Marketing Intelligence – Exam

Laurens Doedes Breuning ten Cate – MBD'18 – 01/03/2018

Introduction

This report pertains with the website analytics data and goes quite in-depth to figure out what information is valuable for the insurance company Verti. It tries to answer questions regarding effectiveness for lead attainment of channels, devices, age of target customers and gender of target customers. Also, it tries to identify models that predicts leads based on each of the before mentioned variables. And finally, it gives a better overview of the importance of each of these variables and the highest value combinations with the help of an optimized decision-tree model.

Current measurements

The company Verti currently measures a few variables:

- Channels: Organic Search', '(Other)', 'Direct', 'Paid Search', 'Display', 'Referral', 'Social', 'Email'
- Devices: desktop, mobile & tablet
- Age group of target customer
- Gender of target customer

The dataset is created around combinations of these grouped by the date. From each unique combination of channel, device, age group and gender some aggregate measures of website visits are then collected.

- Users: the number of users that form this specific aggregate
- Sessions: the aggregate amount of sessions each category of customer has visited the website
- Page Views: the number of pages the customer clicks through on the website.
- Bounces: the number of customers that leave (bounce) the website immediately.
- Time on Page: The actual time spent for each customer on the website.
- Leads

The analysis will focus on these metrics and create some more.

Methodology

Extra metrics

Due to varying amount of user adoption of categories it is necessary to use ratios to effectively compare channels, devices etc. This is why we make new metrics to compare these channels. Below in the table these metrics are provided.

Metric	Calculation
Conversion Rate	Leads / Sessions
Bounce Rate	Bounces / Sessions
Pageviews per session	Page views / Sessions
Self-created Metrics	
Time per session	Time on page / Sessions
Uniques	Customer variable combination of Channel + device + gender + age group

Precautions

For each of the importance or effectiveness question to be answered a few key facts need to be kept in mind at all times. For instance, to clearly identify what the most effective device is we can't just look at the average (or mean) over the whole year because perhaps a few outliers are creating some inconsistencies. Due to this I used both means & medians. But no big differences were found.

Unknown variables

One interesting data point is the sub-category of the channel variable '(Other)'. It's quite unclear what this means at a first glance and as we will see in the results it's actually one of the most important channels. A little bit of digging, under the assumption that this data-set is obtained through Google Analytics (as the channel name conventions seem to be the same) turned up a Regular Expression filter for the '(Other)'.

Other Advertising	Medium matches regex ^(cpv|cpa|cpp|content-text)$

Now these four words mean quite interesting things:

- **Cpv**: cost per view, tends to mean video advertising where the viewer actually viewed for > 30 sec.
- **Cpa**: cost per acquisition, a metric used in content marketing so likely related to content-text
- **Cpp**: cost per rating point, a metric used to measure cost necessary to reach a certain % of your audience, again often used for content marketing.
- **Content-text**: most often refers to content marketing which are for instance blogs or articles.

It seems that '(other)' would squarely fall in a paid marketing category which is extremely interesting and should be kept in mind.

Effectiveness

To measure the effectiveness cross-tables were used to analyze, per 'first stage' variable, the various metrics that are collected. All of these should be somewhat take with a small grain of salt as some differences might seem big but are in reality not as different when applying a more rigorous similarity test like a student t-test. All 'second stage' metrics *and* all the created extra metrics will be used to both have a scaled and non-scaled look at the data. Because, sometimes a high ratio is seen in conversion rate but the actual size of the group is very small. Thus, both are necessary.

Modelling

First model to be used is a standard linear regression model. Various combinations of variables have been implemented. First a standard fit with just the 'first stage' measurements (channel, device, age, gender) plus the Date variable due to the fact that the data is grouped by it. However, since the Data variable does not predict leads it won't have any coefficient impact. The reasoning for just using the 'first stage' variables is because the 'second stage' variables already have an almost direct relationship with leads. As of course, someone who stays on the page for a long time is more likely to become a lead and thus pageviews or sessions is more of a result variable itself than a predictor.
A second model is run with the variable 'unique' also included. This allows me to basically take each and every combination of channel, age, device and gender possible to see which specific combination most often results in a lead.

Finally, a decision tree model was ran to create an easier to interpret model of what variables decide most strongly whether a target customer will convert or not. This will in a sense approach the second model in efficacy but will give us, beside relationship strength, also amounts of people that belong to each of these categories.

Results

To keep this section more organized all tables are in the annex that will be used for the discussion. For clarity of results all notes regarding each metric are put in a table. It should also be noted that in my analysis the metric 'Session Pageviews' and 'Session time' have been discounted because in a separate linear model (annex 3.3 – linear model 3) the prediction value for these metrics is extremely close to zero. Due to this I left them out of my overviews.

Most effective Channels

Metric	Best channel	Details
Conversion rate	Email, but take this with a grain of salt since if we look at counts (last column) we see only 82 email occurrences. A more interesting high-conversion channel is Paid-Search.	(other) and Social also have relatively high impacts.
Bounce rate	The best bounce rates, perhaps unsurprisingly, come from Direct, Organic Search and Referral. All are unpaid channels. The person actually wanted to go to the website on their own accord.	All the paid channels except Social seem to do very bad Bounce rate wise.

The analysis finds that for 'free' channels dominate but the relevant channels to look at are paid since those are directly influenced by our marketing efforts. Thus, the most effective channel goes to paid-Search conversion wise with Social and (other) being second due to relatively high conversion rates.

Most effective devices

Metric	Best Device	Details
Conversion rate	Conversion rate wise Desktop is the strongest but with mobile following suit. The graph below our table shows also that over time the importance of mobile seems to be increasing though more data is necessary to confirm that.	Counts are very similar for each device though tablet is falling somewhat short.
Bounce rate	Clearly bounce-rate is lower for desktops which makes sense as it's easier to stay on a website longer on a full computer.	Interestingly enough tablet bounce rate is slightly lower than for mobile.

It's clear from the analysis that the desktop still is the most important platform by far. However, looking at the date-based graph we can see that mobile is catching up, though I'd need more data to confirm this suspicion. Tablet is also interesting due to its better performance bounce rate. Though it's only better in engagement (session time/pageviews) which aren't very relevant.

Most effective ages

Clearly, over all categories, the age-group 25 to 34 years old is the most important with 35 to 44 being a close second. This distribution is constant over time too as we can see in the graph. In-depth table of analysis in Annex.

Most effective gender

Females are a slightly better performing group although the difference is not massive. Looking at the graph the difference also seems constant. In-depth table of analysis in Annex 2.4.

Lead pipeline modelling

Linear model 1

To optimize the way leads are managed and found it's necessary to get a clear picture of what kind of customers are most likely to become leads. In the previous questions we looked at individual variable importance but to combine this is a lot harder. One way to do this is through using linear regressions. In this case the variable 'Leads' was used as the target variable and Age, channel, device and gender were used as predictor variables. To the left is a table of the coefficient important of each variable.

As we can see customers belonging to the age group of 24-34 are the strongest predictors for becoming a lead. After that, we see that the Organic search and Paid search channels are important. The Desktop as a device and the age group 35-44. Another interesting point is the (other) channel that pops up high, this makes sense as it is also paid advertising like paid search. To really figure out, beside what the strength of each individual variable is we need to look at combinations of terms.

Linear model 2

To really figure out what combination of variables has the most effect on lead production the 'uniques' variable was used. In the table to the left only the top and bottom 5 combinations are displayed, see annex 3.2 for a bigger table.

Immediately these results are extremely helpful for our team at the call center as this table displays the combinations of target customer with the highest propensity to turn into a lead. For instance, we see that a female, between 25-34-year-old, using a desktop pc and found the website through organic search has the highest propensity to become a lead.

Tree model

See Annex 3.4 for the visualization. First is a tree cut off at 3 depth, second is optimal at 9 depth. Decision trees should be interpreted as a question and answers method by going from top-to-bottom and each node asking the question that is written in the top. If the answer is 'Yes', you move to a lower-left node, if the answer is 'No', to a lower right node. Finally, when you reach the bottom we see that the left-hand side turns into a lead and right-hand side doesn't.

It confirms a few things we already saw in our linear models. However, the interesting thing about trees is that we can find a combination of factors with the highest amount of predicted leads. In this case, if we look at the optimal tree with depth of 9 (see annex 3.4.2), the path to the highest bottom node (leaf) is revealing. Age_25-34 > Organic_search_channel > Not-Tablet > Gender_female > Desktop_device. This path leads to the highest predicted Leads of 26.841 out of 396 occurrences.

Conclusion

Key Observations

Verti should in the future refrain from making business decisions on metrics that measure pure engagement (Pageviews per session and Time per session) as they are negligible in their impact on lead attainment. Next to that I believe all good recommendations have underlying actionability built into them so an analysis of the Organic Search channel being very valuable is quite irrelevant as that can't be boosted by our current marketing methods directly. Finally, one key piece of information that is missing and might be quite important is what the '(other)' channel truly entails as it is not the most important but definitely up there and no action can be taken on that channel if it is categorized as such.

Key Recommendations

Based on this dataset I've come to a two-fold tactic for increasing lead generation through the website. First tactic is a broader marketing focus on the highest performing groups and the second tactic is to increase specific attention to the top 5 best performing 'paths'. Now this specific attention can take the forms of any of the four P's (Product, Price, Promotion, Place – Product and Promotion being most effective in this scenario) though because they are already high-performing I'd recommend running some tests to see if these segments are perhaps at maximum saturation (or close to), in which case Verti is better of targeting lower-performing groups with new Products or specific Promotions to avoid having diminishing returns on these best-performing segments. The groups to target are below. The first group should be tested but if metrics do not improve than most likely these groups are already saturated and we are running into diminishing returns. In this case I recommend some other groups that might not be saturated yet.

If Unsaturated	If Saturated
Age group: **25 to 34**	Channel: **(other)**
Channel**: Paid Search**	Gender: **Female**
Device**: Desktop**	Device: **Mobile**
Age group: **35 to 44**	-

The path specific recommendations are below, again, only actionable recommendations have been chosen from the models.

Profile 1	Profile 2	Profile 3	Profile 4	Profile 5
Age**: 25-34**	Age**: 25-34**	Age**: 25-34**	Age**: 35-44**	Age**: 35-44**
Gender: **Female**	Gender**: Female**	Gender**: Female**	Gender**: Female**	Gender**: Female**
Device**: Desktop**	Device**: Mobile**	Device**: Mobile**	Device**: Mobile**	Device**: Mobile**
Channel**: Paid Search**	Channel**: Paid Search**	Channel**: (Other)**	Channel**: (Other)**	Channel**: Paid Search**

Annex

Annex 1: Variable overview

	Date	Channel	Device Category	Age	Gender	Users	Sessions	Pageviews	Bounces	Time on Page	Leads	Datetime	ConversionRate	BounceRate	PageViewsSession	TimeSession	uniques
22544	20170101	Organic Search	desktop	18-24	male	82	117	1874	19	46099	4	2017-01-01	0.034188	0.162393	16.017094	394.008547	Organic Searchdesktop18-24male
31396	20170101	(Other)	mobile	45-54	male	45	49	50	46	775	2	2017-01-01	0.040816	0.938776	1.020408	15.816327	(Other)mobile45-54male
36966	20170101	(Other)	desktop	55-64	female	32	36	124	27	1130	0	2017-01-01	0.000000	0.750000	3.444444	31.388889	(Other)desktop55-64female
52307	20170101	Referral	desktop	25-34	female	13	19	346	4	8385	0	2017-01-01	0.000000	0.210526	18.210526	441.315789	Referraldesktop25-34female
52306	20170101	Organic Search	tablet	45-54	male	13	16	79	3	3099	1	2017-01-01	0.062500	0.187500	4.937500	193.687500	Organic Searchtablet45-54male

Annex 2.1.1: Channel importance - Table

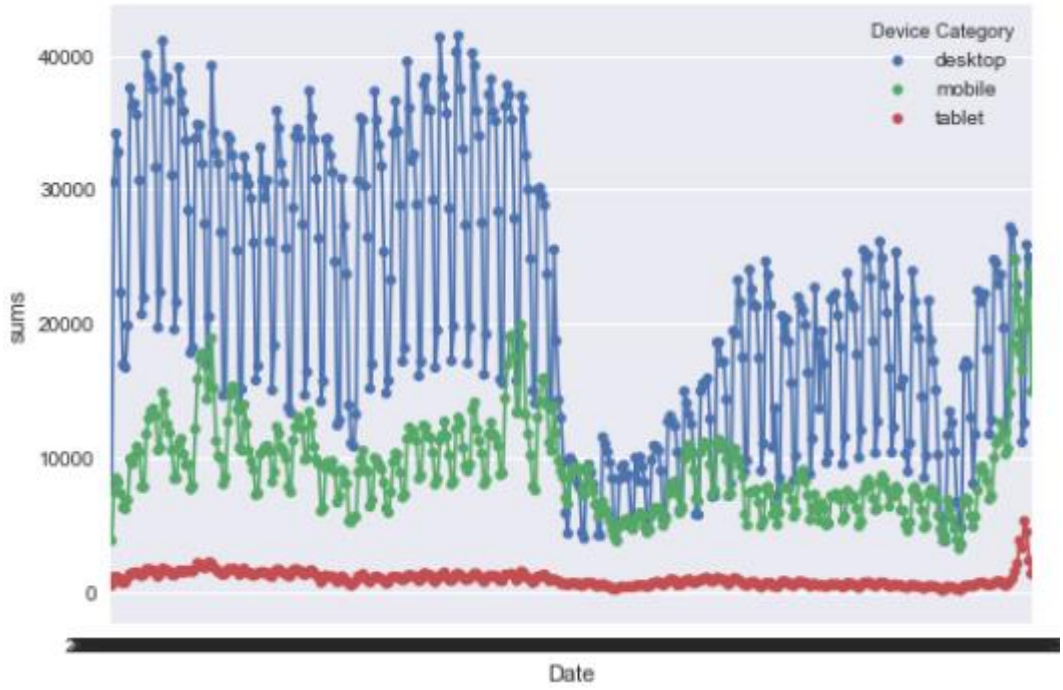
Channel	Leads	ConversionRate	Bounces	BounceRate	Pageviews	PageViewsSession	Time on Page	TimeSession	counts
(Other)	3.513751	0.025046	74.095422	0.699454	399.971323	2.993331	15247.798004	110.702257	11926
Direct	1.206436	0.005274	27.780727	0.122721	8415.550004	16.301601	209194.921502	366.033261	11529
Display	1.235487	0.017870	43.008786	0.551557	901.352455	9.930518	23791.229113	255.905342	5805
Email	4.451220	0.123173	20.256098	0.489211	79.024390	2.148591	7577.243902	209.029270	82
Organic Search	5.495956	0.018656	71.186359	0.275434	5471.744101	11.819871	151682.954557	325.441530	11993
Paid Search	4.546556	0.035211	77.750456	0.687730	310.887208	2.341727	11978.657799	93.907369	10976
Referral	0.827708	0.017992	12.154156	0.277484	783.746096	12.665579	24199.012846	398.233497	3970
Social	0.664846	0.027779	12.300894	0.440183	172.814796	6.453296	6082.126614	228.834649	2014

Annex 2.1.2: Channel importance – over time (as session sums)

Annex 2.2.1: Device importance - table

Device Category	Leads	ConversionRate	Bounces	BounceRate	Pageviews	PageViewsSession	Time on Page	TimeSession	counts
desktop	4.068401	0.022559	63.339124	0.387898	5574.771808	11.798924	151126.919730	342.046281	26710
mobile	3.413122	0.019763	68.910490	0.497216	1330.516894	5.420076	33230.829096	139.460916	20333
tablet	0.508887	0.018356	12.662727	0.477048	317.335229	7.718147	7023.759332	181.905088	11252

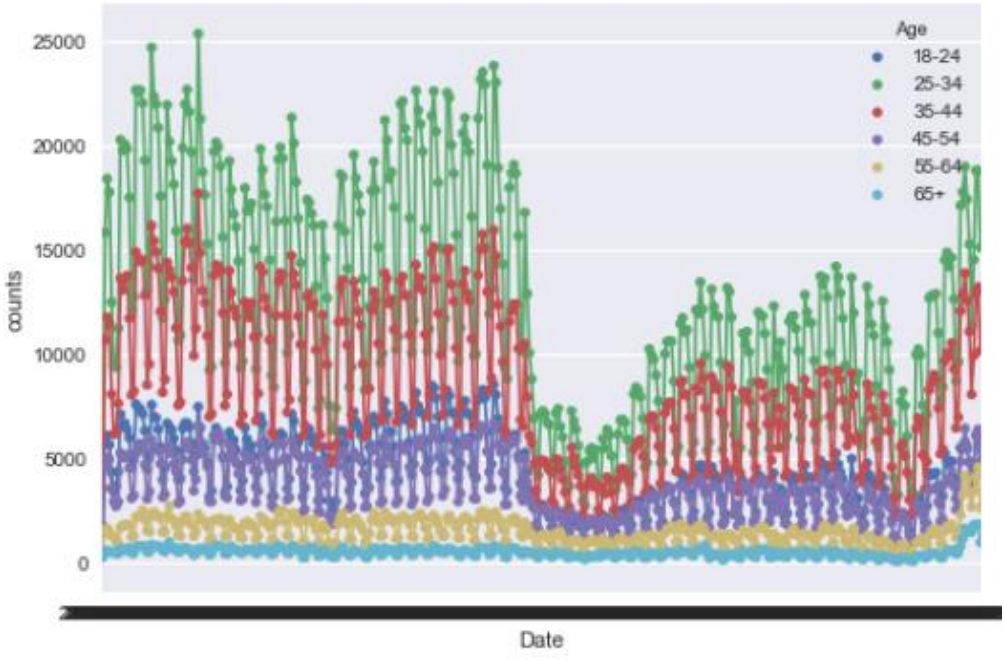
Annex 2.2.2: Device importance – graph



Annex 2.3.1: Age importance

Age	Leads	ConversionRate	Bounces	BounceRate	Pageviews	PageViewsSession	Time on Page	TimeSession	counts
18-24	2.659051	0.020579	56.404593	0.458613	3048.377666	8.912004	82344.751136	247.577006	8579
25-34	6.348255	0.025121	91.984619	0.393124	6067.883105	9.468047	164279.312475	266.205808	12353
35-44	4.179370	0.023564	67.443667	0.399307	4084.121858	9.464081	106743.194338	255.897410	12293
45-54	1.902572	0.019501	40.657383	0.437751	1818.355292	8.845033	47752.295892	236.898358	10808
55-64	0.928396	0.016683	28.070042	0.504846	711.634508	7.404149	18980.324448	201.971206	8966
65+	0.433535	0.013979	17.949207	0.544063	411.293429	7.639871	10773.669940	205.643765	5296

Annex 2.3.2: Age importance - Graph



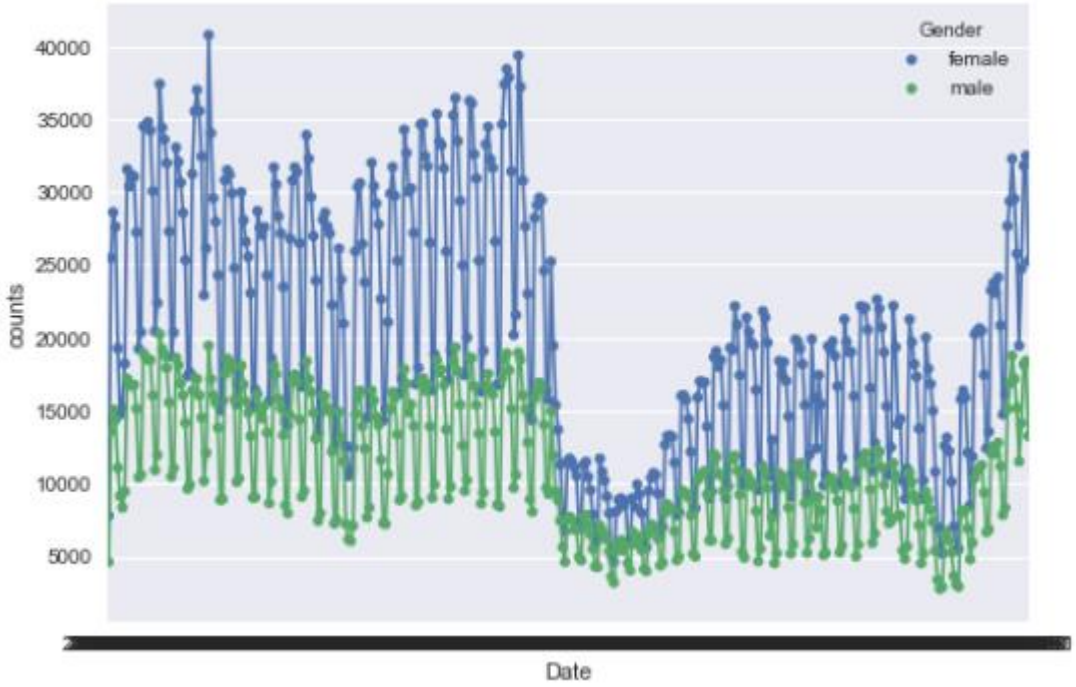
Annex 2.3.3: Age importance - analysis table

Metric	Most important age group	Details
Conversion rate	Conversion wise the age-group 25 to 34 is the most valuable with 35-44 a close second.	Count wise this is also the biggest group.
Bounce rate	Again, 25-34 with 35-44 as a close second.	-
Session pageviews	Similar results with 25-34 being the most important.	-
Session time	We see a similar distribution again.	-

Annex 2.4.1: Gender importance - Table

Gender	Leads	ConversionRate	Bounces	BounceRate	Pageviews	PageViewsSession	Time on Page	TimeSession	counts
female	3.804242	0.021408	62.721729	0.428565	3999.568264	9.421353	105372.466645	254.140114	30880
male	2.419004	0.020057	47.367463	0.459760	2043.390224	8.071089	56079.228306	225.083506	27415

Annex 2.4.2: Gender importance - Graph



Annex 2.4.3: Gender importance – Analysis table

Metric	Most important gender	Details
Conversion rate	The genders are nearly tied but females seem to convert slightly more.	The distribution is quite gender equal with very similar counts.
Bounce rate	Again, females have slightly better bounce rates.	-
Session pageviews	Obviously, these metrics are somewhat correlated and also here females have slightly more page views per session.	-
Session time	Again, slightly more time per session for females.	-

Annex 3.1: Linear Model 1 – Coefficient Table

	Coeffs
Age_25-34	4.377648
Channel_Organic Search	3.966557
Channel_Paid Search	2.686305
Device Category_desktop	2.410459
Age_35-44	2.212799
Channel_(Other)	2.192900
Gender_female	0.882428
Device Category_mobile	0.800717
Channel_Email	0.756524
Date	0.000067
Channel_Direct	-0.502664
Age_45-54	-0.514373
Age_18-24	-0.672876
Gender_male	-0.882428
Channel_Display	-1.462472
Age_55-64	-2.105510
Device Category_tablet	-3.211176
Age_65+	-3.297688
Channel_Referral	-3.520508
Channel_Social	-4.116642

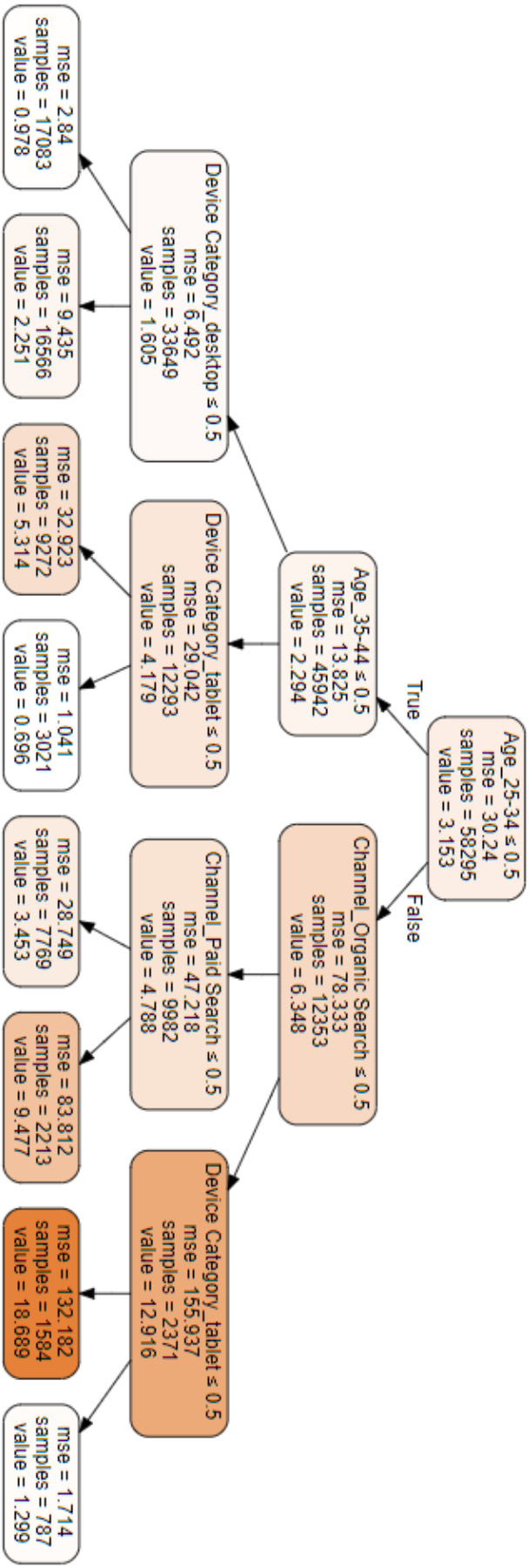
Annex 3.2: Linear Model 2 – Coefficient Table

	Coeffs
uniques_Organic Searchdesktop25-34female	17.020119
uniques_Organic Searchmobile25-34female	14.533369
uniques_Paid Searchdesktop25-34female	10.831501
uniques_Paid Searchmobile25-34female	8.077074
uniques_Organic Searchdesktop25-34male	8.007025
uniques_(Other)mobile25-34female	6.867760
uniques_(Other)mobile35-44female	5.979810
uniques_Organic Searchmobile35-44female	5.428247
uniques_Organic Searchdesktop35-44female	5.356917
uniques_Paid Searchmobile35-44female	4.805286
uniques_Organic Searchdesktop18-24female	4.409091
uniques_Organic Searchdesktop35-44male	4.293317
uniques_Displaytablet65+male	4.043886
uniques_Socialtablet55-64male	3.942170
uniques_Paid Searchdesktop18-24female	3.869463
...	...
uniques_Organic Searchmobile65+female	-3.077187
uniques_Displaymobile25-34male	-3.106384
uniques_(Other)tablet25-34male	-3.165617
uniques_Organic Searchtablet35-44female	-3.213325
uniques_Paid Searchtablet25-34male	-3.264460
uniques_Socialdesktop25-34male	-3.315307
uniques_Organic Searchmobile55-64female	-3.369059
uniques_Organic Searchdesktop55-64female	-3.518672
uniques_Paid Searchtablet25-34female	-3.828144
uniques_Organic Searchdesktop65+female	-3.982891
uniques_Displaymobile25-34female	-4.142149
uniques_(Other)tablet25-34female	-4.151196
uniques_Socialdesktop25-34female	-4.260574
uniques_Organic Searchtablet25-34male	-4.264379
uniques_Organic Searchtablet25-34female	-4.467699

Annex 3.3: Linear model 3 – Coefficients

	Model 3 - Coeffs
Channel_Organic Search	4.520498
Age_25-34	4.512561
Device Category_desktop	2.635970
Age_35-44	2.365043
Channel_Paid Search	2.072144
BounceRate	1.992295
Channel_(Other)	1.608034
Gender_female	0.950103
Device Category_mobile	0.665598
Channel_Direct	0.632724
Channel_Email	0.221540
TimeSession	0.001803
Date	0.000069
PageViewsSession	-0.079516
Age_45-54	-0.474789
Age_18-24	-0.725208
Gender_male	-0.950103
Channel_Display	-1.594160
Age_55-64	-2.221961
Channel_Referral	-3.224987
Device Category_tablet	-3.301567
Age_65+	-3.455647
Channel_Social	-4.235793

Annex 3.4.1: Tree – depth = 3



Annex 3.4.2: Tree – Zoomed-in - depth = 9