and Machine Learning

Uplift Modeling: Identifying Optimal Treatment Group Allocation and Whom to Contact to Maximize Return on Investment

Henrik Karlsson

Spring 2019

Word Count: 18340



Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Examiner
Krzysztof Bartoszek

Supervisor
Linda Wänström

Industry Supervisors
Kim Wedenberg and Elin Magnusson

Abstract

This report investigates the possibilities to model the causal effect of treatment within the insurance domain to increase return on investment of sales through telemarketing. In order to capture the causal effect, two or more subgroups are required where one group receives control treatment. Two different uplift models model the causal effect of treatment, Class Transformation Method, and Modeling Uplift Directly with Random Forests. Both methods are evaluated by the Qini curve and the Qini coefficient. To model the causal effect of treatment, the comparison with a control group is a necessity. The report attempts to find the optimal treatment group allocation in order to maximize the precision in the difference between the treatment group and the control group. Further, the report provides a rule of thumb that ensure that the control group is of sufficient size to be able to model the causal effect. *If* has provided the data material used to model uplift and it consists of approximately 630000 customer interactions and 60 features. The total uplift in the data set, the difference in purchase rate between the treatment group and control group, is approximately 3%. Uplift by random forest with a Euclidean distance splitting criterion that tries to maximize the distributional divergence between treatment group and control group performs best, which captures 15% of the theoretical best model. The same model manages to capture 77% of the total amount of purchases in the treatment group by only giving treatment to half of the treatment group. With the purchase rates in the data set, the optimal treatment group allocation is approximately 58%-70%, but the study could be performed with as much as approximately 97% treatment group allocation.

Keywords: Causal Effect, Uplift Modeling, Class Transformation Method, Model Uplift Directly, Random Forest, XGBoost, Qini Curve, Qini Coefficient, Optimal Control Group Allocation

Acknowledgments

I want to thank my supervisor Linda Wänström, who have helped me and guided me through the thesis. Her help and guidance have been essential to me in order to complete this report.

I would also like to thank my supervisors at If, Kim Wedenberg and Elin Magnusson for their help and enormous patience to answer all my questions over and over. Tack If.

Many thanks to Erika Anderskär, my opponent that helped me identifying several aspects that should be clarified to improve and increase the reader understanding.

I would like to thank Björn Benzler, who has helped me reading through the thesis and given me valuable feedback.

Last but not least, I would like to show my gratitude to my partner Ebba Vidén that have proofread the thesis and continuously helped me to stay motivated.

Key Concepts

Treatment
Treatment is what is given to change an individuals behavior. In this report, treatment is defined as getting a phone call from *If*.

Treatment group
Every individual who is given treatment is assigned to treatment group.

Control group
Every individual in control group has received control treatment, which is no treatment in this report. The purpose of the control group is to be a reference to the treatment group. In other words, control treatment is no treatment, which is in no phone call from *If* in this report.

Uplift
A measure to capture the causal effect, or in other words, the difference in behavior between the treatment group and control group. It is measured by subtracting the outcome of treatment group by the outcome of the control group.

Uplift score
The output from the uplift model. It is used to rank individuals in the order that they should be given treatment to maximize the uplift.

Campaign
In this report, a campaign is considered being a selection method for individuals that have the possibility of receiving treatment. In a campaign, individuals are called by *If*, where they assure that the individual has insurances that cover her needs.

Purchase
In this report, a purchase is when an individual has bought any insurance during the campaign. Therefore a purchase in the treatment group can be performed during the actual phone call or up to approximately 30 days after the call. Purchase within the control group is a purchase when the individual in question is selected for a campaign but have not received treatment.

Contents

1. Introduction

1.1 Background

The purpose of marketing is to make people aware of your brand and ultimately influence peoples' purchasing behavior. In a world where people are constantly exposed to marketing, the competition to sell is fierce. Marketing is costly, and therefore, marketing departments continuously try to evaluate and optimize the marginal effect of their marketing efforts (Kotler and Armstrong, 2010). Where and how should a company spend its marketing budget in order to best influence the behavior of potential customers? The purpose of this project is to evaluate and maximize the causal effect from a marketing treatment compared to giving no treatment.

This report is written in collaboration with *If*, where *If* has provided the objective and the data. *If* is one of the major insurance companies in the Nordic region with more than 3.6 million customers across Scandinavia and the Baltic's. They provide insurances for both private and commercial customers in a wide range of categories.

If is today using several marketing channels to stay in contact with their customers, were telemarketing is one of them. Telemarketing enables *If* to talk to their customers to ensure that their insurance need is fulfilled and potentially get the possibility to sell additional insurances. *If* have categorized all customers into strategic segments which are used in the process of selecting whom to contact (give treatment). All customers in the strategic segment are then contacted in a random order for a fixed period of time, which is equivalent of the campaign length.

Historically, *If* have used purchase rate (proportion of sales in the targeted customer sample) as one of their primary key measures to evaluate how well a marketing campaign was. Now, *If* is interested in taking this further by evaluating and model the causal effect of treatment. The causal effect of a treatment is the change of behavior caused by the treatment. The gain of knowing the causal effect is twofold, firstly it provides a more accurate measure on marginal effect of their marketing and secondly, it enables *If* to identify customers that are more likely to be persuaded to purchase due to receiving treatment before giving it.

In order to measure the causal effect, a control group is required, whose members have not been exposed to the treatment. The effect of treatment would then be the difference in outcome after a certain amount of time between the treatment group and control group, which will be referred to as causal effect or uplift.

Insurances are not a consumable product and therefore, are the purchase rates in the data material low. It is hard to model small changes in a data set, and a slight change is of high interest. Further, it is a challenge to conclude that a difference in purchase

rates between the treatment group and control group is due to natural variation in the data material or if it is an actual difference (Wang and Chow, 2014). To investigate if an identified difference is an actual difference, requires an extensive data material, as the uncertainty in the proportions reduces as the size of the data material increases. This study has access to an extensive data set of approximately 630,000 customer interactions, of which approximately 50,000 belong to the control group.

By modeling which customers who would have a positive response because of treatment, enables *If* to distribute treatments more efficiently than random selection within the treatment group. Today, treatment is distributed by random selection in the sense that each individual that has been selected for treatment have the same probability of being treated (called by *If*); thus there is no prioritization or order of whom to treat within the treatment group. By ranking individuals who are more likely to purchase **because**, of treatment and providing treatment in that order would, in theory, increase earnings simultaneously as marketing costs are reduced. Because individuals that are convinced to purchase because of treatment is targeted to a greater extent, meanwhile individuals who would have purchased without treatment no longer receive treatment but continues to purchase. Therefore, the total number of treatments given can be reduced without affecting the sales, in the best of worlds (Radclifte and Simpson, 2008).

It is also of interest for the individual that companies have an efficient methodology for identifying individuals who are more likely to purchase. If they can identify who is more likely to purchase, it implies that they also know who are less likely to purchase, so each individual can be exposed to fewer commercials that are not of interest.

An uplift model tries to predict which individuals that will have a positive outcome from treatment before giving treatment. Uplift is a general methodology that can be applied to several types of problems with different data sources. E.g., uplift can be used to identify who is more likely to purchase because of treatment (like in this report) or to identify who is most likely to churn (quit being a customer) and could be retained if treatment is given. Any data containing customer information on individual level combined with group belonging, where at least one group received some form of treatment, and the outcome is captured, can be used to train an uplift model. Uplift models have also been used within the medical field, to discover groups of patients where a treatment are most beneficial (Jaśkowski and Jaroszewicz, 2012).

1.2 Objective

The objective of this project is to model the causal effect of treatment from compared to giving no treatment, i.e., the uplift from a marketing campaign. This report will evaluate two different methodologies to model uplift by using data provided by *If*. In order to measure uplift, it is a necessity to have a control group of sufficient size. Objective 1 is to develop uplift models with a binary target feature that capture the causal effect of treatment. Objective 2 of this report is to provide a recommendation of what would be an

optimal treatment group allocation given a certain uplift methodology and classification algorithm without considering the cost of giving treatment.

1.3 Ethical Consideration

By being a customer, the individual has accepted that the data is used within *If* to improve their services to their customers. The data used in this report is anonymized before it has been given to the author. No part of the analysis is performed on subgroups so small that there is a risk of being able to identify a single individual. Moreover, the data contain as little individual information as possible to ensure that no one can be identified. For instance, only the year of birth instead of birth date is kept in the data set to reduce the chance of identifying a single individual. Further, if an individual is older than 80 years, the data is stored as 80+ years. As individuals grow older, the fewer amount of individuals share the same age and could potentially identify a person due to the extraordinary age. Be adjusting the birth year to 80+ makes this impossible.

If is in general careful with how personal information is handled and who has access to it. For more information on how the data is being used, see *If*'s web page (If, 2019).

2. Theory

2.1 Causal Inference

The causal effect of a treatment given a time interval t_1 to t_2, is the difference in outcome at t_2 given that a unit would have been exposed to the treatment initiated at time t_1 compared to if the unit would have been exposed to the control treatment initiated at time t_1 (Rubin, 1974, Rubin and Waterman, 2006). When the two treatments are mutually exclusive, and the experiment cannot be repeated, it is of interest to evaluate the outcome in the treatment group and the control group. The difference in outcome would then be the causal effect of the treatment. The causal effect is defined as:

$$\tau_i = Y_i(1) - Y_i(0) \tag{2.1}$$

where $Y_i(1)$ denotes the response of individual i, that have been exposed to treatment and $Y_i(0)$ denotes the response of individual i, that have been exposed to control treatment.

Since a single individual i, cannot be subject to both treatment and control treatment, the true treatment effect for an individual, can never be observed. This is called *the fundamental problem of causal inference* (Holland, 1986, p.947). As a consequence, general supervised learning algorithm (where the outcome for all possible treatments is known for each subject in the training data) cannot be applied. In order to try to capture the treatment effect, individuals in the treatment group and control group are compared. For this comparison to be valid, it is of importance that these two groups are as similar as possible. The expected causal effect of treatment is estimated by Conditional Average Treatment Effect (CATE), which is defined as:

$$CATE := \tau(X_i) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i] = \mathbb{E}[Y_i(1)|X_i] - \mathbb{E}[Y_i(0)|X_i] \tag{2.2}$$

where X_i denotes a vector of features (Athey and Imbens, 2016).

By defining an indicator variable, $W_i \in \{0, 1\}$, evaluating to 1 if the individual i belongs to treatment group and 0 if the individual i belongs to control group, then the observed outcome becomes:

$$Y_i^{obs} = W_i Y_i(1) + (1 - W_i) Y_i(0) \tag{2.3}$$

where Y_i^{obs} is the observed outcome for individual i. In the case of a binary target feature, such as a purchase or not purchase, Y_i^{obs} would evaluate to 0 or 1.
Additionally, if the assignment of group belonging is random conditional on X_i, the Conditional Independence Assumption (CIA or unconfoundedness) holds.

$$CIA : \{Y_i(1), Y_i(0)\} \perp W_i|X_i \tag{2.4}$$

When the CIA assumption holds, it implies that there are no confounding features that affect the assignment of group belonging. As mentioned, it is of importance that the

treatment group and control group is as similar as possible, but for the CIA to hold it also requires that there are no unmeasured features that block or cause the causality between the target and the features, which is true if the treatment assignment is random conditional on all features. When the CIA assumption holds, CATE can be estimated by equation 2.5.

$$\mathbb{E}(CATE) = \mathbb{E}[Y_i^{obs}|X_i = x, W_i = 1] - \mathbb{E}[Y_i^{obs}|X_i = x, W_i = 0] \tag{2.5}$$

Uplift modeling aims to model CATE. For an overview of causal inference, see Gutierrez and Gérardy (2017) and for a more detailed description, see Rubin (1974), Morgan and Winship (2015) and Athey and Imbens (2015).

This report estimate causal effect by Rubin (1974) work with propensity scores and potential outcomes. Another possible method to estimate causal inference is by the work of Pearl (2009), which uses graphs theory to model causality. Both methodologies investigate "what-if" scenarios. Within the potential outcome framework, the "what-if" is usually mentioned as *counterfactuals* while the "what-if" is known as *antecedent* within the causal graphs framework (Morgan and Winship, 2015, Pearl, Glymour, and Jewell, 2016).

2.2 Uplift Modeling

Uplift modeling started to show up in the literature a few years before the millennium shift, and have had many names since then. What today is called uplift models have been called Differential response analysis, incremental value modeling, and true lift models. Uplift models try to model second-order phenomena as it tries to model the conditional average treatment of two or more different mutually exclusive groups. It is achieved by subtracting the outcome from the respective treatment group by the outcome from the control group. The literature have consistently used Rubin (1974)'s framework of modeling causality by the help of propensity scores and have on top of that used a wide variety of algorithms to model uplift.

The explicit goal in uplift modeling is to model the conditional average treatment effect, which is measured as the difference between the treatment group and the control group. The conditional average treatment effect or uplift, estimate the **increase** of purchase probability given that a customer receives treatment compared to if no treatment is given (Radcliffe and Surry, 2011). By being able to identify which customers who are more likely to purchase before treatment is given, would ideally let a company target a smaller part of the sample and thus reduce marketing costs meanwhile they maintain or even increase their earnings (Siegel, 2011).

The uplift model returns a score for each customer, where a higher score means a higher chance of positive outcome. This score should be seen as a priority list of whom to

give treatment first (Naranjo, 2012). The score is then used to partition the individuals into segments for the treatment group and control group and the uplift is computed per segment. For a more detailed description of how to evaluate the result of a uplift model, see section 4.2.

When the entire sample have received an uplift score, the task for the marketing department is to find the optimal proportion of the population to give treatment to maximize profit. What is considered as the optimal proportion depends on the specific campaign, e.g. the cost of giving treatment, if there are fixed staffing costs, costs of having too few subjects to call and the risk of acquiring a negative effect. Further, the report is limited to uplift models with a binary response feature, which only model the action *to purchase* or *not purchase*.

There are four possible outcomes in a binary uplift model; first, the *persuadeables*, customers that would be convinced to purchase because they received treatment. They are the optimal customers to target since the response changes from no purchase to a purchase when treatment is given. Secondly, the *sure things*, customers that would purchase the product with or without treatment and the opposite third group, the *lost causes*, customers who would not purchase the product with or without treatment. Both the *sure things* and the *lost causes* are considered as a waste of money to give treatment to, because the treatment will not affect their response. Finally, the fourth outcome, the *sleeping dogs* which are customers who are convinced not to purchase when they receive treatment but would have purchased if no treatment were given. For *sleeping dogs*, the treatment has the opposite effect than intended and the customer is lost (Siegel, 2011, Rzepakowski and Jaroszewicz, 2012b). Table 2.1 shows a summary of the different outcomes. Keep in mind that an individual cannot be in both treatment group and control group and therefore is it impossible to state in which of the four region an individual belong, only the column or row is known.

			Control group	
			Purchased?	
			Yes	No
Treatment group	Purchased?	No	*Sleeping dogs*	*Lost causes*
		Yes	*Sure things*	*Persuadeables*

Table 2.1: Four archetypes of Uplift

3. Data

3.1 Raw Data and Features

The data consist of approximately 630,000 customer interactions with approximately 60 features collected from *Ifs* telemarketing system from their Swedish branch. The total purchase rate is approximately 4%, purchase rate in the treatment group is approximately 5% and purchase rate in control group 2%. The purchase rate is computed by n/N, where n is the number of purchases, and N is the sample size. The control group is approximately 8% of the data set. The total uplift in the data is therefore approximately, $0.05 - 0.02 = 0.03$.
The features in the data are summarized in the list below:

- Purchase - (Binary) - Indicator if a customer purchased within the campaign window (30 days) from treatment. This is the feature that is modeled.
- Control group indicator - (Binary) - Indicator if the customer is in the treatment group or control group
- Basic demographic variables - Gender (Categorical) and birth year (Date)
- Total number of active insurances at *If* (Numeric)
- Insurance groups - (numeric) - The number of insurances the customer has in each branch of insurances that *If* offer; Motor, Property and Personal insurances.
- Insurance types household level - (Numeric) - number of insurances the customers household have, e.g., number of Car, Child, Pet, and Travel insurances.
- Last product bought - (Categorical) - indicating what was the last insurance the customer purchased
- Date for last product bought - (Date)
- Payment method - (Binary) - indicating whether the customer pays with electronic payments or not
- Payment frequency - (Numeric) - how often the customer is billed in months
- Last interaction - (Date) - Date for last interaction with the customer per channel, phone, email, and text message.
- Opened email - (Numeric) - an indicator of how many emails the customer has opened and in the past 90 days
- Clicked in email - (Numeric) - an indicator of how many links in emails the customer has clicked on the past 90 days

- Received Telemarketing - (Numeric) - how many phone calls the customer has received past 540 days.
- Answered Telemarketing - (Numeric) - how many phone calls the customer has answered past 540 days.
- Inbound calls - (Numeric) - the number of calls the customer have made to *If* past 90 days
- Last product bought household - (Categorical)
- Date for last product bought household - (Date)
- Anchor Date - (Date) - The date the customer was selected for the campaign. Also, the date all other customer features were extracted from the database

3.1.1 Selection Method for Campaigns

An example of a typical campaign targets "all middle-aged customers who have a Villa insurance but no car insurance and have been a customer between 2 to 5 years". The individuals that match the condition and are not in quarantine are selected for the campaign and have the chance of receiving treatment. The data consist of 113 different campaigns that have been run continuously between January 2017 to November 2018. As a consequence, all individuals have not been exposed to an identical campaign. However, the procedure during the call is identically disregarding the specific campaign. The customer is asked several questions to ensure that her insurance needs are fulfilled disregarding what the specific campaign offer was and then the campaign offer is given, e.g, "We can see that you have a villa insurance but no car insurance, do you or your family have a car and is it insured properly?". A campaign can, therefore, be considered more like a selection process of whom to target rather than a campaign in the more traditional context, where an item is sold with a special offer for a limited time. As a consequence, a purchase of *any* insurance within the campaigns time frame is considered being because of the campaign. Since the campaign itself is not trying to sell specific insurance and that the methodology across campaigns is identical, it can be argued that the effect of treatment is captured disregarding the purpose of the specific campaign which enables campaigns to be aggregated into a single data set.

If have defined treatment as receiving a phone call from them. By receiving the phone call, the customer is classified as *used* and is assigned to the treatment group. If the customer answers the phone call, she is classified as *contacted* and remain in the treatment group. As a consequence, there are customers in the treatment group that has not answered the call from *If*. This introduces noise in the data, as it is reasonable to assume that customers that are *used* behaves differently than customers that are *contacted*. Customers that are *used* (did not answer the call), might even behave more similar to customers in the control group compared to *contacted* customers in the treatment group. Approximately, one-third of the customers in treatment group is only *used* and

not *contacted*. The data set in this report is, unfortunately, missing information whether an individual in the treatment group is *used* or *contacted*.

Marketers at *If* design campaigns for different strategic target groups. Once a campaign is launched, the relevant customers are selected and put in quarantine. Being in quarantine locks the customers to the campaign so they cannot be selected for any other campaign or marketing activity for a fixed period of time. From the selection date, the date when the customer was selected to a campaign, the actions of selected customers are being tracked for 30 days. For each day the campaign runs, some the of selected customers are **randomly** given treatment (receive a phone call). When a customer has received a phone call, disregarding if the customer answered or not, the customer is considered *used*. Once the customer is *used*, she is put to the treatment group, and the actions of the customer are tracked for the following 30 days. If the customer purchases any insurance within this time window, it is considered to be a purchase because of the treatment. If the customer is called (*used*) and answered the call, the customer is considered being *contacted* and will not receive any further calls. If the customer is selected to the campaign but not used (never received a phone call), the customer is put to control group, for further discussion see section 3.1.2. Figure 3.1 shows an made up example of the campaign selection funnel.

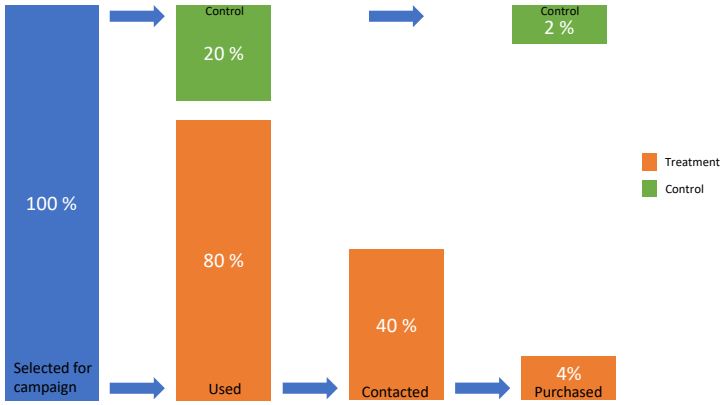


Figure 3.1: Example of the campaign selection funnel with made up numbers

From the example in figure 3.1, all individuals are selected to the campaign. From the sample, individuals are called randomly, which transforms the individual to *used* and assigns her to the treatment group. In the example, 20 % the campaign sample is *unused* and therefore is assigned to the control group. Half of all individuals that have been *used*

answered the phone call and then became *contacted*. For a more extensive discussion, see section 6.1.5.

3.1.2 The Control Group

In this project, the control group consist of customers that have been selected for the campaign but have not been *used* (called). The selection of whom to call is random, so no bias is introduced in the selection of whom to put to control group. The reason why a part of the selected customers of a campaign is not used is that the call center does not want to risk that they run out of customers to call during the campaign; therefore a larger amount of customers than what is needed is selected. As a consequence of this, the proportion of the control group differs between campaigns. Also, by not setting aside a true control group, there is a risk that a customer purchases insurance meanwhile she is in the campaign but before she is *used*, which would be considered as a purchase within the control group. If the same customer then receives a call during the campaign, suddenly she is no longer in the control group instead the treatment group and the earlier purchase does not count, because it was performed before she was *used*. So, by not having a true control group, the unlikely but still possible event that a control group purchase is turned into a treatment group non-purchase could happen. This would introduce errors in the data.

Part of the objective of this report is to find how much control group that is needed in order to evaluate uplift efficiently. The data contain a control group, but it is not by definition a true control group that is set aside before the experiment is performed, rather it just consists of the leftovers that were not randomly selected for treatment. However, in this project, the control group is used as if it were a true control group. It can be argued that the objective of this report still can be reached even though the control group is not true. The previously mentioned risk of control group purchases that is turned to treatment group non-purchases is considered being so small that the data material is still valid for giving recommendation on what would be considered sufficiently large control group. However, the frequency of this potential error is unknown.

3.2 Data Cleaning and Preprocessing

All date features have been recalculated as the number of months since the anchor date except feature birth year, which has been transformed to the number of years from the anchor date.

Before training each model, the data have been split up into two sets. The first set, called *train*, contain 75% of the data and have been used to train all models. The other set, called *test*, contain the last 25% of the data and have been used to evaluate how well the models performed. The models have never "seen" the test data before the model

evaluation. Once the model has evaluated the test data, the model has not been re-trained again to improve the score. The assignment of each data point to respective set have been performed randomly with seed 0.

3.2.1 Missing Values

Some of the features in the data set contain missing values. The number of missing values per feature and how they have been handled can be seen in table 3.1.

Table 3.1: Amount of Missing Values and Imputation method

Feature	**Number of Missing Values**	**Proportion of data set**	**Reason**	**Imputation Method**
Gender	65	0.001 %	Unknown	Removed from data set
Age	23	0.0004%	Unknown	Removed from data set
Months since last Telemarketing Offer	276,880	43.7%	No telemarketing offer have been given	Imputed with 36
Months since last email	220,175	34.8%	No email have been sent	Imputed with 36
Months since last text message	310,636	49.1%	No mail have been sent	Imputed with 36

Respondents who have not received a telemarketing offer, email och text message have been imputed with a value of 36, which is the equivalent of three years without contact. According to *If*, it is reasonable to assume that customers that have not received offers within the last three years have similar behavior as the customer who never received an offer.

4. Method

4.1 Three Different Ways to Model Uplift

Uplift can be modeled by three different methods, two model approach, class transformation method or model uplift directly. The first method does not strictly model uplift, as it builds two separate models, one for treatment group and one for control group and uplift is estimated by subtraction the result of these two models. The two last models are built on top of Rubin (1974) causality work, but estimate uplift differently. Modeling uplift directly with tree-based method such as decision trees and random forests have been and still is the dominating method in literature, and have been used by a wide range of authors (Radcliffe and Surry, 1999, Chickering and Heckerman, 2000, Lo, 2002, Hansotia and Rukstales, 2002, Radcliffe, 2007, Jaśkowski and Jaroszewicz, 2012, Rzepakowski and Jaroszewicz, 2012a, Rzepakowski and Jaroszewicz, 2012b, Jaroszewicz and Rzepakowski, 2014, Guelman, Guillén, and Pérez-Marín, 2015, Wager and Athey, 2017).

Although the dominance of the tree-based methods to estimate uplift, there exist some alternative methods such as k-nearest neighbor (Jaroszewicz and Rzepakowski, 2014) and support vector machines (Zaniewicz and Jaroszewicz, 2013) to estimate uplift. Nassif et al. (2012) who have developed a method to improve the diagnosis of breast cancer for women that resembles uplift. Instead of relying on a treatment group and control group, Nassif et al. (2012) have modeled breast cancer with Logical Differential Prediction Bayesian Net, which uses graphs to model the causality.

4.1.1 The Two Model Approach

The two model approach is the simplest method for modeling uplift. The idea is that two models are trained separately to predict the result for the treatment group and the control group. After that, the uplift is computed by subtracting the outcome of those two models or subtracting the coefficients of the two models. This methodology is flawed in the context of predicting uplift since no part of the fitting process tries to capture the actual uplift. If both models predicted the outcome perfectly, it would capture uplift correctly. However, no model performs perfect predictions. Nothing guarantees that the subtraction of the result from two good models (for treatment and control group respectively) would become a good model for uplift, (see Radcliffe, 2007, Siegel, 2011, Rzepakowski and Jaroszewicz, 2012b, Jaśkowski and Jaroszewicz, 2012, Gutierrez and Gérardy, 2017). Empirically, the two model approach has shown to perform badly (Radcliffe and Surry, 2011).

4.1.2 Class Transformation Method

The class transformation method was introduced by Jaśkowski and Jaroszewicz (2012). It transforms the two features, *group belonging* (treatment group and control group) and

purchase into a single transformed feature, Z_i by:

$$Z_i = \begin{cases} 1, & \text{if } W_i = 0 \text{ and } Y_i^{obs} = 0 \\ 1, & \text{if } W_i = 1 \text{ and } Y_i^{obs} = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

where W_i is the group belonging and Y_i^{obs} is the observed outcome for each individual. Because of the *fundamental problem of causal inference*, it is impossible to compare the outcomes for a single individual in both the treatment group and control group. If $W_i = 0$ and $Y_i^{obs} = 0$ then the individual is in control group and did not perform a purchase, therefore the individuals is either a *Lost Causes* or a *Persuadeables*. If $W_i = 1$ and $Y_i^{obs} = 1$, the individual is in treatment group and performed a purchase, and therefore the individual is either a *Sure thing* or a *Persuadeable* (see table 2.1). In both of these outcomes, Z_i evaluates to 1. Then the user can be sure that this group only contain individuals that will not yield a negative effect (no *Sleeping dogs*), and there is no risk of approaching them with treatment.

Jaśkowski and Jaroszewicz (2012) proved that uplift could be modeled with the transformed variable Z_i and equation 4.2 if the outcome is binary and that the treatment group and control group is balanced in size.

$$\tau(X_i) = 2P(Z_i = 1|X_i) - 1 \tag{4.2}$$

Athey and Imbens (2015) have developed the methodology further by constructing an algorithm that manages to estimate uplift with the transformed variable, Y_i^*, even if the treatment and control group is unbalanced as long as the CIA property holds and the outcome is binary.

$$Y_i^* = Y_i^{obs} \cdot \frac{W_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))} \tag{4.3}$$

where $e(X_i) = P(W_i = 1|X_i = x)$ is the propensity score. If the group assignment is completely random, the CIA assumptions hold, and the propensity score becomes constant $e(X_i) = p$ for all x, where p is the probability of being assigned to treatment group. Y_i^* is the uplift score, and if the probability of being assigned to treatment group is 0.5, then a purchase in the treatment group is evaluated to 2, purchase in control group is evaluated as -2 and no purchase in treatment and control group is evaluated to 0. Further, the estimation of uplift can be simplified to

$$\tau(X_i) = \mathbb{E}[Y_i^*|X_i = x] \tag{4.4}$$

This implies that any consistent estimator of $\mathbb{E}[Y_i^*|X_i]$ is also an consistent estimator of $\tau(X_i)$ (Gutierrez and Gérardy, 2017). The class transformation method returns a score, Y_i^* for each individual. This score is the order the individuals should be contacted in order to maximize uplift.

The class transformation method has a strong practical advantage, the transformed class becomes a single target feature and can be modeled by any standard regression algorithm to predict uplift. This enables the user to take advantage of already existing and highly optimized packages to model uplift. The python package *Pylift* by Yi (2018) has been used to model uplift by class transformation method. Another great advantage with the Pylift package is that it is built on top of the python package *Scikit-learn* (Pedregosa et al., 2011), which easily lets the user change which algorithm that is being used to predict the transformed feature. The package also utilizes the possibility to search for optimal hyperparameters with cross-validated hyperparameter grid search and cross-validated randomized search from *Scikit-learn*. This is a great aid for the user when it comes to parameter tuning in the modeling process.

Extreme Gradient Boosting (XGBoost)

In this report, the prediction of the class transformation score has been performed with the Extreme Gradient Boosting algorithm, commonly known as XGBoost. The algorithm is a boosted tree algorithm which learns and remember the parameter between each iteration, which enables the algorithm to learn faster. The algorithm can be used for both classification and regression problems. XGBoost was initially developed by Chen and Guestrin (2016) who had the ambition to build a fully scalable algorithm that would become the new state-of-the-art predictor. When XGBoost is compared to other algorithms, it tends to be up to 10 times faster and has been part of several winning solutions in the later Kaggle competitions. Kaggle is a website that hosts online machine learning competitions for everyone interested.

4.1.3 Model Uplift Directly

As tree-based algorithms are built to divide and evaluate the data into subregions or subpopulations, which makes them useful for modeling differences, such as the difference between the treatment group and control group (Radcliffe and Surry, 2011). Tree-based methods are common in the uplift literature and have been utilized by Hansotia and Rukstales (2002), Radcliffe and Surry (2011), Rzepakowski and Jaroszewicz (2012a), Sołtys, Jaroszewicz, and Rzepakowski (2015), Guelman, Guillén, and Pérez-Marín (2015) and Athey and Imbens (2015) to mention a few.

General Tree-based Methods

Tree-based methods have two main steps, splitting and pruning. The splitting part tries to find the most optimal split to separate the data and generate as *pure* nodes as possible. The pruning removes nodes (or branches) that do not improve the generalization of the tree.

When a node is pure, it implies that all data points in the node should be as similar as possible. Depending on the type of tree algorithm, each node can either be split into two

nodes (such as CART trees) or multiple nodes (such as CHAID trees). The algorithm will grow the tree (continue to split its nodes) until all nodes are completely pure or the stopping criterion is met. If the tree grows until all nodes are entirely pure, the model will classify the training data with full accuracy, but the result would not generalize well to a new data set, and the model would be overfitted. In order for the tree to decide where to do the split, it computes the information gain and the potential split with the highest information gain will be conducted. The information gain is estimated by multiplying the proportion of data that was assigned to the child node with the purity of the node (Quinlan, 1986 and Breiman et al., 1984)

Many tree-based methods, such as CART and Quinlan's C4.5 tree, use a two-step approach where the first phase is to split data into nodes in a top-down fashion. The second phase is pruning, where unhelpful splits are removed. This approach is utilized because the tree methods are highly non-linear and therefore, strongly depend on the interaction between the selected features. For example, a given split can seem meaningless when it is evaluated in the current node, but splits further down in the branch can be of great importance in conjunction with the current split. Since each split is evaluated at the current node disregarding possible future splits further down in the tree, it is of interest to build a deep tree with many splits and thereafter prune away the split that did not contribute enough.

Overfitting can be avoided, either by inserting a stopping criterion that limits how deep the tree can grow, to apply pruning to the tree or both in conjunction. The pruning can be performed before the split based on a significance test or after the tree have finished growing, depending on what type of tree is being used. The latter method is more common (Breiman et al., 1984).

Generally, decision trees that have been allowed to grow deep have low bias and high variance between different trees even though they were trained on identical data sets. It is desirable to build deep trees because a deep tree can classify the data well, but the high variance between different decision tree creates unrobust results. In order to reduce the variance, the random forest algorithm has been developed. It builds many decision trees and averages the results of each tree to achieve a more robust result. This comes with a cost of slightly more bias in the model, but the variance is reduced. The random forest uses bagging and trains each tree with a different subset of features, which helps to average out the effect of deep trees even further (Hastie, Tibshirani, and Friedman, 2008). In this report, the uplift random forest uses the square root of the number of features to train each tree.

Tree-based Methods for Estimating Uplift

When a tree-based algorithm is used to model uplift, the splitting criterion is modified in order to make sure uplift is captured. The literature suggests several methods of how to adjust the splitting criterion to best estimate uplift, and it seems that a consensus in

which method is best has not yet been reached. Hansotia and Rukstales (2002) proposed a splitting criterion which tries to maximize the difference between the differences in treatment group and control group probabilities for the left and right subnodes. Rzepakowski and Jaroszewicz (2012a) introduced the concept of divergence from information theory as a splitting criterion, where a tree-based algorithm capture uplift by trying to maximize the distributional difference between treatment and control group. This report will use the work with statistical divergence from Rzepakowski and Jaroszewicz (2012a) as the splitting criterion.

A distributional divergence is a measure of how much information that is lost when q(x) is used to approximate p(x), where q(x) normally represent sampled data and p(x) is theoretically derived from a distribution. The divergence is the "distance" between two probability distributions, but it is a weaker measure than a metric distance because it is not symmetric and does not have to satisfy the triangular inequality (Chodrow, 2017). When the divergence measure is used as a splitting criterion, it tries to maximize the divergence between the treatment group and the control group in every node. When using an ensemble of trees, the predicted uplift is obtained by averaging the uplift predictions of the individual trees in the ensemble (Guelman, Guillén, and Pérez-Marín, 2015). The four different divergence measures that is used as splitting criterions can be seen in equation 4.5, 4.6, 4.7 and 4.8.

Rzepakowski and Jaroszewicz (2012a), list three criteria's that the splitting criterion should satisfy in order to capture uplift:

1. If the class distributions in the treatment group and control group are the same in all branches, it should evaluate to the minimum value.

2. The value of the splitting criterion should evaluate to zero if a test is statistically independent of the outcomes in both the treatment group and the control group.

3. The splitting criterion should reduce to standard splitting criteria used by decision trees if the size of the control group is zero.

Since uplift is captured by achieving the biggest possible distributional divergence between the treatment group and control group, it is reasonable that the splitting criterion evaluates to the minimum when the two groups are identical, as the first criterion states. The second criterion state that a split that is statistically independent should not be used as a splitting criterion since it would not improve the tree in a general decision tree. However, when modeling uplift, a split can make the distributions more similar than before, and it is, therefore, possible to receive a negative splitting value. This implies that an independent split may not be the worst possible split in a given situation (Rzepakowski and Jaroszewicz, 2012a).

This report will evaluate the results of four different splitting criterions in conjunction with uplift random forests, Kullback-Leibler divergence, Euclidean distance, Chi-square

divergence, and L1-norm divergence. Kullback-Leibler divergence, Euclidean distance, and chi-square divergence were introduced in the uplift literature by Rzepakowski and Jaroszewicz (2012b) and L1-norm divergence was introduced in uplift literature by Guelman, Guillén, and Pérez-Marín (2015).

Kullback-Leibler divergence:

$$KL(P:Q) = \sum_i p_i log \frac{p_i}{q_i} \tag{4.5}$$

Euclidean distance:

$$ED(P:Q) = \sum_i (p_i - q_i)^2 \tag{4.6}$$

χ^2-divergence:

$$\chi^2(P:Q) = \sum_i \frac{(p_i - q_i)^2}{q_i} \tag{4.7}$$

L1-norm divergence:

$$L1(P:Q) = \sum_i |p_i - q_i| \tag{4.8}$$

where the divergence is computed between two distributions $P = (p_1, ..., p_n)$ and $Q = (q_1, ..., q_n)$.

The random forest model returns the conditional probability of purchase given group belonging. The uplift score is computed by equation 4.9. The score should be sorted in descending order, assuming that the individual with the highest score should be the most likely to make a purchase given treatment, the second individual the second most likely to purchase because of treatment and so on. The interpretation of a negative score is that the probability of purchase for that individual is higher if treatment is not given.

$$\text{Uplift score} = P(\text{purchase}|\text{treatment group}) - P(\text{purchase}|\text{control group}) \tag{4.9}$$

One advantage with *modelling uplift directly* is that the method can model uplift with both a binary and continuous target feature (Radcliffe, 2007). The R-package *uplift* by Guelman (2014) has been used to model uplift directly with random forests.

4.2 Evaluation of Uplift Models

The conditional average treatment effect can be estimated in a variety of methods, and all method face the challenge of how to evaluate the result, as uplift models suffer from the fundamental problem of causal inference. There are three different methods to evaluate uplift models in the literature, bins of uplift, uplift curves, and Qini measures. Bins of uplift is limited to evaluating uplift for a single model and cannot be used to compare models. Uplift curve visualizes the cumulative gain from the model and can be used as an aid to decide how big proportion of the treatment group that should be given treatment. From the uplift curve, the Gini coefficient can be computed, which is used as a measure to compare different uplift models. Qini curve is a further development of the uplift curve, which manages to capture a potentially negative effect from individuals in the sample, which give a better view of how the model performs. The Qini coefficient is similar to the Gini coefficient, which enables model comparison. Bins of uplifts is described by Ascarza (2018), uplift curve used by Rzepakowski and Jaroszewicz (2012a), Sołtys, Jaroszewicz, and Rzepakowski (2015), Jaśkowski and Jaroszewicz (2012) and the Qini measures that were developed by Radcliffe (2007). For a comparison of methods, see Naranjo (2012).

4.2.1 Bins of Uplift

Bins of uplift takes the uplift score and sort the individuals in descending order for both the treatment group and control group separately and then split them up into k segments. Usually into ten splits (deciles) but it depends what is most appropriate for the given problem. The bins of uplift methodology assume that all individuals in segment k have the same probability for any given outcome. Uplift is evaluated by subtracting purchase rate in treatment group and control group per segment k, by equation 4.10 (Naranjo, 2012):

$$u_{kp} = \frac{r_k^t}{n_k^t} - \frac{r_k^c}{n_k^c} \tag{4.10}$$

where u_{kp} is the predicted uplift per segment, r_k^t and r_k^c is the number of purchases in treatment group and control group respectively per segment, n_k^t and n_k^c is the number of individuals in treatment group and control group respectively per segment.

The bins of uplift methodology do not provide any metrics to compare different uplift models to each other; rather, it only visualizes the uplift per segment k. An example of bins of uplift can be seen in figure 4.1.

The bins of uplift in figure 4.1 is constructed by an uplift model that has sorted the data according to the model score, split up the data into k segments and estimate the uplift for each segment with equation 4.10. The blue bars represent the predicted uplift for each segment, and the red bars represent the actual uplift per segment in the data set. The first segments manage to capture individuals that are more likely to purchase because

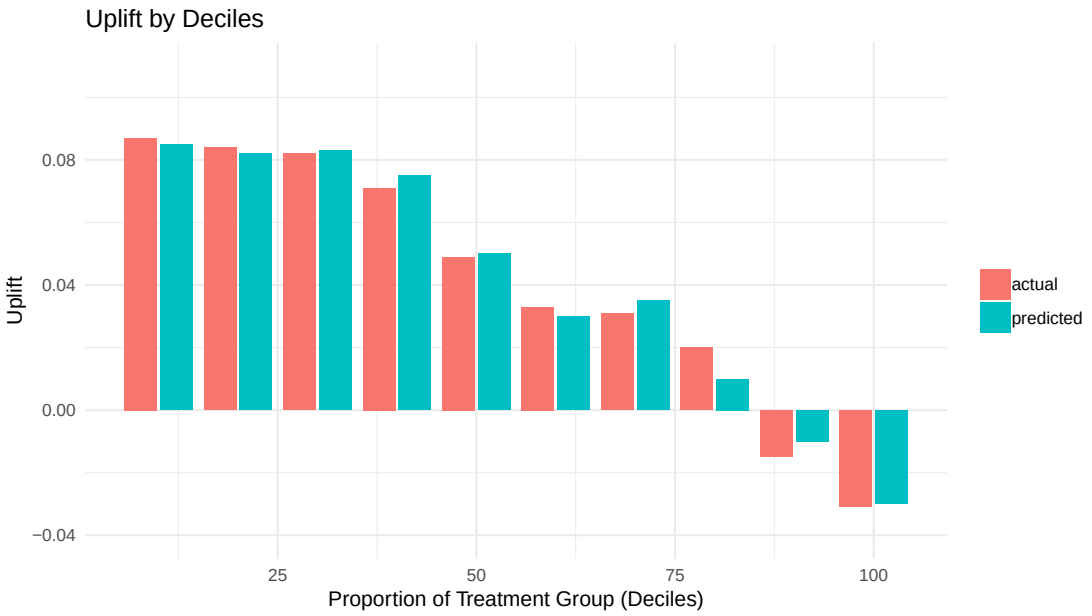


Figure 4.1: Example of Bin of Uplift Graph

of treatment as the two last segment capture a negative effect, which implies that the chance of purchase is higher if no treatment is given to individuals in these segments.

4.2.2 Uplift Curve

The uplift curve requires a model with a binary target feature, and an uplift score for each individual where a higher value implies a higher chance of purchase given treatment. The individuals are sorted by the uplift score in descending order, and cumulative sum of purchase is computed. The uplift curve assumes that the individual with the highest score is contacted first. An example of how to construct the gain chart can be seen in figure 4.2.

Individual	Is purchase	Score given by model	Number of cumulative purchases	Number targeted
x_1	Yes	1.2	1	1
x_2	Yes	1	2	2
x_3	No	0.9	2	3
x_4	Yes	0.8	3	4
x_5	Yes	0.75	4	5
x_6	No	0.7	4	6
x_7	No	0.5	4	7

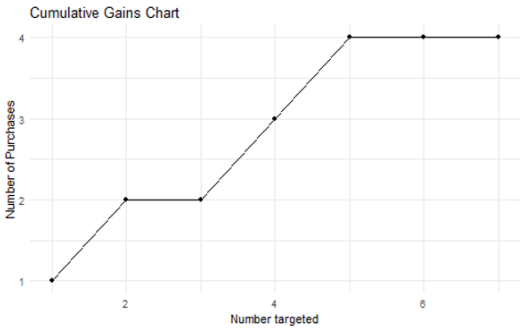


Figure 4.2: Cumulative table of gain and uplift curve

The left side of figure 4.2 shows a table where each row represents an individual. The table is sorted in descending order by the score returned from the model. The right side of figure 4.2 shows how the graph that is constructed from the table on the left-hand side. The number of individuals targeted or the proportion of the treatment group is on the horizontal axis, and the cumulative purchases are visualized on the vertical axis. As the vertical axis only visualize the cumulative sum of a purchase or no purchase, the uplift curve cannot capture a potential negative effect.

In figure 4.3, an example of an uplift curve for a sample with a purchase rate of 5 % can be seen. The graph visualizes the number of purchases as a function of the number of individuals treated. The linear diagonal line, *random*, shows the effect of treatment if the selection of whom to treat within the treatment group is random. If the entire sample is targeted, the 5 % that purchases would be found. In contrast, the *optimal* model, manages to identify all the 5 % of the sample that will purchase because of treatment. Therefore, the curve has a steep increase until all purchasers due to treatment have been identified, then the curve flattens out horizontally since no other individual in the sample will purchase because of treatment. A typical uplift model will be somewhere in between the *random* and *optimal* curves and an example is visualized as *model 1*. The closer *model 1* is to the optimum, the better model. If *model 1* would be below the *random* curve, it implies that the model has the opposite effect, it captures the individuals that would not purchase because of treatment. See section 4.2.3 to see how the optimal curve is computed.

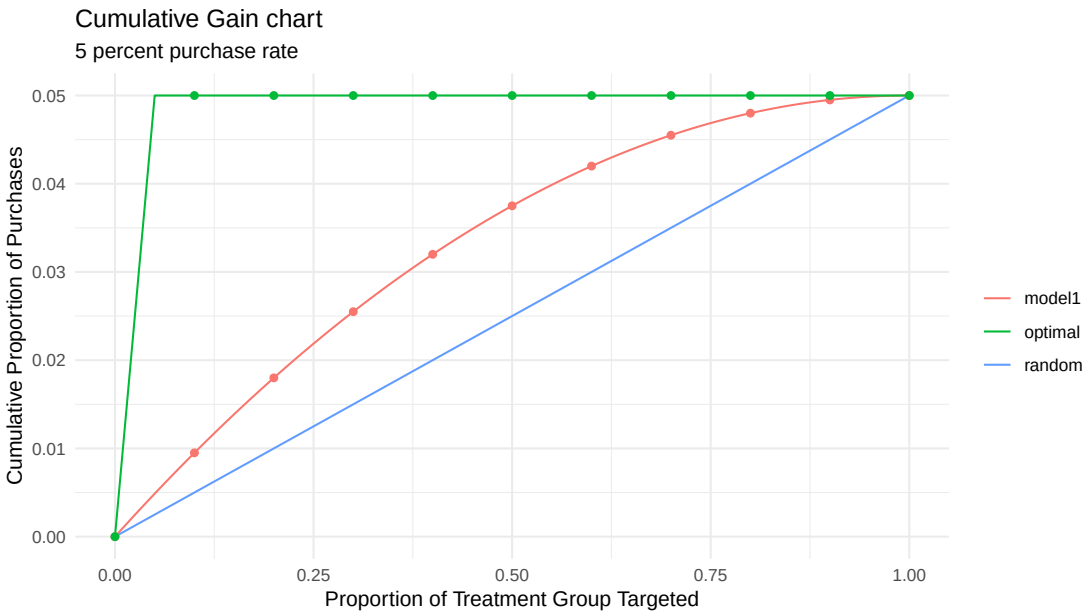


Figure 4.3: Uplift Curve/ Gains Chart

From the gains chart in figure 4.3, the Gini coefficient can be computed, which is a measure to compare how well a model performs. It is a ratio of the area between the *random* diagonal and the actual curve (*model 1*) to the corresponding area above the *random* diagonal and the optimum curve (Radcliffe, 2007). The perfect model has a Gini coefficient of 1, and a model without predictive power has a Gini coefficient of 0. The Gini coefficient is similar to the receiver operating characteristic curve (ROC-curve), with the difference that the horizontal axis in uplift curve plots the number of treated individuals rather than the non-responders as the ROC-curve does.

4.2.3 Qini Measures

The Qini measures, Qini coefficient and Qini curve are a generalization of the uplift curve and the Gini coefficient developed by Radcliffe (2007). The difference between the Qini curve and the gain curve is that the Qini curve plots the *incremental* purchases instead of the cumulative number of purchases. The incremental purchase is computed per segment and group, by

$$u_{kp} = r_k^t - \frac{r_k^c n_k^t}{n_k^c} \tag{4.11}$$

where u_{kp} is the predicted uplift per segment, r_k^t and r_k^c is the number of purchases in treatment and control per segment respectively, and n_k^t and n_k^c is the group sizes per segment for treatment group and control group. The Qini coefficient is computed in the same manner as the Gini coefficient, but it is computed from the Qini curve instead of the uplift curve.

An example of how this computation is performed can be seen in table 4.1. The model returns a score for each individual; the data is sorted in decreasing order and then split into k segments. The number of individuals and the number of purchases in each segment is computed for the treatment group and control group. In example table 4.1, Segment 1 has 11 purchases in the treatment group and 3 purchases in the control group. The cumulative extra purchases for segment 1 is therefore $11 - \frac{3 \cdot 100}{100} = 8$.

Table 4.1: Example of computation of incremental purchase

	Treated		Control		Treated - Control
Segment k	Cumulative purchases (r_k^t)	Cumulative targeted (n_k^t)	Cumulative purchases (n_k^c)	Cumulative targeted (r_k^c)	Cumulative extra purchases (u_{ka})
1 (10%)	11	100	3	100	8
2 (20%)	30	200	10	200	20
3 (30%)	44	300	22	300	22
...	...	...	...	...	...

Worth mentioning is that the uplift estimates are not strictly additive. Therefore, it is usually more accurate to estimate the cumulative uplift at each point from zero rather than accumulating a set of uplifts (Radcliffe and Surry, 2011). The cumulative number of purchases for the treatment group and the control group is subtracted to compute uplift per segment.

The incremental purchase enables the Qini curve to capture the potential negative effect of treatment. An example can be seen in figure 4.4. The graph should be interpreted in the same manner as the uplift chart, the *optimal curve* represents the theoretically

best possible outcome where every individual that can be persuaded by treatment is identified and given treatment first. The diagonal *random* line represent the outcome if all individuals were given treatment in random order. Model 1 and Model 2 are two uplift models, where model 2 outperforms model 1. The entire sample has an uplift of 5 % in this example, but when the entire treatment group receives treatment, both potential *persuadeables* and *sleeping dogs* are given treatment. If the uplift model works as intended, *persuadeables* are given the highest score and the *sleeping dogs* the lowest. By identifying the optimal amount of subjects in the treatment group to give treatment, would in theory, enable a company to treat a smaller proportion of the sample, meanwhile maintaining the earnings (Radcliffe, 2007). The highest possible uplift a model can capture is computed by the total uplift minus the negative effect of treatment. So in this example, the model 2 suggest that approximately 65 % of the treatment group should be given treatment to gain approximately 6.5% uplift.

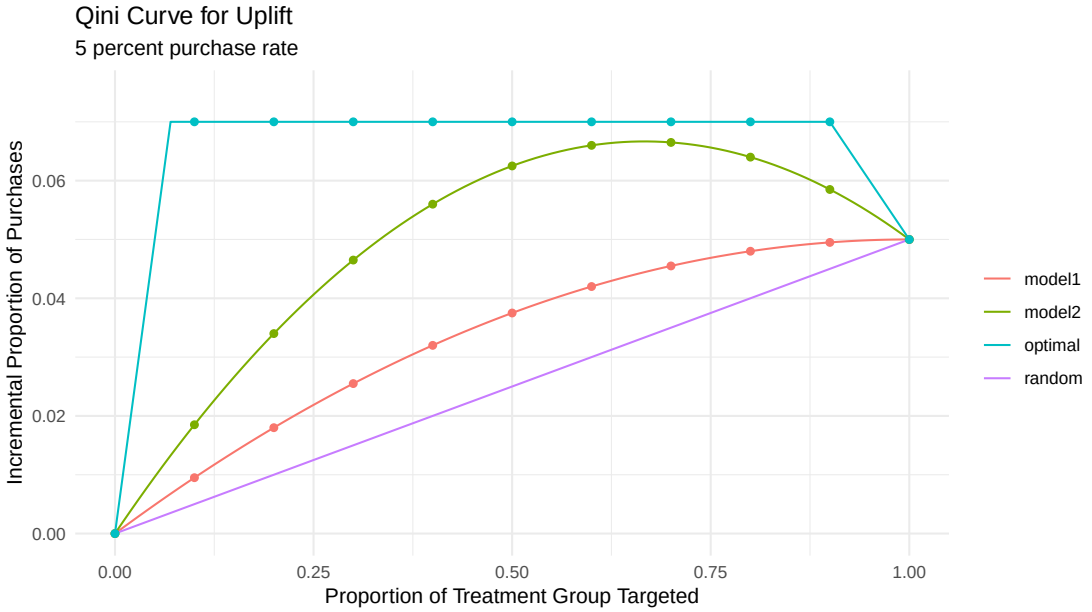


Figure 4.4: Example of Qini curve

When the total proportion of purchases is 5%, it is possible that only 5% of the individuals in the treatment group purchases and no negative effect is present, but it is also possible that 6.5% of the individuals in the treatment group purchases and 1.5% of the individuals decide not to purchase because of the treatment. So by reducing the number of individuals that receive treatment enables the uplift model to return a higher proportion of purchases if a smaller part of the treatment group is targeted, i.e., no *sleeping dogs* are given treatment.

The optimum curve is a theoretical curve which is computed under the assumption that

all the individuals in the treatment group are persuaded to purchase because of the given treatment. This assumption is the theoretically best possible outcome and can be estimated by assuming that all the *sure things* are *persuadeables*. The theoretical maximum curve is computed in the same manner for the uplift curve (Yi, 2018).

The Qini measures can be easily modified to evaluate the result from an uplift model with continuous target feature. It is performed by replacing the incremental purchase on the vertical axis in the Qini curve with the total value of the incremental purchase (Radcliffe, 2007).

4.3 Project Implementation

In this report, two methodologies will be used to model and compare uplift, *class transformation method*, and model *uplift directly*. The *two model approach* will not be considered due to two reasons. Firstly, it does not try to model uplift directly; rather, it assumes that uplift is the outcome from the subtraction of two models trained for the treatment group and control group separately. Secondly, several authors mention that the two model approach performs poorly in practice (see section 4.1.1). The results in this report will be evaluated with the Qini measures.

There is no method that is consistently outperforming the other, therefore it is of importance to test several methodologies which utilize different algorithms to find which is most appropriate for the specific problem (Guelman, Guillen, and Perez-Marin, 2012, Jaśkowski and Jaroszewicz, 2012), e.g., uplift random forests can outperform class transformation method with XGBoost in one problem but the vice versa can happen on another problem. Therefore, this report test and evaluates different methodologies to find which performs is best in this particular data set.

4.4 Statistical Power of a Test

The statistical power of a test is the probability of capturing a significant result, given that there is a difference in the population, or in other words, the ability to detect a specific effect size. Effect size is defined as the difference between the two groups, here treatment group and control group, so the effect size in this data set is approximately $5\% - 3\% = 2\%$. In order to conclude that the effect size from two proportions in the same sample is significant, one can use a binomial test of equal proportions, also known as *two proportion Z-test*. When performing a study like this, it is of interest to ensure that enough sample is available to ensure that a significant difference between groups can be identified. The required sample size per group can be estimated by equation 4.12 (Wang and Chow, 2014).

$$n = \frac{(z_{\alpha/2} + z_{\beta})^2 \cdot (p_t(1-p_t) + p_c(1-p_c))}{(p_t - p_c)^2} \tag{4.12}$$

Where n is the minimum required sample, $z_{\alpha/2}$ is the critical value for the normal distribution for $\alpha/2$ (two-sided confidence level), z_β is the critical value of the normal distribution for β (the power of the test), p_t is the proportion for treatment group and p_c is the proportion for the control group. Equation 4.12 assumes that the groups are of equal size.

4.5 Experimental design

4.5.1 What is the optimal control group size?

Commonly when the optimal design of a study is discussed, one refers to how the study should be planned to ensure that the design in the experiment is set up in a way that enables the researcher to measure and capture the effect she is interested in. For instance, the researcher can be interested in maximizing precision in the difference $p_1 - p_2$, maximize the precision in the ratio p_1/p_2 or maximize the power to detect a group difference (Brittain and Schlesselman, 1982).

Usually, three areas are considered in the planning of a study. First, how many treatments should be tested in the study? Is it only one treatment group vs. control treatment or several treatments that should be compared? Secondly, how strong should the dosage be in (each) treatment? For instance, in a medical study, the dosage is the amount of medicine that is given. Thirdly, how many subjects should be assigned to each group (Begg and Kalish, 1984)?

There is only one type of treatment in this report, which is the act of calling a customer. This is a binary action, which limits the study to a single treatment group and a control group that does not receive treatment. The dosage of treatment is not considered in this report as the same offer is given to everyone in the campaign, and it is only given once. The last thing to consider is the allocation of customers to the treatment group.

This study, try to find the optimal allocation to treatment group. In order to evaluate what is optimal, data will be simulated with different purchase rates for the treatment group, control group, and treatment group allocation. Thereafter, the standard deviation for uplift per segment will be computed. The model parameters that yield the smallest standard deviation will be considered as the optimal parameter setting for this application. How the parameters have been chosen can be seen in section 5.6. The result will not give the global optimum, just which of the tested settings that perform best since the entire function space is not evaluated. The result should be considered more as a guidance of how to plan future studies. How these simulations have been performed can be seen in detail in section 4.5.2.

The uplift literature has not mentioned what is considered the optimal size for the control group. However, Radcliffe and Surry (2011) suggest two rules of thumb for specifying a sufficient size of the control group. The first rule is that the control group needs to be at

least ten times larger when predicting uplift compared to when uplift is to be measured after a campaign. Secondly, modelling a binary outcome, the product of the overall uplift and the size of each sample should be at least 500, so if uplift is 2% then each group should at least contain $\frac{500}{0.02} = 25000$ individuals.

4.5.2 Optimal Allocation for Treatment Group

Brittain and Schlesselman (1982) have developed equation 4.13 to compute the optimal allocation of group belonging for a binary experiment when the researcher wants to maximize the precision in estimating a difference in proportions. Their equation returns the optimal proportion of the sample that should be assigned to the treatment group in a binary experimental setting given that the proportions are known. Generally, the proportions are unknown before collecting the data, but when the data is collected, it can be too late to assign group belonging. Therefore, it is hard to compute the optimal allocation for treatment group beforehand. The following equation shows how the optimal treatment group allocation is computed.

$$W_t = \frac{(p_t q_t)^{1/2}}{(p_t q_t)^{1/2} + (p_c q_c)^{1/2}} \tag{4.13}$$

where, W_t is the optimal proportion of the collected data that should be in treatment group, p_t is the rate in treatment group and p_c is the rate in the control group. q_t and q_c respectively is computed by $q_x = 1 - p_x$. When the optimal allocation is computed, one can easily compute the number of observations for the treatment group by $n_t = N \cdot W_t$ where N is the total number of individuals in the data set. The size of the control group in an experiment with one treatment group and the control group is computed by subtracting the treatment group size from the total number of individuals in the sample.

Equation 4.13 return the optimal proportion for uplift if the uplift is computed as the subtraction of two proportions (the two model approach). As uplift is computed by class transformation method and model uplift directly with uplift random forests, a simulation study will be performed to indicate which treatment group allocation that is optimal using another method.

Simulations to Find Optimal Allocation to Treatment Group

The scope of the simulations is to identify the optimal treatment group allocation when uplift is modeled by uplift random forests and compare the result from equation 4.13. The simulation generates data and train uplift models with different purchase rates and treatment group allocation to identify which allocation that is best. To achieve this, data is simulated by a truncated multivariate normal distribution as it manages to capture both the dependencies between the explanatory features and the dependency between the target feature and the explanatory features (Breslaw, 1994). The distribution is truncated to ensure that only values which are feasible in this application are generated. In order to find reasonable parameters to estimate the truncated normal distribution,

five features from the real data have been chosen to estimate the parameters used to generate new data.

The chosen features are purchase, age, duration as a customer in months, the number of insurances the customer have and the number of months since the last purchase. The purchase feature is truncated between 0 and 1 and is considered to represent the probability of purchase for the customer. Each simulated value for purchase rate is then transformed by the binomial distribution of size 1, with the simulated value as a probability of receiving 1. The purchase features are required to be binary as the model is limited to a binary response feature. The age feature has been truncated in the data generation process to be within the range of 18-80, which is the age limit for giving treatment. The number of months as a customer and the number of months since last purchase have been truncated in the data generation process to the range 0 and 300. The number of insurances is truncated in between 1 and 20. From the purchase feature and the four explanatory features in the real data, the mean vector, standard deviation vector, the correlation matrix is estimated. The truncated normal distribution is simulated by the help of an R package *tmvtnorm* which utilize an Gibbs sampler (Stefan and Manjunath, 2015).

When performing a simulation study, it is essential that the parameter settings are identical except for the parameter in question that is tested for different levels; otherwise, the user faces the challenge to prove that the change was caused by the change in the parameter and not because of the change in the dependency structure. To avoid this, the same mean vector and covariance matrix should be used when data is generated for each model (apart from the parameter in question).

When generating new data with the mean vector and covariance matrix, an issue arises with the purchase feature. As the truncation limited is between 0 and 1 simultaneously as the variance in the feature is relatively high, a proportion of the generated distribution is outside the truncation limit. As the truncation only accepts values within the limit, the mean of the distribution becomes different than specified, see figure 4.5. In order to avoid the this from happening, the variance in the purchase feature is reduced, and the specified mean is lowered slightly, to shift the entire distribution sideways, so the correct proportion of the simulated data is within the limit. The reason why the variance is not reduced to the extent that the entire distribution fits entirely within the truncation limit is because some of the models have a purchase rate very close to the lower bound of the truncation. By letting the variance be slightly higher and instead shift the distribution sideways when the desired mean is close to the limit lets the user have a somewhat higher variance which resembles the real data set used in this report to a greater extent. The required amount to adjust the mean value for purchase have been found by generating 1,000 truncated normal distribution with 1,000 values in each iteration with different mean parameters.

The truncated multivariate normal distribution requires the mean vector and covariance to generate new data. Since dependencies between the features should not change for

different purchase rates or treatment allocations, the same correlation matrix is used for all simulated models. However, as the covariance represents the dependency structure, a single value in the covariance matrix cannot be changed in order to reduce variance in a single feature. Therefore, is the covariance matrix computed from the correlation matrix and the standard deviation. The standard deviation vector and correlation matrix are estimated from the data set, the standard deviation for purchase is reduced, and the new adjusted standard deviation vector is used in conjunction with the correlation matrix to compute the covariance matrix, by equation 4.14. Figure 4.6 shows an example of simulated data when the mean vector has been adjusted to a lower value, and the variance for the purchase feature has been reduced.

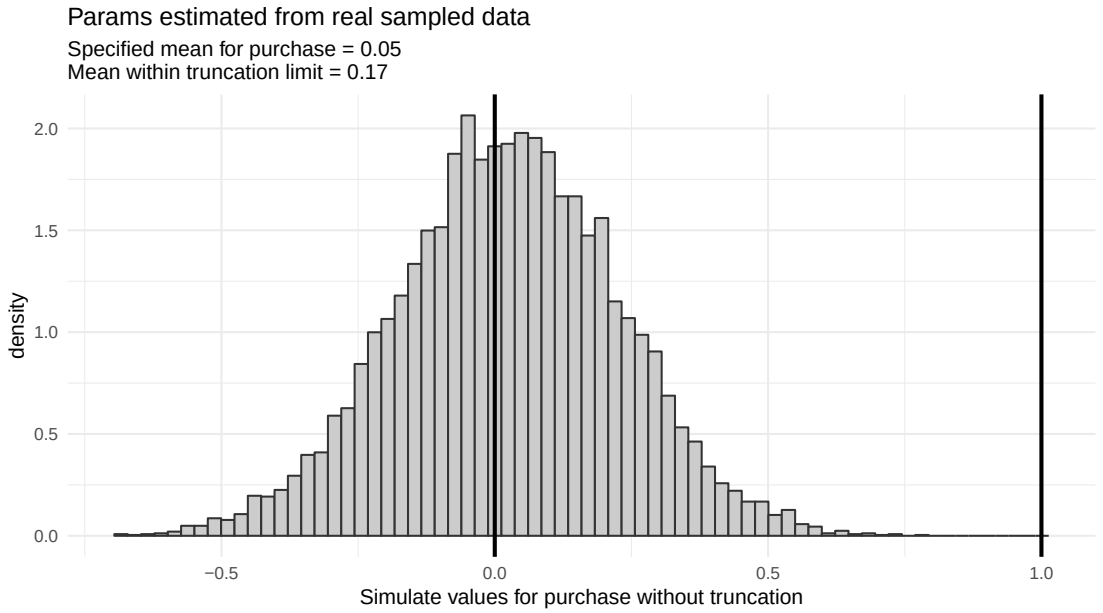


Figure 4.5: Example of simulated data from a multivariate normal distribution with parameters for model 0.05 purchase rate with 60% treatment group allocation, with truncation limits visualized. The data have been simulated using the covariance matrix estimated from real data. The mean of all data points within the truncation limit is approximately 0.17, but the mean is specified to 0.05.

$$\mathbf{cov} = (\mathbf{sd} \cdot \mathbf{sd}^T) \cdot \mathbf{cor} \tag{4.14}$$

Once the mean vector and covariance matrix are computed, the simulation can start. Sixteen different models are tested, with parameters from table 5.4. Each model generates a new data set and fit a uplift random forest in each of the 1,000 replications. Each model generates a single test data set, which is used to evaluate the 1,000 models trained in each replication. When data is simulated for a new replication, two subsets of data is generated, one data set for the treatment group and one set for the control group. These

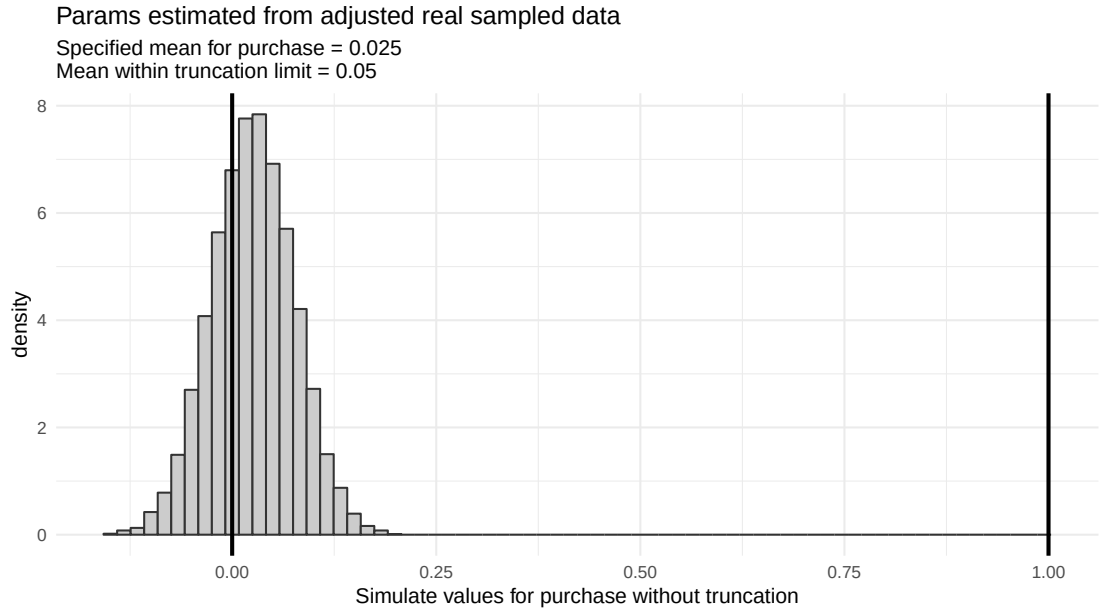


Figure 4.6: Example of simulated data from a multivariate normal distribution with an adjusted mean for model 0.05 purchase rate with 60% treatment group allocation, with truncation limits visualized. The mean of all data points within the truncation limit is approximately 0.05 and when purchase rate is specified as 0.025. The variance has also been reduced so a bigger proportion of the distribution will fit within the truncation limit.

two simulated data sets have different purchase rate and size, so the two data sets together have the correct purchase rates and treatment group allocation. The model trained in each replication is used to estimate the uplift in the test data set. When all replications have been run, the mean uplift and standard deviation per segment are computed. The model with the smallest standard deviation will be considered as the best because a low standard deviation implies a higher certainty in the parameter estimate.

To better capture the possible deviations within each model, a small sample is generated in each replication, and many replications are performed. Each replication generates 500 individuals, and as a consequence, the data is evaluated in five segments, as the size of the generated data set is not enough to be evaluated on more segments.

The parameter settings have been chosen to cover the optimal parameter setting according to the results from equation 4.13, which is approximately 60% allocation to the treatment group given the purchase rate in the data set. The treatment group allocation increases incrementally by 0.1 until it reaches 0.9, which is close to the proportion that is in the real data. The treatment allocation covers the space between the optimal allocation to the allocation that is in the data set. The purchase rates have been chosen to capture both what happens if the purchase rate increases and what happens if the

difference between the purchase rate in the treatment group and control group increases. It starts from a rate of 0.05 for treatment group and 0.025 for the control group, which is close to the purchase rate in the data material and then increases to a maximum of 0.20 purchase rate in the treatment group and 0.175 in control group, see table 5.4.

5. Results

5.1 Comparison of Results from each Model

Table 5.1 shows the trained models and the Qini score for each model. The random uplift forest with Euclidean splitting criterion is outperforming the other uplift random forests and the class transformation method with XGBoost. The class transformation method is the second best model, with a Qini score close to the winning model. Five models that capture the conditional average treatment effect have been built with two different methods.

Table 5.1: Model results

Method	Model	Qini
Model Uplift Directly	Random Forest Euclidean Distance	0.150
Class Transformation Method	XGBoost	0.142
Model Uplift Directly	Random Forest Kullback-Lieber Divergence	0.128
Model Uplift Directly	Random Forest Chi-square Divergence	0.125
Model Uplift Directly	Random Forest L1-norm Divergence	0.123

5.2 Class Transformation Method

Once the data is prepared, equation 4.3 is applied to the treatment and outcome feature to transform them into one continuous variable, Y_i^*, which is the uplift score. The new transformed feature is used as a target to train the XGBoost regressor. The result of the trained model has been evaluated with a test data set, where XGBoost have predicted the score, and the result has been evaluated by Qini measures. The XGBoost regressor has been trained by using a cross-validated randomized grid search algorithm, that searches through the hyperparameter space to identify optimal parameter setting.

Figure 5.1 shows a histogram of the predicted scores from XGBoost. The scores seem to be approximately normally distributed with a slight right-skewness. Scores below 0 can be interpreted as individuals that would have a negative effect from treatment and should therefore preferably not be given treatment.

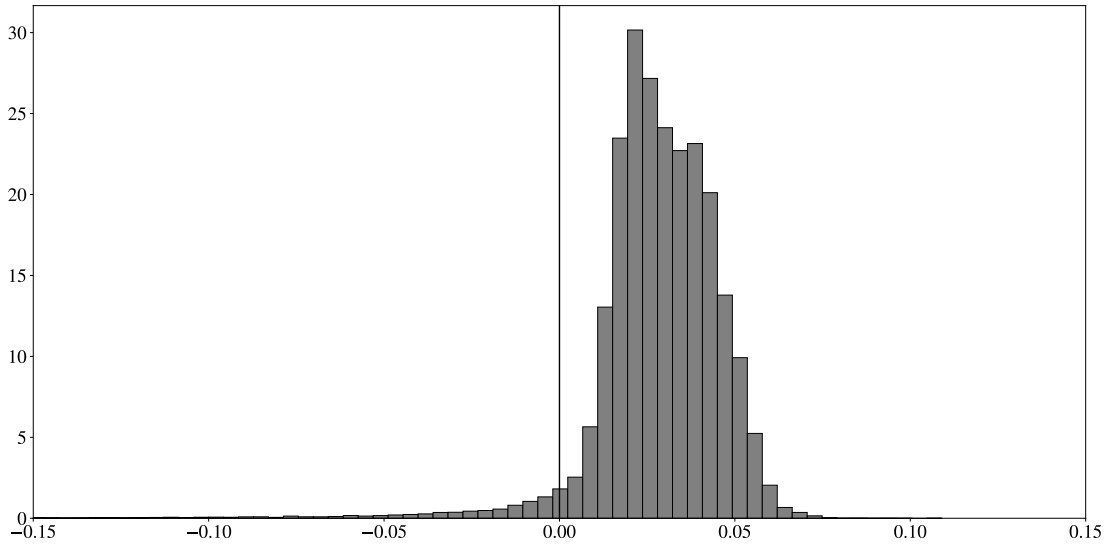


Figure 5.1: Histogram of predicted score for class transformation method with XGBoost

Table 5.2 shows the estimated uplift per segment. The data set has been portioned into 20 segments. The number of segments has been chosen arbitrary, the more segments, the more detailed evaluation can be performed; however, it requires a large data set.

Column n_t and n_c contain the number of individuals belonging to the respective group per segment. Column $n_{t,1}$ and $n_{c,1}$ is the number of purchases performed in each group per segment. Column $r_{t,1}$ and $r_{t,1}$ is the purchase rate per group and segment, computed from $n_{t,1}/n_t$ and $n_{c,1}/n_c$. Uplift is the result of $r_{t,1} - r_{c,1}$. An ideal model would have the highest uplift in the first segment, second to largest uplift in the second segment, and so on. It can be seen that the second segment and not the first that captures the most uplift, but apart from that, almost all segment capture a bit less uplift than the previous one. The 19th segment captures a negative effect, which implies that the purchase rate of individuals in this segment is higher for the control group than in the treatment group; therefore individuals in this segment should not be given treatment.

The Qini curve has been computed and can be seen in figure 5.2. The green dashed line is the theoretical maximum line. The grey dashed line represents the uplift at a given proportion of the sample if the allocation of treatments was given randomly. The computed Qini curve shows the uplift at a given proportion of the sample if the individuals are targeted according to the uplift score. If the entire sample is targeted, the uplift model and random allocation will yield the same result, since everyone is given treatment. The highest point in the uplift curve gives the optimal uplift, and from that point, it can be seen what the optimal proportion of the treatment group that should be given treatment. In this case, without considering the cost of giving treatment, the optimal proportion is at 90% of the treatment group and would yield approximately 2.4%

Table 5.2: Performance table for class transformation method with XGBoost, uplift per segment

Bin	n_t	n_c	$n_{t,1}$	$n_{c,1}$	$r_{t,1}$	$r_{c,1}$	Uplift
1	7390	518	715	20	0.096752	0.038610	0.058142
2	7343	565	591	11	0.080485	0.019469	0.061016
3	7357	551	472	8	0.064157	0.014519	0.049638
4	7314	594	422	9	0.057698	0.015152	0.042546
5	7257	651	443	17	0.061045	0.026114	0.034931
6	7354	554	374	15	0.050857	0.027076	0.023781
7	7318	590	307	12	0.041951	0.020339	0.021612
8	7347	561	300	10	0.040833	0.017825	0.023008
9	7305	603	250	7	0.034223	0.011609	0.022615
10	7263	645	263	12	0.036211	0.018605	0.017606
11	7247	661	201	10	0.027736	0.015129	0.012607
12	7286	622	231	10	0.031705	0.016077	0.015627
13	7222	686	206	2	0.028524	0.002915	0.025609
14	7203	705	180	6	0.024990	0.008511	0.016479
15	7235	673	194	5	0.026814	0.007429	0.019385
16	7237	671	159	12	0.021970	0.017884	0.004087
17	7182	726	174	10	0.024227	0.013774	0.010453
18	7213	695	210	12	0.029114	0.017266	0.011848
19	7135	773	237	26	0.033217	0.033635	-0.000419
20	6949	963	397	52	0.057131	0.053998	0.003133

uplift, which is higher than if the entire treatment group was given treatment. This is equivalent of targeting segment 1 to 18. The Qini coefficient, which is the area between the Qini curve and the random selection divided by the area between the theoretical maximum and the random selection curve is 14.2%. Ideally, the model should capture a greater part of the volume "earlier" in the population.

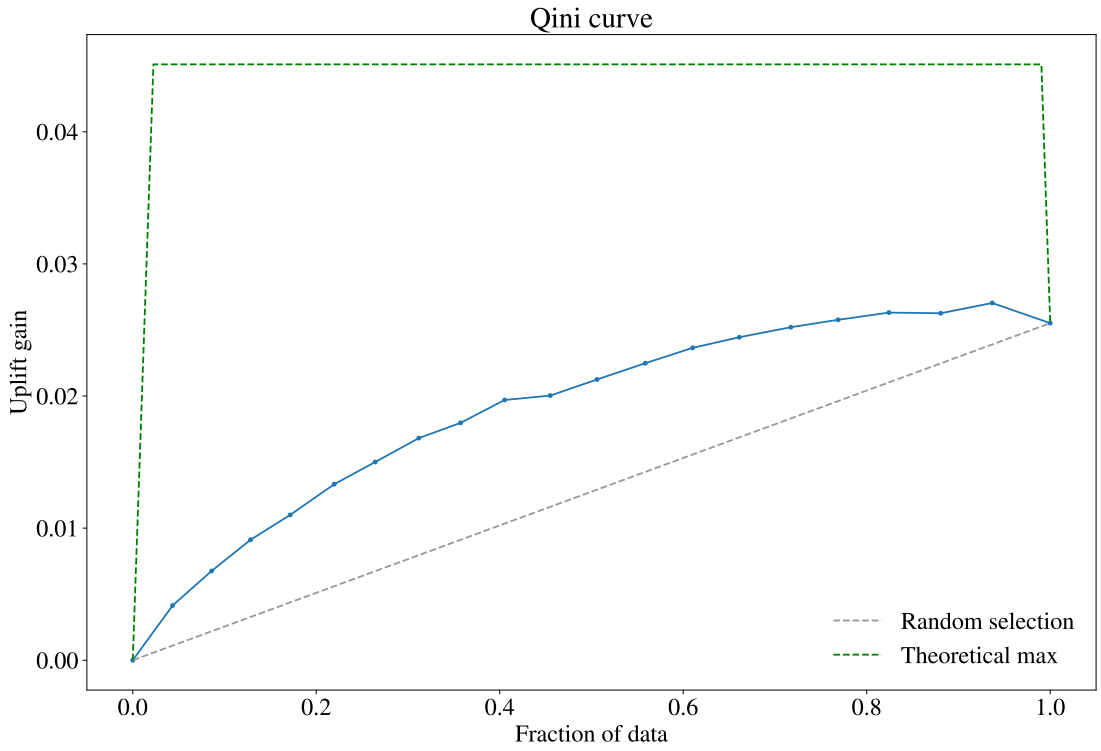


Figure 5.2: Qini Curve for Extreme Gradient Boosting with class transformation method

5.3 Model Uplift Directly

An approach based on random forests has been utilized to model uplift directly. For comparability, each model has been trained and evaluated with the same train and test data set containing the same features. Four different splitting criterion for the random forest algorithm has been tested to evaluate which capture uplift best. Each random forest model has had the same parameters except the splitting criterion. The number of trees in each forest has been selected to 200 and nodes with less than 1,000 observations have not been subject for splitting, which enables the model to build deep trees. The uplift random forests have not been subject to hyperparameter optimization.

One advantage with random forests is the interpretability of the model and that the relative feature importance can be estimated. See figure 5.3, 5.4, 5.5 and 5.6. In short, all four different splitting criterion's rank similar features as the top most important for the model.

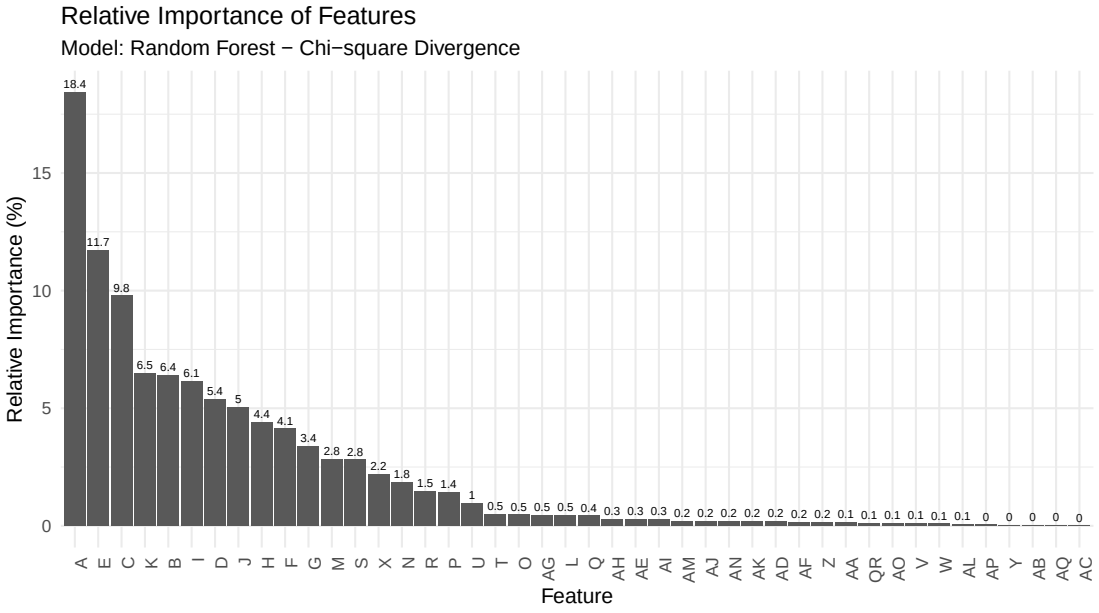


Figure 5.3: Feature Importance for Random Forest with Chi-square Divergence as splitting criterion

The original features are hidden by a request of the commissioner, so each feature is replaced with a letter. The most important features for the χ^2-divergence splitting criterion (figure 5.3) is age (A), time as a customer in months (E) and time since last bought product (C). For the Euclidean distance splitting criterion (figure 5.4), the most important features are age (A), time since the last call in months (B) and time since last bought product in months (C). Kullback-Liebler divergence (figure 5.5) important features are age (A), time since last purchase in months (C) and time since the last

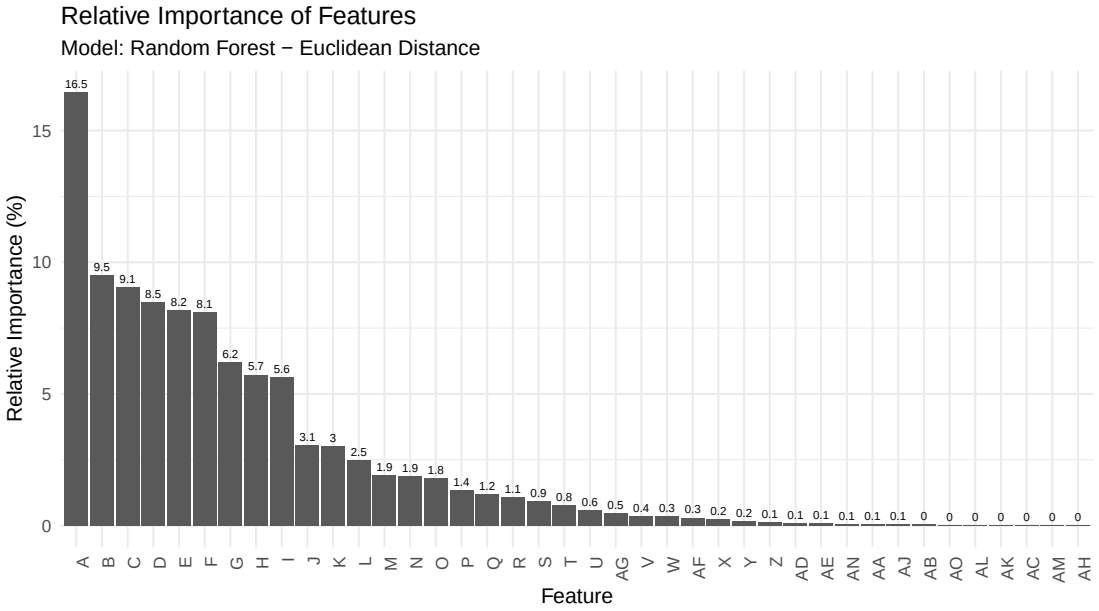


Figure 5.4: Feature Importance for Random Forest with Euclidean Distance as splitting criterion

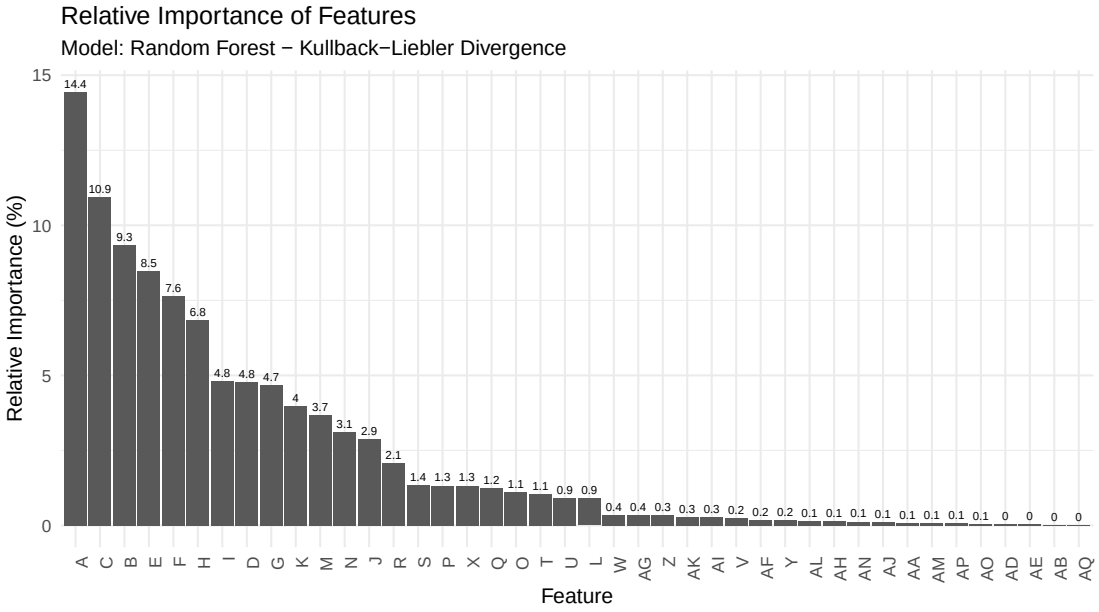


Figure 5.5: Feature Importance for Random Forest with Kullback-Liebler Divergence as splitting criterion

call in months (B). L1-norm divergences (figure 5.6) best features are the time since last purchase in months (C), time since the last call in months (B) and the number of insurances within the category "household" in the individuals household (G). All of the splitting criterion's have several features with very low or no relative importance for the model, such as the binary indicators for which month of the year for last purchase was performed (AD - AO). One could argue that features with very low or no contribution should be removed, and the model should be trained once again without those features. Each splitting criterion's top feature stand out as substantially stronger than the second feature and χ^2, Euclidean and Kullback-Liebler's splitting criterion's, the top feature is age (A). Those splitting criteria have 18-20 features that contribute with 1% or more to the model; meanwhile, L1-norm divergence has fewer contributing features, only 15.

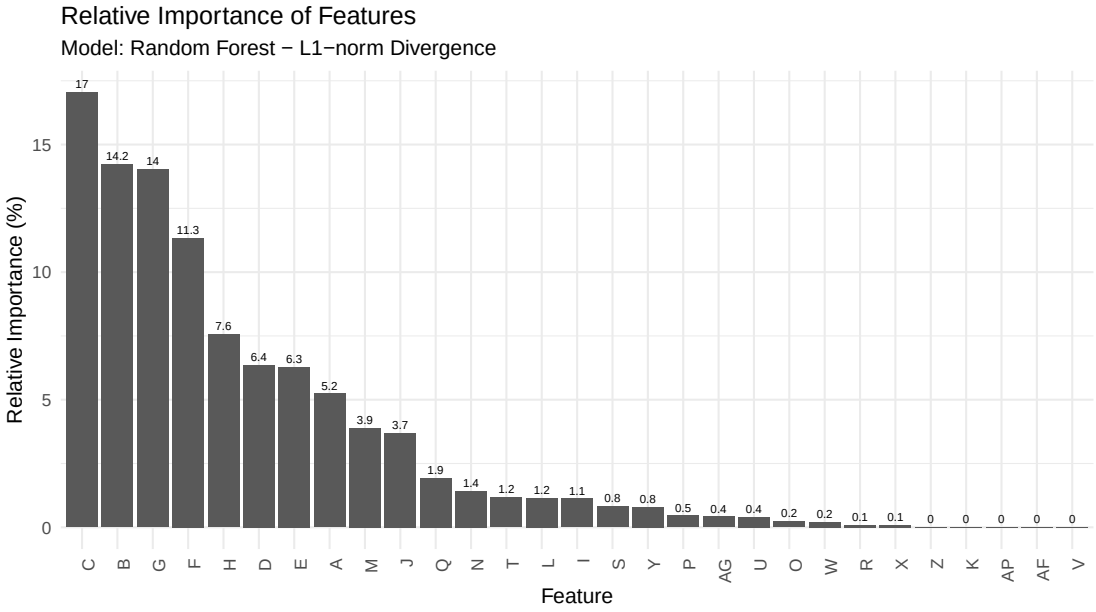


Figure 5.6: Feature Importance for Random Forest with L1-norm Divergence as splitting criterion

The uplift random forest returns the probability of purchase given group belonging, which then is transformed into the uplift score (see equation 4.9). All individuals should be sorted according to the score to maximize the uplift. A negative score is interpreted as an individual that has a higher probability of purchase in the control group than in the treatment group. Therefore, an individual with a negative score should not be given treatment. Figure 5.7 shows a histogram of the scores for each splitting criterion. For a more detailed view, see appendix figure 1, 2, 3 and 4.

The score for χ^2-divergence splitting criterion seems to yield a distribution similar to a normal distribution. The Euclidean distance splitting criterion has a lower mode and fat-

ter tail than χ^2-divergence, which looks similar to a t-distribution. The Kullback-Liebler divergence splitting criterion yield a higher mode than the χ^2-divergence, but still has what seem to be somewhat symmetric tails. The L1-norm divergence splitting criterion has an even higher peak than the Kullback-Liebler divergence, and the distribution is skewed to the right.

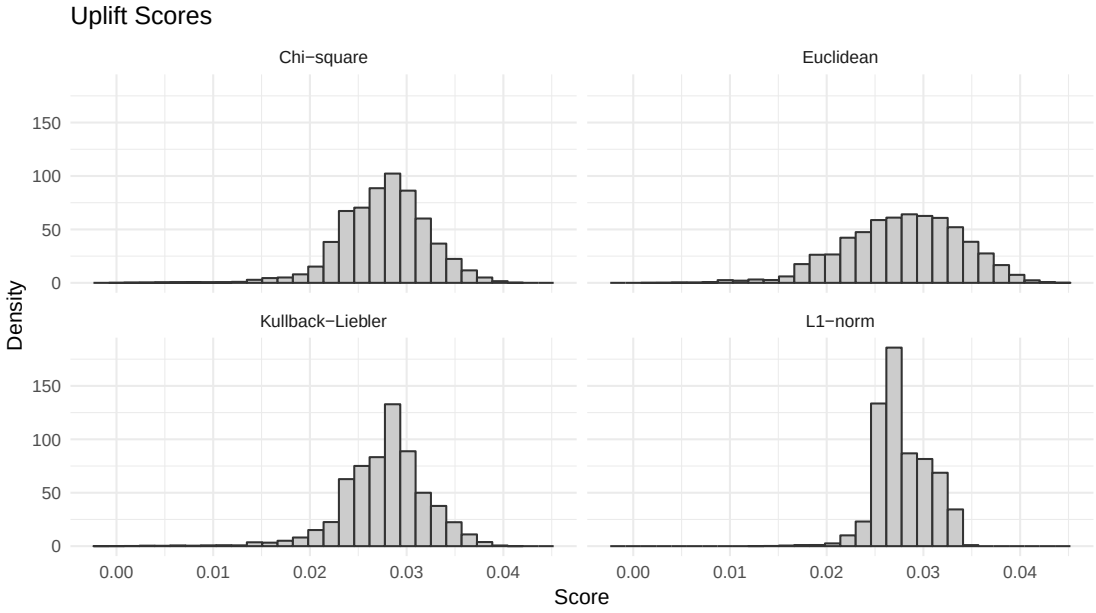


Figure 5.7: Histogram of Uplift Score for each splitting criterion

Figure 5.8 show the Qini curve for each uplift random forest. It can be seen that the Euclidean splitting criterion outperform the others slightly, as it manages to identify more purchasers "earlier" in the sample. For a more detailed view of uplift per segment per model, see the performance table 5.3 for Euclidean distance or appendix table 1, 2 and 3 for the other splitting criterion's. It is noteworthy that the Euclidean splitting criterion manages to identify more persuadeables in the "beginning" of the sample, but it does not manage to identify any sleeping dogs, individuals that get a lower change to purchase given treatment. The other splitting criterions manages to identify one or two segments in the bottom where treatment has a negative effect.

The outcome of the models are similar, where all of them approximately manage to capture 2 percent-units of the total 2.6 percent-units uplift when half of the treatment group is given treatment. In other words, $\frac{2}{2.6} = 0.77$ of the total amount of purchases can be found in the first half of the treatment group.

Table 5.3: Performance table for Random Forest Euclidean Distance, uplift per segment

Bin	n_t	n_c	$n_{t,1}$	$n_{c,1}$	$r_{t,1}$	$r_{c,1}$	Uplift
1	7256	652	693	12	0.095507	0.018405	0.077102
2	7178	729	602	23	0.083867	0.03155	0.052317
3	7218	700	548	15	0.075921	0.021429	0.054493
4	7279	619	479	20	0.065806	0.03231	0.033496
5	7277	629	411	13	0.056479	0.020668	0.035812
6	7315	593	461	12	0.063021	0.020236	0.042785
7	7286	621	387	19	0.053116	0.030596	0.02252
8	7283	624	362	17	0.049705	0.027244	0.022461
9	7237	671	340	17	0.046981	0.025335	0.021645
10	7253	654	295	12	0.040673	0.018349	0.022324
11	7294	613	301	8	0.041267	0.013051	0.028216
12	7276	632	261	9	0.035871	0.014241	0.021631
13	7281	632	221	11	0.030353	0.017405	0.012948
14	7235	666	216	6	0.029855	0.009009	0.020846
15	7289	620	186	5	0.025518	0.008065	0.017453
16	7287	620	184	5	0.02525	0.008065	0.017186
17	7191	715	153	8	0.021277	0.011189	0.010088
18	7325	583	127	5	0.017338	0.008576	0.008762
19	7293	614	132	7	0.0181	0.011401	0.006699
20	7435	473	149	8	0.02004	0.016913	0.003127

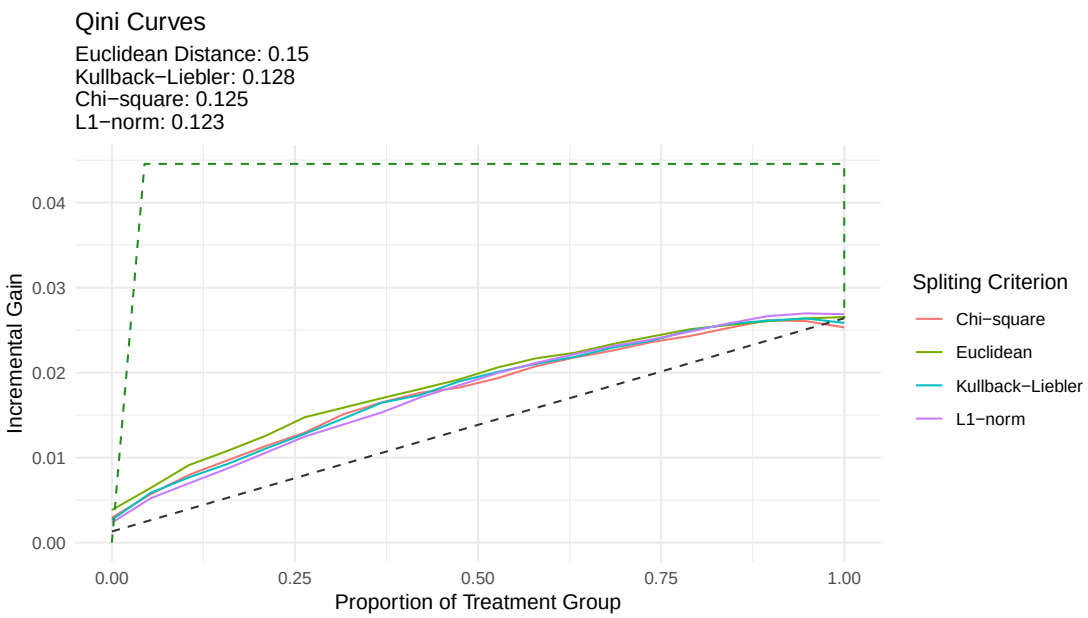


Figure 5.8: Qini Curve for Random Forest which model uplift directly

5.4 Statistical Power of a Test

If significance level $\alpha = 95\%$, power $1-\beta = 80\%$, a sample proportion of 5% for treatment group and 3% for the control group, equation 4.12 then states that the study would require a minimum of 1504 individuals in the respective group to detect a significant difference of magnitude 2%. If the desired effect size were even smaller a larger sample would be required, for instance, is the purchase rate in the treatment group was 5%, and control group was 4.5%, then the required sample size to find a significant effect would be 28406 individuals in each group. As the data used in this study have approximately 580,000 individuals in the treatment group and 50,000 individuals in control group, the sample size is by far more extensive than what is required in order to capture a significant result which the specified significance level and power.

5.5 Exhaustive Search for Optimal Design

The optimal allocation depends on the purchase rate in the treatment group and the control group. Figure 5.9 shows how the optimal allocation shifts as the purchase rate changes, which is the result from equation 4.13.

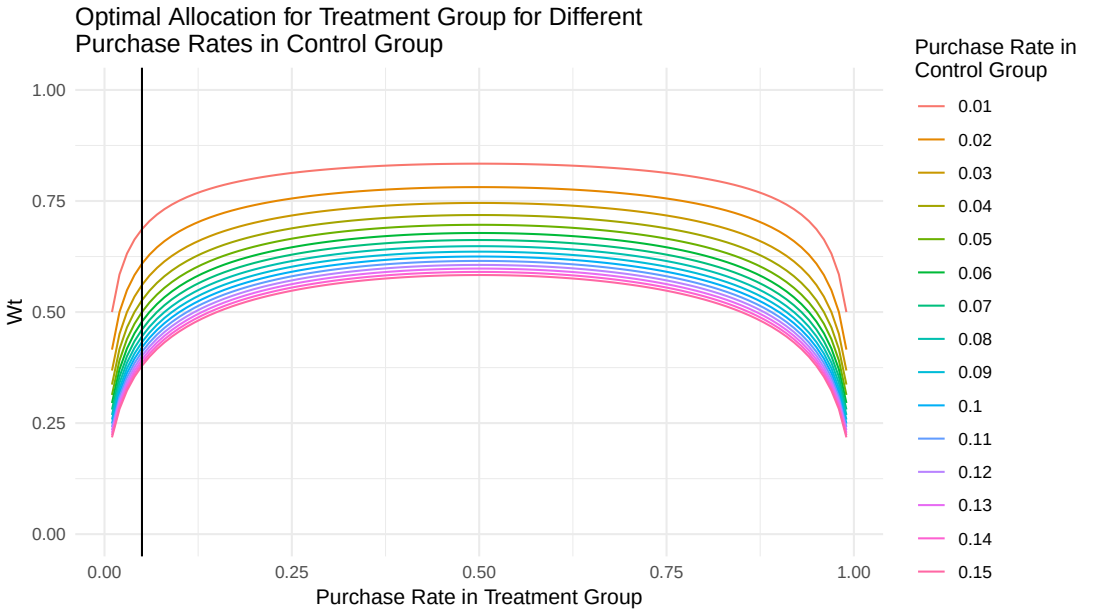


Figure 5.9: Optimal Allocation given different purchase rate in treatment and control group, computed by equation 4.13.

The horizontal axis in figure 5.9 is the purchase rate for the treatment group, from 0 to 1. The vertical axis shows the optimal allocation of data points to the treatment group (W_t). Each curve presents a fixed value of the purchase rate for the control group. The

fixed values for the control group purchase rate have been chosen since they are close to the original values in the data set. The vertical line in the figure is a visual aid to easier identify the optimal allocation when the purchase rate in the treatment group is 5%, which is close to the purchase rate in the treatment group in the data material. Since equation 4.13 multiplies each term with both $p_x q_x$, the curve becomes symmetric, so a purchase rate of 10% in control group have the same position as the curve for purchase rate 90% in the control group.

If the purchase rate in the treatment group (horizontal axis) would be 0.5, and the purchase rate in the control group was 0.03, then the optimal allocation for the treatment group would be 0.75. Instead, if the purchase rate in the treatment group would be 0.05, and the purchase rate in the control group would remain at 0.03, the optimal treatment allocation would be approximately 0.56.

5.6 Simulation Studies for Optimal Control Group Size

The algorithm that has been developed to estimate the optimal control group size is utilizing the best model, uplift random forest with Euclidean distance as the splitting criterion. It is assumed that the optimal control group size would yield the smallest standard deviation for uplift per segment. The process of how the data is simulated is described in detail in section 4.5.2.

The first numbers in the model name represent the purchase rate in the treatment group, the second number is the purchase rate in the control group, and the third number is the proportion of samples allocated to the treatment group. Each number in the model name is separated with a _. So mod_005_0025_06 is trained with purchase rate 5% in the treatment group, 2.5% in control group, and 60% of the customers were assigned to the treatment group.

Sixteen different models with different combinations of the three parameters have been tested with 1,000 replications for each model. The uplift and standard deviation of uplift is estimated for each segment. The used parameter settings can be seen in table 5.4.

Table 5.4: Parameter settings for 16 models to compare uplift and standard deviation per segment

#	Model Name	Purchase rate Treatment group	Purchase rate Control group	Allocation to Treatment group	Recoded purchase rate Treatment group	Recoded Purchase rate Control group
1	mod_005_0025_06	0.05	0.025	0.60	0.025	-0.05
2	mod_005_0025_07	0.05	0.025	0.70	0.025	-0.05
3	mod_005_0025_08	0.05	0.025	0.80	0.025	-0.05
4	mod_005_0025_09	0.05	0.025	0.90	0.025	-0.05
5	mod_01_005_06	0.10	0.05	0.60	0.10	0.025
6	mod_01_005_07	0.10	0.05	0.70	0.10	0.025
7	mod_01_005_08	0.10	0.05	0.80	0.10	0.025
8	mod_01_005_09	0.10	0.05	0.90	0.10	0.025
9	mod_02_01_06	0.20	0.10	0.60	0.20	0.10
10	mod_02_01_07	0.20	0.10	0.70	0.20	0.10
11	mod_02_01_08	0.20	0.10	0.80	0.20	0.10
12	mod_02_01_09	0.20	0.10	0.90	0.20	0.10
13	mod_02_0175_06	0.20	0.175	0.60	0.20	0.175
14	mod_02_0175_07	0.20	0.175	0.70	0.20	0.175
15	mod_02_0175_08	0.20	0.175	0.80	0.20	0.175
16	mod_02_0175_09	0.20	0.175	0.90	0.20	0.175

Table 5.5 show the mean vector and the standard deviation vector estimated from real data. The mean vector is used to simulate data for each model for the treatment group and the control group. For each model, the purchase means have been replaced with the corresponding mean values for the treatment group and control group for that model.
Table 5.6 shows the correlation matrix that is used to compute the covariance matrix to generate data from the truncated multivariate normal distribution. It is noticeable

Table 5.5: Mean and Standard deviation vector. Mean vector is used to generate data and the standard deviation vector is used to transform the correlation matrix to a covariance matrix by equation 4.14.

	Purchase Probability	Age	Duration as customer (months)	Number of Insurances	Months since last bought Insurance
Mean	0.05	53.42	65.03	4.91	34.64
Standard Deviation	0.05	15.80	65.64	2.09	44.14

that the correlations to purchase are low, which implies that the explanatory features do not contribute much in estimating the uplift for each model. The correlation matrix is estimated from the original data set. The covariance matrix is computed from the correlation matrix and standard deviation vector by equation 4.14. The covariance matrix can be seen in table 5.7.

Table 5.6: Correlation matrix used to compute covariance matrix

	Purchase Probability	Age	Duration as customer (months)	Number of Insurances	Months since last bought Insurance
Purchase Probability	1.0000	0.0074	-0.0046	0.0614	-0.0264
Age	0.0074	1.0000	0.2652	0.0478	0.1991
Duration as customer (months)	-0.0046	0.2652	1.0000	0.2472	0.5853
Number of Insurances	0.0614	0.0478	0.2472	1.0000	-0.1288
Months since last bought Insurance	-0.0264	0.1991	0.5853	-0.1288	1.0000

In order to ensure that the simulated data have the intended property, the total uplift for each data set is computed for each replication. The average uplift in the data set should be equal to treatment group - control group. Histogram of the total uplift for each combination of purchase rates can be seen in figure 5, 6, 7 and 8 in the appendix. Each histogram has its mode at the intended level, which is at the subtraction of treatment group purchase rate and control group purchase rate.

The average uplift per segment has been computed for each model from the 1,000 replications. For more details of how the computation has been performed, see section 4.5.2. The result can be seen in table 5.8. It is noticeable that the models do not seem to be able to identify the *persuadeables* nor the *sleeping dogs* as the first segment do not capture the most uplift, the following segments should capture a lower amount of uplift and the

Table 5.7: Covariance matrix used to generate data

	Purchase	Age	Duration as customer (months)	Number of Insurances	Months since last bought Insurance
Purchase	0.00	0.01	-0.02	0.01	-0.06
Age	0.01	249.68	275.10	1.58	138.89
Duration as customer (months)	-0.02	275.10	4309.12	33.97	1695.93
Number of Insurances	0.01	1.58	33.97	4.38	-11.91
Months since last bought Insurance	-0.06	138.89	1695.93	-11.91	1948.59

last should capture the lowest uplift or even a potential negative effect. No matter what treatment group allocation that is assigned to the model 5-12, the average uplift seem to increase for each segment, which indicates that the model has captured some effect but it is reversed. For instance, model 8 has a negative effect in segment 1 which implies that the chance of purchase is higher in the control group than the treatment group. This is a indication that the group consist of *sleeping dogs*. For models 1-4 and 13-16 the average uplift is approximately the same in each segment and there is no apparent trend if the average uplift increase or decrease between segments.

The standard deviation for uplift has been computed for each model and segment. The result can be seen in table 5.9. It is noticeable that the standard deviation is similar in all segments for each model, which indicate that the uncertainty is approximately the same in the segments.

To easier identify the smallest the standard deviations from table 5.9, each value has been replaced by the rank within each segment, where lowest rank, 1, means smallest standard deviation and highest rank 12 mean highest standard deviation. The result can be seen in table 5.10.

In order to conclude which model that performs best, the standard deviation per segment has been added together for each model. When the standard deviations for each segment are added together it removes the possibility to interpret the value, and it can merely be used to rank different models. The model with the lowest sum has the smallest standard deviation across all segments and is therefore considered to be the best of the tested treatment group allocations. The result can be seen in table 5.11. By adding the standard deviations for each segment together, the result considers each segment in the model equally important.

Table 5.8: Average uplift for each model per segment based on 1,000 replications

#	Model Name	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
1	mod_005_0025_06	0.0173	0.0056	0.0089	0.0215	0.0318
2	mod_005_0025_07	0.0067	0.0084	0.0080	0.0076	0.0026
3	mod_005_0025_08	0.0056	-0.0033	-0.0058	-0.0222	-0.0341
4	mod_005_0025_09	0.0520	0.0519	0.0448	0.0422	0.0383
5	mod_01_005_06	0.0385	0.0303	0.0314	0.0352	0.0506
6	mod_01_005_07	0.0052	0.0041	-0.0018	-0.0060	-0.0102
7	mod_01_005_08	0.0475	0.0539	0.0629	0.0752	0.0876
8	mod_01_005_09	-0.0774	-0.0128	0.0025	0.0128	0.0313
9	mod_02_01_06	0.0899	0.0928	0.0927	0.0956	0.1074
10	mod_02_01_07	0.0864	0.0899	0.0856	0.0775	0.0546
11	mod_02_01_08	0.0621	0.0752	0.0863	0.1169	0.1546
12	mod_02_01_09	0.0864	0.1428	0.1743	0.2017	0.2194
13	mod_02_0175_06	0.0332	0.0231	0.0307	0.0408	0.0446
14	mod_02_0175_07	0.0630	0.0724	0.0685	0.0701	0.0691
15	mod_02_0175_08	0.0449	0.0440	0.0482	0.0560	0.0675
16	mod_02_0175_09	-0.0029	0.0336	0.0387	0.0471	0.0369

In table 5.11, the optimal treatment allocation is 70% for two combinations of purchase rates, when the purchase rate is 0.05 for treatment group and 0.025 for the control group (1); and when the purchase rate for treatment group is 0.20 and 0.10 for the control group (3). The best treatment group allocation is 60% when purchase rate is 0.10 in the treatment group and 0.05 in the control group (2); and when 0.20 purchase rate in the treatment group and 0.175 purchase rate in the control group (4). It is noteworthy that the two model settings which have 0.025 difference between the treatment group and control group (1 and 4) have different optimal treatment group allocation, (1) has optimal at 70% treatment group allocation meanwhile (4) has optimal at 60% treatment group allocation.

The result is hard to interpret, as both the lowest and highest purchase rate in the treatment group (1 and 3) is optimal at 70% treatment group allocation meanwhile purchase rate for treatment group between the extremes has the optimal allocation at 60% (2). When the subtraction between treatment group and control group is the highest tested combination, 0.10 units (3), the optimal treatment group allocation is 70% simultaneously as the lowest subtraction, 0.025 (1 and 4), return different optimal treatment group allocation, 70%, and 60%. The optimal treatment group allocation never exceeds 70%.

In order for the results to be consistent with equation 4.13, each purchase rate should

Table 5.9: Standard deviation in uplift for each model and segment

#	Model Name	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
1	mod_005_0025_06	0.0464	0.0458	0.0467	0.0459	0.0425
2	mod_005_0025_07	0.0364	0.0429	0.0426	0.0406	0.0324
3	mod_005_0025_08	0.0440	0.0486	0.0498	0.0576	0.0505
4	mod_005_0025_09	0.0300	0.0400	0.0488	0.0532	0.0539
5	mod_01_005_06	0.0460	0.0425	0.0460	0.0447	0.0460
6	mod_01_005_07	0.0555	0.0561	0.0558	0.0566	0.0522
7	mod_01_005_08	0.0610	0.0553	0.0593	0.0614	0.0659
8	mod_01_005_09	0.1116	0.1043	0.0955	0.0822	0.0711
9	mod_02_01_06	0.0646	0.0648	0.0637	0.0635	0.0592
10	mod_02_01_07	0.0536	0.0557	0.0624	0.0602	0.0622
11	mod_02_01_08	0.0830	0.0794	0.0761	0.0740	0.0670
12	mod_02_01_09	0.0734	0.0792	0.0661	0.0507	0.0421
13	mod_02_0175_06	0.0737	0.0702	0.0685	0.0717	0.0696
14	mod_02_0175_07	0.0696	0.0701	0.0715	0.0743	0.0724
15	mod_02_0175_08	0.0800	0.0856	0.0850	0.0797	0.0673
16	mod_02_0175_09	0.1097	0.1119	0.1104	0.1150	0.1158

have lowest standard deviation with 60% treatment group allocation. Though, it worth to keep in mind that the formula do not consider any explanatory features so a comparison is not possible, but the result from equation 4.13 is still interesting and can be used as an indication when deciding treatment group allocation prior to a study is performed. Further, it should be mentioned that adding the standard deviations for each segment is a naive evaluation method for the models, as it assumes that all segment has equal importance, which is not the case. The first and last segment are ideally the ones that are most interesting as they should capture the *persuadeables* and *sleeping dogs*.

Table 5.10: Rank of standard deviation per segment

#	Model Name	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
1	mod_005_0025_06	4	4	2	2	2
2	mod_005_0025_07	2	5	5	5	12
3	mod_005_0025_08	3	2	1	1	1
4	mod_005_0025_09	5	1	4	12	5
5	mod_01_005_06	1	3	3	4	3
6	mod_01_005_07	10	7	6	6	6
7	mod_01_005_08	6	10	7	3	4
8	mod_01_005_09	7	6	10	10	9
9	mod_02_01_06	9	9	9	7	10
10	mod_02_01_07	14	14	12	9	7
11	mod_02_01_08	12	13	13	13	11
12	mod_02_01_09	13	12	14	11	15
13	mod_02_0175_06	15	11	11	14	13
14	mod_02_0175_07	11	15	15	15	8
15	mod_02_0175_08	16	8	8	8	14
16	mod_02_0175_09	8	16	16	16	16

Table 5.11: Sum of standard deviation for each model

#	Treatment Group Allocation	60	70	80	90
1	mod_005_0025	0.2273	0.1950	0.2504	0.2259
2	mod_01_005	0.2252	0.2761	0.3029	0.4647
3	mod_02_01	0.3158	0.2941	0.3796	0.3114
4	mod_02_0175	0.3537	0.3580	0.3976	0.5628

6. Discussion

6.1 Treatment and Uplift Modeling

Both methods, class transformation and model uplift directly, manages to capture some uplift in the data. The best model, random forest with Euclidean distance as the splitting criterion, captures 15% of the area from the theoretically best possible model, which must be considered quite bad. The worst model manages to capture 12.3%, so the spread from the best to the worst model is quite small. However, random forest with Euclidean distance as the splitting criterion, manage to capture 77% of the total amount of purchases by giving treatment to 50% of the treatment group. That is a quite substantial increase in efficiency compared to capturing 100% of the total purchases by giving treatment to the entire treatment group.

The class transformation method with XGBoost is performing close to as well as the uplift random forest with Euclidean distance as the splitting criterion. The class transformation model captures a somewhat smaller share of the best theoretical curve and is therefore considered slightly worse, but the model manages to capture a higher uplift than the best model, as it identifies a negative effect for the last segment.

The class transformation method is faster to train compared to the uplift random forests. However, it depends highly on how deep the random forest trees are allowed to grow and how many trees should be trained. Though, by utilizing the cross-validated grid search for parameter tuning in class transformation method, increases the time of training from a few seconds to a few days.

Since the two methods perform similarly well, one can select the model which best suit one's need. This report has used a python implementation for the class transformation method and an R implementation for the uplift by random forest. The authors' opinion is that the class transformation method is easier and faster to implement and especially easier when it comes to parameter tuning, thanks to the Pylift package.

6.1.1 No Negative Effect?

All models return a Qini curve that is continuously increasing as the proportion of the sample that is given treatment increases, except for the last and second to last segment for a few models. The expected result was that the Qini curve would have a clear identifiable mode "early" in the treatment group and from that point, the curve would flatten out or decrease, similar to model 2 in the example figure 4.4. All models capture very little to no negative effect, which either means that the treatment itself causes no negative effect or that the model does not manage to capture it. If the first statement is true, *If* should continue to call as many customers as possible because the treatment is appreciated and the need for an uplift model would be minimal. If the second statement is true, the model

performs bad and does not manage to separate the *persuadeables* and the *sleeping dogs* from the others.

It is reasonable to believe that the potential negative effect is small in the treatment since the service of assuring that one's insurance needs are fulfilled is a comfortable knowledge to have for the customer. If it is considered that the service is provided for free, it makes it even more attractive. Though it is not a fact that can be taken for granted, but the low negative effect from every model could be considered as an indication that it is the case.

When reviewing the dependencies between the features, it is noticeable that the correlation between features and purchase is close to zero. It is always hard to predict a feature, but the task becomes even harder when the relationship between features is close to non-existing. The correlation matrix of the data has been left out of this report by request of *If*.

It would be of interest to once again review the literature to see if there are methods to estimate uplift by data mining techniques such as frequent patterns or some other methodology that have proven itself robust when the signal in data is weak and plenty of data is available.

6.1.2 Feature selection

This report has not fully explored the possibilities of feature selection and whether it might improve the result. The result in this report relies on a machine learning approach where feature selection is barely considered. Instead, almost all features are used to train the model in the training set, and the result is evaluated on another test data set. This approach has been encouraged by *If*. This in conjunction with that the algorithms used in both methodologies is tree-based, both random forests, and XGBoost, which uses a different subset of features to train each tree to prevent overfitting (Chen and Guestrin, 2016). Though, it must be pointed out that before modeling with all features, controls have been performed to ensure that there are no features with strong correlations between each other.

Radcliffe and Surry (2011) mention that the feature selection is, in their experience, crucial in uplift modeling because uplift models try to model a second-order phenomena as uplift models model the difference between two outcomes instead of modeling the result directly. As a consequence uplift models are even more unstable than traditional models.

Moreover, Radcliffe and Surry (2011) mention three desirable properties for features, *Predictiveness*, *Stability (robustness)* and *Independence*. The first property is a general key property that applies to all models, without predictiveness is it hard to build a well functioning model. Stability means that the model should be robust, so if there is a

small change in the input data the model output should also have a small change, this is important as uplift is a second-order phenomena. Independence refers to the explanatory features used to model uplift, it is desirable that the feature used are independent of each other so they capture different aspects of the relationships of the target feature. Independence is the least important of the three according to Radcliffe and Surry (2011), though it is still of importance. Having strong correlations in a data set is not recommended (except for the target feature).

The data set lack features with strong predictiveness. One can argue that it causes the poorly performing models. Contrasting to what Radcliffe and Surry (2011) states, many features have been used, relying on the fact that the tree-based methods have feature selection procedure. As the feature selection within the tree-based method is not too refined, it makes the assumption that the tree-based algorithm will handle the feature selection in an optimal way naive. A potential consequence could be that poor features are selected which in turn could deteriorate the parameter tuning and the final result. Therefore, it would be of interest to apply a more restrictive feature selection and train the models once again. Re-training the models with fewer features might also reduce the required time to train the models. Figure 5.3, 5.4, 5.5 and 5.6 could be used as indication of what features to include when the models are re-trained when a more restrictive feature selection is applied.

6.1.3 Uplift Modeling in Different Levels of the Business

Uplift models can be utilized to model different objectives in the purchase situation. In this report, uplift models have been used to increase or at least maintain the number of purchases; meanwhile, fewer customers are given treatment. Uplift models can also be used to increase the number of visits to a store/website or even to increase the total spending of customers. By focusing on the number of purchases does not necessarily increase the revenue for a company, though it often does give that earnings are maintained; meanwhile, the cost of treatment is reduced. If instead uplift were applied on customer spending, the model would have been able to identify customers that spend more because of treatment that would not have spent as much otherwise (Radcliffe, 2008). This could have been achieved by using a continues target feature such as spending rather than the binary purchase feature that has been used in this report. Though, the continues target feature violates the requirements for the class transformation method and leaves the user with modeling uplift directly.

By modeling a continuous target feature, could potentially have a stronger impact on the business than modeling a binary outcome. A continuous feature, such as spending can identify not only who might be persuaded to purchase because of treatment but also model the value of the purchase.

Most companies have a diversified product portfolio with more and less profitable products. It is naive to assume that all purchases are considered equal in value for the

company. Any company that has marketing and is interested in maximizing the profit should be interested in uplift models and especially with a continuous target feature.

6.1.4 Find the Optimal Cut-off in Treatment Group

While reviewing the Qini curves the models have returned, the user needs to decide how big proportion of the treatment group that actually should receive treatment. Neither of the models returns a Qini curve that has a distinct mode early in the treatment group. In order to find what is optimal for each company, they must consider the cost of giving treatment and the cost of having staff that gives treatment without having subjects to treat. Those costs are unknown in this project and therefore will no recommendation be given of what proportion of treatment group that should be given treatment.

Nevertheless, if the cost of treatment is not considered, the optimal uplift would be at the mode of the Qini curve. For class transformation method with XGBoost, it would be to give treatment to segment 1-18, approximately 90% of the treatment group. By limiting treatment to those individuals would have a higher uplift than giving treatment to everyone. The best model, model uplift directly with random forest and Euclidean splitting criterion, capture no negative effect in the treatment group. Therefore, the optimal amount of treatment without considering the cost of treatment would be to give treatment to the entire treatment group. Though, it should be pointed out that giving treatment to 50% of the treatment group would account for approximately 2 percent units gain out of the total 2.6 percent units. In other words, by only giving treatment to half of the treatment group would account for approximately 77% of the purchases.

6.1.5 What Should Be Considered as Treatment and When Should It Be Given?

Today, *If* have defined treatment as receiving a phone call from them. The implication of this is that there is a group in the data that have received a call (is *used*) and therefore qualify for the treatment group, but they have not answered the call (become *contacted*). It would be of interest to test if the model would perform better if every customer who is *used* but not *contacted* would be removed from the training data. By definition, everyone that is *used* belongs in the treatment group, but one can suspect that there is a difference in behavior between individuals that are just *used* compare to *contacted* (that answered the call). By training the model on data only containing individuals in the treatment group that are *contacted* would potentially simplify for the model to identify *persuadeables* and *sleeping dogs*. There is also a risk that the uplift is underestimated with the current data material, as one can suspect that the causal effect of treatment would be stronger if individuals who have not spoken to *If* (who are only *used*) would be moved to the control group.

Since the data set used in this report do not have a feature indicating the status of the individuals in the treatment group, it is impossible to re-train the model and exclude

used but not *contacted* and compare the result. However, if the definition of treatment were changed to *contacted*, it would imply that all *used* but not *contacted* would be moved from the treatment group to the control group. It is known that approximately 1/3 of the individuals in the treatment group is only *used*, which is equivalent to 194,000 individuals. Then, the most extreme possible changes would be if all of the moving individuals made a purchase or all of them made no purchase. As the treatment group only contain approximately 26,000 purchases, the worst possible outcome would be if all who performed a purchase and additionally $194,000 - 26,000 = 168,000$ individuals would move from the treatment group to control group. The total uplift would then be -10.90%. The best possible outcome would be if everyone who moved from the treatment group to the control group made no purchase, the total uplift would then be 6.27%.

If all *used* but not *contacted* were removed from the training data, it would most likely reduce the noise in the training data set, and it would be easier for the model to learn. However, if the individuals that only is *used* would be removed from the training data; the model would only apply to customers that answer the call. This would require another model which model whether the customer would answer the call or not. If the number of customers who do not pick up the phone is random, perhaps the second model would not be needed in order to capture uplift. However, it is also possible that the model still performs better when it is trained without individuals that are only *used* but the results are evaluated on a test data set containing individuals that are only *used*.

Moreover, *If* have reason to believe that some individuals have a systematic behavior to not answer the phone when an unknown number is calling. Therefore, by excluding individuals who are only *used* from the training data, there is a risk that a systematic bias is built into the model as their behavior will never be part of the training process. This potential systematic bias is the main argument for *If* to define treatment as calling an individual rather than talking to her.

6.1.6 Value of Uplift Modelling

If the value of having a customer is known, it is possible to compute the impact of the uplift model and translate this to increased earnings for the company. The result of the uplift models shows that targeting more customers is not always better and less efficient. It could be tough to convince the management at a company to set aside customers for the control group, which is required to measure and model uplift. If the management does not agree that the increased efficiency in identifying customers that can be persuaded for purchase is enough for reserving some customers for the control group, it could be of interest to show the potential value that can be gain if the value of a customer is known.

6.2 Control Group Size Simulations

6.2.1 Truncation Issues

The estimated parameters for the purchase rate from the real data have a variance that is too wide, considering what is an accepted value. In order to avoid values outside the accepted range, the multivariate normal distribution has been truncated. Since only approximately half of the distribution for the generated data is within the accepted interval if the covariance matrix estimate from real data is used, the mean of the generated data becomes higher than what is specified, see figure 4.5. In order to make the distribution fit better within the truncation limits for purchase, the variance has been reduced and in cases where the desired mean is close to the truncation limit, the specified parameter mean have been adjusted. It is necessary to adjust the parameters to ensure that the generated distribution have the desired mean because the simulated value is the probability of purchase, which is what is investigated. As the simulation is supposed to be general, it is possible to adjust the parameters. It is, of course, preferable if the simulated data correspond to the distribution in the real data so inference easily could be made, but it is not a necessity.

6.2.2 Control Group Size

According to the simulation study, the parameter setting with 60% and 70% treatment group allocation yield the smallest standard deviation for uplift for two purchase rate combinations each. This result is different from equation 4.13, which states that the optimal control group size would be 58.3% for purchase rate of 0.05 and 0.025, 57.9% for purchase rate of 0.1 and 0.05, 57.1% for purchase rate 0.2 and 0.1 and 51.3% for purchase rate 0.2 and 0.175. Equation 4.13 and the simulation study both try to identify the optimal treatment group allocation but in different situations, and therefore is the results not comparable. The equation returns the optimal proportion to maximize the precision in a difference; meanwhile, the simulations try to identify which of the tested treatment group allocation that is best for an uplift model.

Surprisingly, the result of the simulation study generated inconsistent results. As the purchase rate increases, the optimal treatment group allocation either increase or decrease; and as the subtraction between models increases, the optimal treatment allocation either increase or decrease. The result becomes hard to interpret and it is hard to draw conclusions from it. One can conjecture that the reason for the inconsistent result could be explained by the fact that explanatory features used barely had any correlation with the purchase feature which causes prediction of the purchase feature to be close to random. The simulated data lack *predictiveness*. That in conjunction with the fact that each replication has trained a model on a small sample, only 500 individuals, which makes the simulations sensitive. By supplying a correlation matrix with higher correlations between the target feature and the explanatory features, could potentially yield a more robust result that would be easier to interpret.

When evaluating the total uplift for each replication, it can be seen in figure 5, 6, 7 and 8 in the appendix, that the intended uplift is on average correct. This indicates that the simulated data is unbiased and capture what is intended. Preferably, the evaluation of uplift should be performed per segment and not on the total uplift, but as the true uplift per segment is unknown in each replication, it is not possible to evaluate it per segment.

One can also question the method used to evaluate which model performs best. The standard deviation per segment capture the variability of the estimated uplift for that segment but adding the segments together does not necessary becomes a good evaluation measure. By summing together the standard deviations assume that each segment is equally important in evaluating uplift. This is a naive assumption, as the first and the last segments are by far the most important segments for the model, as the top is the *persuadeables*, and the bottom is the *sleeping dogs* if the model work as intended.

The question remains, what is the required control group size? Considering the second rule of thumb from Radcliffe and Surry (2011); when the outcome is binary, the product of the overall uplift and the size of each sample should be at least 500. The rule depends on the magnitude of uplift, so if the total uplift in the data is 2%, each group would at least be required to contain 25,000 individuals. If the uplift were 20%, each group would require 2,500 individuals.

The second rule of thumb does not consider the proportion of individuals to the respective group; rather, it only considers the absolute amount of individuals needed in each group to measure and model uplift. The data set used in this report contains approximately 630,000 customers, with a total uplift of 2.6%. To capture uplift of magnitude 2.6% would require at least 19230 individuals in both groups according to the rule.

The proportion of individuals that is within the control group is approximately 8%, which is the equivalent of approximately 51,000 customers. This is far more than what is needed according to the rule of thumb. By taking the rule of thumb to the extreme, given the data set in this report, it would be enough with assigning 3% of all customers to control group. This result is very different from the result of equation 4.13 and the simulation study. Equation 4.13 states that the optimal treatment group allocation when the purchase rate is 5% in the treatment group, and 2.5% in the control group should be 57.9%. In absolute numbers, that would be assigning 364,770 individuals to control group, compared to 19,230 given by the second rule of thumb.

Though, one must remember that both equation 4.13 and the simulation study try to identify the optimal allocation to the treatment group meanwhile, Radcliffe and Surry (2011) second rule of thumb tries to estimate the minimal required amount in each group in order to be able to model uplift. So the measures have two different purposes.

Given the data set that has been used in this study, it leaves the user to assign somewhere between 57% - 97% of the individuals to the treatment group in a new study of similar

size and expected the magnitude of purchase. However, Radcliffe and Surry (2011) rule of thumb is just a measure of what is required, so one can argue that the closer to 57% - 70% allocation to treatment group would be better, though it is possible to assign as much as 97% to treatment group and still be able to measure uplift with magnitude of 2.6%.

The levels of allocation mentioned above do not consider the cost of giving treatment. The cost of treatment can be extensive, so the treatment group has to be kept to a minimum in order to afford to make the study or the opposite case, the cost of treatment is close to non-existing, and it is desired to provide as many treatments as possible to the subjects. Therefore, is it hard to provide a general rule of thumb that can be applied to all types of studies.

6.2.3 Continuous Collection for the Control Group

If uplift models are to be used continuously within the organization, it is of great importance to continue to set aside a control group as the campaigns continue. It is not recommended to launch a pilot study, where the uplift model is trained and return satisfactory results, and apply that model for countless future studies. The model had proven itself to perform well when it was trained; however, this is no guarantee that the model will continue to perform well as time passes. In order to ensure that the uplift model is continuing to perform, it is of importance that the control group is continuously set aside, so there is data to re-train and update the model parameters.

6.3 Statistical Power of a Test

On should always keep in mind that it is hard to model small changes in data, as it is likely that a small observed difference can be a consequence of normal variation in the sample that has been collected. Before a study, it is possible to estimate the minimum required sample size to capture a significant effect given the significance level and the power of the test, which can be performed by equation 4.12. When the objective is to model an effect size of magnitude 2%, approximately 1,500 individuals are required in both the treatment and control group to capture a significant result. If the effect size would be even smaller, more data is required. As the sample size increases, the uncertainty in the estimates decreases and yield a greater precision. This effect is captured by Radcliffe and Surry (2011) second rule of thumb which state that, if the uplift changed from a magnitude of 2% to 1%, the minimum required number of individuals in each group would change from 25,000 to 50,000.

6.4 Further Research

It would be of great interest for *If* to try to further develop the results and model uplift with a continuous target feature, but it would limit *If* to model uplift directly with

random forests. By modeling the continuous feature would most likely have a closer and stronger impact on the business than modeling just the binary purchase feature. Although if the binary target feature is replaced with a continuous, the Qini measures can still be utilized to evaluate and compare models.

To improve the received results, it would be of interest to see if a model trained on treatment group were all individuals who have not answered the phone call were excluded from the train data set. This report relies on the assumption that many features will improve the results in the model and the risk of overfitting the model is nothing to worry about as long as the result is evaluated on another data set than the one used to train the model. It would therefore be of interest to test to apply a more restrictive feature selection to the data set to see if it would improve the results as some authors have suggested (Radcliffe and Surry, 2011). As uplift is a second-order phenomenon, robustness in estimations is even more desirable than normal.

It would be interesting to perform the simulation studies once again with stronger correlations between the features to see if the results differ. As the correlations are close to zero between the purchase feature and the explanatory features, the impact of the multivariate distribution is close to none. If the correlation were stronger, it could potentially produce more consistent results which might be easier to interpret.

Another promising approach would be to rely on the causal inference work of Judea Pearl and Bayesian networks (Pearl, 2009) instead of Rubin's work with potential outcomes. By using graphs, causal inference of treatment could potentially be estimated and modeled without the use of the treatment group and control group. This would be an interesting path to explore as it potentially could yield the same or similar results without the need of having a control group.

7. Conclusion

During the last two decades, uplift models have been invented, further researched, and proven its value in several real-world applications to increase the efficiency in identifying whom to give treatment for optimal effect.

The first objective of this report was to develop uplift models with binary feature that capture the causal effect of treatment. This has been performed by two different methodologies, class transformation method, and uplift random forest. The class transformation method transforms the binary outcome and group belonging to a single feature that has been estimated by an extreme gradient boosting algorithm. The uplift random forest has been tested with four different splitting criterion's that try to maximize the distributional divergence between treatment group and control group, a concept taken from information theory. The divergence measures are χ^2, Kullback-Liebler, L1-norm divergence, and Euclidean distance.

The best model, uplift modeled directly with random forests using Euclidean distance as the splitting criterion manage to identify approximately 77% of the total number of purchases by giving treatment to the 50% of the highest scoring individuals in the treatment group. If the cost of giving treatment increases linearly, the cost of treatment could be cut in half, meanwhile approximately 3/4 of the earnings is maintained. Though, the model presented in this report only manage to capture 15% of the theoretically best model. All tested models perform similar results and are lacking predictiveness for the purchase feature in the data. If the cost of treatment and the earnings of a sell is known, this result can be translated into monetary terms and impact on the business can be measured.

In order to measure and model uplift, it is a necessity to have a control group who are not given treatment. Uplift is estimated by subtracting the results from the treatment group and the control group. To model uplift before the study is performed requires at least as ten times as many individuals in the control group compared to "just" measuring uplift after the study. The second objective of this report was to provide a recommendation of what would be an optimal treatment group allocation given a certain uplift methodology and classification algorithm without considering the cost of treatment. The optimal treatment group allocation to maximize the precision to detect a group difference have been computed, and a simulation study has been performed to evaluate what is optimal treatment group allocation in uplift modeling. The simulation study has used uplift random forest with Euclidean splitting criterion, where the best treatment group allocation is 70% if the purchase rate is 0.05 for the treatment group and 0.025 for the control group. The result from the simulation study is inconsistent, and conclusions drawn from it should be used with caution. The inconsistent result is believed to be because of the low correlation between the target feature and explanatory features in the simulated data. The general optimal treatment group allocation when the precision of a proportion should be maximized is approximately 58% according to equation 4.13.

There is a difference in what is considered the optimal to detect a difference between groups and what is required to be able to perform the study. A rule of thumb to ensure that the control group is of sufficient size is that the product of the overall uplift and the size of each sample should be at least 500. So, if the overall uplift would be 2%, the data requires at least 25,000 individuals to treatment group and control group. A new study with similar size and purchase rates could be performed with as little as 25,000 individuals in the respective group, which equal a treatment group allocation to approximately 97%.

Bibliography

Ascarza, Eva. "Retention Futility: Targeting High-Risk Customers Might Be Ineffective." In: *Journal of Marketing Research (JMR)* 55.1 (2018), pp. 80 –98. ISSN: 00222437.

Athey, Susan and Guido W Imbens. "Machine learning methods for estimating heterogeneous causal effects". In: *stat* 1050.5 (2015).

Athey, Susan and Guido Imbens. "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7353–7360.

Begg, Colin B. and Leslie A. Kalish. "Treatment Allocation for Nonlinear Models in Clinical Trials: The Logistic Model." In: *Biometrics* 40.2 (1984), p. 409. ISSN: 0006341X.

Breiman, L., C.J. Stone, J.H. Friedman, and R.A. Olshen. *Classification and regression trees.* (1)University of California: CRC Press, 1984. ISBN: 9781351460491.

Breslaw, J.A. "Random sampling from a truncated multivariate normal distribution". In: *Applied Mathematics Letters* 7.1 (1994), pp. 1 –6. ISSN: 0893-9659. DOI: `https://doi.org/10.1016/0893-9659(94)90042-6`.

Brittain, Erica and James J. Schlesselman. "Optimal Allocation for the Comparison of Proportions". In: *Biometrics* 38.4 (1982), pp. 1003–1009. ISSN: 0006341X, 15410420.

Chen, Tianqi and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM. 2016, pp. 785–794.

Chickering, David Maxwell and David Heckerman. "A decision theoretic approach to targeted advertising". In: *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. 2000, pp. 82–88.

Chodrow, Philip. "Divergence, Entropy, Information: An Opinionated Introduction to Information Theory". In: *CoRR* abs/1708.07459 (2017). arXiv: `1708.07459`.

Guelman, L., M. Guillén, and A.M. Pérez-Marín. "Uplift random forests." In: *Cybernetics and Systems* 46.3-4 (2015), pp. 230–248. ISSN: 10876553.

Guelman, Leo. *uplift: Uplift Modeling*. R package version 0.3.5. 2014.

Guelman, Leo, Montserrat Guillen, and Ana M. Perez-Marin. "Random Forests for Uplift Modeling: An Insurance Customer Retention Case." In: *Modeling and Simulation in Engineering, Economics, and Management: International Conference, MS 2012, New Rochelle, NY, USA, May 30-June 1, 2012 Proceedings*. Royal Bank of Canada: Lecture Notes in Business Information Processing, vol. 115. New York and Heidelberg: Springer, 2012, pp. 123 –133.

Gutierrez, Pierre and Jean-Yves Gérardy. "Causal Inference and Uplift Modelling: A Review of the Literature". In: *International Conference on Predictive Applications and APIs*. 2017, pp. 1–13.

Hansotia, Behram and Brad Rukstales. "Incremental value modeling". In: *Journal of Interactive Marketing* 16.3 (2002), pp. 35–46.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2008.

Holland, Paul W. "Statistics and Causal Inference". In: *Journal of the American Statistical Association* 81.396 (1986), pp. 945–960. ISSN: 01621459.

If. *Handling Personal Data*. Accessed May 17, 2019. URL: `https://www.if-insurance.com/about-the-website/handling-of-personal-data`.

Jaroszewicz, Szymon and Piotr Rzepakowski. "Uplift modeling with survival data". In: *ACM SIGKDD workshop on health informatics (HI-KDD'14), New York City*. 2014.

Jaśkowski, Maciej and Szymon Jaroszewicz. "Uplift modeling for clinical trial data". In: 2012.

Kotler, Philip and Gary Armstrong. *Principles of marketing*. Pearson education, 2010.

Lo, Victor S. Y. "The True Lift Model: A Novel Data Mining Approach to Response Modeling in Database Marketing". In: *SIGKDD Explor. Newsl.* 4.2 (Dec. 2002), pp. 78–86. ISSN: 1931-0145. DOI: `10.1145/772862.772872`.

Morgan, Stephen L. and Christopher Winship. *Counterfactuals and Causal Inference*. 2nd ed. Cambridge University Press, 2015.

Naranjo, Oscar Mesalles. "Testing a New Metric for Uplift Models". In: (2012).

Nassif, Houssam, Yirong Wu, David Page, and Elizabeth Burnside. "Logical Differential Prediction Bayes Net, improving breast cancer diagnosis for older women". In: *AMIA Annual Symposium Proceedings*. Vol. 2012. American Medical Informatics Association. 2012, p. 1330.

Pearl, Judea. *Causality*. Cambridge university press, 2009.

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. 1st ed. Wiley, 2016.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python ". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

Quinlan, J. Ross. "Induction of decision trees". In: *Machine learning* 1.1 (1986), pp. 81–106.

Radcliffe, NJ. "Hillstrom's MineThatData email analytics challenge: An approach using uplift modelling". In: *Stochastic Solutions Limited* 1 (2008), pp. 1–19.

Radcliffe, Nicholas J and Patrick D Surry. "Differential response analysis: Modeling true response by isolating the effect of a single action". In: *Credit Scoring and Credit Control VI. Edinburgh, Scotland* (1999).

Radcliffe, Nicholas J. and Patrick D. Surry. "Real-World Uplift Modelling with Significance-Based Uplift Trees." In: (2011).

Radcliffe, Nicholas. "Using control groups to target on predicted lift: Building and assessing uplift model". English. In: *Direct Marketing Analytics Journal* (2007), pp. 14–21.

Radclifte, Nicholas J and Rob Simpson. "Identifying who can be saved and who will be driven away by retention activity." In: *Journal of Telecommunications Management* 1.2 (2008).

Rubin, Donald B. "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of educational Psychology* 66.5 (1974), p. 688.

Rubin, Donald B. and Richard P. Waterman. "Estimating the Causal Effects of Marketing Interventions Using Propensity Score Methodology." In: *Statistical Science* 21.2 (2006), p. 206. ISSN: 08834237.

Rzepakowski, Piotr and Szymon Jaroszewicz. "Decision trees for uplift modeling with single and multiple treatments". In: *Knowledge and Information Systems* 32.2 (2012), pp. 303–327.

— "Uplift modeling in direct marketing". In: *Journal of Telecommunications and Information Technology* (2012), pp. 43–50.

Siegel, Eric. "Uplift modeling: Predictive analytics can't optimize marketing decisions without it". In: *Prediction Impact white paper sponsored by Pitney Bowes Business Insight* (2011).

Sołtys, Michał, Szymon Jaroszewicz, and Piotr Rzepakowski. "Ensemble methods for uplift modeling". In: *Data Mining and Knowledge Discovery* 29 (Nov. 2015). DOI: 10.1007/s10618-014-0383-9.

Stefan, Wilhelm and B G Manjunath. *tmvtnorm: Truncated Multivariate Normal and Student t Distribution*. R package version 1.4.10. 2015.

Wager, Stefan and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests". In: *Journal of the American Statistical Association* just-accepted (2017).

Wang, Hansheng and Shein-Chung Chow. "Sample size calculation for comparing proportions". In: *Wiley StatsRef: Statistics Reference Online* (2014).

Yi, Robert. *Pyift*. https://github.com/wayfair/pylift. [Online; accessed 25-January-2019]. 2018.

Zaniewicz, Łukasz and Szymon Jaroszewicz. "Support vector machines for uplift modeling". In: *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*. IEEE. 2013, pp. 131–138.

Appendices

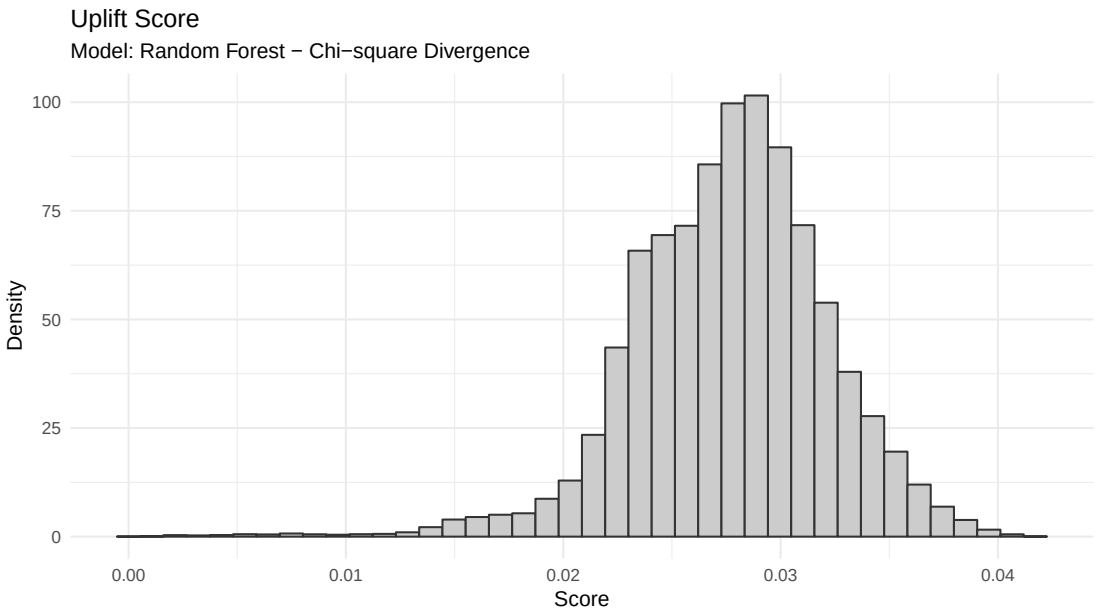


Figure 1: Histogram of Scores for Random Forest with Chi-square Divergence as splitting criterion

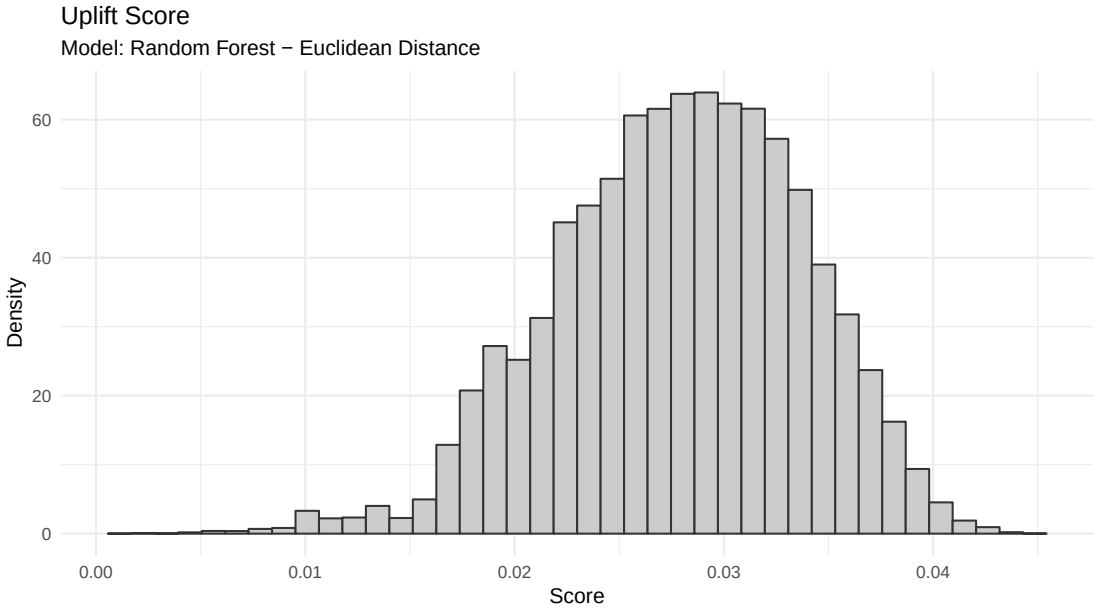


Figure 2: Histogram of Scores for Random Forest with Euclidean Distance as splitting criterion

Table 1: Performance table for Random Forest Chi-square Divergence, uplift per segment

Bin	n_t	n_c	$n_{t,1}$	$n_{c,1}$	$r_{t,1}$	$r_{c,1}$	Uplift
1	7200	708	571	14	0.079306	0.019774	0.059532
2	7197	710	541	14	0.07517	0.019718	0.055452
3	7241	667	496	15	0.068499	0.022489	0.04601
4	7311	596	483	19	0.066065	0.031879	0.034186
5	7315	592	455	17	0.062201	0.028716	0.033485
6	7327	581	404	14	0.055139	0.024096	0.031042
7	7326	581	364	4	0.049686	0.006885	0.042801
8	7312	595	315	9	0.04308	0.015126	0.027954
9	7294	614	295	11	0.040444	0.017915	0.022529
10	7307	601	295	17	0.040372	0.028286	0.012086
11	7251	655	249	8	0.03434	0.012214	0.022126
12	7232	676	249	4	0.03443	0.005917	0.028513
13	7266	641	227	7	0.031241	0.01092	0.020321
14	7180	727	221	10	0.03078	0.013755	0.017025
15	7169	739	192	5	0.026782	0.006766	0.020016
16	7203	704	206	10	0.028599	0.014205	0.014395
17	7186	721	217	8	0.030198	0.011096	0.019102
18	7284	624	202	6	0.027732	0.009615	0.018117
19	7424	483	232	16	0.03125	0.033126	-0.00188
20	7463	445	294	24	0.039394	0.053933	-0.01454

Table 2: Performance table for Random Forest Kullback-Liebler Divergence, uplift per segment

Bin	n_t	n_c	$n_{t,1}$	$n_{c,1}$	$r_{t,1}$	$r_{c,1}$	Uplift
1	7160	748	536	15	0.07486	0.020053	0.054807
2	7178	729	548	9	0.076344	0.012346	0.063999
3	7158	750	506	25	0.07069	0.033333	0.037357
4	7186	721	487	26	0.067771	0.036061	0.03171
5	7277	630	451	16	0.061976	0.025397	0.036579
6	7368	540	372	9	0.050489	0.016667	0.033822
7	7369	538	371	8	0.050346	0.01487	0.035476
8	7362	547	325	4	0.044146	0.007313	0.036833
9	7357	550	293	12	0.039826	0.021818	0.018008
10	7318	588	293	5	0.040038	0.008503	0.031535
11	7273	634	243	7	0.033411	0.011041	0.02237
12	7279	629	244	9	0.033521	0.014308	0.019213
13	7191	716	239	12	0.033236	0.01676	0.016476
14	7161	746	228	7	0.031839	0.009383	0.022456
15	7169	739	214	10	0.029851	0.013532	0.016319
16	7193	714	233	6	0.032393	0.008403	0.023989
17	7242	665	190	7	0.026236	0.010526	0.01571
18	7326	582	169	9	0.023069	0.015464	0.007605
19	7450	457	283	15	0.037987	0.032823	0.005164
20	7471	437	283	21	0.03788	0.048055	-0.01018

Table 3: Performance table for Random Forest L1-norm Divergence, uplift per segment

Bin	n_t	n_c	$n_{t,1}$	$n_{c,1}$	$r_{t,1}$	$r_{c,1}$	Uplift
1	7115	793	576	26	0.080956	0.032787	0.048169
2	7033	875	561	18	0.079767	0.020571	0.059195
3	7133	775	451	21	0.063227	0.027097	0.03613
4	7159	747	464	22	0.064814	0.029451	0.035362
5	7148	759	452	19	0.063234	0.025033	0.038202
6	7167	741	417	15	0.058183	0.020243	0.03794
7	7160	747	399	20	0.055726	0.026774	0.028952
8	7238	669	380	16	0.052501	0.023916	0.028584
9	7318	590	357	8	0.048784	0.013559	0.035224
10	7410	507	367	11	0.049528	0.021696	0.027831
11	7470	428	298	5	0.039893	0.011682	0.028211
12	7459	450	256	5	0.034321	0.011111	0.02321
13	7439	470	275	8	0.036967	0.017021	0.019946
14	7381	524	198	4	0.026826	0.007634	0.019192
15	7274	632	145	3	0.019934	0.004747	0.015187
16	7228	702	167	3	0.023105	0.004274	0.018831
17	7261	674	171	3	0.02355	0.004451	0.019099
18	8316	742	180	5	0.021645	0.006739	0.014906
19	6254	455	131	6	0.020947	0.013187	0.00776
20	7525	380	263	14	0.03495	0.036842	-0.00189

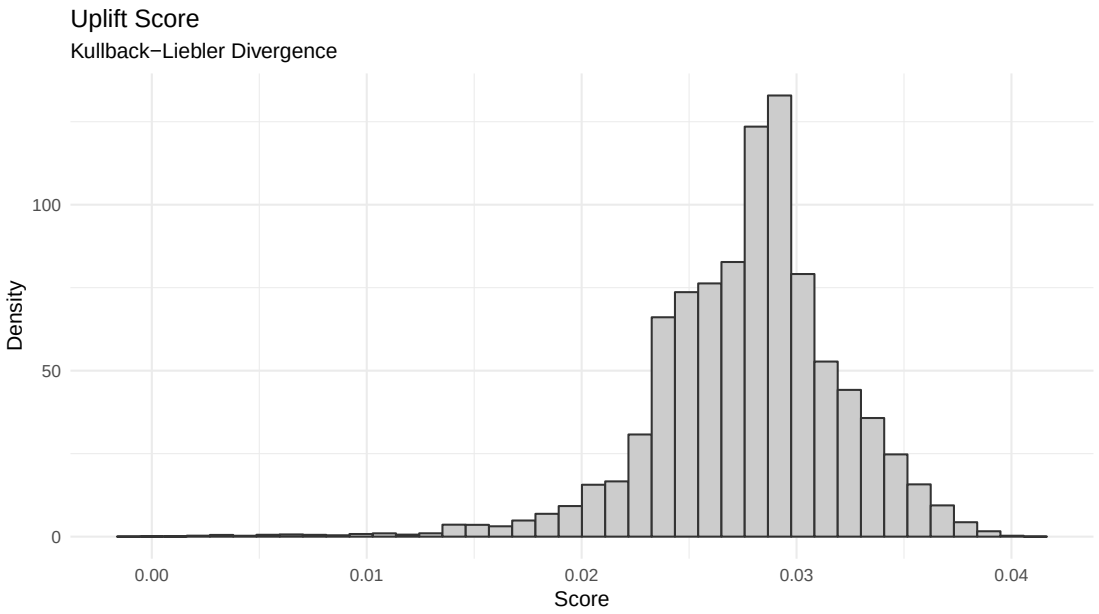


Figure 3: Histogram of Scores for Random Forest with Kullback-Liebler Divergence as splitting criterion

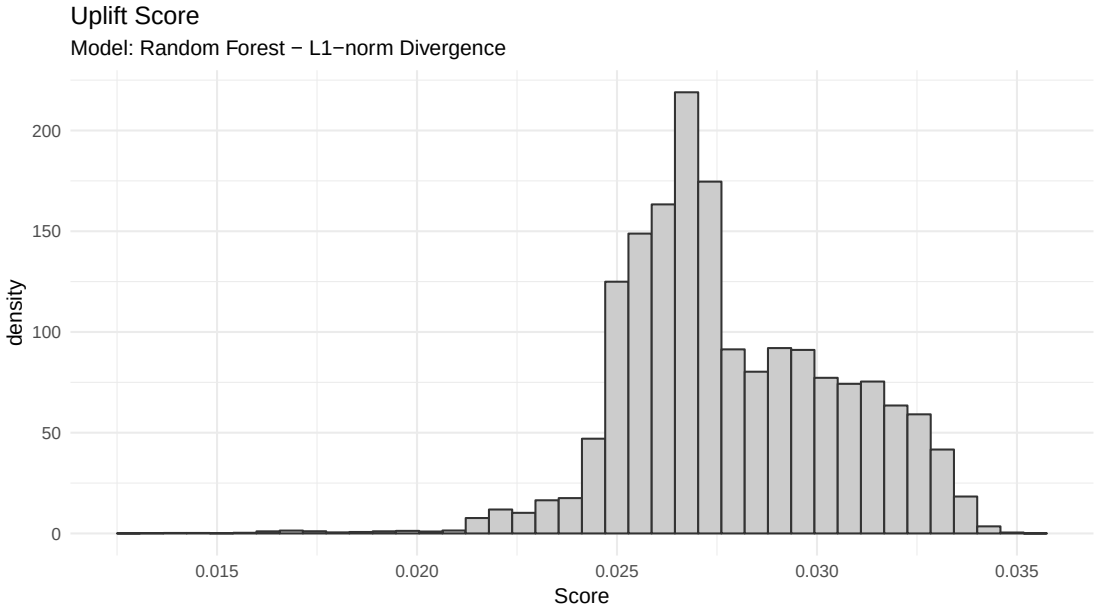


Figure 4: Histogram of Scores for Random Forest with L1-norm Divergence as splitting criterion

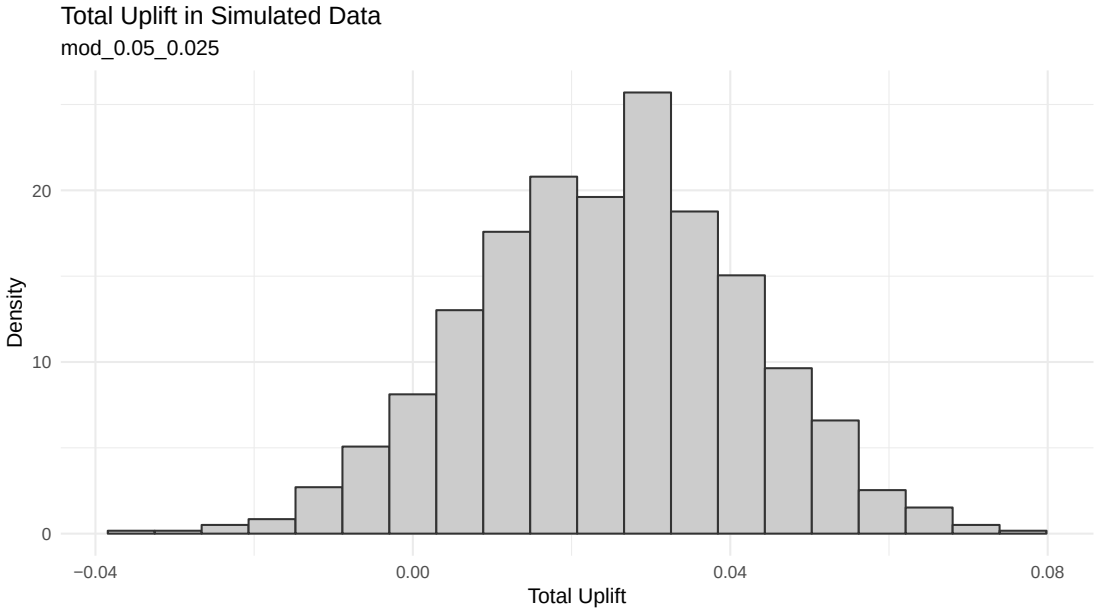


Figure 5: Total uplift in from each replication with purchase rate of 5% in the treatment group and 2.5% in the control group

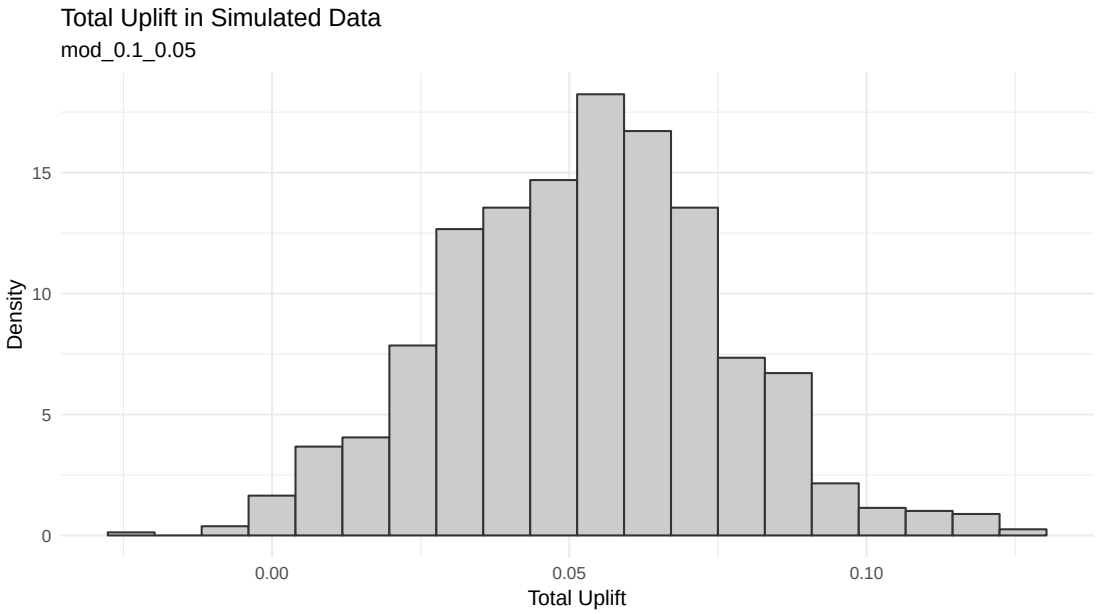


Figure 6: Total uplift in from each replication with purchase rate of 10% in the treatment group and 5% in the control group

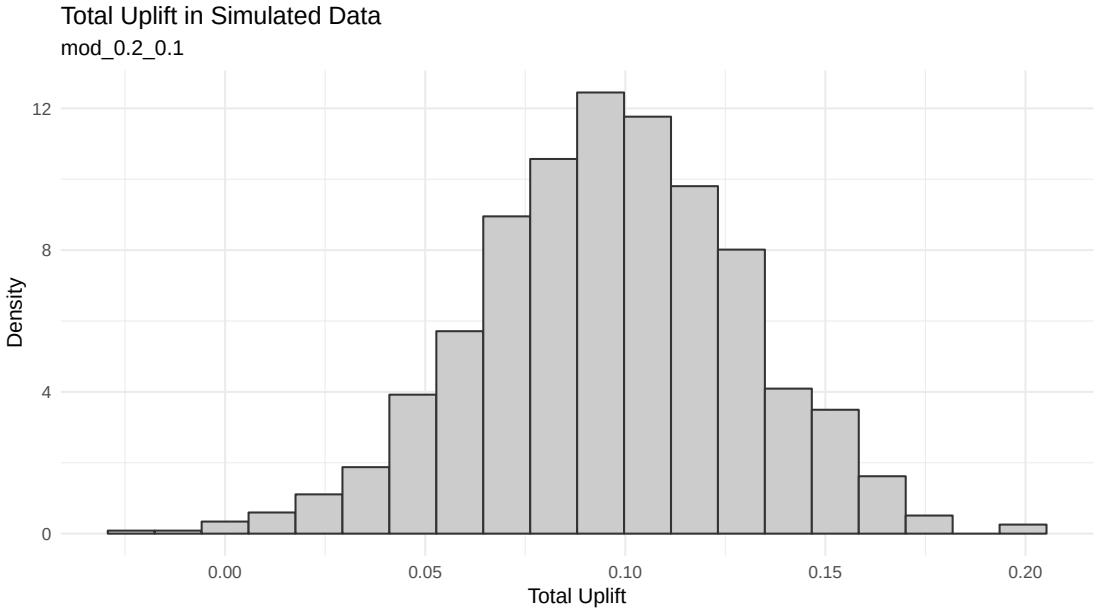


Figure 7: Total uplift in from each replication with purchase rate of 20% in the treatment group and 10% in the control group

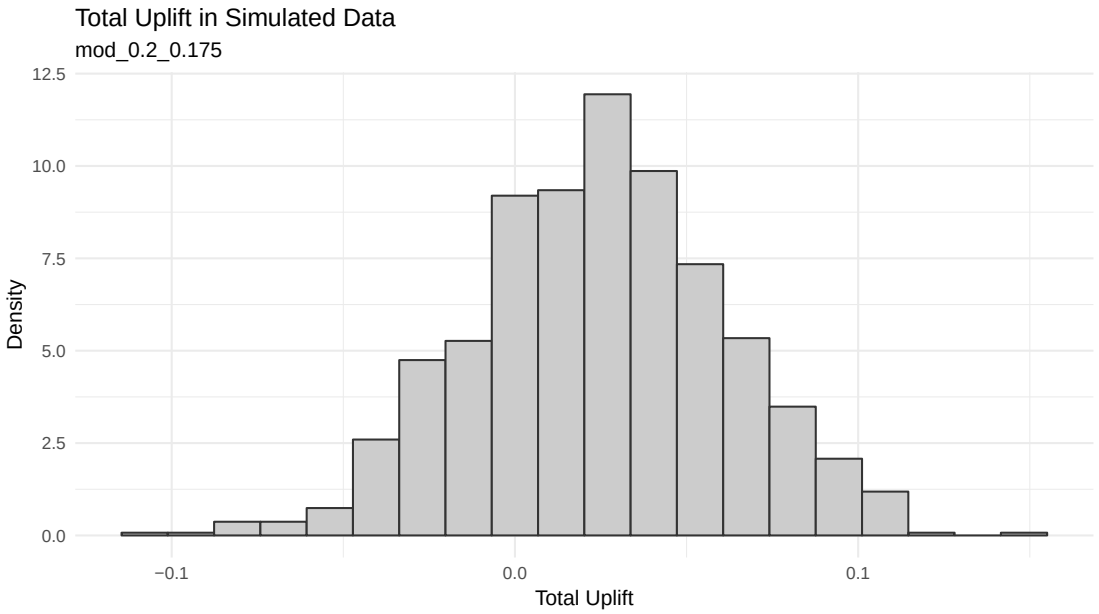


Figure 8: Total uplift in from each replication with purchase rate of 20% in the treatment group and 17.5% in the control group