



Company Project # 2015-1428

Martin Nicklas Jørgensen tzk173@alumni.ku.dk

User Behavior Analysis Using Decision Trees

Supervisor: Mikkel Rønne Jakobsen mikkelrj@di.ku.dk

December 12, 2015

Abstract

Write something sensible, requirements is on the intranet page.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec a diam lectus. Sed sit amet ipsum mauris. Maecenas congue ligula ac quam viverra nec consectetur ante hendrerit. Donec et mollis dolor. Praesent et diam eget libero egestas mattis sit amet vitae augue. Nam tincidunt congue enim, ut porta lorem lacinia consectetur. Donec ut libero sed arcu vehicula ultricies a non tortor. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean ut gravida lorem. Ut turpis felis, pulvinar a semper sed, adipiscing id dolor. Pellentesque auctor nisi id magna consequat sagittis. Curabitur dapibus enim sit amet elit pharetra tincidunt feugiat nisl imperdiet. Ut convallis libero in urna ultrices accumsan. Donec sed odio eros. Donec viverra mi quis quam pulvinar at malesuada arcu rhoncus. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. In rutrum accumsan ultricies. Mauris vitae nisi at sem facilisis semper ac in est.

Contents

1	Introduction	2
2	Simplesite ApS	2
2.1	Product	2
2.2	Departments	2
2.3	Company Structure	3
3	Problem Description	3
3.1	Requirements	3
3.2	Success Criteria	3
4	Problem Analysis	4
4.1	Available Data	4
4.2	Pruning Datasets	5
5	Results	6
6	Competencies and Methods	6
7	Conclusion	6

1 Introduction

2 Simplesite ApS

2.1 Product

Simplesite produce and manage an inhouse website Content Management System (CMS) and operate a hosting location where this CMS runs. Customers can register for a free account allowing them to host a website that is edited through the CMS. The free accounts are under restrictins on number fo pages, images and videos that can be added to the site. Customers can then change to a paid subscription which allows them to have more pages, images and videos on their website.

Simplesite also offer additional services to their customers that can be bought for an extra fee if you are already a paying customer, this includes the ability to have a domain attached to your website, a webshop as well as more additional pages, images and videos.

The product is hosted partially on Simplesites own hardware in an offsite location, and using a number of cloud services to provide faster response times for certain data types.

2.2 Departments

- **Administration** - Situated at the ground floor of the Copenhagen office, this department contains the HR functions as well as finance.
- **Sales** - Also on the ground floor of the Copenhagen office, sales consists of full time employeeess and student helpers. The primary task is to sell the product, currently through localized ad management.
- **P & C¹** - The department sits on the upper floor of the Copenhagen office and is responsible for planning new features in cooperation with the developers as well as manage content on Simplesites own websites. P & C also manages communication such as newsletters and localized ad texts.

¹Product and Communication.

- **In-House Development** - Sitting next to P & C on the upper floor the developers are responsible for implementing new features as well as maintenance, analytics and bug fixing of the product.
- **Operations** - Daily operations are handled primarily by the company CTO, Thomas, as well as 2 part time students, operations sits on the upper floor in Copenhagen next to the developers.
- **Support** - Since the product is offered in several languages, a supporter is hired per language in a part time basis. All supporters work from home but get together once a month for a status meeting and to make sure new knowledge is shared and that relevant information can be given from the regular departments.
- **Remote Dev 1** - A small number of developers are hired in Bulgaria and have their own office in the Sofia, they are offered machines in the Copenhagen office they can VPN into and use for work. This allows the operations department to work from Copenhagen and still service the remote developers.
- **Remote Dev 2** - A number of developers are also hired from Serbia, like the remote developers in Bulgaria, they also VPN into the Copenhagen office and work on machines maintained by the regular operations employees.
- **Miscellaneous** - SimpleSite also occasionally employs external specialists or consultants. Depending on the need, they will either work in the Copenhagen office or from some remote location using VPN.

2.3 Company Structure

Insert finished graphml figure.

3 Problem Description

SimpleSite changed their subscription service to a so called freemium model during 2015. This model means that everyone can have a website for free, and it will never be closed. This way of doing subscriptions means that more people sign up, and become potential customers, however a trend is that most customers that sign up either don't become paying customers, or simply make a site and stop using it after a few days.

SimpleSite wishes to map the life-cycle of customers in an attempt to find out how good users (users that are *retained*, ie. have logins 3-4 weeks after they are created) use their site, as well as look at the life-cycles of the many free or abandoned customers. They hope to discover what, if any, the significant differences are in the different lifecycles. The goal is to attempt to guide new customers down the paths that are known to be "good" and hopefully detect if customers are stuck or have forgotten about their website.

3.1 Requirements

SimpleSite have already created some code in R² so any continued work should also be done in the R language. I have also been added to a Github repository with already populated with analytics code and a preferred structure and coding style that should be adhered to.

Two datasets are also available containing different information gathered from the live system. These two datasets should form the basis for all the dataanalysis performed. An additional dataset is being constructed during the project, but due to the time it takes to populate it with enough observations to be meaningful it might not be finished before the project is over.

3.2 Success Criteria

The project have a number of success criteria that should all be fulfilled to some degree.

²<https://www.r-project.org/>

1. A model for classifying customers are created.
2. New knowledge about customers lifecycles are acquired from the model. More preferably, do *retained* customers have something in common.
3. A prototype R script that can automatically build/create the model from new customer data should be created.
4. A method for using the model should be designed or reasoned about.

4 Problem Analysis

4.1 Available Data

The basis for the analysis is 2 datasets created by SimpleSite: **EngagementData** & **CustomerJourney**. Table 1 and Table 2 names the features of each dataset and what they represent. Both datasets contain users created between September 1st 2015 and September 15 2015.

Attribute Name	Attribute Data
<i>islogins1</i>	Bool, true if: one or more logins for the user.
<i>islogins2</i>	Bool, true if: two or more logins for the user.
<i>islogins3</i>	Bool, true if: three or more logins for the user.
<i>islogins4</i>	Bool, true if: four or more logins for the user.
<i>isedit30m</i>	Bool, true if: User edited site within 30 minutes of creation.
<i>isedit24h</i>	Bool, true if: User edited site within 24 hours of creation (excluding the first 30 minutes).
<i>isaddpage30m</i>	Bool, true if: User added a new page within 30 minutes of creation.
<i>isaddpage24h</i>	Bool, true if: User added a new page within 24 hours of creation (excluding the first 30 minutes).
<i>isimgupload30m</i>	Bool, true if: User uploaded their own image within 30 minutes of creation.
<i>isimgupload24h</i>	Bool, true if: User uploaded their own image within 24 hours of creation (excluding the first 30 minutes).
<i>iseditdesign30m</i>	Bool, true if: User edited site within 30 minutes of creation.
<i>iseditdesign24h</i>	Bool, true if: User edited site within 24 hours of creation (excluding the first 30 minutes).
<i>customerid</i>	Integer value with the customers unique ID.
<i>marketname</i>	String with the market the user came from (US, TR, DK etc.)
<i>siteverkey</i>	String with what version of the site the user is created in (US, TR, DK etc.)
<i>ispayer</i>	Bool, true if: The customer have a paid subscription.
<i>culturekey</i>	String with language information for the site (en-US, fr-FR etc.)
<i>iso14</i>	Bool used by marketing.

This window should maybe be expanded a little for more data.

Table 1: Features found in the **EngagementData** dataset.

Attribute Name	Attribute Data
<i>customerid</i>	Integer value with the customers unique ID.
<i>logins14</i>	Integer, number of times the customer logged in the first 14 days (week 1-2 after creation).
<i>logisnw2w4</i>	Integer, number of times the customer logged in in week 3-4 after creation.
<i>edits14</i>	Integer, number of times the customer edited a page within the first 14 days.
<i>iscjtrial</i>	Bool, true if: Always true, everyone starts as a trial.
<i>iscjonboarded</i>	Bool, true if: $\text{edits14} \geq 1$.
<i>iscjactivated</i>	Bool, true if: $\text{edits14} \geq 3$.
<i>iscjengaged</i>	Bool, true if: $\text{edits14} \geq 6$ and $\text{logins14} \geq 2$.
<i>iscjinvested</i>	Bool, true if: $\text{edits14} \geq 15$ and $\text{logins14} \geq 6$.
<i>iscjretained</i>	Bool, true if: $\text{logisnw2w4} \geq 1$.
<i>isimgupload1d</i>	Bool, true if: Customer uploaded an image within the first 24 hours of being created.
<i>iseditdesign1d</i>	Bool, true if: Customer edited the design within the first 24 hours of being created.
<i>isaddpage1d</i>	Bool, true if: Customer added a new page within the first 24 hours of being created.
<i>isedit1d</i>	Bool, true if: Customer edited a page within the first 24 hours of being created.

Table 2: Features found in the **CustomerJourney** dataset.

4.2 Pruning Datasets

The initial goal is to find customers who are retained (*iscjretained* = **True**) and see if there is some pattern, that SimpleSite can try to guide other customers down in order to increase the number of retained customers. With this in mind there is some attributes of the datasets that will not be helpful, either because they cannot be controlled/changed, or because they do not make sense. The following is a list of attributes removed from the **EngagementData** dataset during work, along with the reason for the removal.

- *islogins1* : Removed since it is always true for all customers.
- *islogins2* : Removed because the definition of a retained customer requires one or more logins, so this must always be true.
- *islogins3* : Same as *islogins2*.
- *islogins4* : Same as *islogins2*.
- *marketname* : Removed since we are unable to get a customer from a different market, we are interested in variable we can change for each customer.
- *siteverkey* : Same as *marketname*.
- *ispayer* : Removed because it is an alternative target variable, it does not say anything about how the user behaves, other than they are indeed a good customer.
- *culturekey* : *marketname*.
- *iso14* : Value used by marketing.

The following is a list of attributes removed from the **CustomerJourney** dataset during work, along with the reason for the removal.

- *logins14* : Removed because initial tests showed high bias. For a ctrees³ of depth 4, the three top levels was occupied with choices regarding logins14.
- *logisnw2w4* : Removed since this attribute is in the definition of our target variable *iscjretained*.

³Conditional Inference Tree.

is o14
usefull?
Clear
up with
Morten.

- *iscjtrial* : Removed since it is always true.
- *iscjonboarded* : Removed since it serves as an alternative target variable, and is set by us, it does not say anything about the user behaviour that is not already present.
- *iscjengaged* : Same as *iscjonboarded*.
- *iscjinvested* : Same as *iscjonboarded*.

In both datasets the *customerid* attribute is kept in each dataset, even though it cannot be used as a feature for analysis since each customer have a unique ID and thus will not yield any patterns, since it can be used to join the two datasets together.

5 Results

6 Competencies and Methods

7 Conclusion