

MASTER

Behavior-based website redesign

how can data mining and process mining be applied to become more relevant to the end user on websites promoting complex products?

Penders, J.A.

Award date:
2015

[Link to publication](#)

Disclaimer

This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain

Behavior-based website redesign

How can data mining and process mining be applied to become more relevant to the end user on websites promoting complex products?

Master Thesis
Business Information Systems

Joris Penders (0657445)
j.a.penders@student.tue.nl

4-3-2015

Supervisors

B.F. van Dongen

b.f.v.dongen@tue.nl

F. van Geffen

f.geffen@rn.rabobank.nl

T. van den Berg

t.berg@rn.rabobank.nl

TABLE OF CONTENTS

Table of Contents	1
List of Tables	4
List of Figures	6
Terminology	8
Abbreviations	8
Preface	9
Management Summary	10
1 Introduction	11
1.1 Problem Description	12
1.2 Research Questions	13
1.3 Framework	13
1.4 Running example	14
2 How can we identify different information needs on a website?	16
2.1 Related work	16
2.2 Overview	17
2.3 Obtain the information need of one visitor	17
2.4 Apply segmentation on the set of visitors	19
2.5 Assign visitor into a cluster based on web data	20
3 How can we identify which visitor needs which information?	21
3.1 Related work	21
3.2 Train a model	22
3.3 Assess model	24
4 How do you present this information on a website?	25
4.1 Related work	25
4.2 Process Mining on webdata	26
4.3 Data Preparation for Process Mining	26
4.4 Process mining with buckets	27
4.5 Behavior on a page	28

5	Preparation phase - Business Understanding	30
5.1	General Introduction	30
5.2	Background	30
5.3	Business objectives	32
5.4	Inventory of Resources	32
5.5	Data Mining Goals	33
6	Preparation phase - Data understanding	34
6.1	Initial Data Collection	34
6.2	Description of data	34
6.3	Connection of profile information and visitorid	35
7	Case Study 1 - Optimize Conditional Content	36
7.1	Problem Description	36
7.2	Deliverables	39
7.3	Data Preparation	39
7.4	Current Performance CCF	42
7.5	Effect of classification	44
7.6	Modeling	45
7.7	Validation	49
8	Case Study 2 – Add Intelligence to FAQ	51
8.1	Problem Description	51
8.2	Deliverables	51
8.3	Data Preparation	51
8.4	Data Selection	52
8.5	Relevance of FAQ	53
8.6	Modeling	55
8.7	Behaviour on a page	59
9	Results	64
9.1	Conclusions for University	64
9.2	Conclusions for Rabobank	64

9.3	Recommendations for Rabobank	65
10	References	67
I	Appendix 1 – ETL Netinsight	68
II	Appendix 2 – Meta-data about tables Case Study	69
III	Appendix 3 – Overview of FAQ	70
IV	Appendix 4 – Conversion with FAQ	71
V	Appendix 5 – Decision tree FAQ on profile information	73
VI	Appendix 6 – Cluster with k-means	74
VII	Appendix 7 – KNIME models for clustering	75
VIII	Appendix 8 – Behavior on page	77

LIST OF TABLES

Table 1: Terminology.....	8
Table 2: Abbreviations	8
Table 3: Regular versus complex customer journey for purchasing one product	12
Table 4: Example of log data of a website	18
Table 5: Visitors and their questions.....	18
Table 6: Expected views of the pages	19
Table 7: Ranking of clusters	19
Table 8: Visitor 1005	20
Table 9: Visitor 1006	20
Table 10: Explanation of terms	22
Table 11: Visitors with extended information MyEnergy	23
Table 12: Buckets of MyEnergy.....	27
Table 13: Events on a page	28
Table 14: Views and events on a page	29
Table 15: Data sources	32
Table 16: Overview of data tables	34
Table 17: Current groups on Mortgage Home	37
Table 18: Target group and number of visitors.....	40
Table 19: Number of target groups per visitor	40
Table 20: Count of portlet modes	41
Table 21: Number of portlet modes per visitor	41
Table 22: Classification Mapping	42
Table 23: Scorer for as-is situation.....	43
Table 24: Accuracy statistics as-is for whole population	44
Table 25: Accuracy statistics as-is for targeted population	44
Table 26: Bucket analysis	45
Table 27: Relevant variables in VW_VisitorProfileTable	45

Table 28: Split on age 33 for people who live in an area where houses are bought and do not have a Rabobank mortgage.....	49
Table 29: Scorer of live validation.....	50
Table 30: Differences between old and new situation for targeted visitors.....	50
Table 31: Conversion and FAQ.....	54
Table 32: T-test conversion.....	54
Table 33: Sweet spot for conversion.....	54
Table 34: Sweet spot with visitors who saw at least one question.....	54
Table 35: Cluster algorithms.....	55
Table 36: Cluster 0 - house owner (top 6).....	56
Table 37: Cluster 1 – starter (top 6).....	56
Table 38: Scorer for FAQ section.....	58
Table 39: Accuracy matrix for FAQ section.....	58
Table 40: Proposal for CCF.....	58
Table 41: Proposal for CCF (only profile).....	58
Table 42: Statistics of data tables.....	69
Table 43: Overview of questions and shortcodes.....	70
Table 44: Cluster 0 - house owner (top 6) – k-means.....	74
Table 45: Cluster 1 – starter (top 6) – k-means.....	74

LIST OF FIGURES

Figure 1: Flow of generating insights	12
Figure 2: Framework	13
Figure 3: Available data of a visitor	15
Figure 4: Website of MyEnergy - Pricing page	15
Figure 5: Flow of creating clusters	17
Figure 6: Classification model (Tan, Steinbach, & Kumar, 2006)	21
Figure 7: Area of process mining (van der Aalst, 2011)	25
Figure 8: Process model obtained from log file MyEnergy	27
Figure 9: Process model of web behavior	27
Figure 10: MyEnergy Bucket Process Model	28
Figure 11: Behavior on page - MyEnergy	29
Figure 12: Screenshot RHD.....	30
Figure 13: Process view of visitor	31
Figure 14: Process view of bank	31
Figure 15: Flow of connecting profile information	35
Figure 16: Mortgage homepage.....	37
Figure 17: Pages and variables	38
Figure 18: Classification model	39
Figure 19: Mined decision tree for classification BT Home with eigendom_cd	47
Figure 20: Mined decision tree for classification BT Home with eig_hs_ind	48
Figure 21: Time line of data	49
Figure 22: Front page of FAQ	51
Figure 23: View and event.....	52
Figure 24: Views in the FAQ	52
Figure 25: Classification engine predicting cluster	56
Figure 26: Decision tree on clusters	57
Figure 27: Behavior on repayment, savings, donations and change interest rate.....	59
Figure 28: Behavior on “Can I get a mortgage in my situation?”	60

Figure 29: Behavior on “How does the mortgage interest tax deduction works?”	61
Figure 30: Behavior on “Which issues are involved when selling my house?”	62
Figure 31: Behavior on “What is the current state of the housing market?	63
Figure 32: ETL NetInsight	68
Figure 33: Make an appointment.....	71
Figure 34: RHD Start.....	71
Figure 35: RHD Make an appointment.....	72
Figure 36: Decision tree for FAQ only on profile table.....	73
Figure 37: Data preparation phase	75
Figure 38: Clustering FAQ.....	75
Figure 39: Classification FAQ.....	76
Figure 40: Can I get a mortgage in my situation - visualization with other pages in mortgage section	77

TERMINOLOGY

Some business terminology is elaborated in Table 1.

Table 1: Terminology

Term	Description
Behavioral Targeting	Technique to select relevant advertisements on web sites based on web browsing behavior
Conversion rate	Percentage of visitors that reach a desired point
Customer	Client of the Rabobank
Customer Journey	All experiences that a customer have with a company before buying a product
Data mining	Knowledge discovery in databases
Doorstomer	Person who want to move from one house to another house
IBM Interact	Current Behavioral Targeting engine
IBM Netinsight	Current web analytics software
Lamellae	Part on a web page that can be opened and closed through a button (click)
Mortgage	Loan to buy a house with
Ophoger	People who want to raise their mortgage loan
Oversluter	People who want to transfer their mortgage to the Rabobank
Process mining	Extracting information out of event logs
Starter	People who are orienting to buy their first house
View	Page view, belonging to a visit and occurring at only one page
Visit	Session on the website, consisting of one or more views.
Visitor	Unique browser that is used to visit the website. A visitor has one or more visits.

ABBREVIATIONS

An overview of the abbreviations is shown in Table 2.

Table 2: Abbreviations

Abbreviation	Description
BT	Behavioral Targeting
CCF	Conditional Content Filter
FAQ	Frequent Asked Questions
NN	Nearest Neighbor search
RHD	Rabobank Hypotheek Dossier ("Rabobank Mortgage File Application")
UX	User Experience Center

PREFACE

This master thesis contains my final work to complete the study Business Information Systems at the University of Technology in Eindhoven. The last six months I worked at this thesis at Rabobank Nederland and provided a methodology to become more relevant online using data mining and process mining techniques.

I would like to thank my supervisor Boudewijn van Dongen, who always had time to discuss my results. He gave me inspiration to look at new possibilities for solving problems and generating insights. I would also like to thank Tim van den Berg and Frank van Geffen for their time and valuable comments on the process and report.

MANAGEMENT SUMMARY

In this thesis, a methodology is developed to become more relevant on websites promoting complex products. The thesis provides a method to develop a model that predicts the information need of an individual website user. Such a model can be used to target content that has the best fit with the information need.

Data mining and process mining play an important role in this methodology. The thesis provides a method to segment users into groups with a similar information need, and create a model to predict these information needs.

The classification engine of Rabobank is assessed in a case study. Through data mining, a proposal is given to improve the current classification rules and these improvements are validated in a field test.

In a second case study, different information needs in the FAQ of the mortgage sections are identified and classification rules are mined to predict the information need of an individual user. A clustering algorithm obtained two clusters: a cluster with questions for people who own a house and a cluster with questions for people who not own a house.

The developed methodology is suitable to support webmasters in designing their *conditional content pages*. The content on such pages depends on the profile of a visitor. The thesis also provides methods to assess the performance of the classification engine. Besides, this thesis suggests how to balance between recall and precision and shows the effect of a wrong and correct classification on a conditional content page.

In order to design meaningful web pages, we have also researched how to analyze behavior on pages with process mining. The click logs of a website and appended profile information of visitors are used to generate these insights.

It is worth investigating how process mining is an addition to current User Experience research. Process mining is able to visualize the behavior on a page.

1 INTRODUCTION

Much research is conducted how to design a web page that is suitable and relevant to visitors. Many articles are written about content design, page design and ease of use navigation design (Nielsen, 1999). Eye tracking is also a method that is used to explore web page viewing behavior and web page design. A study in 2004 shows that gender, the viewing order of a web page and the interaction with the page influences oracular behavior (Pan, Hembrooke, Gay, Granka, Feusner, & Newman, 2004). Another article considers the web as a cognitive landscape to obtain a guideline for effective page design (Rosen & Purinton, 2004). This thesis focusses on designing web pages based on the behavior of individual visitors on a page. This behavior is extracted from click logs.

In this thesis, a methodology is developed to configure a website using data and process mining techniques in such a way that it becomes more relevant to the end user. The area of research is a website, promoting a complex financial product. The product itself does not differ from competitors products in terms of properties, but does so in terms of service around this product. For example, a banking company provides mortgages. The mortgage itself is a loan of money, but services around this mortgage differ between banks. Another property of this product is that the product is not a simple, small, product which you easily buy from a store. Before buying this product, the consumer needs to understand the product and several documents must be checked and several forms are needed to be filled in. This makes the product more *complex*.

1.1 PROBLEM DESCRIPTION

Nowadays, many companies record online behavior of web site visitors. This information is saved in the web server log files. The log files contain the surfing behavior of their visitors and in some cases, additional data is available. This information can be used to become more relevant to the end user. A flow of this process is depicted in Figure 1.

This thesis investigates how to use this data to configure a website in the area of complex products that is more relevant to the end user. A complex financial product differs from regular (small) products. The main differences are shown in Table 3. *The customer journey*¹ is longer for complex products, consists of multiple visits and has a longer throughput time.

Table 3: Regular versus complex customer journey for purchasing one product

Regular web shop / recommender system	Website promoting complex products
Short throughput time	Long throughput time
One or a few visits	Multiple visits
Simple product	Complex product
Low need for additional information before buying	High need for additional information before buying
Unique product	Commodity product
Buying frequency is high	Buying frequency is low

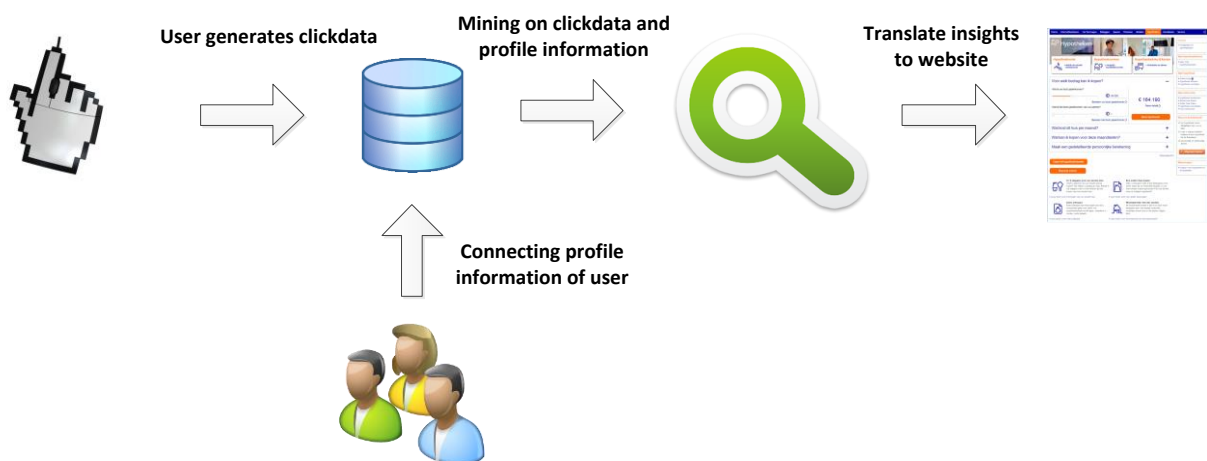


Figure 1: Flow of generating insights

¹ All experiences that a customer have with a company before buying a product

1.2 RESEARCH QUESTIONS

The research questions addressed in this thesis are elaborated in this section.

How can data mining and process mining be applied to become more relevant to the end user on websites promoting complex products?

1. How can we identify different information needs on a website?
2. How can we identify which visitor needs which information?
3. How do you present this information on the website?

An information need is described as a set of pages, ranked by *relevance* for the visitor. Through the first question, we try to obtain different information needs for groups of visitors. In a group the information needs are similar, but the information needs between groups differ. In the second question, we develop a method to train a model that predicts the information need of a visitor based on the information that is available of the visitor. In addition to the first two questions, we also investigate what the best way is to present information on a website. We use the behavior of the visitor on the web site for this analysis.

In two case studies performed at Rabobank, we test the developed methodology. Also a field test is executed. The Cross-Industry Standard Process for Data Mining is followed to obtain the results in this case study (Chapman, Clinton, & Kerber, 2000).

1.3 FRAMEWORK

This project follows the framework shown in Figure 2. In the preparation phase, we conduct the business understanding and data understanding parts. When these two steps are completed, two case studies are conducted. In the first case study we try to improve the current classification engine on the website of Rabobank. Only the methodology of research question 2 is used in this case study. In case study 2, we try to identify different information needs, classify visitors in one of these information needs and perform an analysis on how to present information on the website. The methodology of all three research questions are used in this second case study. After the two case studies are conducted, we draw conclusions and give recommendations.

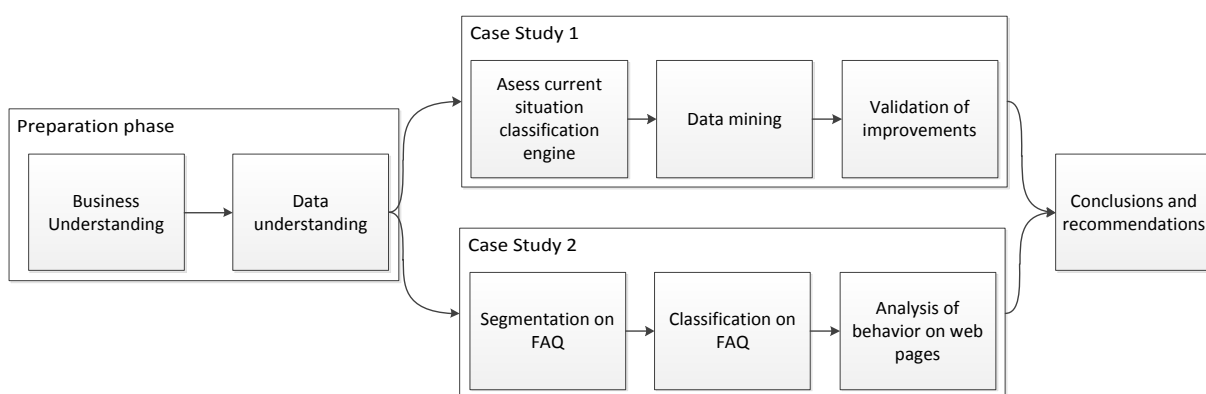


Figure 2: Framework

1.4 RUNNING EXAMPLE

We use a running example for better understanding of the developed methodology. Our running example is a fictive company that provides electricity for consumers. We call the company *MyEnergy*. MyEnergy provides an energy service to customers. Although they have a call center, MyEnergy is a company with a focus on online services. MyEnergy delivers energy, but they do this in different forms, either grey energy or green energy. The website of MyEnergy consists of both an information section on grey and green energy. It also allows a customer to log in into their personal environment. People can get online advice through a calculator on the website and search through frequent asked questions. A screenshot of the website is shown in Figure 4, where a small section of the web page is shown.



MyEnergy tracks the online *journey* of the visitor. The company uses a logging system that records behavior information of a user after a cookie has been accepted. It is also possible to add more information about the customer, such as demographic and profile information. This data is connected to a cookie after logging into the secure area of MyEnergy. An overview of these data sources is depicted in Figure 3. MyEnergy aims to increase their sales on the website and to become more relevant to the website visitor.

MyEnergy also wants to optimize the *customer journey*. A customer journey is the set of all the interactions a customer has with the company before buying a product. MyEnergy aims to make the customer journey better in terms of quality and time. A part of this customer journey takes place on the website. Therefore, making the web site more relevant also improves the customer journey.

MyEnergy wants to increase the *conversion rate* on the website. A *conversion page* is the page or action where we want a visitor to end, such as submitting a form or applying for a service. The conversion rate is the percentage of the visitors that reach this page. In the screenshot of Figure 4, the conversion rate is the percentage of people that send the contact form.

The behavior on a web page is also logged. In Figure 4, two *lamellae* are shown. Lamellae are elements on a page that can be expanded and collapsed. More information is hidden in lamellae. The content of the lamellae is made visible by clicking on the "plus symbol". The expanding and collapsing actions are recorded in the event log. Filling in the contact form and sending the contact form are also actions that are recorded in the event logs of MyEnergy.

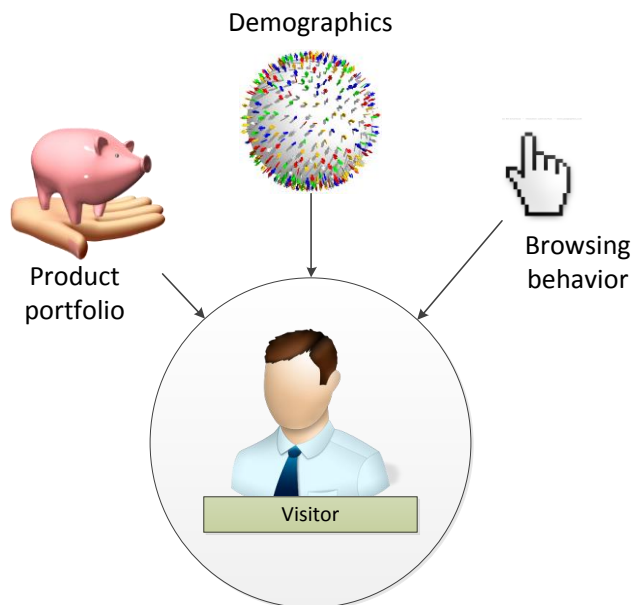


Figure 3: Available data of a visitor

The screenshot shows the 'My Energy' website's pricing page. At the top left is the 'My Energy' logo. To the right is a search bar with the text 'zoeken' and a magnifying glass icon. Below the logo is a navigation menu with links: Home, Prices, Carbon Footprint, Green Energy, Grey Energy, and Application. The main content area is divided into two columns. The left column is titled 'Prices' and contains two articles:

- One love, one price:** Dated 06 feb 2015. Text: 'We have one price - All of our customers are always on our latest best price – no matter when they joined us and regardless of how they pay (including customers on pre-payment meters)'. Includes a plus icon.
- Check our prices:** Dated 06 feb 2015. Text: 'It's really easy to get a price estimate for our Green Electricity and Green Gas – just enter your postcode below.'. Includes a plus icon.

 The right column is titled 'Contact us now' and contains a form with the text 'Receive new information about MyEnergy'. The form has two input fields labeled 'Name' and 'E-mail', and a 'Send form' button.

Figure 4: Website of MyEnergy - Pricing page

2 HOW CAN WE IDENTIFY DIFFERENT INFORMATION NEEDS ON A WEBSITE?

Different visitors have different information needs on the website. In the case of our running example, one can imagine that some visitors are environmentally aware and other visitors are not. We obtain different sets of visitors that have the same interests and the same information needs. This is also called user-to-user clustering. We assume that the click behavior of a visitor represents the information need of a visitor.

2.1 RELATED WORK

In literature, several approaches are described to be more relevant online. Collaborative filtering (Ricci, Rokach, & Shapira, 2011) is a method of making predictions about the interests of a user by collecting preferences of many users. Collaborative filtering can be performed item-based or user-based. Item-based filtering finds rules that predict only on the items that are shown: users who bought product 1, are also interested in product 2. These rules are obtained with use of an item-item matrix. The second technique is user-based filtering (also known as people-to-people correlation) (Ricci, Rokach, & Shapira, 2011). User-based filtering search for users that are similar to predict their information needs. The rating patterns are taking into account to predict which users are similar.

A problem with collaborative filtering is the cold start problem (Ricci, Rokach, & Shapira, 2011). The cold start problem is the problem that insufficient data is available for new items or new users. This collaborative filtering method recommends items or pages to other users based on the past preferences. The cold start problem is bypassed by taking only users who have visited enough pages to make a reliable prediction. This enables the collaborative filtering algorithm to capture their preferences more accurate and thus provide reliable recommendations.

To find similar user, we perform a similarity search. Nearest Neighbor Search is used to perform this similarity search (Ricci, Rokach, & Shapira, 2011). The nearest neighbor search problem is defined as follows: given a set V of points in a space W and a query point, find the closest point in V to the query point. The distance function is defined as Euclidian distance, Manhattan distance or other distance function. The Euclidian distance is defined as $d(p, q) = d(q, p) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$ where p and q are vectors.

Through this user-based filtering, we obtain the information needs of individual users. What we want is to cluster these information needs to one or more clusters that show the *different* kinds of information need. We need a form of unsupervised learning (Manning, Raghavan, & Schütze, 2008). Unsupervised learning is the problem to find hidden structure in unlabeled data. Cluster analysis is a method to group items that are similar. The similarity in a cluster is high, while the similarity between clusters is low.

Several clustering techniques are available, which can be divided in two methods: *hard clustering* and *soft clustering* (Manning, Raghavan, & Schütze, 2008). In hard clustering, the data is divided into distinct clusters where each data element belongs to exactly one cluster. In soft clustering, data elements often belong to more than one cluster. An example of hard clustering is k-means clustering. The aim of this algorithm is to divide the observations into k clusters. A new observation is added to the cluster with the nearest mean. In K-medoids clustering, the center of the cluster must be a member of the data set.

In soft clustering (also known as fuzzy clustering), an element does not belong to exactly one cluster, but has a membership in one or more clusters. The FCM algorithm (Bezdek, Ehrlich, & Full, 1984) is a wide used algorithm to perform fuzzy clustering and is the *soft* variant of k-means clustering. After soft clustering, we are able to obtain the winner cluster. The user has the highest membership in the winning cluster.

In practice, one must obtain the best cluster method by trial and error (Estivill-Vastro, 2002). One tries to obtain the best cluster in terms of relevance and number of clusters. This trial and error method is solution

oriented. Relevance is how useful the clusters are to implement on the web site: do the clusters clearly indicate a different information need. By vary the number of clusters, changing between *k-medoids* and *k-means* and allowing fuzzy cluster assignment, we obtain the best cluster method and the best clusters.

2.2 OVERVIEW

In this chapter, we describe the steps that are used to obtain these clusters. We follow the following steps (also displayed in Figure 5):

1. Data filtering
2. Searching Nearest Neighbors
3. Apply segmentation on the set of visitors
4. Obtain the information needs in the obtained clusters

After each step, we obtain an intermediate data set that is used in the next step. After the data filtering step, we obtain a filtered set where only the visitors are included that saw at least a number of questions. For each user in this set, the nearest neighbors are obtained to construct the information need of a visitor. The output of this nearest neighbor search is a prepared data set with the information need of each visitor. This data set is used for segmentation. The output of this segmentation step is a set with visitors and their assigned cluster. This set is used to obtain a set of pages for each cluster, ranked by relevance. These rankings are used to assign a visitor into a cluster based on web data. These steps are further explained in the next sections.

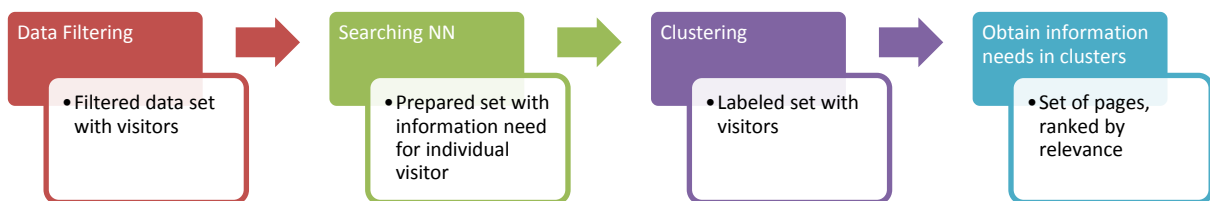


Figure 5: Flow of creating clusters

2.3 OBTAIN THE INFORMATION NEED OF ONE VISITOR

The website of *MyEnergy* consists of multiple pages that contain different information and provides answers on a variety of questions. The website of *MyEnergy* provides information pages with information about the application process (page 1), their Carbon Footprint (page 2), grey (page 3) or green (page 5) energy and a page with their prices (page 4).

The behavior of the visitors on the website is recorded in a log file. Initial log data is recorded in a format that is displayed in Table 4. The data is transformed in such a way that the visitors are listed with the pages that they have seen, as is shown in Table 5. Note that *time* and *viewid* are no required fields for this analysis.

Table 4: Example of log data of a website

Viewid	Visitor	Page	Time
000	1001	Page 1	01/09/14 8:01:01
001	1002	Page 5	01/09/14 8:02:42
002	1002	Page 1	01/09/14 8:03:26
003	1003	Page 4	01/09/14 8:03:33
004	1001	Page 2	01/09/14 8:04:04
005	1003	Page 3	01/09/14 8:04:07
006	1003	Page 1	01/09/14 8:05:07
007	1001	Page 5	01/09/14 8:06:10
008	1002	Page 5	01/09/14 8:07:11
009	1004	Page 3	01/09/14 8:15:11
010	1004	Page 4	01/09/14 8:16:33
011	1004	Page 3	01/09/14 8:16:41
012	1004	Page 1	01/09/14 8:18:45

Table 5: Visitors and their questions

Visitor	Page 1	Page 2	Page 3	Page 4	Page 5
1001	1	1	0	0	1
1002	1	0	0	0	2
1003	1	0	1	1	0
1004	1	0	2	1	0

2.3.1 DATA FILTERING

When we cluster visitors, a problem occurs due to the fact that some pages are visited very frequent and some pages are viewed less often. When some pages are popular and all other pages are unpopular, the problem arises that the information needs of the visitors are clustered in visitors who saw popular pages and visitors who saw unpopular pages. We consider this effect as an undesired effect.

To solve this issue, we apply filtering to obtain a more reliable idea of the information need of a visitor. We only keep the visitors in our data set that have enough views in the section to determine their information needs. This cut-off point is determined by trial and error, but one must use at least 3 or 4 views. We try to obtain clusters that are relevant and this cut-off point is one of the parameters we can tune. The method is therefore solution-oriented. Relevant clusters are clusters that clearly indicate different information needs. We have too little information to obtain the information need of a visitor if we do not have more than 2 page views belonging to that visitor.

2.3.2 NEAREST NEIGHBOR SEARCH

Another method to obtain the information need of a visitor, is looking at other visitors who have seen approximately the same pages. We apply Nearest Neighbor search with the *Euclidian distance* as distance function. In our running example, the space is a 5-dimensional space with each page as a dimension. Visitor 1001 has a distance of $\sqrt{2}$ with visitor 1002, a distance of 2 with visitor 1003 and a distance of $\sqrt{7}$ with visitor 1004. Therefore, visitor 1001 and 1002 are nearest neighbors. When we take the average of the page views of the visitors and the nearest neighbors of this visitor, we obtain the expected page views for each page for that visitor. The data of our running example is shown in Table 6, where the expected number of page views is displayed.

Table 6: Expected views of the pages

Visitor	NN	E[Page 1]	E[Page 2]	E[Page 3]	E[Page 4]	E[Page 5]
1001	1002	1	0,5	0	0	1,5
1002	1001	1	0,5	0	0	1,5
1003	1004	1	0	1,5	1	0
1004	1003	1	0	1,5	1	0

For example: page 2 (information about the carbon footprint) could also be interesting for visitor 1002, although the visitor has not seen that page in its original visit. When we have a larger data set, we increase the number of Nearest Neighbors.

2.4 APPLY SEGMENTATION ON THE SET OF VISITORS

After determining the information need of an individual visitor, we apply clustering. The goal of this clustering is to identify groups of visitors who have similar information needs. In our running example, we obtain two clusters: *cluster 0* with visitors 1001 and 1002 and *cluster 1* with visitors 1003 and 1004. By taking the average of the expected page views per cluster, we obtain the expected page view for a visitor in that cluster. With this information, we make a ranking of pages for each cluster. In our running example, the top 3 pages in Cluster 0 is {Page 5, Page 1, Page 2} and the top 3 pages in Cluster 1 is {Page 3, Page 4, Page 1}. The rankings are shown in Table 7.

Since the data set of our running example is small and the clusters are distinct, we use a hard clustering algorithm. When the data set becomes larger, it is possible that visitors appear who do not fit completely into one of the clusters. A *soft clustering* algorithm could perform better. The best cluster and clustering algorithm is obtained through trial and error. We try to find the most relevant clusters: the clusters that clearly show a different information need. In our running example, cluster 0 and cluster 1 clearly show a different information need: a cluster with “Green” information and a cluster with “Grey” information.

Table 7: Ranking of clusters

PageID	Page title	Cluster 0 (“Green”)	Cluster 1 (“Grey”)
Page 1	Information about application	1	1
Page 2	Our carbon footprint	0,5	0
Page 3	Grey energy	0	1,5
Page 4	Prices of energy	0	1
Page 5	Green energy	1,5	0

2.5 ASSIGN VISITOR INTO A CLUSTER BASED ON WEB DATA

With the obtained clusters, we are able to assign a visitor to a cluster based on the log of the web data. In our running example, we classify new visitors 1005 (Table 8) and 1006 (Table 9) into cluster 0 “Green” or cluster 1 “Grey”. For each page that a visitor saw, we obtain the expected value of the view in both clusters. We see that the most appropriate cluster for visitor 1005 is cluster 0 and the most appropriate cluster for visitor 1006 is cluster 1.

However, we want to classify the visitor before his/her visit and not after the visit has taken place. In the next section we develop a method to identify which visitor needs which information based on other data than web log data from this section.

Table 8: Visitor 1005

<i>Visitor 1005</i>	Expected view “Green”	Expected view “Grey”
Page 1 (application)	1	1
Page 2 (carbon footprint)	0,5	0
Page 5 (green energy)	1,5	0
Average	1	0,33

Table 9: Visitor 1006

<i>Visitor 1006</i>	Expected view “Green”	Expected view “Grey”
Page 1 (application)	1	1
Page 3 (grey energy)	0	1,5
Page 4 (prices)	0	1
Average	0,33	1,17

3 HOW CAN WE IDENTIFY WHICH VISITOR NEEDS WHICH INFORMATION?

In this section, we train a model to classify visitors to one of these information needs. We have already identified the information needs of the visitors. The classification is done with the use of various information, such as click data or other profile information. Through this model, we are able to provide the visitor with a more relevant web site. In our running example, we use the political preference, age and gender to train our model with.

3.1 RELATED WORK

3.1.1 CONFIGURATION

Identifying which visitor needs which information is a classification problem. It is, contrary to segmentation, a form of supervised learning. The different categories are defined through segmentation or clustering or are pre-defined by a domain expert. A classification model uses an attribute set x to predict the category or class label y (Tan, Steinbach, & Kumar, 2006). A number of classification algorithms (classifiers) are mentioned in Crisp-DM (Chapman, Clinton, & Kerber, 2000), such as discriminant analysis, rule induction methods, decision tree learning, neural networks, k-Nearest Neighbor, case-based reasoning and genetic algorithms.

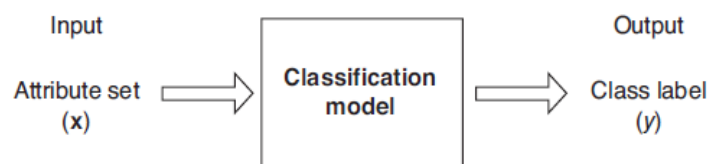


Figure 6: Classification model (Tan, Steinbach, & Kumar, 2006)

Not all classification algorithms suit to this classification problem. One must check the requirements of the environment to choose the most suitable classifier. The requirements that must be evaluated are the data requirements and software requirements in the systems.

A decision tree learner is a classifier that is commonly used in data mining. The decision tree learner predicts a variable (dependent variable) with a set of independent attributes. The C4.5 algorithm (Quinlan, 1993) chooses the attribute that most effectively split the set into two subsets. The criterion that is used for the splitting of the set is information gain. Information gain is a method to express the difference between two groups. The attribute with the highest information gain is used to create a split (Tan, Steinbach, & Kumar, 2006). For each subset, the same procedure is performed.

A method to avoid overfitting is *pruning* the decision tree. The decision tree is pruned with the MDL approach (Mehta, Rissanen, & Agrawal, 1995). Pruning is done with two reasons: reduce the complexity of the decision tree (remove unnecessary nodes) and avoid overfitting

Artificial neural networks and support vector machines are also widely used classifiers. Neural networks are presented as a system of connected neurons that are able to train themselves on a set of data (Zhang, 2000). Support vector machines (Hearst, Dumais, Osman, & Platt, 1998) are also supervised learning methods that recognize complex patterns in data. Although the predictive power is high, the parameters and model are hard to interpret and understand.

3.1.2 EVALUATION

After building a classification model, we need a method to assess the quality of this model. Precision, recall, accuracy and confusion matrices are commonly used methods to evaluate these classification models (Tan, Steinbach, & Kumar, 2006). A prediction can generate a True Positive (TP), True Negative (TN), False Positive (FP) or a False Negative (FN). Is it only possible to calculate accuracy throughout the whole population. Precision and recall can be calculated for each segment. In the ideal situation, one would have 100% recall and 100% precision, which would result in 100% accuracy. In practice, there is a tradeoff between recall and precision (Jibza, 2007). More details are shown in Table 10.

Table 10: Explanation of terms

Prediction / result	Cluster 0	Cluster 1	
Cluster 0	True Positive (TP)	False Positive (FP)	Precision = $(TP / (TP+FP))$
Cluster 1	False Negative (FN)	True Negative (TN)	
	Recall = $TP / (TP + FN)$		Accuracy = $(TP+TN) / (TP+TN+FP+FN)$

To test if a new model works better than an old model, a random experiment must be conducted (Montgomery & Runger, 2007). An experiment that can result in different outcomes, even though it is repeated in the same manner every time, is called a random experiment. The current version is considered as the h_0 hypothesis and the new version is considered as the h_1 hypothesis. Half of the users get the current version and half of the users get the new version. After collecting the data, a two-sample hypothesis test is performed. We see if the new variant performs better in terms of accuracy, precision and recall.

3.2 TRAIN A MODEL

For the classification, we want to have a model that classifies quickly and is understandable for the stakeholders. We want to have a model that works with training and test data and creates classification rules that are mutually exclusive. A number of classification algorithms are available (Chapman, Clinton, & Kerber, 2000). For this classification, we use the decision tree miner because of the following reasons:

- ✓ The created decision rules are expressed as (Boolean) expression
- ✓ The expressions are complete and mutually exclusive
- ✓ The derived expressions are easy to implement
- ✓ The decision rules make decisions very quick
- ✓ Decision trees are easy to understand

We extend the visitor information of our running example with profile information, shown in Table 11. We train a decision tree on our running example.

3.2.1 SETUP OF MODEL TRAINING

When training a classification model, we must avoid overfitting. Overfitting is the issue of having a model that describes the random error or noise in the data instead of the underlying relationship. Therefore, the model only performs well on the training set. To avoid overfitting, we apply cross-validation and pruning. Both techniques are implemented in our test design.

For the Decision Tree model building, we made use of the C4.5 algorithm (Quinlan, 1993). This algorithm is a statistical classifier and generates a decision tree. The pseudo-code is:

1. Check for base cases
2. For each attribute a
 - a. Find the normalized information gain ratio from splitting on a
3. Let a_{best} be the attribute with the highest normalized information gain
4. Create a decision node that splits on a_{best}
5. Recur on the sublists obtained by splitting on a_{best} , and add those nodes as children of node

- **Information gain ratio** is used as the quality measure.
- The decision tree is also pruned with the **MDL approach**. We want to prune the tree because of two reasons: reduce the complexity of the decision tree (remove unnecessary nodes) and avoid overfitting.
- We set a **minimum node size**. The nodes must be large enough to prevent the overfitting of the model. This also reduces the complexity of the model.

In our running example, we want to have leafs with at least two items. When the training set becomes larger (with thousands of items), we want to have leafs with at least a few hundred items.

3.2.2 RUN DECISION TREE MINER

Table 11: Visitors with extended information MyEnergy

Visitorid	Cluster	Age	Gender	Political preference
1001	Cluster 0	23	Men	Green party
1002	Cluster 0	25	Woman	Green party
1003	Cluster 1	51	Men	Conservative party
1004	Cluster 1	51	Woman	Conservative party

We train a decision tree to predict the cluster of a visitor. A decision tree trained on this data split on the variable *Political Preference* or *Age*. The age 38 is a decent split point between cluster 0 and cluster 1, even as the difference in people who have the *Green Party* or the *Conservative party* as their political preference. With this model, we decide in which category the user falls and which content is relevant to the user.

Possible ruleset (1)

Cluster 0 "Green": Age < 38
Cluster 1 "Grey" : Age >= 38

Possible ruleset (2)

Cluster 0 "Green": Political Preference = "Green Party"
Cluster 1 "Grey" : Political Preference = "Conservative party"

3.3 ASSESS MODEL

After building or implementing the model, we want to have a method to judge if the model works and what the quality of the model is. To assess the performance of a prediction, accuracy, precision and recall are used. A prediction can generate a True Positive (TP), True Negative (TN), False Positive (FP) and a False Negative (FN). In our running example, classify every visitor in Cluster 1 would give a recall of 100% for Cluster 1, but a low precision (50%). Precision can also be seen as relevance, where recall can be seen as the reach.

$$Accuracy = \frac{(TP + TN)}{P + N}$$

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

4 HOW DO YOU PRESENT THIS INFORMATION ON A WEBSITE?

With the use of process mining, the behavior of a visitor on the website can be visualized. A distinction can be made between the “highways” in the process and the routes that are taken with a lower frequency. Process mining is usually conducted in the area of Business Process Management, but is, with some adaptations, also suitable for customer journey visualizations. We introduce process mining in the *related work* section.

4.1 RELATED WORK

Process mining techniques allow for extracting information from *event logs* (van der Aalst, 2011). There are three types of process mining tasks: discovery, conformance and enhancement (Figure 7) (van der Aalst, 2011). Process discovery is used to construct a process model from event logs. Web behavior of users is also recorded into event logs. With process mining, we are able to generate process models of the online behavior on the web.

A process mining algorithm is a function that maps a log file into a *marked petri net*. Several process mining algorithms are developed, such as the alpha-algorithm, genetic algorithms and fuzzy algorithms.

Real life processes are less structured than an ideal perception of the processes. A miner that deals well with these unstructured processes is the fuzzy miner. The fuzzy miner is configurable to display only the most frequent paths in the process and hide the less frequent paths (Gunther & van der Aalst, 2007).

Even when the fuzzy miner is used for process discovery, one could obtain “spaghetti” processes. These processes are unstructured and it is hard to obtain any useful information out of these processes. Filtering of activities is a method to deal with this structure (van der Aalst, 2011). When applying filtering, only activities are included when these activities have a high frequency of occurring. Another method is dividing the complete process model in smaller processes (van der Aalst, 2011). Smaller processes have fewer activities. This results in simpler processes. A guideline for process models proposes that a maximum of 50 elements are allowed to come to a comprehensible model (Mending, Reijers, & van der Aalst, 2009).

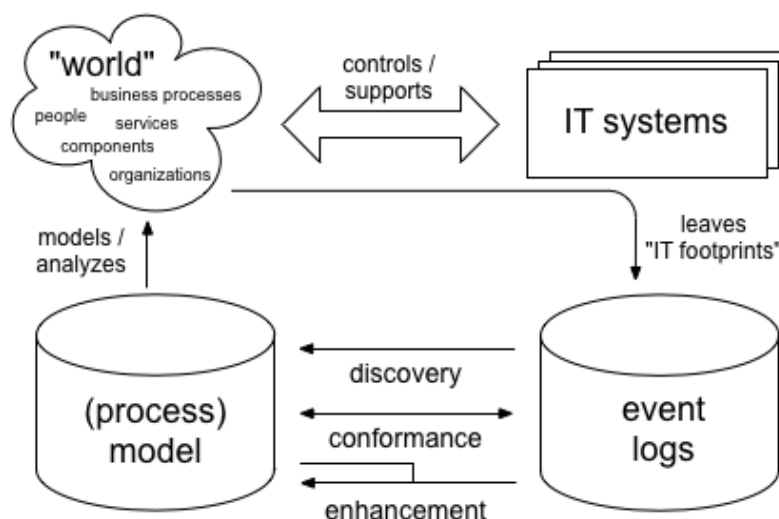


Figure 7: Area of process mining (van der Aalst, 2011)

4.2 PROCESS MINING ON WEBDATA

Visitors on a website are free to do whatever they want and are not obligatory to follow a predefined route. A *Customer Journey Manager*, however, has a desire line of the route to a conversion. With process mining, one can see to what extent a journey is followed. Due to the large amount of web pages and the large freedom of visitors on the web site, one should prepare the web data in such a way that compact and comprehensible process models can be made. To mine an understandable model, a fuzzy miner can be used. The fuzzy miner can deal with the complexity of unstructured processes and can be configured how much detail must be showed (Gunther & van der Aalst, 2007). An improved version of the Fuzzy Miner is implemented in the software tool Fluxicon Disco (Rozinat, 2013), that is suitable for this analysis.

If a model contains more than 50 elements, hierarchy can be used or the scoping of the model must be adapted (according to (Mendling, Reijers, & van der Aalst, 2009)). To mine understandable models, we can reduce the number of events by grouping pages into *buckets* or *categories*. Another method is adjusting the scoping of our mining project. Process mining can also be used to check the behavior of the visitor within a single page, when *event tracking* on the website is active.

4.3 DATA PREPARATION FOR PROCESS MINING

Process mining uses traces of individual visitors to create process maps. The data that can be used for this, have a case ID and a number of events ordered in time (van der Aalst, 2011). When we look at our log file from MyEnergy (Table 4), we can map our *Visitorid* as *CaseID*. We also have an event name, which is the name or identifier of the page. The ordering of the events can be obtained by sorting the log file by the time stamp. Note that we only use the time stamp to determine the sequence of the views. The process map obtained from this logging, is shown in Figure 8. One sees clearly the entry pages, exit pages and in which order the visitors visit the pages. The process map is annotated with an explanation of the different elements.

The entry pages are the pages that have an incoming dotted grey arrow from the *start* event at the top: the green circle with the triangle inside. The exit pages have a dotted grey *outgoing* arc from the page to the stop event: the red circle with the square inside. The arrows between the pages (blue rectangles) are the transitions between the current and the next page. The size of the arrows represents the frequency of the transition. The blue color of the rectangle represents the frequency of the transition: dark blue is a high frequent activity, where a light blue transition occurs less frequent.

The process map of our event log looks very simple because of the small amount of pages. However, if the number of pages increases, the complexity of our mined model also increases. We see an example process map in Figure 9. It is hard to obtain useful information out of this figure. Therefore, we apply *bucketing* to create clearer process models and apply process mining to assess the behavior on pages. These techniques are clarified in the next two sections.

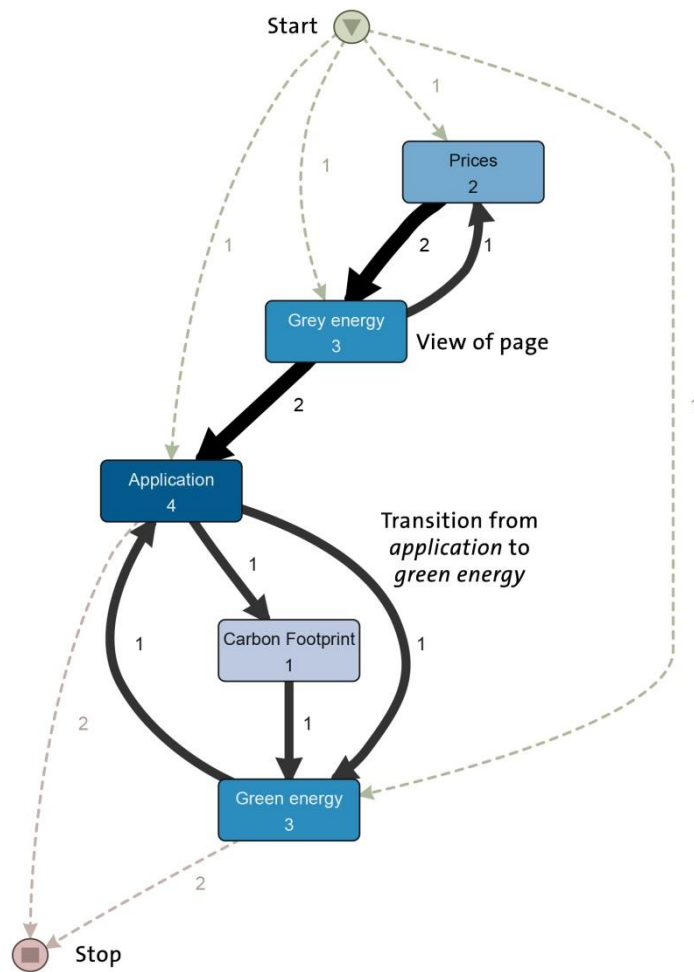


Figure 8: Process model obtained from log file MyEnergy

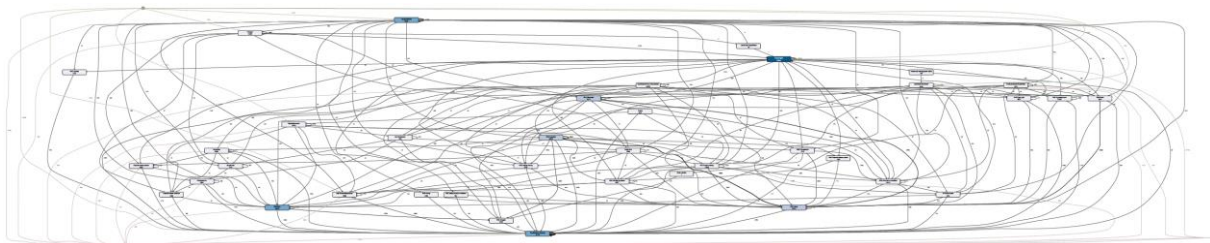


Figure 9: Process model of web behavior

4.4 PROCESS MINING WITH BUCKETS

We categorize the pages into a number of categories to keep the number of unique events small. In our running example, we create three categories and put the pages in one of these categories (Table 12).

Table 12: Buckets of MyEnergy

PageID	Page title	Category
Page 1	Information about application	General information
Page 2	Our Carbon Footprint	General information
Page 3	Grey energy	Energy information
Page 4	Prices of energy	Prices
Page 5	Green energy	Energy information

The process model that is obtained from the log file is shown in Figure 10. We see that most visitors of “energy information” are coming from the general information pages or start their visit in this section. All visitors of the page “prices” continue on the website and go to the energy information. We also see that most visitors end their visit in the “energy information” or “general information” section.

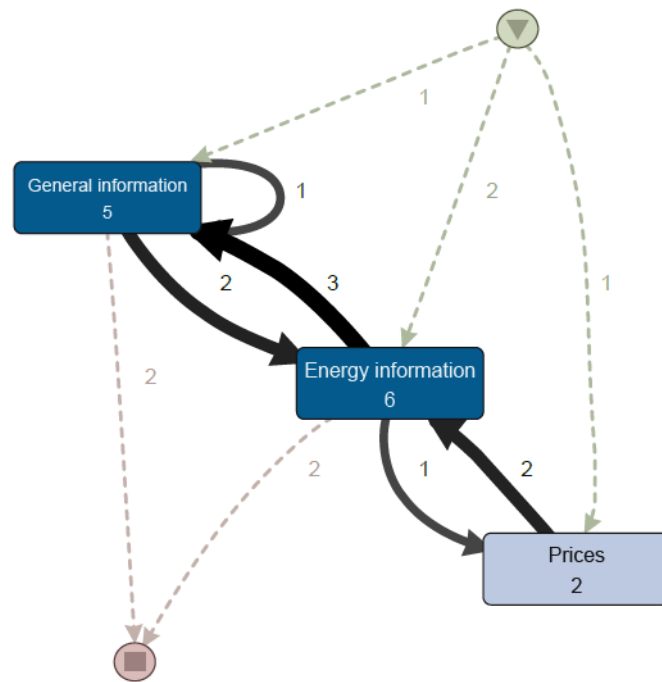


Figure 10: MyEnergy Bucket Process Model

4.5 BEHAVIOR ON A PAGE

On the website of *MyEnergy*, events that occur on a page are also logged. The opening of a contact window, the opening and closing of a tab or lamellae and the actions in a form are also recorded in a log file. This log is a separate log, but is coupled to the original log through *viewid*. This log file is shown in Table 13.

Table 13: Events on a page

Eventid	Viewid	Page	Action	Time
001	003	Page 4	Open lamellae 1	01/09/14 8:03:35
002	003	Page 4	Open lamellae 2	01/09/14 8:03:38
003	003	Page 4	Open contact form	01/09/14 8:03:55
004	010	Page 4	Open lamellae 2	01/09/14 8:16:58
005	010	Page 4	Open lamellae 1	01/09/14 8:17:02
006	010	Page 4	Open contact form	01/09/14 8:17:22
007	010	Page 4	Send contact form	01/09/14 8:18:33

We extend this logging with the *view* of the page. Only the information about the events occurring on *page 4* is included in this logging. The result of appending the original log (Table 4) with the logging of the events (Table 13) is shown in Table 14.

Table 14: Views and events on a page

Source	Eventid	Viewid	Page	Action	Time
Original log		003	Page 4	Open prices page	01/09/14 8:03:33
Events on page	001	003	Page 4	Open lamellae 1	01/09/14 8:03:35
Events on page	002	003	Page 4	Open lamellae 2	01/09/14 8:03:38
Events on page	003	003	Page 4	Open contact form	01/09/14 8:03:55
Original log		010	Page 4	Open prices page	01/09/14 8:16:33
Events on page	004	010	Page 4	Open lamellae 2	01/09/14 8:16:58
Events on page	005	010	Page 4	Open lamellae 1	01/09/14 8:17:02
Events on page	006	010	Page 4	Open contact form	01/09/14 8:17:22
Events on page	007	010	Page 4	Send contact form	01/09/14 8:18:33

Now, the data is prepared from which we mine a process model of the behavior of the visitors on *Page 4*. Through the mined process model, we are able to answer the following questions:

- Which part is the most popular part of the page?
- In which sequence the different parts are viewed?
- Are the sub questions ordered on the page in the same way as the sub questions are viewed?
- How many questions are visited by one visitor on average?

The obtained process model of our running example is shown in Figure 11. We see that the opening of the page is the start event, followed by opening at least one of the lamellae before opening the contact form. Only one visitor sends the contact form and ends the view. The other visitor does not send the contact form, but ends the view directly.

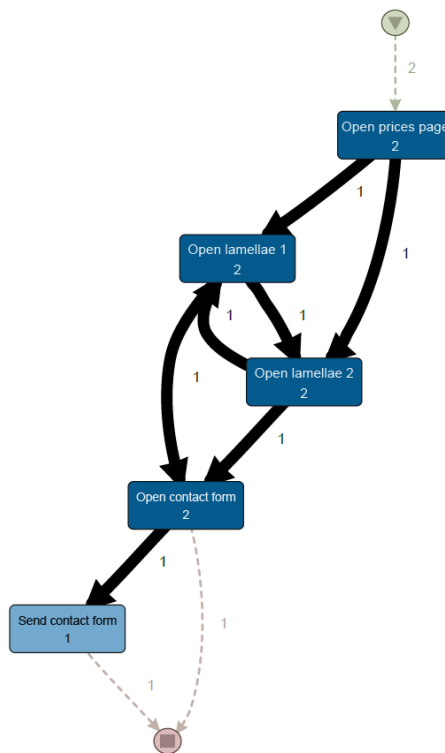


Figure 11: Behavior on page - MyEnergy

5 PREPARATION PHASE - BUSINESS UNDERSTANDING

5.1 GENERAL INTRODUCTION

Two case studies are conducted at Rabobank. Rabobank is a large financial institution with a cooperative organization model, consisting of about 100 different local branches and a central supporting office located in Utrecht. Rabobank provides payment accounts, savings accounts, loans, mortgages, credit card services, insurances and investment services to both consumers as well as businesses. In this case study we focus on the online web pages of the retail mortgages section. Rabobank would like to find out whether or not real time process control on the website is beneficial for them in order to become more relevant for their website visitors in order to increase the conversion and customer satisfaction.

5.2 BACKGROUND

Mortgages are important for banks. For the Rabobank, a new mortgage yields large revenue. Beside this yield, the bank is more likely to become the *main bank* of a customer once the customer has a mortgage with Rabobank. This means that bank sells additional products to this customer, such as savings accounts, credit cards, insurances and other products. The bank puts much effort in optimizing the *customer journey* of obtaining a mortgage. The customer journey is the experience of orienting, advising and signing a mortgage offer and eventually obtaining a mortgage. For many customers, some parts of this journey take place on the website of the Rabobank.

In order to make the online customer journey as easy as possible, Rabobank offers an online mortgage calculator and a mortgage dossier application. This is called the “Rabobank Hypotheek Dossier”, also abbreviated with RHD. A screenshot is shown in Figure 12. In this online application visitors can enter information about their current financial situation and gets insights on the possible mortgage amount as well as their monthly payments. Through this dossier, visitors can also upload documents about their situation, such as income statements. When a visitor fills the dossier in a proper way and uploads all the required documents before the actual physical appointment takes place, he/she gets a discount up to 450 euros. This discount is given because the amount of work for the local branch employees is reduced.

Rabobank Hypotheekdossier

Hypotheekdossier opslaan

Help Contact

Introductie

1. Persoonlijke berekening

Inkomsten

Schulden en kredieten

Nieuwe woning

Verbouwen en extra lenen

Eigen geld

Rente

Hypotheekvorm

Terugbetalen

Afspraak maken

Ik heb al een afspraak

Er is nog geen berekening gemaakt. Vul eerst links uw gegevens in. Daarna ziet u hier hoeveel u kunt lenen, hoeveel u nodig heeft en hoeveel u per maand betaalt. Hoe meer u invult, hoe uitgebreider de persoonlijke berekening wordt.

2. Afspraak maken

3. Documenten toevoegen

Figure 12: Screenshot RHD

5.2.1 PROCESS OF BUYING A MORTGAGE

The desired process of buying a mortgage is elaborated in Figure 13 and Figure 14. The journey consists of a number of steps, such as orienting, visiting the website, and logging in into RHD. The process is divided in a customer view and a bank view.

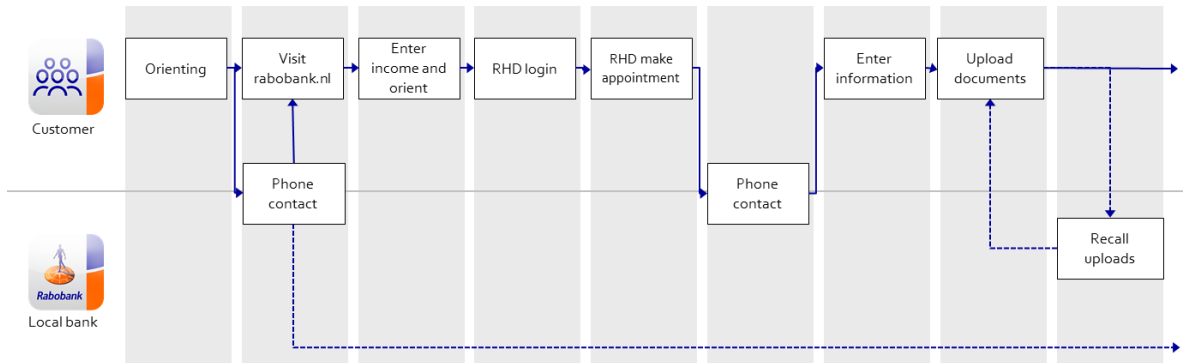


Figure 13: Process view of visitor

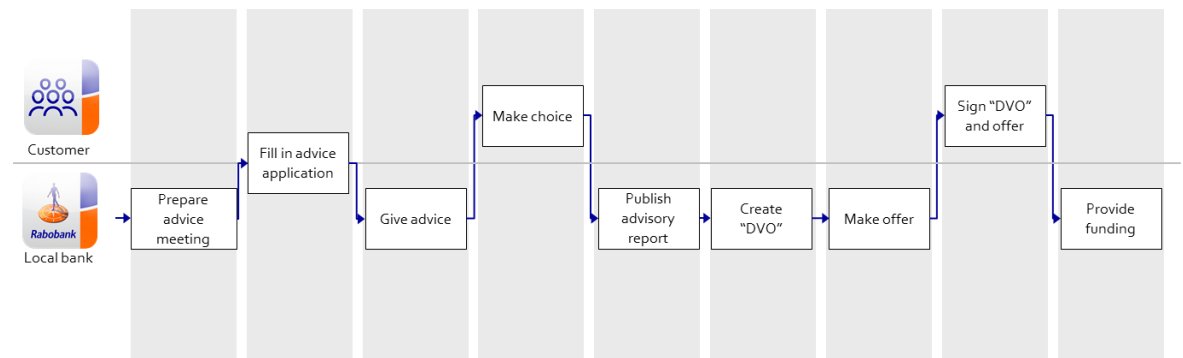


Figure 14: Process view of bank

5.2.2 MORTGAGE SECTION ON RABOBANK

The mortgage section on the Rabobank website consists of two major parts: the mortgage information pages and the "Rabobank Hypotheek Dossier". The RHD is an application inside the website and is implemented as a so-called *portlet*. A portlet is a module that is active on the website. A visitor can use the RHD in different stages of his/her mortgage: orienting for their first mortgage, raising their existing mortgage, adapting the mortgage for moving or transferring their mortgage from another bank to the Rabobank.

Mortgage section Rabobank.nl	
Mortgage Information Pages <ul style="list-style-type: none"> • Customer Journey leading to the RHD • Information pages about mortgages • Home page already adapted to different target groups 	Rabobank Hypotheek Dossier (RHD) <ul style="list-style-type: none"> • Advanced calculator for mortgages • Extensive online orientation • New customers can get a discount by filling in the dossier and uploading documents to the dossier

5.3 BUSINESS OBJECTIVES

The main business objectives are to: increase the sales of mortgages and be relevant online for the customers. For the website, this objective is translated in several sub objectives:

- Let visitors enter the mortgage calculator application (RHD)
- Let visitors send a request for a mortgage advice appointment with their local bank
- Let visitors find their information more easily and be relevant to the visitor

The business success criteria are to increase the website conversion as well as to become more relevant to the end user. Becoming more relevant can be achieved by presenting the most suitable information to the visitor in a prominent way and by hiding (or presenting less prominently) less suitable information for the visitor.

5.4 INVENTORY OF RESOURCES

5.4.1 SOFTWARE SOURCES

The current website of the Rabobank already has some *intelligence* active on their website. A start is made with Behavioral Targeting, mainly with banners on the home page and the logout page of internet banking. Behavioral Targeting is a method to select the most relevant advertisements on a web page, based on web browsing behavior. On a few pages, conditional content is active. This is a method to present different content to different groups of people based on previous website visits and/or customer profile. For this classification, the *IBM Interact* solution is used. The bank uses *IBM Netinsight* as their web analytics solution. Furthermore, *Tridion* is their Content Management System. A conditional content filter (CCF) takes care of showing the correct information to the correct target group.

5.4.2 DATA SOURCES

Click data and profile data is linked together using a reference table, based on cookie information. An overview of the available data sources is shown in Table 15.

Table 15: Data sources

Data source	Contains
IBM NetInsight	Click log data of the website
IBM Interact	Profile information about the visitors

5.5 DATA MINING GOALS

The goal of this data and process mining project is to deduce new rules that can be used to adapt the web site of the Rabobank in a better way. As a data source for the mining tasks, web data is used. Also, this project provides useful insights in the current behavior of visitors on the website. This project is divided in two parts: a visitor classification part and a content segmentation and classification part. Therefore, two case studies are conducted.

1. Assess current classification method on the home page of the mortgage section and search for improvements
2. Analyze and create new clusters by using segmentation on the FAQ section and apply classification to visitors into these clusters

5.5.1 DATA MINING SUCCESS CRITERIA

The data mining project is executed successfully when relevant insights are gathered from the data. These relevant insights are obtained through:

- Insight in the performance of the current conditional content filter
- Improvement proposal for the Conditional Content Filter, based on web data
- Field test: validation or rejection of the proposed improvements
- Proposal of new clusters for the FAQ section based on web data
- Classification of individual users into the new proposed clusters
- Insight in effect of classification on the homepage of the mortgage section
- Insight in behavior on web site with use of process mining

6 PREPARATION PHASE - DATA UNDERSTANDING

In this chapter, a deeper look is taken into the data. We answer questions such as: which data is available, what does the data look like and what is the data quality?

6.1 INITIAL DATA COLLECTION

Because of the large amount of traffic on the Rabobank web site, a small time period is chosen to perform the analysis on. Besides, the profile data source of the visitors refresh every month. The time horizon of the data chosen therefore is September 1, 2014 till October 12, 2014. In this period we have a consistent data set available. The queries used to retrieve this data are located in Appendix 1. A visual representation of these queries is shown in Talend Open Studio, an ETL (extract, transform, and load) tool.

6.2 DESCRIPTION OF DATA

Since the used applications contain over thousands of tables and relations, only the relevant tables and fields are discussed. The *views* are the views of a page, belonging to a *visit*. A *visit* is performed by a *visitor*. The information of a *visitor* is, in some cases, extended with the profile and other information in the *vw_VisitorProfileTable*. Each view has a related *page* and sometimes a target group (*targrpf*) or *events* that occurred during that *view*. The actual target group and event label are retrieved by using the lookup tables *eventlabel* and *targrpid*. In Table 16 and Table 42 (Appendix), more information about the used tabled is provided. A description for each table is given and the number of fields and entries are shown.

Table 16: Overview of data tables

Source	Table	Description
Netinsight	Visitorid	Unique visitors
Netinsight	Visits	Visits, belonging to a visitor
Netinsight	Views	Views of a page, belonging to a visit
Netinsight	Pageid	Unique pages on the website
Interact	VW_VISITORPROFILETABLE	Profile and click data information of visitors
Netinsight	Targetgroup	Behavioral Targeting Classification
Netinsight	Targetgroupid	Label of Behavioral Targeting Classification (lookup)
Netinsight	Events	Events on a page, such as a link click or tab open
Netinsight	Eventlabel	The connection between an event and a label
Netinsight	Eventlabelid	The actual label (lookup)

6.3 CONNECTION OF PROFILE INFORMATION AND VISITORID

In the Dutch law, it is not permitted to place targeting cookies without explicit permission of the visitor. So only visitors that explicitly accepted the cookie allowance, are used for this project. Therefore, a part of the visitors cannot be tracked. It is also not possible to track a visitor through different devices. Each device has its unique cookie.

More advanced demographic and customer profile data is connected to the cookie identifier after a visitor has logged in into the secure area of internet banking. The information about the visitor, such as age, and gender as well as their current product portfolio, is available after this connection is made. If a visitor has accepted cookies, but has not logged in yet, only click data behavior is recorded. This flow is depicted in Figure 15. The distribution of visitors that cannot be recognized, only recognize on click data and both recognized on click and profile data, is about 1/3 for each group. Both click data and profile data is located in *vw_visitorprofiletable*.

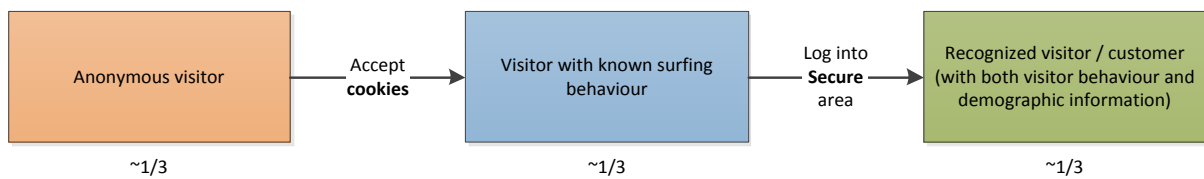


Figure 15: Flow of connecting profile information

7 CASE STUDY 1 - OPTIMIZE CONDITIONAL CONTENT

7.1 PROBLEM DESCRIPTION

Behavioral Targeting is active on the mortgage section of the Rabobank website. For different groups of people, specific content is provided. These groups are defined by the customer journey manager of the mortgage section. The classification of visitors into one of these sub groups is based on profile information, used by IBM Interact (as elaborated on in *data understanding* section). In this case study, we assess the current performance of the classification engine, propose improvements and run a validation. We use the methodology of research question 2 (How can we identify which visitor needs which information).

7.1.1 ASSESS CURRENT SITUATION

At the moment of writing, Rabobank targets three groups. The homepage of the mortgage section is adapted to each group. The groups and corresponding classification rules are shown in Table 17. Also, a fourth group is active. This is the *default* group and this page is shown when there is no information available about the visitor. The classification engine cannot make a decision or visitors did not accept the cookie allowance. When the classification engine is unreachable at the moment of the view, a *target group* with label *overig* is recorded and the content of the *default* page is shown.

The adaptation of the mortgage homepage consists of several elements. In Figure 16, a screenshot of the *default* mortgage homepage is shown. This *default* page is not adapted to a certain target group. For each of the three specific groups, the upper three white elements are changed, some information on the right hand side of the page is modified and the four blocks at the bottom contain different content, suitable to the group of the visitor.

Figure 16: Mortgage homepage

Table 17: Current groups on Mortgage Home

Group	Classification Rule	Definition
Starters	$eig_hs_ind = 0$ and $dnst_hyp_afn_ind=0$ and $leeftijd>18$	Visitors looking for their first house and mortgage
Oversluiters	$dnst_hyp_afn_ind=0$ AND ($eigendom_cd=5$ OR ($eigendom_cd=4$ and $eig_hs_ind<>0$)) and $leeftijd>18$	Visitors who already own a house, but do not have a mortgage at Rabobank
Doorstromers (also called "Hypotheekbezitters")	$dnst_hyp_afn_ind=1$	Rabobank Mortgage owners

It is not clear to the bank what the performance of the classification really is. This performance is measured in accuracy, recall and precision. The classification performance is assessed by checking the classification of a page later in the funnel. Once a visitor enters RHD (mortgage application on the web), they choose their "group" explicitly by answering 4 questions. Based on the answer of these questions, the visitor continues to

the RHD application in a certain mode, also called a *portlet mode*. Once we know both variables of a visitor, we can assess the performance. The variables together with the associated web pages are displayed in Figure 17. Also the possible values of the *target group* and *portlet mode* variables are depicted in Figure 17.

To clarify this concept, we use two examples of a visitor that flow through the situation in Figure 17. When a visitor is recognized by the classification engine, he/she gets a home page with conditional content. Suppose this visitor is a young female adult that rents a house. The conditional content filter presents her a page with information about buying your first house. After orienting in the mortgage section, she wants to make a calculation about her monthly costs. She enters the RHD and on the *RHD Start* page, she explicitly states that she does not have a house and that she is orienting to buy one. The prediction on the mortgage homepage was correct: suitable content was shown to her.

Another visitor, a young male adult that still lives at his parents, is targeted as *Oversluiter*. The mortgage homepage provides him content about moving his current mortgage to Rabobank. After orienting online, he enters the RHD and he explicitly states that he is looking for his first house. There is a mismatch between the content that he saw at the mortgage homepage (*target group*) and the RHD modes (*portlet_mode*). The prediction was wrong: the information shown at the home page did not fit his information need.

Classification prediction: classification done by the Interact engine on the Mortgage homepage (*target group*)

Classification result: *portlet_mode* in RHD

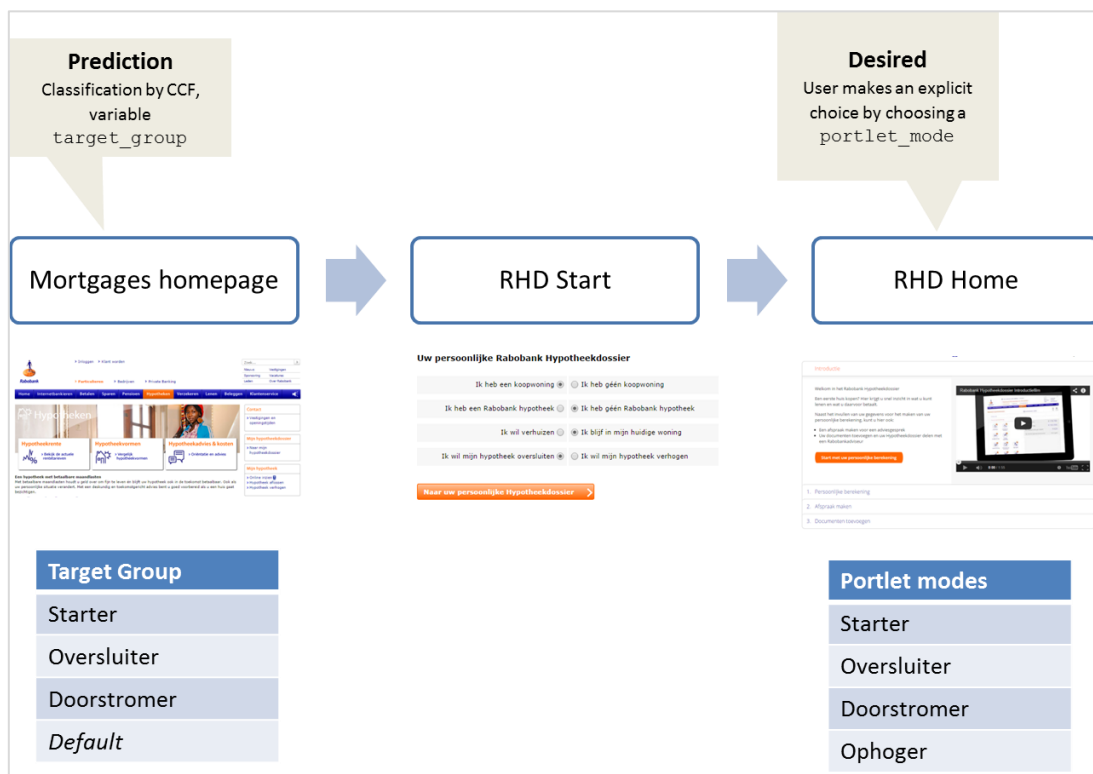


Figure 17: Pages and variables

7.1.2 IMPROVE SITUATION

For a part of the visitors, profile information is available. Once we know this information, and we know the *portlet mode* of a visitor that enters RHD, we have the dependent (target) variable and the independent variables to train a classification model with. The classifier with inputs and outputs is shown in Figure 18. This classification model has as a requirement that the output of this model can be implemented in the current classification engine.

After building the model, we check if we expect an increase in accuracy, precision and/or recall. After implementing the new rules into the classification engine, we re-run the performance analysis and investigate if the quality of the engine is increased.

Target variable: portlet mode
Independent variables: all data in vw_VisitorProfileTable

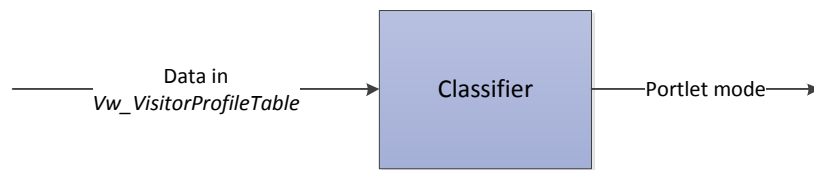


Figure 18: Classification model

7.2 DELIVERABLES

We answer the following questions in this case study:

- What is the quality of the classification in terms of accuracy, precision and recall?
- What effect does the conditional content in visiting relevant pages?

Furthermore, we propose a new model for the classification of visitors, based on the data in the *vw_visitorprofiletable*. Possible improvements are implemented and validated in a field test.

7.3 DATA PREPARATION

Before assessing and modelling with the retrieved data, we need to prepare and clean this data. Since the data has the web as a source, several filtering steps must be conducted.

7.3.1 SELECT DATA

A data export is made from September 1st, 2014 till October 12th, 2014. On this data, we want to run analysis only including visitors who have *both* a unique target group *and* a unique portlet mode. This means that they have seen the mortgage homepage with conditional content (target group) and have entered RHD with a portlet mode.

In our data set, we want only have visitors who have seen a *consistent* website. It is possible that a visitor has multiple target groups. In this situation, a visitor has seen multiple versions of the mortgage homepage. We do *not* want to take these visitors into account.

However, we do consider visitors who have seen an *overig* target group. The failure of the classification engine is temporary. If a visitor has two target groups where one of the target groups is *overig*, we assume that the visitor has seen one consistent target group. E.g., if the set of target groups of a visitor is {*starter*, *overig*}, we

assume that the corresponding target group is *{starter}*. We filter out visitors with target group sets such as *{starter, oversluiter}*, *{default, starter}*, etc.

Also for the *portlet* mode, we want to have consistent visitors. This means that we want to have visitors with one (and only one) portlet mode. The RHD enforces that visitors must have one target group. The combination of visitors with one consistent target group and one consistent portlet mode is the final set for the analysis.

7.3.1.1 IMPACT OF FILTERING

To know which impact this filtering has on the total set of data, we investigate how many visitors we filter out. In this part of the case study, we investigate the number of target groups per visitor and the number of portlet modes per visitor.

Table 18: Target group and number of visitors

Target group	Number of visitors
Starter	15.262
Doorstromer	51.160
Oversluiter	13.093
Default	114.094
Overig	16.776

In Table 18, we see that the most of the visitors are not targeted at all (*default*). The *doorstromers*, also described as *Rabobank mortgage owners*, occur the second most. The *overig* group has a share of 8%, which implies that in many cases the classification engine could not be reached.

In Table 19, the numbers of target groups in combination with the number of visitors are shown. 91% of the visitors have seen one target group. We omit the other 9% of the visitors who have seen more than one target group.

Table 19: Number of target groups per visitor

Nr. Of target groups	Nr. Of visitors	% Visitors
1	174.224	91%
2	15.691	8%
3	1.504	1%
4	63	0%
5	3	0%

We omit the visitors who have more than one portlet mode. In Table 20, the statistics of the portlet modes are shown. The portlet modes per visitor are displayed in Table 21. In total, an inner join is performed on the set of visitors having one *target group* and the visitors having one *portlet mode*. The join resulted in a data set with 9.100 visitors.

Table 20: Count of portlet modes

Portlet mode	Count
Starter	12.002
Doorstromer	12.533
Oversluiter	2.571
Ophoger	5.289

Table 21: Number of portlet modes per visitor

Number of portlet modes	Number of visitors	% visitors
1	31.751	99%
2	317	1%
3	2	0%
4	1	0%

7.4 CURRENT PERFORMANCE CCF

With the obtained set of 9.100 visitors, we analyze the current performance. In this part of the report, we first elaborate of the classification mapping between the mortgage homepage and the portlet modes in RHD. Subsequently, we provide the scorer and accuracy measurements. We also take a look at the impact of a correct, wrong or no (*default*) classification.

7.4.1 CLASSIFICATION MAPPING

Terms in the target group and terms in the RHD do not always have the same meaning. Because terminology between the mortgage homepage and the mortgage dossier differ, a mapping is created. The mapping is shown in Table 22 and is drafted with a web analyst of Rabobank. We distinguish a difference between a good classification (correct), wrong classified (wrong) and not classified (default).

Table 22: Classification Mapping

Portlet Mode (actual)	Target Group (prediction)	Quality	Argumentation
Starter	Starter	Correct	Visitor is searching for his/her first house.
Starter	Oversluiter	Wrong	Visitor is searching for information for his first house, but gets information about transferring his/her mortgage to the Rabobank.
Starter	Doorstromer	Wrong	Visitor is searching for information for his/her first house, but gets information about his/her mortgage (which he does not have).
Starter	Default	Default	Visitor is not recognized and gets default content.
Doorstromer	Starter	Wrong	Visitor wants to move to another house, but gets content about a first house.
Doorstromer	Oversluiter	Correct	The visitor wants to move (and does not have a mortgage at the Rabobank). He/she gets information about transferring his mortgage to the Rabobank.
Doorstromer	Doorstromer	Correct	Visitor wants to move and he/she gets information about his/her current mortgage. On this page, also information about moving is displayed.
Doorstromer	Default	Default	Visitor is not recognized and gets default content.
Ophoger	Starter	Wrong	Visitor wants to raise his/her mortgage, but gets information about his/her first mortgage.
Ophoger	Oversluiter	Wrong	Visitor wants to raise his/her Rabobank mortgage, but gets information about people who want to transfer their mortgage to the Rabobank
Ophoger	Doorstromer	Correct	Visitor wants to raise his/her mortgage and gets information about his/her mortgage. On this page, also information about raising is displayed.
Ophoger	Default	Default	Visitor is not recognized and gets default content.
Oversluiter	Starter	Wrong	Visitor wants to transfer his/her mortgage to the Rabobank, but gets information about their first house.
Oversluiter	Oversluiter	Correct	Visitor wants to transfer his/her mortgage to the Rabobank and gets adequate information.
Oversluiter	Doorstromer	Wrong	Visitor wants to transfer their mortgage to the Rabobank, but gets information about mortgage owners.
Oversluiter	Default	Default	Visitor is not recognized and gets default content.

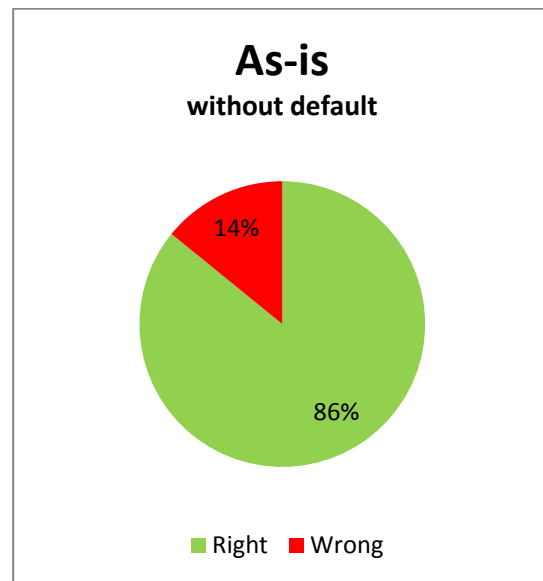
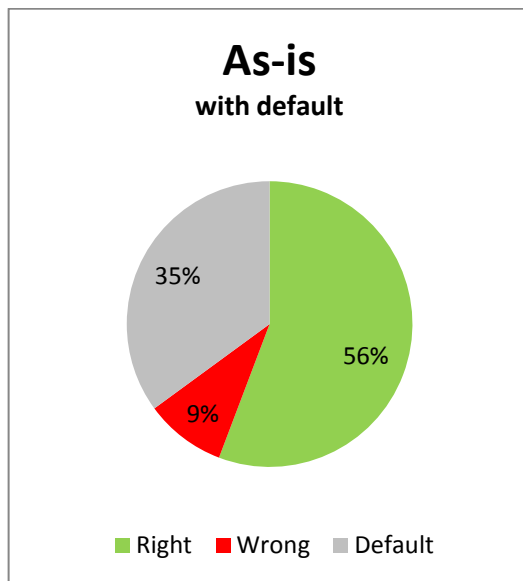
7.4.2 SCORER

Table 23 shows a scorer table for the current situation. At the top, the classification of the Behavior Targeting engine is shown (*target group*). At the left, the portlet mode in the mortgage dossier is shown. With color coding, the *correct (green)*, *wrong (red)* and *default (grey)* classifications are shown. The majority of the visitors are classified correct.

Table 23: Scorer for as-is situation

	Oversluiter	Starter	Doorstromer	Default
Oversluiter	371	66	90	308
Starter	267	1.881	229	1.354
Doorstromer	434	96	1.532	1.289
Ophoger	76	11	857	239

We assess the performance in two separate groups: the total population and the population without the *default* classification. This second group is a subset of the total population, but gives a more fair performance of the classification. The total classification is 56% correct, but in the subset (without *default*), the performance is for 86% correct. 9% of the visitors have seen a page that did not suit to their situation.



7.4.3 ACCURACY STATISTICS

The accuracy metrics are calculated and shown in Table 24 and Table 25. In this situation, not for every class proper metrics can be calculated. This is because the mapping, showed in Table 22, is not one to one.

When looking at *starters*, the precision and recall are defined as follows:

Precision: The percentage of the predicted *starters* that are actual *starters*

Recall: The percentage of actual *starters* that are predicted as *starters*

When precision for starters is low, many visitors saw content about starters that did not fit by their information need. When the recall for starters is low, many visitors did not see the starters content while they were interested in this content. A more detailed explanation about accuracy, precision and recall can be found in the methodology (Assess Model).

The precision of *starters* is very high with 92%. The precision of *Oversluiters* is somewhat lower with 70%. This means that 30% of the visitors who have seen the page about *Oversluiters*, saw wrong content. The recall of *ophogers* is high, compared to the other categories. This makes sense, because these *ophogers* are already customers of Rabobank and are recognized easily. In Table 25, the recall within the targeted population is depicted.

Table 24: Accuracy statistics as-is for whole population

Category	Recall	Precision
Starter	50%	92%
Hypotheekbezitters	n/a	88%
Oversluiters	44%	70%
Ophogers	72%	n/a

Table 25: Accuracy statistics as-is for targeted population

Category	Recall	Precision
Starter	79%	92%
Hypotheekbezitters	n/a	88%
Oversluiters	70%	70%
Ophogers	91%	n/a

7.5 EFFECT OF CLASSIFICATION

In this section, we investigate if the quality of the classification influences the behavior of a visitor on the website. We check the difference in visits of relevant pages. Three groups are distinguished: visitors that are classified correct, wrong or are not classified at all. Note that we only know if a classification is correct or not after people have entered the RHD. Therefore, we only do the analysis on the visitors having both a *target group* and a *portlet mode*.

The Mortgage section on the website consists of a number of categories. These categories are defined by Rabobank² and are used in this *bucket analysis*. In Table 26 is shown that for each group (portlet mode) and each bucket, there is an influence in visiting the relevant pages. The percentages are the fractions of the total views visited in the mortgage section. The views in the mortgage application (RHD) are excluded from this analysis.

² Used in an internal research of Carpaij and Cox: "RHD Klantpadenonderzoek fase 3"

Table 26: Bucket analysis

Group	Bucket	Percentage of views from visitors that are correct classified	Percentage of views from visitors that are incorrect classified	Percentage of views from visitors that are not classified
Starters	Eerste huis	8,9%	2,1%	7,2%
Doorstromer	Ander huis	5,8%	2,3%	5,5%
Ophogers	Huis verbouwen	3,0%	1,9%	4,4%
Oversluiters	Oversluiten	16,0%	9,2%	13%

We see that the visitors who are classified correct visit more *relevant* pages on average than visitors who are classified wrong. We also see that visitors who are correctly classified, visit more relevant pages than visitors who are not classified (except *ophogers*). The difference between the visits of relevant pages of a visitor who is targeted default or correct, is not that big. The biggest difference is the difference between a wrong targeted visitor and the other (default, correct) targeted visitors.

In general we say that targeting has a positive influence in visiting relevant pages. We also see the drop in the visiting of relevant pages when the targeting is wrong. Therefore we say that is it good to target visitors, but one should be sure about the classification. If the probability of a certain class is not high enough, it is better to not target them at all.

7.6 MODELING

In this part, we check if the classification performance can be increased. The overall accuracy of 86% is quite decent, however, the precision of *Oversluiters* and the recall of *Starters* have the lowest performance.

We use a decision tree learner as the classifier, as proposed in our methodology. We develop a test design, build the model and finally assess the model. After the assessment, the improvements are ready for implementation.

7.6.1 GENERATE TEST DESIGN

Rabobank uses two ways to determine if someone is a house owner. This is done through a Boolean indicator (*eig_hs_ind*) that contains the information if someone is a house owner, or by looking at the geographical location of their address (*eigendom_cd*). If the houses in the area of their address are all owner-occupied houses, Rabobank assumes that the customer is a house owner. Therefore, two decision trees are trained, one with the geographical information and one with the Boolean indicator. Other variables are depicted in Table 27.

Table 27: Relevant variables in VW_VisitorProfileTable

Variable	Description
Eig_hs_ind	Boolean indicator if the user owns a house (0,1)
Eigendom_cd	Indicator of distribution between rental houses and owner-occupied houses in the area where the user lives (0-5)
Dnst_hyp_afn_ind	Boolean indicator if the user has a mortgage of Rabobank (0 = false, 1 = true)
Leeftijd	Age of customer (0-125)
Won_1e_fase	Indicator if the user has clicking behavior on the pages about your first house (0-2)
Huish_Type_cd	Household type / family composition (1-17 or 99)

7.6.2 BUILD MODEL

For the Decision Tree model building, we made use of the C4.5 algorithm. This algorithm is a statistical classifier and generates a decision tree. The implementation in KNIME 2.10.3 is used. The implementation works with the C4.5 decision tree modeler. Through the x-partition and x-validation nodes a cross-validation is executed. Furthermore, we require a minimum of 200 records per node. We also apply a 10-fold Cross-Validation with random sampling

7.6.3 ASSESS MODEL

The two models that are mined are depicted in Figure 19 and Figure 20. The variable that contributes the most to the information gain, creates the first split. The variable that contributes the second most to the information gain, creates the second split, etc. One sees that the variable *dnst_hyp_afn_ind* is the variable that contributes the most in terms of information gain, in both models. The prediction in the left and right leaves have a higher certainty. For example: when no information is available, the decision tree predicts that a visitor is *doorstromer*, but this prediction is only true in 52,1% of the cases. When we know that the visitor has a mortgage of Rabobank (*dnst_hyp_afn_ind* = 1), the prediction probability rises to 86,8%. Also the *eig_hs_ind* variable does contribute to the information gain.

An interesting point is shown in Figure 19: the geographical information (*eigendom_cd*) does contribute to the information gain, but the age (*leeftijd*) of a visitor is much more important. In this decision tree it is shown that people below the age 33 should be targeted as *Starter* instead of *Oversluiter*. We use this observation in our implementation recommendation.

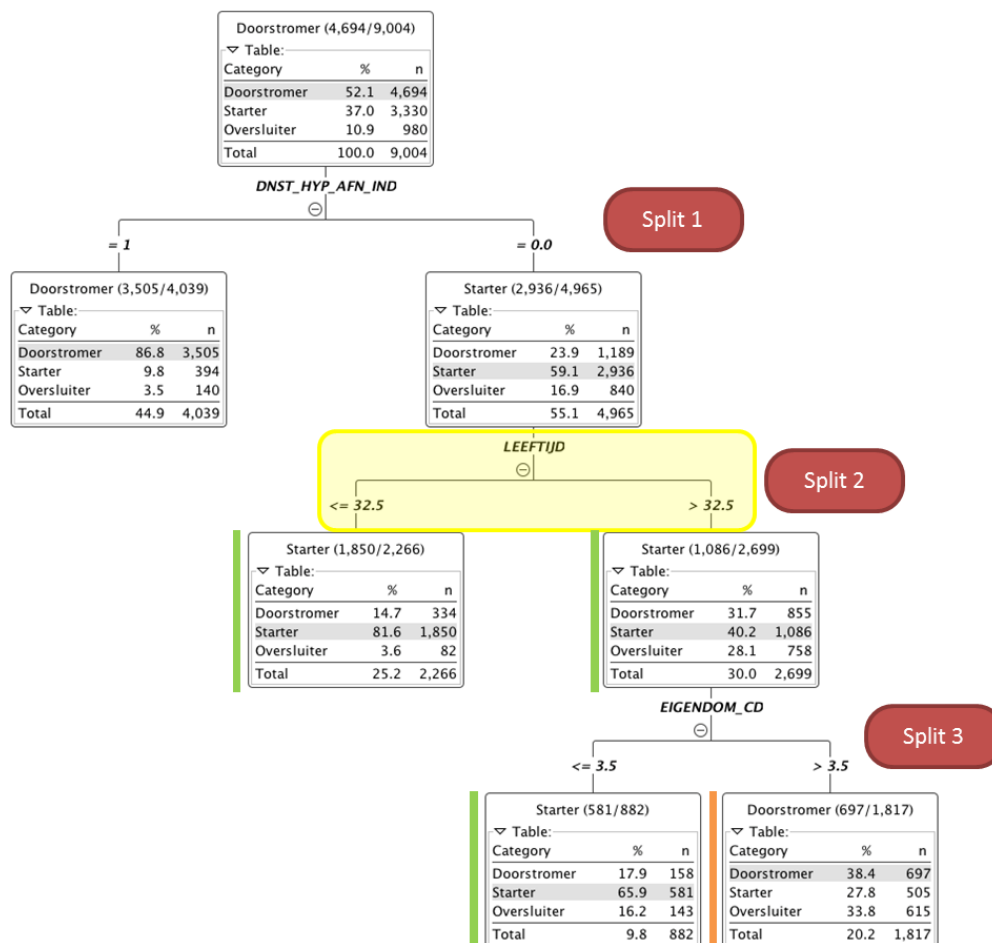


Figure 19: Mined decision tree for classification BT Home with eigendom_cd

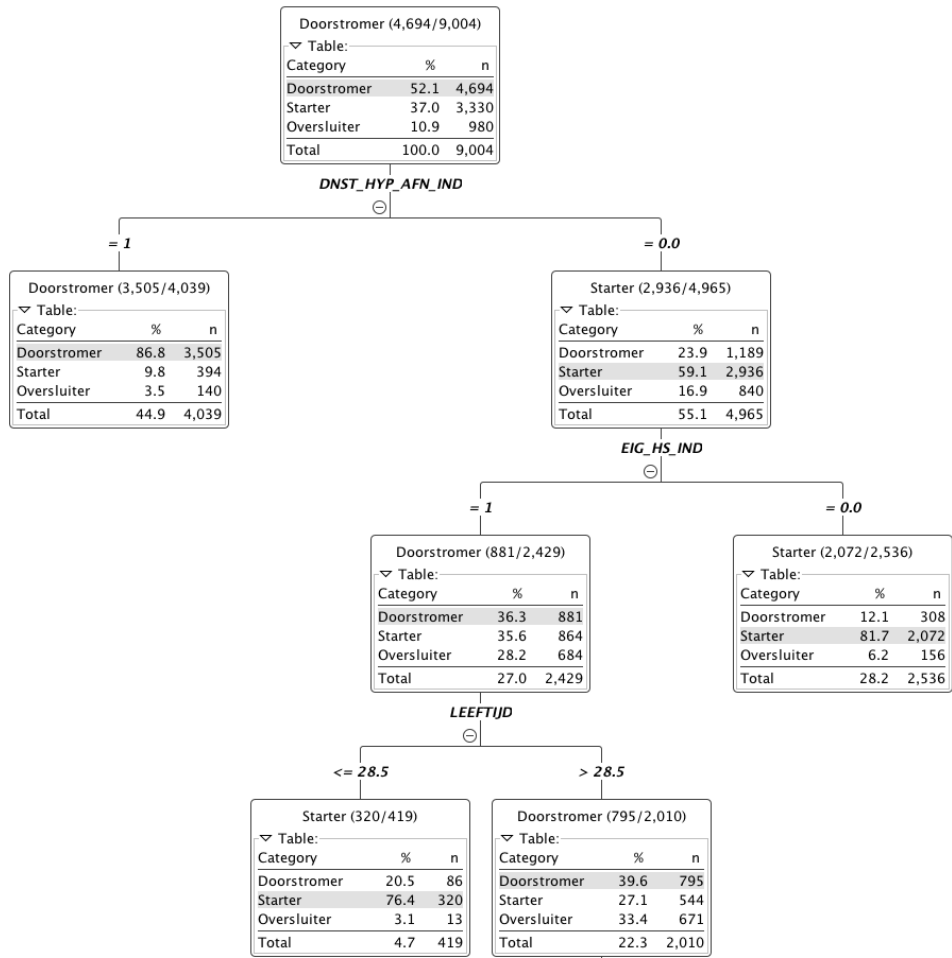


Figure 20: Mined decision tree for classification BT Home with eig_hs_ind

7.6.4 IMPLEMENTATION RECOMMENDATION

When we look at the mined decision tree, we see that the split of age at 33 is an important change to the current classification rules. We look at the group of visitors who:

- Live in an area where the houses are bought (owner-occupied houses)
- Do not have a Rabobank mortgage

The age of 33 should be added as the split point. In the old situation, this group would be classified as *Oversluiter (house owner)*. In Table 28 we see that the visitors below 33 are *starters* in 79% of the cases and do not own a house. The visitors who are 33 years old or older, do *own* a house in 74% (33% + 41%) of the cases. The classification rules should be changed with taking this new information into account. Note that we only assess people who do live in an area where the houses are bought and do not have a mortgage by the Rabobank.

Table 28: Split on age 33 for people who live in an area where houses are bought and do not have a Rabobank mortgage

Group	Share (LEEFTIJD (age) < 33)	Share (LEEFTIJD (age) >= 33)
Doorstromer	12,5%	33%
Ophoger	4%	5%
Oversluiter	4,7%	41%
Starter	79%	21%

7.7 VALIDATION

The improvements to the classification rules were implemented on the Rabobank website. The change of this rule went live at November 12th, 2014. Since November 27th, the Conditional Content Filter has been disabled³. For the validation, the click log of these 15 days is retrieved. The data is processed in the same way as the data for the *current situation* is processed. In this chapter, the new situation is assessed and a validation is performed.

Ideally, we would run an A/B test. An A/B test is a random experiment where the current classification rule set is associated with h_0 and the new classification rule set is associated with h_1 . This is a concurrent statistical test in which 50% of the visitors are shown content that is selected based on the original classification rules (control variant) and 50% of the visitors are shown content that is selected based on modified classification rules.

Due to limitations of the A/B testing functionality, we perform a t_0/t_1 test (Figure 21). This is an uncontrolled experiment. The rules are set live in a moment in time. Situation 1 represents our h_0 hypothesis and situation 2 represents our h_1 hypothesis.



Figure 21: Time line of data

³ Due to a new layout of the mortgage section, the classification engine is disabled since November 27th

7.7.1 SCORER AND ACCURACY

Also for the validation data, a scorer table is created. On the horizontal axis, the prediction of the BT engine (*target* group) is shown. On the vertical axis, the *portlet mode* is shown. The scorer table is shown in Table 29. With this data, accuracy statistics are calculated.

Table 29: Scorer of live validation

	Oversluiter	Starter	Doorstromer	Default
Oversluiter	334	42	62	691
Starter	182	1.887	207	4.090
Doorstromer	392	129	1.576	2.829
Ophoger	64	11	865	668

The accuracy statistics for the new situation (without default) and the difference between the assessed As-is situation (1) and new situation (2) are displayed in Table 30.

We expect that the proposed changes have a small effect on the classification engine. The proposed changes in the model should have effect on the precision and recall of *starters* and *oversluiters* and on the overall accuracy. Note that the effect in the section *Implementation Recommendation* seems huge, but that the selected data in this recommendation is a small subset of the total group.

The total accuracy is increased with 2%. We see that the precision of the *oversluiters* rose with 5%. More people saw the page *Oversluiter* when they should have. The recall of *starters* is also increased with 4%. More people who are starter, are correct targeted at the homepage.

The 1% loss of the precision of starters is the downside of this change. Some *Oversluiters* are now targeted as *Starter*. Comparing to the increase of the precision of *Oversluiters* and the increase of recall of *Starters*, this decision was worth the change. The increase of recall of *Oversluiters* is a positive, but not an expected effect. Although the rule set for *Oversluiters* should classify fewer visitors as *Oversluiters*, the recall is still increased.

Table 30: Differences between old and new situation for targeted visitors

Category	Recall Old	Recall New	Precision Old	Precision New	Difference Recall	Difference Precision
Starter	79%	83%	92%	91%	4%	-1%
Hypotheekbezitter			88%	90%		2%
Oversluiters	70%	76%	70%	75%	6%	5%
Ophogers	91%	92%			1%	

8 CASE STUDY 2 – ADD INTELLIGENCE TO FAQ

8.1 PROBLEM DESCRIPTION

The FAQ page on the mortgage section on Rabobank.nl (Figure 22) does not contain any form of intelligence towards the end user. In this case study, we take a closer look how intelligence can be configured with help of data and process mining techniques. The current section contains 11 question pages, with 24 (sub) questions on the question pages. This makes a total of 35 questions within the FAQ. The goal is to reduce the bounce rate or exit rate on the FAQ section by showing only relevant questions to the web site visitor. Rabobank also wants to increase the conversion of RHD and “make appointment” forms. Furthermore, Rabobank can be more relevant to the individual end user by adapting this FAQ. We use the methodology of research questions 1, 2 and 3 in this case study. First, we obtain the different information needs. Subsequently, we obtain a model to classify the visitor in the correct cluster. Finally, we perform an analysis what on how the information should be displayed on the website.



Figure 22: Front page of FAQ

8.2 DELIVERABLES

The goal of this case study is to obtain clusters with visitors who tend to have different questions. Visitors in a cluster must have the same need for a certain type of questions, while the visitors belonging to another cluster must have another need for another type of questions. Also classification rules are mined belonging to the obtained clusters. Furthermore, insights are given how important the FAQ is to the total conversion.

8.3 DATA PREPARATION

The view statistics of the questions are not located in one data source. The pages with questions are located in the *views* table, but the opening of the lamellae on the pages are recorded in *events*. The situation is displayed in Figure 23. Through a data merge operation, the views and events are located in one new data source. The total views per question or sub question is obtained from this set and are shown in Figure 24.

U kunt ook als starter op de huizenmarkt, zzp'er, werknemer met een tijdelijk contract of met een studieschuld een hypotheek afsluiten. We kijken welk bedrag u verantwoord kunt lenen, nu en in de toekomst. Bekijk hier welke opties u heeft.

Kan ik als starter een hypotheek krijgen?	+
Wat is een starterslening?	+
Hoe kunnen mijn ouders een hypotheek krijgen?	+
Kan ik een hypotheek afsluiten met een (studie)schuld?	+
Kan ik een hypotheek krijgen zonder vast contract?	+
Kan ik een hypotheek krijgen als zzp'er?	+

Figure 23: View and event

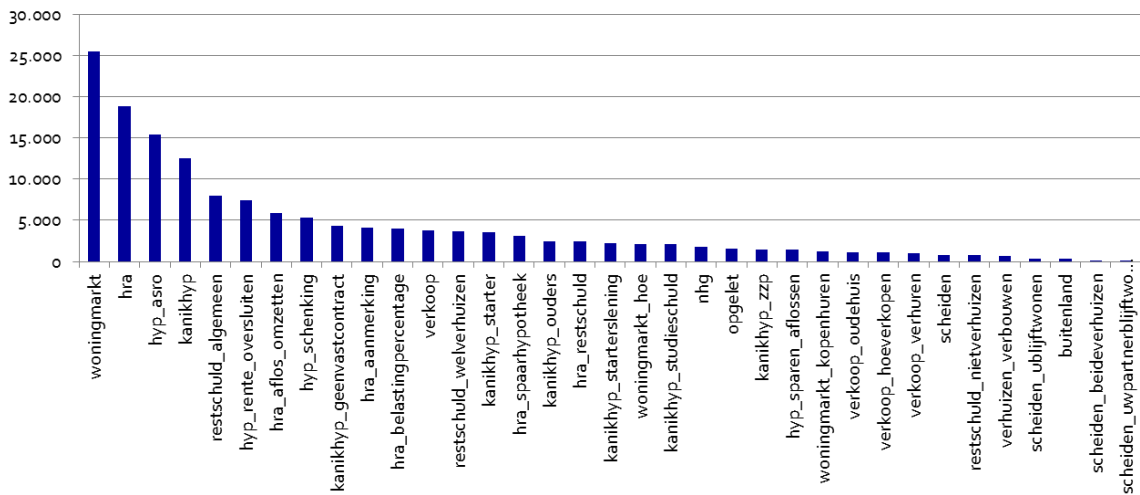


Figure 24: Views in the FAQ

8.4 DATA SELECTION

We only considered the questions that are accessible from the FAQ homepage (Woonvragen home). An overview of all the selected questions is added in the Appendix. The questions and shortcodes are shown in Table 43.

The most visitors only visit one question and often open one of the sub questions on that page. For the cluster analysis, only the *question pages* are taken into account, and not the individual questions on that page. The clusters that would be mined are the *question pages* with the corresponding sub questions We consider this as an undesired outcome.

8.5 RELEVANCE OF FAQ

Before diving into the modeling of clusters and classification rules, we first take a look at the relevance of the FAQ section. We define a few pages as end points and check the conversion to these pages. We also perform a sweet spot analysis that investigates how many questions are viewed on average by the visitors before reaching the conversion. The end points are defined in Appendix 4 – Conversion with FAQ:

- Make an appointment: “afspraak maken openen” (Figure 33)
- Enter RHD: “RHD Start” (Figure 34)
- Make appointment in RHD: “RHD afspraak maken versturen” (Figure 35)

8.5.1 CONVERSION

In this section, we answer the question: what is the contribution from the mortgage FAQ to the conversion? We divide the population of the visitors of the mortgage section in two groups and compare these two groups in terms of conversion into the end points. We distinguish the groups by the following formulas:

$$\text{No Questions Seen} := \text{Count}(\text{view in FAQ}) = 0$$

$$\text{Questions Seen} := \text{Count}(\text{view in FAQ}) > 0$$

We only check *if* a visitor reached a conversion, but do not consider *how many times* a visitor reached a conversion. A *t-test* is performed on the data. The confidence interval probability is set to 95%. The differences between the conversion ratios are displayed in Table 31.

The t-test is performed in Table 32. In a t-test, we set two hypotheses (Montgomery & Runger, 2007). The h_0 hypothesis says that there is no difference between the two groups. The h_1 hypothesis says that there *is* a difference. The table shows that every h_0 hypothesis is dismissed, because every *p-value* is below 0.05. Therefore, we conclude that the differences shown in Table 31 are significant.

A visitor who visits the FAQ section is more likely to make a (regular) appointment at the bank. The conversion ratio is still low. The most conversions from the FAQ section are coming from the pages: closing your mortgage against another interest percentage (101 cases), mortgages and donating (52 cases) and the question about the current state of the housing market (31 cases).

The visitors who visit the FAQ section have a lower conversion to *RHD Start* and making an appointment in the RHD. One of the reasons for this could be that visitors of the FAQ section have another goal on the website. Some visitors are only seeking information about their current mortgage and are not interested in the calculations in the RHD. The questions with the most conversions are: the current state of the housing market (883 cases), closing your mortgage against another interest percentage (268 cases) and moving while having a remaining debt (159 cases).

Table 31: Conversion and FAQ

Group	RHD Start	RHD Make appointment	Make appointment
No questions seen	9,32%	0,60%	1,01%
Questions seen	4,91%	0,14%	1,30%

Table 32: T-test conversion

Test Column	p-value (2-tailed)	Mean Difference	Conf. Interval	Lower bound difference	Upper bound difference
RHD Start	0,00	-4,41%	95%	-4,61%	-4,20%
RHD Make appointment	0,00	-0,47%	95%	-0,51%	-0,43%
Make appointment	0,00	0,29%	95%	0,19%	0,39%

8.5.2 SWEET SPOT ANALYSIS

The *sweet spot* is the average number of questions a visitor has seen when reaching a conversion. In this thesis, sweet spots are calculated for two groups:

- Table 33: All visitors who reached a conversion
- Table 34: Visitors who have reached a conversion and saw at least one question

We see that only a small fraction of the visitors that have reached a conversion, have seen a question before this conversion (5% for *RHD Start* and 7% for the regular *Make Appointment*). Therefore, the average number of question viewed when reached a conversion is low. For *make appointment*, the number is the highest with an average of 0,22 views per conversion.

The visitors who *have* seen at least one question *and* reached a conversion, all have seen more than two views on average (respectively 2,3, 5,2 and 2,9). We conclude that people who reached a conversion *and* viewed a question, are interested in de FAQ section. This group of visitors is probably the group that orient extensive online before reaching a conversion.

Table 33: Sweet spot for conversion

Goal	Case freq.	Views in FAQ	Expected views in FAQ per visitor
RHD Start	37.007	3.497	0,09
RHD Make appointment	2.242	385	0,17
Make appointment	4.437	962	0,22

Table 34: Sweet spot with visitors who saw at least one question

Goal	Case freq.	Views in FAQ	Expected views in FAQ per visitor
RHD Start	1.512	3.497	2,3
RHD Make appointment	74	385	5,2
Make appointment	330	962	2,9

8.6 MODELING

8.6.1 SEGMENTATION

In this part, we try to find new clusters of visitors who have the same information needs. The method that is used to obtain these clusters is described in the methodology. As a cut-off point, we use 4 questions. Visitors who have less than 4 views in the FAQ section are omitted. Besides this filtering, we apply a Nearest Neighbor search with searching for 24 neighbors. These parameters are obtained by trial and error. By inspection of individual visitors, we see by *which number of neighbors* we obtain a clear ranking of relevant questions for each visitor. This ranking is used by the clustering algorithm.

Table 35: Cluster algorithms

	K-means	c-means (fuzzy)
Number of clusters	2	2
Distribution between clusters	6% / 94%	40% / 60%
Relevance	-	+
Total	-	+

Two cluster algorithms are used. In Table 35 is depicted which algorithm performs best. The *number of clusters* is the number of distinct clusters that are obtained. Distinct clusters are clusters that clearly have another ranking of questions than the other clusters. The *distribution of clusters* represents which share each cluster has. The relevance of the clusters is assessed by inspection and represents to what extend the clusters are useful to implement on the web page of Rabobank. The *soft clustering* algorithm c-means performs best in terms of relevance and the distribution between the clusters. The two clusters of *k-means* do not clearly show another information need, while the cluster of *c-means* do show another information need. Therefore, the outputs of the c-means clustering are used.

The clusters obtained by c-means (fuzzy clustering) are depicted in Table 36 and Table 37. The clusters obtained through k-means clustering are depicted in Appendix tables Table 44 and Table 45. The questions are ranked according to the expected number of visits and the top 6 of most relevant question is shown. The questions with the highest *score* are the ones that are most relevant to the visitor. One sees that the question about repayment, savings, donations and changing the interest rate of the mortgage is the most interesting question for house owners. The question about "Can I get a mortgage in my situation?" is the most important question for *starters*. The other questions in this cluster are less relevant for the visitor.

Table 36: Cluster 0 - house owner (top 6)

Question	Score (Expected number of views)
Hoe kan ik aflossen, sparen, rente oversluiten op mijn hypotheek?	1,99
Hoe staat het met de woningmarkt?	1,77
Hoe is de hypotheekrenteaftrek geregeld?	0,57
Een restschuld, en dan?	0,1
Wat komt er kijken bij de verkoop van mijn huis?	0,1
Wat is de Nationale Hypotheek Garantie?	0,03

Table 37: Cluster 1 – starter (top 6)

Question	Score (Expected number of views)
Kan ik een hypotheek krijgen in mijn situatie?	3,8
Een restschuld, en dan?	0,56
Waar wordt op gelet bij het afsluiten van een hypotheek?	0,088
Wat komt er kijken bij de verkoop van mijn huis?	0,027
Nationale Hypotheek Garantie	0,023
Scheiden	0,0065

8.6.2 CLASSIFICATION

After the mining of the clusters, we assign a visitor into one of the two clusters based on their web data. We create a data set with all visitors having visits in the FAQ section and a profile in the *VW_visitorprofiletable*. The methodology is described and explained in the beginning of this thesis (assign visitor into a cluster based on web data). The most suitable cluster is assigned to the visitor, based on the highest average expected number of views.

8.6.2.1 TRAIN DECISION TREE BASED ON DATA INCLUDED IN BT PROFILETABLE

With a new trained decision tree, we train a model that predicts in which cluster a visitor belongs. The setup of this classifier is depicted in Figure 25. We take the assigned cluster as the dependent variable and other variables in the profile table as the independent variables. The minimum node size is set to 500, since we have 30.720 records available for mining. The decision tree is shown in Figure 26.

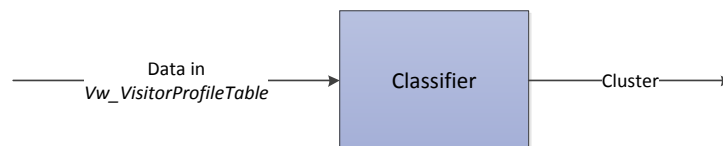


Figure 25: Classification engine predicting cluster

The result of the decision tree is that the clicking behavior on the website has the most predictive power. If a visitor visited web pages within the category *WON_1E_FASE*, he or she is classified in the *Starter* cluster. The variable represents that the visitor visited pages in the section “First house”. The possible values are {0,1,2} and represent the phase in the process of buying a mortgage. The values of *WON_1E_FASE* have an influence in the certainty, but the *Starter* cluster is the best choice for all values. If there is no click behavior in *WON_1E_FASE*, the best prediction is that the visitor is a house owner.

If no click data is available, the engine can make a decision on only the profile information about the customer itself. In Appendix 5 – Decision tree FAQ on profile information, the mined decision tree is added for this situation. The prediction is less powerful (in terms of accuracy) than the prediction with use of the web data. When a visitor do not own a house and has an age under 35, the website should show the content of the cluster Starter. In all the other cases, the website should show the content for “House Owners”.

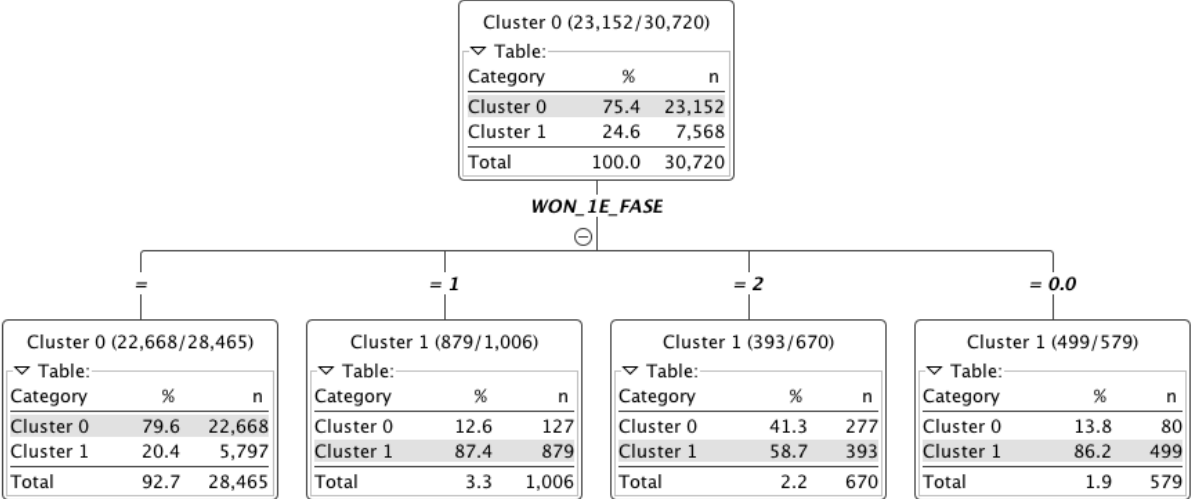


Figure 26: Decision tree on clusters

8.6.2.2 STATISTICS

In Table 38, the score matrix is shown of the mined decision tree in Figure 26. The prediction of the classification engine is shown at the left (rows) and the actual cluster is shown at the top (columns). The accuracy information (precision, recall) is shown in Table 39. One sees that cluster 0 (“house owner”) is assigned in the most of the cases (23.512). Therefore, the recall is very high. Cluster 1 (“starters”) has a low recall, but has a decent precision. The recall is low due to the fact that the click data variable is not available for the most visitors.

Table 38: Scorer for FAQ section

	House Owner	Starter
House Owner	22.638	514
Starter	5.776	1.792

Table 39: Accuracy matrix for FAQ section

	Recall	Precision
House Owner	97,8%	79,9%
Starter	23,7%	77,7%

8.6.3 PROPOSAL FOR CONDITIONAL CONTENT FILTER

Rabobank can add intelligence on the FAQ homepage by implementing the two clusters, elaborated in Table 36 and Table 37. The corresponding classification rules are shown Table 40. When no click behavior is recorded, the engine falls back on only profile information. The rules that apply in that situation are shown in Table 41.

Table 40: Proposal for CCF

Cluster	Rule
House Owner	WON_1E_FASE = (null)
Starter	WON_1E_FASE = 0 OR WON_1E_FASE = 1 OR WON_1E_FASE = 2

Table 41: Proposal for CCF (only profile)

Cluster	Rule
House Owner	EIG_HS_IND = null OR EIG_HS_IND = 1 OR (EIG_HS_IND =0 AND LEEFTIJD >= 35) OR (EIG_HS_IND =0 AND LEEFTIJD < 35 AND HUISH_TYPE_CD >= 11)
Starter	(EIG_HS_IND =0 AND LEEFTIJD < 35 AND HUISH_TYPE_CD < 11)

8.7 BEHAVIOUR ON A PAGE

We use process mining to visualize and generate insight of the behavior on a single FAQ page. In the FAQ, there are pages with multiple questions, displayed in lamellae. The process always starts with opening the page and continues with opening a number of lamellae (≥ 0). The fuzzy miner (Gunther & van der Aalst, 2007) is used to create comprehensible process models. Further, the methodology of research question 3 is used. Therefore, only paths with a high frequency are shown in the figures. With this analysis technique, we answer several questions:

- Which part is the most popular part of the page?
- In which sequence the different parts are viewed?
- Are the sub questions ordered on the page in the same way as the sub questions are viewed?
- How many questions are visited by one visitor on average?
- Do visitors end their visit or continue their visit after seeing a question?

8.7.1 PAGE: REPAYMENT, SAVINGS, DONATIONS AND CHANGE INTEREST RATE

On the page about repayment, savings, donations and changing the interest rate, three sub questions are shown. The process of the behavior on this page is shown in Figure 27. The view of the *question page* is the activity *hyp_asro*. All visitors first open the page, before visiting a sub question. The arrows from the main question are going to the questions about donations, change the interest rate of your mortgage and the question about savings and repayments. The two arrows that occur the most frequent, are the arrows opening the questions about donating and changing the interest rate of the mortgage.

The visitors seem to have two main goals: donate and changing the interest rate of their mortgage. Transitions between the sub questions do not occur very frequent. Therefore, it is better to split up these questions. Visitors with different goals landed on the same page.

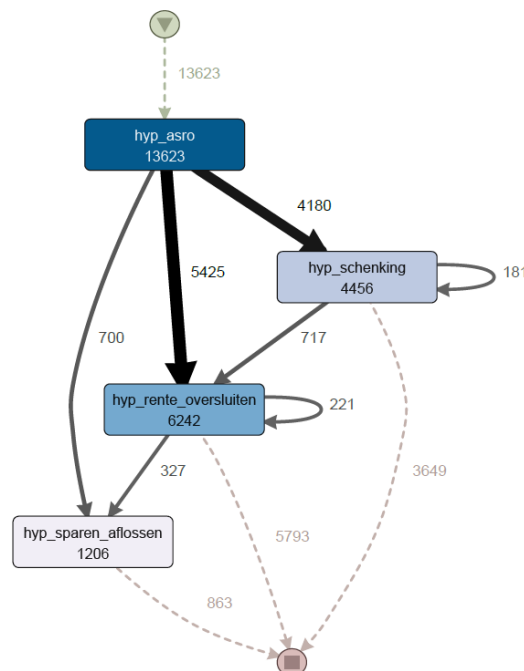


Figure 27: Behavior on repayment, savings, donations and change interest rate

8.7.2 PAGE: CAN I GET A MORTGAGE IN MY SITUATION?

The page about “Can I get a mortgage in my situation” is visualized in Figure 28. Two questions are visited very frequent: *Can I get a mortgage as a starter* and *Can I get a mortgage if I don’t have a permanent contract*. The two arrows that occur the most are going to these two questions. The other questions are visited less frequent and mostly indirect. In Figure 28, it also can be seen that a part ($3.415 / 11.014 = 31\%$) leave the page without opening any lamellae. However, only half of them really leave the website (1.737) (Figure 40 - Appendix).

The two sub questions *Can I get a mortgage as a starter* and *Can I get a mortgage if I don’t have a permanent contract* are the most popular questions. The other questions are less popular and are often visited indirect. It is advised to split up the two most visited questions to separate pages or reorder the questions on the page. The reordering of the questions is based on the frequency of the direct visits (the visit is a direct followed of the main question). The new proposed order of the questions:

1. Can I get a mortgage as a starter (2.512 direct visits)
2. Can I get a mortgage if a do not have a permanent contract (1.866 direct visits)
3. How can my parents help with the purchase of a house? (1.138 direct visits)
4. What is the loan specific for starters (“Wat is een starterslening”)? (942 direct visits)
5. Can I get a mortgage if I have a student loan? (692 direct visits)
6. Can I get a mortgage if I am self-employed? (449 direct visits)

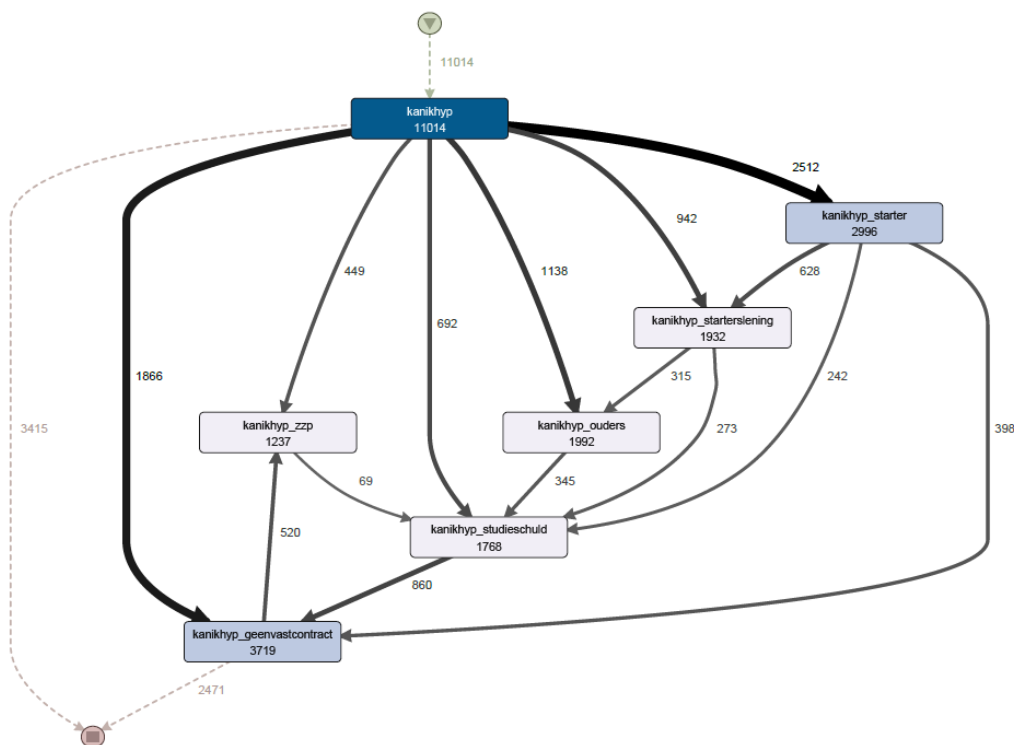


Figure 28: Behavior on “Can I get a mortgage in my situation?”

8.7.3 PAGE: HOW DOES THE MORTGAGE INTEREST TAX DEDUCTION WORKS?

The main question on this page is *HRA* and contains content about the mortgage interest tax deduction. The first thing we notice in Figure 29, is that 49% of the visitors do not open any lamellae. The arrow from *hra* to the stop event has a frequency of 8.334. Furthermore, the *outstanding debt* question has the lowest number of visits. Also, the question about the tax rate is visited less often, but is not located at the bottom.

We propose to reorder the questions on this page. The reordering is based on the ranking of the direct visits. These are the visits of the sub questions coming directly from the *main question*. The new ordering is implemented on the website of Rabobank.

1. Repayment and change mortgage (3.455 direct visits)
2. Am I eligible for mortgage interest tax deduction? (2.398 direct visits)
3. Endowment mortgage (“Spaarhypotheek”) (1.286 direct visits)
4. Tax percentage (1.010 direct visits)
5. Outstanding debt (375 direct visits)

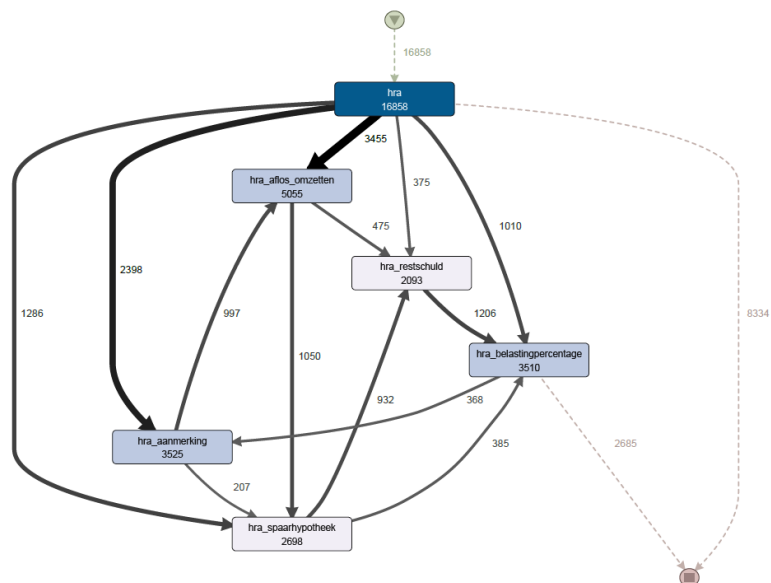


Figure 29: Behavior on “How does the mortgage interest tax deduction works?”

8.7.4 PAGE: WHICH ISSUES ARE INVOLVED WHEN SELLING MY HOUSE?

The behavior of the visitors on the question page about selling your house is displayed in Figure 30. The majority of the views of the sub questions are a direct follower of the main question. A smaller part of the visitors is “clicking through” all lamellae by the display order on the website. 38% of the visitors leave the page without opening lamellae.

Visitors clearly know what they are looking for: the most visitors directly go to one of the sub questions. The ordering of the sub questions on the page is the same order as the most visitors viewed these questions. There is no indication that this ordering is incorrect.

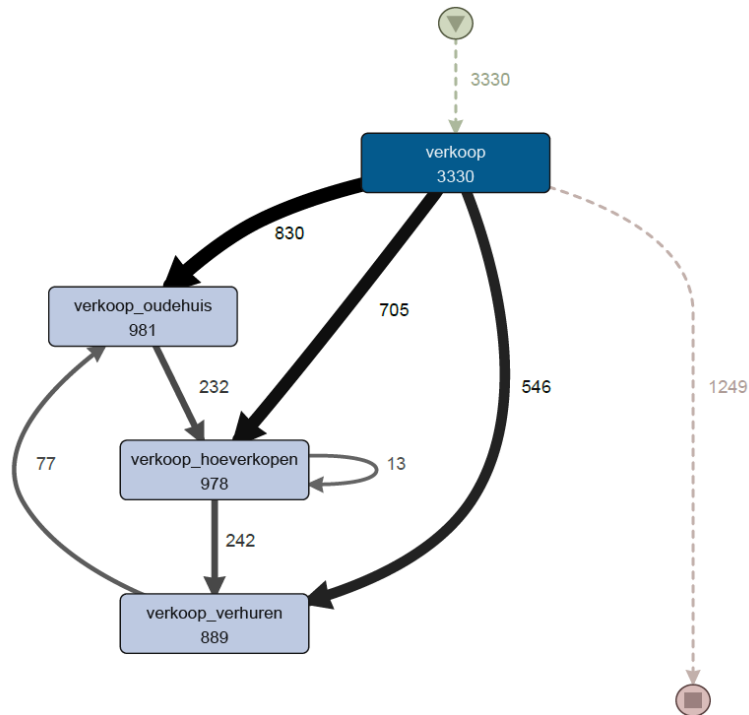


Figure 30: Behavior on “Which issues are involved when selling my house?”

8.7.5 PAGE: WHAT IS THE CURRENT STATE OF THE HOUSING MARKET?

The behavior on the page of the questions about the *housing market* is displayed in Figure 31. On this page, the most information is placed outside the lamellae. In 88% of the cases none of the lamellae is opened. Half of this 88% leaves the website after visiting this question. The most visited lamina is the current situation of the housing market (*hoe*) lamellae. There is not much interaction between the lamellae.

Although the most information on this page is placed outside the lamellae, still a substantial amount, 88%, does not open these lamellae. It is worth investigating why the visitors do not open these lamellae.

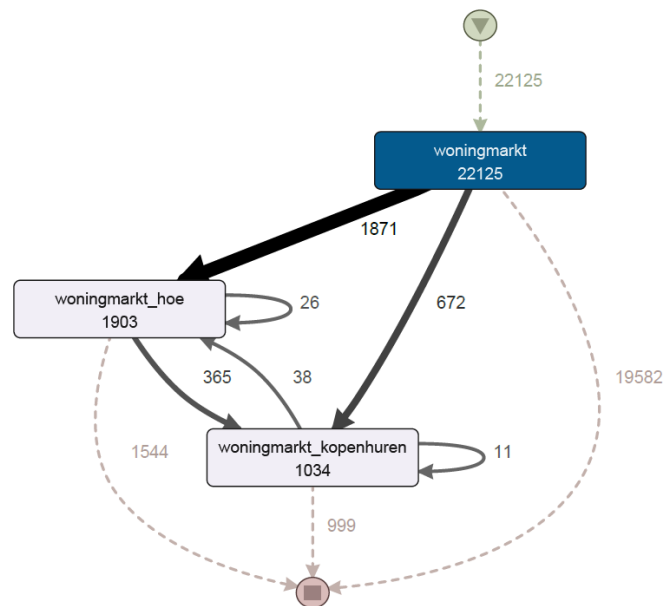


Figure 31: Behavior on “What is the current state of the housing market?”

9 RESULTS

The main question in this thesis is how data mining and process mining can be applied to become more relevant to the end user on websites promoting complex products. The thesis investigates how we can identify different information needs, how we can identify which visitor needs which information and how you present this information. A methodology is provided for each question.

The different information needs of visitors are obtained through segmentation. Much data preparation is needed to obtain relevant clusters. Classification is used to identify which visitor needs which information. Process mining is applied on web data to visualize the behavior of visitors on the website. The most important conclusions for university and Rabobank are listed below.

9.1 CONCLUSIONS FOR UNIVERSITY

Standard process mining techniques result in process models that are hard to interpret when applied on click logs of a website with high volume traffic. This thesis provides two solutions to make process mining more useful when it is used with this web visit behavior data.

Two options to increase understandability and readability of the process models are:

Scoping: To make process mining more useful, the process miner can decrease the scope of the project. For example: by changing the scope to only the behavior on a single page, the number of events decreases. The complexity of the model decreases and this makes the process model more readable.

Group Events: By grouping events into several categories, the number of events decreases. The number of transitions between the events in the model also decreases. This aggregation helps to make the process model more readable.

Collaborative filtering techniques, like segmentation and classification, can be applied on websites promoting a complex product. A website promoting complex product has a customer journey that consists of more steps, visits and has a longer duration than a regular web shop or recommender system. Collaborative filtering is able to deal with large data sets and is suitable to be applied on such websites.

A field test validates that an improvement is possible when applying the decision tree as the statistical classifier. Data filtering and Similarity search with the Nearest Neighbor approach help in preparing the data set to apply segmentation and identify different information needs.

In general, in large organizations, the required data is stored in a variety of systems. Therefore, a significant amount of data preparation is needed to generate useful insights, such as joining data, mapping data, cleaning data and filtering data.

9.2 CONCLUSIONS FOR RABOBANK

This thesis assessed the performance of the *conditional content page* of the mortgage section (mortgages home) and concludes that the accuracy is 86%. The classification engine generates an *overig* event in 8% of the cases, which imply that the classification engine is not reachable at the moment in time.

Accuracy, precision and recall are metrics that are suitable to assess the performance of a classification engine. Accuracy is a measurement of the total performance of the classification engine. Precision and recall can help Rabobank balancing between relevance and market reach: do we want to be relevant to a small group of people or do we want to be less relevant for a large group of people.

The effect of conditional content pages strongly depends on the degree to which the classification engine classifies correctly. The *conditional content pages* on the website have a positive influence in redirecting the visitors to the relevant pages when the classification engine targets the visitor correct. However, a wrong decision of the classification engine causes a decrease in the visits of relevant pages. It is better to not classify at all then classify a visitor in a wrong category. This implies that Rabobank must have a classifier with a high precision.

Data mining is suitable to create or improve classification rules for new or current *conditional content pages*. This research resulted in two insights: (1) The current classification engine is improved with the use of a decision tree learner and this improvement is validated with a field test; (2) New segments and related classification rules are obtained through clustering and a decision tree learner. More specific, in the FAQ section of the mortgage section, two clusters are identified with related classification rules. The most relevant predictors in these classification rules are the variables that contain the click behavior of a visitor. Click behavior has a high predictive power in this situation.

To mine process models that are understandable and useful, data preparation is needed. Applying process mining on the click logs of the mortgage section, without sufficient data preparation, results in complex process models. To increase the usefulness of these process models, the process miner can use two methods: scoping or categorizing page. A small scope increases the usefulness of the obtained model. For example: process mining is suitable to analyze the behavior of a visitor on only one page. The events on a page can be visualized in a process model. Another method to increase the usefulness of process mining is categorizing pages. This categorization decreases the size of the mined process model.

9.3 RECOMMENDATIONS FOR RABOBANK

It is recommended performing an A/B test on the conditional content pages to answer the question if these conditional content pages increase conversion rates. In this thesis, it is shown that a correct classification leads to more visits in relevant sections and that an incorrect classification reduces the visits to relevant pages. This does not imply that this causes an increased conversion rate. An A/B test will clearly show the added value of a conditional content page to the conversion rate. Rabobank should further investigate the 8% *overig* events, which indicates that the classification engine was not reachable at the time of classification.

Rabobank should use data mining for future web page design. When designing new conditional content pages, data mining helps in defining the segments and classification rules. Clustering and classifiers help domain experts in designing their web pages and can support their design decisions with data. Through the connection between visitors and profile information, Rabobank is also able to assess the performance of the classification engine.

Although it can help in supporting the decisions with data, data mining is not needed for designing *every* conditional content page. The current classification engine performs well, even before the improvement. It can also be used for auditing purposes and for validation of the clusters and classification rules.

Rabobank should use real time web data to classify visitors. The current classification engine classifies real time, but bases its decision on offline data and uses offline deduced classification rules. The behavior of visitors on the Rabobank webpage has a high predictive power. Now, each 24 hours this behavior is processed before the classification engine can use it to classify visitors. Other data about their customers, such as profile information or use of their services, is refreshed only once a month. Rabobank should investigate how to process web behavior data quicker, ideally within the same visit. Rabobank should also investigate which extensions are possible with respect to web behavior data. The date since the last visit to a web site section could be of a high predictive value as well.

Since the classification performs well, Rabobank can also use the classification engine to prefill forms or guide customers to certain pages. Every form that must be filled in a customer journey cause a risk that a visitor leaves the website. When information is already prefilled correctly, it could be that visitors are less likely to leave the customer journey. It is recommended to investigate if prefilled forms on the website decrease exit rates. It is also recommended to increase the amount of prefilled forms on the website.

Rabobank should investigate how process mining is an addition to current User Experience research. In current User Experience research, the web behavior of a small group of people is analyzed extensively. With process mining, a much larger group can be analyzed more objective. The website logs are not *biased*, but show the real behavior of *all* web site visitors. Through the *events* on a page, a process map can be generated how a visitor behaves on a certain page.

To apply process mining in User Experience research, the detail level of measurements must be increased. In a User Experience research, one sees the behavior of a small group of persons very detailed. When the detail level of measurements is increased, the process mining analyses and current User Experience research are more comparable and are complementary to each other.

10 REFERENCES

- Bezdek, J., Ehrlich, R., & Full, W. (1984). FCM: The Fuzzy c-Means Clustering Algorithm. *Computers & Geosciences*.
- Chapman, P., Clinton, J., & Kerber, R. (2000). *Crisp-DM 1.0*.
- Estivill-Vastro, V. (2002). Why so many clustering algorithms: a position paper. *ACM SIGKDD Explorations Newsletter*, 65-75.
- Gunther, C., & van der Aalst, W. (2007). Fuzzy Mining – Adaptive Process Simplification Based on Multi-Perspective Metrics.
- Hearst, M., Dumais, S., Osman, E., & Platt, J. (1998). Support Vector Machines.
- Jibza, R. (2007). *Recall and Precision*. Retrieved from Creighton:
https://www.creighton.edu/fileadmin/user/HSL/docs/ref/Searching_-_Recall_Precision.pdf
- Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Mehta, M., Rissanen, J., & Agrawal, R. (1995). MDL-Based Decision Tree Pruning. *KDD*.
- Mendling, J., Reijers, H., & van der Aalst, W. (2009). Seven Process Modeling Guidelines (7PMG).
- Montgomery, D., & Runger, G. (2007). *Applied Statistics and Probability for Engineers*. John Wiley & Sons, Inc.
- Nielsen, J. (1999). *Designing Web Usability: The Practice of Simplicity*. New Riders Publishing.
- Pan, B., Hembrooke, H., Gay, G., Granka, L., Feusner, M., & Newman, J. (2004). The Determinants of Web Page Viewing Behavior: An Eye-Tracking Study.
- Quinlan, J. (1993). *Programs for Machine Learning*.
- Ricci, F., Rokach, L., & Shapira, B. (2011). *Recommender Systems Handbook*.
- Rosen, D., & Purinton, E. (2004). Website design: Viewing the web as a cognitive landscape. *Journal of Business Research*, 787-794.
- Rozinat, A. (2013, June 28). *Disco Tour*. Retrieved January 26, 2015, from Fluxicon Disco:
<http://fluxicon.com/disco/files/Disco-Tour.pdf>
- Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*.
- van der Aalst, W. (2011). *Process Mining*.
- Verbeek, H. (2014). Retrieved 2015, from Process Mining: <http://www.processmining.org/>
- Zhang, G. (2000). Neural Networks for Classification: A Survey. *IEEE Transactions On Systems, MAN, and Cybernetics*.

I APPENDIX 1 – ETL NETINSIGHT

In this screenshot, the initial query is shown for the import of the web data from Netinsight.

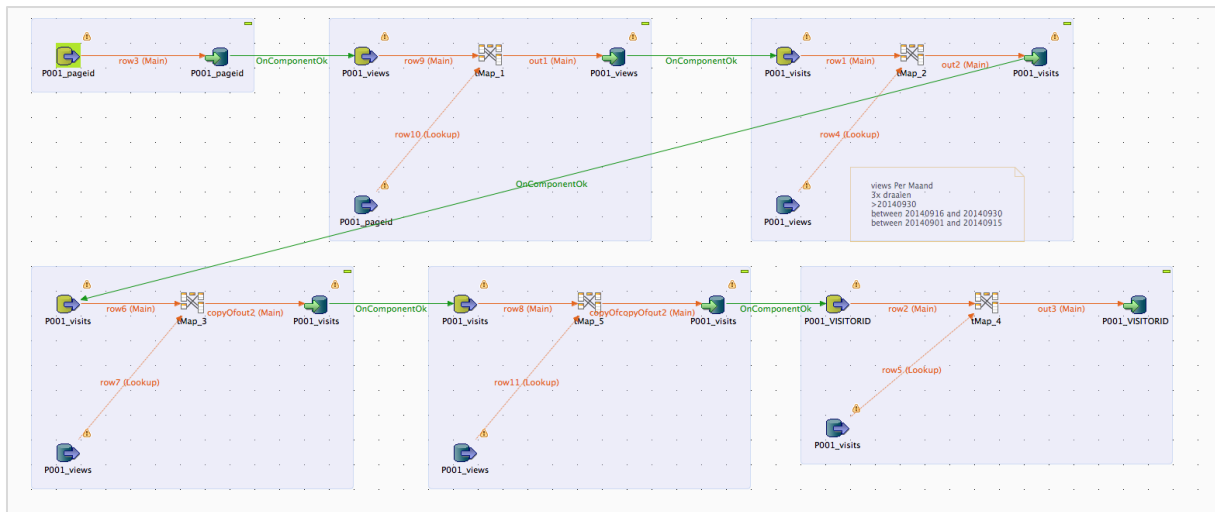


Figure 32: ETL NetInsight

II APPENDIX 2 – META-DATA ABOUT TABLES CASE STUDY

Table 42: Statistics of data tables

Table	Number of fields	Number of Entries	Time horizon
Visitorid	6	406.479	1-9 / 12-10
Visits	105	560.495	1-9 / 12-10
Views	64	2.298.865	1-9 / 12-10
Pageid	6	19.559	All data
Targetgroup	9	1.840.800	All data
Targetgroupid	6	14	All data
Events	8	1.207.786	1-9 / 12-10
Eventlabel	3	1.207.489	1-9 / 12-10
Eventlabelid	2	216	All data
VW_VisitorProfileTable	1.356	3.808.595	Month to date, retrieved on 12-10

III APPENDIX 3 – OVERVIEW OF FAQ

Overview of FAQ

Table 43: Overview of questions and shortcodes

Question	Shortcode
Een restschuld. En dan?	restschuld_algemeen
Ik ga niet verhuizen en verwacht een restschuld	restschuld_nietverhuizen
Ik ga wel verhuizen en verwacht een restschuld	restschuld_welverhuizen
Hoe kan ik aflossen, sparen, rente oversluiten op mijn hypotheek?	hyp_asro
Sparen of aflossen op mijn hypotheek?	hyp_sparen_aflossen
Kan ik aflossen met een schenking?	hyp_schenking
Kan ik mijn rente oversluiten naar de huidige rentetarieven?	hyp_rente_oversluiten
Hoe staat het met de woningmarkt? Ga ik kopen of huren?	woningmarkt
Hoe staat het met de woningmarkt?	woningmarkt_hoe
Nu een huis kopen of huren?	woningmarkt_kopenhuren
Kan ik een hypotheek krijgen in mijn situatie?	kanikhyp
Kan ik als starter een hypotheek afsluiten?	kanikhyp_starter
Wat is een starterslening?	kanikhyp_starterslening
Hoe kunnen mijn ouders helpen met de aankoop van een huis?	kanikhyp_ouders
Kan ik een hypotheek afsluiten met een studieschuld?	kanikhyp_studieschuld
Kan ik een hypotheek afsluiten zonder vast contract?	kanikhyp_geenvastcontract
Kan ik een hypotheek krijgen als ZZP'er?	kanikhyp_zzp
Hoe is de hypotheekrenteaftrek geregeld?	hra
Wanneer kom ik in aanmerking voor hypotheekrenteaftrek?	hra_aanmerking
Kan ik mijn aflossingsvrije hypotheek omzetten?	hra_aflos_omzetten
Kan ik mijn spaarhypotheek ophogen of veranderen?	hra_spaarhypotheek
Mag ik de hypotheekrenteaftrek over mijn restschuld aftrekken?	hra_restschuld
Tot welk belastingpercentage kan ik de hypotheekrente maximaal aftrekken?	hra_belastingpercentage
Wat komt er kijken bij de verkoop van mijn huis?	verkoop
Wel of niet eerst het oude huis verkopen?	verkoop_oudehuis
Hoe krijg ik mijn huis verkocht?	verkoop_hoeverkopen
Mijn huis staat te koop. Is tijdelijk verhuren iets voor mij?	verkoop_verhuren
Waar wordt op gelet bij het afsluiten van een hypotheek?	opgelet
Wat is Nationale Hypotheek Garantie?	nhg
Ga ik verhuizen of verbouwen?	verhuizen_verbouwen
Ik ga scheiden, wat doe ik met het huis en de hypotheek?	scheiden
U blijft zelf in het huis wonen	scheiden_ublijftwonen
Uw partner blijft in het huis wonen	scheiden_uwpartnerblijftwonen
U verhuist beiden	scheiden_beideverhuizen
Wat komt er kijken bij verhuizen naar België of Duitsland?	buitenland

IV APPENDIX 4 – CONVERSION WITH FAQ

Ik wil een afspraak over hypotheek

* is verplicht veld

Welk onderwerp wilt u bespreken tijdens het gesprek? (meerdere keuzes zijn mogelijk)

- Hoeveel kan ik lenen (maximale hypotheekbedrag)
- Eerste woning kopen
- Andere woning kopen
- Woning verbouwen
- Financiële situatie doornemen
- Overig

Aanhef* Meneer Mevrouw

Voorletters* en achternaam*

Geboortedatum* (dd-mm-jjjj)

Heeft u een rekening bij de Rabobank?* Ja Nee

Rekeningnummer*

Telefoonnummer*

E-mailadres*

Stuur mij een bevestiging van deze aanvraag per e-mail.

[Wat doet de Rabobank met uw persoonsgegevens?](#)

Verzenden

Figure 33: Make an appointment

Uitgebreide berekening in het Rabobank Hypotheekdossier

Maak nu uw berekening in het Rabobank Hypotheekdossier als u:

- plannen heeft om een huis te kopen
- zich afvraagt of oversluiten van uw hypotheek interessant is
- verbouwingsplannen heeft en wilt weten hoeveel u extra kunt lenen



Start uw Hypotheekdossier door de onderstaande stellingen te kiezen die het meest van toepassing zijn op uw situatie. Dit zijn er maximaal 4.

Uw persoonlijke Rabobank Hypotheekdossier

Ik heb een koopwoning Ik heb géén koopwoning

Naar uw persoonlijke Hypotheekdossier >

Figure 34: RHD Start

 **Afspraak maken**  Help


U bent van harte welkom voor een adviesgesprek bij de Rabobank. Wij willen ons goed voorbereiden zodat wij u een passend advies kunnen geven tijdens het gesprek. Voor nu, maar ook in de toekomst.

Uw adviseur bekijkt daarom graag de gegevens in uw persoonlijke Rabobank Hypotheekdossier. Wilt u hiervoor uw gegevens in de veilige omgeving bewaren? U bewaart uw gegevens door u in te loggen. Heeft u al een dossier? Logt u dan in.

[Wat doet de Rabobank met uw gegevens?](#)

Bent u klant en heeft u Rabo Internetbankieren?

Log in met uw bankpas en Random Reader of Rabo Scanner.



Heeft u géén Rabo Internetbankieren?

Log in met uw persoonlijk account.

Heeft u geen persoonlijk account? Maak deze dan eenvoudig aan met uw 06-nummer.

Let op: U verlaat dit scherm. Na het inloggen komt u automatisch terug in uw Rabobank Hypotheekdossier.

Figure 35: RHD Make an appointment

V APPENDIX 5 – DECISION TREE FAQ ON PROFILE INFORMATION

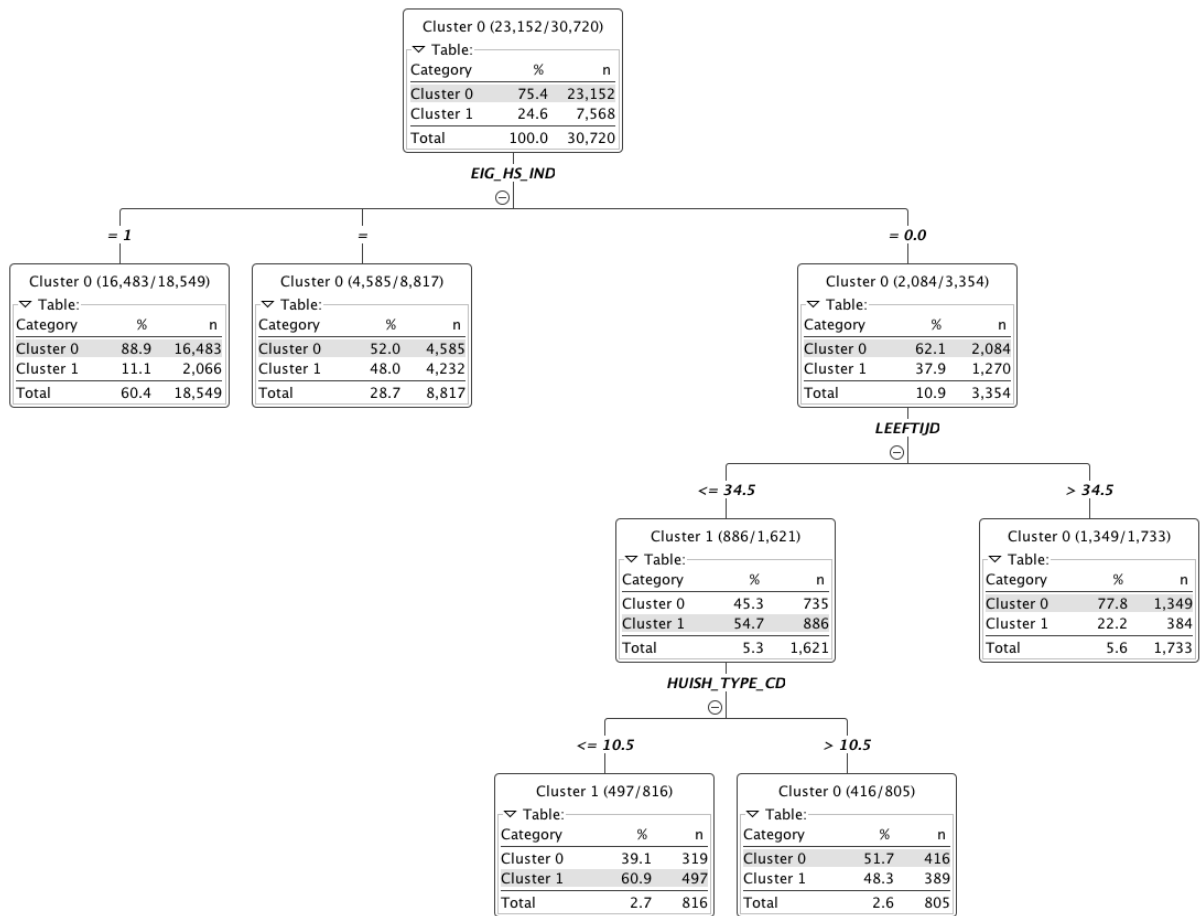


Figure 36: Decision tree for FAQ only on profile table

VI APPENDIX 6 – CLUSTER WITH K-MEANS

Table 44: Cluster 0 - house owner (top 6) – k-means

Question	Score (Expected number of views)
Hoe is de hypotheekrenteaftrek geregeld?	4,03
Hoe kan ik aflossen, sparen, rente oversluiten op mijn hypotheek?	0,31
Hoe staat het met de woningmarkt?	0,21
Een restschuld, en dan?	0,05
Hoe is de NHG geregeld?	0,03
Wat komt er kijken bij de verkoop van mijn huis?	0,02

Table 45: Cluster 1 – starter (top 6) – k-means

Question	Score (Expected number of views)
Hoe kan ik aflossen, sparen, rente oversluiten op mijn hypotheek?	1,25
Hoe staat het met de woningmarkt?	1,13
Kan ik een hypotheek krijgen in mijn situatie?	0,99
Een restschuld, en dan?	0,57
Wat komt er kijken bij de verkoop van mijn huis?	0,21
Wat wordt opgelet bij het afsluiten van een hypotheek?	0,15

VII APPENDIX 7 – KNIME MODELS FOR CLUSTERING

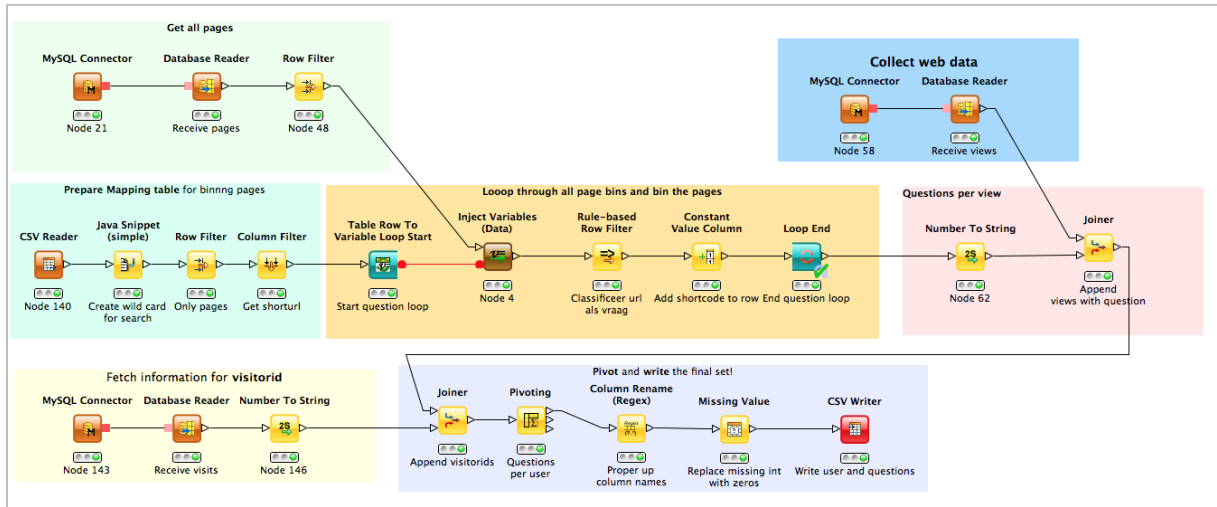


Figure 37: Data preparation phase

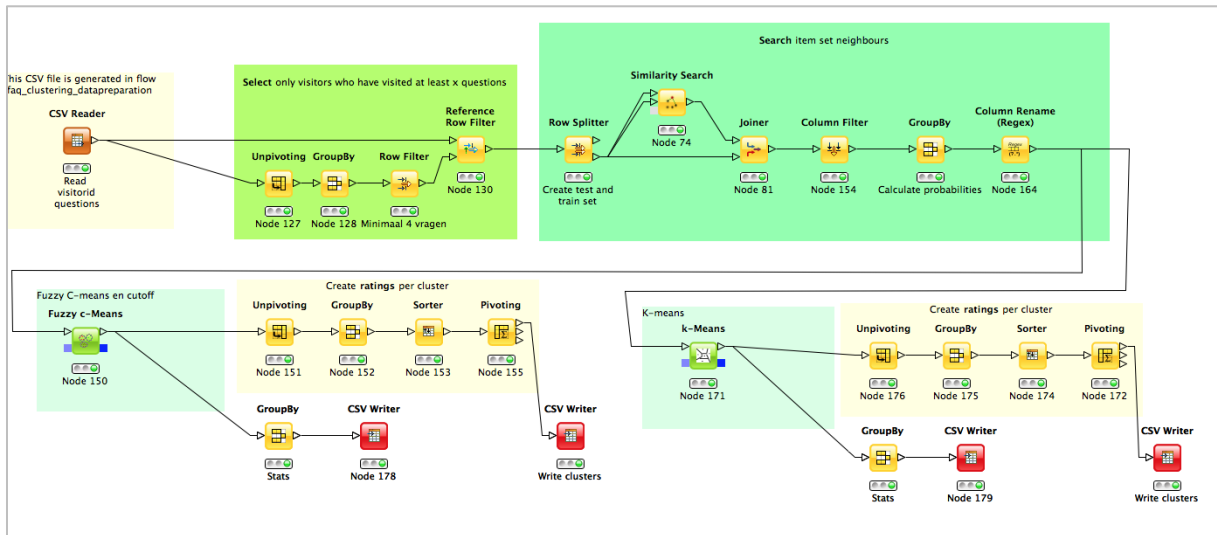


Figure 38: Clustering FAQ

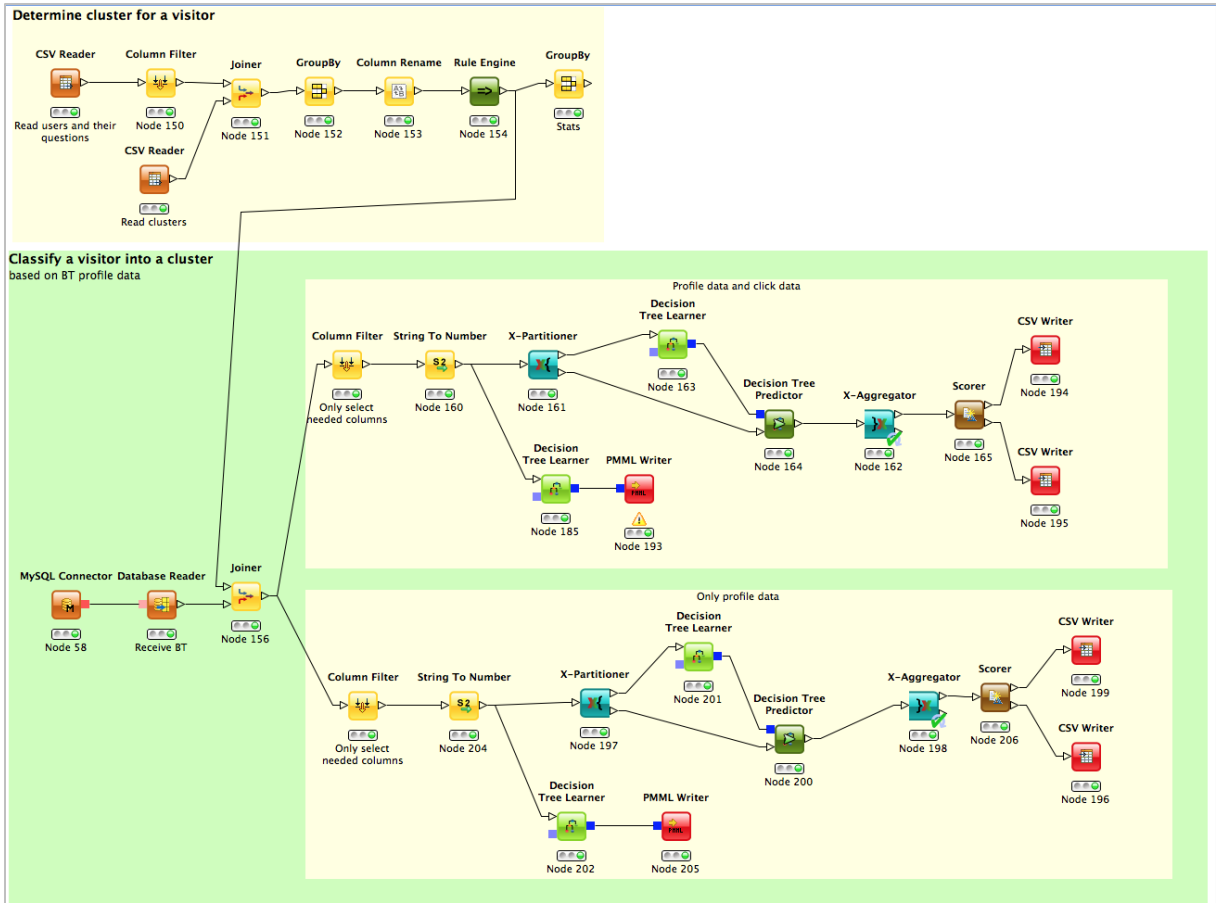


Figure 39: Classification FAQ

VIII APPENDIX 8 – BEHAVIOR ON PAGE

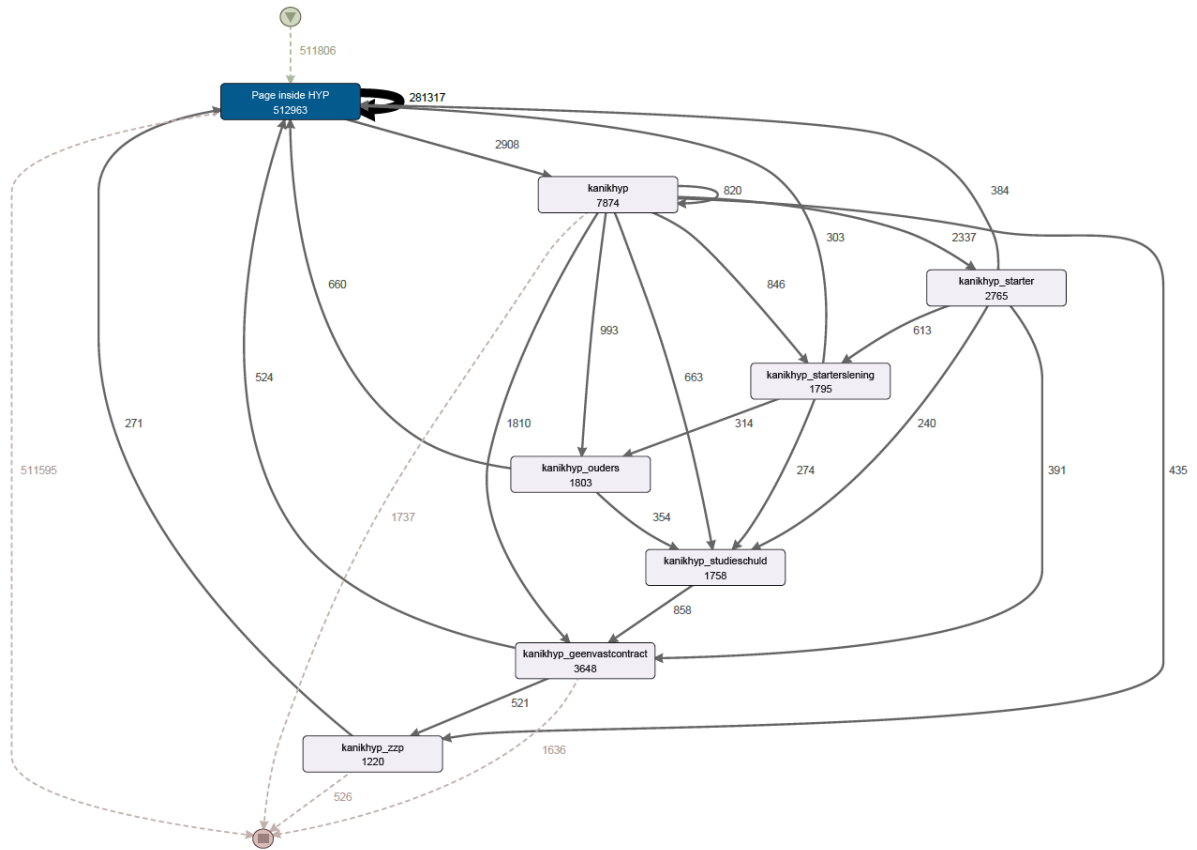


Figure 40: Can I get a mortgage in my situation - visualization with other pages in mortgage section