



# PREDICTION OF CONVERSION RATES IN ONLINE MARKETING

A study of the application of logistic regression  
for predicting conversion rates in online  
marketing.

## 1 Abstract

This thesis was written in collaboration with an anonymous European automotive company, Company X, which uses online marketing as a part of their business model. In online marketing it is of interest to estimate conversion rates, that is the quota of a population at an initial state that will go on to perform a certain action. The action could be, but is not limited to, clicking on an advertisement, interacting in a certain way with the advertisers webpage, or buying a product.

If the advertiser can estimate the value of the performed action, and the conversion rate to the action, the advertiser can then calculate the value of the initial state. In extension, this means that if a company knows the life time value of a customer, and can estimate the conversion rate from someone clicking on one of their advertisements to becoming a customer, they can calculate the value of that click.

Generally online marketing space is sold through auctions. Different companies bid for the same given advertising space depending on the expected value of the space and pay for exposure. Exposure is either measured in how many users that has seen the ad (impressions) or how many users that have interacted with the ad (usually measured in clicks). Due to this, if a company can improve the precision of how they estimate the value of an impression or click they can spend their online marketing budget more effectively. Considering the size and rapid growth of the online marketing market, this is of high interest.

In this thesis a logistic regression modeling approach was compared to a group average approach for predicting conversion rates. The group average approach is based on grouping different advertisements that have few observations into bigger populations and then using the average of the bigger population. The thesis finds that in most cases logistic regression models seems preferable. However, when the variance of the conversion rates is large, the group average model can be preferable.

## 2 Keywords

Online marketing, conversion rates, logistic regression, latent variables, machine learning.

## Table of Contents

1	Abstract .....	1
2	Keywords .....	1
3	Introduction.....	5
3.1	Research questions .....	5
3.2	Company background .....	6
3.3	Outline.....	6
4	Online marketing at a glance.....	7
4.1	What is an online advertisement?.....	7
4.2	The acquisition funnel.....	7
4.3	Types of online advertising.....	9
4.3.1	Search Engine Marketing .....	9
4.3.2	Display.....	9
4.4	Networks.....	10
4.4.1	SEM – Google AdWords.....	10
4.4.2	GDN – Google display network .....	10
4.4.3	FBA - Facebook Ads .....	10
4.4.4	Other display networks .....	11
4.5	Intent.....	11
4.6	Account hierarchy.....	11
4.7	Active interventions.....	12
5	Theory.....	13
5.1	Logistic regression.....	13
5.1.1	Iteratively Reweighted Least Squares.....	14
5.1.2	Interpretation of variables .....	15
5.1.3	Feature selection with Lasso .....	16
5.2	Cross validation .....	17
5.3	K-fold cross validation .....	18
5.3.1	One standard error rule.....	18
5.4	Model evaluation.....	19
5.5	Chi-squared test for independence of variables .....	20
5.6	Pearson residuals.....	20
6	Methodology.....	21
6.1	Data description .....	21
6.2	Data pre-processing.....	22
6.3	Dividing data into modeling and test set.....	23
6.4	Models .....	23

6.4.1	Group Average.....	23
6.4.2	Logistic regression on contextual data .....	26
6.4.3	Logistic regression with early conversion rates .....	27
6.4.4	Logistic regression with latent variables .....	28
6.5	Modeling procedure.....	28
6.5.1	Pre-processing.....	28
6.5.2	Fitting the group average model .....	30
6.5.3	Fitting the logistic regression models .....	31
6.5.4	Results .....	42
6.6	Ad type .....	42
6.6.1	Group average .....	43
6.6.2	Univariate analysis .....	43
6.6.3	Multivariate analysis.....	43
6.7	Asses performance .....	45
7	Results.....	46
7.1	Performance of models.....	46
7.2	Final models .....	47
7.2.1	Display .....	47
7.2.2	Facebook.....	47
7.2.3	Google display network .....	48
7.2.4	Search Engine Marketing .....	48
7.3	Performance due to ad type covariate.....	49
8	Discussion and conclusions .....	50
8.1	How accurate are the models? .....	50
8.2	Group average vs logistic regression models.....	50
8.3	Using latent variables.....	50
8.4	Why group average fails for Facebook.....	50
8.5	Variability when using observed values .....	51
8.6	Further improvements .....	51
8.6.1	Grouped data.....	51
8.6.2	Burn out effect.....	51
8.6.3	Features from the ad creative .....	52
9	Further research .....	53
10	References .....	54
11	Appendix 1.....	56
12	Appendix 2.....	57

# Vocabulary

**User** – An internet user. A person who is accessing the internet.

**Device** – A machine that a user is accessing the internet from, i.e. desktop, tablet or mobile.

**Impression** – An ad being shown on a users device.

**Click** – A user that has clicked on an advertisement.

**Lead** – A user that has performed a desired action on the advertisers webpage. Could be but not limited to placing items in a shopping cart or submitting the email address.

**Acquisition** – A user that has become a customer.

**Conversion** – A user that has gone from one predetermined step in the customer journey to a later step.

**Conversion rate (CVR)** – The quotient of users that go from one predetermined step in the customer journey to a later step.

**Click through rate (CTR)** – The quotient of users that go from impression to click. This is a type of conversion rate.

**Ad creative** – The image/text/rich media shown to users in an advertisement.

**Ad placement** – A type of address on the internet. Defines where on the internet the ad is shown and to whom it is shown.

**Ad group** – A set of ad placements

**Campaign** – A set of ad groups.

**Account** – A set of campaigns.

**Ad network** – A collection of websites where advertisers can publish their ads.

**Search engine marketing (SEM)** – A form of advertising where ads are shown in combination with the results for a search query.

**Display marketing** – A form of advertising where ad media is shown on a webpage.

### 3 Introduction

Since the first clickable online advert was published in 1993 [1] there has been a rapid development in online marketing. Today there are numerous types of online marketing, and methods for using them are getting more sophisticated. Two important things that set online marketing apart from offline marketing are targeting and feedback.

The difference in targeting stems from that in the offline world an advertiser cannot filter whom the ad is showed to. If an ad is published in a magazine, the ad is shown to all readers and the advertiser has to pay for the total exposure. If the advertiser is only interested in a subset of the readers, the rest of the exposure represents a type of waste.

Online however, there is a possibility of going beyond the placement, and target on other attributes such as age, gender, income and interests. This means that advertisers can find their target group more effectively, and don't need to pay for exposure outside of that target group. It has also brought about a shift where advertisers can target individuals rather than big groups. Advertisement has gone in to a new paradigm where advertisers can buy single impressions rather than billboards.

Online marketing allows for an entirely different feedback system than offline marketing. Advertisers can get exact information on how many people has viewed an ad, how many of them has clicked, and which users moved on to making a purchase in real time. This makes it possible to perform more granular customer analyses. Such analyses can support increased targeting efforts and lead to a better understanding of existing and potential customers.

In the new paradigm where advertisers are bidding for single impressions or clicks, a crucial part of an effective marketing strategy is to estimate the value of an impression or click. This estimation is done through predicting how likely it is that the user that the ad is shown to is going to convert into a customer and predicting how much that customer will be worth. This thesis will focus on the former of those two predictions.

#### 3.1 Research questions

This thesis aims to answer three questions:

1. Can logistic regression be used to improve predictions of conversion rates in online marketing relative to group average models?
2. Can an approach using latent variables improve predictions of conversion rates for logistic models?
3. Is there a difference in how well suited different models are for different platforms?

To answer the first question a group average model is created. The group average model uses the mean for the category of ad placements that the ad placement being predicted belongs to. This base-case model is then compared to models created through logistic regression.

To answer the second question latent variables are introduced to the model by predicting earlier conversion steps. The residuals from those predictions are then used

as input to the new model. The logic behind the procedure is that the residuals can contain information about the quality or features of the ad creative.

Different online marketing platforms have different types of traffic and provide different types of data. It is thus possible that some models will work better on some platforms and worse on others. To answer the third question the performance of the different models is compared across different online advertising platforms.

### 3.2 Company background

This thesis was conducted in cooperation with a European automotive startup that wishes to be anonymous. The company, hereafter referred to as Company X, has shared their online marketing data for the purposes of this thesis.

The company's online marketing efforts are aimed at acquiring new customers. In this process potential customers, called users, go through a customer journey. That journey starts with them interacting with an advertisement. Then they go through several steps on the company's online platform until potentially converting into a customer. These steps are more thoroughly described in chapter 4.2. The data provided by Company X is described in chapter 6.1.

Since the conversion rates are sensitive information for company X, conversion rates are standardized in the plots of the report.

### 3.3 Outline

In chapter 4 some background about online marketing is given as a frame to the problem of predicting conversion rates. Chapter 5 gives the statistical background used to solve the problem and in chapter 6 the method behind the modeling is described.

The results are presented in chapter 7, followed by a discussion and conclusions in chapter 8, and suggestions for further research in chapter 9.

## 4 Online marketing at a glance

A basic understanding of online marketing is necessary to interpret the results of this report as well as understanding the methodology. This chapter aims to give a brief but sufficient background for reading the thesis.

### 4.1 What is an online advertisement?

An online advertisement is a paid message published online. An online advertisement consists of two parts:

1. Ad creative – the text/image/video etc. that contains the message that the advertiser wants to convey.
2. Ad placement – The location on the internet where the ad is published. This could be on the top of a certain web page.

An online advertisement is normally clickable, redirecting users that click on the ad to a landing page.

### 4.2 The acquisition funnel

A good model for describing online marketing is the acquisition funnel. In the acquisition funnel potential customers are going through different steps in the customer journey, before becoming customers. Different companies can have funnels that vary slightly. Figure 1 presents a general acquisition funnel that describes the acquisition process at Company X well.

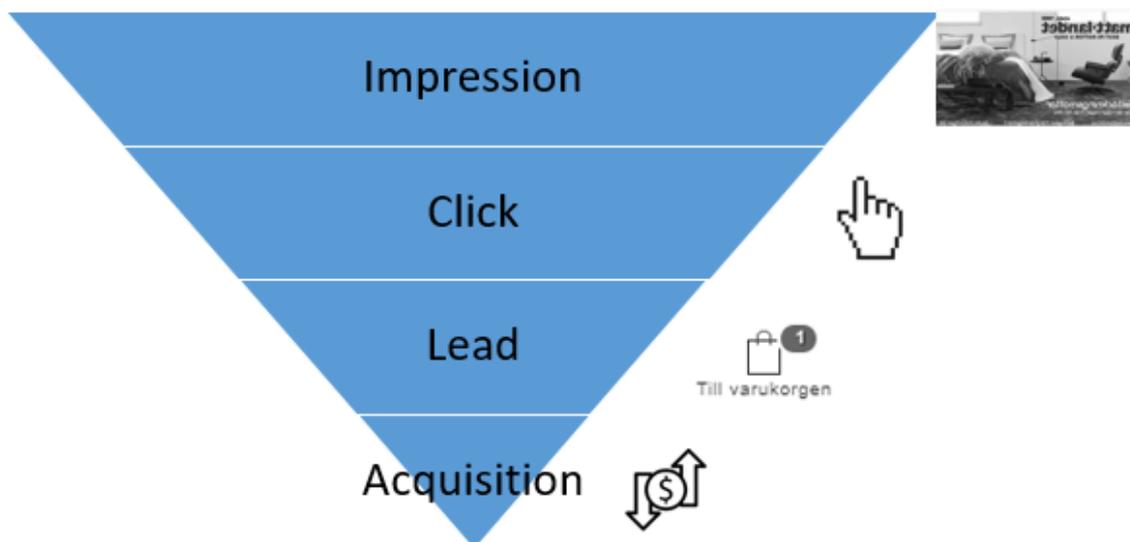


Figure 1, the online marketing funnel of Company X. A potential customer starts of by seeing an ad, clicks on the advertisement, creates a lead on the company's platform and becomes a customer.

The customer journey through the funnel can be described in the following way:

- Impression – The user sees the advertisement
- Click – The user clicks on the advertisement and is re-directed to the advertisers web-page.
- Lead – The user performs a pre-determined task on the advertisers web-page.

- Acquisition – The user becomes a customer by buying the product or service that is advertised.

When a user goes from one step to another in the acquisition funnel it is called a *conversion*. The quota of users converting from one step to another step is called the *conversion rate* (CVR). The first conversion rate in the funnel, from impression to click, is called *click through rate* (CTR). The formula for calculating the conversion rate is

$$CVR_{State\ 1,State\ 2} = \frac{Size\ of\ population_{State\ 2}}{Size\ of\ population_{State\ 1}}$$

where

$$Size\ of\ population_{State\ 2} \leq Size\ of\ population_{State\ 1}$$

In the acquisition funnel model, a user cannot get to a later step of the funnel without going through the earlier steps. Users can also leave the funnel at any point. This means that conversion rates will always be in the interval [0,1].

Between each step in the funnel there is also a time difference. The time differences for Company X are described in figure 2 together with typical number of users at each step.

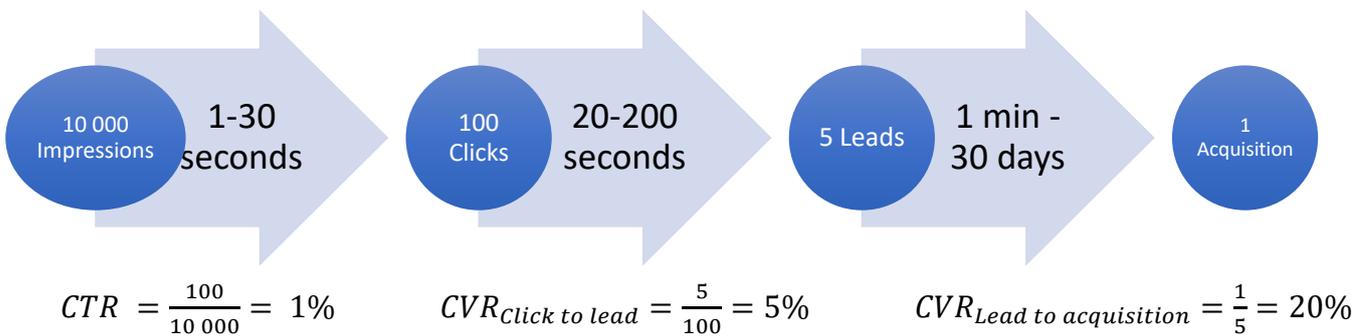


Figure 2, timespans in the customer journey are presented on the arrows between the different steps. An example of the number of users at each step is presented within the circles.

The time before Company X receives feedback in the first two conversion steps is relatively short compared to the third. Provided that there is a big enough amount of traffic, the measured conversion rates can be used to estimate future conversion rates. However as there is a long time-span before Company X receives feedback for the third conversion step, it's possible that the feedback that it has received from 30 days ago doesn't represent the current environment. Considering also that the number of users creating a lead is only a fraction of the clicks, which is a fraction of the impressions, there is also significantly less data to support the later conversion rates.

These two factors makes it especially interesting to make predictions for the last conversion rate, from lead to acquisition. This conversion rate will be of focus for the predictions in this report.

### 4.3 Types of online advertising

There is a myriad of different types of online marketing. This thesis will be limited to Search Engine Marketing (SEM) and display marketing.

#### 4.3.1 Search Engine Marketing

SEM is paid advertisement that appears when a user is conducting a web search on a search engine. These ads are normally related to the users search query, where advertisers bid for keywords that are related to their business. The biggest players in SEM are Google, Yahoo and Bing [2].

The image shows a screenshot of a search engine results page for the query "Pay Per Click". The results are numbered 1 through 7. Items 1-3 are advertisements, while items 4-7 are organic search results. Item 4 is another advertisement. The ads include titles, URLs, and short descriptions. The organic results include a Wikipedia entry and a Trada article. The text "Ads - Why these ads?" is visible above the ad items.

- 1. Pay-Per-Click Advertising** 1 (877) 490 2353  
[www.google.com/AdWords](http://www.google.com/AdWords)  
Reach Your Customers With Google. Get Your Ad Online In Just 15 Mins!
- 2. Pay Per Click (PPC) - Get More Leads and Sales with PPC.**  
[www.hanaginmarketing.com](http://www.hanaginmarketing.com)  
Contact the **Pay Per Click** Experts.
- 3. Pay Per Click Service - 20 Billion Impressions Per Month.**  
[www.advertise.com/Pay-Per-Click-Service](http://www.advertise.com/Pay-Per-Click-Service)  
Get \$100 in Free Advertising.
- 4. Easily Manage PPC Ads**  
[www.doubleclick.com](http://www.doubleclick.com)  
Manage all paid search ads from one interface with DART for Search.
- 5. Pay per click advertising**  
[www.wischoiceadco.com](http://www.wischoiceadco.com)  
Professional Adwords PPC Manager  
Flat Monthly Fee / 1-877-729-7007
- 6. Pay Per Click Marketing**  
[www.lqemarketing.com/Pay\\_Per\\_Click](http://www.lqemarketing.com/Pay_Per_Click)  
Here to Help You Succeed. We Offer Complete Transparency On Our Work!
- 7. Sick of Pay Per Click?**  
[www.trada.com/Learn-More](http://www.trada.com/Learn-More)

**Pay per click - Wikipedia, the free encyclopedia**  
[en.wikipedia.org/wiki/Pay\\_per\\_click](http://en.wikipedia.org/wiki/Pay_per_click)  
**Pay per click (PPC)** (also called **Cost per click**) is an Internet advertising model used to direct traffic to websites, where advertisers pay the publisher (typically a ...  
Determining cost per click - History - See also - References

Figure 3, example of SEM advertisement. This one is from Google AdWords and the ad positions are marked out with numbers. [3]

Figure 3 presents an example of a SEM advertisement. The ads are displayed together with the results of the users search query. There can be different numbers of ads displayed for a certain search, and the displayed ads are associated with a position. A lower position number is general associated with a higher click through rate [4]. It has been shown that the position can have an influence on conversion rates at later stages of the customer journey as well [5].

#### 4.3.2 Display

Display advertising is defined as advertising on webpages through different forms such as banner ads, rich media and more [6]. Targeting in display is primarily done through finding webpages with an audience that is interested in the product that is advertised. Beyond just targeting through finding a relevant webpage, some display platforms (such as Google display network and Facebook ads) can also target by user information, device information, location etc. An example of targeting on location is presented in figure 4.

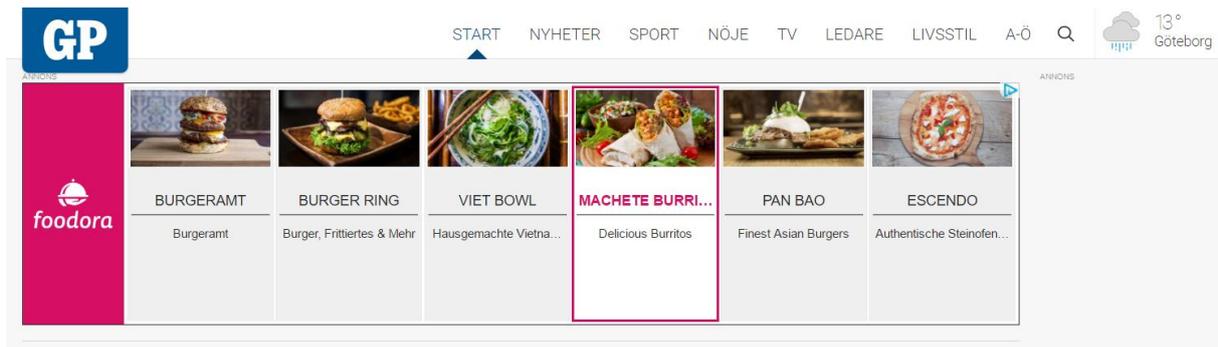


Figure 4, example of a display advertisement. In this case, the network providing the webpage with advertisements have additional information about the user. Through this information, it can display suggestions of restaurants close to the user's location in Germany on the page of a Swedish newspaper. [7]

A major difference between display and SEM is that in the latter the user is actively writing search queries, and looking for certain information. The search query carries valuable information since it tells advertisers what the user is interested in. Advertisers can then show their ads to users that are interested in things related to their product, and actively looking for information online.

In display advertisement, the user is not necessarily looking for information and the advertisements are shown in a more general context. This suggests that the users would be less interested in advertisements in general, since the advertisements are not what the user is on the webpage for.

#### 4.4 Networks

An online advertising network or ad network is a company that connects advertisers to web sites that want to host advertisements [8]. For both SEM and display marketing there are multiple networks. Each network provides advertisers with different types of data and can also have different characteristics in terms of audience. The data for this thesis was collected from the Google Adwords network, the Google display network, Facebook Ads and a selection of other smaller display advertisement networks.

##### 4.4.1 SEM – Google AdWords

Google AdWords is a search engine marketing network. It allows the advertisers to bid for exposure to users that have typed in different search queries. They also allow for geographical targeting.

One special feature in Google AdWords is that bids in auctions for ad placements are weighted by a quality score. The quality score is a measure of how relevant Google thinks that the advertisers web page is to a search query. Google shares this measure with the advertisers [9].

##### 4.4.2 GDN – Google display network

The Google display network also runs on the Google AdWords network, but it shows the advertisements on websites that are not search engines. It provides the same type of data as Google AdWords.

##### 4.4.3 FBA - Facebook Ads

Facebook Ads is a display network showing display ads on Facebook and Instagram. It has rich targeting features such as gender, age, location, interests etc.

#### 4.4.4 Other display networks

Company X is also advertising on many smaller advertisement networks. For these networks, the amount of feedback data tends to be lower. These providers generally have less targeting possibilities, as they do not have access to the same amount of user information as Google and Facebook does. For the purposes of this thesis these different smaller networks have been grouped together for the analysis.

#### 4.5 Intent

An important concept when discussing online advertising is intent, referring to why users interact with a certain advertisement. Consider a banner ad on a webpage, if that ad is for selling a shoe and the webpage is a page with shoe reviews it's possible to imagine that a visitor of that webpage has the intent to buy shoes. However, if the same ad is showed on a fashion blog it's not as clear that a visitor of that page is interested in buying shoes, a click on the ad in that case might be because the person is interested in fashion but has limited interest of buying shoes at that time.

Just as the context of the ad placement has an impact on which intent one could expect so does the ad creative. If a text of the ad on the fashion blog reads "get your new pair of shoes in 24 hours!" people who click on such an ad probably have a different intent than someone clicking an ad saying, "View the new fashionable shoes of the summer". It is not difficult to imagine how these ads could have different conversion rates, even if they would be placed on the exact same website and exposed to the same audience.

#### 4.6 Account hierarchy

Advertisers typically have a large amount of ad placements. To manage all placements, they need some type of structure to get an overview of how their marketing activities are performing. To achieve this, placements are sorted into a hierarchical structure, see figure 5.

In this structure each ad belongs to an ad group, a campaign, an account which is managed on the ad network. Accounts, campaigns and ad groups are all sets of ad placements. Further, each ad group consists of one or more ad placements, each campaign consists of one or more ad groups, each account consists of one or more campaigns, and each network with an ad placement has one or more accounts.

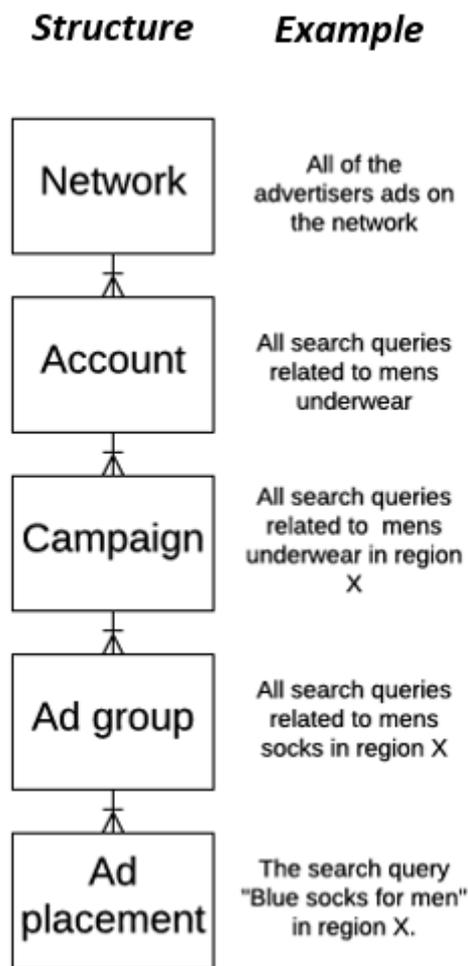


Figure 5, general account structure in networks. The boxes contain the different levels in the structure and the text on the right side of the boxes is an example of how the structure is used. All networks do not use all layers in the structure and the names of the sets differ between the different ad networks. The notation used for the boxes in the figure is the one used in google ad words and will for convenience be used for all other networks as well [22].

#### 4.7 Active interventions

Online marketing is a process with constant intervention from the marketers. Ad placements that perform poorly will get shut down and ad placements that perform well will get bigger budgets.

These interventions are intended to increase traffic for higher performing ads (ads that convert impressions to acquisitions) and decrease it for low performing ads. When introducing a new ad, this means that the historic data can have a strong representation of ads with high conversion rates. When introducing new ads, ads that have low conversion rates have not yet been filtered away (since it requires data to see which ads are good/bad), this could lead to the conversion rates of new ads being systematically over estimated.

## 5 Theory

In this chapter, the necessary statistical background for the thesis is covered.

### 5.1 Logistic regression

Logistic regression can be thought of as a development of linear regression. In linear regression data points are fitted to a model by using linear relationships between the dependent variable (the variable which is the target of the prediction) and the independent variables (the variables used for predicting). The relationship is described through

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon \quad (1)$$

where  $y$  is the dependent variable,  $x_i$  are the independent variables,  $\beta_i$  are the coefficients for the corresponding dependent variables and  $\varepsilon$  is the error of the model. The parameter  $\beta_0$  is called the intercept as it is the value of the dependent variable when all independent variables are zero.

To fit the model to the data in linear regression the most common procedure is OLS (ordinary least squares), which minimizes

$$\sum (y - (\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p))^2$$

OLS has a closed form solution and has been shown to be equivalent to the maximum likelihood estimation of the coefficients, as well as being BLUE (Best Linear Unbiased Estimator) for the coefficients provided that the errors are gaussian.

While linear regression is a good and simple model it's not very applicable for modeling conversion rates as it makes predictions on the interval  $[-\infty, \infty]$  while conversion rates are on the interval  $[0,1]$ . To solve this problem a conversion rate  $\mu$  can be transformed through the logit function

$$\text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

and the log odds of  $\mu$  becomes continuous on the  $[-\infty, \infty]$  interval. The logit of  $\mu$  can be modeled through linear regression

$$\log\frac{\mu(x)}{1-\mu(x)} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (2)$$

and the conversion rate  $\mu$  can be modeled through using the inverse of the logit function

$$\mu(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}}$$

resulting in a probability of a conversion on the  $[0,1]$  interval.

For equation 2 OLS could be applied to estimate the coefficients, but the problem of interest is not to minimize the residuals of the log odds, but rather to find the most probable model for the originally measured probabilities. To find the coefficients that gives the highest probability of the observed data, the maximum likelihood estimation can be used.

To make a maximum likelihood estimate it is necessary to know the probability distribution. Since conversions in the marketing funnel is analogous to a Bernoulli process the maximum likelihood estimation can be applied to estimate the parameters of the independent variables.

The likelihood function that is to be maximized is:

$$l(\beta) = \prod_{i=1}^n \mu(x_i)^{y_i} (1 - \mu(x_i))^{1-y_i}$$

where  $y_i$  is the outcome of the Bernoulli process for event  $i$ , with  $y_i=1$  for a success and  $y_i=0$  for failure.  $y_i$  and  $y_{i+1}$  are independent of each other which makes the total likelihood of the outcome the product of the likelihood of all individual events.

As the likelihood function is convex it has the same maxima as the log likelihood function and the maximum likelihood estimates are the same as the maximum log likelihood estimates.

$$\max(L(\beta)) = \max(\ln[l(\beta)]) = \sum_{i=1}^n \{y_i \ln[\mu(x_i)] + (1 - y_i) \ln[1 - \mu(x_i)]\}$$

Differentiating the expression with respect to  $\beta$  gives the likelihood equations

$$\sum_{i=1}^n [y_i - \mu(x_i)] = 0$$

$$\sum_{i=1}^n x_{ij} [y_i - \mu(x_i)] = 0$$

These equations don't have any closed form solution for logistic regression (as they do in linear regression). To solve this problem the iteratively reweighted least squares (IRLS) approach can be taken. IRLS has been shown to be equivalent to the Newton-Raphson method for solving the maximization [10].

### 5.1.1 Iteratively Reweighted Least Squares

The idea behind IRLS is to update the prediction each time by giving the poorly predicted values more weight in the next iteration. The iteration continues until no more improvements in terms of likelihood can be made by reweighting the observations.

For calculations it's convenient to write the dependent variable  $y$  from the  $n$  observations on vector form

$$y^T = [y_1, y_2, \dots, y_n]$$

and the independent variables  $X$

$$X = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ 1 & x_{2,1} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{bmatrix}$$

and the coefficients  $\beta$  at iteration k

$$\beta^k = [\beta_0^k, \beta_1^k, \dots, \beta_p^k]^T$$

on matrix form.

The fitted values  $\mu$  given the independent variables at iteration k is written as

$$\mu^k = [\mu_1^k, \mu_2^k, \dots, \mu_n^k]^T$$

where

$$\mu_i^k = \frac{1}{1 + e^{-x_i \beta^k}}$$

is the fitted value for observation i at iteration k with  $\mu_i = \mu(x_i)$ .

The expected variance for each prediction is contained in the diagonal matrix S

$$S^k = \begin{bmatrix} \mu_1^k * (1 - \mu_1^k) & 0 & \dots & 0 \\ 0 & \mu_2^k * (1 - \mu_2^k) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mu_n^k * (1 - \mu_n^k) \end{bmatrix}$$

which is used to reweight the influence of the observations for the next iteration.

The updating procedure for IRLS is

$$\beta^{k+1} = (X^T S^k X)^{-1} X^T (S^k X \beta^k + y - \mu^k)$$

which can also be written as

$$\beta^{k+1} = \beta^k + (X^T S^k X)^{-1} X^T (y - \mu^k) \quad (3)$$

From equation 3 it is possible to see that the coefficients are updated through multiplying the residuals with the independent variables, and dividing by the expected variance of the prediction multiplied with the corresponding independent variables squared.

### 5.1.2 Interpretation of variables

In linear regression it is straight forward to interpret the coefficients of the independent variables. From equation 1 it is clear that a unit increase in  $x_i$  will generate a  $\beta_i$  increase or decrement in the expected  $y$  depending on the sign of  $\beta_i$ , assuming that all other covariates are held constant.

For logistic regression interpretation of the variables is more complicated. From equation 2 it's clear that a unit increase in  $x_i$  would generate a  $\beta_i$  increase or decrease in the expected log odds of  $\mu$  depending on the sign of  $\beta_i$ , with all other covariates held constant. To get an understanding of how the log odds of the probability relates to the probability figure 6 can be studied.

**Relationship between log odds and probability**

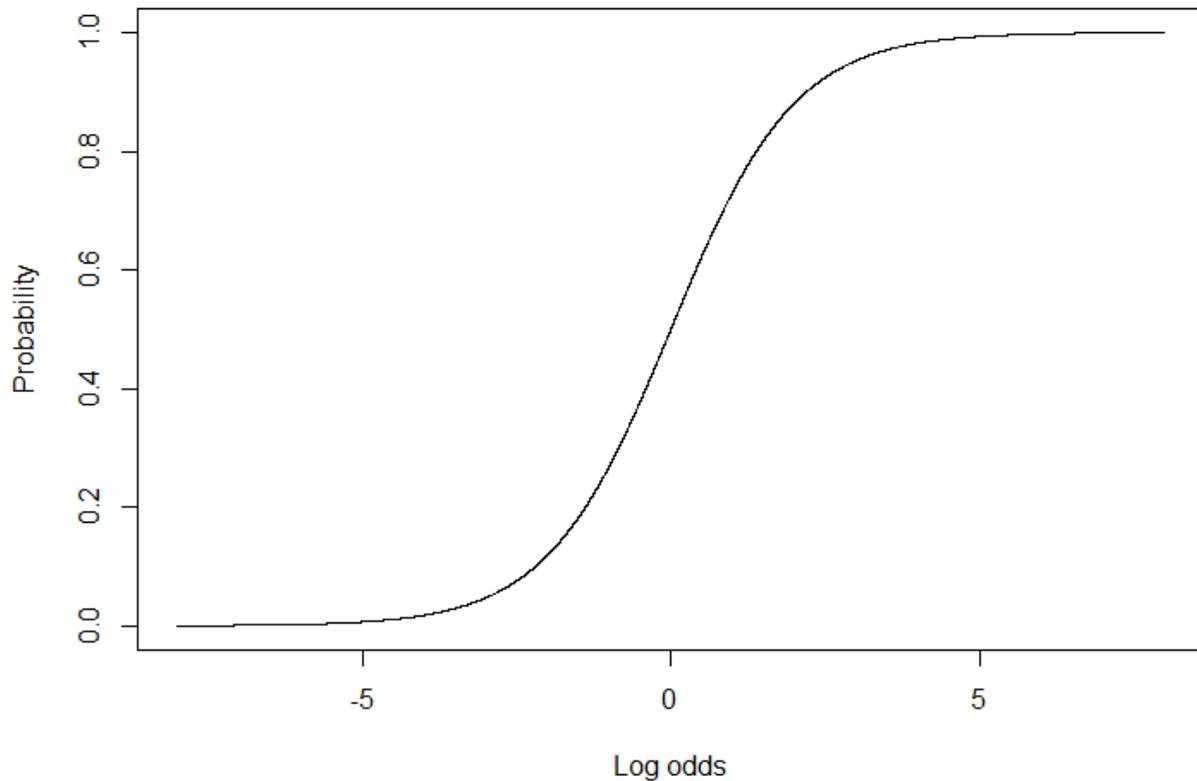


Figure 6, relationship between probability and log odds.

As can be seen in figure 6 an increase in log odds corresponds to an increase in probability. This means that a unit increase in the variable  $x_i$  would generate an increase of the estimated probability  $\mu$  if  $\beta_i$  is positive and a decrease if  $\beta_i$  is negative, assuming all other covariates are held constant. How big the increase/decrease would be depending on what the pre-change probability is.

### 5.1.3 Feature selection with Lasso

Overfitting, that is using variables that uses structures in the modeling set that are not signals but just noise, can easily occur if no measure is taken against it. If there is no cost associated with including more variables to a model, the performance of the model on the modeling set can only increase from including a new variable. This can lead to the inclusion of variables to the model that have no actual relation to the outcome. Such variables are likely to worsen the performance of the model when it is applied to a new set of data.

One measure against overfitting is introducing a penalty function to the maximum likelihood equation. The penalty function puts a cost to adding more variables, thereby preventing the model from including variables that don't have a big enough improvement on performance. There are multiple penalty functions each with its own properties. One popular function is the LASSO (Least Absolute Shrinkage and Selection Operator) as introduced by R. Tibshirani [11]. Lasso penalizes the coefficients by multiplying the  $L_1$  norm of the coefficients (except for the intercept) with a cost  $\lambda$

$$\max \sum_{i=1}^N \left\{ y_i(\beta_0 + x_i^T \beta) - \ln \left( 1 + \frac{1}{e^{-(\beta_0 + x_i^T \beta)}} \right) - \lambda \|\beta\|_1 \right\} \quad (4)$$

to shrink parameters and prevent overfitting from noise.

By penalizing the absolute value of the coefficients, it shrinks coefficients towards zero. For coefficients with limited, if any, impact this results in the coefficient becoming zero. Putting unimportant coefficients to zero doesn't only improve out of sample performance, but also makes the results more interpretable as it is possible to see directly from the coefficients if the independent variable has any influence or not. It's important to note that the intercept is not penalized (see equation 4) unless there is reason to believe that the mean should be zero.

To decide on what value of  $\lambda$  k-fold cross validation and the one step prediction error approach as described below can be used.

## 5.2 Cross validation

Another way of preventing overfitting is cross validation. A model that is trained on a dataset might find structures in that data set that aren't results of the process but rather just a coincidence. When the model is applied to new data the predictions can be made worse by accounting for such coincidental structures, compared to if they were not accounted for at all.

Cross validation is a procedure meant to prevent this by dividing the data into a modeling set and a validation set. The model parameters are fitted on the modeling set, then the performance is assessed on the validation set. By checking the performance on a different set than the models parameters were fit to, only coefficients that rely on structures that are present in both data sets will boost performance on the validation set. Variables that do not boost the performance on the other set can then be removed from the model.

Often when a model is created it's also of interest to know how well a model will perform on completely new data. This requires another set of data which has not been a part of the model building. These sets are often called training and test set where the training set can be divided into a modeling set and a validation set (see figure 7). The modeling set is the data that the parameters are estimated with, the validation set is where the model created on the modeling set is cross validated, and the test set is where the final model is tried to determine its performance.

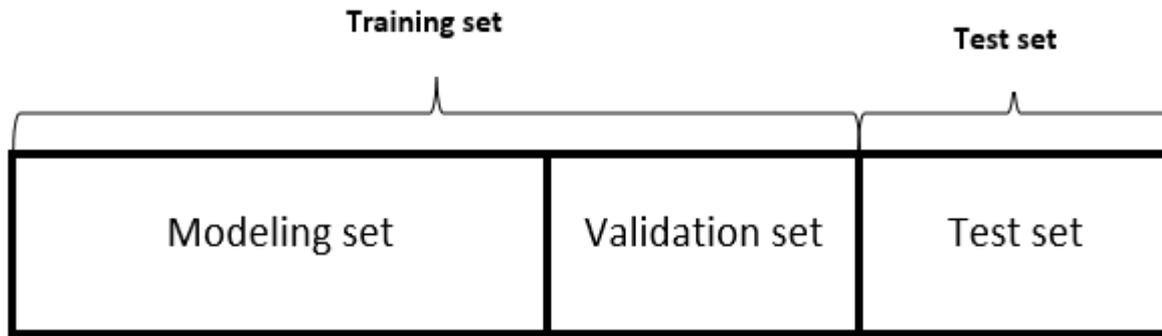


Figure 7, illustration of data divided into different sets used at different stages in the model building and evaluation.

When deciding on the proportions of testing and modeling sets it's important to consider what the purpose of the study is. A bigger test set will give a more reliable estimate of what performance the model will deliver. A bigger modeling set will give more reliable parameter estimates.

### 5.3 K-fold cross validation

Dividing up the data into a modeling and a validation set certainly helps in the problem of overfitting, but the test set uses a considerable amount of data that could be used to build the model into account. The estimate of the prediction performance is not very robust as it is only one estimate of the performance. A way of getting a more robust estimate of the predicting power of the model while having a bigger modeling set is using K-fold cross validation.

The idea behind K-fold cross validation is dividing the data into k different equally sized parts called folds. Out of these folds k different pairs of training and test sets are created, with one of the folds being the validation set and the other folds being the modeling set. Modeling is then carried out on each one of these pairs and performance can be calculated for each of the k models.

As Elements of statistical learning points out it is important to consider each of the k pairs to be isolated from the others, and not to use information from the other pairs in model design. If used for modeling design, it would mean that the validation set is no longer independent from the modeling set [12].

There is no general rule on how to decide the value of k but k=10 is commonly used [13]. There is also the special case where k is equal to the number of samples available which is called leave-one-out cross validation.

#### 5.3.1 One standard error rule

While cross validation decreases the risk of overfitting it does not eliminate it. Especially when considering a large number of explanatory variables, it is likely that some variable will appear to have explanatory power on the validation set while it is completely random. This phenomenon was named selection bias by McLachlan et al [14].

As a measure against selection bias, the one standard error rule can be used [15]. The idea is to find the number of parameters that gives the best average prediction through k fold cross validation (or another resampling technique). The parameters used in the

final model is then the parameters in the model that has an average expected prediction error no more than one standard deviation higher than the expected prediction error of the best performing model.

Figure 8 shows an example of the one standard error rule in action. The model that is optimal on the cross-validation set has 162 degrees of freedom and the model that is within one standard error away from it has 23. This shows that model size can be reduced by the one standard error rule.

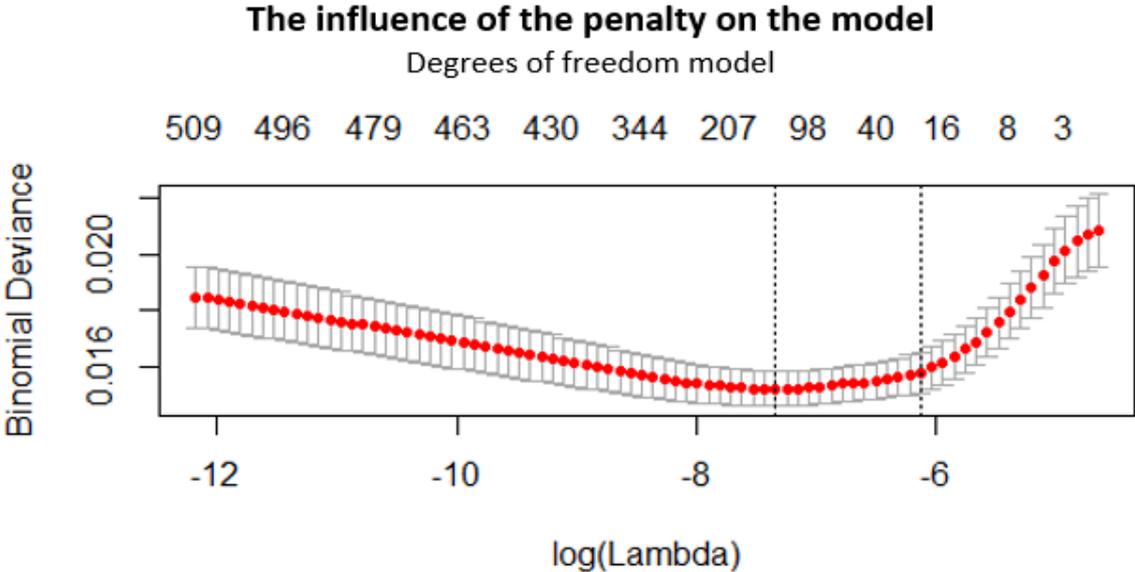


Figure 8, plot showing the binomial deviance of a logistic regression model with LASSO depending on the choice of lambda with the optimal lambda being the first line from the left the greatest lambda value that gives a performance less than one standard deviation higher than the optimal performance and the models degrees of freedom the top of the plot.

#### 5.4 Model evaluation

When evaluating different models, it is important to have a performance measure that allows for comparison between different models and is somewhat intuitive. In ordinary least squares regression R squared is normally used as an evaluation metric. It is a measure of how much of the total variance has been explained by the model and is calculated through the formula

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

Where  $y_i$  is the outcome,  $\hat{y}_i$  the predicted outcome,  $\bar{y}$  is the average of all outcomes  $y_i$  and N is the total number of observations.

In OLS regression R-squared has several nice properties such as:

1. Explaining variability
2. Improvement compared to null model
3. Being the square of the correlation

However, R squared is not as good of a measure for logistic regression since it doesn't aim to minimize variance but rather maximize the likelihood. As the properties of R-squared are very helpful several pseudo R-squares has been created for logistic regression. One of the more widely used is McFadden's R-squared [16] calculated through:

$$R^2_{McFadden} = 1 - \frac{\ln \hat{L}(M_{Full})}{\ln \hat{L}(M_{Intercept})}$$

McFadden's R-squared can be interpreted as a measure of how much better the model is compared to just using the average as a predictor. It ranges from 0 to 1 and is applicable as a performance measure for any model that maximizes likelihood.

### 5.5 Chi-squared test for independence of variables

To see if two variables are independent a Chi-squared test can be conducted. With  $n$  observations of an event that has  $s \cdot r$  different possible outcomes  $B_i A_j, i = 1, \dots, s, j = 1, \dots, r$  with the probability  $p_{i,j}$  ( $\sum_{i=1}^s \sum_{j=1}^r p_{i,j} = 1$ ).  $x_{i,j}$  is the frequency of the outcome ( $\sum_{i=1}^s \sum_{j=1}^r x_{i,j} = n$ ).

The probabilities  $p_i$  and  $p_j$  can be estimated as

$$p_{.j}^* = \frac{\sum_{i=1}^s x_{i,j}}{n} \quad p_{i.}^* = \frac{\sum_{j=1}^r x_{i,j}}{n}$$

If the different outcomes are independent, then

$$p_{i,j} = p_i * p_j$$

To control this the test variable

$$Q_{obs} = \sum_{i=1}^s \sum_{j=1}^r \frac{(x_{i,j} - n * p_{.j}^* * p_{i.}^*)^2}{n * p_{.j}^* * p_{i.}^*}$$

is created. Independence can be rejected at a  $\alpha$  level of confidence if

$$Q_{obs} > \chi^2_{\alpha}((r - 1)(s - 1))$$

As a rule of thumb, it should be checked that

$$n * p_{.j}^* * p_{i.}^* > 5 \forall i, j$$

when using the test [17].

### 5.6 Pearson residuals

One way of understanding the residuals of a logistic regression is looking at the Pearson residual. It takes the difference between the outcome and the expected outcome and divides it by the expected standard deviation [18].

$$p_i = \frac{y_i - n_i \hat{\mu}_i}{\sqrt{n_i \hat{\mu}_i (1 - \hat{\mu}_i)}}$$

(5)

## 6 Methodology

This chapter describes the method that was used for the study.

### 6.1 Data description

First the data was collected from the reporting system of Company X in a CSV file.

The data consists of 65.000 rows where each row represents a combination of an ad creative with an ad placement during a calendar week for every week of 2016. This means that all impressions with the same attributes has been aggregated into one row. Example data is provided in appendix 1.

The data can be divided into four different categories. *Contextual data* which describes where on the internet the ad is placed, *targeting data* which is information about the specific user attributes, *ad specific data* which describes the ad and *feedback data* which is how the ad has performed.

In table 1 the different data columns are presented.

Variable name	Description	Data category	Display	Facebook	GDN	SEM
Network	The advertising network that the ad placement was active on.	Contextual data				
Account	The account that the ad placement is placed under.					
Account type	The type of campaigns that are placed in the account.					
Campaign	The campaign that the ad placement was placed under.					
Campaign type	The type of ad placements that are placed in the campaign.					
Ad group	A group of ad placements. All placements in an ad-group uses the same ad creative.					
Device	Describes if the ad placement is shown on a desktop, mobile or tablet.	Targeting				
Country	Which country the user viewing the ad is located in.					
Gender	The users gender (man, woman or unknown)					
Targeting	What type of group affinity the user belongs to (Similar to current customers, interests etc.)					
Creative id	A unique identity number for the ad creative.	Ad specific data				

Ad type	What type of message the ad creative contains.					
Dimensions	The dimensions of the ad-creative.					
Brand	Which one of Company X brands the ad placement promoted.					
Image/Text ad	If the ad creative consists of an Image, Text or a combination.					
Impressions	The number of impressions that the ad placement has gotten in the time period.	Feedback data				
Clicks	The number of clicks that the ad placement has gotten in the time period.					
Leads	The number of leads that the ad placement has gotten in the time period.					
Acquisitions 30 days	The number of users that has become customers within 30 days after the impression.					
Average position	SEM specific, the average position among the paid search advertisements shown for the searches that it is present in (See figure 3).					
Average Quality Score	SEM specific, the quality score that the ad placement is assigned by the search engine. It depends on expected click through rate, ad relevance, and landing page experience [19].					

Table 1, attributes of the observations in the data, the columns to the right indicates if the data is available for the different channels where a grey box indicates that it is not used. It should be noted that while SEM has campaigns and ad groups these were not used as the size of the data set would then be too large to manage.

The data is collected from four sources (Facebook Ads, Google Display Network, Google Ad Words and a consolidation of other display networks). There is a difference in which data is available between different networks. Gender and targeting data is only available for Facebook and GDN while average position is only applicable for SEM where there are multiple slots where the ad placement can be shown.

## 6.2 Data pre-processing

The collected data had a large amount of NA values amongst its numerical values. In this case an NA for the numerical values should be interpreted as 0 since it simply means that there are no users at that point in the funnel, so all NA values were changed to zero.

In the data there was also instances where the later stages in the funnels had a larger number than earlier stages. This is an obvious error since the conversion rate cannot be

larger than 100%. This error seems to come from some type of bug in the reporting system. These data point could either be adjusted down so that they have the same value as the previous conversion step or removed from the data. Since it's unclear from the reporting system how later conversion rates are affected from this bug it was considered safer to remove the observations from the data set completely, especially since it didn't have a large impact on the size of the data set.

6.3 Dividing data into modeling and test set

The data is randomly divided into modeling or test set with 20% of the rows going into the test set and 80% going into the modeling set.

6.4 Models

Four different models are created for each online advertising network. The different models are:

1. Group average – Predicts by finding the most granular group that has enough data to make a prediction for the conversion rate, then predicts the future conversion rate as the weighted average conversion rate of the group.
2. Logistic regression on contextual data – A logistic regression model that uses information about the placement (such as webpage, device etc.) and ad specific information.
3. Logistic regression with early conversion rates – This model uses the same information as previous model together with early conversion rates (click through rate and click to lead).
4. Logistic regression with latent variables – This model uses the same information as previous model and adds two covariates by using the residuals from predictions of earlier conversion rates as input for the model.

By design, the different models make use of different information. An overview off which information that is available for which model is given in table 2.

Model	Contextual data	Early conversion rates	Latent variables
Group average	Yes		
Logistic regression on contextual data	Yes		
Logistic regression with early conversion rates	Yes	Yes	
Logistic regression with latent variables	Yes	Yes	Yes

Table 2, schematic of data available to the different models.

6.4.1 Group Average

The group average approach is how Company X currently predicts conversion rates. The idea is to start at the most granular level in a predetermined hierarchy, and check if there is enough data to make a prediction of the conversion rate. This is done through

determining if the number of conversions are higher than a certain cut off criteria, the cut off point. If not, it goes up one level in the hierarchy and checks if there is enough data on that level, and so on, until it is able to make a prediction.

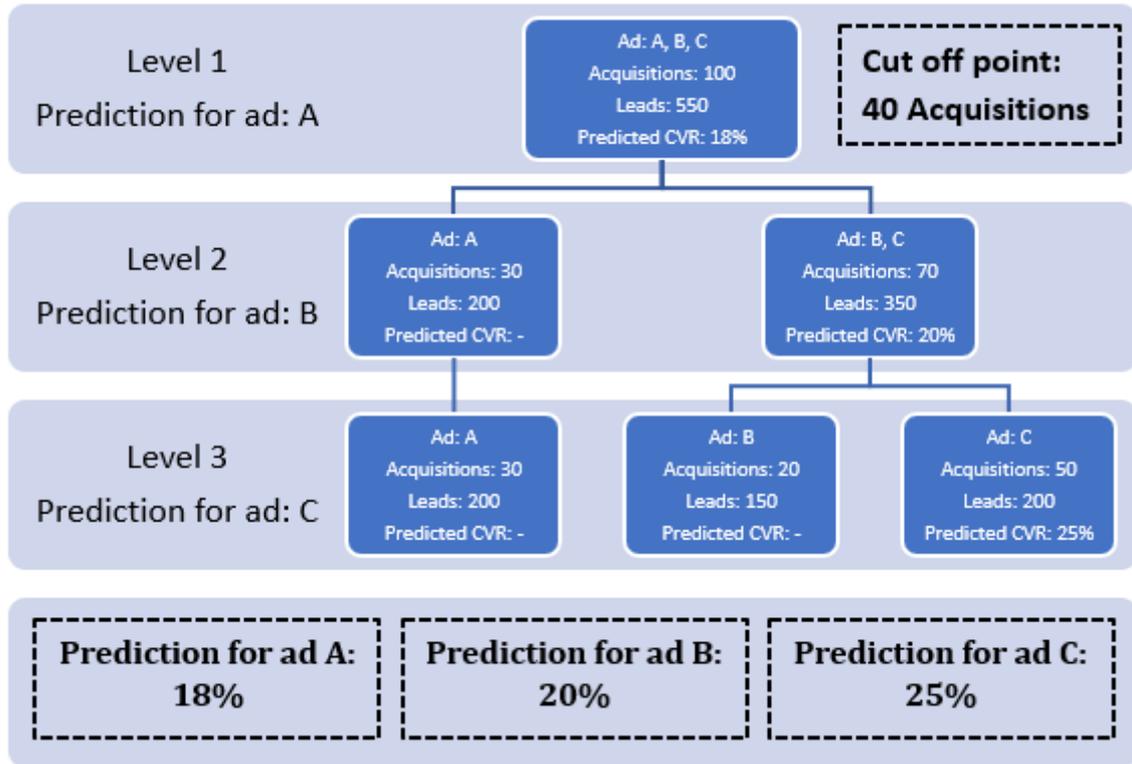


Figure 9, example of hierarchy structure for making predictions with the group average model.

To check which level in the hierarchy the conversion rate should be predicted at for ad  $i$  the variable  $\theta_{lvl,i}$  is introduced.  $\theta_{lvl,i}$  is calculated through the formula

$$\theta_{lvl,i} = \begin{cases} 1 & \text{if } \sum_{j \in \Omega_{lvl,i}} A_j \geq \lambda \cup \sum_{k=1}^{lvl-1} \theta_{k,i} = 0 \\ 0 & \text{if } \sum_{j \in \Omega_{lvl,i}} A_j < \lambda \cap \sum_{k=1}^{lvl-1} \theta_{k,i} = 1 \end{cases} \quad (5)$$

Where  $A_j$  is the number of acquisitions for ad placement  $j$ ,  $\lambda$  is the cut off point.  $lvl$  is the level in the hierarchy and  $\Omega_{lvl,i}$  is a set containing the index number of all the ad placements that are in the same prediction group as ad placement  $i$  on level  $lvl$ .  $\Omega_{lvl-1,i}$  is a subset of  $\Omega_{lvl,i}$ .

It's important to note that

$$\sum_{lvl=1}^m \theta_{lvl,i} = 1 \forall i$$

where  $m$  is the number of levels in the hierarchy, meaning that predictions will only use one of the levels in the hierarchical structure. The predicted conversion rate  $\hat{\gamma}_i$  is then calculated through

$$\hat{\gamma}_i = \frac{\sum_{lvl=1}^n \sum_{j \in \Omega_{lvl,i}} \theta_{lvl,i} * A_j}{\sum_{lvl=1}^n \sum_{j \in \Omega_{lvl,i}} \theta_{lvl,i} * L_j} \quad (6)$$

where  $L_j$  is the number of leads for ad placement  $j$ .

Since

$$\theta_{lvl,i} = 0$$

for all levels except the most granular level that has enough data, only the ad placements that belong to the set  $\Omega_{lvl,i}$  for that level will influence the prediction.

#### *Search procedure for the group average model*

Company X have a hierarchy that they currently use with a certain cut off point. However, in a more general situation that might not be available so instead of using the already existing hierarchy a search procedure is created to find a good hierarchy and cut off point.

Before predicting with the group average model there are two different things that needs to be decided, the hierarchy levels and the cut off point. The hierarchy levels are which covariate that is used to split the data at the different levels.

The cut off point is used to decide when there is sufficient data in a group to make a prediction on that level. Since the choice of cut off point is influenced by the choice of hierarchy levels and vice versa, a search to find the optimal value is conducted.

Finding the combination of cut off point and hierarchical model with the highest likelihood could be solved by using a grid search and calculating the likelihood of all possible combinations of cut off points and hierarchical model. This would however be very computationally heavy.

To use less computational power than a grid search, a forward selection algorithm is used. The algorithm starts by only using the global average for the prediction and then adds the covariate that will increase the performance of the model the most at the next level of the hierarchy. The algorithm continues to add layers to the hierarchy until there are no covariate than will improve the performance of the model by being added to the hierarchy left.

The goal of the search is to find the model that maximizes the likelihood equation. Maximizing the likelihood function is analogous to maximizing the log likelihood. The log-likelihood function is defined as

$$L(y, F(x)) = \ln[l(y, F(x))] = \sum_{i=1}^n \{y_i \ln[F(x_i)] + (1 - y_i) \ln[1 - F(x_i)]\}$$

Where  $F(x)$  is equation 6

The algorithm for deciding which attribute should be at which level is as follows. The input to the algorithm is a list of cut off points and all categorical covariates.

For cut off in  $cut.off.list$

1. All possible grouping attributes belonging to the set  $\pi$ .
2.  $lvl=0$ 
  - a. Divide data into 10 folds
  - b. Do:
    - i.  $lvl=lvl + 1$
    - ii. For each  $cov$  in  $\pi$ 
      1.  $\Omega_{lvl} = cov$
      2. For  $k= 1:10$ 
        - a. Hold out one 10<sup>th</sup> as a test set and the rest as a training set
        - b. Make predictions on the test set based on the training set data through:
 
$$LogLikelihood_{co,lvl,cov,k} = \sum_i L \{y_i, F_{lvl}(x_i, cut\ off)\}$$
    3.  $LogLikelihood_{co,lvl,cov} = mean(LogLikelihood_{co,lvl,cov,k})$
    - iii.  $cov_{best} = \max_{cov}(LogLikelihood_{co,lvl,cov})$
    - iv. Remove  $cov_{best}$  from  $\pi$
    - v.  $\Omega_{lvl} = cov_{best}$
    - c. While  $\pi \neq \emptyset \wedge \max(LogLikelihood_{co,lvl,cov}) > \max(LogLikelihood_{co,lvl-1,cov})$
3.  $model_{co} = F_{lvl-1}(x, cut\ off)$
4.  $model\_performance_{co} = \max(LogLikelihood_{co,lvl-1,cov})$

$$cut\ off = \max_{co}(model\_performance_{co})$$

Final model =  $model_{co}$  ,  $co = cut\ off$

#### 6.4.2 Logistic regression on contextual data

The idea behind this model is to see if a logistic regression model can outperform the group average model if it is built from the same available data. The main difference between the models is that a logistic regression model can use information across different groups (for example the difference between desktop and mobile users) while the group average is limited to using the data that is in its category to make predictions. This difference is illustrated in figure 10.

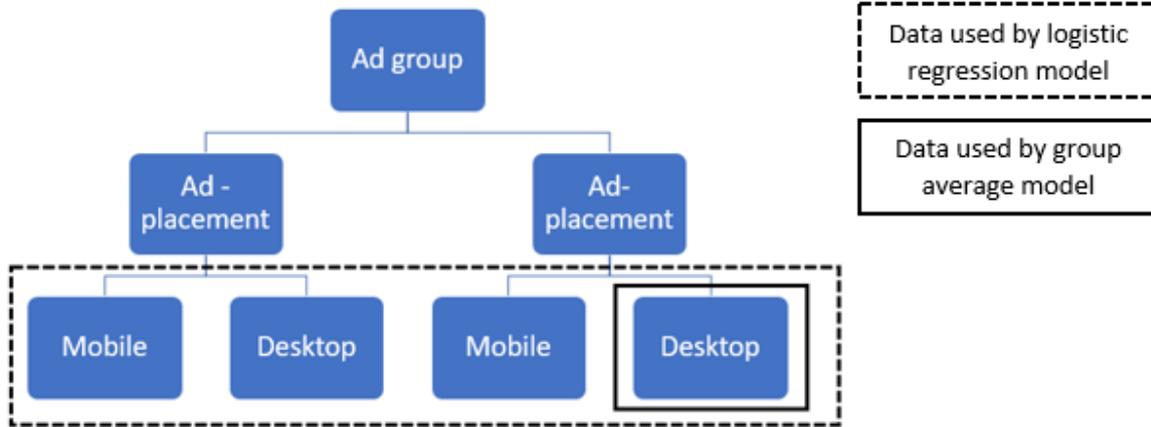


Figure 10, When predicting the conversion rate of an ad-placement displayed on a desktop the group average model only uses the data available on that specific ad placement displayed on desktops. The logistic regression model can however use all the information from the ad placement and adjust for the difference between mobile and desktop devices that exists in other ad placements as well to make a prediction.

Predictions are calculated through

$$\hat{\mu}_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i})}} \quad (7)$$

Where  $\hat{\mu}_i$  is the predicted conversion rate and  $\beta$  is estimated through iteratively reweighted least squares, as described in chapter 5.1.1. The covariates are all contextual, targeting and ad specific variables listed in table 1.

#### 6.4.3 Logistic regression with early conversion rates

The logistic regression model with early conversion rates also contains the information from the earlier conversion steps. Since it's built from more data than previous models it should be more accurate provided that the early conversion rates are correlated to the conversion rates from lead to acquisition.

The conversion rate is estimated through equation 7 where  $\beta$  is estimated through iteratively reweighted least squares. The covariates are all contextual, targeting and ad specific variables listed in table 1. Together with the conversion rates for the two earlier conversion steps, estimated as the average conversion rate for the ad placement with its specific ad creative, i.e.

$$\overline{CTR}_i = \frac{C_i}{I_i}$$

and

$$\overline{CVR}_{Click\ to\ lead,i} = \frac{L_i}{C_i}$$

where  $C_i$  is the number of clicks,  $I_i$  the number of impressions and  $L_i$  the number of leads for ad placement i.

An important consideration when applying this model is that for ad placements with small amounts of data there is expected to be a larger estimation error for the conversion steps since there are fewer observations to base the conversion rate on. Because of this it is interesting to see if predictions are better when the conversion rates are included for low volume ads compared to when excluded from the model.

#### 6.4.4 Logistic regression with latent variables

The logic behind adding latent variables is that the residuals from predictions of earlier conversion rates could hold information about the quality of the ad. If the CTR is a lot higher than what would be expected for an ad placement, that could be interesting information. A possible explanation for an unexpectedly high CTR is that the ad is a “click bait”, an ad that is created to create as many clicks as possible. Then a bigger residual would correlate to lower conversion rates for the later conversions, since the users that clicked the ad were not really interested in the product, but clicked for other reasons.

To get the residuals of earlier conversion steps logistic regression models are created for these conversion rates as well. The models are created following the same procedure as the logistic regression with early conversion rates.

The residuals are then calculated as the difference between the estimation and the measured conversion rate which is the same as the average.

$$r_{CTR,i} = \overline{CTR}_i - \widehat{CTR}_i$$

And

$$r_{CVR (click to lead),i} = \overline{CVR}_{click to lead,i} - \widehat{CVR}_{click to lead,i}$$

These residuals are then used as input for the logistic regression with latent variables model.

### 6.5 Modeling procedure

The modeling procedure is described in detail for the display platform. The same procedure was followed for the other platforms and the results for those platforms are presented in the next chapter.

#### 6.5.1 Pre-processing

The same pre-processing steps are used for all models. The group average model and the logistic regression on contextual data model does however not use the continuous variables.

##### *Transformations of continuous covariates*

For continuous variables three different transforms were created. The idea behind this is that the model can choose the transform that best fits the data. The different variables are then standardized by subtracting the mean and dividing by the standard deviation.

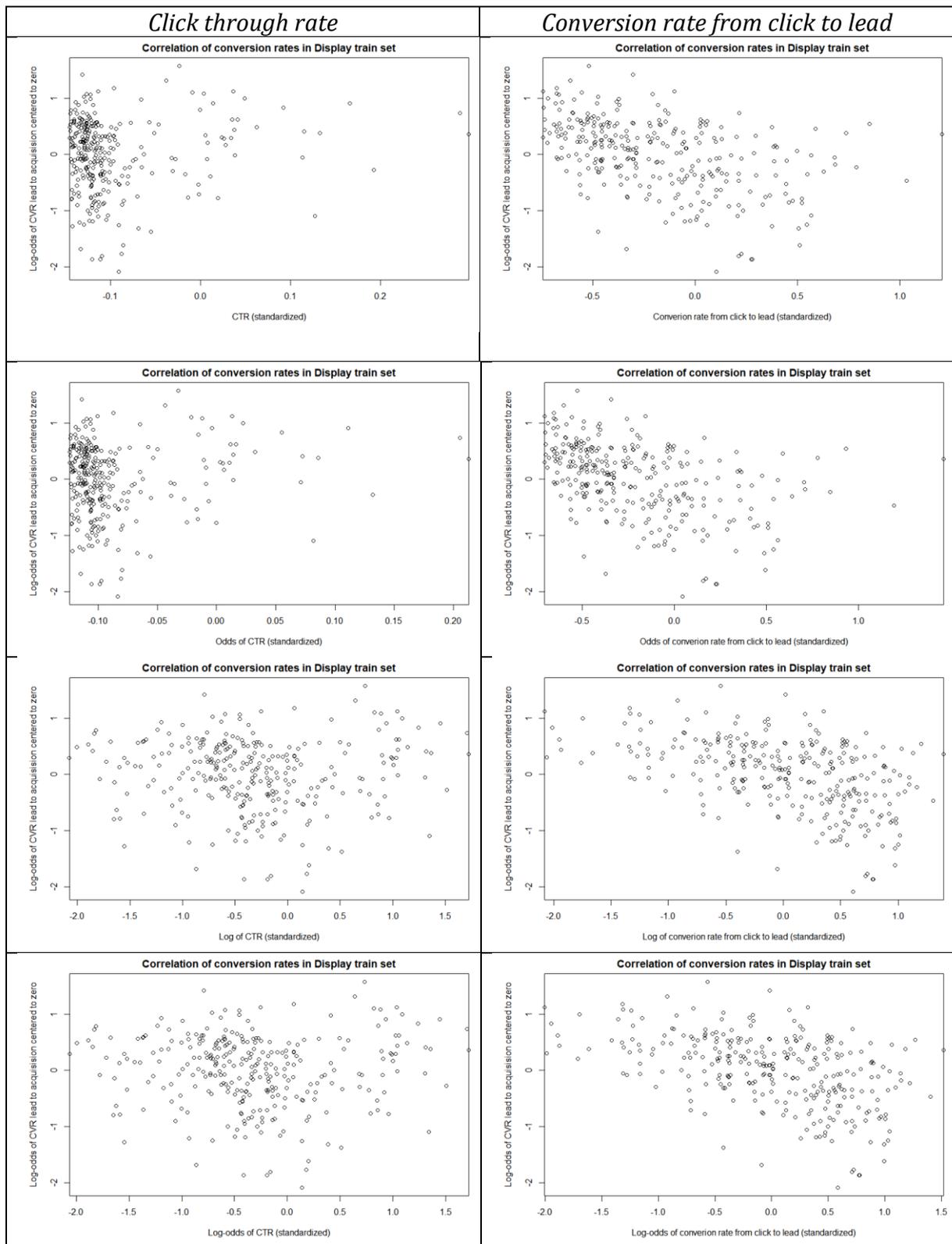


Figure 11, Scatter plots showing different transformations of the continuous variables.

### Re-grouping categorical covariates

Some covariates contain very small groups. If a group contains a zero cell in the contingency table that risks the stability of the model [20]. Such categories can either be removed or combined into a bigger group if there is a sensible combination. There is

also a case for combining small groups even if there isn't a zero cell in the contingency table.

Since a model coefficient for a small group gets punished just as hard as a model coefficient from a larger group by the LASSO, the influence of a variable only affecting a small population runs a high risk of being excluded from the model, even if it is important. By combining such a variable with other variables that has a similar influence, it is more likely that it will get included into the model.

The dimensions variable had both zero cells in the contingency table and small groups. The marketing department at Company X was consulted, and it was determined that the most sensible grouping would be to combine the smaller dimension groups based on where on the webpage they would appear. See the grouping in figure 12.

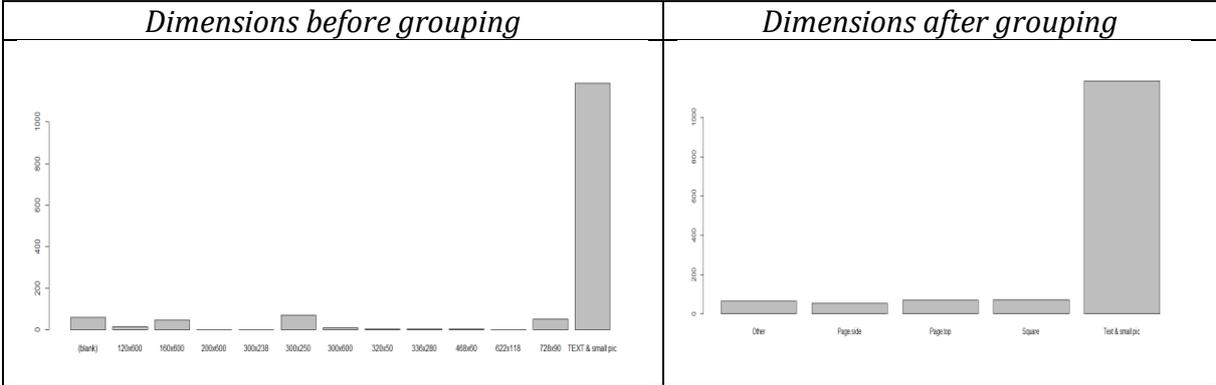


Figure 12, showing the different levels of dimension.

The Campaign variable also had several zero cells in the contingency table. All the zero cells were in the Acquisitions column, and hence had an observed conversion rate of 0. To put them in a group with a low and stable conversion rate they were added to the group that had lowest conversion rate of all the groups that had more than 1000 leads. This had a minor impact on the total size of that campaign.

There are multiple small ad networks provided by the same companies. Many of these have zero cells in their contingency tables. To adjust this the different networks are grouped based on the company that maintains them. This grouping is based on the knowledge of the online marketers at company X who claim that the same providers use a lot of the same algorithms for deciding on which ad to show where, and have similar models for bidding on ad placements. This reduces the number of different ad networks from 76 to 8 providers and a "Other" category which contains smaller providers. The 8 biggest companies account for 98% of all leads.

6.5.2 Fitting the group average model

The algorithm described in section 6.4.1 was used to find the optimal model for group average. First a broader search was started with wide ranging cut off points.

McFadden R-squared	Levels	Cut off point
0.007217548	-Network2-Brand	7.389056
0.006510932	-Network2	20.085537
0.007465628	-Network2-Brand-Device	54.598150
0.007266235	-Network2-Brand-Campaign.type	148.413159
0.007201739	-Network2-Campaign.type-Dimensions2	403.428793

Figure 13, output from group average model search

After the range in where the optima exist a new search was conducted.

McFadden R-squared	Levels	Cut off point
0.007198912	-Network2-Device-Brand	10.000000
0.007127243	-Network2-Dimensions2-Brand	20.000000
0.007296687	-Network2-Brand-Dimensions2	30.000000
0.007042419	-Network2-Brand	40.000000
0.007121337	-Network2-Brand	50.000000
0.006892275	-Network2-Device-Brand-Campaign.type	60.000000
0.006980155	-Network2-Device	70.000000
0.007108315	-Network2-Brand	80.000000
0.007048644	-Network2-Brand-Dimensions2-Device	90.000000
0.006659843	-Network2-Device-Brand	100.000000

Figure 14, output from group average model search with smaller range for cut off point values

Finding that the optimal combination of cut off point and categories was cut off = 30 and categories Network, Brand and dimensions. This model was then used to make predictions on the train and test set. Results were recorded and are presented in chapter 7.

### 6.5.3 Fitting the logistic regression models

For logistic regression the overall procedure was inspired by chapter 4 in Applied logistic regression [20], the process can be described in four steps:

1. First a univariate analysis is carried out. The idea of the univariate analysis is to filter away variables that has little explanatory power.
2. After the univariate analysis is conducted a first model is created from all variables that passed the univariate analysis. From this model the residuals are analyzed to see if there are patterns that can be used to create new variables. These new variables could both be from transformations of continuous variables as well as interactions between different variables.
3. A new model is created containing all the variables that were included into the previous model and the new variables created through transforms of continuous variables or interactions between variables. All variables that got included into this model is then used in the final model.
4. Features that has earlier been discarded are re-introduced to the model to see if any improvements can be made to the models performance. The model that has the highest performance on the validation sets is chosen as the final model.
5. The final model is fitted 10 times and the performance of the predictions on the training set and the test set were recorded.

This procedure was followed for all three logistic regression models.

#### *Univariate analysis*

After adjusting the variables, a chi-squared test was used for all categorical values. If the p-value is lower than 0.25 then the variable would continue to be used in the multivariate analysis. Otherwise it would be disregarded, as suggested by Hosmer and Lemeshow [20].

Variable	Degrees of freedom	P-value
Network	8	2.2e-16
Device	3	2.2e-16
Brand	1	5.69e-14
Campaign	49	2.2e-16
Dimensions	4	1.025e-07
Image/Text	1	2.555e-09

*Table 3, univariate analysis for categorical covariates*

For continuous variables a model was created with only the variable. A likelihood ratio test was conducted to determine if the univariate model was significantly better than the null model at a p=0.25 significance level.

Covariate	P-value
CTR	0.06862553
CTR Odds	0.2907403
CTR Log	0.02454817
CTR Log-odds	0.02584231
CVR to Lead	0
CVR to Lead Odds	0
CVR to Lead Log	0
CVR to Lead Log-odds	0

*Table 4, univariate analysis for continuous covariates*

As CTR Odds was not significant at a 0.25 significance level it was discarded while the other transformations of the early conversion rates were kept for the multivariate analysis.

#### *Logistic regression on contextual data - Multivariate analysis*

All the covariates deemed significant in the univariate analysis, except for early conversion rates, were included into a logistic regression model. After using 10-fold cross validation to find the optimal value for the penalty variable lambda, a model was fit using the optimal lambda and predictions were made on the training set.

To understand how big the influence of each variable that was included into the model was, the absolute value of the parameter multiplied with the covariate was summed up over the entire data set. This was then plotted in a bar chart, see figure 15.

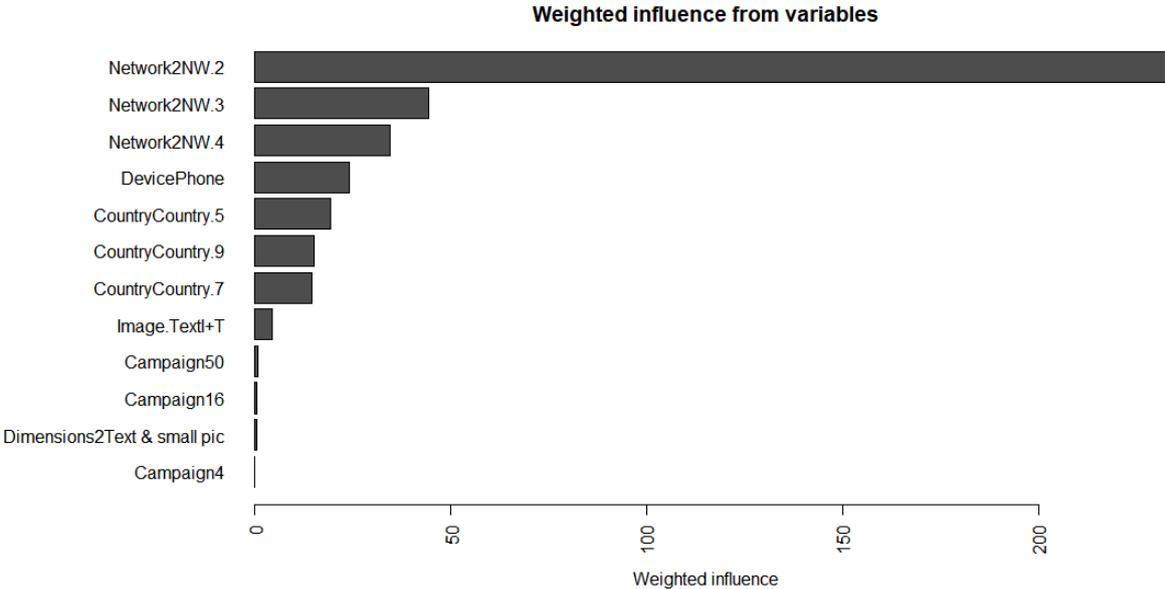


Figure 15, weighted impact of each variable level.

The residuals from these predictions were then analyzed.

Residual analysis

Combinations of different factors were plotted to see if there was interaction between the different variables. The Pearson residuals of the different observations on the training set was plotted in a box plot to see how the distribution looks. In figure 16 it seems possible that there is some interaction between the two different factors brand and network. A dummy variable is created for the interaction and added to the model.

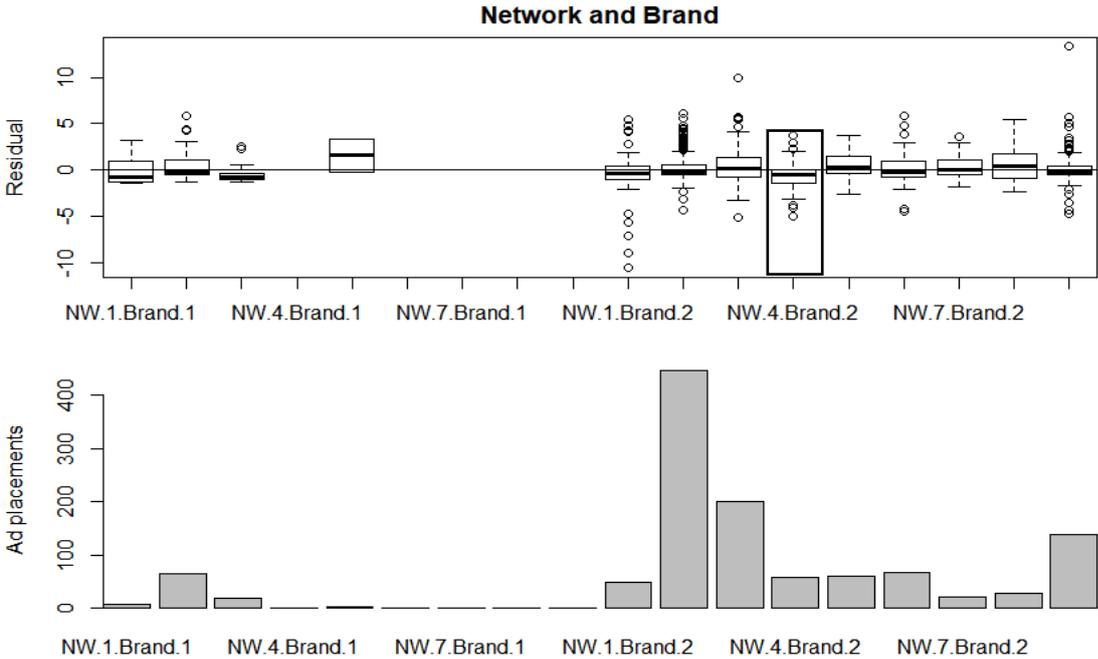


Figure 16, boxplot of residuals showing possible interactions between the covariates network and brand.

Possible interactions were also found between network and device and the covariates country and Image/text. These also had dummy variables created for them that was added into the model.

To analyze if there is a connection to the continuous variables the ad placements are grouped into groups containing a large amount of leads each (minimum 5000) to see if there is a trend in the predictions. This is done according to the following algorithm

1. Sort by continuous variable in ascending order.
2. Set  $n = 1$ ,  $group.prediction = 0$ ,  $group.acquisitions = 0$ ,  $group.leads = 0$
3. For each *ad placement*
  - a.  $group.prediction_n = group.prediction_n + prediction_{ad\ placement}$   
 $group.acquisitions_n = group.acquisitions_n + acquisitions_{ad\ placement}$   
 $group.leads_n = group.leads_n + leads_{ad\ placement}$
  - b. if ( $group.leads_n > 5000$ )
    - i.  $group.residual_n = (group.acquisitions_n - group.prediction_n) / group.leads_n$
    - ii.  $n = n + 1$

The different group residuals are then plotted to see if there is a correlation between the residual and the variable used for the grouping.

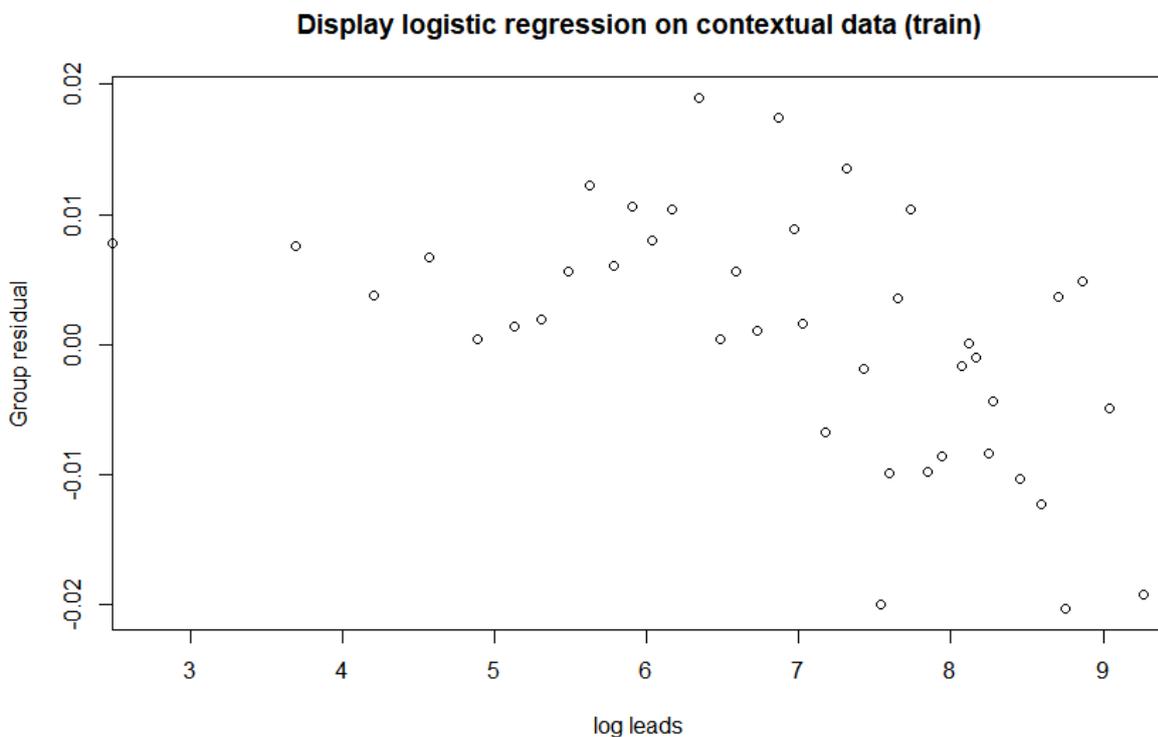


Figure 17, group residual plot showing a possible correlation between the residuals and volume of leads for an ad placement.

When analyzing figure 17 we can see that ad groups where the ad placements has had more leads tend to have a lower conversion rate than the model predicted. There seems to be a trend that the model has not been included into the model. Possible driving factors for this trend are:

- The errors from different ad-placements don't cancel each other out when fewer unique ad-placements are included into each group making the variation bigger.
- Ads "burn out" after a while. When they have been exposed to long in the same ad-placement users that have already seen the ad and learnt to ignore it.
- Larger traffic sources are less niched and appeal to a bigger audience where a smaller portion of the audience is interested in the ad content compared to more specialized smaller sources.

If there would be a measure of how big the total traffic is for a certain ad-placement a measure could be created for "burn out rate" by dividing the total impressions by the unique users. There is however no such measure available which makes it harder to create a sensible covariate simulating the effect. Simply using the number of leads is not a good option, since if the model would be implemented there is a high risk that an automatic algorithm would simply stop the ads from being shown once they have reached a certain number of leads.

If it is a "burn out effect" that is shown in figure 17, it doesn't make sense to have a global stop criterion since it should be highly related to the specific ad placements number of unique users. If it is a matter of niched sites vs larger, less niched traffic sources, an algorithm capping the number of leads would not separate between these two but simply put a lower amount of traffic to the larger sources.

Since there is no obvious/desirable variable that can model the trend we simply note that it exists at this stage and continue to observe it as the more advanced models are created.

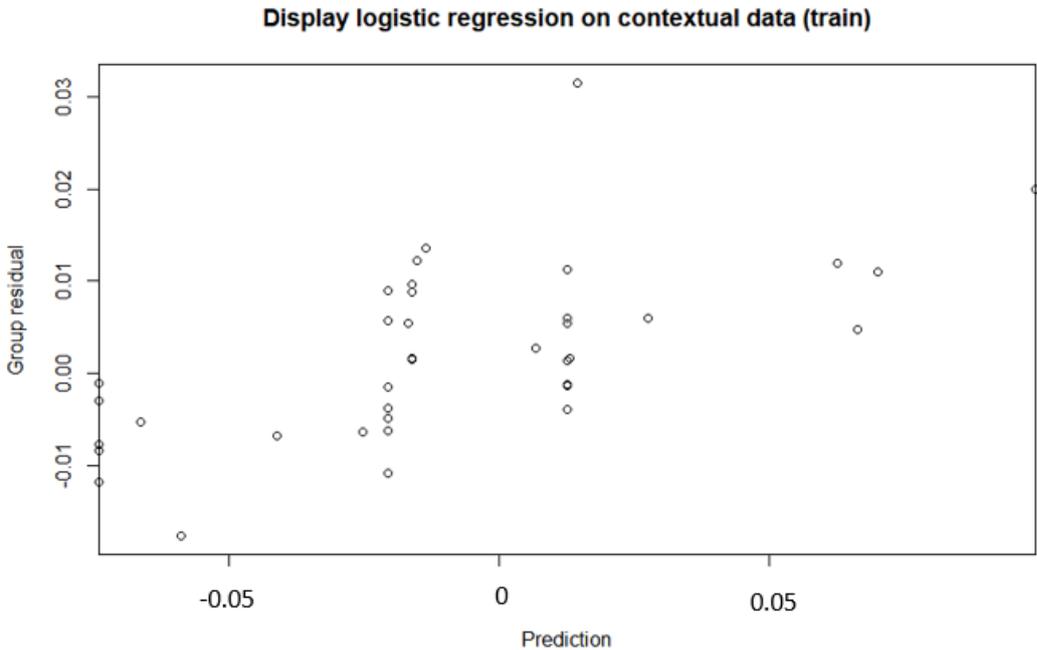


Figure 18, group residual plot showing correlation between the prediction and residuals.

In figure 18 it is possible to see a positive correlation between the group residuals and the predicted conversion rate. This is natural since LASSO punishes more extreme predictions so that predictions are shrunk closer to the average. It does however seem like there could be potential gains from having a smaller lambda and allowing a bit more extreme prediction.

The reason that lambda is big is however to cancel out other noisy variables and prevent overfitting. It is possible that lambda will be made smaller when removing covariates that were not used in the model before fitting. If that is not sufficient another possibility is to relax the 1 standard error rule and use the lambda that maximizes the out of set performance instead, that would however increase the risk overfitting due to selection bias.

[Fitting the final model](#)

A new model, containing the interactions is fit to the data in the same way as previous models. Figure 19 displays the different covariate levels included and their importance to the model.

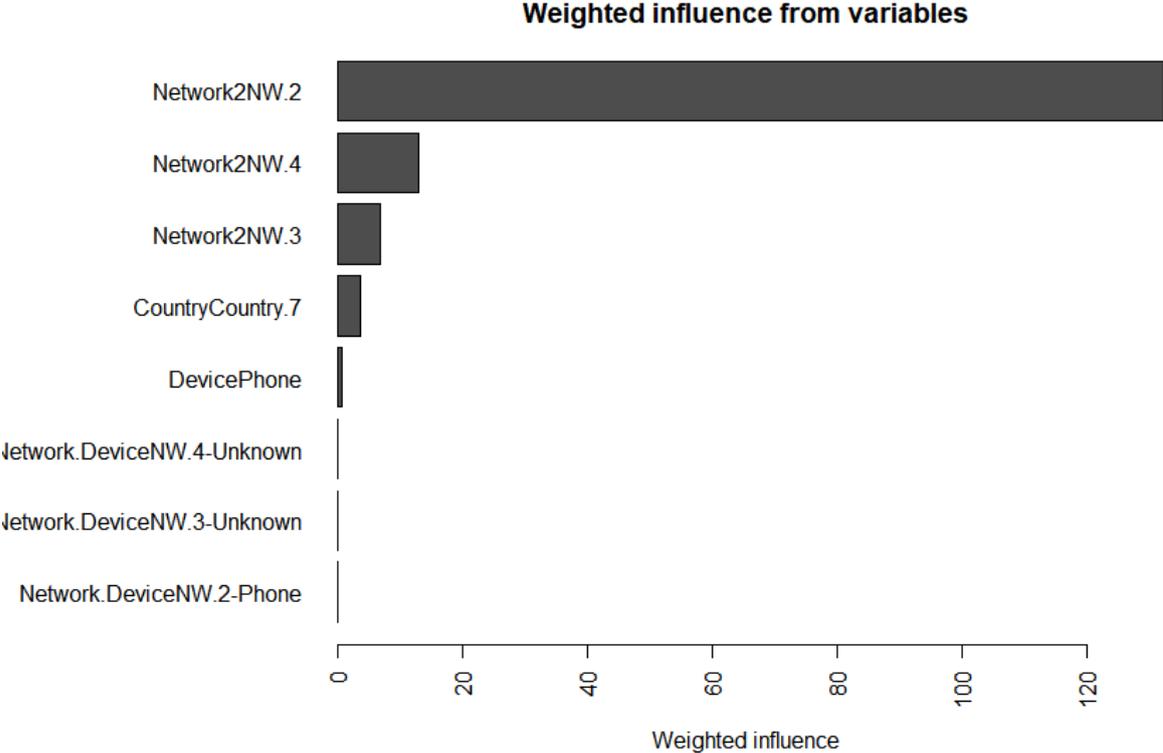


Figure 19, variable importance model with interactions.

Covariates that were not included into this model was removed from the selection to reduce noise. The new subset of covariates should only contain the most important variables and a final model was fitted with these covariates following the same procedure as before. Figure 20 displays the different covariate levels included and their importance to the final model.

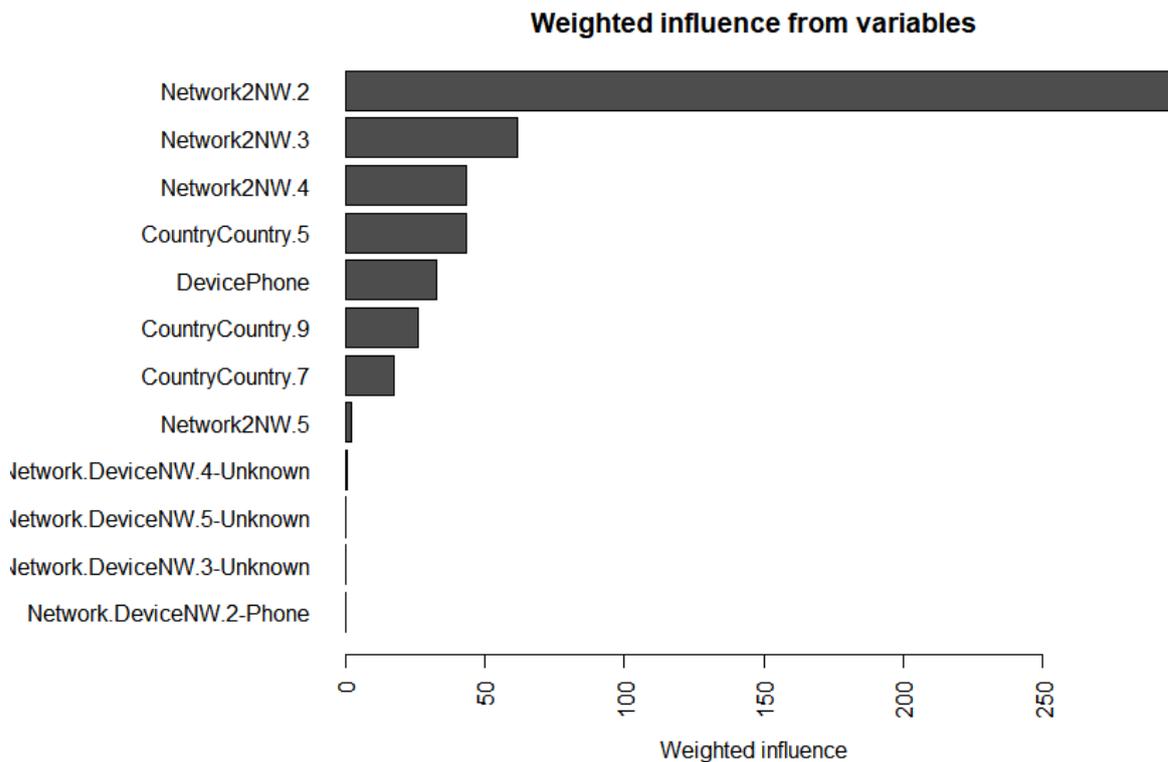


Figure 20, influence of variables for final model.

Comparing the number of variables included into the models in figure 15 and 19, it's clear that once variables that only adds noise are removed that the lambda gets lower and more levels are included.

#### Logistic regression with early conversion rates - Multivariate analysis

The same procedure that was used for finding the final model for the logistic regression model on contextual data was used. First differences in the residual analysis is described and then the final models are presented.

#### Residual analysis

When looking at possible interactions, a new candidate appears. From figure 21 it seems possible that there is an interaction between network and device that has an impact on the outcome. The interaction does get into the final model and thus, seems to have an actual influence.

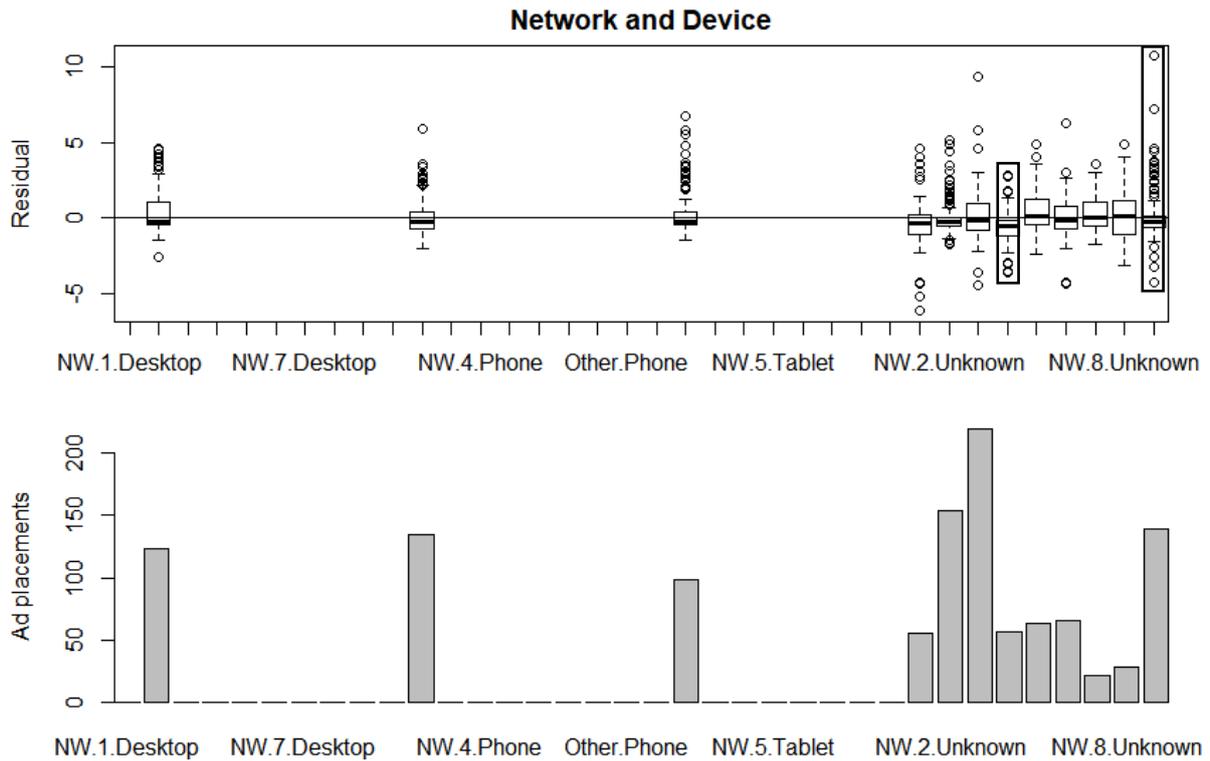


Figure 21, boxplot showing a possible interaction between the covariates network and device.

Figure 22 shows the grouped residuals related to the conversion rate from impression to lead, that is

$$CVR_{Impression\ to\ Lead} = CTR * CVR_{Click\ to\ Lead}$$

It seems like there could be a connection between when both CTR and  $CVR_{Click\ to\ Lead}$  is high and the conversion rate from lead to acquisition (the target variable).

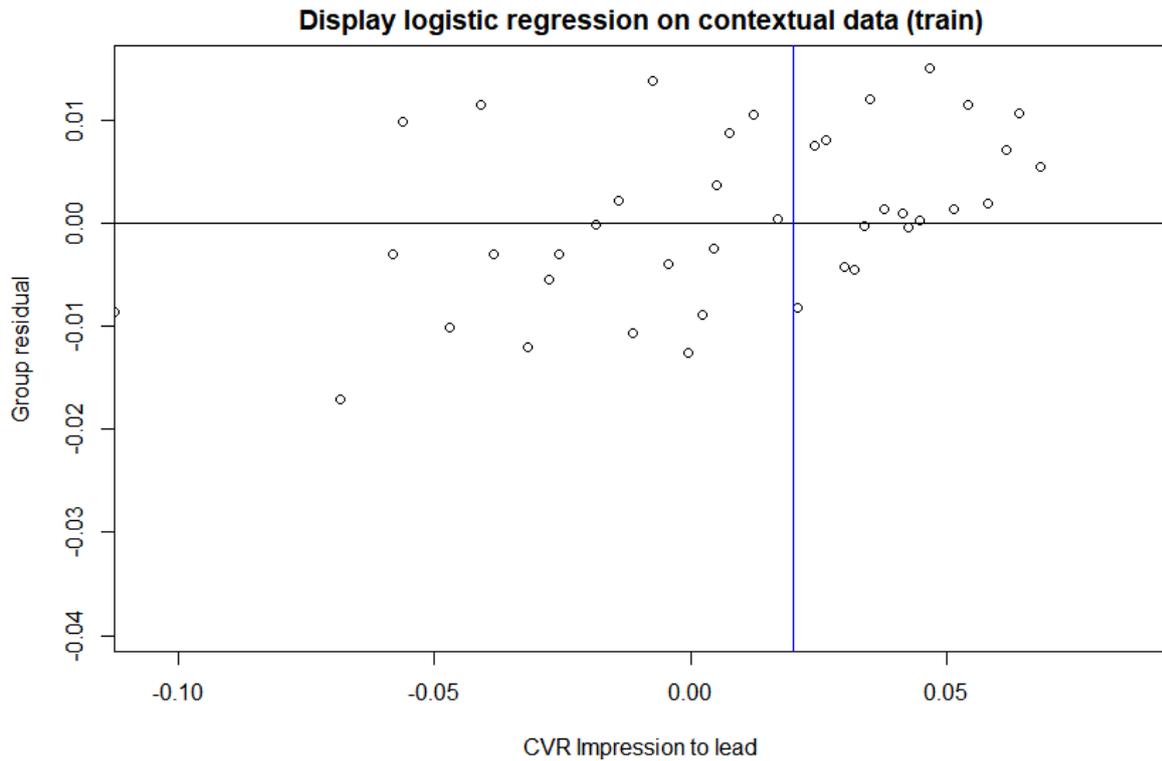


Figure 22, grouped residuals depending on the conversion rate from impression to lead. A line is added to the plot to show that there seems to be different behaviors for conversion rates smaller than 0.02 compared to the ones that are larger than 0.02.

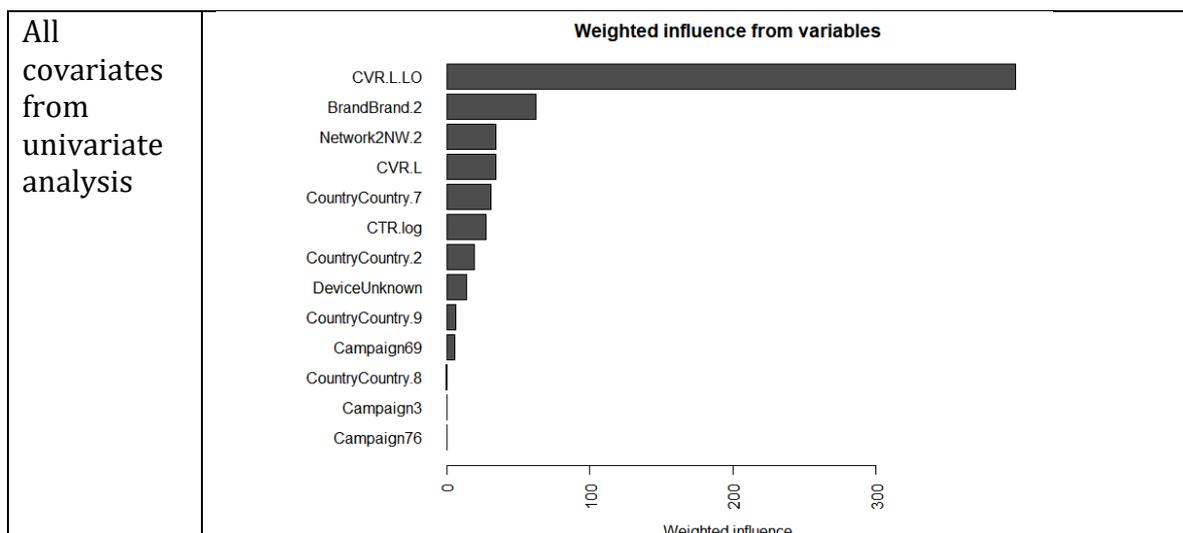
A variable was created

$$CVR'_{Impression\ to\ Lead} = \begin{cases} CTR * CVR_{Click\ to\ Lead} & \text{if } CTR * CVR_{Click\ to\ Lead} > 0.02 \\ 0 & \text{if } CTR * CVR_{Click\ to\ Lead} \leq 0.02 \end{cases}$$

This variable did however not make it into the final model.

### Results

The covariates included in the models are presented in figure 23



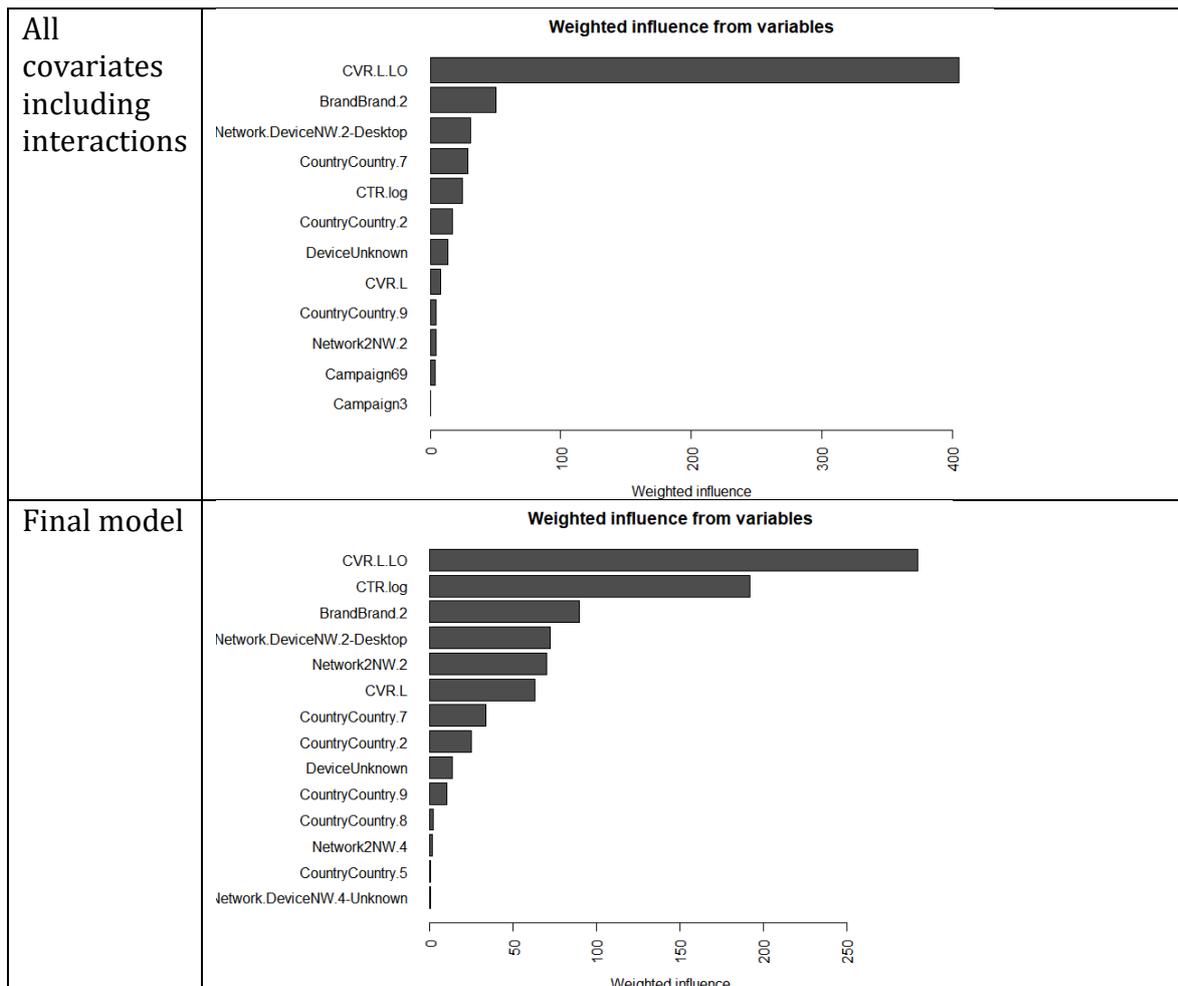


Figure 23, importance of variables in the final models.

*Logistic regression with latent variables - Multivariate analysis*

The earlier conversion rates were predicted and the residuals from these predictions were used as covariates for a new model, as described in chapter 6.4.4. In figure 24 the new covariates are plotted against the residuals from the logistic regression with early conversion rates model.

From looking at figure 24 it the relationship seems to have a V-shape. To handle this the variable is split into two pieces.

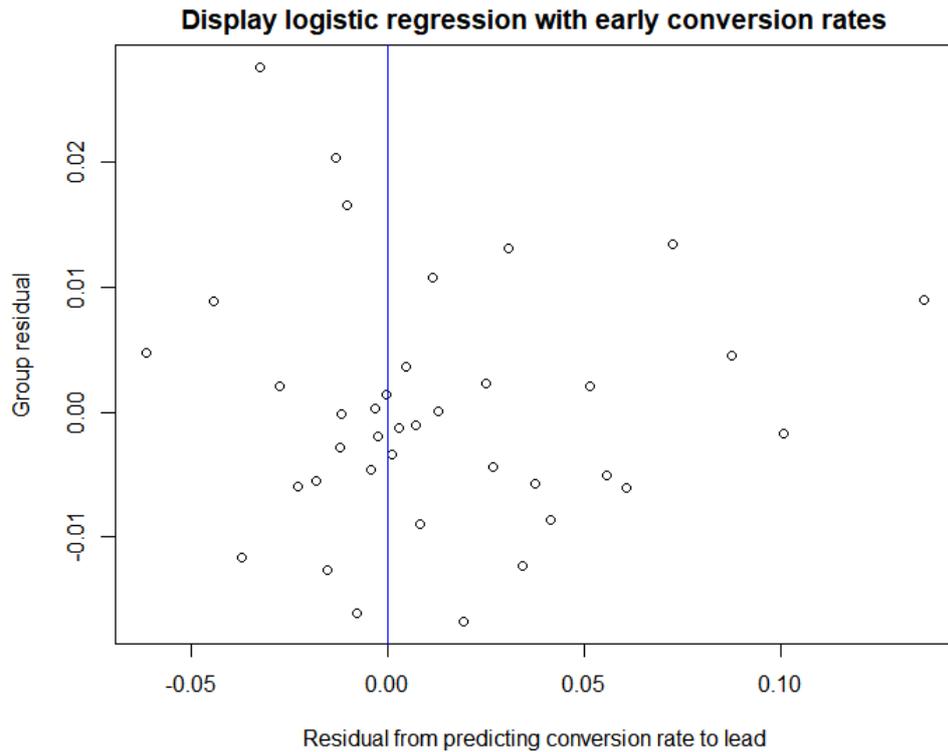


Figure 24, grouped residuals from the logistic regression with early conversion rates model grouped by the residuals from predicting the conversion rate from click to lead.

3 new variables are created from the residuals and these together with the variables from the final logistic regression with early conversion rates model are used to fit a new model.

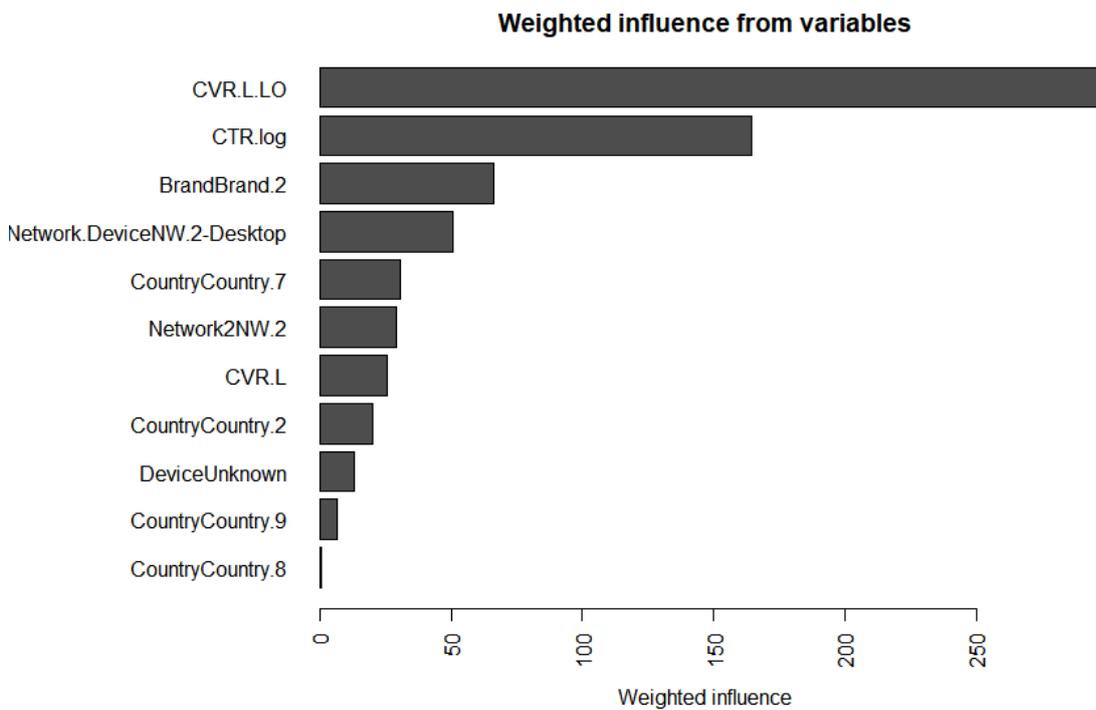


Figure 25, influence of different covariates in the logistic regression with latent variables. CVR.L.LO is the log-odds of the conversion rate from click to lead.

As is clear from figure 25, none of the latent variables got included into the final model. This means that the logistic regression with latent variables ends up being the very same model as the logistic regression with early conversion rates.

#### 6.5.4 Results

The performance of the different models is then tried on the test set. Since the choice of lambda is stochastic and has a lot of influence on the final model it is interesting to get a more stable estimate of the performance as well as of the variance. To get this, the final models are refitted 10 times each.

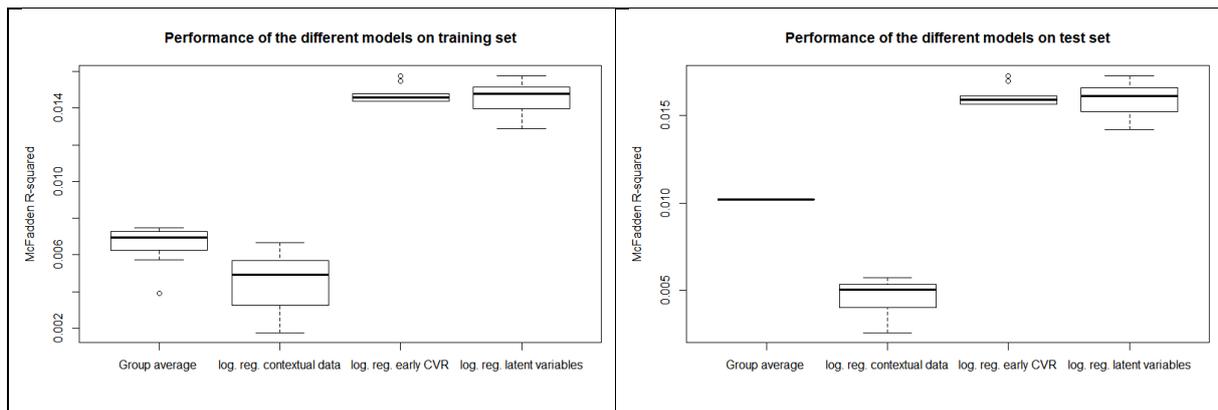


Figure 26, Performance of the different models on the training set and the test set.

Since the logistic regression with latent variables did not include any of the latent variables it is the same as the logistic regression with early conversion rates model.

#### 6.6 Ad type

In the data used so far, no information about the ad creative is used. However, according to the experience of the marketers at company X the ad creative has a big influence on conversion rates. The most important factor according to them is the message of the ad creative. To make use of this information a new variable was created.

The new covariate was created through categorizing the ads as directed to create interest or sales. A list of words associated with creating interest and another with words associated with being more sales driven were created. These lists were then translated into all the different languages that the company uses in their ads. All texts belonging to the different ad creatives were then categorized into one of six categories, based on the occurrence of the different words from the lists in the ads. The different categories are

- Sales driven – The ad creative contains only words associated with creating sales.
- Mixed with focus on sales – The ad creative contains words both associated with creating sales and interest. The majority of the words are associated with creating sales.
- Mixed – The ad creative contains words both associated with creating sales and interest. There are just as many word associated with creating sales as creating interest.
- Mixed with focus on interest – The ad creative contains words both associated with creating sales and interest. The majority of the words are associated with creating interest.

- Interest – The ad creative contains only words associated with creating interest.
- Other – The ad creative doesn't contain any words associated with creating sales or interest.

After the new model was created the same process as before was followed to create the four different models.

### 6.6.1 Group average

First a wide search is started with many different lambdas to see where a potential optimum for the cut off point could be.

McFadden R-squared	Levels	Cut off point
0.01414689	-Ad.type-Network2-Brand	7.389056
0.01501946	-Ad.type-Network2-Brand-Dimensions2	20.085537
0.01401961	-Ad.type-Network2-Brand	54.598150
0.01279203	-Ad.type-Dimensions2-Brand-Device	148.413159
0.01194466	-Network2-Ad.type-Dimensions2-Brand	403.428793
0.01196600	-Ad.type-Dimensions2	1096.633158

Figure 27, output from group average search with the ad type feature included as a covariate.

Once an approximate range is discovered a new grid search is conducted to find the optima.

McFadden R-squared	Levels	Cut off point
0.01462642	-Ad.type-Network2-Brand	5.000000
0.01509922	-Ad.type-Network2-Brand	10.000000
0.01488690	-Ad.type-Network2-Brand	15.000000
0.01493337	-Ad.type-Network2-Brand	20.000000
0.01456028	-Ad.type-Network2-Brand-Dimensions2	25.000000
0.01466561	-Ad.type-Network2-Brand-Dimensions2	30.000000
0.01435401	-Ad.type-Network2-Brand	35.000000
0.01435697	-Ad.type-Network2-Brand-Dimensions2	40.000000
0.01415125	-Ad.type-Network2-Brand-Dimensions2	45.000000
0.01438967	-Ad.type-Network2-Brand	50.000000

Figure 28, output from group average search with ad type feature for a smaller range of cut off points.

After the second grid search the model with the highest performance is used as a final model.

### 6.6.2 Univariate analysis

The ad type feature was determined to have some explanatory power in the univariate analysis and was hence used as a candidate for the multivariate model.

Variable	Degrees of freedom	P-value
Ad type	4	2.2e-16

Table 5, univariate test to determine if the new feature has any explanatory power.

### 6.6.3 Multivariate analysis

There were no new interactions that appeared when introducing the new variable. In figure 29 the importance of the different variables are presented.

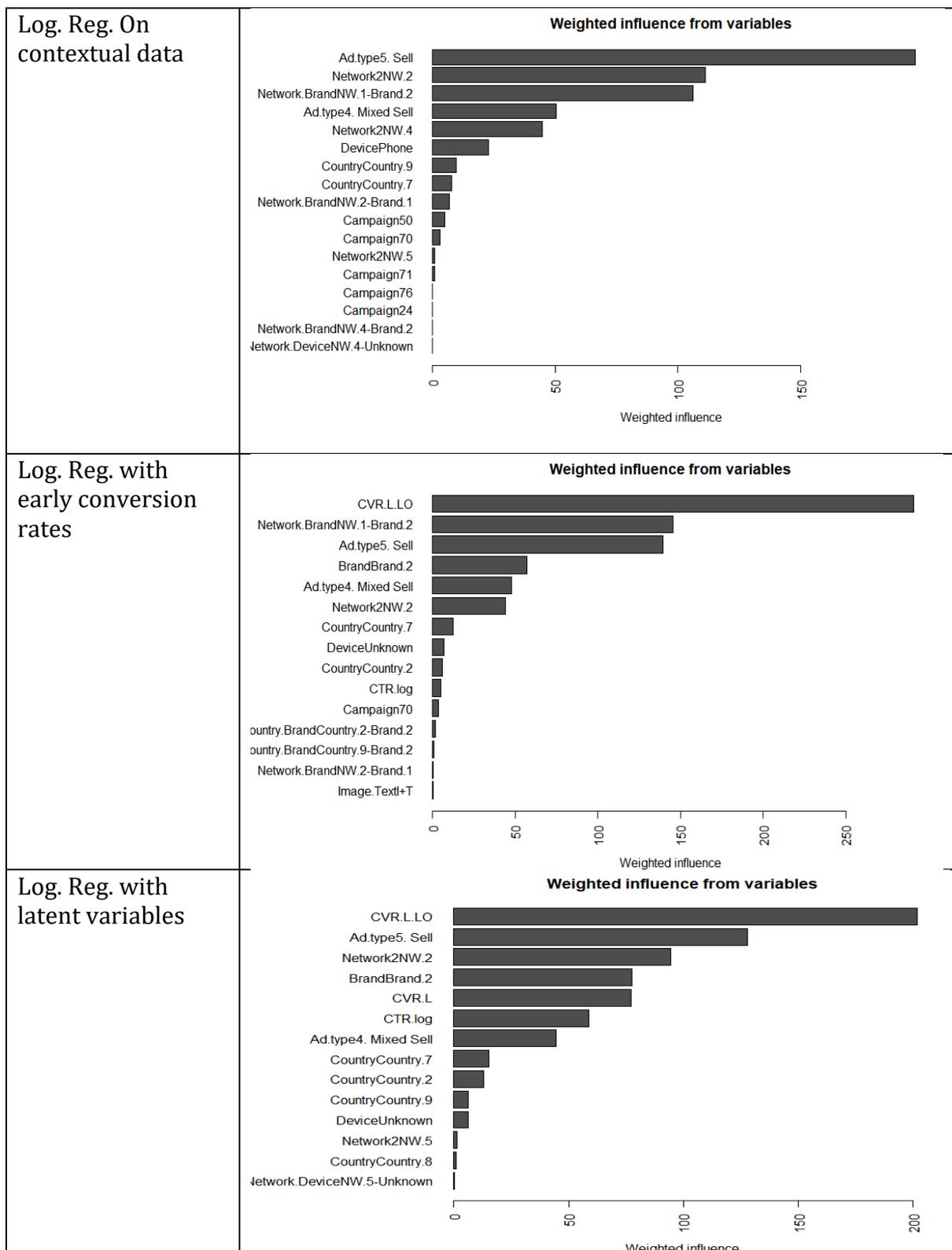


Figure 29, variable importance in the different models.

The final models are used on the training and test set. To get a more stable estimate and an understanding of the variance the final model was refitted 10 times. The result is displayed in figure 30.

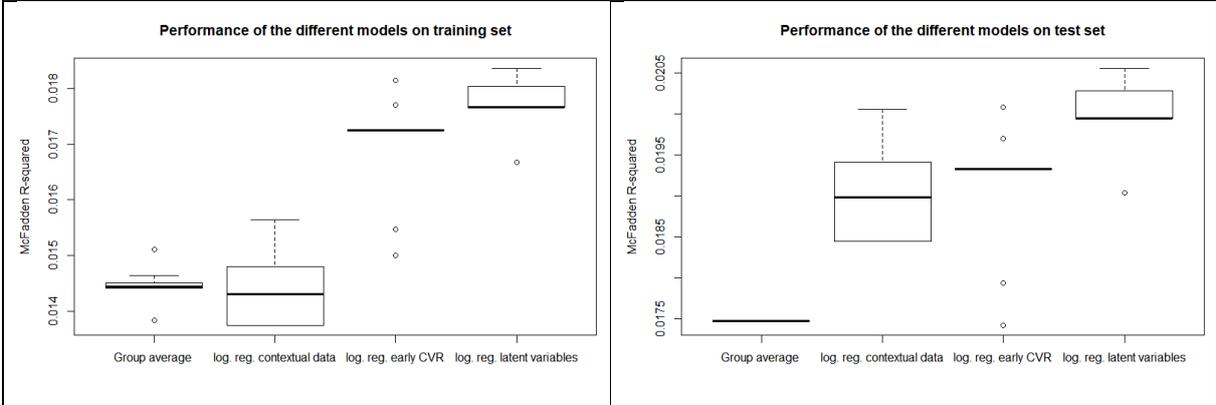


Figure 30, Performance of the different models on the training set and the test set.

It is worth noting that the final logistic regression with latent variables does not use the latent variables and should have the same performance as the logistic regression with early conversion rates model.

6.7 Asses performance

For grouped data the highest possible McFadden R-squared is lower than 1. To find out what the highest obtainable McFadden R-squared that is possible to obtain on the different data sets, the McFadden R-squared was calculated for the observed conversion rates from lead to acquisition. This was then used as a reference for which level of accuracy could be obtained on the data.

Ad network	Display	
Data set	Train	Test
Optimal performance	0,039	0,049

Table 6, optimal performance for predictions on the data set.

While such a procedure gives an upper bound of the accuracy, it highly unlikely that that level of performance will be obtained in a stochastic environment. Even if the exact conversion rates for each category was known, there would still be noise due to that the process is random. The noise of course lowers the performance.

## 7 Results

In this chapter, first the performance of the different models are described and then the covariates used in the models are presented. Lastly the influence of the ad creative is briefly discussed through highlighting the influence of the ad type feature.

### 7.1 Performance of models

The performance of the different models are presented in table 7.

Ad network	Display		Facebook		GDN		SEM	
Data set	Train	Test	Train	Test	Train	Test	Train	Test
Group average model	0,014	<b>0,017</b>	0,014	<b>0,001</b>	0,009	<b>0,006</b>	0,039	<b>0,038</b>
Logistic regression on contextual data	0.014	<b>0.019</b>	0,016	<b>0,008</b>	0,008	<b>0,006</b>	0,039	<b>0,036</b>
Logistic regression with early conversion rates	0,017	<b>0,019</b>	0,030	<b>0,018</b>	0,008	<b>0,005</b>	0,039	<b>0,037</b>
Logistic regression with latent variables	0.018	<b>0.020</b>	0,031	<b>0,019</b>	0,008	<b>0,005</b>	0,039	<b>0,037</b>

Table 7, performance of the different models on training and test set.

While McFaddens R-squared theoretically ranges between 0 and 1, in reality with grouped data, even the exact prediction of the conversion rate for each grouped data point would not generate a score of 1. To get a better sense of how accurate the predictions are in relation to how accurate they could be, the McFadden R-squared from using the observed conversion rates as predictions is calculated. The McFadden R-squared from the models are then divided by this value and presented in table 8.

Ad network	Display		Facebook		GDN		SEM	
Data set	Train	Test	Train	Test	Train	Test	Train	Test
Group average model	38%	<b>36%</b>	27%	<b>2%</b>	17%	<b>11%</b>	56%	<b>55%</b>
Logistic regression on contextual data	38%	<b>39%</b>	31%	<b>20%</b>	15%	<b>11%</b>	57%	<b>54%</b>
Logistic regression with early conversion rates	44%	<b>39%</b>	56%	<b>45%</b>	15%	<b>11%</b>	56%	<b>53%</b>

Logistic regression with latent variables	46%	<b>41%</b>	58%	<b>47%</b>	15%	<b>11%</b>	56%	<b>53%</b>
---	-----	------------	-----	------------	-----	------------	-----	------------

Table 8, performance of the different models on training and test set relative to optimal performance.

## 7.2 Final models

In this section the different final models are presented on a high level. For further details of the importance of each covariate in the different models see appendix 2.

### 7.2.1 Display

Group average model:

	With Ad type feature
Cut off point	10
Level 1	Ad type
Level 2	Network
Level 3	Brand

Table 9, summary of group average model for the display ad networks.

Logistic regression models:

	Log. Reg. on contextual data	Log. Reg. with early conversion rates	Log. Reg. with latent variables
Network	X	X	X
Country	X	X	X
Device	X	X	X
Campaign	X	X	X
Ad type	X	X	X
Image/text		X	X
Network*Brand	X	X	X
Network*Device	X		
Country*Brand		X	X
Log CTR		X	X
Log-odds $CVR_{Click,Lead}$		X	X

Table 10, summary of the logistic regression models for the display ad network.

### 7.2.2 Facebook

Group average model:

Cut off point	60
Level 1	Country
Level 2	Device
Level 3	Targeting
Level 4	Brand
Level 5	Network

Table 11, summary of group average model for the Facebook ad network.

Logistic regression models:

	Log. Reg. on contextual data	Log. Reg. with early conversion rates	Log. Reg. with latent variables
--	------------------------------	---------------------------------------	---------------------------------

Country	X	X	X
Device	X		
Campaign type	X	X	X
Dimensions		X	X
Brand		X	X
Device*Country	X	X	X
Dimensions*Country	X		
Ad type*Country	X	X	X
CTR		X	X
$CVR_{Click,Lead}$		X	X
$CVR_{Impression,Lead}$		X	X
$r_{CVR_{Impression,Lead}}$			X

Table 12, summary of the logistic regression models for the Facebook ad network.

### 7.2.3 Google display network

Group average model:

Cut off point	20
Level 1	Account
Level 2	Image/text
Level 3	Country
Level 4	Brand
Level 5	Network

Table 13, summary of group average model for the google display ad network.

Logistic regression models:

	Log. Reg. on contextual data	Log. Reg. with early conversion rates	Log. Reg. with latent variables
Country	X	X	X
Brand	X	X	X
Account	X	X	X
Campaign type	X	X	X
Image/text	X	X	X
Ad type	X	X	X
Country*Brand	X	X	X
Country*Image/text	X	X	X
Country*Network		X	X
$CVR_{Click,Lead}$		X	X

Table 14, summary of the logistic regression models for the google display ad network.

### 7.2.4 Search Engine Marketing

Group average model:

Cut off point	15
Level 1	Account
Level 2	Device
Level 3	Account type
Level 4	Country

Table 15, summary of group average model for the search engine marketing ad network.

Logistic regression:

	Log. Reg. on contextual data	Log. Reg. with early conversion rates	Log. Reg. with latent variables
Country	X	X	X
Device	X	X	X
Account	X	X	X
Account type	X	X	X
Country*Device	X	X	X
Country*Account type	X		
Device*Account type	X	X	X
Ad type		X	X
$CVR_{Impression,Lead}$		X	X
$r_{CTR}$			X

Table 16, summary of the logistic regression models for the search engine marketing ad network.

### 7.3 Performance due to ad type covariate

In most channels the ad type feature that was created did not have a profound influence on results but for display it improved the performance. From table 18 it is clear that without the ad type feature, earlier conversion rates improved the performance of the model a lot. However, with the new feature the increase in performance is almost gone. This indicates that the different features might have modelled the same thing.

Display	With ad type		Without ad type	
	Training set	Test set	Training set	Test set
Group average model	0,017	<b>0,014</b>	0,010	<b>0,007</b>
Logistic regression on contextual data	0.019	<b>0.014</b>	0,005	<b>0,005</b>
Logistic regression with early conversion rates	0,019	<b>0,017</b>	0,016	<b>0,015</b>
Logistic regression with latent variables	0.020	<b>0.018</b>	0,016	<b>0,015</b>

Table 17, comparison of performance of prediction models with and without the ad type feature in the display ad networks.

## 8 Discussion and conclusions

### 8.1 How accurate are the models?

None of the created models are very close to the highest possible McFadden R-squared. This indicates that there is room for improvement. To improve the predictions further it will be necessary to either use new features or a different modeling approach. Possible improvements are discussed in section 8.6. It's however worth noting that all models seem to be a lot better than simply using the global average.

### 8.2 Group average vs logistic regression models

The group average model outperforms the logistic regression model on the SEM platform. For GDN the performance is more or less equal. For Facebook and Display the logistic regression models are performing better than the group average model, especially when leveraging earlier conversion rates and latent variables.

What seems to have made the difference is that in SEM the range of conversion rates is much larger than in the other platforms. When using a penalty function, such as in lasso, it shrinks the predictions towards the mean. The penalty is chosen in such a way that the trade off between overfitting and shrinking the predictions too much is balanced. In group average however, overfitting is prevented through the cut off point, and the predictions are not shrunken. This could be the reason why the group average performs better on SEM.

When comparing the range of the predictions between the logistic regression models and the group average for SEM the range is much tighter for the logistic regression models (the standard deviation of the predictions is about 15% lower for the logistic regression models). This indicates that it could indeed be the shrinkage which causes the logistic regression model to perform worse on SEM. For the other platforms the logistic regression model is performing better or similarly to the group average model.

There does not seem to be a silver bullet, but the model has to be chosen based on the problem. Starting with a logistic regression model and looking at if there seems to be too much shrinkage could however be a good general approach.

### 8.3 Using latent variables

In both Facebook and SEM the latent variable, the residual from predicting earlier conversion steps, was included into the final model. The increase in performance was however quite small. It should be considered that it is a rather computationally heavy and time consuming to create two extra models for doing predictions on the earlier conversion rates to be able to create the features.

Comparing the rather insignificant gains of using latent variables to the extra complexity of the model and the computer power necessary to build it, it's hard to say that using latent variables is worth the effort. It is probably possible to pick lower hanging fruit in other areas when it comes to improving the performance of the predictions.

### 8.4 Why group average fails for Facebook

As can be seen in table 7, the performance of the group average model is very low on the test set compared to the training set. After some analysis it is clear that this is because

the performance on the high volume ads is very low. If the 4 ads with the highest volume would be removed from the test set the performance, measured as the McFadden R-squared, would go from 0,001 to 0,007.

Facebook ad placements tend to be high volume during a short period of time. With this in mind it, is not optimal to aggregate the data points once per week, since this could mean that predictions for an ad placement is not updated even when there is a large amount of data that could support a new prediction.

Having more granular data points would have distributed the observations more evenly between the test and training set. This would have made it possible to make predictions based only on the observed performance of the biggest ad placements which should have improved the performance rather drastically.

### 8.5 Variability when using observed values

In chapter 6.4.3 the impact of the variance of the observed early conversion rates was discussed. There was however not a drop in performance for the low volume ad placements when using these features, which could have been expected. A possible explanation for why it did not happen, is that since the volume for the earlier conversion rates is so much higher than for the later, the observed values are quite stable.

### 8.6 Further improvements

There could be other modeling approaches that are better suited for making the predictions. However, no model is better than the underlying data so this section gives some examples of features that could be used to improve the predictions.

#### 8.6.1 Grouped data

Aggregating data points is a necessity to be able to handle the amounts of information in the data set. However, when aggregating information is also lost which reduces the possibility to do accurate predictions. To balance computational convenience against information loss, tradeoffs needs to be made.

A more granular data set can reveal details that are important, an example is that it's known that the weekday can have an influence on conversions for company X. However, aggregating on days instead of weeks would mean that the data set would be about 5 times bigger making some heavy computations even heavier. If the models are created through cloud computing using big data techniques this might not be a problem. On a stationary 8 core computer it is.

In this thesis aggregation was used at an early stage to make the project computationally feasible. A different approach could have been sampling out a smaller data set to explore the data, find interesting features and then aggregating to reduce the dimensions of the problem by only aggregating on features that look promising.

#### 8.6.2 Burn out effect

As described in chapter 6.5.3 there might be a "burn out" effect such that when an ad has been published in the same environment for a long time, users learn to ignore it. A possibly useful measure for this would be the average of how many times the ad placement has been shown to a user.

### 8.6.3 Features from the ad creative

In this thesis only one feature was created from the ad creative, the ad type feature. There are of course many possible features that could be created from it. It's not too farfetched to think that the ad creative might influence users behavior, and that such information can be useful for predicting conversion rates. However, the usefulness of such features depends on multiple things.

While ad type was included to almost all models, and there was a strong belief amongst in the online marketing department that it has a big influence on the conversion rate, it was only on the display platform that it made a big difference in performance. The reason for the impact being so small on the other platforms was that in most cases there was only one ad type used for a certain campaign type or account type. When the feature correlates so strongly with another feature it will of course have a limited impact including both of the features into the model.

Before putting the time and effort into creating new features from the ad creative, it can thus be well worth looking into if there is a strong correlation to another feature. If that is the case, it might be easier to simply use the other feature as a proxy. It should be mentioned though that there might be other benefits of mapping what features the ad creatives have and where they are used outside the realm of predicting conversion rates.

## 9 Further research

The group average model can be further explored. A constraint used in this thesis was that all the levels of the model have the same cut off point, this constraint can be loosened. If each level could have its own cut off point, the performance of the model could possibly be improved, however it would also become more prone to overfitting through selection bias.

The final group average model did sometimes vary between different iterations, indicating that what data gets sorted into what set impacts the results. To get a more stable model, multiple group average models could be created, the prediction of the total model being the average predictions of the different sub models.

This thesis has focused on comparing logistic regression models to a model that company X currently uses. There are of course many other machine learning algorithms that can be applied to the problem such as gradient boosting, neural networks and KNN to mention a few. Since the different models made very different predictions for the same ad placements it could also be interesting to look at ensemble models.

## 10 References

- [1] R. Briggs and N. Hollis, "Advertising on the Web: Is there Response Before Clickthrough?," *Journal of Advertising Research*, nr April, p. 33–45, 1997.
- [2] "Statcounter," [Online]. Available: <http://gs.statcounter.com/search-engine-market-share/desktop-tablet-console/worldwide/#monthly-201607-201707-bar>. [Använd 11 08 2017].
- [3] bluecaribu, "Las métricas del ego de los anunciantes en Google Adwords: visibilidad y posición," bluecaribu, [Online]. Available: <http://www.bluecaribu.com/metricas-ego-adwords/>. [Använd 17 May 2017].
- [4] M. Richardson, E. Dominowska and R. Ragno, "Predicting Clicks: Estimating the Click-Through Rate for New Ads," i *16th International Conference on the World Wide Web*, 2007.
- [5] A. Agarwal, K. Hosanagar and M. D. Smith, "Location, location, location: An analysis of profitability in online advertising markets.," *Journal of Marketing Research* 48(6), pp. 1057-1073., 2011.
- [6] marketingland, "marketingland," 17 April 2017. [Online]. Available: <http://marketingland.com/library/display-advertising-news>.
- [7] Göteborgs Posten, "www.gp.se," Göteborgs Posten, [Online]. Available: <http://www.gp.se>. [Använd 17 May 2017].
- [8] "https://en.wikipedia.org/wiki/Advertising\_network," [Online]. Available: [https://en.wikipedia.org/wiki/Advertising\\_network](https://en.wikipedia.org/wiki/Advertising_network). [Använd 21 10 2017].
- [9] "Quality Score: Definition," [Online]. Available: <https://support.google.com/adwords/answer/140351?hl=en>. [Använd 02 12 2017].
- [10] T. Hastie, R. Tibshirani and J. Friedman, "Fitting Logistic Regression Models," i *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer, 2016, pp. 120-121.
- [11] T. Robert, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B*, vol. 58, nr 1, pp. 267-288, 1996.
- [12] T. Hastie, R. Tibshirani and J. Friedman, "Bias, Variance and Model Complexity," i *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, Springer, 2016, p. 222.

- [13] G. J. McLachlan, K.-A. Do and C. Ambroise, *Analyzing microarray gene expression data*, Wiley, 2004.
- [14] G. J. McLachlan and C. Ambroise, "Selection bias in gene extraction on the basis of microarray gene-expression data," i *National Academy of Sciences of the United States of America*, 2002.
- [15] T. Hastie, R. Tibshirani and J. Friedman , "Cross-Validation," i *he Elements of Statistical Learning - Data Mining, Inference, and Prediction*, Springer, 2016, p. 244.
- [16] J. Freese and L. J. Scott , *Regression Models for Categorical Dependent Variables Using Stata*, College Station: Stata Press, 2006.
- [17] G. Blom, J. Enger, G. Englund, J. Grandell and L. Holst, "Chi2 test," i *Sannolikhetsteori och statistikteori med tillämpningar*, Lund, Studentlitteratur, 2005, pp. 347-348.
- [18] . L. Fahrmeir and G. Tutz, *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer, 1994.
- [19] Unknown, "Quality Score: Definition," Google Adwords, [Online]. Available: <https://support.google.com/adwords/answer/140351?hl=en>. [Använd 24 06 2017].
- [20] S. Lemeshow and D. W. Hosmer, "Model building strategies and methods for logistic regression," i *Applied logistic regression*, New York, John Wiley & Sons, INC, 2000, p. 93.
- [21] "Las métricas del ego de los anunciantes en Google Adwords: visibilidad y posición," [Online]. Available: <http://www.bluecaribu.com/metricas-ego-adwords/>. [Använd 25 6 2017].
- [22] Google, "About your account organization," 21 June 2017. [Online]. Available: <https://support.google.com/adwords/answer/1704396?hl=en>.
- [23] Unknown, "FAQ: WHAT ARE PSEUDO R-SQUAREDs?," University of California Los Angeles, 20 October 2011. [Online]. Available: <https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/>. [Använd 10 04 2017].
- [24] Unknown, "Om Googles Display-nätverk," Google Adwords, [Online]. Available: <https://support.google.com/adwords/answer/2404190?hl=sv>. [Använd 18 06 2017].
- [25] P. Breheny, "GLM Residuals and Diagnostics," 26 03 2013. [Online]. Available: <https://web.as.uky.edu/statistics/users/pbreheny/760/S13/notes/3-26.pdf>. [Använd 02 12 2017].

## 11 Appendix 1.

### Example data:

This data is made up but resembles the data provided by Company X that has been used for the analysis.

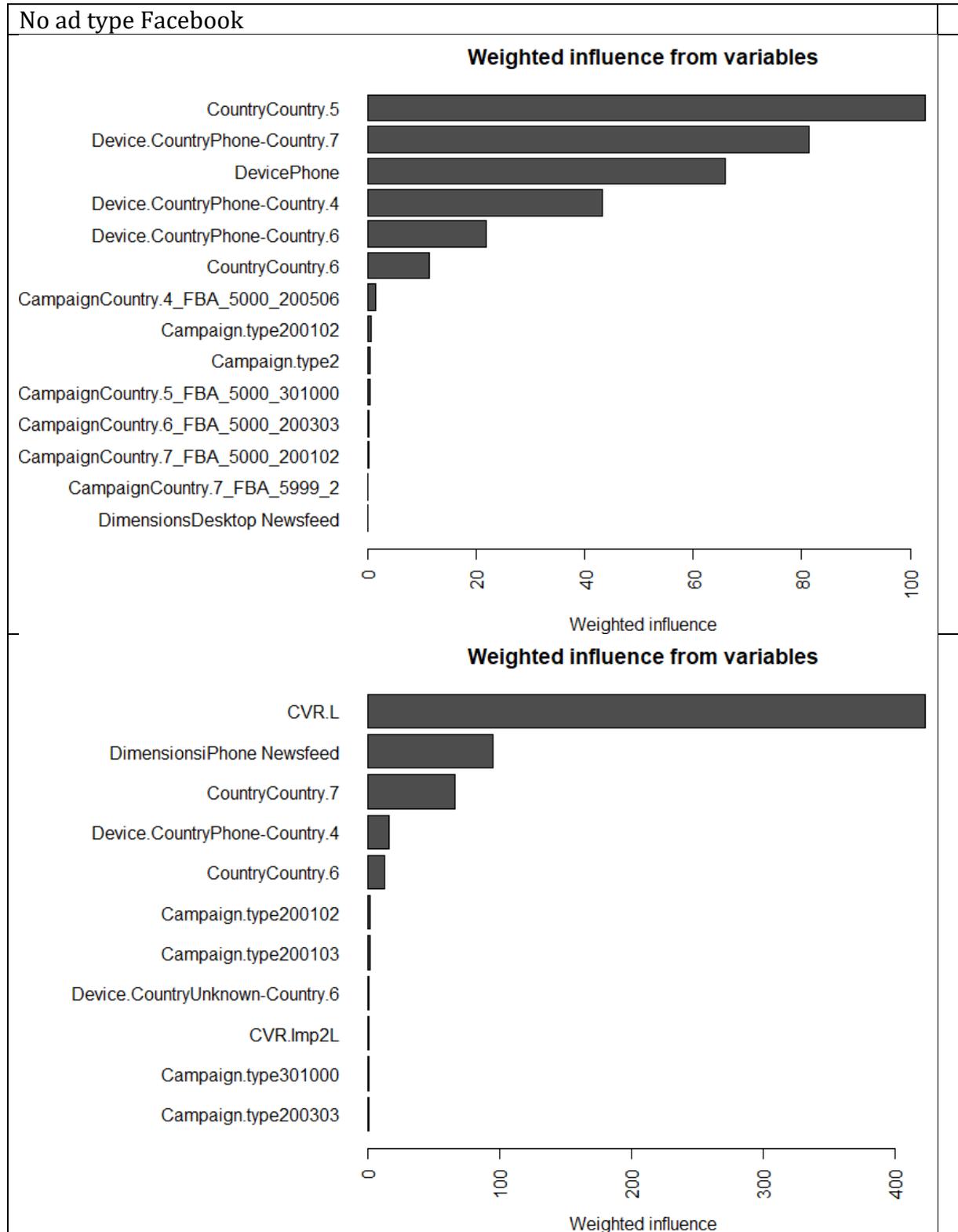
CHANNEL	NETWORK	COUNTRY	DEVICE	BRAND	ACCOUNT	ACCOUNT TYPE	CAMPAIGN	CAMPAIGN TYPE
DISPLAY	(blank)	Country.1	Tablet	WKDA	C1-001-02-13-543	Sport pages	13	Re-target
GDN	(blank)	Country.5	Phone	WKDA	C5-003-123-3-221	News	3	Seasonal
FACEBOOK	Instagram	Country.3	Desktop	WKDA	C3-002-17-4-779	Youth	4	Trends
DISPLAY	(blank)	Country.3	Desktop	WKDA	C3-002-17-5-779	Gaming	5	Tech

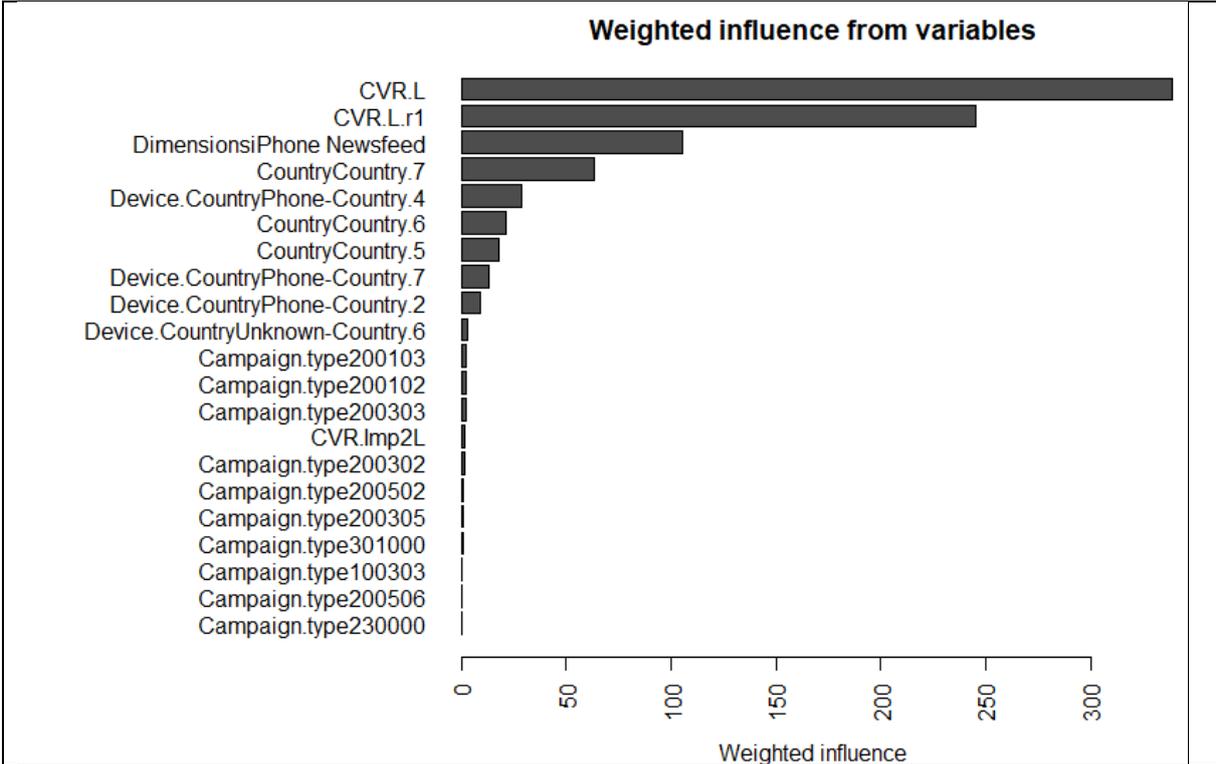
AD GROUP	DIMENSIONS	AD ID	AD TYPE	WEIGHTED AVGPPOSITION	WEIGHTED QUALITYSCORE	IMAGE/TEXT
Local, north	TEXT & small pic	214	3. Mixed	(blank)	(blank)	I+T
Small papers	300x250	12	Other	(blank)	(blank)	T
Male 14-18	Sidebar	478	Other	(blank)	(blank)	I
(blank)	TEXT & small pic	965	Other	(blank)	(blank)	I

	IMPRESSIONS	CLICKS	LEADS	BOOKINGS 30DAYS
	2133	3	0	0
	1999	876	52	20
	200	6	1	1
	20000	20	3	1

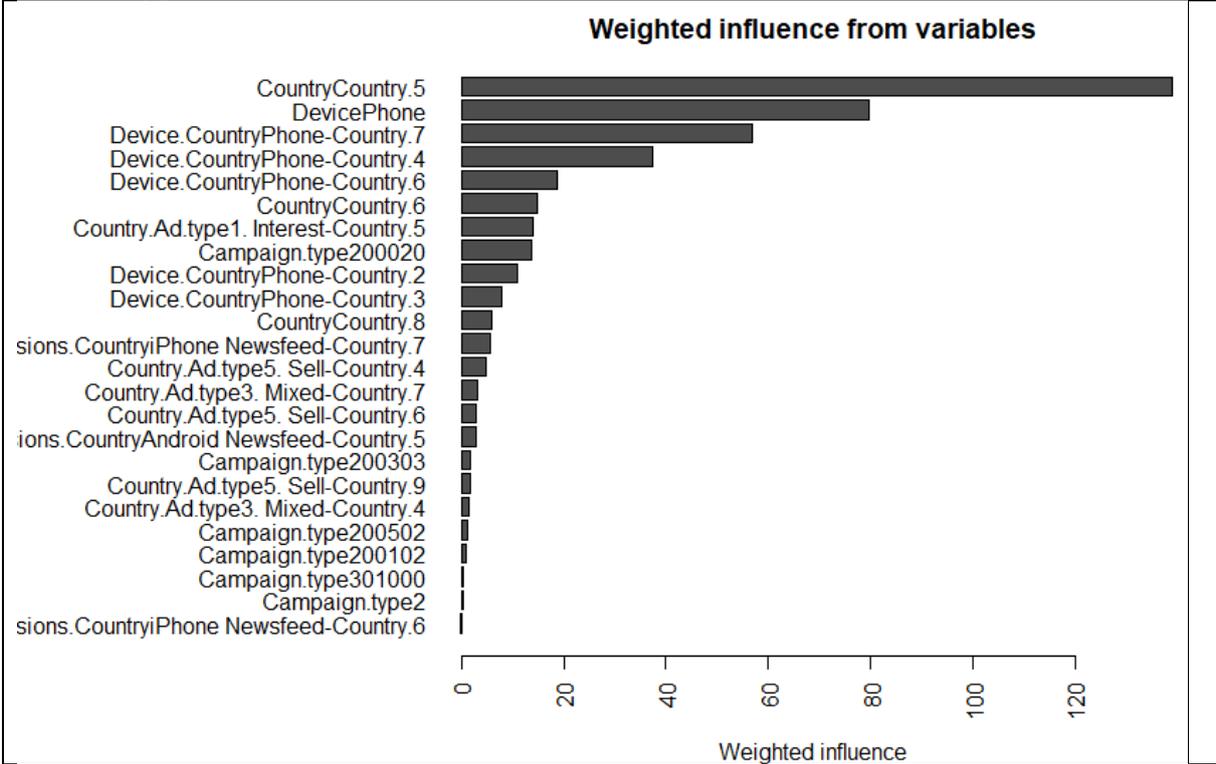
## 12 Appendix 2.

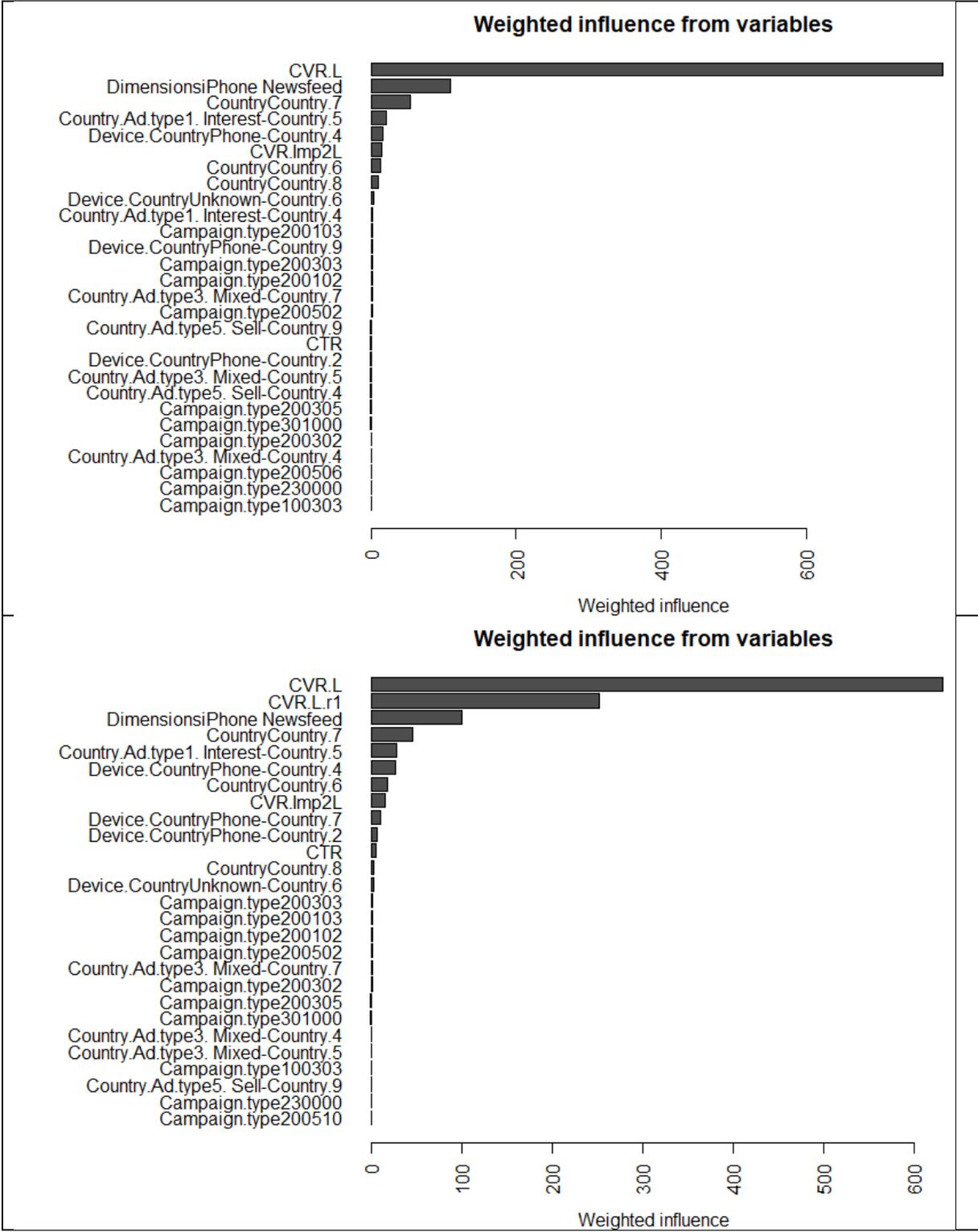
### 12.1 Facebook





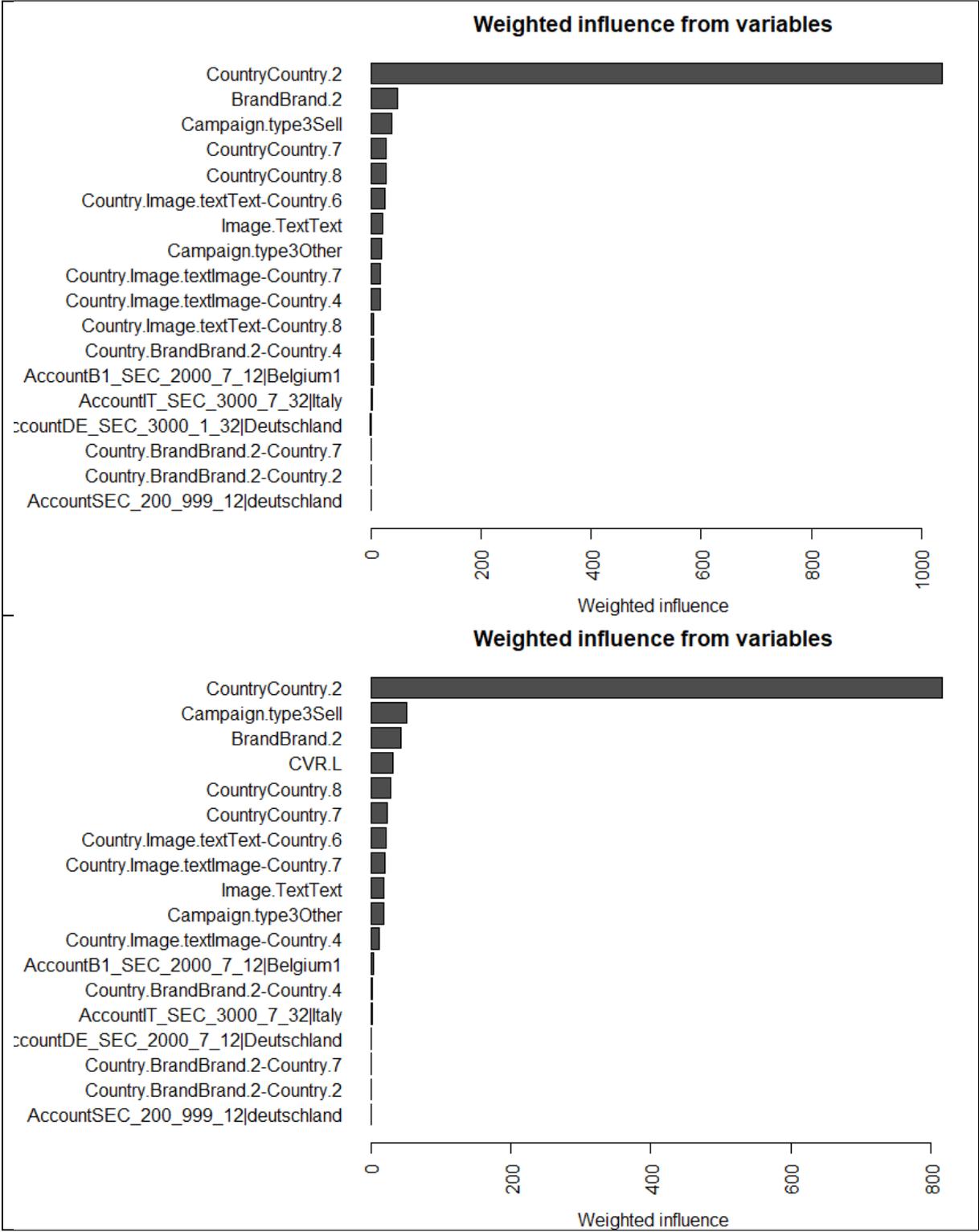
**With ad types**

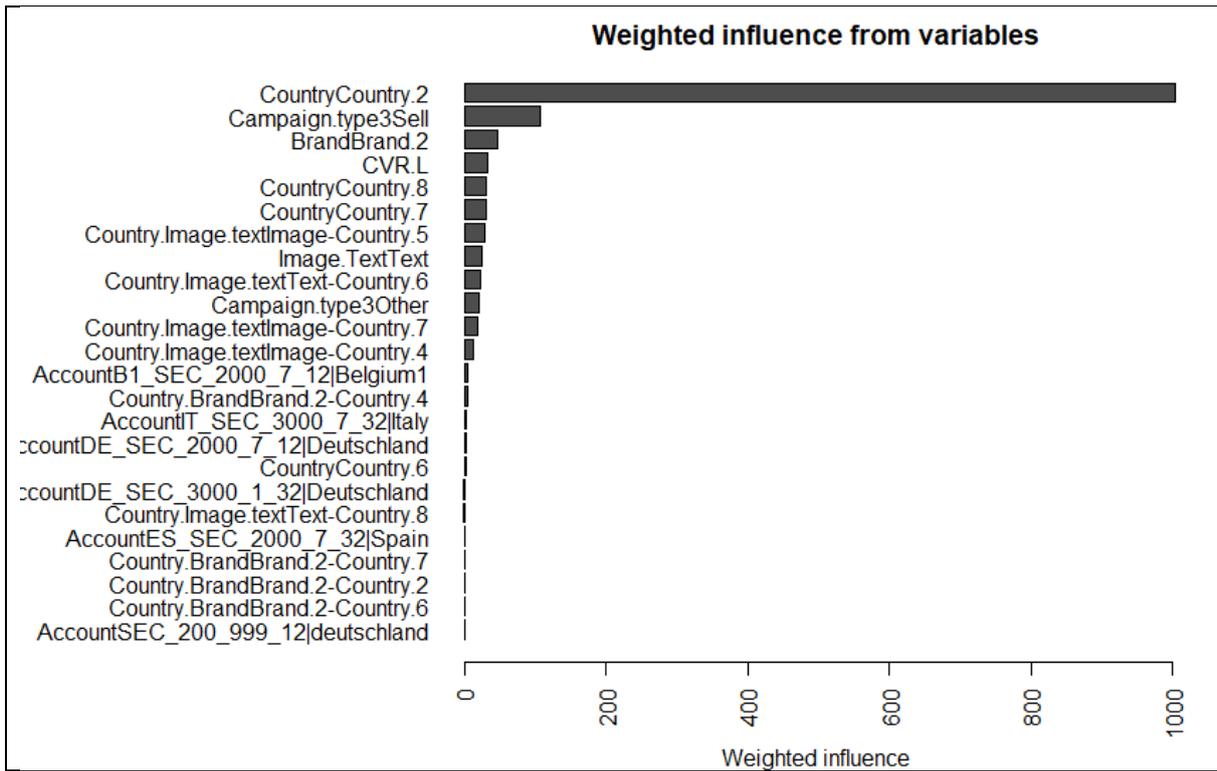




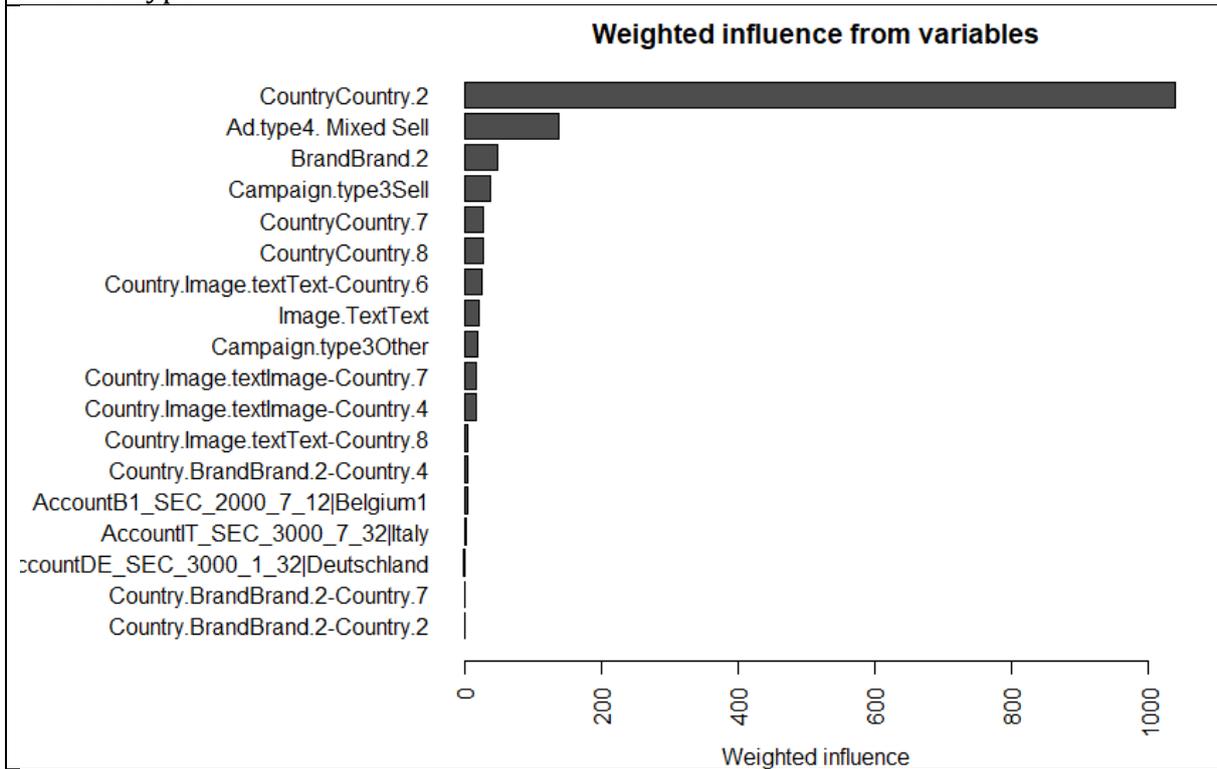
12.2 GDN

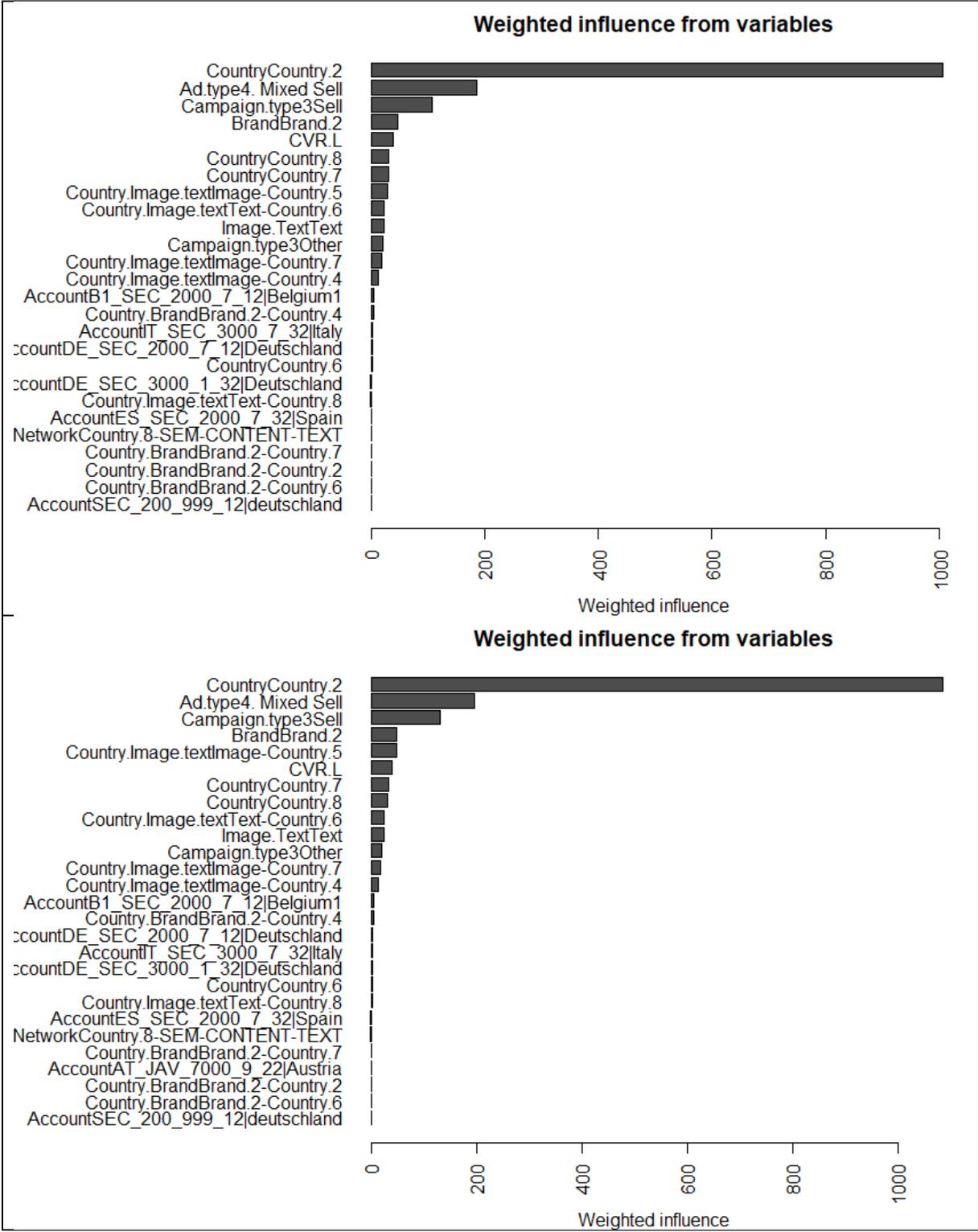
No ad type



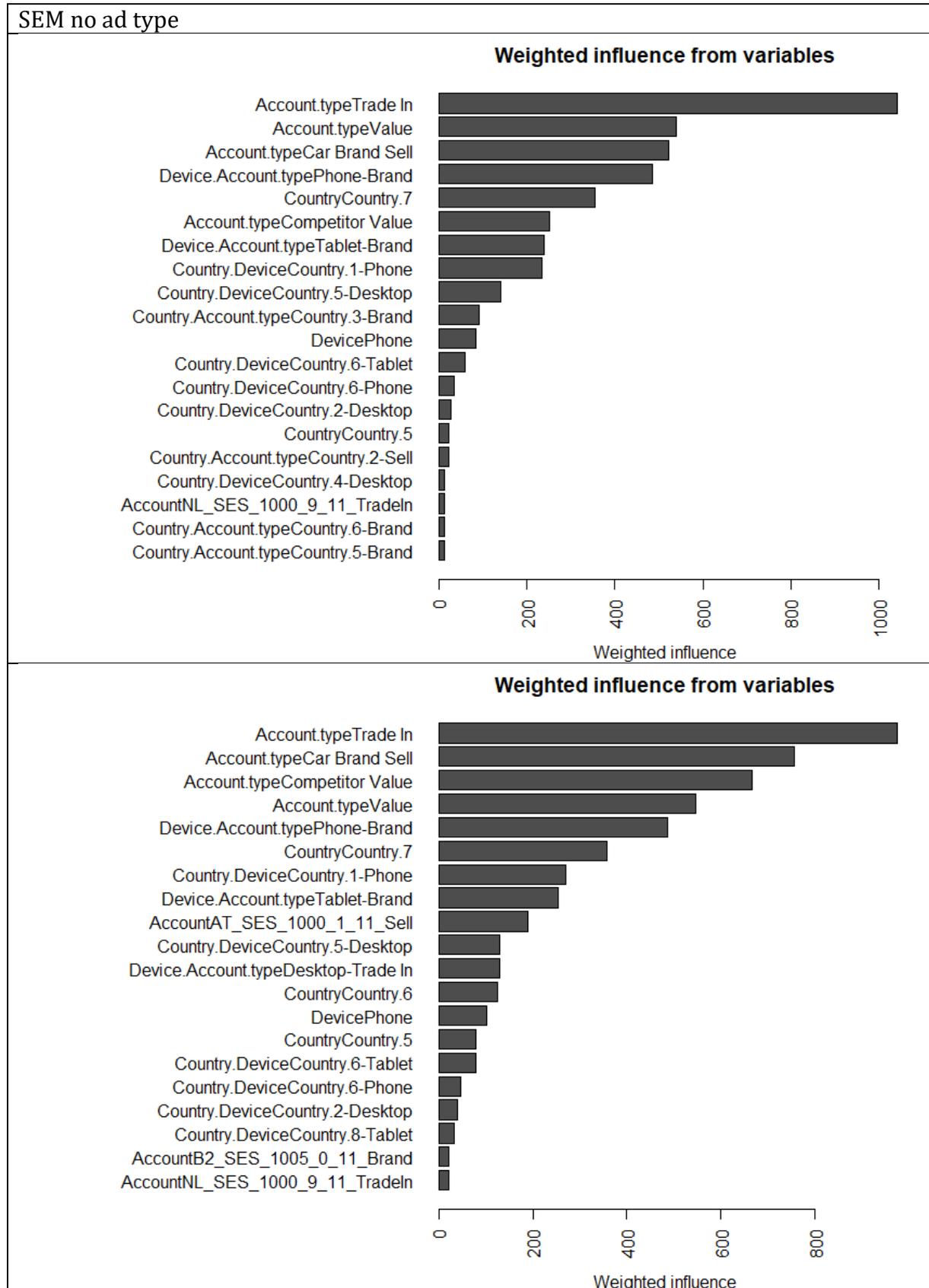


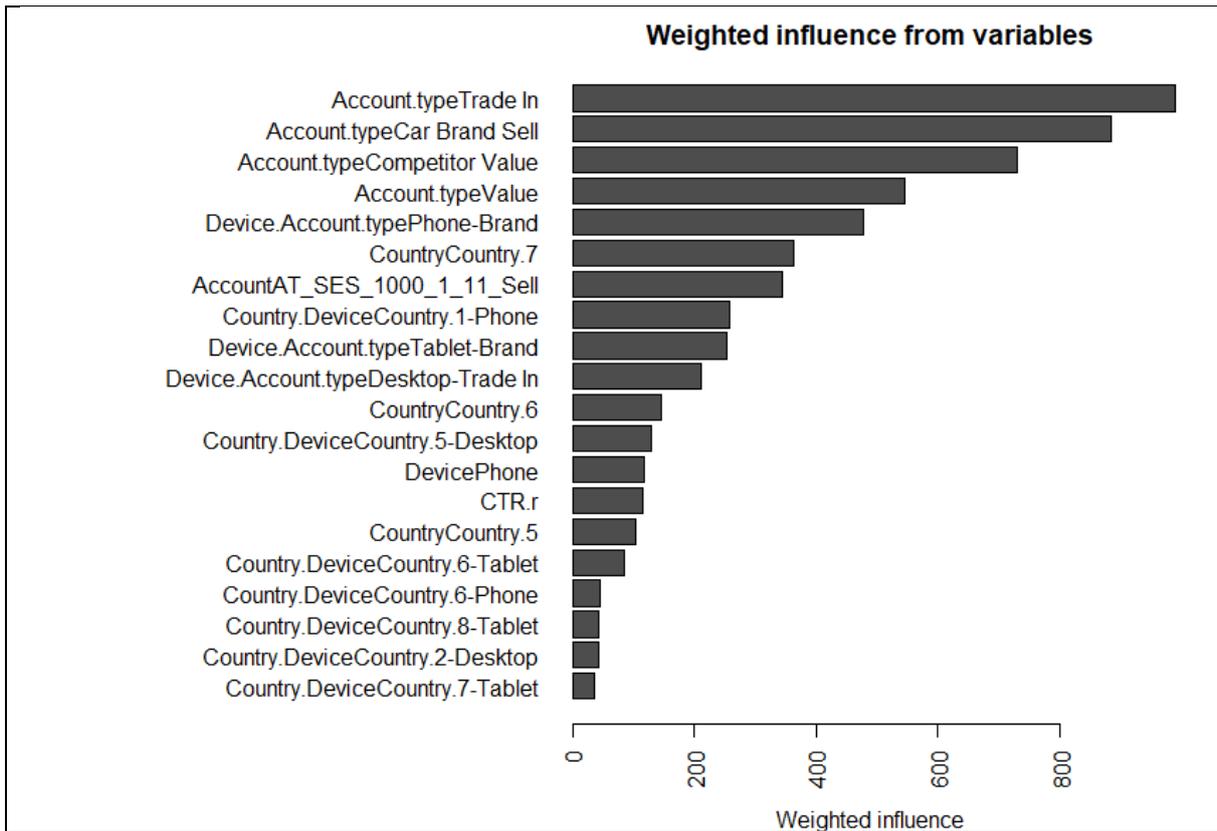
With ad type





12.3 SEM





With ad type

