PREDICTING THE CONNECTION BETWEEN MOBILE DEVICES AND COOKIES

RESEARCH PAPER BUSINESS ANALYTICS

Author: Nap, Hanna
Supervisor: Eiben, Guszti

FACULTY OF SCIENCES, VU UNIVERSITY De Boelelaan 1081, 1081 HV Amsterdam

3-10-2016

Preface

This research is part of the program of the master Business Analytics. This paper is written to share background information about the current topic *cross-device tracking*, to discuss my research and to make the reader enthusiastic and aware of the possibilities in the future regarding the connection of mobile devices.

This research focusses on the development of a model that predicts which mobile devices belong to the same person. In this way marketers gain a better understanding of the behavior of individuals and are able to show and send more effective personalized advertisements and emails.

I would like to thank my supervisor Guszti Eiben for his enthusiasm, his knowledge and his tips concerning the literature study and defining my research goal.

Summary

Connecting mobile devices for online marketing purposes, better known as *cross-device tracking*, becomes increasingly important, due to the growth of mobile devices and data traffic. Where nowadays marketers focus on cookie-based targeting, a shift is needed towards people-based targeting.

A firm that is prepared for this change, is Drawbridge. Drawbridge provided a platform for marketers where cross-device tracking is the cornerstone and where a predictive model, that predicts which mobile devices belong to the same person, is the basis. However, at the start of this research, June 2015, the known predictive models for cross-device tracking had a relatively low precision, from 60% up to 90%, Drawbridge asked for help by setting up a Kaggle competition, where the firm provided anonymized data (concerning mobile devices and cookies). Therefore, this research focuses on the development of a model that predicts which mobile devices and cookies belong to which person, by using the dataset provided by Drawbridge. In other words, the research goal is the following: building a predictive model with an accuracy of 90% or more, that predicts which mobile devices and cookies belong to the same individual.

The data that is obtained by Drawbridge consists of six tables and contains information about the users, devices, cookies and associated IP addresses and properties. On the one hand, Drawbridge provided data concerning mobile devices and on the other hand the firm provided data concerning cookies stored on a desktop browser, like on a computer for example. In the relational database this information is linked by a person ID. The goal of this research is to predict this link by the use of a classification decision tree.

The approach to build a predictive model for this linkage problem consists of the following steps: *step 1.* exploratory data analysis, *step 2.* transformation of the tables, *step 3.* realization of the data set and *step 4.* realization of a predictive model.

The exploratory data analysis resulted in multiple findings, including the fact that more than 90% of the devices and cookies belong to country_146 and the fact that on average each device is traced on 14 IP addresses and each cookie on 4 IP addresses. Based on these findings, the tables were transformed and joined in such a way that the data set consists of one instance for each device-cookie combination and that a training set was realized were 5,000 records contain a match between device and cookie and 5,000 records a non-match. Finally, three columns were added that contain additional information: IP_match, property_match and country_match. The columns inform about matches between IP addresses, properties or countries of the device and cookie.

After importing this data set in SPSS Modeler, the feature selection option was run, resulting in several columns, containing IDs and IP addresses, that were indicated as non-influencing factors concerning the dependent variable. These were removed as inputs for the three decision tree algorithms: the C&RT, C5.0 and CHAID algorithm, which were run afterwards by using ten-fold cross validation. Measuring the accuracy, sensitivity and specificity showed that the CHAID algorithm has the highest average accuracy and sensitivity. Besides, because of the low variation of the measure values, the algorithm seemed stable. Therefore, the final model was built based on the CHAID algorithm.

This resulted in a classification decision tree with an accuracy of 0.9160, a sensitivity of 0.8526 and a specificity of 0.9794. Looking at the accuracy value, the research goal is achieved.

Table of Contents

Preface 1

Summary 2

1. Introduction 4

2. Background 6

    2.1. Online marketing 6

        2.1.1. Cookie-based targeting 7

        2.1.2. From device-based marketing to people-based marketing 8

3. Data 10

4. Methods 13

    4.1. Exploratory data analysis 13

    4.2. Transformation of the tables 13

    4.3. Realization of the data set 14

    4.4. Realization of a predictive model 14

        4.4.1. Feature selection 15

        4.4.2. Building classification decision trees 16

        4.4.3. Determining the best algorithm and building the final model 18

5. Results 19

    5.1. Exploratory data analysis 19

    5.1. Transformation of the tables 23

    5.2. Realization of the data set 23

    5.3. Realization of a predictive model 24

        5.3.1. Feature selection 25

        5.3.2. Building classification decision trees 26

Conclusion and recommendations 30

References 31

Appendices 33

    Appendix I: Results from researches performed by Cisco and KPMG 33

    Appendix II: The data tables 35

    Appendix III: The final decision tree 39

1. Introduction

Nowadays, it is hard to imagine a world without mobile devices. Devices, such as tablets, mobile phones and laptops, are important tools when it comes to communication and obtaining information. Besides, these devices are often used for entertainment purposes and online shopping. They are simply incorporated in the society.

On average each individual owns 4 mobile devices, which will keep growing in the future. According to a research performed by Cisco[1], the number of mobile devices globally was around 7.9 billion in 2015, which will increase to 11.6 billion in 2020. Furthermore, Cisco predicts that the data traffic will increase as well. Where the data traffic was equal to 3.7 exabytes per month in 2015, it will probably become 30.6 exabytes by 2020, which is almost 8 times as big.[2]

The awareness of using these data to improve processes and seize business opportunities is present. Based on data, companies improve the functionalities of an application or website, marketers focus on personalized marketing and providers like Ziggo change their business plan. However, this will be taken to the next level. According to the concept of the *Internet of Things*, not only humans will communicate in the future, but also devices to humans and devices to devices. These devices are called 'smart' devices. One may think of smart cars, and smart home and smart city devices.

Although the number of smart devices is growing fast, Internet of Things is still in its infancy. There are a lot of growth opportunities in different industries. The 'Global technology innovation survey of 2015' of the firm KPMG points out that the industry with the biggest growth opportunities is the retail industry. According to this survey[3], when adopting the Internet of Things the sales of the global retail industry will increase with 22% within three years.[4]

Looking at this industry, there is certainly room for improvement. Nowadays, information about a customer's behavior is lacking, due to the fact that we are not able to share information. For example, suppose a person is traveling home by bus and is looking at a certain product at a certain site by using his iPhone, but decides not to buy the product yet. When arrived at home, he looks at the same product at the same site, but now using his laptop. He decides he first wants to see the product before buying it, so the next day he goes to the store and buys the product.

According to the data obtained, the activities performed on the iPhone of customer X and the laptop of customer Y did not result in a purchase. Only the activities of customer Z, registered by the customer's card, leaded to an acquisition. However, it is unknown that person X, Y and Z are the same person. This results in a lack of information about the customer and its customer journey. Although, when these devices would be connected, so when adopting the Internet of Things, more information of a customer would be shared. In this way a better picture of the customer can be formed and marketers are better able to respond to the customer's behavior, his needs and demands.

---

[1] Source: (Cisco, 2016)
[2] See appendix I for the findings that resulted from the research performed by Cisco.
[3] Source: (KPMG, 2016)
[4] See appendix I for the findings that resulted from the research performed by KPMG.

Connecting mobile devices for online marketing purposes is called *cross-device tracking*. Cross-device tracking becomes increasingly important, due to the growth of mobile devices and data traffic. A firm that focusses on cross-device tracking is Drawbridge[5]. Drawbridge is a leading company when it comes to building tools for marketers where cross-device tracking is the cornerstone. This firm built a platform where marketers can get an overview of their customers and can be helped with the execution of media campaigns across devices. The basis of the platform is a predictive model that predicts which mobile devices belong to the same person.

At the start of this research, June 2015, the known predictive models for cross-device tracking had a relatively low precision, from 60% up to 90% [6]. Because of that, Drawbridge asked for help by setting up a Kaggle competition. Drawbridge provided anonymized data (concerning mobile devices and cookies) and offered a reward for the best predictive model[7]. Therefore, this research focuses on the development of a model that predicts which mobile devices and cookies belong to which person, by using the dataset provided by Drawbridge. In other words, the research goal is the following: building a predictive model with an accuracy of 90% or more, that predicts which mobile devices and cookies belong to the same individual.

This research paper discusses some background information, including the current marketing approach and cross-device tracking approaches, followed by an overview of the data, the methods used and the resulting model. The paper ends with a conclusion and discussion.

[5] Sources: (Drawbridge, About us - We're Perfecting the Art of Connecting Brands with Consumers, 2015) (Drawbridge, Solutions - We're Powering a More Personalized Internet for Everyone, 2015)
[6] Source: (Tradedoubler, 2016)
[7] Currently, the predictive model of Kaggle has a precision of 97.3%.

2. Background

The rise of the internet had a major impact on the retail industry. Together with the growth of mobile devices, it leaded to the emergence and increase of web shops and online shopping. This also had an effect on marketing. Marketers were able to obtain more and more data and could use these data to gain a better understanding of their clients, whereupon they were able to react better to their clients. Because of that, it makes sense that over time the focus has been shifted from offline marketing to online marketing.

2.1. Online marketing

Where there is no internet involved in offline marketing (one may think of advertisement in magazines, billboards and brand images on plastic bags), this is exactly the case when it comes to online marketing. The four types of online marketing that are mostly used, are the following [8]:

- E-mail marketing;
- Online advertising;
- Marketing through social media;
- Search engine marketing, where websites are shown that match a search term the best.

One remark should be made. Two definitions are often mixed up, namely: marketing and advertising. However, advertising is a component of marketing, so advertising is marketing, but marketing is not always advertising. Besides, in literature different terms with the same meaning appear:

e-/internet/online/digital marketing and *e-/internet/online/digital* advertising. In this paper the terms online marketing and online advertising will be used. Online advertising is a form of online marketing that is mostly used to reach (potential) customers.

As said, due to the growth of available data, marketers are able to gain a better understanding of (potential) customers and react on their customers by sending relevant e-mails and showing relevant advertisements, for example. In other words, the data is being used for targeting. Targeting is an important aspect when it comes to CRM (Customer Relationship Management), because knowing your clients, and knowing their demands and wishes, will lead to a better relationship with them.

In online marketing there are two types of targeting: semantic targeting and behavioral targeting.[9] Semantic targeting, also known as semantic advertising, uses semantic techniques to determine the context of a website in order to place suitable advertisements. Behavioral targeting however looks at the behavior of an individual to place advertisements or send emails for example. This is the type of targeting that is applied the most. The fact is that several studies have shown that individualized advertisements significantly increase the effectiveness of online advertising.[10] To determine the behavior of an individual, data has to be obtained and analyzed. These data can be obtained from the database, for example for information about purchases, but most often it involves cookies.[11] Therefore, behavioral targeting is also known as cookie-based targeting.

[8] Source: (eMarketing: The Essential Guide to Online Marketing)
[9] Source: (Schmücker, 2011)
[10] Source: (Goldfarb & Tucker, 2011)
[11] Source: (Groep, 2012)

2.1.1. Cookie-based targeting

Cookies are small text files that are stored on the device when a website is being visited, based on a request of the server behind the website. These text files contain the domain, the IP address, the duration of storage of the cookie, and information about the visit. Besides, often a session ID is added. In this way the server of a website can link requests (buying a certain book for example) to the corresponding web browser, without using the IP address. Also the next time when visiting the website, the stored cookie on the device will be send to the server in order to recognize the client and see the history of the client, based on the old cookies. (This doesn't apply to session cookies, which are removed when the visitor leaves the website. That is, these cookies are only needed to let the site function properly.) However, due to a security policy , it is not allowed to send cookies to other domains.[12]

2.1.1.1. First party cookies vs third party cookies

There are different types of cookies, with different purposes, namely: first party cookies and third party cookies. First party cookies are cookies that are set for the domain the client is viewing. These cookies are mostly functional cookies, in the way that they improve the functionality of the website. These cookies are used to remember a username and password, or to remember the selected items in the shopping cart for example. However, third party cookies are set for external domains, the "third party". If different domains accept to set cookies for this external domain, like an advertiser, this advertiser can track the behavior of a client over multiple websites. Therefore, third party cookies are better known as tracking cookies.[13] These cookies are used to track customers. In other words, tracking cookies are used for behavioral targeting, a.k.a. cookie-based targeting.

2.1.1.2. Setting cookies

The way cookies are set, depends on three aspects, namely:

- The use of a mobile application or a mobile web browser;
- The type of cookie;
- The type of browser (if a browser is being used).

When using a laptop, it is common to use one of the mobile web browsers, but when using a mobile phone, there are more options. Besides using a web browser, one may also choose for one of the applications to view the content of a website. All these web browsers and apps have a different way in storing cookies.

When a site is being viewed through a web browser, the cookies are set on the device by this browser. It depends on its settings whether the storage of a cookie will be accepted. Generally all first-party cookies are stored, but the browser can limit the storage of third-party cookies. It is possible that only view-based or click-based conversions are set. (View-based conversions are sales after seeing an advertisement and click-based conversions are sales that are attained directly via clicks.)

When a site is being viewed through an app, the cookie is stored in the "webview", in the same way it is stored in web browser. Webview is the technology of an app that is being used to show the content of a website for example. However, applications are standalone products with unique webviews. Therefore, it is not possible to share their information of cookies. The same holds for web browers. When a site is

[12] Source: (Schmücker, 2011)
[13] Source: (Mayer & Mitchell)

visited twice, first through browser A and second through browser B, the history of the viewer can't be seen. In other words, the use of cookies to track the behavior of people is limited.[14]

2.1.1.3. The cookie law

As said, by using tracking cookies advertisers are able to place individualized advertisements, depending on the behavior of the client. However, the use of tracking cookies caused a discussion about privacy issues. The fact is that tracking cookies contain personal information and visitors of sites were, in most cases, not able to decline the setting of tracking cookies on their device. And when they were able to reject it, the website blocked some functionalities.[15] On the one hand the visitors of websites wanted to keep their privacy and on the other hand the advertisers wanted to obtain as much information as possible, so they could show individualized advertisements that attract the attention of the viewer, so that are effective. This discussion resulted in an European directive, that was adopted by the European countries in the month May of 2011 in the form of a law. [16] This law is known as the 'cookie law' and differs per country. However, it should include the following: all websites should inform visitors about the meaning and use of cookies and give the visitors the opportunity to reject non-functional, so tracking, cookies (opt-out) or give permission to set them on the browser of the device (opt-in).[17] The Unites States however do not have a national cookie law.

This cookie law and the possibility of deleting cookies from your browser, resulted in a decrease of information available for advertisers. Consequently, less effective advertisements were shown.[18] This got even worse with the growth of mobile devices. That is, when a client owns multiple devices and uses these devices to visit websites, the advertiser does not know that these devices belong to the same person. It is not known that the behavior of the owner of device A and the behavior of the owner of device B are the behavior of the same person. This results in more loss of information and so less effective advertisements.

In short, only using cookies to obtain information about the behavior of clients is not enough anymore. To decrease the lack of information, it's important to connect devices by using more information and shift from device-based marketing to people-based marketing.

2.1.2. From device-based marketing to people-based marketing

People-based marketing focusses on tracking the behavior of people across devices and other sources, like a database. Determining which devices belong to a specific person is better known as cross-device tracking. To be able to connect devices, cookies information is not enough. The needed information to connect devices depends on the approach: the deterministic approach or the probabilistic approach.

2.1.2.1. Deterministic approach

The deterministic approach, also known as deterministic tracking, is based on user data when logged in on a system. That is, by logging in, it is known which person logged in, using which device. When a user log in with different devices, the data of these devices can be connected by using unique identifiers. An

---

[14] Source: (IAB, 2015)
[15] Source: (Jegatheesan)
[16] Source: (Optanon, sd)
[17] Source: (Groep, 2012)
[18] Source: (Schmücker, 2011)

advantage of this approach is that the data is extremely reliable. However, a disadvantage is the fact that is hard to obtain a large amount of data. Besides, the firm that obtained the data, is not allowed to share it, due to privacy reasons.
Well known tools that use deterministic tracking for marketing purposes are Facebook Atlas, Apple IDFA and Google Analytics.[19]

2.1.2.2. Probabilistic approach

The probabilistic approach is less accurate, but has a bigger reach than the deterministic approach. Here an algorithm is used to predict the connection between devices based on anonymized data from these devices. Besides cookies, also information like the IP address, device type and operating system are being used.[20] Using this approach, a user does not have to be logged in.
Well known tools that use probabilistic tracking for marketing purposes are the tool created by Drawbridge, Tapad and Adobe Audience Manager.[21]
As said, the data that is used for this research is provided by Drawbridge. Drawbridge asked for help to create a predictive model by setting up a Kaggle competition. However, due to the fact that probabilistic tracking is a new approach in marketing, no research has been done to determine the best algorithm. Even there is hardly no literature that mentions the predictive models and corresponding algorithms that are used to connect devices. Therefore it is not possible to base the choice for a predictive model on this literature study.

[19] Sources: (Leune, 2016), (Tradedoubler, 2016)
[20] Sources: (IAB, 2015), (Signal), (Leune, 2016)
[21] Sources: (Leune, 2016), (Tradedoubler, 2016)

3. Data

The data that is obtained by Drawbridge contains information about the users, devices, cookies and their behavior. These data consists of six tables[22]:

- A device table, containing high-level information about the devices, such as the device ID, device type and OS version.
- A cookie table, containing high-level information about the cookies, such as the cookie ID, computer OS type and browser version.
- An IP table, containing information about the behavior of a cookie or device on a certain IP address, such as the cookie or device ID, the (anonymous) IP address and the number of appearances on the IP address.
- An IP aggregation table, containing high-level information about an IP address, such as the IP address and the total number of appearances.
- A property observation table, containing information about a website (for cookie) or an application (for device) that a user visited, such as the device or cookie ID, the property ID (i.e. website/app ID) and the number of appearances.
- A property category table, containing the category a website or application belongs to. In other words, it contains the property ID and a category.

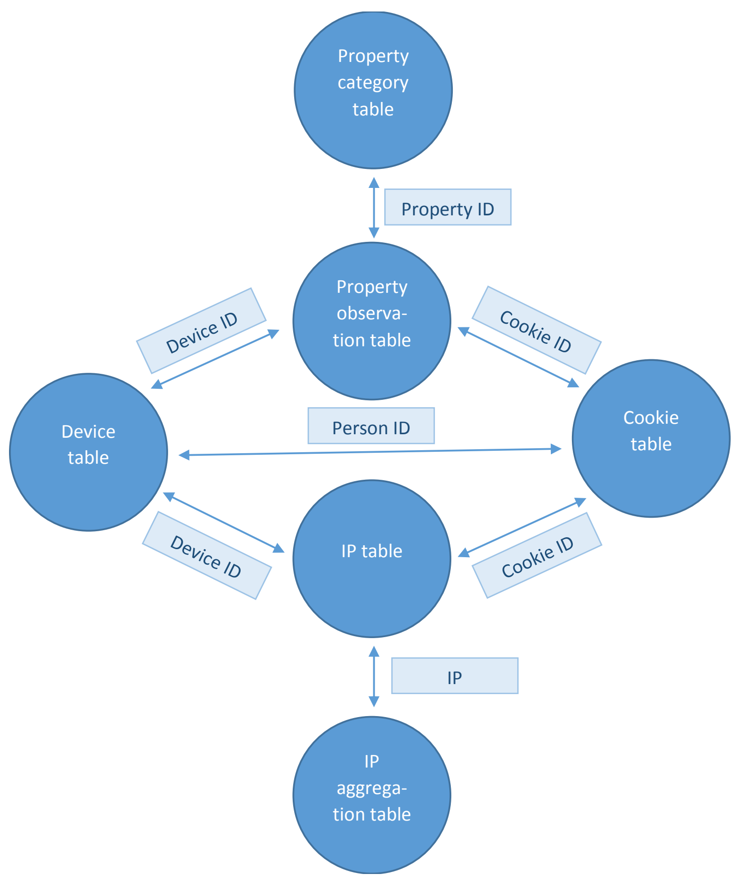

Figure 1. The data tables

The data of these tables can be merged by joining the tables on an ID or anonymized IP address. See figure 1. The objective of probabilistic cross-device tracking is creating a predictive model that connects a user which his/her devices. However, Drawbridge did not provide the necessary data to predict this link directly. On the one hand, Drawbridge provided data concerning mobile devices and on the other hand the firm provided data concerning cookies stored on a desktop browser, like on a computer for example. In the relational database this information is linked by a person ID. However, because the given cookies are cookies stored on a desktop browser, and not on a browser of a mobile device, the cookies don't belong to one of the given mobile devices, but to a computer for example. In other words, when devices and cookies are linked by a person ID, it just means

[22] See appendix II for an overview of the tables and their features.

that they belong to the same person. This case is a fundamental component of cross-device tracking. Based on the information concerning devices and cookies and the behavior of persons on their devices and cookies, the objective of this research is to create a predictive model that connects devices with cookies that belong to the same person. See figure 2. In other words, for each device in the test set the corresponding cookies must be provided. Because the cookies don't belong to a specific device, the resulting list with cookies is the same for all devices of the same person. [23] See table 1 for an example.

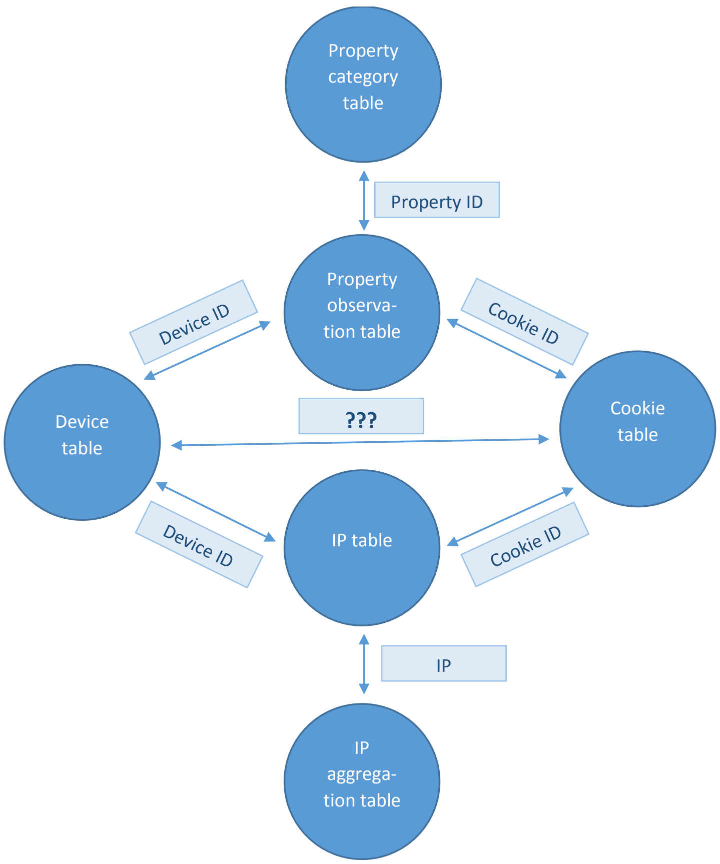

Figure 2. The test set

Person ID	Device ID	Cookie ID's
Person_1	Device_110	Cookie_13
Person_2	Device_29	Cookie_5, Cookie_40
Person_2	Device_54	Cookie_5, Cookie_40

Table 1. The linkage

An approach to realize this list is to build a predictive model that predicts whether a device and cookie belong to the same person, and afterwards build an algorithm that predicts the relation for each

[23] Sources: (Drawbridge, ICDM 2015: Drawbridge Cross-Device Connections - Description, 2015), (Drawbridge, ICDM 2015: Drawbridge Cross-Device Connections - Data files, 2015)

plausible combination of device and cookie. Due to the limit of time, this research focusses on the first part of the approach, namely the creation of a predictive model.

4. Methods

The approach to build a predictive model for this linkage problem consists of the following steps:

1. Exploratory data analysis
2. Transformation of the tables
3. Realization of the data set
4. Realization of a predictive model

The first three steps are performed by using SQL Server and the last step is performed by using SPSS Modeler. SQL Server is a data base management system, where the programming language SQL is used to obtain and transform the data. SPSS Modeler is a software application from the firm IBM that people use to build predictive models.

4.1. Exploratory data analysis

After importing the tables in a SQL database, the first step is performing an exploratory data analysis, consisting of different counts to obtain first insights in the data and to determine the data quality. A number of the analyses are:

- The number of missing values per column
- The number of device IDs and cookie IDs
- The number of devices per person
- The number of cookies per person
- The number of IP addresses per device ID/cookie ID
- The number of properties per device ID/cookie ID
- The number of different device types, device OS and device countries
- The number of different computer OS types, browser versions and cookie countries

The results of this analysis provides insights concerning the tables, but also concerning the many to many relationship between the tables. In other words, the results are a detrimental factor concerning the transformation and realization of the data set.

4.2. Transformation of the tables

The transformation of the tables consists of the following steps:

1. Deleting columns with more than 50 % of missing values.
2. Creating reference tables for nominal and categorical variables. For example, device ID 'device_110' has reference number 10.
3. Replacing the values of nominal and categorical variables by their reference numbers. In this way SPSS Modeler is able to work with these attributes.
4. Transforming the tables in such a way that, when joining the tables, the resulting data set contains one instance for each device-cookie combination. See table 2 for an example. (All columns in this example are nominal or categorical variables, so consist of reference numbers.)

Person ID	Device ID	Device type	Device IP 1	...	Cookie ID	Cookie IP 1	...
1	33	1	100	...	10	111	...
1	33	1	100	...	13	222	...
2	5	2	233	...	15	333	...

Table 2. Example of joined tables

4.3. Realization of the data set

After preparing the tables, one data set is created that will be used as input in SPSS Modeler. This final data set consists of 10,000 instances, where 5,000 instances have a device and cookie that belong to the same person, thus where the binary dependent variable has the value 1, and 5,000 device-cookie combinations where the device and cookie belong to a different person, thus where the dependent variable has the value 0. This rate of 50-50 ensures the model predicts well for both possible outcomes. Besides, the sample size of 10,000 instances is large enough to represent the total dataset obtained.

The following steps are performed to create this final data set:

1. Joining the tables, in order that the relational database transforms into a flat file database. The resulting data set only consists of device-cookie combinations that belong to the same person.
2. Selecting 5000 instances of the data set randomly.
3. Creating 5000 instances where the device and cookie do not belong to the same individual.
4. Combining the two tables into one table and adding a column that will function as (binary) dependent variable. The 5.000 instances with a matching device-cookie combination obtain the value 1 and the instances with non-matching combinations obtain the value 0.
5. Adding columns that contain additional information, based on combining existing columns. For example, a match between the country of the device and the country of the cookie.

4.4. Realization of a predictive model

When the final data set is created, the following steps are performed to build a predictive model in SPSS Modeler:

- Running the feature selection option to identify and remove the attributes that don't influence the prediction of the target.[24] In other words, the features without any correlation with the dependent variable. Besides, it provides insights into the most important attributes. The output, so the remaining attributes, is input for step 2.
- Building classification decision trees based on three different algorithms by using 10-fold cross validation, namely:
    a. The C&RT algorithm
    b. The C5.0 algorithm
    c. The CHAID algorithm
- Determining the best algorithm based on measurements and building the final model based on the total dataset, so based on the 10,000 instances.

[24] Sources: (IBM, Feature selection node, 2012), (IBM, Feature selection options, 2012), (IBM, Feature selection model settings, 2012)

A more detailed explanation of the steps is given below. Furthermore, the reason for choosing these methods and the underlying techniques are discussed.

4.4.1. Feature selection

Feature selection is an option in SPSS Modeler that suggests which attributes to select as inputs, based on their importance. The option consists of three steps:

- Screening
- Ranking
- Selecting

Screening

The attributes are screened based on one or more of the following criteria (depending on the selection of the user):

- The percentage of missing values is above the maximum
- The percentage of records in one single category is above the maximum
- The number of categories is above the maximum, compared to the total number of records.
- The variation coefficient is under the minimum. (This measure is only applicable to categorical attributes).
- The standard deviation is under the minimum. (This measure is only applicable to continuous attributes).

The minimum or maximum, so the boundaries, are set by the user. The attributes that satisfy one or more of these criteria are removed as inputs for the next step, because they are not influencing factors concerning the dependent variable. Besides, records with a missing value for the target attribute, or missing values for all independent variables, are excluded from the next step.

Ranking

In this step, first the independence between the input attribute and the target are tested for all attributes. In this case, where the dependent variable is a categorical variable and the inputs consists of both categorical and ordinal attributes, the independency can be tested by using the Pearson chi-square test or the Likelihood ratio chi-square test. Last, the attributes are ranked based on their resulting p-values.

Selecting

Based on the results of the independency test, the option will select important attributes, based on one of the following options:

- Labeling the attributes as ‘important’, ‘marginal’ or ‘unimportant’ and selecting the attributes with the label ‘important’.
- Selecting the top *x* attributes based on their p-values.
- Selecting all attributes with a p-value above a certain value.

When the attributes are selected, you are able yourself to adjust this selection.

Feature selection is used to provide insights into the importance of the attributes and to remove the attributes that resulted as unimportant attributes according to the screening. Even though the decision tree algorithms are able to handle unimportant attributes, removing them in advance guarantees that the predictive model doesn't make decisions based on the values of these attributes.

4.4.2. Building classification decision trees

A classification decision tree is a model (looking like a tree) that classifies each instance based on their attribute values. The model starts with a dataset (the first node), after which it splits the dataset into subsets based on a decision rule. The subsets are placed in a so-called child node. Thereafter, each subset is split again in smaller subsets based on another decision rule. This process is repeated until the maximum number of levels (tree depth) is reached or when a split causes a worse accuracy.

The reason of creating classification decision trees, is because of the following characteristics[25]:

- It is easy to understand, because it is rule based;
- it is able to handle a binary dependent variable;
- it is able to handle categorical and numerical variables;
- it is not necessary to determine which attributes will be used as input, based on their relevance, because the generated decision tree model will point out which attributes are most important based on the ranking of decision rules. In other words, decision trees are able to deal with a high amount of (relevant and irrelevant) attributes as input. The model itself will indicate which attributes are most relevant.

Therefore, the input for all algorithms[26] will be the output of the feature selection option.

The C&RT algorithm

C&RT stands for 'classification and regression trees', since the C&RT algorithm is able to build decision trees for both continuous dependent variables (regression trees) and categorical dependent variables (classification trees).

The algorithm is as follows:

1. Start in the root node, with the full data set.
2. For each attribute, find a split into two subsets (the child nodes) that minimizes the sum of impurities, by minimizing the sum of Gini indexes. (See the formula below.)
3. Choose the attribute with the lowest sum of indexes and create these two child nodes.
4. Repeat step 2 and 3 until the final nodes only contain records from one class (in this case, only matches or non-matches).
5. Prune the decision tree to a smaller size that has the same or better estimate of the misclassification error than the maximum model, to avoid overfitting.

$$Gini\ index = \sum_{i=1}^{j} f_i * (1 - f_i) = 1 - \sum_{i=1}^{j} {f_i}^2 = \sum_{i \neq j} f_i f_j$$

[25] Source: (Pekelis, 2013)

[26] Source: (Loh, 2011), (IBM, C5.0 node, 2012)

The C5.0 algorithm

The C5.0 algorithm is almost identical to the C&RT algorithm. The only difference is that C5.0 uses the Information gain to choose the best split, instead of the Gini index. The information gain is the decrease in entropy:

$$information\ gain = property(parent\ node) - weighted\ average\ \big(property(children\ nodes)\big)$$

with $property\ of\ a\ node = \sum_{i=1}^{j} -p_i \log_2 p_i$

The attribute with the highest information gain, will be chosen.

One remark should be made. The C5.0 algorithm is an improvement of the C4.5 algorithm, in such way that it offers a boosting method that increases the accuracy and it allows to weight misclassification types.

The CHAID algorithm

The CHAID algorithm is as follows:

1. Start in the root node, with the full data set.
2. For each attribute, create a number of subsets. When the attribute is an ordinal variable, split the values into ten intervals, otherwise assign each value to one child node.
3. Use significance tests to merge pairs of these notes iteratively.
4. For each attribute, find the subsets, so the splits, that minimize the sum of Gini indexes.
5. Choose the attributes with the lowest sum of the indexes and create the child nodes.
6. Repeat step 2 and 3 until the final nodes only contain records from one class (in this case, only matches or non-matches).
7. Prune the decision tree to a smaller size that has the same or better estimate of the misclassification error than the maximum model, to avoid overfitting.

Because this algorithm merges attributes based on significance, it is possible that one node is split in more than two nodes. This is different compared to the C&RT and C5.0 algorithm.

The approach

After running the feature selection option, ten different datasets are created by using 10-fold cross validation.[27] This method splits the dataset into ten sub datasets, were nine subsets form the training set (90.000 instances), and the remaining subset forms the test set (10.000 instances). This is repeated ten times, in order that each dataset has a different subset as test set. Afterwards, for each of the ten resulting data sets, the different algorithms will be run on the training set and tested on the test set. At last, the accuracy, sensitivity and specificity of the prediction of a model will be measured for each run. These measures are well known measures concerning the prediction of a binary target. This will result in three measurements per algorithm and run. The formulas of the measures are given below.

The accuracy is a measure that indicates the correctly predicted matches and non-matches, where sensitivity only indicates the correctly predicted matches and specificity the correctly predicted non-matches.

---

[27] Source: (Tang, 2008)

$$accuracy = \frac{\#\, true\ positives + \#\, true\ negatives}{total\ \#\ of\ predictions}$$

$$sensitivity = \frac{\#\, true\ positives}{\#\, true\ positives + \#\, false\ negatives}$$

$$specificity = \frac{\#\, true\ negatives}{\#\, true\ negatives + \#\, false\ positives}$$

4.4.3. Determining the best algorithm and building the final model

Because all measures are equally important, the comparisons between the measurements have the same weight. The best algorithm will be determined by:

- Calculating the average accuracy, sensitivity and specificity per algorithm and comparing them.
- Determining the algorithm that creates the most stable models, i.e. with the least variation in the measurements.

Based on the findings, the best algorithm will be selected to build the final model. The input for this model will be the full data set, so all 10,000 records.

5. Results

In order to build a good predictive model, the four steps of the approach are executed. The results are discussed below.

5.1. Exploratory data analysis

First, the percentage of missing values was calculated for each attribute. However, because none of the attributes exceeded the maximum percentage of 50%, it was not needed to exclude attributes from further research. (The attribute 'anonymous_c0' from the device table had the biggest percentage of missing values, that is 35%.)

Besides calculating the proportion of missing values, also some calculations were performed to provide insights in the data. Some of the findings are listed below. One remark should be made. All variables consist of anonymized values. Therefore it is not known which mobile devices are listed in the dataset for example. Besides, some attributes have an anonymous meaning.

Findings concerning the devices

- On average each individual owns 1 mobile device. The minimum is 1 and the maximum is 4. See table 3.
- The data contains eight different device types, which are anonymized. -1 indicates that the type of the device is unknown. Most devices belong to devtype_2 or devtype_4. Possibly these values refer to iPhones and android phones. See table 4.
- The number of different operating systems (OS) per device type differ from 1 to 37. Devtype_2 and devtype_4 have the largest number of different operating systems. See table 5.
- A device belongs to one of the 82 device countries, including the unknown value: -1. However, the greatest part belongs to country_146. See table 6. Since the firm Drawbridge is located in the United States, there is a great chance that country_146 refers to the United States.

Number of devices per person	Percentage
1	97.78%
2	2.16%
3	0.06%
4	0.01%

Table 3. Number of devices per person

Device type	Percentage
devtype_4	50.18%
devtype_2	44.50%
devtype_5	2.81%
devtype_7	1.58%
devtype_6	0.79%
devtype_1	0.10%
-1	0.03%
devtype_3	0.01%

Table 4. Device type

Device type	Number of OS	Percentage
devtype_4	37	26.81%
devtype_2	31	22.46%
devtype_5	20	14.49%
devtype_7	19	13.77%
devtype_6	18	13.04%
devtype_1	10	7.25%
devtype_3	2	1.45%
-1	1	0.72%

Table 5. Number of different operating systems per device type

Country	Percentage
country_146	91.85%
-1	1.96%
country_169	1.91%
country_201	1.91%
Remaining 78 countries	2.37%

Table 6. The country of a device

Findings concerning the cookies

- On average each individual is linked to 4 cookies. The minimum is 1 and the maximum is 4. See table 7.
- A cookie is linked to one of the 84 computer OS types, including the value -1. Most cookies are related to computer_os_type_133 or computer_os_type_203. See table 8. One remark should be made. Because these cookies are desktop cookies and the devices are mobile devices, it is not possible to the OS attribute of the device table with the OS attribute of the cookie table.
- A cookie is linked to one of the 300 browser versions, including the value -1. Mainly used browser versions are computer_browser_version_683, computer_browser_version_36, computer_browser_version_1158 and computer_browser_version_875. See table 9.
- A cookie belongs to one of the 80 device countries, including the unknown value: -1. Again, the main part belongs to country_146. See table 10.

Number of cookies per person	Percentage
1	95.80%
2	3.93%
3	0.26%
4	0.01%

Table 7. Number of cookies linked to a person

Computer OS type	Percentage
computer_os_type_203	43.31%
computer_os_type_133	24.79%
computer_os_type_109	4.39%
computer_os_type_149	3.23%
Remaining 80 computer OS types	24.27%

Table 8. Computer operating system types

Computer browser version	Percentage
computer_browser_version_875	29.32%
computer_browser_version_1158	11.92%
computer_browser_version_36	9.81%
computer_browser_version_683	8.43%
computer_browser_version_377	3.78%
computer_browser_version_1238	3.71%
computer_browser_version_1421	3.33%
computer_browser_version_897	3.14%
computer_browser_version_978	2.79%
Remaining 291 browser versions	23.76%

Table 9. Computer browser versions

Country	Percentage
country_146	90.50%
-1	3.52%
country_201	1.94%
country_169	1.90%
Remaining 76 countries	2.13%

Table 10. Country of a cookie

Findings concerning the many to many relationship between tables

- Although in reality a mobile device can belong to multiple individuals, like a household, in this dataset each device belongs to one person.
- On average each individual owns 1 mobile device and is linked to 4 cookies.
- On average each device is traced on 14 IP addresses and each cookie on 4 IP addresses.
- On average each device is linked to 42 apps that are visited and each cookie to 48 websites that are visited. A remark should be made. When these properties (app or website) have the same property ID, there is a link. In other words, the content of the website can be seen by visiting the website of by using the application.
- On average each property (app or website) belongs to 23 categories.

See figure 3 for the averages linked to the tables. (The IP aggregation table consists of one record per IP address.)

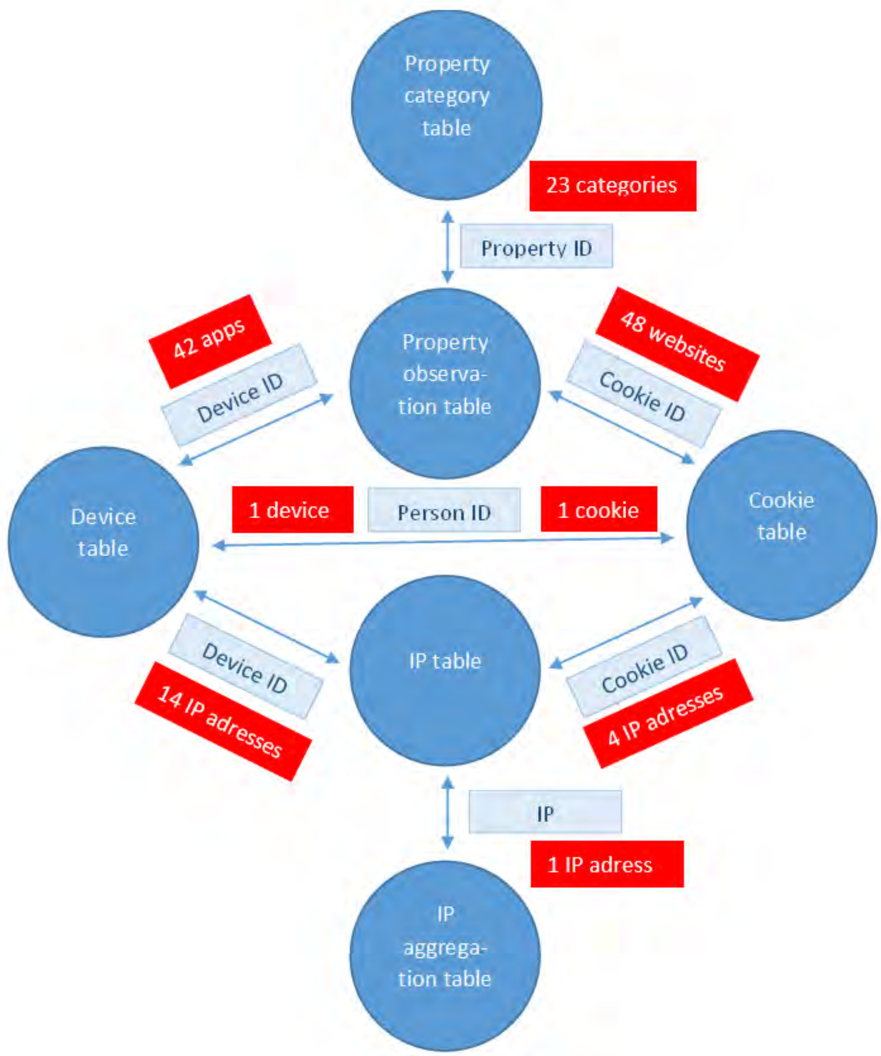

Figure 3. Relational data base with findings

Joining the tables in figure 3 would result in a high number of records for one device-cookie combination. See table 11 for an example. This is not desirable. Therefore it is needed to transform the property observation table, IP table and property category table in such way that they contain one row per ID or IP address.

Device ID	Cookie ID	IP address device	IP address cookie	Property ID device (app)	Property ID cookie (website)
Dev_1	Cookie_23	IP_13	IP_23	Prop_6	Prop_7
Dev_1	Cookie_23	IP_13	IP_23	Prop_6	Prop_8
Dev_1	Cookie_23	IP_13	IP_23	Prop_6	Prop_39
…	…	…	…	…	…
Dev_1	Cookie_23	IP_13	IP_23	Prop_6	Prop_32
Dev_1	Cookie_23	IP_13	IP_23	Prop_6	Prop_55

Table 11. Joined tables by using the original tables

5.1. Transformation of the tables

After creating reference tables and replacing the values of nominal and categorical variables by their reference number, the following tables were created based on transformations:

- The IP top 2 table
- The property top 2 table
- The property category count table

The IP top 2 table

The behavior of individuals through their devices is an important element when predicting the connection between these devices. An indication of this connection could be matching IP addresses or property IDs (so, websites and corresponding apps). Exploratory data analysis on a sample of matching devices and cookies showed that frequently there is a match of IP addresses when selecting only the top 2 IP addresses of devices and the top 2 IP addresses of cookies (based on frequency). Therefore, the IP table is transformed into a table that contains one record for each (device or cookie) ID together with their top 2 IP addresses. Furthermore a column is added with the total number of IP addresses per ID. In that way not all information will be lost.

The property top 2 table

The same table is created for properties. However, because of the low percentage of property matches when having a device-cookie match, 0.18%, there is decided to calculate the number of matches over all property IDs instead of only looking at the top 2. See paragraph 5.2 for the creation of this count.

The property category count table

The original table, the property observation table, only consists of two columns: the property ID and property category. Because of the large number of categories, it is not desirable to use the original table. Therefore the tables is transformed into a new table where each records contains the property ID and the total number of categories it belongs to.

5.2. Realization of the data set

After transforming the three tables, the next step was transforming the relational data base into a flat file data base. The tables were joined on matching IDs and IP addresses, whereupon 5000 records were selected randomly. Afterwards, 5000 records were created with non-matching device-cookie combinations. Last, these two tables were combined and a binary dependent variable, called 'MATCH', was added. When a record contains a device and cookie that belong to the same person, the target MATCH has value 1, and otherwise it has value 0.

Finally the following three columns were added that contain additional information:

- IP_match
- Country_match
- Property_match

The IP_match variable has the values 0, 1 and 2. A records gets the value 0 when there is no match between the top 2 IPs from the devices and the top IPs from the cookies. The value 1 is obtained when there is exactly one match and the value 2 when there are two matches. See table 12 for some findings based on this variable. The values in the tables show that when there are one or two IP matches, there is always a match between the device and cookie.

MATCH	# records IP_match = 0	# records IP_match = 1	# records IP_match = 2	Total # records
0	5000	0	0	5000
1	999	3698	303	5000

Table 12. IP_match counts

The country_match variable also has the values 0, 1 and 2. The value 0 is obtained when there is no match between the device and cookie country, 1 when there is a match and both have value country_146 and 2 when there is a match, but concerning another country than country_146. There's made a distinction between country_146 and the other countries, because the results of the exploratory data analysis showed that 90% of the cookie IDs and device IDs belong to country_146. See table 13 for some findings based on this variable. It makes sense that when there is a match between the countries, it doesn't have to mean there is a match between cookie and device, especially concerning country_146. However, because only 10 per cent of the cookies and device belong to another country, a match between these countries can certainly be important.

MATCH	# records country_match = 0	# records country_match = 1	# records country_match = 2	Total # records
0	805	4192	3	5000
1	230	4477	293	5000

Table 13. Country_match counts

The property_match is a continuous variable, because it consists of the number of matches on property ID. This attribute is created by joining the properties of a device an cookie and counting the number of matches. Finally, the column with the number of matches is added to the data set. See table 14 for some findings based on this variable. Looking at the percentages, the proportions for MATCH = 0 and MATCH = 1 are almost the same, namely: 60%-40%. Therefore, it is reasonable that the attribute property_match is not an important attribute and will not be shown in the decision tree.

MATCH	Minimum nr of matches	Average nr of matches	Maximum nr of matches	Percentage without matches	Percentage with matches
0	0	2.31	76	60.82%	39.18%
1	0	2.49	58	59.94%	40.06%

Table 14. Property_match counts

After adding the three attributes, the creation of the data set is finished and can be imported into SPSS Modeler.

5.3. Realization of a predictive model

After importing the prepared data, the feature selection was executed, followed by building prediction models based on three different algorithms, measuring the quality of the models, choosing the best algorithm and finally building the final model.

5.3.1. Feature selection

To decide which attributes to remove before running the algorithms, the attributes were screened on all five criteria, with their standard settings. That is:

- a maximum percentage of missing values of 70% (all fields)
- a maximum percentage of records in a single category of 90% (categorical fields)
- a maximum number of categories as a percentage of records of 95% (categorical fields)
- a minimum coefficient of variation of 0.1 (continuous variables)
- a minimum standard deviation of 0.0 (continuous variables)

Screening the attributes resulted in fourteen attributes (see *table 15*) that did not satisfy one or more of these conditions.

The number of times that a device or cookie appeared on the IP address.

Attribute	**Meaning**	**Reason**
device_country	The country to which a device belongs	single category too large
device_anonymous_7	Anonymous meaning (info concerning the device)	coefficient of variation below threshold
dev_ip1	IP address where a device appeared the most on	single category too large
dev_ip2	IP address where a device appeared the second most on	single category too large
dev_app1_id	Application ID that is visited the most	single category too large
dev_app2_id	Application ID that is visited the second most	single category too large
cookie_country	The country to which a cookie belongs	single category too large
cookie_anonymous_7	Anonymous meaning (info concerning the cookie)	coefficient of variation below threshold
cookie_ip1	IP address where a cookie appeared the most on	single category too large
ip1C_is_cellular_ip	Whether the IP address is a cellular IP address (1) or not (0).	single category too large
cookie_ip2	IP address where a cookie appeared the second most on	single category too large
ip2C_is_cellular_ip	Whether the IP address is a cellular IP address (1) or not (0).	single category too large
cookie_website1_id	Website ID that is visited the most	single category too large
cookie_website2_id	Website ID that is visited the second most	single category too large

Table 15. Resulting attributes of the screening

Looking at the attributes of table 1, most attributes are IDs or IP addresses. This is reasonable, because one ID or IP address only occurs a couple of times in the data set. The opposite applies to the countries and the attributes concerning cellular IP addresses. As seen before, more than 90% of the devices and cookies belong to country_146. Besides, more than 90% of the cookie IP addresses are not cellular IPs. Because these attributes are not influencing factors concerning the dependent variable, they were removed as inputs for the next step, but also as inputs for the algorithms.

Afterwards, the remaining attributes were ranked using the standard setting of feature selection, namely the Pearson chi-square test. The top 10 attributes, so the most important attributes based on their dependency, are listed in *table 16*.

Attribute	**Meaning**	**P-value**
ip_match	Match between IP address of device and cookie	1,000
country_match	Match between country of device and cookie	1,000
cookie_computer_os_type	The type of computer operating system	1,000
cookie_anonymous_6	Anonymous meaning (info concerning the cookie)	1,000
cookie_computer_browser_version	The cookie browser version	0,994
cookie_anonymous_5	Anonymous meaning (info concerning the cookie)	0,993
cookie_ip1_anonymous_c5	Anonymous meaning (info concerning the IP that is most seen on the cookie)	0,982
ip1C_anonymous_c1	Anonymous meaning (info concerning the IP)	0,978
ip2C_anonymous_c0	Anonymous meaning (info concerning the IP)	0,969
cookie_ip1_anonymous_c3	Anonymous meaning (info concerning the IP that is most seen on the cookie)	0,958

Table 16. Top 10 attributes

The top 2 attributes, ip_match and country_match, was expected. Although a match between the countries with both value 'country_146' does not have to indicate a match between a device and cookie, a match were the countries have another value, could be important information. It was also expected that the attribute 'property_match' is not in list of top 10 attributes, because of the similar distributions (see table 14). However, the other eight attributes in the list were unexpected, because no patterns in the explanatory data analysis were found concerning these attributes.

Lastly, the feature selection option selected this top 10 of attributes. However, this selection was changed into all attributes excluding the attributes that resulted from the screening.

5.3.2. Building classification decision trees

After selecting the inputs for the three algorithms, the data set was split into two subsets: a data set where all records have ip_match = 1 or ip_match = 2 (4001 records in total) and a data set where all records have ip_match = 0 (5999 records in total). The reason for this division is the fact that all records with an match between IP addresses, also have a match between the device and cookie. As a result, these two subsets are the first two child nodes. Because, 4001 records of the subset with IP matches are all matches between the device and cookie, no further decision rules are needed. Therefore, after splitting the data set into these two subsets, the algorithms were only run on the subset of 5999 records.

First ten different data sets, each of approximately 600 records, were created by using 10-fold cross validation. Afterwards, the three algorithms were run and the quality of the models was measured based on the accuracy, the sensitivity and specificity. Obviously, when measuring the quality, also the prediction based on the first decision rule (IP match = 0 or IP match = 1 or 2) was included. The results of these measures are listed in *table 17*.

	C5.0			**C&RT**			**CHAID**		
Model	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Model 1	0.8970	0.8623	0.9568	0.9180	0.8618	0.9686	0.9300	0.8104	0.9827
Model 2	0.8860	0.8557	0.9543	0.8590	0.8371	0.9933	0.9030	0.8606	0.9515
Model 3	0.8950	0.8209	0.9815	0.8980	0.8320	0.9795	0.8850	0.8671	0.9697
Model 4	0.9330	0.8577	0.9492	0.9250	0.8367	0.9677	0.8930	0.8806	0.9542
Model 5	0.9110	0.8602	0.9537	0.9210	0.8499	0.9843	0.9320	0.8750	0.9730
Model 6	0.9390	0.8323	0.9810	0.8730	0.8059	0.9914	0.9230	0.8239	0.9843
Model 7	0.8930	0.8465	0.9834	0.9290	0.8297	0.9942	0.9200	0.8628	0.9720
Model 8	0.8850	0.8665	0.9468	0.9130	0.8600	0.9735	0.9200	0.8770	0.9443
Model 9	0.9230	0.8367	0.9788	0.9290	0.8201	0.9863	0.8780	0.8270	0.9809
Model 10	0.8950	0.8557	0.9360	0.9360	0.8422	0.9924	0.9270	0.8753	0.9559
Average	**0.9057**	**0.8495**	**0.9622**	**0.9101**	**0.8375**	**0.9831**	**0.9111**	**0.8560**	**0.9669**

Table 17. Measurements

Looking at the averages, the values per measure are very close. The CHAID algorithm has the highest average of accuracy and sensitivity (correctly predicted matches) and the C&RT algorithm has the highest mean of specificity (correctly predicted non-matches). Because the overall prediction, but also the correctly predicted matches is important (so the marketeers send effective advertisements), the CHAID algorithm is chosen as best algorithm. Besides, the variation between the measurements of the ten models is low, so the CHAIN algorithm is stable. That is also taken into account. One remark should be made. The specificity is relatively low, compared to the accuracy and specificity. Therefore, the false negatives should be a point of concern for further investigation.

Lastly, the CHAIN model is run over the total subset of 5999 records, so all records without an match between IPs. This resulted in a decision tree with a three depth of 6, including the first decision rule of the IP_match. See table 18 for the first three levels of the model and see figure 4 of a visualization of the first two levels. The full model is described in appendix III.

Looking at table, the first and second decision rule were expected. Choosing for a division between a match of country_146 and matches of other countries, has been a good choice, since it is a decision rule in the model. When there is no country match, the next split is based on the attribute 'ip1D_is_cellular_ip', in other words, whether the IP address that is most seen on a device is a cellular IP address. When this is true, the model predicts that there is no match between device and cookie. This can be related to the fact that less than 10% of all cookie IPs are cellular IPs. (See paragraph *5.3.1. Feature selection*.) However, this should be further investigated. When both device and cookie belong to country_146, the next decision is made based on the values for attribute 'ip1D_anonymous_c2', an anonymous attribute of the IP address that is most seen on a device. This attribute is also a point of concern for further investigation. See *figure 4* for the number of records and number of matches versus non-matches per node.

The measure values of the final model are the following:

- Accuracy: 0.9160
- Sensitivity: 0.8526
- Specificity: 0.9794

```
ip_match = 0 [ Mode: 0 ]
        Country_match_adj in [ "0" ] [ Mode: 0 ]
                ip1D_is_cellular_ip = 0 [ Mode: 0 ]
                        …
                ip1D_is_cellular_ip = 1 [ Mode: 0 ] → 0
        Country_match_adj in [ "1" ] [ Mode: 0 ]
                ip1D_anonymous_c2 <= 13 [ Mode: 0 ]
                        …
                ip1D_anonymous_c2 > 13 and ip1D_anonymous_c2 <= 19 [ Mode: 0 ]
                        …
                ip1D_anonymous_c2 > 19 and ip1D_anonymous_c2 <= 986 [ Mode: 0 ]
                        …
                ip1D_anonymous_c2 > 986 [ Mode: 0 ]
                        …
        Country_match_adj in [ "2" ] [ Mode: 1 ] → 1
ip_match = 1 or ip_match = 2 [ Mode: 1 ] → 1
```

Table 18. Fist three levels of the final model

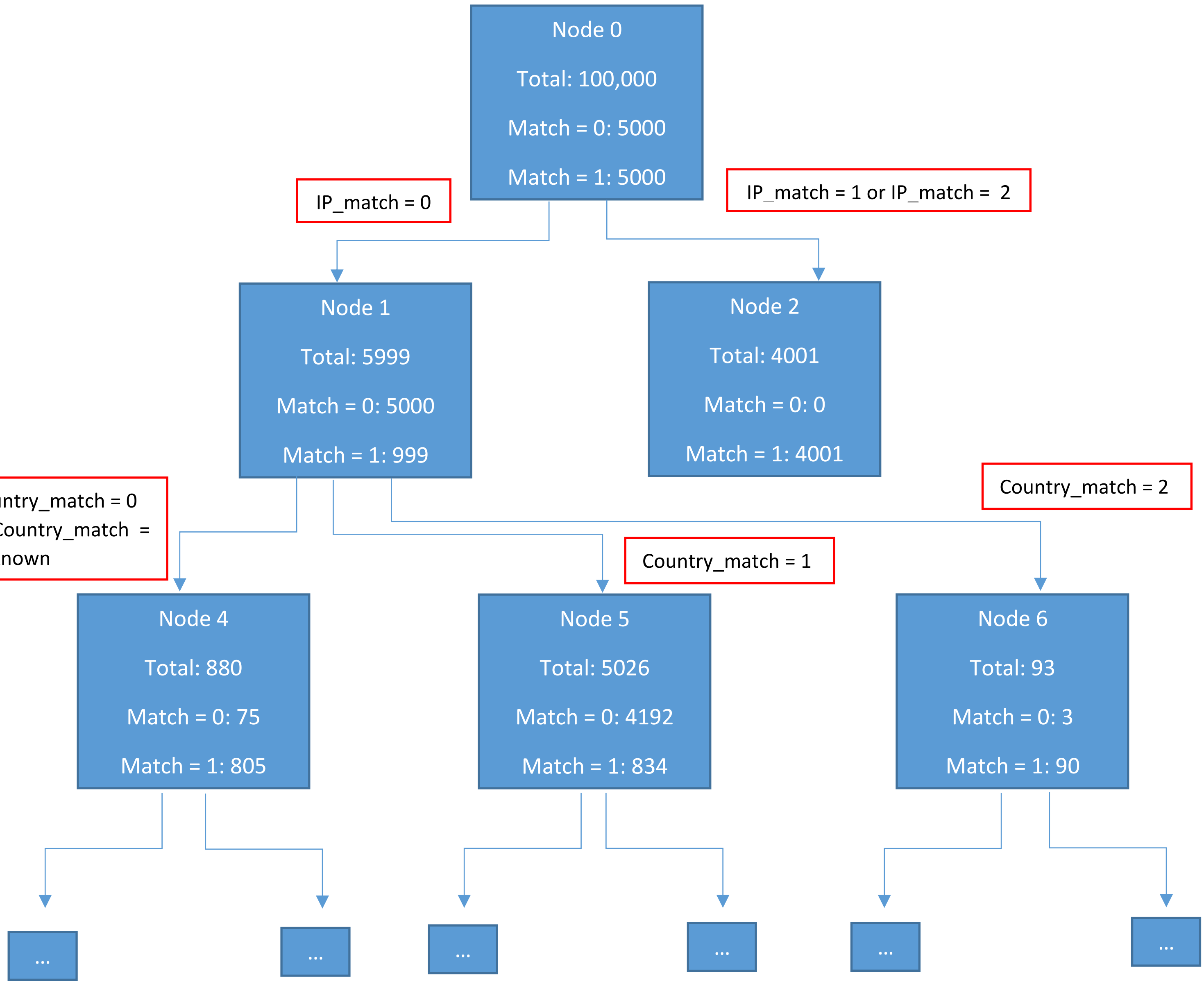

Figure 4. The first two levels of the final model

Conclusion and recommendations

The goal of this research was the following:

building a predictive model with an accuracy of 90% or more, that predicts which mobile devices and cookies belong to the same individual.

This research goal is achieved in the sense that the final model, that predicts whether one device and one cookie belong to the same individual, has an accuracy of 0.9160. (Besides, the model has a sensitivity of 0.8526 and a specificity of 0.9794.) However, this accuracy would become less when the model should predict all devices and cookies belonging to one person correctly. Because the latter is the goal of cross device tracking, further research is needed. There are several points of concern, concerning the explanatory data analysis, the transformation of data and building models. These are listed below:

- Doing more research into the attributes that are the base for the final decision tree (especially the attributes from the third level until the sixth level), such as ip1D_is_cellular_ip, ip1D_anonymous_c0 and ip2C_anonymous_c0.
- Doing more research into the IP addresses, including:
    - The match between IP addresses. In this case only the top 2 IP addresses are selected and matched, but this resulted in the situation where there is no match between the IPs, but there is a match between device and cookie. However, when matching all IP addresses from a device and cookie, this situation could disappear or become less. However, this would result in a longer running time.
    - The distinction between cellular and non-cellular IP addresses. Because less than 10% of the IP adresses that belong to cookies are cellular IP addresses there is no possible match with a cellular IP address from a device. Therefore, this should be further investigated.
- Paying more attention to false negatives, because of the relative low sensitivity
- Keeping in mind that the labels of the country match (0, 1 and 2) are working well in this case, because approximately 90% of the cookies and devices belong to country_146. However,in another situation this could be different.
- Building models based on different decision tree algorithms, like random trees and QUEST.
- Doing more research into other approaches of linkage problems and build models based on these approaches.
- Building an algorithm that predicts the relation for each plausible combination of device and cookie.

All in all, the findings and the final model are a good starting point for further research.

References

Brite, L. (2015, June 25). *Why We Must Move Past Cookies on Mobile*. Retrieved from SpotX: https://www.spotxchange.com/resources/blog/product-pulse/productpulse-why-we-must-move-past-cookies-on-mobile/

Cisco. (2016, February 1). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020 White Paper*. Retrieved from Cisco: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html#Trend_5_Profiling_Mobile_Apps_Use

Drawbridge. (2015). *About us - We're Perfecting the Art of Connecting Brands with Consumers*. Retrieved from Drawbridge: https://drawbridge.com/about-us

Drawbridge. (2015, June 1). *ICDM 2015: Drawbridge Cross-Device Connections - Data files.* Retrieved from Kaggle: https://www.kaggle.com/c/icdm-2015-drawbridge-cross-device-connections/data

Drawbridge. (2015, June 1). *ICDM 2015: Drawbridge Cross-Device Connections - Description*. Retrieved from Kaggle: https://www.kaggle.com/c/icdm-2015-drawbridge-cross-device-connections

Drawbridge. (2015). *Solutions - We're Powering a More Personalized Internet for Everyone*. Retrieved from Drawbridge: https://drawbridge.com/solutions

eMarketing: The Essential Guide to Online Marketing. (n.d.). Retrieved from Saylor Academy: http://www.saylor.org/site/textbooks/eMarketing%20-%20The%20Essential%20Guide%20to%20Online%20Marketing.pdf

Goldfarb, A., & Tucker, C. E. (2011). *ACM.* Retrieved from Online advertising, behavioral targeting, and privacy: http://portal.acm.org/citation.cfm?id=1941498

Groep, L. (2012, August). *Impact van de nieuwe cookiewet op Search Marketing, Web &.* Retrieved from Searchresult - performance marketing: https://www.searchresult.nl/wp-content/uploads/Whitepaper-Cookiewet-LECTRIC-Groep.pdf

IAB. (2015, August). *Understanding mobile cookies.* Retrieved from IAB: https://www.iab.com/wp-content/uploads/2015/08/IABDigitalSimplifiedMobileCookies.pdf

IBM. (2012). *C5.0 node*. Retrieved from IBM knowledge center: https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/c50node_general.htm

IBM. (2012). *Feature selection model settings*. Retrieved from IBM knowledge center: https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/featureselection_settings.htm

IBM. (2012). *Feature selection node*. Retrieved from IBM knowledge center: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/featureselectionnode_general.htm

IBM. (2012). *Feature selection options*. Retrieved from IBM knowledge center: https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/featureselection_options.htm

Jegatheesan, S. (n.d.). *Cookies – Invading Our Privacy for Marketing, Advertising and Security Issues.* Retrieved from Cornell University Library: https://arxiv.org/ftp/arxiv/papers/1305/1305.2306.pdf

KPMG. (2016). *The Changing Landscape of Disruptive Technologies (KPMG Technology Innovation Survey).* Retrieved from KPMG Technology Innovation Center: https://techinnovation.kpmg.chaordix.com/static/docs/TechInnovation2015-Part2.pdf

Leune, L. (2016, January 6). *Aan de slag met cross device tracking: twee manieren*. Retrieved from Emerce: http://www.emerce.nl/best-practice/aan-de-slag-met-cross-device-tracking-2-manieren

Loh, W.-Y. (2011). *Classification and regression trees.* Retrieved from Department of Statistics: http://www.stat.wisc.edu/~loh/treeprogs/guide/wires11.pdf

Mayer, J. R., & Mitchell, J. C. (n.d.). *Third-Party Web Tracking: Policy and Technology.* Retrieved from https://jonathanmayer.org/papers_data/trackingsurvey12.pdf

Optanon. (n.d.). *The Cookie Law Explained*. Retrieved from Optanon: https://www.cookielaw.org/the-cookie-law/

Pekelis, L. (2013). *Classification and regression trees: a practical guid for describing a dataset.* Retrieved from Stanford - Department of Statistics: http://statweb.stanford.edu/~lpekelis/talks/13_datafest_cart_talk.pdf

Schmücker, N. (2011). *Web tracking.* Retrieved from Service-centric Networking - Department of Telecommunication Systems: http://www.snet.tu-berlin.de/fileadmin/fg220/courses/SS11/snet-project/web-tracking_schmuecker.pdf

Signal. (n.d.). *The Advertiser's Guide to People-Based Marketing.* Retrieved from CIO summits: http://www.ciosummits.com/Signal_Advertisers_Guide_to_People_Based_Marketing_www.signal.co_.pdf

Tang, L. (2008). *Cross-validation.* Retrieved from Lei Tang: http://leitang.net/papers/ency-cross-validation.pdf

Tradedoubler. (2016). *Cross Device Tracking – Probabilistic versus deterministic matching*. Retrieved from Tradedoubler - Helping digital marketers succeed: http://www.tradedoubler.com/en/about/resources/cross-device-tracking-probabilistic-versus-deterministic-matching/

Appendices

Appendix I: Results from researches performed by Cisco and KPMG

According to a research performed by Cisco, the number of mobile devices globally was around 7.9 billion in 2015, which will increase to 11.6 billion in 2020. Furthermore, Cisco predicts that the data traffic will increase as well. Where the data traffic was equal to 3.7 exabytes per month in 2015, it will probably become 30.6 exabytes by 2020, which is almost 8 times as big. See figure 5 and 6 for these numbers.

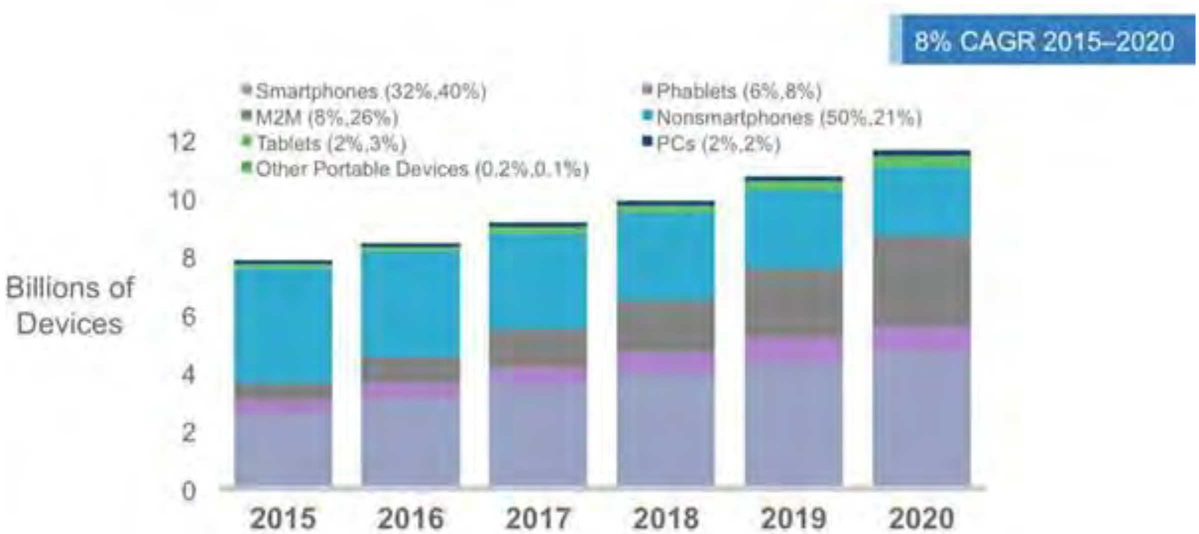

Figure 5. Global mobile devices and connetions growth

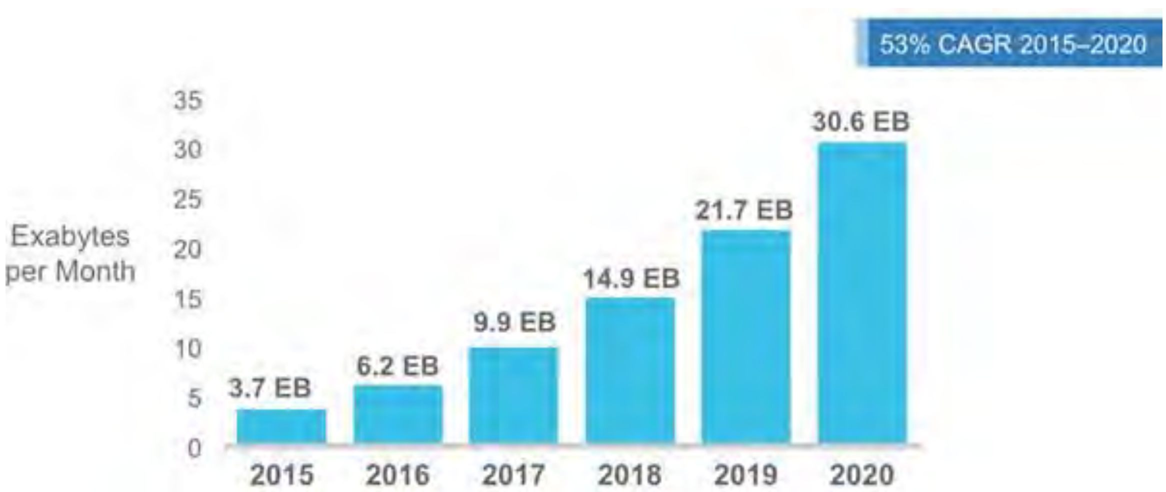

Figure 6. Global growth of mobile data traffic

The 'Global technology innovation survey of 2015' of the firm KPMG points out that the industry with the biggest growth opportunities is the retail industry. According to this survey, when adopting the

Internet of Things the sales of the global retail industry will increase with 22% within three years. See figure 7 for this number.

Q: Which vertical has the greatest monetization potential as a result of the adoption of Internet of Things in the next three years?*

Industry	Global	U.S.	China	Japan	ASPAC	EMEA
Consumer markets/retail	**22%**	**21%**	**34%**	3%	**25%**	**18%**
Technology	13%	20%	8%	10%	10%	10%
Aerospace & defense	10%	5%	18%	**20%**	13%	10%
Education	10%	10%	4%	7%	9%	10%
Automotive/transportation	9%	5%	16%	3%	10%	9%
Healthcare	6%	11%	1%	3%	4%	6%
Financial services	6%	4%	3%	10%	6%	6%
Telecommunications	6%	5%	0%	17%	6%	6%
Energy	5%	4%	3%	13%	4%	6%
Manufacturing	4%	3%	4%	3%	5%	3%
Media	4%	4%	2%	7%	1%	6%
Services	3%	2%	1%	0%	3%	5%

*Partial list of verticals shown

Source: KPMG Tech Innovation Survey Year-End 2015

Figure 7. IoT potentials

Appendix II: The data tables

The data contains of seven data tables, which were provided as text files:

- device_train_basic.csv, containing high-level information about the devices. This dataset will be used as training set. *(142,771 records)*
- device_test_basic.csv, containing high-level information about the devices. This dataset will be used as test set. *(61,157 records)*
- cookie_basic.csv, containing high-level information about the cookies. *(2,175,521 records)*
- id_all_ip.csv, containing information about the behavior of a cookie or device on a certain IP address. *(2,390,247 records)*
- ipagg_all.csv, containing information about an IP address. *(11,037,815 records)*
- id_all_property.csv, containing information about a website (for cookie) or an application (for device) that a user visited. *(2,199,445 records)*
- property_category.csv, containing the category a website or application belongs to. *(368,567 records)*

An overview of the tables consisting of the attribute names, a meaning of the data attributes and the type of the data attributes is given below. This overview is provided by Drawbridge via the Kaggle competition.[28] Drawbridge defined for types of attributes: index, categorical, integer and boolean. (Index and categorial are enumerated types where index is used for IDs.)

device_train_basic.csv

Attribute name	**Data type**	**Meaning**
Drawbridge handle	Index	It can be seen as the ID of a person. Devices and cookies with the same 'Drawbridge handle' belong to the same person.
Device ID	Index	The index number of a device.
Device type	Categorical	Device type, like an android phone, android pad, iphone, ipad, et cetera.
Device OS version	Categorical	Device OS version, like ios 8.0.
Device country info	Categorical	The country to which a device belongs.
Anonymous_c0	Boolean	Anonymous meaning
Anonymous_c1	Categorical	Anonymous meaning
Anonymous_c2	Categorical	Anonymous meaning
Anonymous_5	Integer	Anonymous meaning
Anonymous_6	Integer	Anonymous meaning
Anonymous_7	Integer	Anonymous meaning

[28] Source: (Drawbridge, ICDM 2015: Drawbridge Cross-Device Connections - Data files, 2015)

device_test_basic.csv

Attribute name	Data type	Meaning
Drawbridge handle	Index	It can be seen as the ID of a person. Devices and cookies with the same 'Drawbridge handle' belong to the same person.
Device ID	Index	The index number of a device.
Device type	Categorical	Device type, like an android phone, android pad, iphone, ipad, et cetera.
Device OS version	Categorical	Device OS version, like ios 8.0.
Device country info	Categorical	The country to which a device belongs.
Anonymous_c0	Boolean	Anonymous meaning
Anonymous_c1	Categorical	Anonymous meaning
Anonymous_c2	Categorical	Anonymous meaning
Anonymous_5	Integer	Anonymous meaning
Anonymous_6	Integer	Anonymous meaning
Anonymous_7	Integer	Anonymous meaning

cookie_basic.csv

Attribute name	Data type	Meaning
Drawbridge handle	Index	It can be seen as the ID of a person. Devices and cookies with the same 'Drawbridge handle' belong to the same person.
Cookie ID	Index	The index number of a cookie.
Computer OS type	Categorical	The type of computer operating system, like Windows XP.
Browser version	Categorical	The cookie browser version, like Safari-6.0.
Cookie country info	Categorical	The country to which a cookie belongs.
Anonymous_c0	Boolean	Anonymous meaning
Anonymous_c1	Categorical	Anonymous meaning
Anonymous_c2	Categorical	Anonymous meaning
Anonymous_5	Integer	Anonymous meaning
Anonymous_6	Integer	Anonymous meaning
Anonymous_7	Integer	Anonymous meaning

id_all_ip.csv

Attribute name	Data type	Meaning
Device/cookie ID	Index	The index number of a device or cookie.
Device or cookie	Boolean	This attribute denotes whether it is a device (0) or a cookie (1).
IP	Index	The IP adress
Freq count	Integer	The number of times that a device or cookie appeared on the IP address.
Anonymous count 1	Integer	Anonymous meaning
Anonymous count 2	Integer	Anonymous meaning
Anonymous count 3	Integer	Anonymous meaning
Anonymous count 4	Integer	Anonymous meaning
Anonymous count 5	Integer	Anonymous meaning

ipagg_all.csv

Attribute name	Data type	Meaning
IP address	Index	The IP address
Is cell IP	Boolean	This attribute denotes whether the IP address is a cellular IP address (1) or not (0).
Total freq	Integer	The number of observations that are seen on the IP address.
Anonymous count c0	Integer	Anonymous meaning
Anonymous count c1	Integer	Anonymous meaning
Anonymous count c2	Integer	Anonymous meaning

id_all_property.csv

Attribute name	Data type	Meaning
Device/cookie ID	Index	The index number of a device or cookie.
Device or cookie indicator	Boolean	This attribute denotes whether it is a device (0) or a cookie (1).
Property ID	Index	The index number of a website (for a cookie) or mobile application (for a device).
Property unique count	Integer	The number of times that a cookie or device is seen on this property.

property_category.csv

Attribute name	Data type	Meaning
Property ID	Index	The index number of a website (for a cookie) or mobile application (for a device).
Property category	Categorical	The category to which the website or mobile app belongs.

Appendix III: The final decision tree

ip_match = 0 [Mode: 0]
Country_match in ["0"] [Mode: 0]

ip1D_is_cellular_ip = 0 [Mode: 0]

dev_total_nr_of_ips <= 2 [Mode: 0] => 0

dev_total_nr_of_ips > 2 and dev_total_nr_of_ips <= 4 [Mode: 0] => 0

dev_total_nr_of_ips > 4 [Mode: 0] => 0

ip1D_is_cellular_ip = 1 [Mode: 0] => 0

Country_match in ["1"] [Mode: 0]

ip1D_anonymous_c2 <= 13 [Mode: 0]

ip1C_anonymous_c1 <= 64 [Mode: 0]

cookie_total_nr_of_ips <= 2 [Mode: 0]

ip1D_anonymous_c0 <= 26 [Mode: 0] => 0

ip1D_anonymous_c0 > 26 and ip1D_anonymous_c0 <= 51 [Mode: 0] => 0

ip1D_anonymous_c0 > 51 [Mode: 0] => 0

cookie_total_nr_of_ips > 2 and cookie_total_nr_of_ips <= 9 [Mode: 0]

ip2C_anonymous_c0 <= 17 [Mode: 0] => 0

ip2C_anonymous_c0 > 17 and ip2C_anonymous_c0 <= 37 [Mode: 0] => 0

ip2C_anonymous_c0 > 37 [Mode: 0] => 0

cookie_total_nr_of_ips > 9 [Mode: 0]

ip2D_anonymous_c2 <= 134 or ip2D_anonymous_c2 IS MISSING [Mode: 0] => 0

ip2D_anonymous_c2 > 134 [Mode: 0] => 0

ip1C_anonymous_c1 > 64 and ip1C_anonymous_c1 <= 620 [Mode: 0]

cookie_ip1_anonymous_c1 <= 0 [Mode: 0] => 0

cookie_ip1_anonymous_c1 > 0 [Mode: 0]

ip1D_anonymous_c0 <= 50 [Mode: 0] => 0

ip1D_anonymous_c0 > 50 [Mode: 0] => 0

ip1C_anonymous_c1 > 620 [Mode: 0]

ip2C_total_freq <= 7.821 or ip2C_total_freq IS MISSING [Mode: 0] => 0

ip2C_total_freq > 7.821 [Mode: 0] => 0

ip1D_anonymous_c2 > 13 and ip1D_anonymous_c2 <= 19 [Mode: 0] ip1C_anonymous_c0 <= 51 [Mode: 0]

dev_ip1_anonymous_c4 <= 2 [Mode: 0]

dev_ip2_freq_count <= 22 or dev_ip2_freq_count IS MISSING [Mode: 0] => 0

dev_ip2_freq_count > 22 [Mode: 1] => 1

dev_ip1_anonymous_c4 > 2 [Mode: 0] => 0

ip1C_anonymous_c0 > 51 [Mode: 0] => 0

ip1D_anonymous_c2 > 19 and ip1D_anonymous_c2 <= 986 [Mode: 0]

cookie_total_nr_of_ips <= 2 [Mode: 0]

ip1C_anonymous_c0 <= 47 [Mode: 0] => 0

ip1C_anonymous_c0 > 47 [Mode: 0]

dev_ip1_anonymous_c5 <= 4 [Mode: 0] => 0

dev_ip1_anonymous_c5 > 4 [Mode: 0] => 0

cookie_total_nr_of_ips > 2 and cookie_total_nr_of_ips <= 5 [Mode: 0] => 0

cookie_total_nr_of_ips > 5 [Mode: 0] => 0

ip1D_anonymous_c2 > 986 [Mode: 0]

ip2D_total_freq <= 22.563 or ip2D_total_freq IS MISSING [Mode: 0]

cookie_total_nr_of_ips <= 3 [Mode: 0]

dev_total_nr_of_ips <= 4 [Mode: 0] => 0

dev_total_nr_of_ips > 4 [Mode: 0] => 0

cookie_total_nr_of_ips > 3 [Mode: 0] => 0

ip2D_total_freq > 22.563 [Mode: 0]

ip1C_total_freq <= 2.487 [Mode: 1]

cookie_total_nr_of_ips <= 3 [Mode: 0] => 0

cookie_total_nr_of_ips > 3 [Mode: 1] => 1

ip1C_total_freq > 2.487 and ip1C_total_freq <= 13.296 [Mode: 0] => 0

ip1C_total_freq > 13.296 and ip1C_total_freq <= 62.856 [Mode: 1] => 1

ip1C_total_freq > 62.856 [Mode: 1] => 1

Country_match in ["2"] [Mode: 1] => 1

ip_match = 1 or ip_match = 2 [Mode: 1] => 1