MASTER'S THESIS | LUND UNIVERSITY 2016

Discovering and Inducing Rules to Categorize Sales Personnel

Lisa Stenström, Olof Wahlgren

Department of Computer Science
Faculty of Engineering LTH

ISSN 1650-2884
LU-CS-EX 2016-04

Discovering and Inducing Rules to Categorize Sales Personnel

(Using Salesforce Data and Machine Learning)

Lisa Stenström
dic11lst@student.lu.se

Olof Wahlgren
ama09owa@student.lu.se

January 29, 2016

Master's thesis work carried out at Brisk.

Supervisor: Pierre Nugues, pierre.nugues@cs.lth.se

Examiner: Jacek Malec, jacek.malec@cs.lth.se

Abstract

In sales, it is presumed that the behavior of sales personnel differs depending on what part of sales they are in. However, to the best of our knowledge, there are no studies about conducting a segmentation of sales personnel based on behavioral data from Salesforce, the world's largest Customer Relationship Management platform. Previous research describes how to segment different customers based on their behavioral data, but no one has yet attempted to segment sales personnel. In this thesis, we extracted Salesforce behavioral data about sales staff and clustered them into previously unknown segments. Using a mixture of supervised and unsupervised learning we created six profiles that describe how different sales personnel work in Salesforce. Our findings helped the company Brisk to improve their knowledge about sales personnel.

Keywords: sales, categorization, machine learning, data mining, crm, segmentation

Acknowledgements

We would like to thank Pierre Nugues for all his feedback and invaluable advice. We would also like to thank Hampus Jacobsson for trusting us with the opportunity to work with this thesis at Brisk and Andreas Pålsson for all the great advice and support during our work on the thesis. Finally we would like to thank Siv and John O'Neall for their patient efforts in revising this report.

Contents

Chapter 1
Introduction

This chapter provides the background, context and aim of this thesis.

Customer Relationship Management (CRM) is used for managing a company's interaction with current and future customers. It often involves using technology to organize, automate, and synchronize sales, marketing, customer service, and technical support (Shaw, 1991). One of the most commonly and also globally used products for CRM is Salesforce.

1.1 Salesforce

Salesforce is a cloud-based CRM system, mainly targeted at sales personnel. The system contains data that relate to a company's entire sales process, including its customers, prospects, deals, tasks, meetings etc.

Salesforce contains numerous *objects* that can be altered and used in various ways. The sales personnel who use Salesforce range from salespeople acquiring new customers, to personnel responsible for retaining present ones. Managers and executives also use the platform.

From the large number of objects in Salesforce, the ones of particular interest in this thesis are *leads, opportunities, contacts, accounts, tasks, events* and *cases*. These are illustrated in Fig. 1.1 and described further in the glossaries at the end of this report.

1.2 Brisk

Brisk is a start-up company located in Malmö, Sweden. Founded in 2012, they are developing a system called Brisk.

The Brisk system is a sales-support application that alerts sales personnel to their daily tasks, and presents relevant information about their current or prospective customers. Ac-

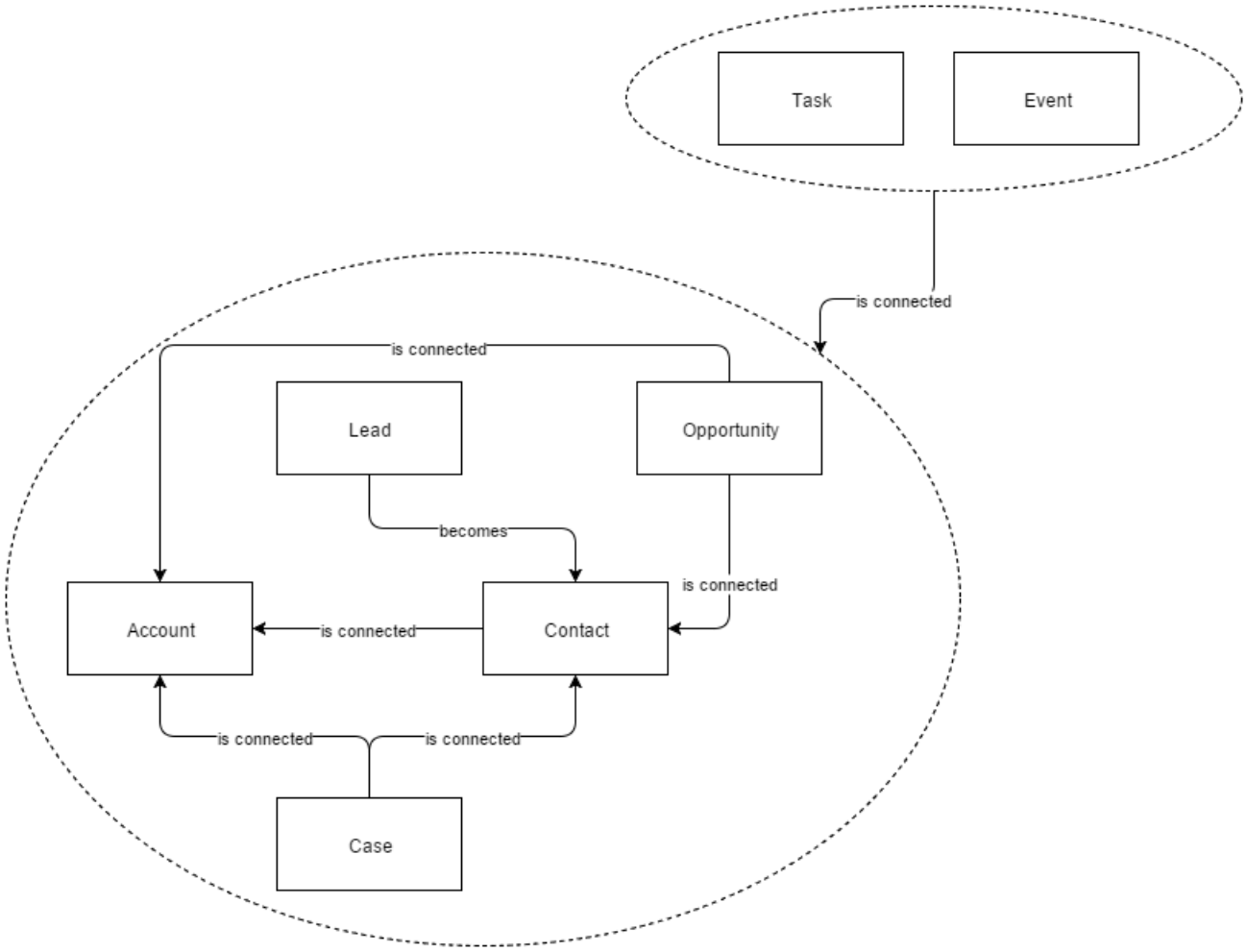


Figure 1.1: The most relevant Salesforce objects and their relation.

tions, tasks, reminders and alerts are some of the features that help the sales person to work more efficiently. Everything is synchronized with Salesforce. The system is implemented as a sidebar application in Google Chrome. It is therefore possible to use it while browsing the web. Figure 1.2 shows the sidebar.

Brisk gathers data from different cloud services, Salesforce being the most important one. In order to use Brisk, a Salesforce account is required. It is also possible to connect accounts from other services, such as LinkedIn, Gmail, Skype, and Evernote. These services add new content to the application.

1.3 Problem Definition

The staff of Brisk wants to gain insights about their customers. Their motive for this is twofold. Insights about customers might facilitate:

Product quality for new users. If user segments are known, new users can be categorized into one of them. A categorized user can be treated better. In the case of the

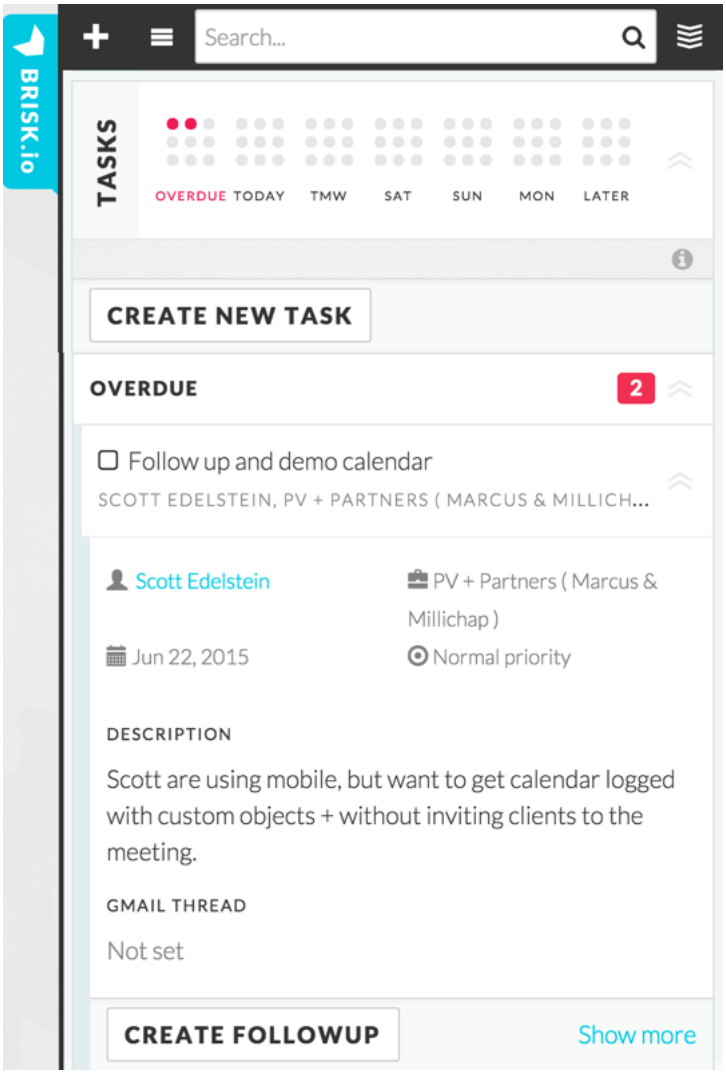


Figure 1.2: The Brisk application.

Brisk application, this means that the user can get customized system settings. If the settings fit the user's needs, the user experience will be better.

Future product development. Segmenting users might lead to improved insight concerning users' needs, problems and behavior. This might improve chances of delivering a product of higher quality or better suited to the customer's needs.

1.4 Related Work

To the best of our knowledge, no investigation concerning the classification of sales staff based on behavioral data from Salesforce has previously been attempted. Kim et al. (2006) performed customer segmentation on CRM data, but they segmented customers based on their lifetime value, which does not relate to behavior. The study aims to determine who would be the most profitable customers.

In a literature study by Ngai et al. (2009), the relation between classification problems

in machine learning and CRM is investigated. The study concludes that data mining is mostly used as a basis in business decisions. It also claims that it is of great importance that the information be easy to interpret.

Research by Ngai et al. (2009) indicates that the two most commonly used data mining algorithms when using CRM data are neural networks and decision trees. Curram and Mingers (1994) compares neural network and decision trees and concludes that the decision tree:

> is more transparent [...] and may give some insight into the relationship of the factors.

This is confirmed by Maimon and Rokach (2008). They claim that as long as the number of leaves are limited in a decision tree, it provides business intelligence. Also the decision tree may be used to create rules (Maimon and Rokach, 2008).

Machine learning algorithms for customer segmentation using CRM data are also described in the book by Tsiptsis and Chorianopoulos (2009). It focuses on segmentation applications in banking, telecommunications and for retailers. It uses clustering algorithms combined with decision tree algorithms to provide a better understanding of customer segments.

1.5 Research Questions

This thesis aims to segment sales personnel into relevant segments. In doing so, previously unknown segments have to be derived and analyzed. A final classifier will also be derived.

Based on Salesforce behavior data this thesis examines:

- *What the relevant segments of sales personnel are.*
- *A good approach for segmentation.*

1.6 Steps of thesis

This thesis consists of several steps. These steps along with important decisions that we made concerning them, are described more thoroughly in Chapter 3. Important steps include:

- Collection of behavioral data from Salesforce. The data is described in 3.3.2.
- Clustering instances, the sales personnel, using two different unsupervised machine learning algorithms. We used DBSCAN and k-means clustering.
- Understanding found clusters by creating a C4.5 decision tree.
- Understanding found clusters by creating segment profiles showing the segments' deviation from average.
- Creating a classifier that can be implemented in production code for Brisk.
- Evaluating the results.

1.7 Limitations

The perhaps most evident limitation in this thesis is time. There are many different approaches, models, and algorithms that we could have examined in every phase of the project. The time limitation of a master's thesis makes the entire project *time boxed*.

The models we have created depend on *feature engineering*. Our modified data resulted in certain types of models. If we had selected other modification techniques, the results would most likely have been different.

The data volume used presents another limitation. We used a total of 3,195 users for creating the models. Extracting this information was a very time-consuming task. An even larger data set might have resulted in even more accurate models.

Other data could have been used to enrich the segments. We examined user behavior in order to find segments with relevant characteristics. In addition, features derived from needs or attitudes could have been used. This is preferably achieved using market surveys (Tsiptsis and Chorianopoulos, 2009). This thesis only uses behavior data from Salesforce, even though other data are available. Other types of behavioral data that could be included is e.g. data from Gmail or data collected from questionnaires.

We examined user behavior during the last 30 days. All behavioral data collected were created within this time frame. Preferably, an even longer time frame could be used. This might result in smoother data with less volatility (Tsiptsis and Chorianopoulos, 2009); but it would have taken more time to collect.

1.8 Contributions

Lisa Stenström was the main contributor to Chap. 2. Olof Wahlgren was the main contributor to Chap. 1. Remaining areas of the thesis were executed entirely in collaboration.

1.9 Structure

Chapter 2 Algorithms *Describes the theory of the different algorithms used in the project.*

Chapter 3 Approach *Describes the workflow used in this thesis.*

Chapter 4 Results *Presents the profiles of different sales staff.*

Chapter 5 Evaluation *Presents an evaluation of the obtained results.*

Chapter 6 Discussion *Discusses the results obtained by the thesis.*

Chapter 7 Conclusions *Discloses the essence of the thesis and the conclusions drawn from it.*

Chapter 2
Algorithms and Mathematical Models

This chapter describes the theory of the different algorithms and mathematical models used in the project. The clustering algorithms are used in Section 3.3 for segmenting the data, and are evaluated in Section 4.1. The decision tree is used for interpreting the segments and can be implemented in production code for Brisk. It is used in Section 3.3 and evaluated in Section 4.3. Principal component analysis (PCA) and Interquartile range (IQR) is used in 3.3.3.

2.1 Clustering Algorithms

The purpose of clustering is to find the natural groups in a data set (Jain, 2010). In the context of this thesis, clustering techniques identify meaningful natural groupings of sales personnel. Segments are based on behavioral data and they group customers with internal cohesion. To create high quality clusters, customers that do not fit into any segment are set apart and therefore not clustered (Tsiptsis and Chorianopoulos, 2009).

2.1.1 *k*-means

K-means clustering is one of the oldest but also one of the simplest clustering methods. It aims to minimize the *squared* Euclidean distance between each instance and its nearest cluster center (MacQueen, 1967). Equation 2.2 shows the equation for the Euclidean distance. An example of the algorithm is shown in Fig. 2.1.

The algorithm partitions an n-dimensional population $X = \{x_1, x_2..x_n\}$ into k subsets $C = \{c_1, c_2, .., c_k\}$(MacQueen, 1967). The objective is to minimize the function:

$$J(C) = \sum_{j=1}^{k} \sum_{n \in S_j} ||x_n - \mu||^2 \tag{2.1}$$

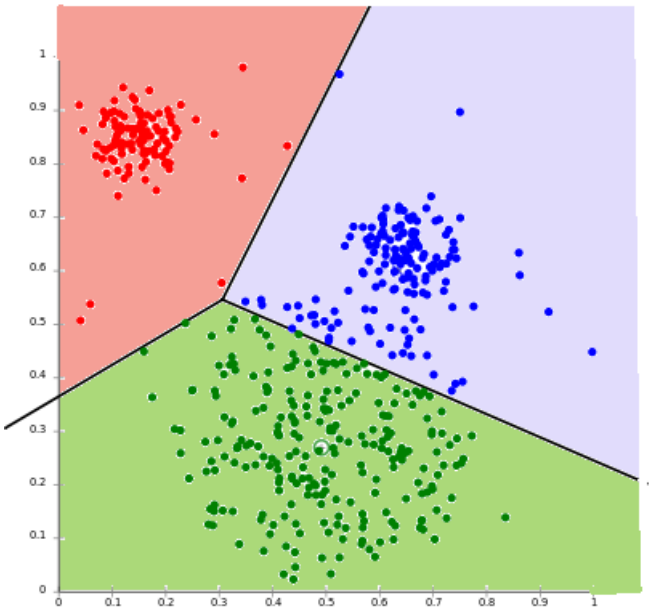


Figure 2.1: An example of the k-means algorithm (Wikimedia-Commons, 2011b).

The algorithm assigns instances and recalculates cluster centers iteratively until their convergence. Cluster centers are calculated as the mean of all instances included in the cluster. The procedure is described below (Jain, 2010):

1. Create k randomly initiated cluster centers $C = \{c_1, c_2 ... c_k\}$.
2. Generate a new partition where each data point is assigned to the closest cluster center.
3. Generate new cluster centers.
4. Repeat step 2 and 3 until the clusters are stable.

The clustering depends on the initial selection of the clusters C. If these are selected poorly, the algorithm may not find global minimums; but only local ones.

In k-means clustering every instance is assigned to a cluster, hence even instances that are considered outliers are grouped into the clusters. As a result, if these are not removed the clustering will not be optimal. Hence extra care has to be put into removing outliers.

2.1.2 DBSCAN

DBSCAN is one of the most popular clustering algorithms. It is frequently used and cited in scientific literature, and relies on a density-based notion. The algorithm is designed to discover clusters of arbitrary shape (Ester et al., 1996). It groups data points that are closely packed together. Points with few neighbors are marked as outliers. These are defined as noise and not clustered. An example of the DBSCAN algorithm is shown in Fig. 2.2.

The DBSCAN algorithm requires the following input parameters:

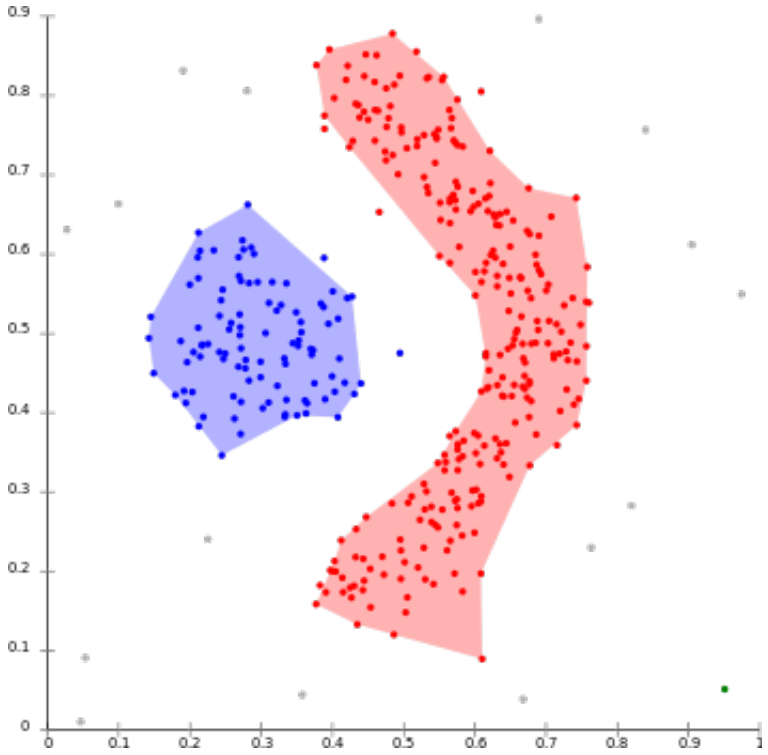


Figure 2.2: An example of the DBSCAN algorithm (Wikimedia-Commons, 2011a).

ϵ: The distance between instances in a cluster.
minPts: The *minimum number of points* required to form a cluster.

The parameter ϵ is closely linked to the distance function selected. In WEKA, the Euclidean distance function is used to calculate the distance. If $p = (p_1, p_2, ..., p_n)$ and $q = (q_1, q_2, ..., q_n)$, and in n dimensions, the function calculated as in Equation 2.2.

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2} \tag{2.2}$$

2.2 Decision Trees

Decision trees can provide an understanding of what characterizes different categories. This is possible because the rules derived are transparent and easy to understand. They do not require any prior data mining knowledge to interpret. Decision trees are therefore good for extracting business insight.

Decision trees operate by recursively splitting the initial population. For each split, they automatically select the most significant predictor. This is the predictor that yields the best separation with respect to the target field. Through successive partitions, their goal is to produce pure sub-segments, with homogeneous behavior in terms of the output (Tsiptsis and Chorianopoulos, 2009).

2.2.1 C4.5 Algorithm

The C4.5 decision tree is a widely used algorithm created by Quinlan (2014). The implementation in WEKA is called J48. The C4.5 tree is created by using the difference in entropy, *information gain*, between attributes. The attribute with the highest information gain is considered the best and is used as a splitting point. Figure 2.3 shows a tree created with the C4.5 algorithm.

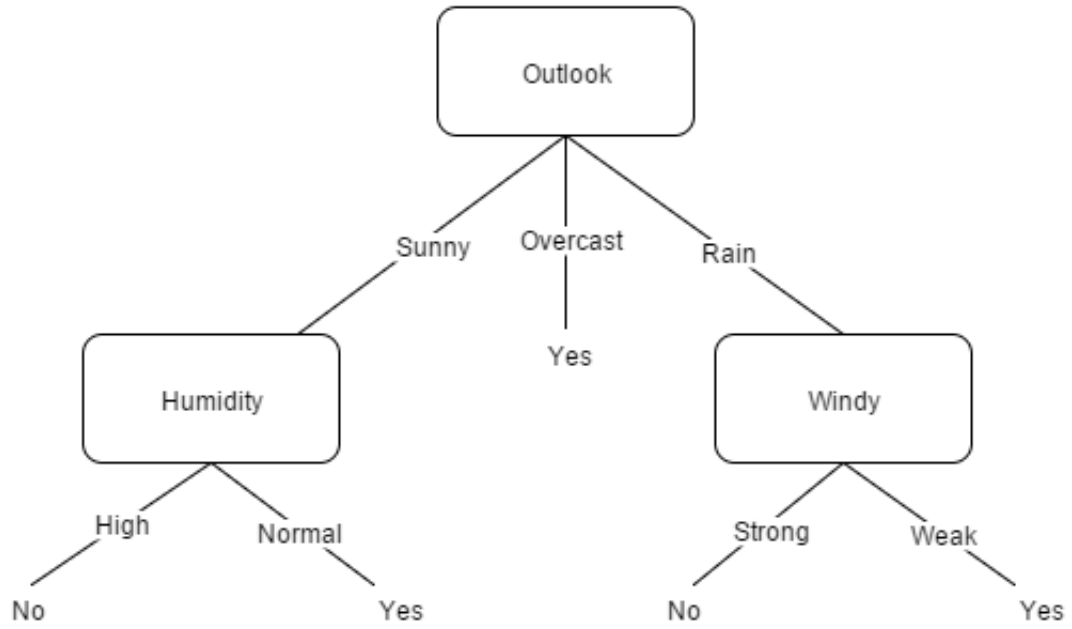


Figure 2.3: An example of the C4.5 algorithm. The data used are from the weather data set, and the tree predicts whether or not to go out and play.

The algorithm checks for different base cases that may occur at a decision node. They can be summarized as:

- All instances have the same class. This results in a leaf with the label of the class.
- All instances have the same input attribute values. Therefore no meaningful split can be performed and the recursion stops.

The actual procedure is performed as follows. The training set $S = \{s_1, s_2, ..., s_n\}$ contains already classified examples. Each vector s_i contains the values of all features and the class of the instance. The algorithm consists of the following steps:

1. Check for base cases.
2. For each attribute a; find the normalized information gain ratio from splitting a.
3. Let a_b be the attribute with the highest normalized information gain.
4. Create a decision node that splits a_b.
5. Repeat the sub-lists obtained by splitting a_b, and add those nodes as children of node.

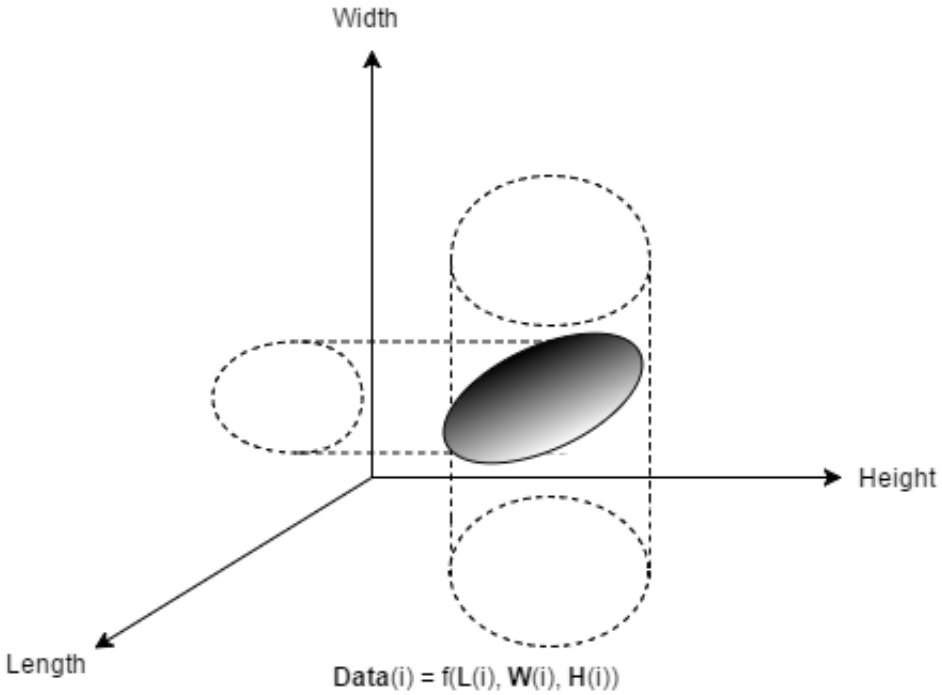


Figure 2.4: First step of PCA. An example of an object described by several features.

2.3 Principal Component Analysis

Principal component analysis (PCA) is a mathematical model created in 1901 by Karl Pearson. It is used for reducing the size of a data set, while maintaining the essential patterns. Figures 2.4, 2.5 and 2.6 visualize the steps of PCA in a simple way.

From a dataset X, PCA aims to find the transformation U that maximizes the variance of these linear transformations Z .

If $Z = XU$ and the number of features is n, the goal is to maximize the function:

$$var(Z) = \frac{U'X'XU}{n-1} = U'SU$$

where $\frac{X'X}{n-1}$ is the co-variance matrix S.

The steps conducted when performing PCA are the following:

1. Standardize or normalize the data set.
2. Obtain the mean value of each column.
3. Calculate the co-variance matrix of X(S).
4. Find the eigenvalues and eigenvectors.

The variance selected will affect the number of eigenvectors that the PCA will result in. In WEKA, the implementation enables a specification of the variance that will be maintained.

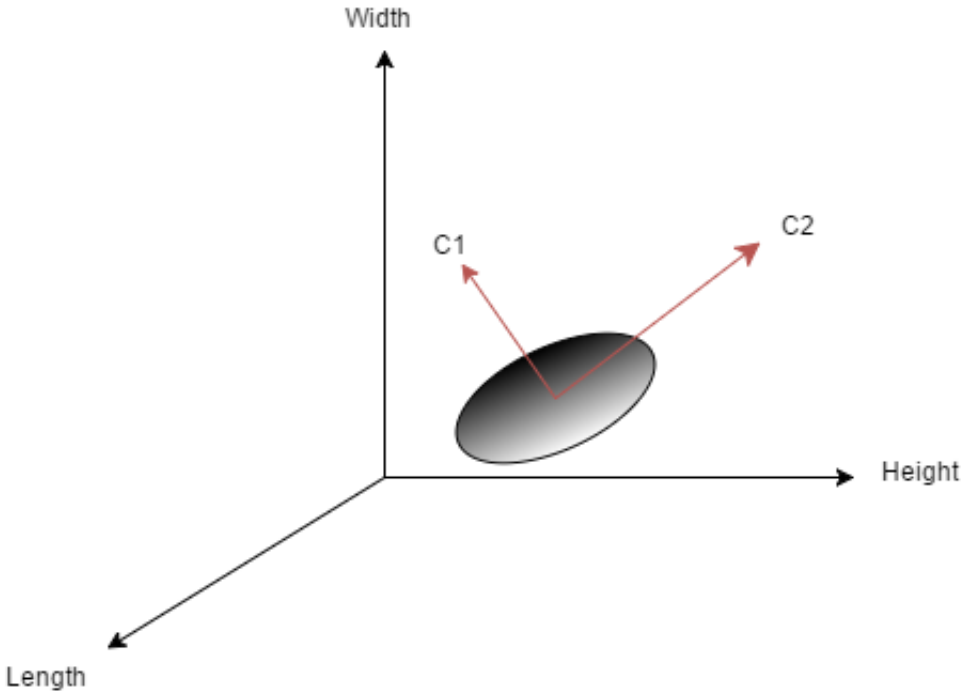


Figure 2.5: Second step of PCA. How the new coordinate axes are found.

2.4 Interquartile Range

Interquartile range (IQR) is an algorithm for measuring statistical dispersion. The algorithm is implemented in WEKA to facilitate detection of outliers and extreme values in a data set. The algorithm takes Extreme Value Factor EVF as input.

EVF Extreme Value Factor
Q1 25% quartile
Q3 75% quartile
IQR Q3 - Q1

Extreme values are defined as in Equation 2.3, 2.4 and are visualized in Fig. 2.7.

$$x > Q3 + EVF \cdot IQR \tag{2.3}$$

$$x < Q1 - EVF \cdot IQR \tag{2.4}$$

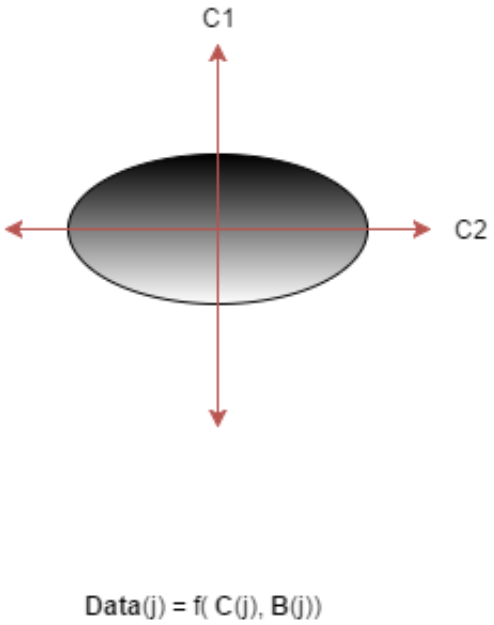


Figure 2.6: Third step of PCA. The new coordinate system, where an object is described by only two features.

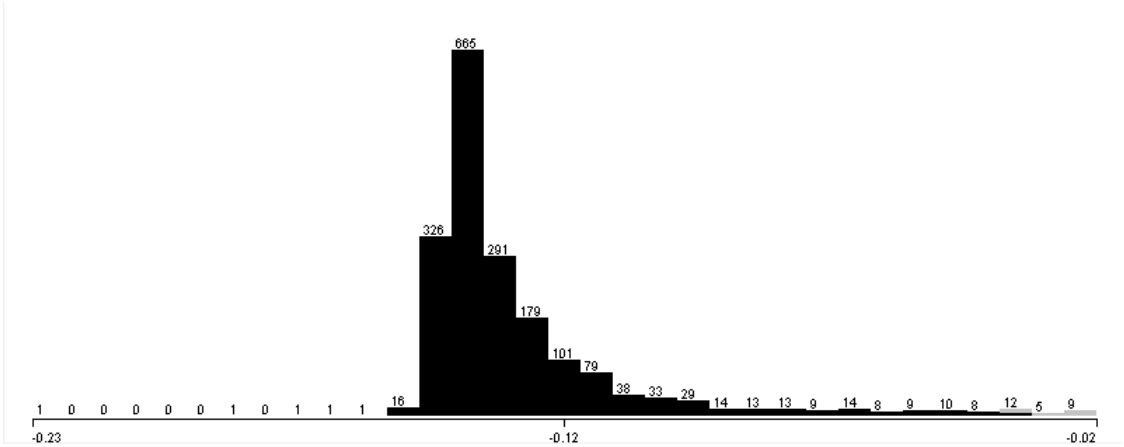


Figure 2.7: Example of how the Interquartile range (IQR) filter defines extreme values in WEKA. The extreme values are grey.

Chapter 3

Approach

This chapter describes all the theoretical background data required and the solution to the research question. The chapter is divided into three parts; Section 3.1, Section 3.2 and Section 3.3. The first section describes the parts of the thesis, the second the selected methodology and the third describes the actual process.

3.1 Machine Learning Challenges

This thesis consists of numerous machine learning challenges. It does not only aim to cluster data. Nor does it merely train a classifier to categorize new instances. In order to create meaningful customer segments and gain insights into them, we use both supervised and unsupervised learning algorithms. Unsupervised learning is used to find hidden structure in unlabeled data. We use both k-means clustering and DBSCAN for unsupervised learning. Supervised learning is used for inferring a function from data with a known output. For this we use the C4.5 algorithm.

3.1.1 Clustering for Segmentation

The first machine learning challenge is a clustering problem. This thesis aims to find different, previously unknown, sales personnel clusters. All features used in the segmentation are constructed from Salesforce behavioral data. We use two different types of clustering algorithms, k-means clustering and DBSCAN, in order to find the unknown segments.

3.1.2 Understanding, Analyzing and Assessing the Clusters

Derived segments must be relevant from a business perspective. This requires that they be understood, analyzed and assessed. Not until then can the segmentation model be validated. We used decision trees in order to create an interpretable model. It was produced by categorizing instances to their respective segment. We determined that the transparent understandable nature of decision trees can provide insights into the different categories.

3.1.3 Training a Classification Model

For valuable insights to be deployed in production code, a classifier is required. It should be easily implemented as code that can be deployed. This enables new users to be categorized into one of the derived segments. As a result, the users will receive a more customized experience, that will be better suited for their needs.

In order for the classifier to be easily implemented, we chose to use decision tree classification also for the classifier. Such a tree can easily be transferred into production code. This is also what the Brisk staff desired as a final classification deliverable.

This classification tree did not necessarily need to be as interpretable as the tree developed for understanding derived segments. Rather, accuracy was the main goal for this final classifier.

3.2 Methodology

Our work flow in this thesis is greatly influenced by two methodologies. These are CRISP-DM and BSM, described in Chapman et al. (2000) and Tsiptsis and Chorianopoulos (2009), respectively. The procedure of these methodologies is described in Sec. 3.3.

3.2.1 CRISP-DM

Cross-industry standard process for data mining (CRISP-DM) is a general, powerful methodology. It forms the basis of the work flow in many different types of data mining challenges. It was created to ensure that data mining was mature enough to be used in business processes. The process is standard in data mining, and is used as an industry tool and in research.

CRISP-DM defines phases for different activities, where the sequence is not rigid. Previous phases are reviewed in each step. This iterative process is illustrated in Fig. 3.1.

3.2.2 BSM

Behavioral segmentation methodology (BSM) (Tsiptsis and Chorianopoulos, 2009) presents a methodology that aims at segmenting customers, based on behavioral aspects. It is similar to the CRISP-DM methodology in many aspects. Like CRISP-DM, it consists of phases to visit and revisit throughout the process.

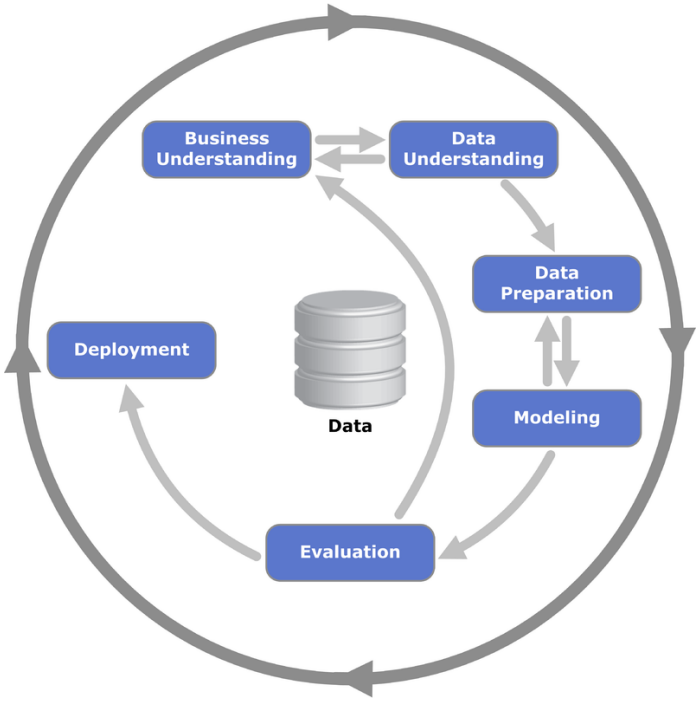


Figure 3.1: Iterative process of data mining using CRISP-DM (Wikimedia-Commons, 2012).

BSM is more specific than CRISP-DM, as BSM describes how to collect and use behavioral data from CRM systems. CRISP-DM formed the basis of our overall thesis workflow. When behavioral segmentation was concerned, we consulted BSM as it adds details to the behavioral segmentation process that CRISP-DM does not possess.

3.3 The Process

This section describes our data mining process. It includes phases and steps from both CRISP-DM and BSM. Important considerations and decisions are described.

3.3.1 Business Understanding

We performed several activities in order to gain business understanding. This phase is an important part of the data mining process, since insights gained will be used throughout the entire project.

Business objectives

We wanted to ensure that the results of this thesis would be relevant to Brisk. In order to achieve this, we decided on business objectives in collaboration with representatives at Brisk.

The business objectives were defined as:

- *Finding relevant user segments based on Salesforce user behavior.*

- *Presenting the segments found in an interpretable way.*
- *Delivering a classifier to classify new users into the right user segment.*

Assessment of situation

We assessed the current situation by examining the different resources, constraints and assumptions. We conducted a lot of discussions with the staff from Brisk. This allowed us to gain further insight into different sales processes. We sketched out the behavior of *assumed* types of sales personnel and their work flows. This rendered an understanding of what type of data would be of interest when extracting data from Salesforce.

Afterwards we read the documentation of the Salesforce data and examined the different objects included in Salesforce. We then evaluated which objects were useful by using the domain knowledge we had already gained from previous steps.

Data mining goals

When defining data mining goals, BSM provided helpful guidance. Tsiptsis and Chorianopoulos state that a successful segmentation scheme:

- Addresses the business objective set by the organization.
- Identifies clearly differentiated segments with unique characteristics and needs.
- Accurately assigns each individual customer to a segment.
- Provides the opportunity for profit in the identified segments.
- Comprises 'identifiable' segments with recognizable profiles and characteristics.

We considered this list throughout the project. When formulating the data mining goals, the basis was constituted by the first, second, third and fifth bullet. We translated the business goals into data mining goals:

- Finding meaningful clusters in user behavior data from Salesforce.
- Creating a decision tree classifier for the clusters.
- Creating a descriptive analysis of clusters.
- Creating an accurate classifier for clusters.

The success criteria for these goals are of both subjective and objective nature. We consider the first three goals to be subjective. They can only be fulfilled if they are proven helpful to the Brisk staff. The CEO and CTO decide if the goals are achieved. We consider the fourth goal an objective success criteria. We selected a relatively arbitrary number for the accuracy, since no baseline was present. The success criteria for the fourth data mining goal was hence set to construct a classifier with an accuracy of more than 80%.

Project plan

We started by creating a preliminary project plan, to facilitate the completion of the project on time. This project plan also made sure all concerned parties agreed on a preliminary

work flow. The plan included the project objective, a schedule containing all the sub-activities, a risk assessment, and a description of the project result. Due to the iterative nature of this thesis, discussed in Sec. 3.3, the project plan consisted of numerous iterations. These were set to last 14 days and to begin with a sync meeting. In the meetings, previous work and future approaches were discussed. For each week we planned separate meetings with the project supervisor concerned.

3.3.2 Data Understanding

During the data understanding phase the goal is to gain insights concerning the data. It is important to collect many types of data and explore where data are missing (Chapman et al., 2000). From this, we could form a hypothesis of what data would be relevant.

Data not related to behavior were omitted, as described by Tsiptsis and Chorianopoulos (2009). Hence, all data used for clustering are directly related to users' behavior in Salesforce.

Collection of data

We started the collection of data by investigating different types of sources. At first we attempted to collect data from Brisk. This was achieved through the use of a third party software that tracks users' behavior on software systems. Every action is logged and can be seen by the company delivering the system. It turned out that the most active users of Brisk used customized variations the system, not accessible to others. As a result user behavior was not comparable. Data from the Brisk application were therefore omitted as a data source in this thesis. After the initial iterations, we decided to use Salesforce as the only source of data.

Different companies have different permission and tracking policies. Therefore, we had to put extra care into deciding what data to fetch. We omitted some features since they contained restrictions. This was disappointing, as some of the omitted features were assumed to relate to user behavior.

We decided that it was important to differentiate the values that could not be fetched from the value of zero. A value that could not be fetched due to e.g. permission restrictions or policies had to be treated accordingly. It should be considered to be a *missing value*. The value zero, on the other hand, should be interpreted as the value zero. The importance of this distinction is supported in a book by Little and Rubin (2014).

The number of queries that could be performed for each user was restricted. Performing too many queries could result in the locking of users' Salesforce accounts. As a result, the final code that fetched the data was optimized to reduce the number of queries.

Description of data

We collected the data from Salesforce using the Salesforce REST API in Java. The data were collected from the Salesforce objects User, Lead, Contact, Opportunity, Case and Task. This provided us with the following data:

- Number of {lead,opportunity,contact,account, case,event and task} objects that a user *owns*.

- Number of {lead,opportunity,contact,account,case,event and task} objects that a user has *created*.
- Number of {lead,opportunity,contact,account and case} objects *edits* that a user has performed.
- Number of lead objects a user has *converted*.
- Number of times a user has *changed the status* of an opportunity.
- Number of opportunities a user has *won* and *lost*.

We collected data from a a total of 3,195 users, and created 33 features. As described in Sec. 3.3.2, only behavioral data were collected for the segmentation. The data collected are a measurement not only of how active each user is, it also describes *what* each user actually works with.

All data fetched were saved in an internal database. This was done since the amount of queries in Salesforce was restricted. As a result queries had to be performed sparingly. Storing data internally led to code that could be altered throughout the entire project. Only when code concerning the queries was altered, did our supervisor have to fetch new data. The database saved user data in JSON-format. This was also the format of user info collected from Salesforce.

Both behavioral and non-behavioral data were collected. For the clustering strictly behavioral data were used. In order to gain insights about the clusters, non-behavioral data were added.

Some new features were created. As mentioned by Tsiptsis and Chorianopoulos (2009); averages, ratios, and percentages are a convenient way to smooth the data that could lead to better models. Ratios were hence created. For the objects lead, contact, opportunity, account, event and case, the *number of edits* and *number of created objects* were related to the *number of owned objects* by the creation of ratios. This is illustrated with two examples:

$$leadEditsPerOwned = \frac{nbrOfLeadsEdited}{nbrOfLeadsOwned}$$

$$opportunityCreatedPerOwned = \frac{nbrOfOpportunitiesCreated}{nbrOfOpportunitiesOwned}$$

If a user did not own any instances of the objects investigated, the denominator for that ratio would get the value zero. These occurrences were given a missing value for the ratio, since it is not possible to divide by zero. This was done, since the ratio did not apply to them. The ratio hence answered the question: *for users that own a certain object; how many of them did he create himself, and how many times on average did he edit each of them?*

Data exploration

The distribution of the data was carefully investigated. We visualized all features, both constructed and original ones, with histograms. The distribution of a typical feature is shown in Fig. 3.2, where the histograms for the number of created leads are shown. The left graph shows the initial distribution, when all instances are used. The right graph shows the distribution when it is zoomed in.

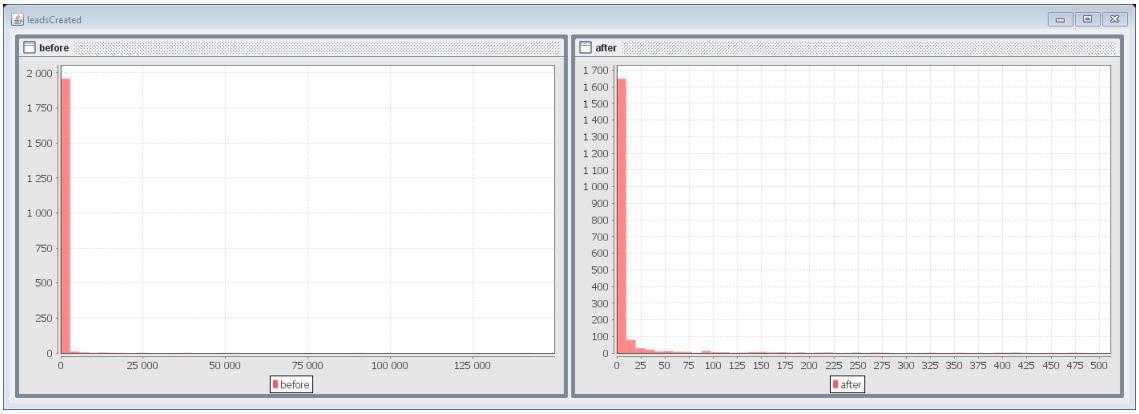


Figure 3.2: Histogram for the distribution of the number of leads that users have created. The figure shows the distribution before and after zooming into the graph. X axis shows the number of users and the Y axis shows the amount of leads created.

We noticed that the data often contained obvious outliers. This can also be seen in Fig. 3.2, in the left graph. All instances but one are so closely distributed, that they end up in the same bin. One single instance deviates a lot. That user has created more than 17,500 leads in the time period of the 30 days used. Instances like these are probably robots or belong to Salesforce accounts that use software as an aid in the sales process.

More often than not, features showed a lot of instances having a relatively low value, often as low as zero. These features often had only a handful of instances with higher values. This can also be seen in Fig. 3.2. The right-side graph is a zoom-in. A lot of instances have created very few leads. The amount of users belonging to each bin decreases as the number of created leads increases.

Outliers could cause poorer models. The huge difference in magnitude between outliers and normal values could result in problems when comparing the different features.

Data quality

Data quality should be high in order to achieve accurate results (Tsiptsis and Chorianopoulos, 2009). We assessed the quality using histograms and the WEKA explorer.

During the initial iterations there was numerous missing values and the data quality was low. This was because the separation between zero and non-available values was poor. As we implemented a better distinction between zero and non available values, data quality improved drastically.

Even with the distinction between zero and missing values, missing values was an issue. Our constructed ratios led to a lot of missing values. Where the denominator was zero, the instance was given a missing value for that ratio. We considered the ratio to be *non applicable* to an instance with a denominator of zero.

We mention in Sec. 3.3.1 that there are access restrictions in Salesforce that a company can enforce. These restrictions are used for limiting the read access for a sales person, and are mainly used as a security precaution. We concluded that the missing values were Missing at random (MAR)Little and Rubin (2014). We guessed that the values were missing

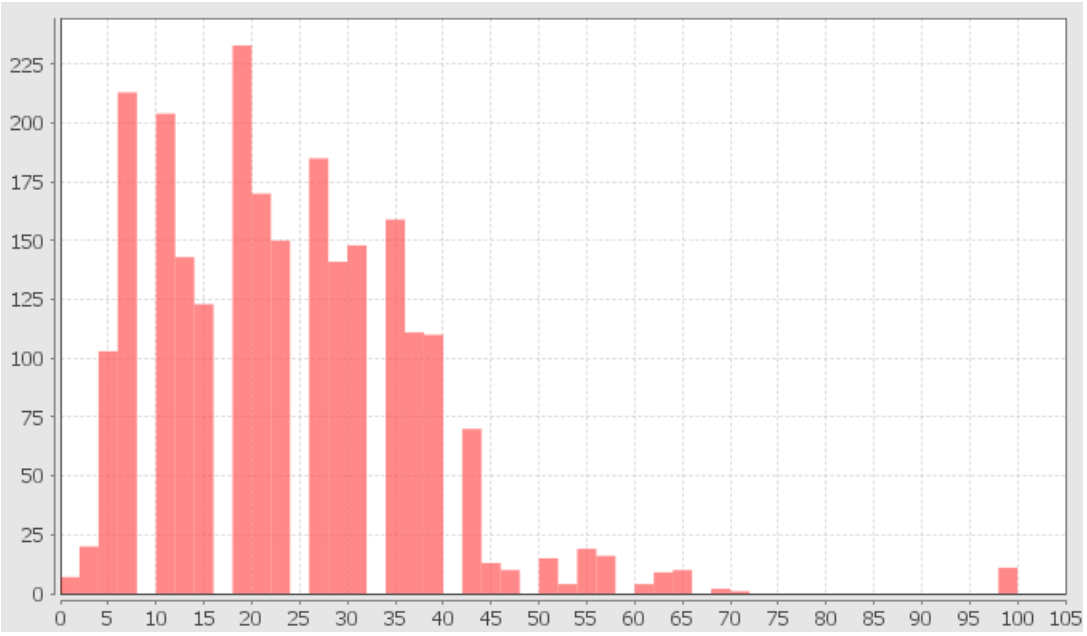


Figure 3.3: Histogram showing the distribution of missing values for users. X axis shows percentage and Y axis shows the number of users.

due to the access restrictions in a company. The missing value percentage for the users is shown in Fig. 3.3. The missing value percentage for the features is shown in Fig. 3.4.

3.3.3 Data Preparation

We chose the correct data, cleaned them, and constructed them into new features during the data preparation phase. This procedure is described in the following section.

Data selection

This thesis aims to segment users based on their behavior, thus all clustering features will relate to user behavior. Although it is tempting to use other data at this stage, features chosen should relate to business goals. We selected a total of 33 features that relate to user behavior.

Data construction

During the data construction phase, new features will enrich the data set (Chapman et al., 2000). We constructed new features using a small Java-program that we wrote. The new features consisted of ratios that injected business knowledge into the data set. We created ratios for e.g. the number of edits per leads that was owned by the user, and how many of the owned opportunities a user had created. We applied this type of feature engineering to the entire data set. The aggregated feature of user activity measure is another example of an engineered feature.

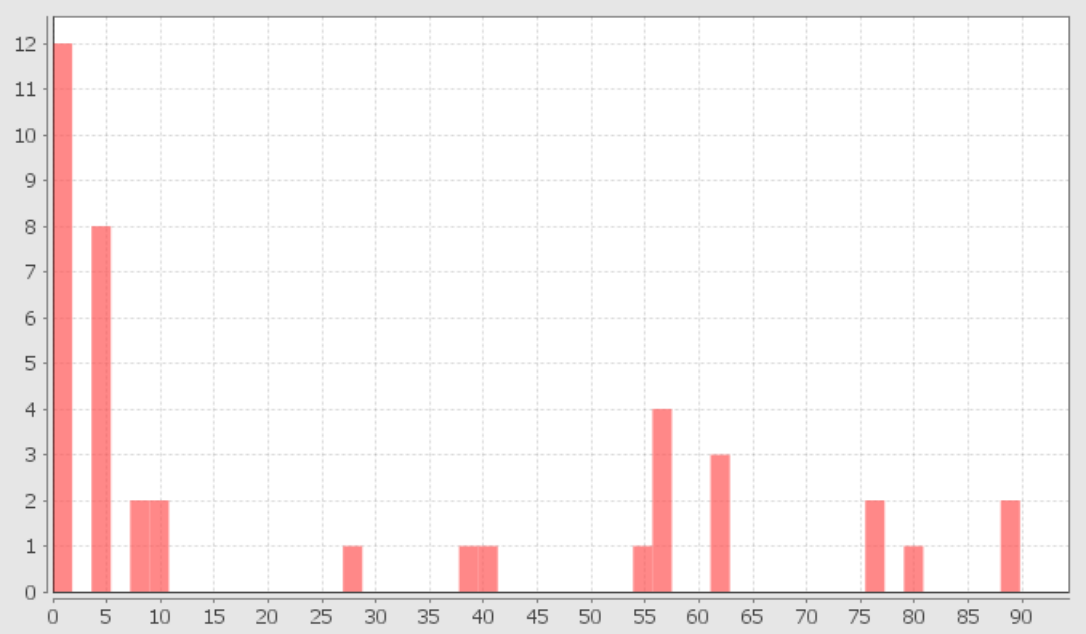


Figure 3.4: Histogram showing the distribution of missing values for features. X axis shows percentage and Y axis shows the number of features.

To maintain a high quality of the data set, we set some rules for the unavailable values when creating new features. All ratios where either the numerator or denominator were non-available, were considered non-available. Ratios having a denominator with the value zero were considered to be *non-applicable*. These ratios were also given a missing value flag.

Data cleaning

Data have to be cleaned e.g. by removing certain features and instances, replacing missing values, and transforming the original data in order to create a good model (Tsiptsis and Chorianopoulos, 2009).

We decided to filter some features out from the selected set, due to insufficient data. We created a histogram that shows how many percent of instances are missing values for each feature. Figure 3.4 shows the histogram. Using this histogram, a reasonable threshold was found. We removed all features lacking values for more instances than that threshold, which was 15%.

In order to achieve an even higher data quality, we decided to remove *instances* lacking more data than a certain threshold. We created a histogram similar to the one described above, that illustrates the percentage of missing values for the different instances. The histogram is shown in Fig. 3.3. All instances lacking more than the threshold selected, which was 40% of their features, were removed.

User activity is another crucial aspect of the data cleaning process. To achieve good clustering, obvious segments such as inactive users, should be removed (Tsiptsis and Chorianopoulos, 2009). We printed histograms that measure a user's activity. This measure was calculated by taking the sum of all object edits, creations, conversion and changes, i.e.

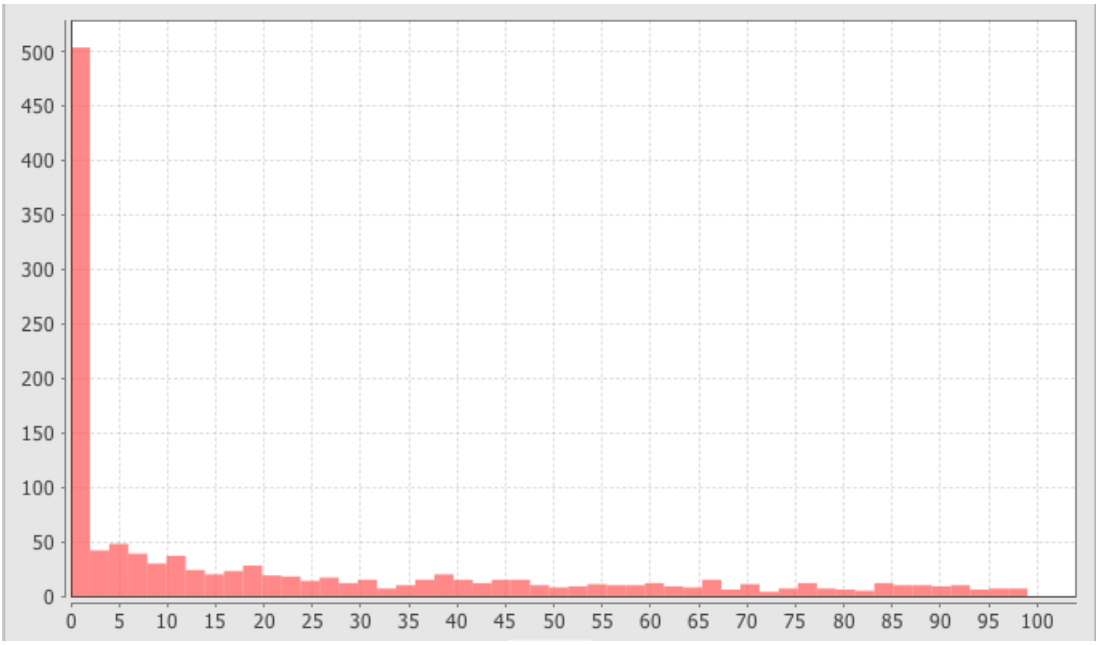


Figure 3.5: A histogram of user activity.

active actions from the user. Figure 3.5 shows a histogram of user activity. It is cropped for illustration purposes. All users with a user activity of less than 20 were removed. This represents an average of one active action per work day.

In order to get relevant data, we only selected users with experience for further analysis. All users with accounts younger than 90 days were hence removed.

After the data cleaning, 21 features and 1,996 users remained.

Finally, missing values were replaced. We wanted to use either mean or median imputation. Normally distributed data are generally imputed using mean values, and median imputation is beneficial when data are skewed (Torgo, 2011). An example of skewness is when there is a small number of very large values. From Fig. 3.2 shows an example of a feature that is far from normally distributed. In fact, almost all of our features showed a similar distribution. We hence used median imputation. This means that every instance that does not have a value for a specific feature, gets the median of the values that are present for that feature.

Data reduction

The data reduction phase aims to reduce the dimensions of the data, while maintaining the information (Tsiptsis and Chorianopoulos, 2009). Once we selected the right population and set of features, the data were exported to WEKA, that is described in 3.4 , for further refinement. First, Principal component analysis (PCA) was performed on the dataset using standardization, as explained in Sec. 2. PCA results in linear combinations of the features. In this thesis, 95% of the variance from the original features is preserved, and the number of features was reduced to 12. These new features looked different from the original ones. Many of them were normally distributed, which facilitated the identification of extreme values.

Secondly, the WEKA filter Interquartile range (IQR) was used in order to detect ex-

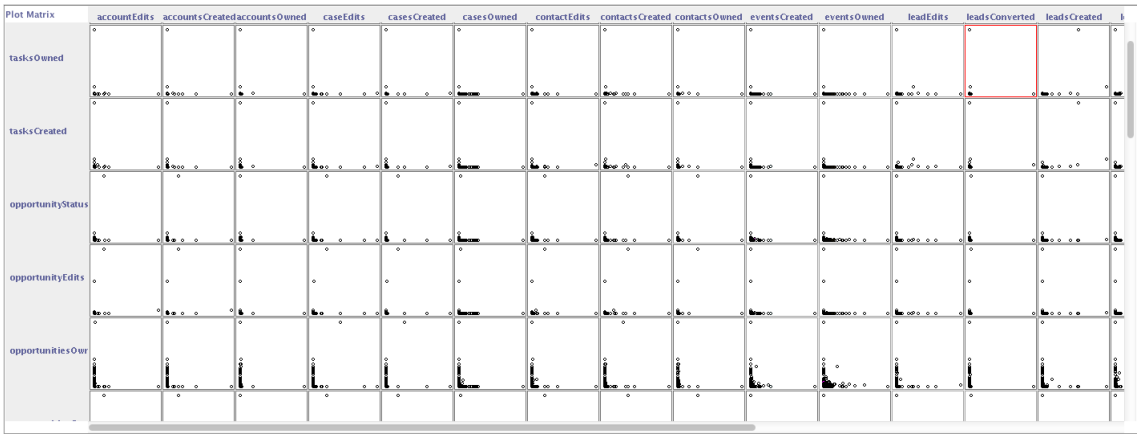


Figure 3.6: Pair-wise plots showing data before preprocessing.

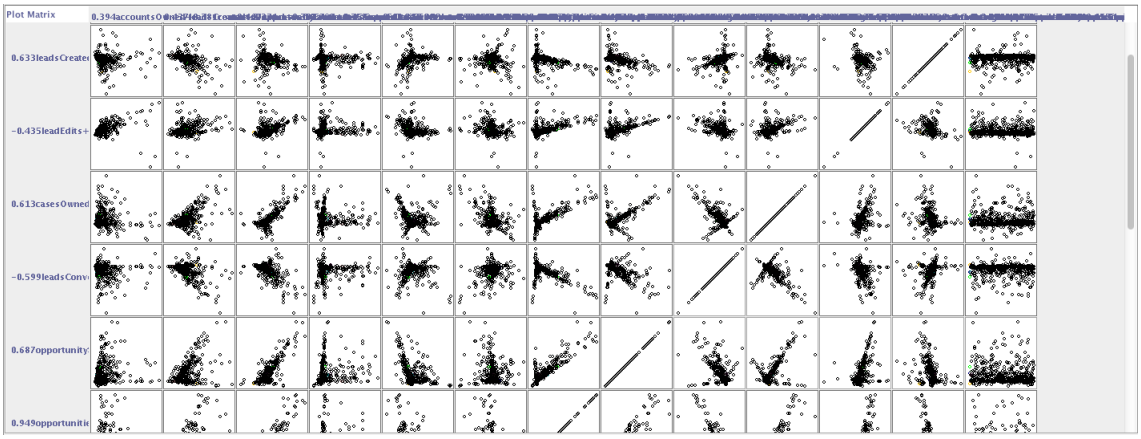


Figure 3.7: Pair-wise plots showing data after preprocessing.

treme values differing too much from the rest of the population, marking them for deletion. We used $EVF = 6$ as input parameter. A total of 1,592 users remained after deletion. Lastly, all features were standardized, leading to all features having a mean value of zero and standard deviation of one. Some of the data before preprocessing is shown in Fig. 3.6. Figure 3.7 shows data after preprocessing.

3.3.4 Modeling

This thesis consists of three stages of modeling; the segmentation phase, the segmentation understanding phase, and the creation of a final classifier. The first stage used machine learning for unsupervised learning by clustering the data set. The latter two stages used models for supervised learning by classifying the data set.

Selection of modeling technique

We tried and assessed many different models for the clustering. For customer segmentation, it is preferable to use a clustering algorithm that does not cluster all instances. The model should determine which instances are non-classifiable outliers (Tsiptsis and Cho-

rianopoulos, 2009). Hence, we made many attempts to use DBSCAN. Not only does the density-based approach fit the needs of the behavior segmentation, it also dismisses outliers as noise. The algorithm did, however, not yield the results we desired. As many instances seemed to have very low values, the density-based DBSCAN often defined one big cluster for these. Instances with values slightly higher were often considered noise. No matter how parameters for the algorithm were altered, no satisfactory clustering could be found. This was considered to be due to the need for a big dataset when using a density-based algorithm. A sparser dataset simply has fewer instances with many neighbors.

As DBSCAN failed to perform satisfactory, we used k-means clustering to segment the data set. We present the results from creating models using DBSCAN and k-means clustering in Chap. 4.

For understanding the customer segments, we created a decision tree. For the final classifier to be delivered, we also created a decision tree, with accuracy as a main goal.

Test design

In this thesis, the testing of the derived cluster models was done manually. We created bar charts, segment descriptions, and a decision tree that we put together into a booklet and showed to the Brisk staff for validation. The approach is iterative and produces valuable feedback from each iteration. Hence, both the cluster- and decision tree models are assessed at the same time.

Construction of model

All models were created using WEKA in an experimental fashion. The parameters used for constructing the different models are presented in Chap. 4.

Identification of the segments with cluster modeling

In order to provide a correct model for the assessment, only interpretable features were selected to be included in the visualization. We determined that the calculated values of ratios were too hard to understand, and were therefore not included. The bar-charts visualizing how the clusters deviated from the mean value are shown in Chap. 4.

Assessment of model

The clusters obtained were initially assessed using our business and data knowledge. To our help, we implemented a Java program that would show bar charts that characterize the derived segment. We also trained a decision tree classifier for derived segments. Following the nodes of an interpretable decision tree, we could gain further insights into the derived user segments. The data mining success criteria served as the main guidance for the assessment.

Once the quality of the obtained segments had increased, the staff of Brisk provided valuable feedback of what was expected of the final product. A few of iterations later, all parties were content with the derived clusters. We then ascertained that the business goals were achieved.

3.3.5 Evaluation

Our workflow included constant evaluation. We deliberately worked in a very iterative manner. Every time a new model was created, it was thoroughly evaluated. This was true in particular for the clustering models. Depending on the results, new strategies for e.g. data collection or processing were formed.

Evaluations of our final segment descriptions were conducted at an early stage. Due to the subjective nature of the success criteria in this thesis, we produced initial versions of our deliverables. These were evaluated by the staff of Brisk. Important feedback was gathered in order to improve the quality of the final deliverables. We also performed a final evaluation sending out a survey to the Brisk staff. The results are presented in Chap.5.

The final classifier was not evaluated to the same extent as the clustering model and the descriptive decision tree. We decided that it would be a waste of time to put effort into creating a relevant classifier for an irrelevant segmentation. We trained and evaluated the final classifier, once the relevant clusters were found.

3.3.6 Deployment

Representatives agreed on receiving the classifying decision tree graphically illustrated. Such a tree can easily be implemented in code.

3.4 Tools

Different tools and programs are used in this thesis. The tools are described below:

Salesforce Sales application and CRM tool that stores data about customers. Required for running Brisk.

Bitbucket Used for configuration management of the written code.

IntelliJ Java development environment. Used for all Java development of data collection, selection and presentation.

WEKA Java based tool for data mining. Used for data visualization, preparation and all machine learning modeling.

Chapter 4
Results

This chapter presents the results from the thesis. It includes the results from the segmentation, profiling and classification.

4.1 Segmentation

This section presents the results from the segmentation phase. From the results in this section it is evident that the DBSCAN algorithm performed less than satisfactory, when segmenting our data set. Out of the models created with k-means clustering, we concluded that it performed the best when $k = 6$. We therefore decided to use that model as the final segmentation model.

4.1.1 DBSCAN

The results from the segmentation using DBSCAN is shown in Table 4.1. The table shows that the results are very poor. Either the number of unclustered instances was too high or the largest cluster too large. We decided that no model performed satisfyingly. Hence, we did not use DBSCAN to produce our final segmentation model.

4.1.2 k-means

We used k-means clustering with different values for k. Each clustering model was closely examined using bar charts displaying clusters and a decision tree that classified the instances into the derived clusters. The within cluster sum of squared errors was also examined, although not used extensively when evaluating models. The sums are shown in Table 4.2. The best clustering was achieved using $k = 6$. These clusters showed clean characteristics that matched our understanding of the sales domain. The derived segments are described in Sec. 4.2.

Table 4.1: Evaluation from DBSCAN. The table shows the percentage of unclustered instances, the resulting number of clusters and the percentage of instances in the largest cluster.

ϵ / minPoints	4	6	8
0.01	Unclustered: 85% Clusters: 13 Largest cluster: 58%	Unclustered: 90% Clusters: 5 Largest cluster: 73%	Unclustered: 92% Clusters: 3 Largest cluster: 82%
0.05	Unclustered: 52% Clusters: 11 Largest cluster: 91%	Unclustered: 56% Clusters: 5 Largest cluster: 96%	Unclustered: 58% Clusters: 2 Largest cluster: 99%
0.1	Unclustered: 26% Clusters: 11 Largest cluster: 93%	Unclustered: 30% Clusters: 8 Largest cluster: 96%	Unclustered: 35% Clusters: 2 Largest cluster: 99%
0.25	Unclustered: 4% Clusters: 2 Largest cluster: 100%	Unclustered: 5% Clusters: 1 Largest cluster: 100%	Unclustered: 5% Clusters: 1 Largest cluster: 100%

Table 4.2: Evaluation of k-means. The table shows the within cluster sum of squared errors with different values of k.

Number of clusters k	Squared error
3	142.8
4	125.3
5	105.9
6	94.7
7	87.7
8	82.0
9	80.0
10	74.7

4.2 Profiles

This section presents the profiles we derived from the customer segmentation. Segments are visualized using an interpretable decision tree and bar charts.

The decision tree was derived by classifying all instances into their derived cluster. The accuracy of the tree was considered to be less important than its interpretability.

The bar charts show how the average values for each clusters deviate from the average values of all the sales personnel examined. A bar of no length does not deviate from the mean of the entire population, and hence illustrate an average type of behavior.

4.2.1 Decision tree for understanding segments

The tree used to understand and assess the derived segments is shown in Fig. 4.1. Again, since the tree is used to understand the derived clusters, its interpretability is crucial. Trees that are too deep and complex are not useful from an understanding point of view.

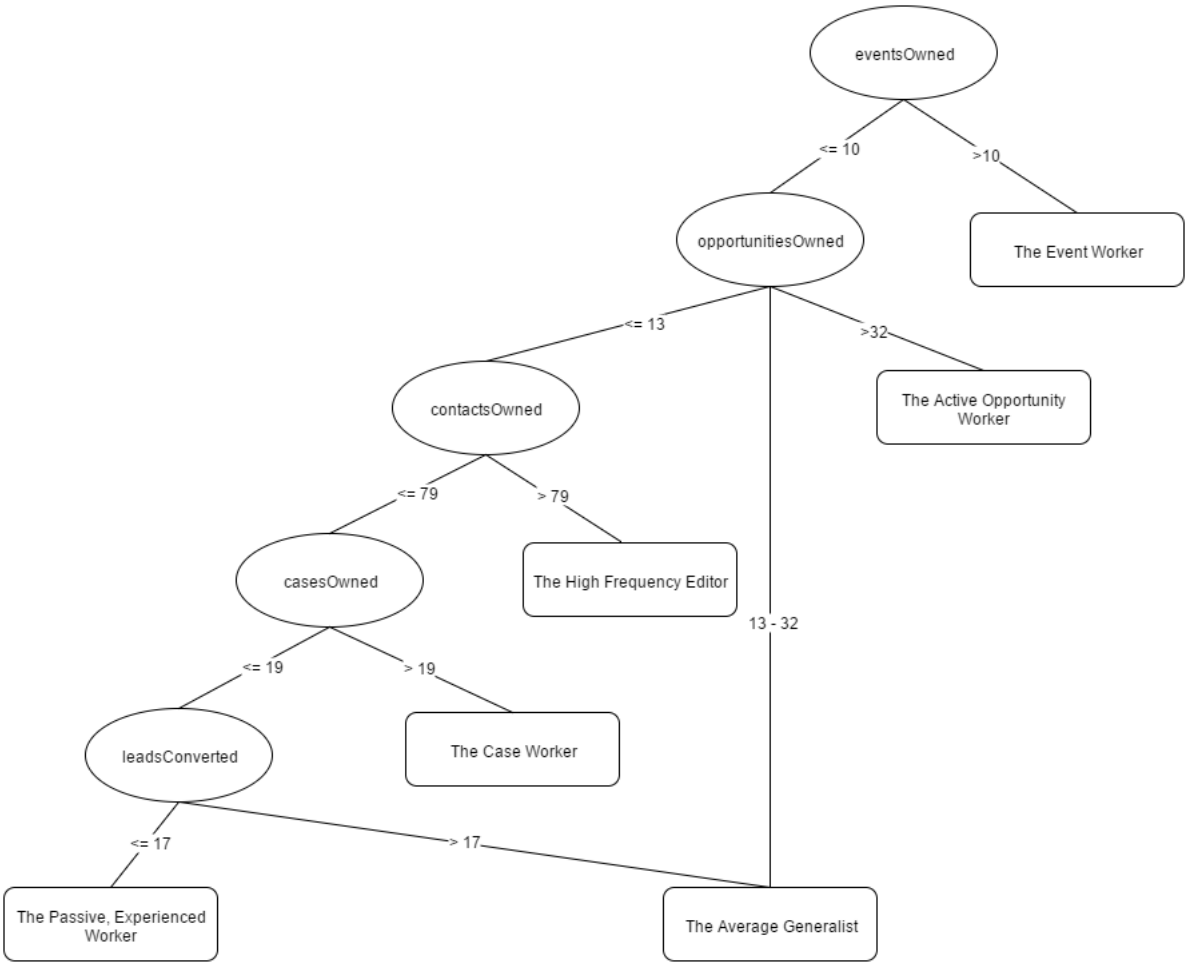


Figure 4.1: The final decision tree used for understanding derived segments.

We derived the tree in an iterative fashion, using the C4.5 algorithm. We did it in steps in order to ensure its interpretability. The input parameter of minimal number of instances of each leaf finally landed at 40. We used 100-fold cross validation in order to get a general classifier and avoid overfitting. The final tree for visualization has the accuracy of roughly 87.6%.

4.2.2 The Passive, Experienced Worker

This largest segment consists of about 59% of the examined users. Figure 4.2 shows this segment. The main characteristic is the low frequency of work. The sales personnel in this segment have a user account that is slightly older than average. They work with various types of work, but with a lower frequency than average. Most common working titles for users in this segment include Account Executive, Account Manager, and Business Development.

Users in this segment might be at a directory level, since they spend less time working

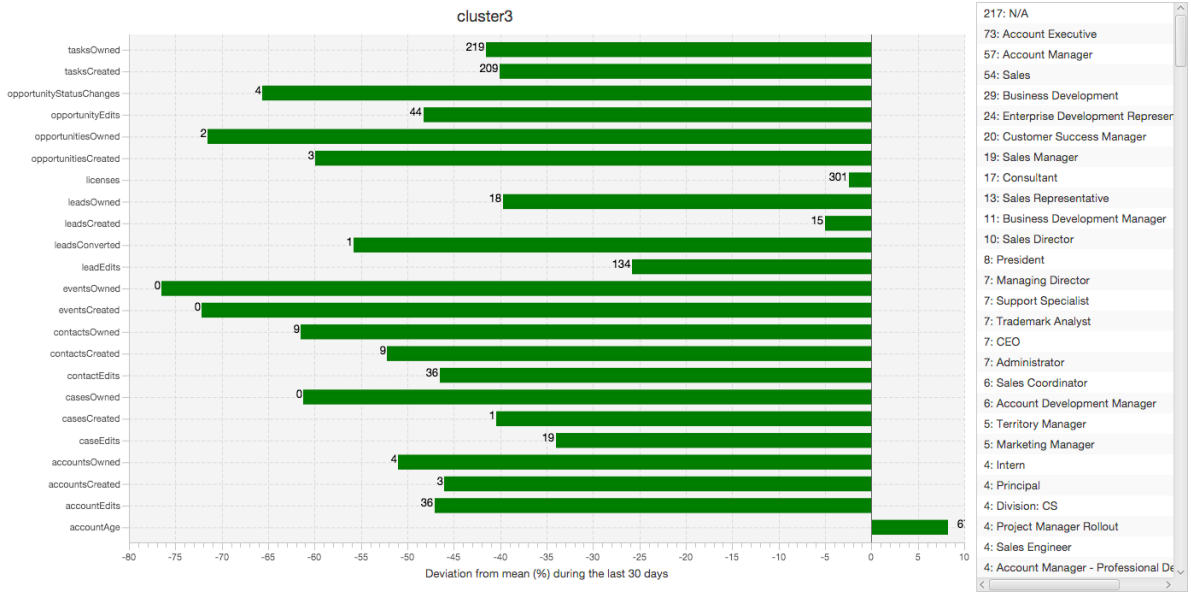


Figure 4.2: The Passive, Experienced Worker. Appendix A shows the image in full size.

in Salesforce compared to other users. The lower activity of this segment may be due to poor adaptation of the CRM tools or less need for one.

4.2.3 The Average Generalist

This second largest segment consists of roughly 16% of all sales personnel examined. Figure 4.3 shows this segment. It is characterized by working with most aspects of sales to an average extent and working with a little bit of everything, except cases and events. The scale on the bar chart indicates that nothing about this segment is extreme. These users have a slightly higher activity on working with opportunities and leads. They work at companies of average size, and user accounts are slightly newer than average. These workers are task-driven but do not work with events. The most common titles for users of this segment are Account Executive, Account Manager, and Enterprise Development Representative.

The general character of these users might be because of an unstructured sales process. The users in this segment seem to be slightly more active than average, but they are not extreme in any way.

4.2.4 The Event Worker

The third largest segment consists of approximately 10% of all users. Figure 4.4 shows this segment. What characterizes this segment is that users work especially with events. Besides working with events the users in this segment are average in all other aspects. They work with everything except with cases, and slightly less with leads than average. The most common titles for users in this segment are Account Executive, Account Manager, and Senior Account Executive.

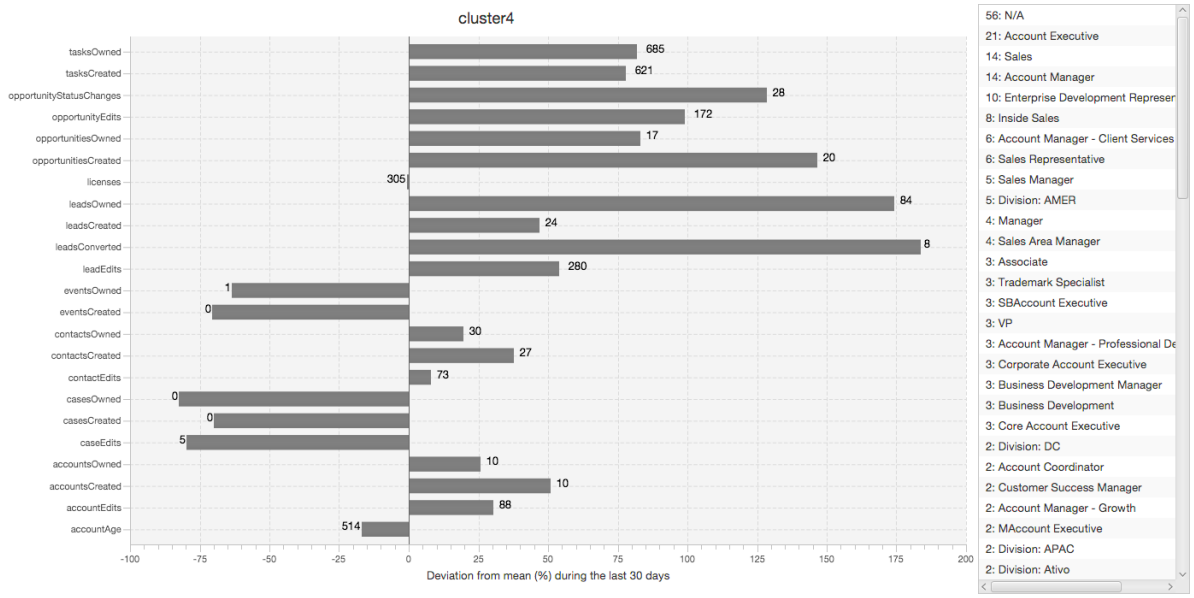


Figure 4.3: The Average Generalist. Appendix A shows the image in full size.

The typical Event Worker probably has a lot of scheduled meetings and are very average in all other aspects.

4.2.5 The Active Opportunity Worker

This segment consists of about 7% of the sales personnel examined. Figure 4.5 shows this segment. The users in this segment work with opportunities to a much higher extent than the other segments. They own many more opportunities than average, and they frequently edit the status of these. These users own a lot of contacts and accounts. Apart from this, their behavior seems to be average. The most common titles of users in this segment are Account Manager, Account Coordinator and Sales Representative.

Users in this segment are likely to be some type of Account Managers, since they own many accounts and contacts and work a lot with opportunities. This is supported by the fact that they work frequently with editing the statuses of their opportunities.

4.2.6 The High Frequency Editor

This segment consists of about 5% of the sales personnel examined. Figure 4.6 shows this segment. The users in this segment are characterized by their extreme edit and creation frequency. Their frequency of working with opportunities is average, but their tendency to edit leads, contacts and accounts is extreme. Interestingly enough, they do not create more leads than an average salesperson. Company size is average and the age of user account is slightly lower than average. Users of this segment are task-driven. The most common titles of users in this segment are Sales Development Representative, Enterprise Development Representative, and Market Development Representative.

The extreme values in this segment may be due to software aiding tools. There are a

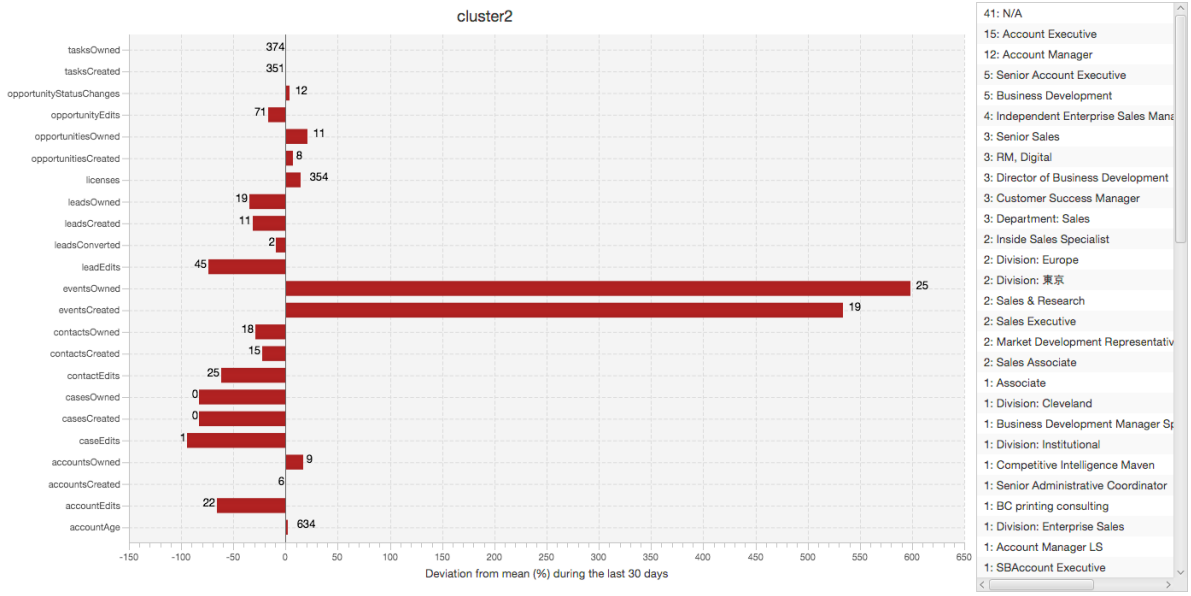


Figure 4.4: The Event Worker. Appendix A shows the image in full size.

number of tools that can be used to aid sales processes. Users of the segment have a process well adapted to Salesforce. Their behavior implies that they work with development of leads, contacts or accounts and are likely to be working with the initial steps of the sales process.

4.2.7 The Case Worker

This smallest segment consists of about 3% of the sales personnel examined. Figure 4.7 shows this segment. Although smallest, this segment has the most extreme user behavior. The users in this segments work with cases to a much higher extent than the other segments. Apart from this, their behavior is very average. The most common titles of users in this segment are Technical Account Manager, Customer Support, and Solutions Engineer.

Users in this segment are likely to work with customers support, since they work with cases a lot more than average. This is supported by the users' titles. Almost all users have a title that implies that they work with support.

4.3 Classification

In order for Brisk to be able to use our derived insights in their production code, new users will be categorized. This requires an accurate classifier. We chose to implement another decision tree. Based on raw, unprocessed features of new instances, it classifies new instances into one of the derived segments with high accuracy. The reason for choosing to use a decision tree for this is their easy transformation into code. Each node could be translated into an *if-statement*. A decision tree was also requested by the company. This final classifying tree has, as opposed to the tree for visualization that we derived earlier,

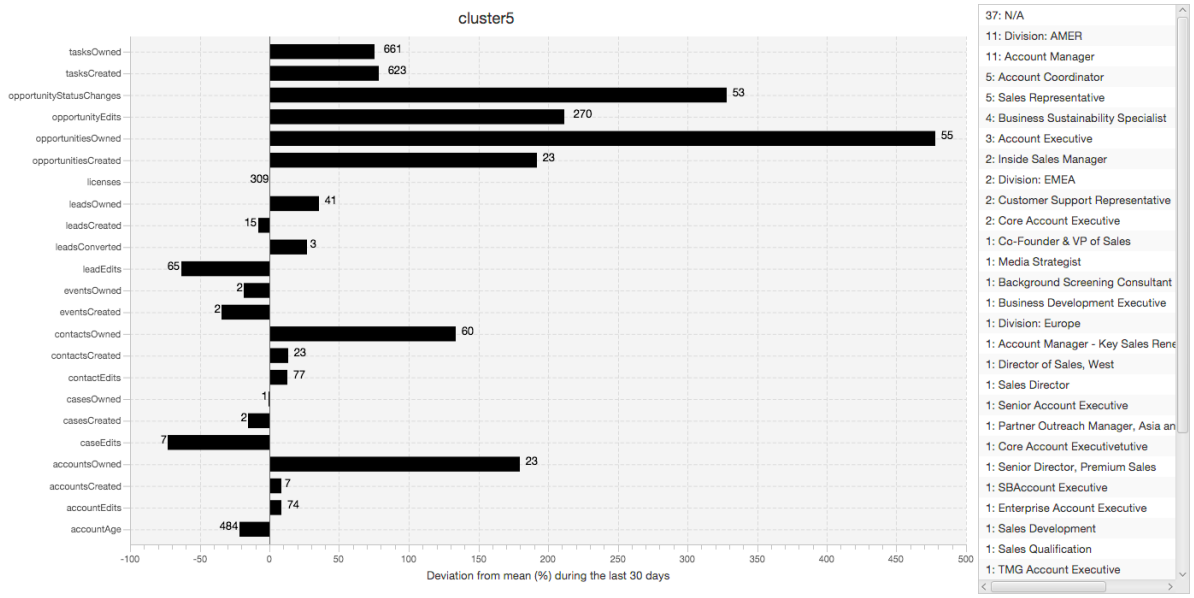


Figure 4.5: The Active Opportunity Worker. Appendix A shows the image in full size.

the goal of categorizing with as high accuracy as possible. It does, however, also need to be simple enough to be implemented into code.

4.3.1 C4.5

When training our final classifier, we agreed with the Brisk CTO to deliver the final classifier as a tree. One goal was derive a tree that could easily be transferred into code. The tree also had to be accurate. This presented a trade-off, since a more complex tree is more accurate but harder to translate into code and vice versa. The development of the final classifier was hence done iteratively with constant assessment.

In order to get a measure of the classifier's accuracy, the entire data set was divided into a training and a test set. We used 66% of the dataset for training and the rest for evaluating the model.

We altered the value of the parameter $MinimumNumberofObjects$ and assessed the resulting tree. The experiments are shown in Table 4.3.

From the experiments, we could conclude that accuracy for this relatively low value of $MinimumNumberofObjects$ was only marginally higher than the much simpler tree derived in Sec. 4.2.1. While slightly more accurate, the trees described in Table 4.3 were substantially more complex. The simplest of them had 23 leaves, while the tree derived for visualization only had 7. The latter was considered to have enough accuracy relative to its complexity. We hence chose to deliver the tree created in Sec. 4.2.1 as our final classifier. The final classifier hence also has the accuracy of 87.6%. This means that the data mining goal of producing a classifier with an accuracy of at least 80% was fulfilled.

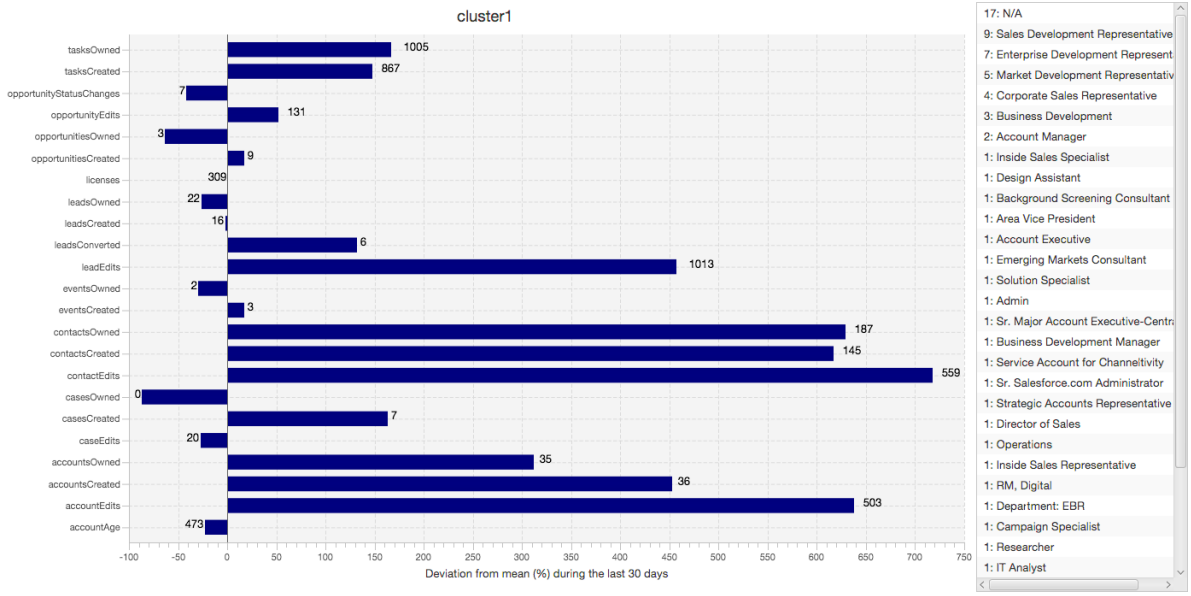


Figure 4.6: The High Frequency Editor. Appendix A shows the image in full size.

Table 4.3: Table presenting the parameters and accuracy from the C4.5 classifier.

Minimum Number Of Objects	Accuracy
3	89.28%
5	90.02%
7	90.57%
9	90.20%
11	88.91%
12	88.91%

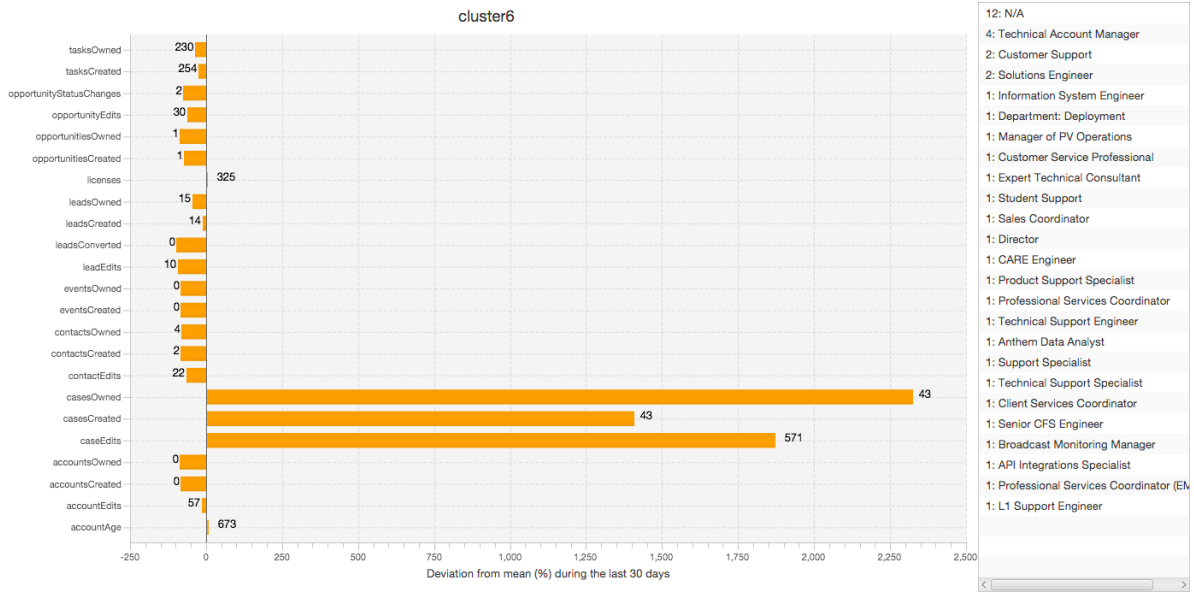


Figure 4.7: The Case Worker. Appendix A shows the image in full size.

Chapter 5
Evaluation

This chapter describes how the results were evaluated.

5.1 Evaluation of Profiles

We presented the segments in a booklet with all the found segments visualized and interpreted. The visualization was similar to the findings in Chap. 4.

The different customer segments were evaluated by the Brisk staff. This ensured both their interpretability, correctness and usefulness. The objectives to be met were the ones defined in Sec. 3.3.1.

We created a survey in order to evaluate the different profiles interpretability, correctness and usefulness. We asked three questions:

- How easy do you think it is to interpret the profiles?
- How correct do you think the profiles are?
- How useful do you think the profiles are?

We asked all 9 of the Brisk staff, and we received 6 answers. The questions were labeled on a numeric scale from very negative to very positive. The answers were summarized and divided by the scale. The result from the evaluation of the profiles are shown in Table 5.1.

The result from the evaluation implies that the Brisk are overall satisfied with the evaluation. The staff thought the profiles were 73% useful and interpretable, and 80% correct.

5.2 Evaluation of Classifier

We discussed how the derived classifier could be evaluated. For instance, this could be achieved by having a person classify instances into one of the derived categories. However,

Table 5.1: Table presenting the results from evaluating the profiles

	Interpretability	Correctness	Usefulness
Average percentage:	73%	80%	73%

we did not perform such an evaluation. Since part of understanding and choosing relevant clusters involves deriving a decision tree classifier, such a manual classification could be performed using said decision tree. This would mean that the manual classification would achieve the same accuracy as the derived tree. This is discussed further in Chap. 6.

Chapter 6
Discussion

In this chapter a thorough discussion and interpretation of the results from the evaluation is presented.

6.1 Data Mining Process

In this section we discuss the decisions we made in the data mining process, from the selection of CRISP-DM combined with BSM, to the creation of models.

6.1.1 CRISP-DM and BSM

The main goal of this thesis was to provide Brisk with an improved understanding of their customers. In our approach, this was translated to business goals. These were then translated into data mining goals. The procedure of making these translations, made our abstract goal more concrete. This helped us to focus on what was really important in the project. It also helped all parties in agreeing on what was expected of the thesis and what was outside the scope of it. Formulating such goals is supported in both CRISP-DM and BSM. Due to the advantages mentioned above, we definitely appreciated this approach.

Our hybrid model proved very successful. It used the general good practices of CRISP-DM and added specifics about behavioral segmentation from BSM. We would recommend any data mining project to use CRISP-DM and any behavioral segmentation data mining project to use BSM. A project consisting of various data mining challenges, like ours, could benefit from using both.

The designated time of one semester is not a long time in the context of data mining projects. We put a lot of effort into avoiding situations where approaching deadlines would compromise the quality of our work. We did not want changes that emerged to ruin our time plan. Neither did we want a time plan based on initial assumptions to cause problems as changes emerged. An iterative and agile work flow proved to be a winning strategy in

working on our thesis. In particular, an early adoption of the CRISP-DM methodology was essential. It is the result of a great deal of experience, and it removes many obstacles. Especially for teams with relatively little experience like ours. CRISP-DM helped us construct and execute according to an efficient strategy and it yielded excellent results.

6.1.2 Data selection

It was a great challenge to determine what data were relevant. We quickly realized that we wanted to collect behavioral data. The objects that a user *owns* may not be considered a behavioral aspect; but it is important to know that in Salesforce a user can have object ownership for various reasons. A user can get assigned object ownership or take it himself. If a user *creates* an object, that user also becomes the owner of the object. It can therefore be argued whether or not the *ownership* of objects should be included in the selected data. We decided that the *ownership* is an interesting feature which, combined with other features, would enrich the data set. The other features selected are more intuitively connected to behavior and do not require further discussion.

There were potentially good features that we could not use. Some features, e.g. information about users' login habits would perhaps have been beneficial for our behavioral segmentation, but could not be used. This was because not enough users were allowed to fetch it, potentially due to permission restrictions. These features were fetched initially, but were then omitted, resulting in fewer queries being performed.

6.1.3 Data preparation

The data preparation part was the most time consuming and demanding part of this thesis. It required knowledge of the business domain in order to perform the preparation correctly. It would, for instance, not be possible to distinguish likely values from unlikely ones without domain knowledge.

When we performed the data cleaning process there was a lot of features that were removed due to numerous missing values. Unfortunately this included all of our created ratios. This was probably because we considered all ratios where the denominator was zero, to be missing values. With a large number of zero values, this led to a lot of missing values for the ratios.

We initially tried a different approach where we replaced the zero denominators with a small value when calculating ratios. We did this in order to get a values for these ratios. Also, where neither one of the numerator or denominator was missing, it seemed intuitive that the ratio should not be missing either. This approach did, however, not yield any good results. When we selected a very small value, the values of the zero value denominator ratios became very differentiated from the values with actual non-zero denominators. Higher values on denominators made the ratios indistinguishable from ratios with a non-zero denominator.

We used median value imputation for the missing values. This was supported in the literature we studied and justified for distributions that are not normally distributed. There are, however, other strategies for this besides mean and median imputation. Another alternative is to create machine learning models to *predict* the missing values. We did, however,

not choose this approach, since it was considered to be too time consuming. We argued that time was better spent on improving the models in other aspects.

PCA was used in order to facilitate the data reduction process. This led to a small loss of information. In the final model, a total of 95% of the variance of the features was preserved. We consider this minor loss of information to be negligible.

6.1.4 Creation of Models

We created all our models in WEKA. We believe that there are more time-efficient libraries for creating machine learning models. WEKA states that its implementation of DBSCAN algorithm should not be used for benchmark testing. In our setting, however, the time it took to create the models was not very important. Most models were created within half a second. The visual interface and ease of algorithm testing in WEKA, made it an excellent tool to work with.

The selection of clustering algorithm was far from trivial. Customer segments are unknown so there is no right answer to be used for validation. Many factors are considered when deciding the best segmentation and hence clustering. The number of clusters and instances in each cluster are two of these. For the algorithms that do not cluster instances considered to be outliers, it is also interesting to see how many instances that were not clustered.

It was recommended in Tsiptsis and Chorianopoulos (2009) to use an algorithm that does not cluster instances that are outliers. These algorithms are based on a density-based notion. More data points increase the probability of more points being close to each other. Although this approach became more beneficial as the dataset grew, it was not used in the final clustering. We believe that our dataset was not big enough to use it.

Although the literature recommends using a density-based algorithm, the results when using DBSCAN were very poor. Either a very small percentage of all users were clustered, or almost all users were put in the same segment. Therefore, we selected k-means clustering as our clustering algorithm.

It would be possible to use another classification method, but we decided that the accuracy from the C4.5 was high enough. Also the classifier was selected on the basis of the ease to it export to executable Java-code.

6.2 Profiles

We spent several months at Brisk. A lot of this time was spent on learning about the domain of sales personnel, Brisk and Salesforce. This knowledge led to the selection of data to be collected. This, in turn, determined the segments found.

6.2.1 Presentation of Profiles

We created a booklet for presenting the segments we found to Brisk. The presentation could have been more complex. It could e.g. have contained more statistics to back up the claims made. It turned out that this was not what the staff of Brisk wanted. The presentation was iterated and subject to feedback from the Brisk staff. The most important

conclusion from the feedback was that the staff desired an interpretation of the data, rather than to be presented with the data itself. We spent many hours looking at the data from different perspectives, created a large number of models, and went back and forth between data and interpretation. Thus we consider ourselves to have a lot of insights to add to the solution. Hence Brisk's desire of interpreted data matched ours. However, even if that had not been the case, we would have met Brisk in whatever desires they might have had expressed. Their satisfaction was part of both our business and data mining goals.

When analyzing the segments, we visualized each segment's deviation from the average of all instances. As a result, segments that are very extreme will affect the visualization. This can have a negative effect on the evaluation of profiles. Naturally, there are other ways of visualizing the differences between segments; one way would be to use the median value instead of mean. We tried this, but concluded that the results were less good. A median comparison means that we just compare the value of different users, specific to a segment, with the median value of a user of the entire data set. Even though a mean value comparison is not optimal, we conclude it is a good enough way to visualize the differences of segments. It is not an easy task to display data for numerous segments, in even more dimensions.

6.2.2 Evaluation of Profiles

We evaluated the satisfaction of the Brisk staff by sending out a survey which evaluated their subjective thoughts about the profiles.

The interpretability of the derived segments was very important in this thesis. Hence, it was positive to get the high score of 73%.

Although interesting, the perceived correctness of the clusters was not crucial. Some of the Brisk staff are not customer-facing, and may have no intimate knowledge of the sales personnel who are their customers. Even the Brisk employees who actually do meet customers, can not possibly have detailed insights into the specific dataset examined. The high score of 80% was very positive, nevertheless.

The most important of our evaluation metrics, was that of usefulness. It resonates perfectly with the goal of the entire thesis. A segmentation that is not perceived as useful, would be considered a failure. Hence, the score of 73% was comforting. We believe that the score would not have been as high, if we had not derived our segment presentation in the iterative fashion that we did. The feedback in every loop was essential.

It is difficult to analyze something as subjective as sales processes. We could have used other methods for evaluating the segments found. E.g. we could have used a set of users very familiar to the Brisk staff, to see in which segment they would end up. This could then have been compared to the knowledge of the staff. This would, however, be a violation of the users' privacy since they would no longer be anonymous. Also, this would require a lot of effort from the Brisk staff. It would also be a relatively narrow evaluation of only a subset of the data set.

We could also have performed a more extensive evaluation of the classifier. This can be performed by sampling e.g. 100 new users from Brisk. An employee from Brisk could manually classify these instances by reviewing the profiles booklet. These instances would afterwards be automatically classified. The results could be compared and the classifier accuracy could be evaluated. This approach allows each user to stay anonymous. However,

this is a very time consuming task for the Brisk employee who would perform the manual classification. Also, we argue that since a decision tree is included in the profile booklet, it is fairly easy to classify an instance with a similar accuracy as the tree. Therefore this approach is redundant.

6.2.3 Analysis of Profiles

Below follows an analysis of each segment derived. In this analysis there are titles that non-available. The titles that are N/A have not been commented in this analysis.

The Passive, Experienced Worker

The Passive, Experienced Worker was by far the largest cluster. It hence seems likely that they would affect the total cluster average greatly. It is therefore very surprising that the average of the users in this cluster deviates from the average of all clusters.

We were surprised that this segment was so large and that the users were so passive. This corresponds, however, to our initial analysis of the data. It suggests that practically all features are centered around low values, having data that are skewed. It is hence not surprising that a large number of users end up in a segment where users are passive.

We assumed that users that had used Salesforce for a long time would be some kind of specialists and work more frequently with some objects. It is therefore surprising that the users in this segment are all very passive and that the account age is higher than average.

The Average Generalist

The Average Generalist segment has characteristics that are supported by the perception of the Brisk staff. When in contact with sales personnel from all over the world, their impression is that sales processes are often unstructured. This might lead to staff working with various parts of the sales process.

The Brisk staff were expecting us to find a segment of sales personnel working a lot with leads. This was part of their feedback on initial segments presented to them. Such a segment was never found. It turns out that staff working with leads belong to *The Average Generalist* cluster.

We find both the average and general nature of this segment interesting. When analyzing the derived interpretation decision tree, this is the only segment that is represented by more than one leaf. Our interpretation of this is that this segment can not be described by only one characteristic.

The Event Worker

The Event Worker segment is very clear. The users in this segment work a lot more than average with events and are fairly average in the other aspects. An event object is associated with a meeting or calendar event. It is therefore similar to tasks and differs from the other objects investigated. The usage of events says a lot about the actual behavior of a sales person.

The fact that these users use events implies that they know how to use Salesforce and that they have a structured sales process. Events can be connected to other objects, and it would therefore be interesting to further investigate which type of object, if any, these users connect their events to. This was something we attempted, but it failed due to the high amount of missing or zero values.

The Active Opportunity Worker

This is a cluster of specialists. During our numerous efforts in creating relevant models, a segment of staff working with opportunities was often present. This implies that there is a large portion of sales personnel that have a niched responsibility. The segment is extreme with its high frequency of work with opportunity objects.

The fact that the sales staff of this segment owns a large portion of accounts and contacts supports our theory of what a sales person working with opportunities does.

The High Frequency Editor

This segment was extreme; users in this segment edit objects with a very high frequency compared to other users. The results might imply that the users in this segment use software as an aid to their usage of Salesforce. This is highly likely since there are many types of software present that can be used in combination with Salesforce.

The Case Worker

Our results imply that there is a very low number of sales personnel who work with cases. The ones who actually do, are easily considered outliers since they are so rare. This means that they are easily removed by mistake during a data reduction phase, if not performed carefully.

The titles of the staff belonging to this segment suggest that the cluster is relevant. Many titles include "technical", "customer" and "support". This indicates that they work with service or support.

We are very pleased with having found a segment of Case Workers. However, we do not encourage too much effort being put into satisfying the needs of sales staff working with cases, since they are so rare.

6.3 Future research

As the importance of CRM grows, we would recommend future research to investigate different applications of performing segmentation with data from CRM systems - both from the sales person's and the customer's point of view. Such analyses could aim to improve current usage of the systems. This was, however, not our goal for this thesis.

One possible application could be to segment sales personnel according to behavior and try to create flow graphs of how different users navigate through the system. This can be used to improve the usability of the system. The systems user profile could possibly be identified. The behavior of a user with a certain profile could provide objective data for usability testing.

Another valuable approach could be to investigate how customers move through the system. By investigating how successful deals ended up in a successful state, conclusions could perhaps be drawn as to how to increase the number of profitable deals. Also, we recommend even further investigation in behavioral segmentation on Salesforce data. Although we are pleased with our own work, we are sure that further research of the same kind would prove fruitful.

We also recommend using additional platforms for collecting data on both customers and sales personnel. If there had been time, we would have used platforms for e.g. email correspondence. We believe that a sales person's email patterns, such as email frequency, might be a strong indicator on that person's professional behavior.

Chapter 7
Conclusions

The project was finished on time and the scope and approach were changed during the project. An agile workflow was therefore crucial. Using the Cross-industry standard process for data mining (CRISP-DM) combined with Behavioral segmentation methodology (BSM) proved to be an excellent approach in deducing and understanding the customer segments.

We extracted behavior data from the Salesforce platform. Gaining business understanding of the sales domain was a very important step in our process. It was necessary in order to make informed decisions of e.g. what data to use and how to process those data. In order to draw conclusions, the knowledge of the domain was also essential. The entire process was therefore guided by both business and data mining goals. We formulated these in collaboration with the Brisk CTO.

The preprocessing of data proved crucial in order to develop relevant clusters in the unsupervised learning stage. The data contained outliers and unrealistic data. These had to be removed in order to create features that could be comparable. Also, replacing missing values using imputation required that we analyzed the distribution of the features. This led us to use median imputation, since some data were skewed and far from normally distributed.

We used k-means clustering to segment our data and found six different segments that were verified to be relevant by the Brisk staff. They could be interpreted using visualization techniques combined with a decision tree. The result implies that most users are inactive in their usage of Brisk and Salesforce, and they are most likely to have a poor sales process. We created a tree classifier by using the C4.5 algorithm, to be implemented into the production code of Brisk.

In the future we would recommend further research into how to deal with data where many values are zero. Also it would be interesting to see more research in how to use behavioral data from CRM systems with data from other platforms. But the field of application is large and the possibilities for future research are numerous.

In conclusion we are satisfied with the result of our work, and believe that even though

the time was limited we achieved good results. Although dealing with very difficult data, our goals were fulfilled.

Acronyms

BSM Behavioral segmentation methodology.

CRISP-DM Cross-industry standard process for data mining.

CRM Customer Relationship Management.

IQR Interquartile range.

MAR Missing at random.

PCA Principal component analysis.

WEKA Waikato environment for knowledge analysis.

Glossary

***k*-means clustering** A simple algorithm for clustering n instances into k clusters.

account An individual account, which is an organization involved with a Salesforce user's business (such as customers, competitors, and partners).

Brisk The name of both a sales staff aiding software and the company that is developing it.

case A customer issue or problem..

contact An individual associated with an account.

DBSCAN A widely used clustering algorithm.

event An event in the calendar.

Google Chrome A web browser developed by Google.

lead A prospect or potential opportunity.

opportunity A sale or a pending deal.

Salesforce A platform for Customer Relationship Management (CRM).

task An activity or to-do item to perform or that has been performed.

Bibliography

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*.

Curram, S. P. and Mingers, J. (1994). Neural networks, decision tree induction and discriminant analysis: An empirical comparison. *The Journal of the Operational Research Society*, 45(4):pp. 440–450.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651 – 666. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR)19th International Conference in Pattern Recognition (ICPR).

Kim, S.-Y., Jung, T.-S., Suh, E.-H., and Hwang, H.-S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert Systems with Applications*, 31(1):101 – 107.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J., editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Maimon, O. and Rokach, L. (2008). *Data mining with decision trees: theory and applications*. USA: World Scientific Publishing.

Ngai, E., Xiu, L., and Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2, Part 2):2592 – 2602.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Shaw, R. (1991). *Computer Aided Marketing & Selling*. Butterworth Heinemann.

Torgo, L. (2011). *Data Mining with R*. Chapman & Hall/CRC.

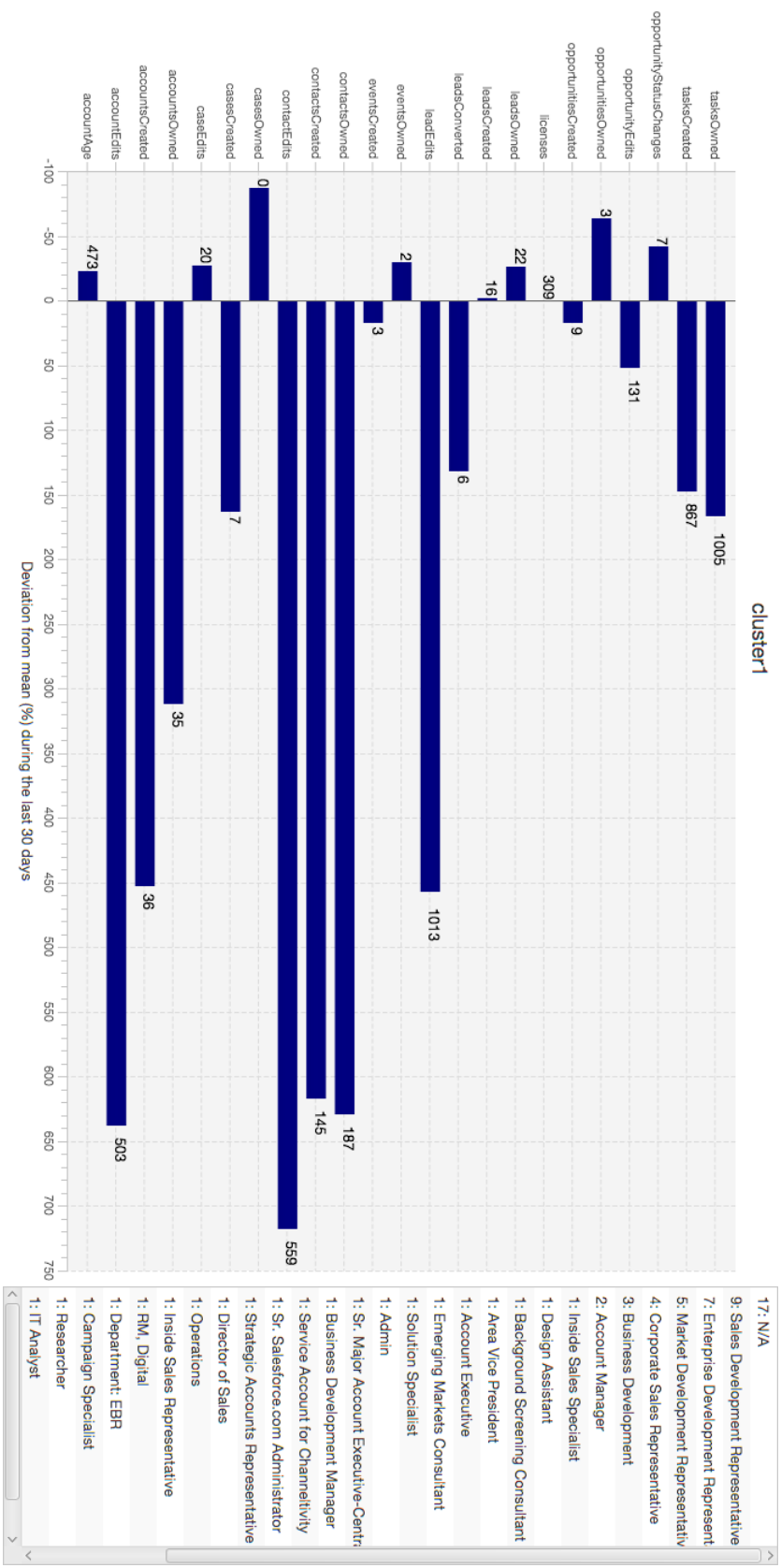
Tsiptsis, K. and Chorianopoulos, A. (2009). *Data Mining Techniques in CRM: Inside Customer Segmentation*. John Wiley and Sons.

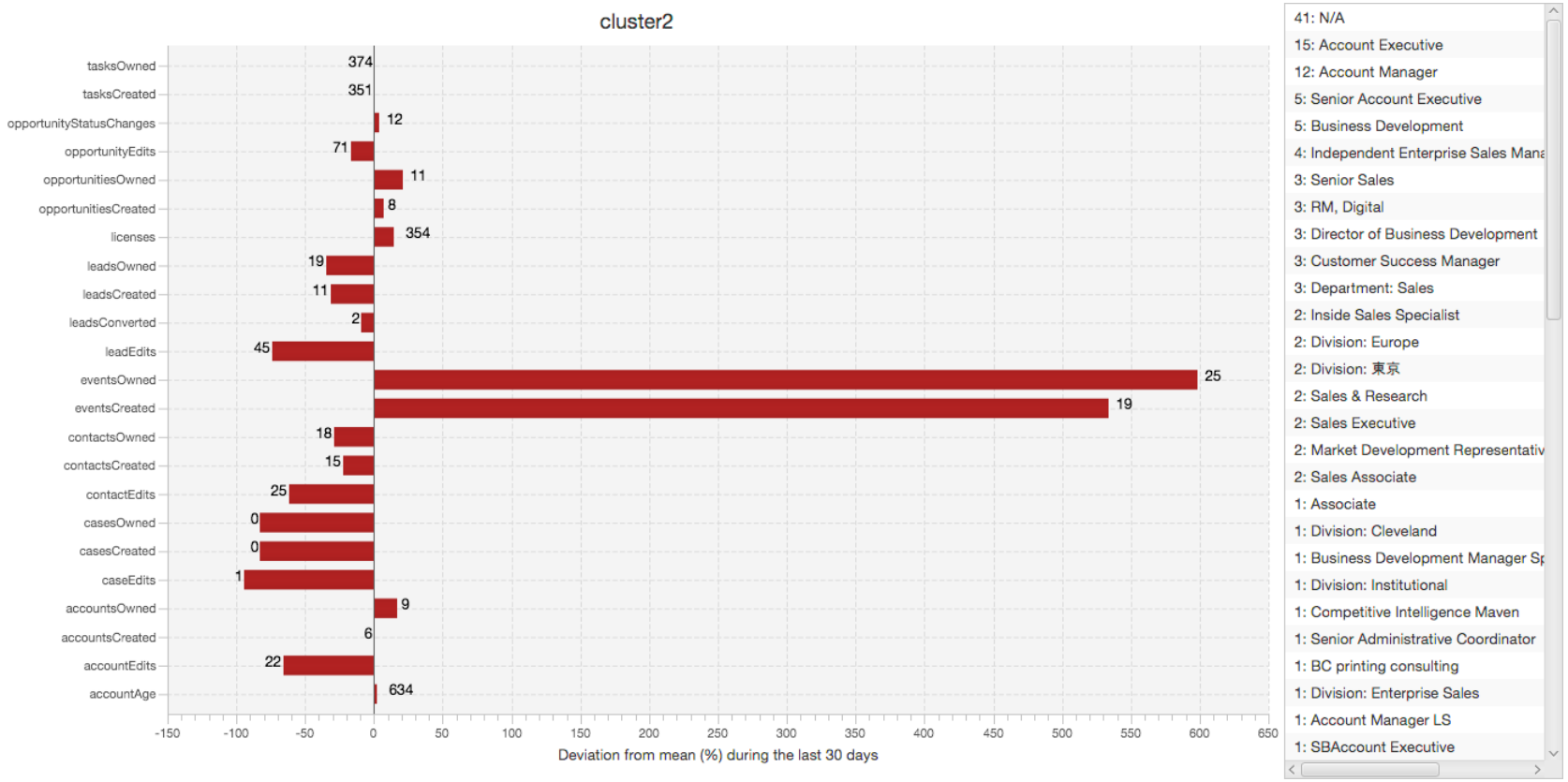
Wikimedia-Commons (2011a). Dbscan density data.

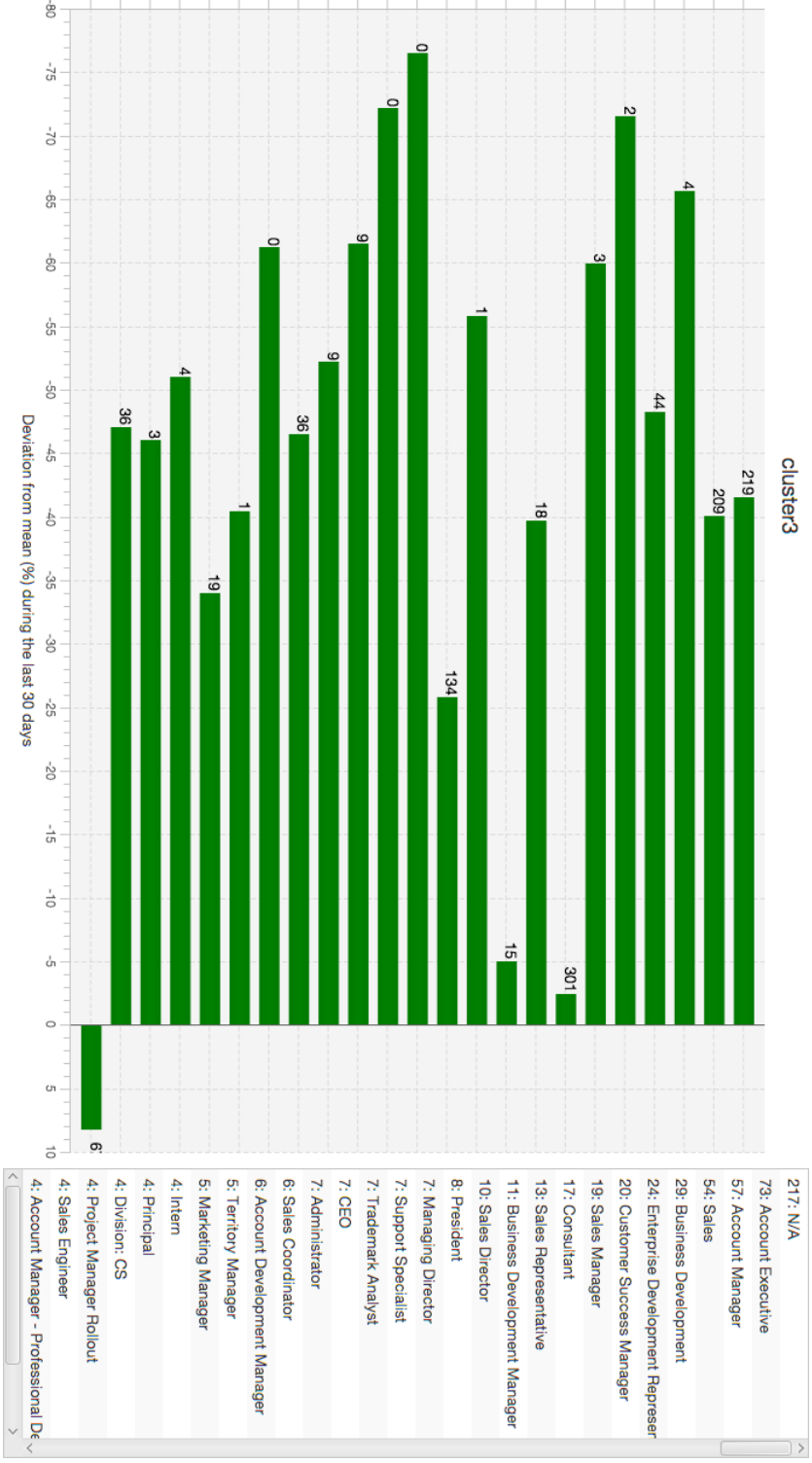
Wikimedia-Commons (2011b). Kmeans gaussian data.

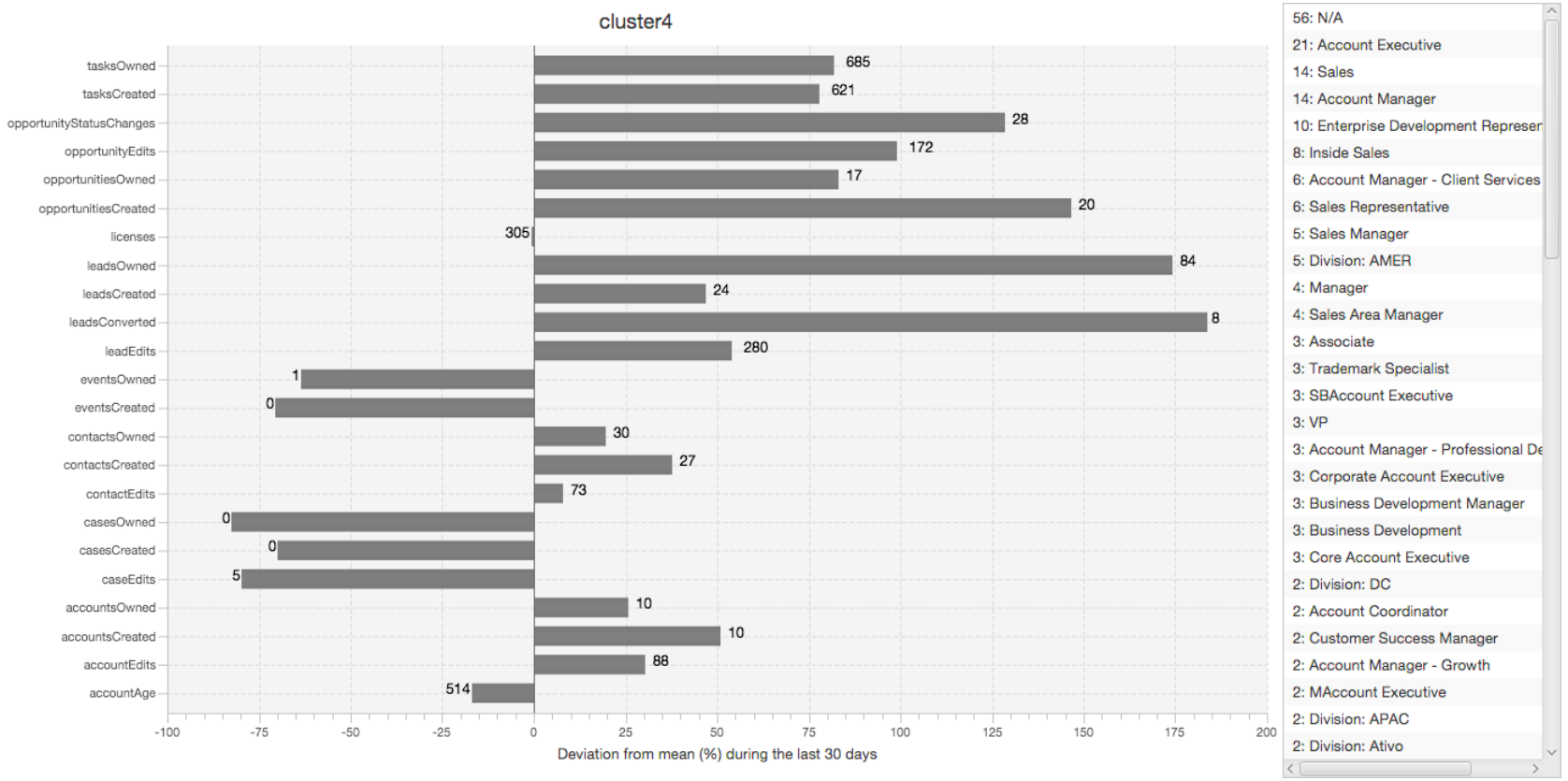
Wikimedia-Commons (2012). Crisp-dm process diagram.

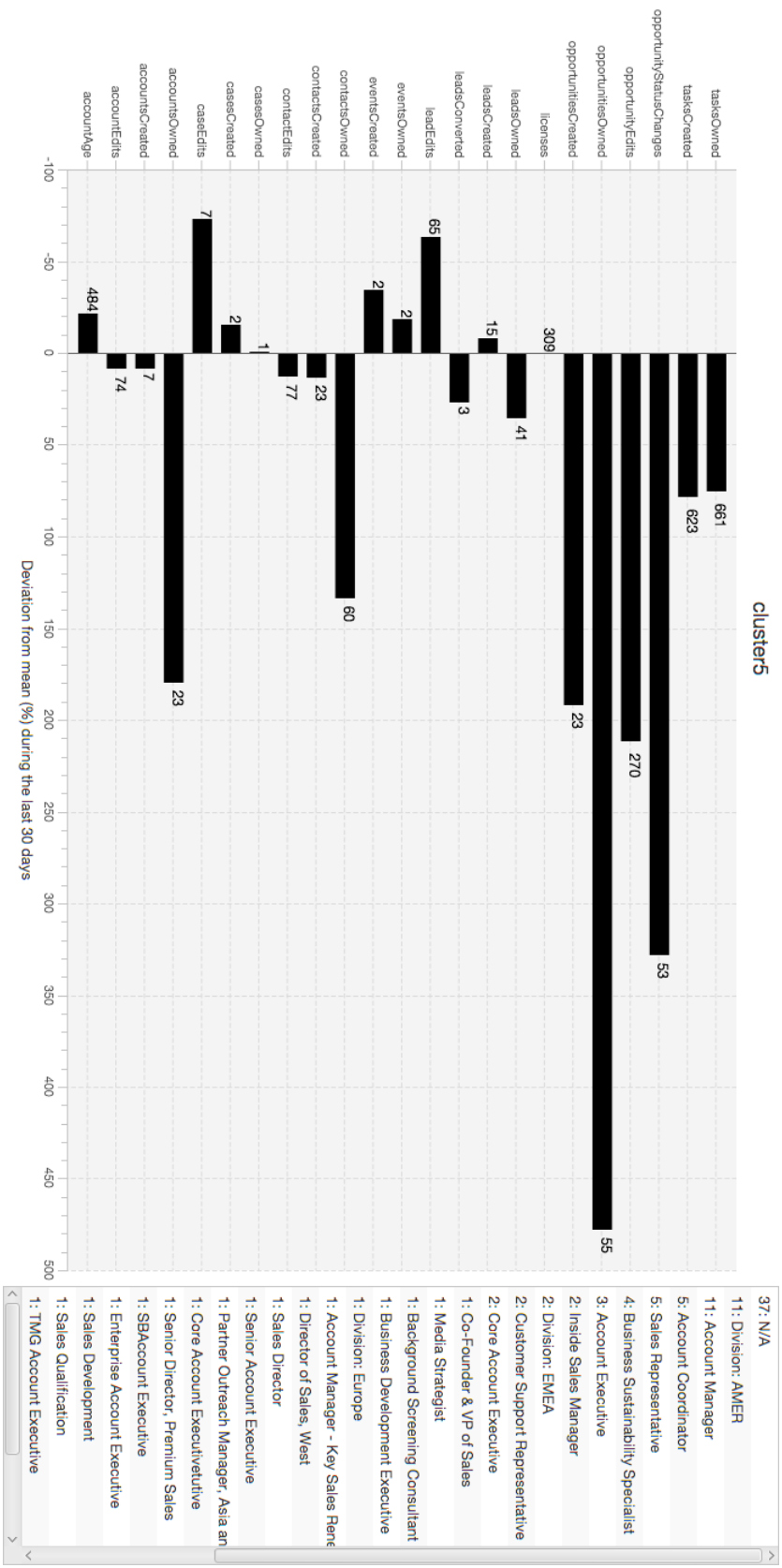
Appendix A

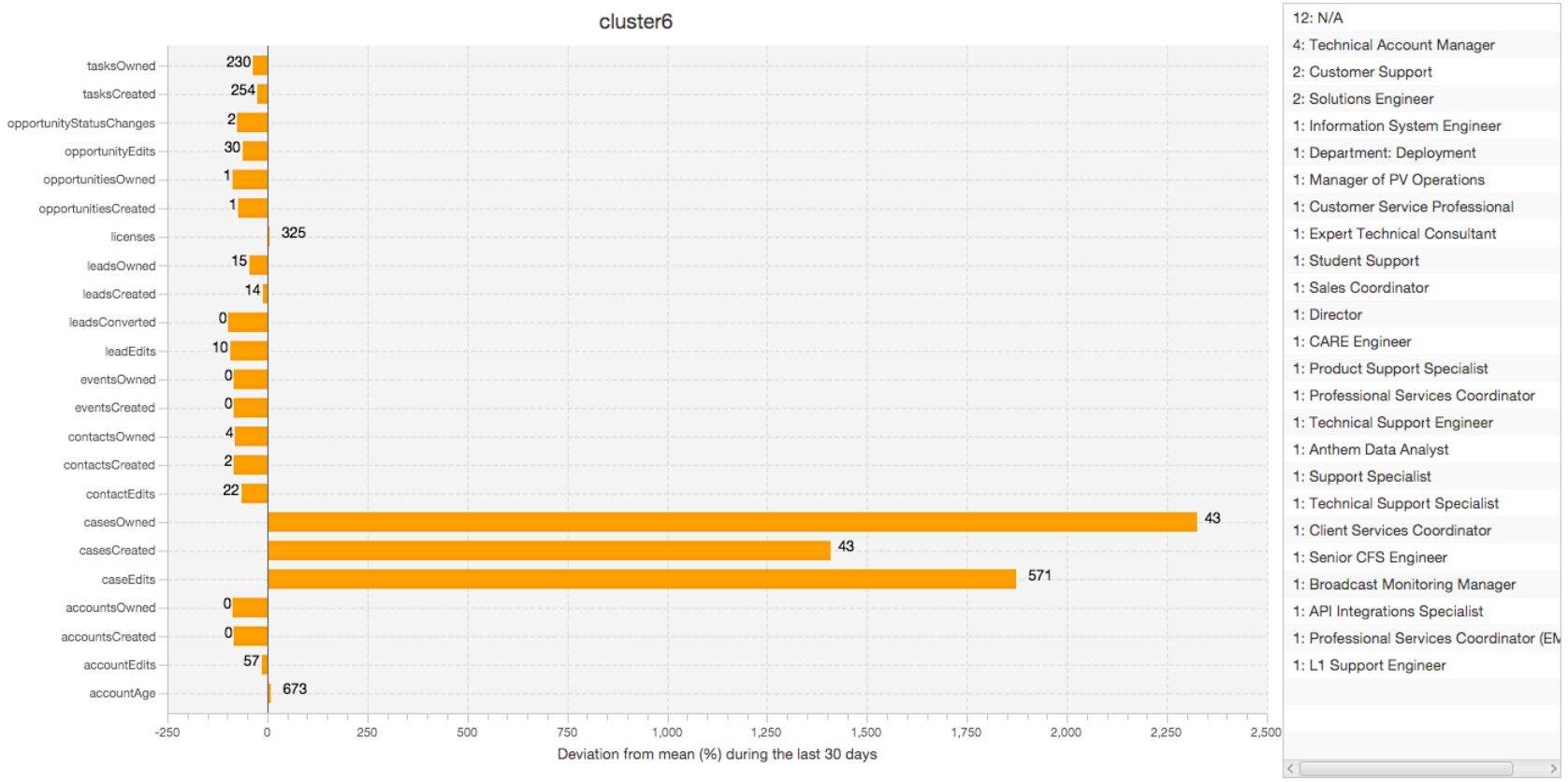

cluster1
tasksOwned
tasksCreated
opportunityStatusChanges
opportunityEdits
opportunitiesOwned
opportunitiesCreated
licenses
leadsOwned
leadsCreated
leadsConverted
leadEdits
eventsOwned
eventsCreated
contactsOwned
contactsCreated
contactEdits
casesOwned
casesCreated
caseEdits
accountsOwned
accountsCreated
accountEdits
accountAge
1005
867
131
7
3
9
309
22
16
6
1013
2
3
187
145
559
0
7
20
35
36
503
473
-100
-50
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
Deviation from mean (%) during the last 30 days
17: N/A
9: Sales Development Representative
7: Enterprise Development Represent
5: Market Development Representativ
4: Corporate Sales Representative
3: Business Development
2: Account Manager
1: Inside Sales Specialist
1: Design Assistant
1: Background Screening Consultant
1: Area Vice President
1: Account Executive
1: Emerging Markets Consultant
1: Solution Specialist
1: Admin
1: Sr. Major Account Executive-Centra
1: Business Development Manager
1: Service Account for Channeltivity
1: Sr. Salesforce.com Administrator
1: Strategic Accounts Representative
1: Director of Sales
1: Operations
1: Inside Sales Representative
1: RM, Digital
1: Depatment: EBR
1: Campaign Specialist
1: Researcher
1: IT Analyst


cluster2
tasksOwned
tasksCreated
opportunityStatusChanges
opportunityEdits
opportunitiesOwned
opportunitiesCreated
licenses
leadsOwned
leadsCreated
leadsConverted
leadEdits
eventsOwned
eventsCreated
contactsOwned
contactsCreated
contactEdits
casesOwned
casesCreated
caseEdits
accountsOwned
accountsCreated
accountEdits
accountAge
374
351
12
71
11
8
354
19
11
2
45
25
19
18
15
25
0
0
1
9
6
22
634
-150
-100
-50
0
50
100
150
200
250
300
350
400
450
500
550
600
650
Deviation from mean (%) during the last 30 days
41: N/A
15: Account Executive
12: Account Manager
5: Senior Account Executive
5: Business Development
4: Independent Enterprise Sales Mana
3: Senior Sales
3: RM, Digital
3: Director of Business Development
3: Customer Success Manager
3: Department: Sales
2: Inside Sales Specialist
2: Division: Europe
2: Division: 東京
2: Sales & Research
2: Sales Executive
2: Market Development Representativ
2: Sales Associate
1: Associate
1: Division: Cleveland
1: Business Development Manager Sp
1: Division: Institutional
1: Competitive Intelligence Maven
1: Senior Administrative Coordinator
1: BC printing consulting
1: Division: Enterprise Sales
1: Account Manager LS
1: SBAccount Executive


cluster3
tasksOwned
tasksCreated
opportunityStatusChanges
opportunityEdits
opportunitiesOwned
opportunitiesCreated
licenses
leadsOwned
leadsCreated
leadsConverted
leadEdits
eventsOwned
eventsCreated
contactsOwned
contactsCreated
contactEdits
casesOwned
casesCreated
caseEdits
accountsOwned
accountsCreated
accountEdits
accountAge
219
209
4
44
2
3
18
15
301
1
134
0
0
9
9
36
0
1
19
4
3
36
6
-80
-75
-70
-65
-60
-55
-50
-45
-40
-35
-30
-25
-20
-15
-10
-5
0
5
10
Deviation from mean (%) during the last 30 days
217: N/A
79: Account Executive
57: Account Manager
54: Sales
29: Business Development
24: Enterprise Development Represer
20: Customer Success Manager
19: Sales Manager
17: Consultant
13: Sales Representative
11: Business Development Manager
10: Sales Director
8: President
7: Managing Director
7: Support Specialist
7: Trademark Analyst
7: CEO
7: Administrator
6: Sales Coordinator
6: Account Development Manager
5: Territory Manager
5: Marketing Manager
4: Intern
4: Principal
4: Division: CS
4: Project Manager Rollout
4: Sales Engineer
4: Account Manager - Professional De


cluster4
tasksOwned
tasksCreated
opportunityStatusChanges
opportunityEdits
opportunitiesOwned
opportunitiesCreated
licenses
leadsOwned
leadsCreated
leadsConverted
leadEdits
eventsOwned
eventsCreated
contactsOwned
contactsCreated
contactEdits
casesOwned
casesCreated
caseEdits
accountsOwned
accountsCreated
accountEdits
accountAge
685
621
28
172
17
20
305
84
24
8
280
1
0
30
27
73
0
0
5
10
10
88
514
-100
-75
-50
-25
0
25
50
75
100
125
150
175
200
Deviation from mean (%) during the last 30 days
56: N/A
21: Account Executive
14: Sales
14: Account Manager
10: Enterprise Development Represer
8: Inside Sales
6: Account Manager - Client Services
6: Sales Representative
5: Sales Manager
5: Division: AMER
4: Manager
4: Sales Area Manager
3: Associate
3: Trademark Specialist
3: SBAccount Executive
3: VP
3: Account Manager - Professional De
3: Corporate Account Executive
3: Business Development Manager
3: Business Development
3: Core Account Executive
2: Division: DC
2: Account Coordinator
2: Customer Success Manager
2: Account Manager - Growth
2: MAccount Executive
2: Division: APAC
2: Division: Ativo


cluster5
tasksOwned
tasksCreated
opportunityStatusChanges
opportunityEdits
opportunitiesOwned
opportunitiesCreated
licenses
leadsOwned
leadsCreated
leadsConverted
leadEdits
eventsOwned
eventsCreated
contactsOwned
contactsCreated
contactEdits
casesOwned
casesCreated
caseEdits
accountsOwned
accountsCreated
accountEdits
accountAge
661
623
53
270
55
23
309
41
15
3
65
2
2
60
23
77
1
2
7
23
7
74
484
-100
-50
0
50
100
150
200
250
300
350
400
450
500
Deviation from mean (%) during the last 30 days
37: N/A
11: Division: AMER
11: Account Manager
5: Account Coordinator
5: Sales Representative
4: Business Sustainability Specialist
3: Account Executive
2: Inside Sales Manager
2: Division: EMEA
2: Customer Support Representative
2: Core Account Executive
1: Co-Founder & VP of Sales
1: Media Strategist
1: Background Screening Consultant
1: Business Development Executive
1: Division: Europe
1: Account Manager - Key Sales Rene
1: Director of Sales, West
1: Sales Director
1: Senior Account Executive
1: Partner Outreach Manager, Asia an
1: Core Account Executivetutive
1: Senior Director, Premium Sales
1: SBAccount Executive
1: Enterprise Account Executive
1: Sales Development
1: Sales Qualification
1: TMG Account Executive


cluster6
tasksOwned
tasksCreated
opportunityStatusChanges
opportunityEdits
opportunitiesOwned
opportunitiesCreated
licenses
leadsOwned
leadsCreated
leadsConverted
leadEdits
eventsOwned
eventsCreated
contactsOwned
contactsCreated
contactEdits
casesOwned
casesCreated
caseEdits
accountsOwned
accountsCreated
accountEdits
accountAge
230
254
2
30
1
1
325
15
14
0
10
0
0
4
2
22
43
43
571
0
0
57
673
-250
0
250
500
750
1,000
1,250
1,500
1,750
2,000
2,250
2,500
Deviation from mean (%) during the last 30 days
12: N/A
4: Technical Account Manager
2: Customer Support
2: Solutions Engineer
1: Information System Engineer
1: Department: Deployment
1: Manager of PV Operations
1: Customer Service Professional
1: Expert Technical Consultant
1: Student Support
1: Sales Coordinator
1: Director
1: CARE Engineer
1: Product Support Specialist
1: Professional Services Coordinator
1: Technical Support Engineer
1: Anthem Data Analyst
1: Support Specialist
1: Technical Support Specialist
1: Client Services Coordinator
1: Senior CFS Engineer
1: Broadcast Monitoring Manager
1: API Integrations Specialist
1: Professional Services Coordinator (EM
1: L1 Support Engineer

INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTATIONSDAG 2016-01-16

EXAMENSARBETE Discovering and Inducing Rules to Categorize Sales Personnel
STUDENT Lisa Stenström, Olof Wahlgren
HANDLEDARE Pierre Nugues (LTH)
EXAMINATOR Jacek Malec

Hur jobbar säljare?

POPULÄRVETENSKAPLIG SAMMANFATTNING **Lisa Stenström, Olof Wahlgren**

Inom sälj antas säljare uppvisa olika beteenden beroende på vilken roll de har. Det visar sig dock att de flesta arbetar på ett mer enhetligt sätt än vad man kan tro.

Vi har analyserat över 3000 säljare utifrån hur de beter sig på Salesforce, en av världens största plattformar för Customer Relationship Management (CRM). Resultaten var överraskande. Så många som 60% av säljarna använde knappt Salesforce överhuvudtaget. 16% använde flera olika delar av systemet och verkar ha en ostrukturerad säljprocess. Vi fann ett fåtal grupper som innehöll säljare med avgränsad spetskompetens. Tillsammans utgjorde dessa specialistgrupper endast 24% av de undersökta säljarna. Det visade sig också att säljarnas titlar sällan avslöjade vad de faktiskt arbetade med.

Examensarbetet skrevs tillsammans med företaget Brisk. De utvecklar programvara, byggd på Salesforce, för att underlätta och effektivisera arbetsflödet för säljare i hela världen.

Det finns en mängd olika roller inom försäljning. Det slängs med olika högtravande titlar hejvilt, den ena mer imponerande än den andra. Vi ville ta reda på hur olika säljare faktiskt arbetar.

Våra efterforskningar tyder på att det finns ytterst få studier kring hur man använder beteendet på en CRM-plattform för att gruppera säljare. Därför kan vårt arbete visa på bra metoder, algoritmer, och visualiseringtekniker, men även andra viktiga tips och tricks för att gruppera säljare utifrån deras beteende i Salesforce. Framförallt kommer kunskapen om säljares beteende att kunna användas i Brisks verksamhet. En större kunskap om säljarna kommer att resultera i en bättre anpassning av företagets tjänster.

Arbetet genomfördes genom att omsorgsfullt analysera vilken data som var relevant för säljarnas beteende. Därefter användes denna data för att skilja de olika säljtyperna åt. Målet var att hitta grupper där de säljare som ingår i samma grupp påminner om varandra, och skiljer sig från säljare i andra grupper. För att ge ett exempel: om en gruppering av ett fotbollslag hade utförts, hade anfallarnas och försvararnas beteenden förmodligen sett olika ut. Anfallarna hade förmodligen haft betydligt fler skott på mål än försvararna, som kanske hade lyckats med fler brytningar. En gruppering av dem, baserat på dessa beteenden, hade förhoppningsvis skiljt de två typerna åt.

I arbetet undersökte vi olika parametrar som karaktäriserar beteende för säljpersonal i Salesforce. Dessa var inte alltid triviala utan fick skapas genom varsam analys av vad som karaktäriserar en säljare. Vi manipulerade datan och använde sedan avancerade algoritmer för att dela upp de undersökta säljarna i relevanta grupper. För att säkerhetsställa att den slutgiltiga grupperingen var relevant, utvärderade vi våra resultat iterativt tillsammans med Brisk under hela arbetet. Vår slutgiltiga utvärdering visar att företaget anser att den funna grupperingen kommer att vara väldigt användbar för framtida produktutveckling.