DEGREE PROJECT IN TECHNOLOGY,
FIRST CYCLE, 15 CREDITS
STOCKHOLM, SWEDEN 2017

Identifying New Customers Using Machine Learning

A case study on B2B-sales in the Swedish IT-consulting sector

PATRIK NORLIN

VIKTOR PAULSRUD

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF COMPUTER SCIENCE AND COMMUNICATION

Abstract

In this thesis, we examine machine learning as a tool for predicting new customers in a B2B-sales context. Using only publicly available information, we try to solve the problem using two different approaches: 1) a naive clustering based classifier built on K-means and 2) PU-learning with a random forests-adapter. We test these models with different sets of features and evaluate them using statistical measures and a discussion of the business implications. Our main findings conclude that the PU-learning could produce results that are satisfactorily for the purpose of improving the sales process, with the best case of being 4.8 times better than a random baseline classifier. However, the clustering based classifier was not good enough, producing only marginally better results than a random classifier in its best case. We also find that using more variables improved the models, even in high-dimensional spaces with over 60 variables.

Acknowledgements

This thesis was written in collaboration with Exsitec AB. From Exsitec, we would therefore like to thank: Disa Törnvall, Anders Uddenberg and Rasmus Toivonen, for their guidance and support. We would also want to thank associate Prof. Johan Boye and data scientist Bure Noréus for their expertise in machine learning. Lastly, we thank our supervisors Joakim Gustafsson and Bo Karlsson.

Contents

1 Introduction

1.1 Background

The digital revolution and the increasing amount of data accumulated by organizations in the last decade has led to a great interest in the fields of artificial intelligence (AI) and machine learning (ML). The techniques of making computers learn by data has proven itself useful for many tasks by finding patterns and making predictions in the data. Researchers are constantly pushing the boundaries of the capabilities of AI and ML, and have made extraordinary progress in areas such as speech recognition, self-driving cars and language translation. Companies on the other hand are trying to incorporate the technology in their businesses to gain competitive advantage, or simply to not lag behind their peers (Ling, 2011). This has led to a scramble for talent, where machine learning and data science experts as well as the technology itself is in an ever increasing demand. In 2015, technology firms including Google, Amazon and Facebook spent a $8.5 billion on acquiring AI companies, an increase of 400% since 2010 (Economist, 2016)

Even outside of technology spheres, the usefulnesses of data is now widely spread. For many companies, it is mainly about making good use of the data that they collect to be able make better decisions in their daily operations (Jordan and Mitchell, 2015). This is often referred to as business intelligence, which builds on aggregating and visualizing data in a user-friendly way that can be easily be incorporated in the business. However, the benefits of machine learning are often overlooked since they require more technical expertise, and it is not always clear how it should be used. As a consequence, many interesting applications of machine learning and AI is left unexplored.

This thesis is written in such a context, as a case study on the Swedish IT-consulting firm Exsitec where we aim to apply machine learning techniques to find potential improvements in their current business. In particular, we look at the marketing process in business-to-business (B2B) sales to see how data-driven machine learning can be used to improve that process.

One of the key tasks in achieving successful organic growth for companies selling B2B is the task of identifying new customers. Compared to business-to-consumer (B2C) products, each sale generally costs more which puts pressure on a successful sales organization as well as giving good service to get a good customer retention rate (Kotler et al., 2006). An important part for B2B companies is therefore decid-

ing which customers to focus on, to avoid wasting resources on developing relations that ultimately will not create opportunities for sales. This is often carried out in a non-systematic way through human estimations in the sales force. It would thus be of interest to explore how this can be improved with a more systematic data driven technique, creating the opportunity to use machine learning as a tool in the sales process.

While there are many factors that affects a company's decision to become a customer, with many of them unobservable for outsiders, there might still be important indicators from publicly available data. The most obvious form of publicly available data is financial statements, which can be viewed as a general fingerprint of a company. This constitutes the foundation for the machine learning algorithms implemented in this paper, where we aim to be able to identify new customers for Swedish IT-consulting firm Exsitec through a classifier.

1.1.1 Exsitec

Exsitec AB is a Swedish IT-consulting firm that provides customized solutions in IT-systems for enterprises and organizations in the Nordic region. It was founded in 1999 in Linköping, Sweden by a professor from Linköping University, where it still has its head office. Today, Exsitec has revenues of SEK 126 million and operates in Stockholm, Linköping, Gävle, Göteborg and Skövde in addition to Linköping, with a total of 100 employees.

Exsitec's core business idea is to provide expertise and support in implementing and maintaining IT-systems that serves to help customers in digitizing their operations, and making better use of their data. They have three main focus areas: Enterprise Resource Planning (ERP) systems, Business Intelligence (BI) and Mobile solutions. Common problems they help simplify includes electronic invoice handling and inventory management (ERP-solutions), internal dissemination of information and budget forecasting (BI-solutions) and mobile applications for logistics workers (Mobile solutions). They do not develop the products themselves (with the exception for mobile solutions), but instead partners with software companies such as Visma and Qlik which allows them to focus on customer specific issues.

The successful growth of Exsitec's business the last few years is in many ways due to a high customer retention rate of 95.5%. Exsitec's sales force focuses on understanding their customers' problems well, allowing them to get good matches and maintaining long relations in their consulting services. However, there is po-

tentially room for improvement in their process of finding new customers. Today, most customers find their way to Exsitec through contacting the providers of a software product (e.g. Visma) which then redirects them to Exsitec for a customized solution, or simply by word-of-mouth (Törnvall, 2017-03-23). Exsitec have realized that there is an opportunity for them to grow sales faster if they could find potential customers that would be interested in buying their products. Therefore, this paper does not solely have an academic intention but also a commercial purpose.

1.2 Purpose

The purpose of this thesis is to investigate and analyze how machine learning can be used to identify potential customers in a business-to-business (B2B) sales context, using only publicly available data as a basis. There is a twofold purpose in this investigation, where the first one is computer scientific. As the boundaries of machine learning are not yet defined, this paper's subject adds to the scientific community by implementing machine learning models in a new real life setting. This leads to discoveries of new possibilities and limitations to the state of the art technology. The second purpose is to analyze the business implications of a machine learning model that predicts potential customers in business markets. In particular, it is of interest to see how such a model would affect the B2B-sales process, discussing improvements and limitation from an industrial marketing perspective.

1.3 Problem Definition

To fulfill the purpose, we intend to answer these questions in this paper:

- How well can machine learning models identify potential customers in a B2B-sales context?
- How would such a model affect the sales process, what are the possible improvements and limitations?

The first question will by evaluated by training machine learning models and evaluate these with quantifiable statistical measures. This will be extended with a qualitative analysis on the business implications of the models, based on theory and interviews.

1.4 Scope and Limitations

The thesis is conducted as a case study of the company Exsitec, which limits the industry and data of the research to that of Exsitec. The scope of the study includes identifying potential customers for Exsitec but excludes other companies, meaning

that the results may not be externally valid. The results will therefore not be general enough to answer the problem definition for every type of industry. However, since the operating strategy, business model and value offering is coherent with most IT-consulting firms, the results will probably be generalizable enough to answer to the problem definition for the Swedish IT-consulting industry as a whole.

The data is limited to the previous and current customers of Exsitec. One implication of this, is that if Exsitec has unexplored customer segments not present among the current customers, identifying such customers will not be covered by the models. Researching if so is the case or discussing the implications of this further is outside of the scope of this paper.

Another limitation regarding the data worth stressing the importance of, is its limitation to publicly and easily accessible company information. A major part of this will consist of financial statement data. Investigating other types of data sources is outside of the scope.

The machine learning aspect of this paper is limited to implementing two approaches. 1) a naive cluster-based classifier and 2) PU-learning with a random forests-adapter. Comparing and tweaking other methods is left out for future research.

1.5 Social and Ethical Aspects

One of the main social and ethical aspects of machine learning research is aspects considering data, which is especially important to reflect on when dealing with personal data. However, in this paper only publicly available company data is used, which does not jeopardize the integrity of individual people or company sensitive data. From a social perspective, one can argue that a better customer identification process is of social good, since it enables success for both the selling and the customer. The seller can focus its resources on providing value to more customers, and customers can access suitable help more easily.

2 Theoretical Framework

The theoretical framework is divided into two parts. The first part explains the theory behind industrial marketing and the B2B marketing and sales process. This aims to give a detailed view of current models for B2B-sales, and thereby acts as a foundation when analyzing in what parts of the process machine learning algorithms can be implemented. The second part explains the concepts and theories of machine learning as well as the implemented algorithms and performance measures in more detail.

2.1 Industrial Marketing

Industrial Marketing is the process of marketing and selling products or services from one business to another. The process of acquiring a new customer can generally be divided into two parts, a *marketing process* and a *sales process*. A main difference between those is that the goal of the marketing process is to generate qualified leads, while the goal of the sales process is to turn the qualified leads into paying customers. (Davies, 2010)

2.1.1 The Sales Funnel

The combined process of the marketing and sales process is often illustrated in a so called sales funnel, as in figure 1.

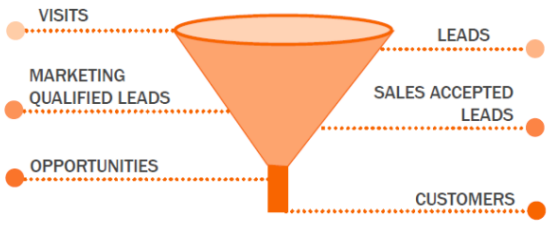


Figure 1: Sales Funnel (Ironpaper)

The sales funnel illustrates that in every step of the process, some potential customers are lost, which means that all prospects does not lead to a deal. The closer a potential customer is to becoming a customer, the more companies is lost until an opportunity finally turns into a customer. Consequently, the process should not be viewed as a mere act of convincing, but rather a process of identifying potential leads and determining their demands and needs.

According to the definition of Davies (2010), the sales process starts with a potential

customer identified. This implicates that the process of identifying new potential customers is within the marketing process.

2.1.2 Market Segmentation

According to Kotler et al. (2006), segmenting businesses builds on the same principles as marketing segmentation in consumer markets; because buyers have different needs, it is important for businesses to divide their markets into smaller segments so that the business can target itself uniquely to each segment. Common segmentation structures includes geographical location, industry, size and buying behavior. As there are many ways to segment a market, it is important for businesses to try different combinations of segmentation variables to find their best market structure in their case.

Market segmentation plays a key role in the marketing process, as shown in figure 2 below.

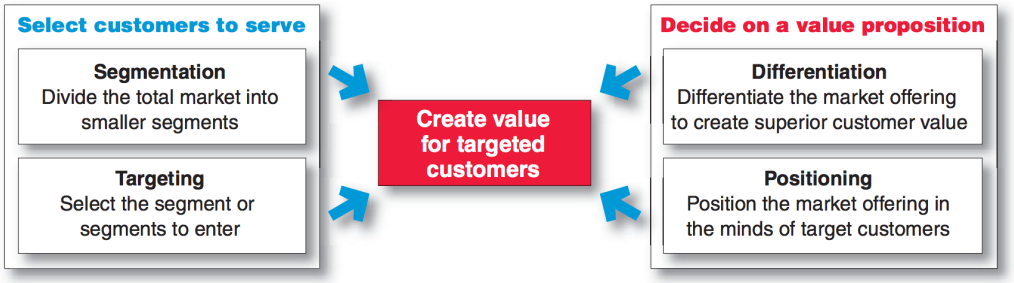


Figure 2: A Marketing Strategy

2.2 Machine Learning

Machine learning is a field in computer science that studies the construction of algorithms that learn from data, and can make predictions on new data. A machine learning process consists of two steps: training a machine learning model, followed by the process of testing the model. The testing process is important to be able to evaluate the performance of the model and compare different models and algorithms.

Because of the training and testing process of machine learning, the data is often split into two separate subsets, a training dataset and a testing dataset. The main reason for this is to be sure that the model can generalize enough to make good predictions on new instances, and avoid so called overfitting.

2.2.1 Supervised and Unsupervised Learning

Machine learning can be divided in *supervised* and *unsupervised* learning. They differ in the overall goal of the model as well as in the requirements of the data used for training and testing the model (Russell and Norvig, 2003).

The goal of supervised learning is to train a model to be able to take a given data point and to map it into a desired output value. Such an output value can be in the form of either a class (a classification problem), or in the form of a continuous variable (a regression problem). Training data in supervised learning consists of instances of data points paired to such output values. In classification, the attribute containing the class identity of a data point is referred to as label. With the instances in the training data, the model generalize and is able to map attribute values to the respective classes, and thereby make predictions of what class a new instance belongs to.

In unsupervised learning, the data does not contain any labels and is referred to as unlabeled data, meaning that the outcome variable is unknown for all data points. The goal of such learning is to cluster the data into groups of similar data without specifically naming each group. This can give new insights about the data and therefore be useful in various applications.

Another area of machine learning is so called semi-supervised learning, which uses techniques of both supervised and unsupervised learning. Techniques of semi-supervised learning are useful when trying to make use of unlabeled data while solving a classification problem. These techniques are often useful since labeled data is often limited, while unlabeled data is generally easier to access.

2.2.2 K-means Clustering

K-means clustering is an unsupervised machine learning algorithm used to cluster data into a number (k) of clusters. Each cluster is assigned a center point, commonly known as a centroid, which is calculated as the mean of all of the data points belonging to the cluster. The K-means algorithm optimize such that each data point belongs to the cluster with the closest centroid. It is an iterative algorithm and consists of two main steps: assigning a cluster to each data point and updating the centroid of each cluster. These are repeated until no data point changes cluster when reassigned (MacKay, 2002). A more detailed description is explained below.

Algorithm Description

1. Randomly select k centroids (one for each cluster).
2. For each data point, calculate the distance between the data point and each centroid.
3. Assign each data point to the cluster with the closest centroid.
4. Calculate the new centroids for each cluster. The new centroid is calculated as the mean of all data points belonging to the cluster. A new centroid c, for a cluster with n data points denoted $\{x_1, x_2, ..., x_n\}$ is calculated as (1).

$$c = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

5. Repeat from step 3 until no data point changes cluster.

The optimal cluster for each data point is found when the algorithm converges.

Using the K-means algorithm requires the prior knowledge of the constant k, the number of clusters. In many real world machine learning implementations, this constant is not known, as it depends on the prior knowledge of the data, which can be limited. There are various approaches for finding the optimal k for a given dataset, such as Gaussian tests (Hamerly and Elkan, 2003) or maximizing the Bayesian information criteria (BIC) (Pelleg and Moore, 2000). One of the most common and easily implemented approaches is the elbow method.

Elbow Method

The basis of the elbow method is to run the K-means algorithm several times with different values of k, i.e. alter the number of clusters for each iteration. For each iteration, the within-clusters sum of square (WCSS), i.e. the variance, is calculated. WCSS is plotted as a function of the number of clusters, k. The idea behind the elbow method is that the marginal gain of adding a new cluster will for some k drop dramatically. This will be clearly visible in the plot as an angle (elbow), and the optimal number of clusters is found at that k. (Bholowalia and Kumar, 2014)

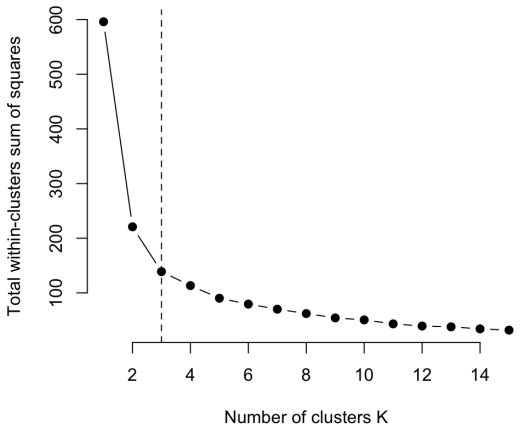


Figure 3: Elbow Method - Example of Plot

2.2.3 Random Forests

Random forests is a supervised machine learning algorithm commonly used in classification and regression problems, built on a simpler machine learning algorithm called decision trees. A decision tree is a binary tree, where each node contains a condition. Each node branches out depending on if the condition is true or false. By following the decision tree and determining each condition, a prediction can be made. This can be graphically represented as in figure 4.

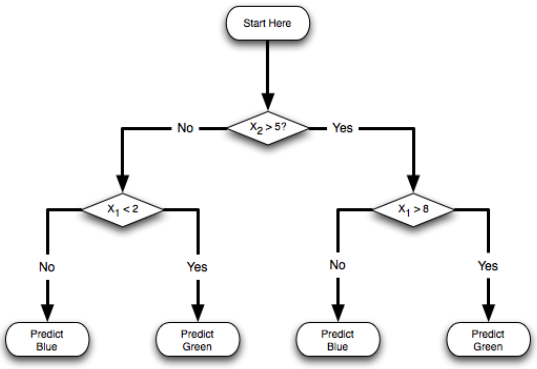


Figure 4: Illustration of a simple decision tree

The main idea behind random forests is to generate many decision trees during training, where each decision tree is trained with a randomly selected sample set of the training data. A prediction is determined by the average prediction of all of the decision trees in the model. Despite the fact that a decision tree easily overfit, random forests counteract overfitting well due to combining many decision trees (Breiman, 2001).

Random forests counteract this phenomenon because each decision tree is trained using a randomly selected sample set of the training data. Sample sets are selected using bootstrapping, which, when performed many times to create several training sets, is referred to as aggregated bootstrapping, or bagging. The algorithm for the random forest algorithm is described below:

Algorithm Description
Let B be the number of times for bagging. Then, for $b = 1$ to B:

1. Sample, with replacement, B examples from the training set.
2. Train a decision tree using the sampled examples.

(Hastie et al., 2001)

2.2.4 PU-learning

PU-learning (= *positive and unlabeled learning*) is a subfield of semi-supervised machine learning, where the available data consists only of positive instances, P, and unlabeled instances, U. An unlabeled set U consists of some combination of positive instances P along with negative instances N. Models are built with the goal of classifying a given instance to either P or N (Li and Liu, 2003). Accomplishing this is generally more difficult than traditional classification techniques in supervised learning, but has proven to be possible (Plessis et al., 2014). PU-learning is relatively new, and a more detailed study of recent research and current techniques is described in 2.3.

2.2.5 Evaluation Measures

To compare and evaluate a machine learning classifier, different measures can be calculated using the predicted and the true label of instances in the test data. In a binary classifier the data points in the test set can be divided into four categories, as shown below:

- True Positives (TP) - Data points correctly classified as positive.
- False Positives (FP) - Data points incorrectly classified as positive.
- True Negatives (TN) - Data points correctly classified as negative.
- False Negatives (FN) - Data points incorrectly classified as negatives.

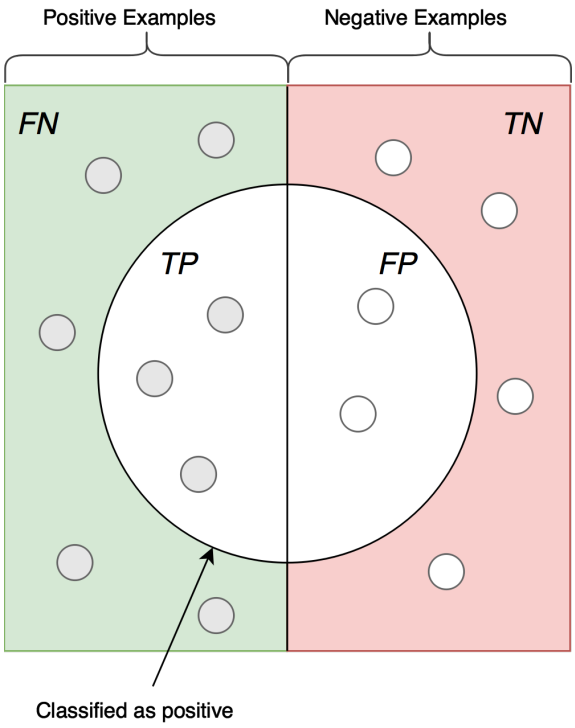


Figure 5: Illustration of evaluation terminology

Precision

Precision, p, is a measure of the number of instances that was correctly classified as positives, relative to all of the positive predictions. It can be interpreted as the probability that a randomly selected positively predicted instance, is truly positive. Precision is calculated as:

$$p = \frac{TP}{TP + FP} \tag{2}$$

Recall

Recall, r, is a measure of the number of correct positive predictions relative to the total number of positives instances. It is a measure of how good the model is at retrieving all of the positive data points. Recall is calculated as:

$$r = \frac{TP}{TP + FN} \tag{3}$$

F1 Score

Since precision and recall measure different aspects of the model, both needs to be considered when evaluating a model. A commonly used measure that considers both the precision and recall is the F1 score. It can be used as a single measure to compare the general performance of different models and is calculated as the

weighted average of p and r, as described in equation (4):

$$F_1 = \frac{2pr}{p+r} \tag{4}$$

2.3 Previous Research

For a further understanding of and to theorize the implementation and evaluation aspects of PU-learning, a few research papers were studied in more detail. The most relevant aspects for our study are presented in this section.

2.3.1 PU-learning Using an Adapted Classifier

One of the more recent approaches in PU-learning, presented and proven successful in a research paper by Elkan and Noto (2008), involves adapting an ordinary classifier algorithm to perform PU-learning. It is explained as follows.

Given a sample set of unlabeled and positively labeled instances, respectively (a PU-set), assume first that the positively labeled instances are selected randomly from *all* of the positive instances (i.e. the sum of labeled and unlabeled positive instances). This is known as the *selected completely at random assumption*. If this assumption holds, the difference in the conditional probabilities in a model based on a regular training set, and a model based on a training set with positive and unlabeled examples is a constant factor. For a full mathematical proof of this, see the article by Elkan and Noto (2008). The consequence of this proof is that a regular classifier model can be adapted to account for the PU-set, using the constant in the conditional probability functions. A method for such an adaptation is described in the paper (Elkan and Noto, 2008). This technique enables a way to implement PU-learning by adapting ordinary classifiers before training, which will be used in this paper.

2.3.2 Evaluating a PU-learning Model

Since the test data in PU-learning lacks negative instances, it is impossible to calculate precision, making it impossible also to calculate the F1 score using the methods described in section 2.2.5. Evaluating a PU-learning model, and in extent comparing the performance of different models, requires an alternative approach. In a research paper by Lee and Liu (2003) a solution to this challenge is presented. A relatively simple alternative measure of performance is presented as in equation (5). A mathematical derivation of how to estimate the measure using only positive and unlabeled instances is also presented in the paper, and described through equations

6 to 9. We start with describing the mathematical denotation:

Denote the input arguments of each instance X and the corresponding label as Y, where Y is either positive ($Y = 1$), or negative ($Y = -1$). Also, denote the classifier as f, for which the performance is measured. Denote precision and recall as p and r. Then:

$$Measure = \frac{p * r}{P(Y = 1)} \tag{5}$$

To see that the measure can be estimated using only positive and unlabeled instances, note that p and r can alternatively be formulated as:

$$p = P(Y = 1|f(X) = 1) \tag{6}$$

$$r = P(f(X) = 1|Y = 1) \tag{7}$$

Using Bayes' theorem, and lastly equation (5) and (6), we get:

$$\begin{aligned} &P(f(X) = 1|Y = 1) * P(Y = 1) = P(Y = 1|f(X) = 1) * P(f(X) = 1) \\ &\Leftrightarrow \frac{P(f(X) = 1|Y = 1)}{P(f(X) = 1)} = \frac{P(Y = 1|f(X) = 1)}{P(Y = 1)} \Leftrightarrow \frac{r}{P(f(X) = 1)} = \frac{p}{P(Y = 1)} \end{aligned} \tag{8}$$

Multiplying both of the sides in (8) by r, an alternative formula of the measure is derived:

$$Measure = \frac{p * r}{P(Y = 1)} = \frac{r^2}{P(f(X) = 1)} \tag{9}$$

Equation (9) is our end result. This formula of the measure can be calculated using only positive and unlabeled instances. Recall can be calculated using the positive labeled data points in the test set, as described in section 2.2.5. $P(f(X) = 1)$ can be calculated by running the test set through the model and calculate how many instances that are classified as positive in relation to the total number of instances.

The measure is proportional to the F1 score, and will be large when both p and r are large and small when both p and r are small. (Lee and Liu, 2003) However, the measure will not always be in a range between 0 and 1, unlike the F1 score. Despite this, it is a good way to estimate the performance of a PU-learning model and compare different models. Hereinafter, this measure will be referred to as *F1 estimator criteria*.

3 Method

This section describe the different steps of the research, in a sequential order of which they were performed.

3.1 Preparatory Work

As a first step, an initial meeting was held with the business intelligence consultant and manager Disa Törnvall at Exsitec. The purpose of the meeting was to understand Exsitec's value proposition and the type of services they deliver. Disa explained the different business areas, the process of delivering the services and what type of companies that usually order from Exsitec. This gave key insights, especially used in the discussion of business implications later.

A longer semi-structured interview was conducted with Rasmus Toivonen, a business developer at Exsitec. The purpose was to gain detailed insight in the current sales process and characteristics of a usual customer. This interview also discussed how a machine learning model could be used by sales personnel and where it would fit in the current sales process.

3.2 Data Gathering

The data used in this study was compiled by gathering data from three different sources: Exsitec, the ITM-school at KTH and Swedish website Allabolag. Here, the process of gathering and selecting the data is described.

Firstly, the data received from Exsitec was a cross-sectional dataset containing information on all of their current and past customers. The information included dates on when they became customers, which of the three business areas they bought services from, and if they had ended their partnership with Exsitec, it also included that year. In total, 1316 observations was included, i.e. information about 1316 different customers that are or have been customers at Exsitec. It was fetched from Exsitec's internal database management system.

Secondly, the ITM-school at KTH provided a dataset with panel data on all companies in Sweden, both listed and unlisted. It contained publicly available information, mostly financial data, on all companies for the time period of 2005-2015. The information was fetched through a database called Serrano created by Bisnode, with data originating from Statistics Sweden (SCB) and the Swedish Companies Regis-

tration Office (Bolagsverket), a Swedish authority that collects financial data from all Swedish firms on a yearly basis (Swedish House of Finance, 2015). The Serrano database has then aggregated these data (see figure 6) and adjusted for broken accounting periods. The dataset contained 6.48m observations in total, i.e. 10 years of data on 648k companies.

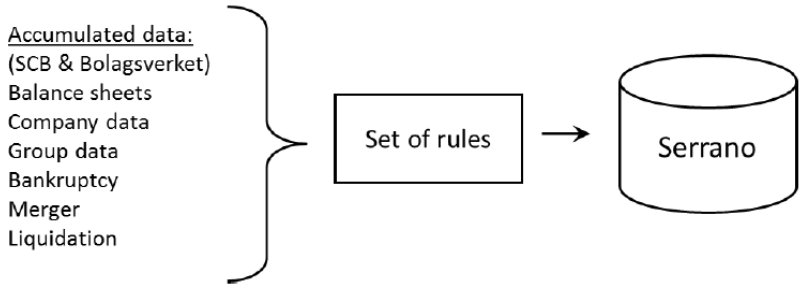


Figure 6: The Serrano Database

Not all data from the Serrano-database was needed. Two selections on companies from the Serrano-database was made which built the foundation for the rest of analysis:

1. Data on Exsitec's customer-base, a set of 1316 companies
2. Data on a randomly selected[1] set of companies, from here on described as *unlabeled* companies, a set of 10,000 companies[2]

Lastly, additional company specific data was fetched by building a web scraper and scraping the Swedish website Allabolag.se on the companies included in the study. Three attributes were fetched: the age of the company's CEO, the mean age of the board and the sector/industry that the company operates in.

It is worth stressing the importance of having good data for this study. A machine learning model's predictions are based on the patterns in the data it was trained with, so the results of this thesis relies on the data being valid and of good quality. Measure errors in the predictors may bias the model downwards (Stock and Watson, 2015) and missing data leads too less observations to train on, giving a less accurate model. It is however reasonable to assume that the quality of all three data sources

[1]The random selection was made in Python through the pseudo-random sample()-method in the Pandas-library

[2]The reason for selecting a larger set of companies here was simply availability

are good enough: the data from Exsitec is used internally and therefore be reliable, the Serrano database should be reliable since Bisnode has an incentive to provide a good product plus the fact that the original data stems from Swedish authorities. Allabolag.se is also a company offering their company-data as a product and falls under the market incentives as well.

3.3 Data Preprocessing

As a first step, the data set from Exsitec containing the list of customers was examined and cleaned up by removing duplicates. Public institutions and municipalities were removed from the list, since only private companies report the type of financial statements that is needed for conducting this study. Only limited liability companies were kept, reducing the number of companies to 894.

The three datasets were merged together into one. Because the Serrano-dataset contained data for a 10-year time period, a decision had to be made on which year to use for the features for each company. The following assumption was made:

- **Assumption 1**: If the company became customer at Exsitec year k, then year $k-1$ contains the most valuable information for deciding why it became customer.

The reasoning behind this assumption is that companies that buy services from an IT-consulting firm probably has some unique characteristics when they decide to become customers. These characteristics are not directly observable so the closest we can come to observe them using public data is by using the latest reported information. When it comes to financial statements, which is the basis for the dataset used in this paper, data is reported annually, and thus it is not possible to get more recent data than last year's financial statement when using the model for predictions.

The merging was done by matching the datasets on their organizational number and by choosing the company specific data from the Serrano-dataset and the Allabolag-dataset on year $k-1$ for companies that became customer at year k. [3] This resulted in a dataset where every row represented one of Exsitec's customers, and the columns all of the company specific data from the data sets. The dataset was cleaned up and cleared from outliers such as inactive companies. Another assumption was made here:

[3]Some information from year $k-2$, $k-3$ ... $k-5$ was also fetched for calculations of e.g. growth numbers.

- **Assumption 2**: All of Exsitec's current and potential customers has > 1 MSEK in revenues, and > 0 employees.

This assumption was made after finding that some of the companies reported as customers was in fact trust companies and holding companies, therefore having extreme values such as 0 employees which is not representable for the underlying operating business.

As in most real world data mining problems, the data contained missing values in some of the instances. A common solution is to replace the missing value by the mean value of that feature. (Tan et al., 2005) This technique is especially useful when the number of instances in the dataset is limited, which was the case for this dataset.

Lastly, the data was scaled and standardized. This is important due to the fact that there is a great variation of magnitude in the variables. For example, revenues for companies are often spanning between 1-500 million while margins usually lies in the range of 0-1, meaning that the revenue variable will weigh heavier in an implemented algorithms. These differences in magnitude should not weigh in on the prediction; most machine learning estimators perform badly if the data is not relatively normally distributed.

For the dataset of unlabeled companies, the same procedures as described above was also conducted.

3.4 Model Building

The goal of the classification models built in this paper was to be able to classify an unlabeled company as *1) potential customer (positive)* or *2) not potential customer (negative)*. The two sets of data described in previous sections, customers and unlabeled companies, were used in the models for training and testing, respectively, as seen in figure 7.

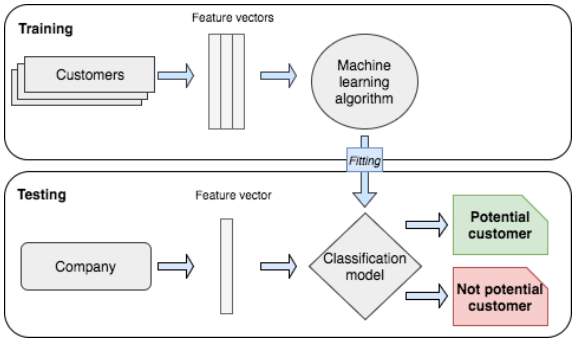


Figure 7: Customer classification

Note that this classifier differs from a textbook example of a supervised binary classifier. In such a case, the classifier would be trained on a mixture of positively and negatively labeled instances to be able to predict which one an unlabeled instance belongs to. Here, the only labeled data is Exsitec's own customer base, meaning that there are no negatively labeled instances. According to research, using a dataset with such a skewed distribution directly as training data on a conventional machine learning algorithm (SVM, logistic regression etc.), would lead to imbalanced training and produce biased results (Kubat et al., 1997). In order to cope with this problem, two different techniques were used: a *clustering based classifier* (CBC) and *PU-learning*.

Because the Serrano-database contains hundreds of features, a process of feature selection was carried out. This resulted in three different sets of features, referred to as Model 1, Model 2 and Model 3 (summarized in table 1). The idea was to start with a small number of features and then scale up to a larger set of features in order to avoid the curse of dimensionality (Keogh and Mueen, 2010). The features included were selected using a business perspective on which types of variables that reveals something about the state of a company. Thus, model 1 starts with a few key features about the size of the company, how profitable it is and how much debt it has. Model 2 and 3 then expands and incorporates features about the companies' growth rates in the last few years, more financials from the balance sheet and income statement, plus a few odd features such as the age of the companies' CEO.

Table 1: Features selected

Feature	Model 1	Model 2	Model 3
Net Sales	x	x	x
# Employees	x	x	x
# Workplaces	-	x	x
CEO age	-	x	x
Board avg age	-	x	x
EBIT	x	x	x
Net Income	-	x	x
Profit margin	-	x	x
Growth	-	x	x
Growth 3 y avg	x	x	x
Growth 5 y avg	-	x	x
EBIT g	-	x	x
EBIT g 3 y avg	-	x	x
EBIT g 5 y avg	-	x	x
*County**	-	x	x
*Industry**	-	x	x
Labor cost	-	x	x
D&A	-	-	x
Total Assets	-	-	x
Leverage	x	-	x
Solidity	x	-	x
Capital turnover	-	-	x
Inventory turnover	-	-	x
Return on capital	-	-	x
Quick ratio	-	-	x
Rev per emp	-	-	x
Increase in emp	-	-	x
Patents	-	-	x
Swe parents	-	-	x
Swe subsidiary	-	-	x
Foreign subsidiary	-	-	x
Property & buildings	-	-	x

** = represented by dummy (binary) variables*

The final representation of the companies after feature selection was in the form of feature vectors. Every company $c_1, c_2, \ldots, c_{644}$ that is customer to Exsitec was represented by a feature vector $x_1, x_2, \ldots, x_{644}$ such that:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{643} \\ x_{644} \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,m-1} & a_{1,m} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,m-1} & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{643,1} & a_{643,2} & \ldots & a_{643,m-1} & a_{644,m} \\ a_{644,1} & a_{644,2} & \ldots & a_{644,m-1} & a_{644,m} \end{bmatrix} \tag{10}$$

where a represents a feature, and $m = 6$ for model 1, $m = 66$ for model 2 and $m = 79$ for model 3.[4]

Analogously, each of the unlabeled companies $u_1, u_2, \ldots, u_{7247}$ was represented by feature vectors $y_1, y_2, \ldots, x_{7247}$ as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{7246} \\ y_{7247} \end{bmatrix} = \begin{bmatrix} b_{1,1} & b_{1,2} & \ldots & b_{1,m-1} & b_{1,m} \\ b_{2,1} & b_{2,2} & \ldots & b_{2,m-1} & b_{2,m} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ b_{7246,1} & b_{7246,2} & \ldots & b_{7246,m-1} & f_{7246,m} \\ b_{7247,1} & b_{7247,2} & \ldots & b_{7247,m-1} & b_{7247,m} \end{bmatrix} \tag{11}$$

where b represents the features and m is defined as above.[5]

An initial data examination was done to notice if any difference could be spotted between a sample of customer and a sample of unlabeled companies. Since it is impossible to visualize a high dimension feature space graphically, the feature spaces were reduced to two dimensions. This was done with linear dimensionality reduction using singular value decomposition, enabling a two dimensional scatter plot.

3.4.1 Clustering Based Classifier

Clustering is a common and frequently used technique for data driven segmentation in marketing research (Wu and Lin, 2005). As Exsitec has three different business areas and a wide variety of customers, it makes sense to use this approach as a basis for classification. The following assumption was made to allow this approach:

- **Assumption 3**: Exsitec's customers can be divided into different segments

[4]This differs from the amount of variables in table 1 due to binary representations of *County* and *Industry*.

[5]The matrices are implemented as NumPy arrays in Python.

The idea is that there is not only one reason why customers buy Exsitec's services, but rather a set of generalizable reasons. This give different market segments of different types of current and potential customers. For example, one segment might be fast-growing companies that need help with their internal IT-systems, while another segment might be companies that are doing badly and therefore makes an effort and hires consultants. According to Kotler and Armstrong (2010), segmentation for business customers can be done in many ways, for example by geographical location, industry or situational factors. The point of the clustering algorithm is thus to combine such factors through the features included in the models, and to find the most fitting segments derived from the companies' feature vectors.

The clustering algorithm chosen in this paper was K-means. It is a partitional clustering algorithm, compared to an hierarchical algorithm that assumes that the clusters can be structured in an hierarchy. This would add unnecessary complexity for our data, and since market segments can be assumed to be independent of each other, a partitional clustering algorithm is preferable. K-means was chosen because it is easy to implement and fast to run (Hastie et al., 2001).

In the implementation of the K-means algorithm [6], the euclidean distance function was used for calculating the distance between two data points q and p, as:

$$distance(q,p) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{12}$$

Where every data point q or p is represented by a feature vector x_k or y_k.

The clustering is however not able to classify instances of companies in itself. Instead, a simple classification model was built on top of the result of the clustering. This was conducted with the following algorithm.

CBC - Algorithm Description

1. Split customers $c_1, c_2, \ldots, c_{644}$ into a training set (80%) and a test set (20%) [7]
2. Cluster the training-data using the K-means algorithm
3. Calculate a threshold distance d_t to each centroid cluster by taking the distance to the 95th percentile data point in the training set

[6] The implementation of K-means from the Scikit-learn library (learn) for python was used.
[7] The split was done to be able to calculate recall later

4. Predict the closest cluster for both the test-set and for the unlabeled companies
5. Calculate the distance to the closest centroid, d_k for the test-set and the unlabeled distance, respectively
6. For every such distance d_k:
   (a) if $d_k < d_t$, classify as *potential customer*
   (b) if $d_k > d_t$, classify as *not potential customer*

The idea behind the algorithm is that if an arbitrarily chosen unlabeled company is similar enough to a company in its nearest cluster centroid, it should be classified as potential customer. To get a measure of similarity, the distance to the nearest centroid can be used, where the smaller the distance, the more similar the company is to a typical Exsitec-customer. An exact threshold for the classification algorithm that translates to "similar enough" is somewhat arbitrarily chosen. The 95th percentile of the distance to nearest centroid for Exsitec's customer was chosen to remove the worst outliers while at the same time trying to minimize the amount of false negatives for the customers. Splitting the customer-data allowed us to test the model on both unlabeled instances as well as on labeled instances.

Lastly, a sanity check was made on this approach where the distances for the test-set on the customers and the unlabeled companies was compared. This was plotted in histograms for the 3 different models, with the associated descriptive statistics for the distances. Because of the difference in sample size on the customer test-data and the unlabeled data, a one-sided students t-test was also conducted on the three different models with the following null hypothesis:

$$H_0 : \mu_c = \mu_u \tag{13}$$

where μ_c is the mean of the centroid-distances for the customer test-data and μ_u is the mean of the centroid-distances to the unlabeled data. The purpose here was to find whether the differences in centroid-distances are statistically significant or not.

3.4.2 PU-learning

As an alternative approach, PU-learning was chosen due to the characteristics of the data and the problem. Identifying new potential customers for marketing purposes has been suggested as a real world problem where PU-learning can be useful. A company generally has data of current customers (positive instances), and data regarding unlabeled instances can often be bought from a database. (Wang et al.,

2011) It is therefore interesting to implement PU-learning in this paper.

The PU-learning algorithm implemented is based on the technique developed by Elkan and Noto (2008), described more thoroughly in section 2.3.1. More specifically, the implementation of such a PU-adapter by Drouin (2013) was used.[8] Random forests was chosen as the classification algorithm to adapt for PU-learning because it is well compatible with the PU-adapter implementation and does not overfit easily. The scikit (learn) implementation of random forests was used with 100 trees in the forest.

Each instance of a customer was assigned the label 1, while each unlabeled instance was assigned the label -1. The data was split into a training and a test set, using 80% of the data for training, and 20% for testing. The number of instances for each label and phase is noted in table 2.

Table 2: Data for PU-learning

	Positive	**Unlabeled**	**Total**
Training	5803	509	6312
Testing	134	1444	1578

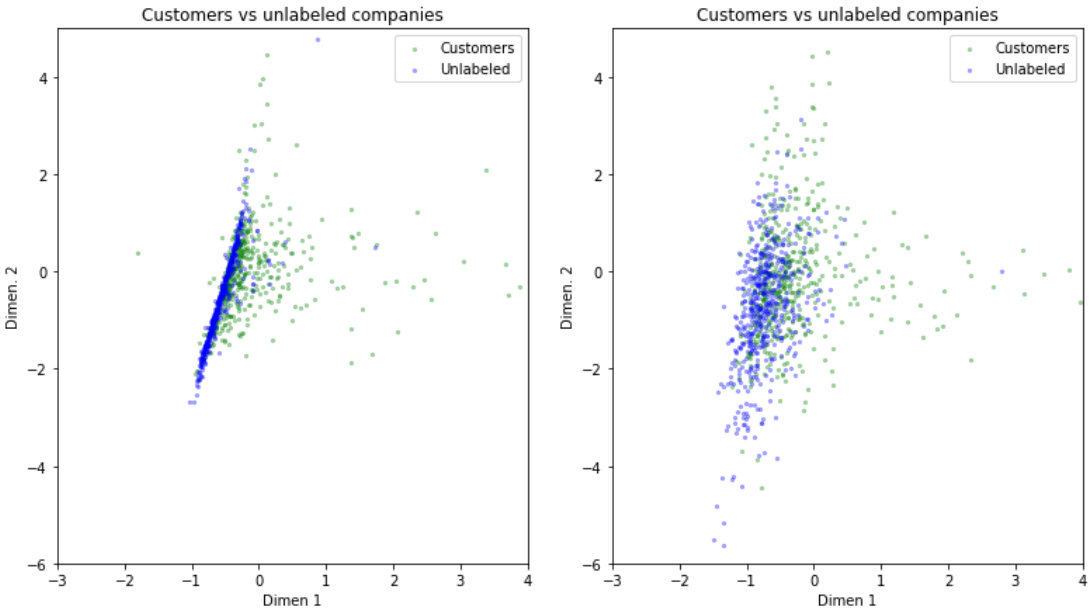
[8] All of this was implemented in Python.

4 Result

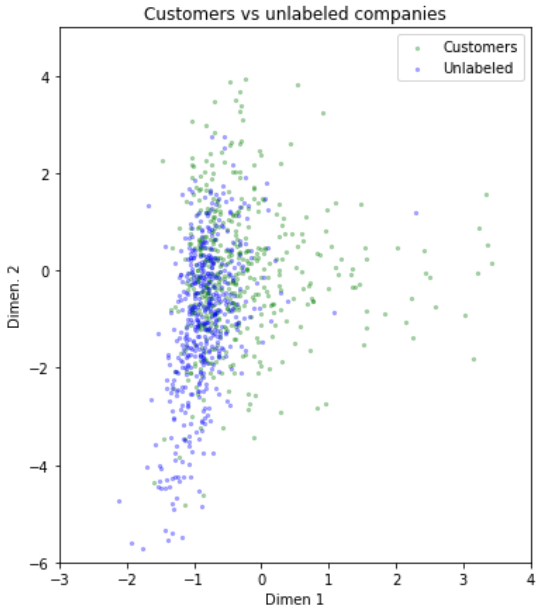
4.1 Feature Vectors

The initial inspection of the feature vectors for Exsitec's customers, X and the unlabeled companies, Y, is presented in scatterplots for each of the three models.

Figure 8: Dimension reduced scatterplot of customers and unlabeled companies



(a) Model 1 **(b)** Model 2



(c) Model 3

The plots above (figure 8) shows that there seems to be a difference in the feature

vectors for customers and a set of unlabeled data for all three models. Thus, Exsitec's customer base differs from a random set of companies in regards of these features, meaning that there is hope for our models.

4.2 Clustering

Below is the results of the Elbow method presented. Three graphs are plotted for each of the three models because of the non-deterministic nature of the K-Means algorithm.

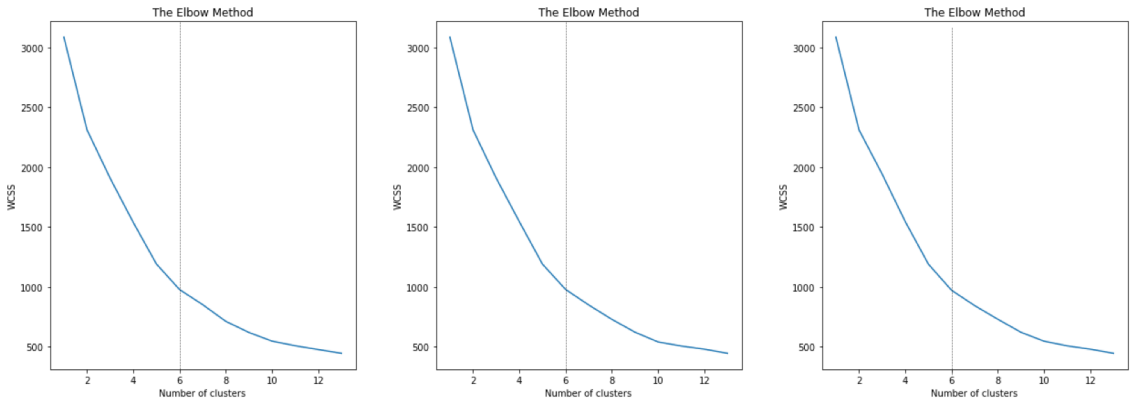


Figure 9: Elbow method - model 1

Model 1 exhibit a somewhat smooth curve with no obvious elbow point. After inspection, the sharpest drop in slope was found at $k = 6$ number of clusters.

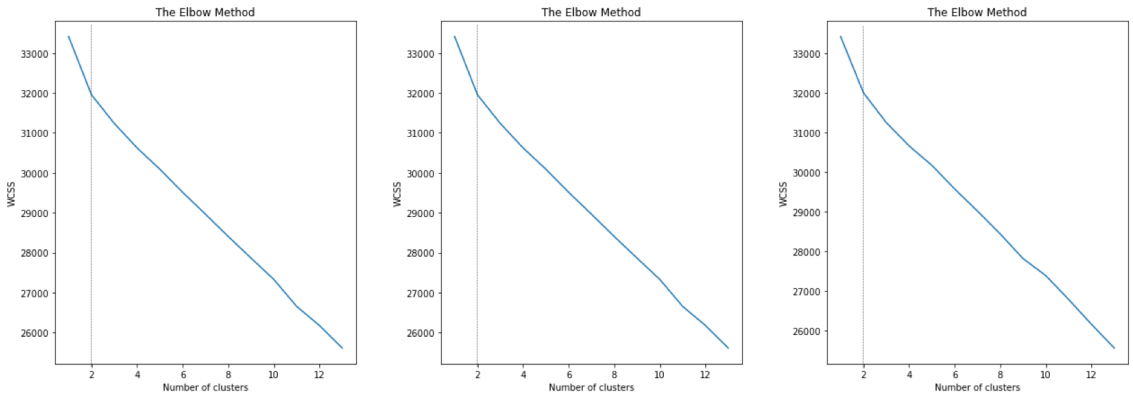


Figure 10: Elbow method - model 2

Model 2 conveys a less smooth and more halting curve. The elbow point in all three examples was best found in $k = 2$.

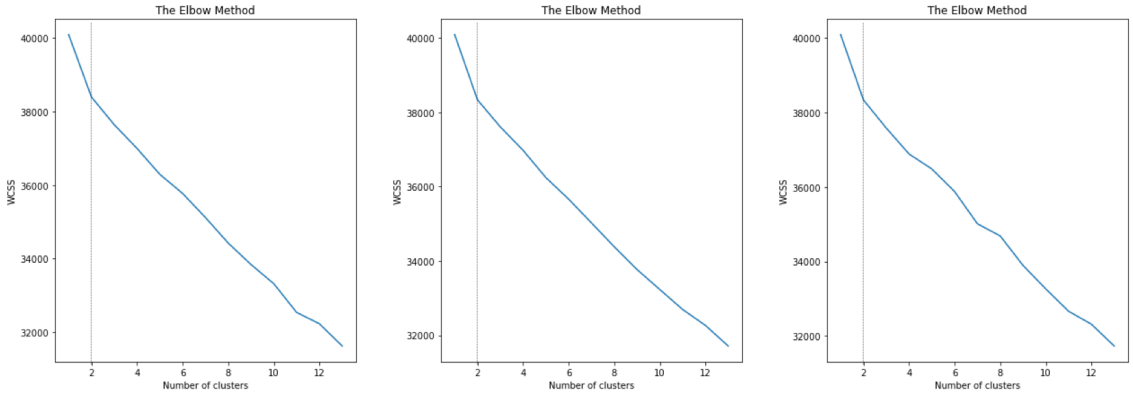


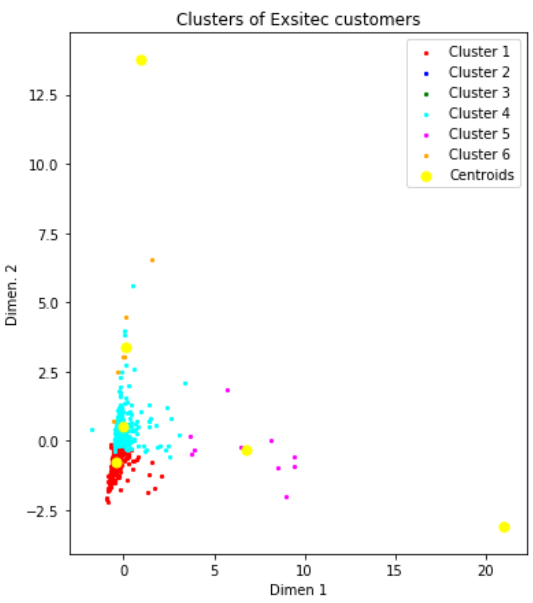
Figure 11: Elbow method - model 3

Model 3 is similar to model 2 when plotting WCSS as a function of number of clusters, but is even more choppy, as can be seen in figure 11. At $k = 2$, the optimal number of clusters seems to be found.

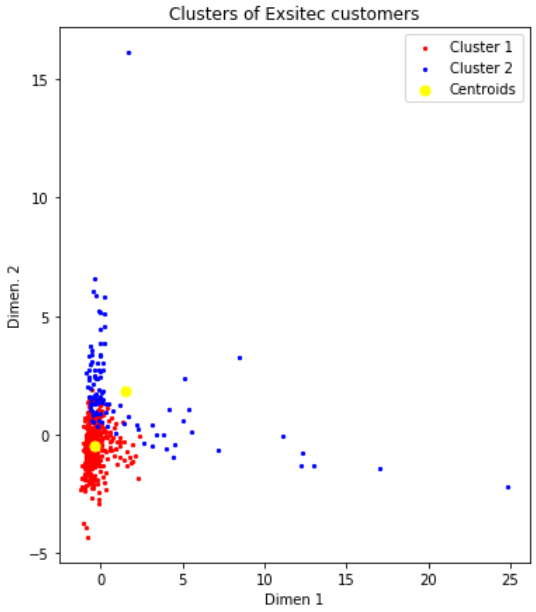
For all three models above (figure 9, 10 and 11), the elbow point, i.e. the point where the marginal gain of adding another K drops dramatically as described by Bholowalia and Kumar (2014), is not obvious by examining the graphs. In model 2 and 3, the rate of change looks almost constant, and while the rate of change decreases with K in model 1, the shape is concave with a second derivative that is similar among all K. This means that the optimal number of clusters is ambiguous and might indicate that there are no clear segments that Exsitec's customers can be divided into.

A visualization of the clusters created with the K-means algorithm is presented in figure 12. The same dimensionality reduction as the feature vector representation has been made, with the difference that only the actual clusters and its corresponding instances for the training dataset are plotted here.

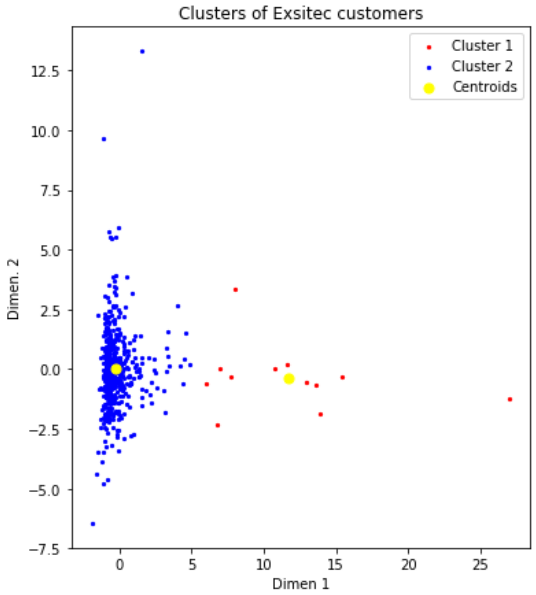
Figure 12: Cluster representation reduced to 2 dimensions



(a) Model 1



(b) Model 2



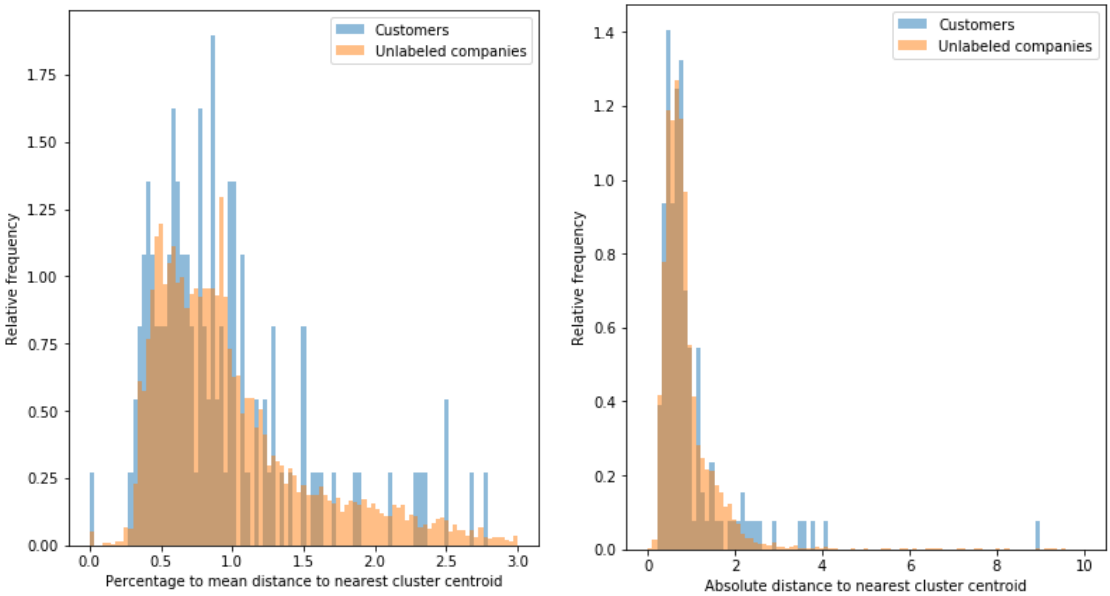
(c) Model 3

4.3 Centroid Distance

The similarity between companies was calculated as the euclidean distance to nearest centroid that each company belongs to. These distances is visualized in histograms (as presented in figure 13 14 and 15), comparing customers with unlabeled companies for each of the three models, and in descriptive statistics in table 3. Distance is

measured as percentage of mean distance to the training-set for each centroid.[9] A company with a distance of 1.00 would thus mean that it lies on the same distance as an average customer of Exsitec, interpreted as equally similar to an average customer to Exsitec.

Figure 13: Model 1 histogram - distance to centroids

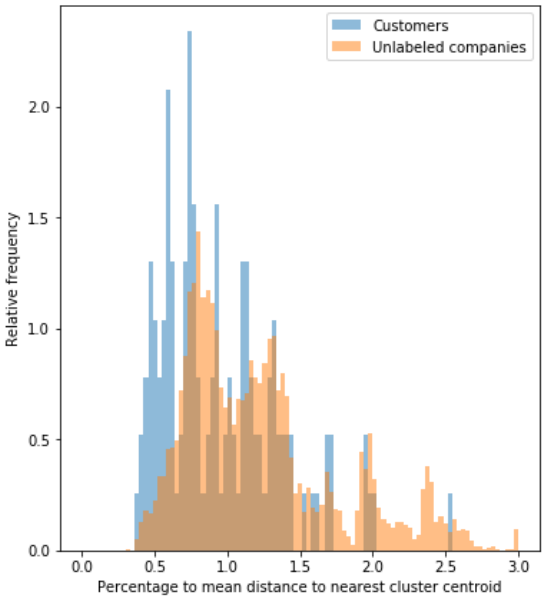


(a) Relative to nearest centroid mean

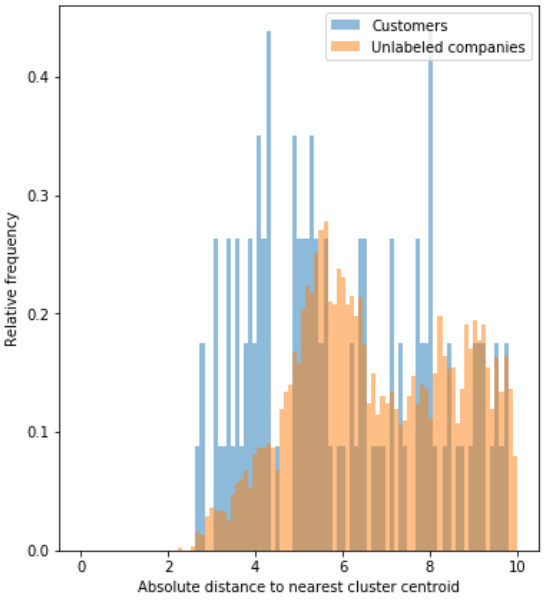
(b) Absolute distance to nearest centroid

[9]Using a relative measure allowed aggregation of the distances to each cluster centroid, as a normalization

Figure 14: Model 2 histogram - distance to centroids

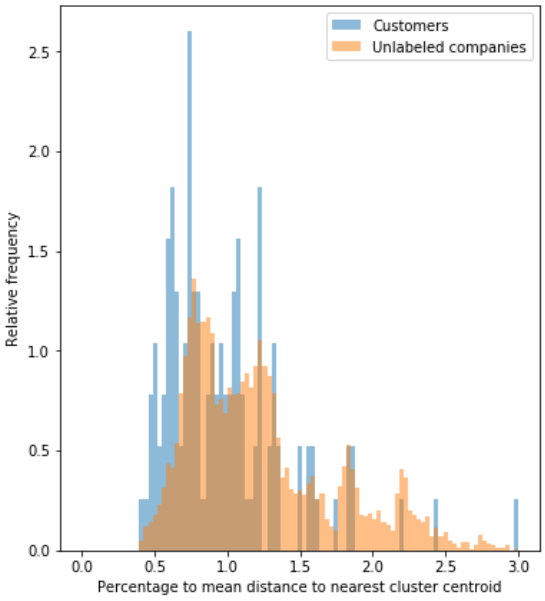


(a) Relative to nearest centroid mean

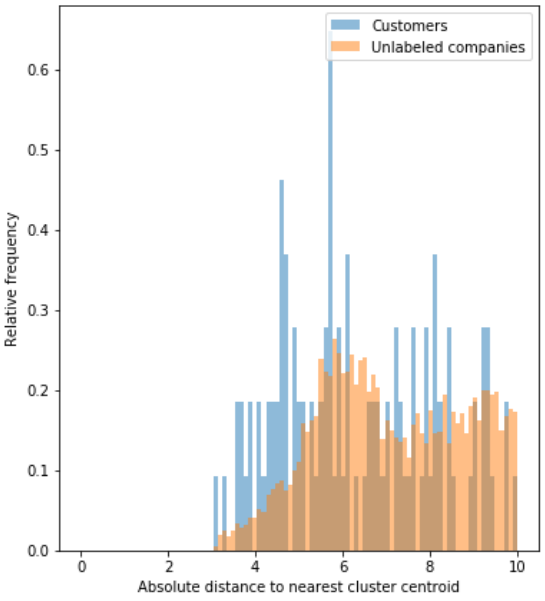


(b) Absolute distance to nearest centroid

Figure 15: Model 3 histogram - distance to centroids



(a) Relative to nearest centroid mean



(b) Absolute distance to nearest centroid

Model 1 that only uses a few variable show that there does not seem to be a large difference in distance to centroids comparing customers with unlabeled companies. However, when including more variables as in model 2 and model 3, the distance gap grows larger in the histograms. This is confirmed by table 3, examining the

Table 3: Percentage of mean distance to centroid - descriptive statistics

	Model 1		**Model 2**		**Model 3**	
	Customers	Unlabeled	Customers	Unlabeled	Customers	Unlabeled
$Count$	129	7247	129	7247	129	7247
$Mean$	1.04	1.03	1.01	1.44	1.01	1.55
Std	0.81	0.70	0.63	3.67	0.54	3.56
Min	0	0	0.39	0.33	0.40	0.41
Q_1	0.56	0.59	0.61	0.83	0.67	0.84
$Median$	0.78	0.85	0.89	1.14	0.89	1.14
Q_3	1.15	1.21	1.60	1.52	1.19	1.60
Max	4.82	12.62	5.31	194.98	3.88	158.48

mean distance between test-data on customers and unlabeled data. In model 1, the unlabeled companies was 1 percentage unit closer than customers to each centroid, conveying that on average, the two datasets was similar in these features. Model 2 shows a mean distance of 1.01 of the test-customers and 1.44 of the unlabeled companies, meaning that there does seem to be a significant difference in terms of euclidean distance to centroids when using more variables. This distance grows even larger to 1.55 compared with 1.01 in model 3 when including the full set of features.

The standard deviation also shows that the test-set of customers generally lies in a small range to the mean, while the unlabeled companies varies quite much in model 2 and 3. This indicates that there does seem to be a difference in a random set of companies compared to a set of Exsitec's customers.

The Student's t-test confirms these results, described in table 4.

Table 4: $t - test$

	Model 1	**Model 2**	**Model 3**
$T - statistic$	0.90	-5.54	-6.43
$p - value$	0.37	$4.92e-8$	$3.40e-10$

For model 1, the p-value of 0.37 tells us that we cannot reject the null hypothesis that the customer test-data and the unlabeled companies have a different mean, on a 95% confidence level. For both model 2 and model 3 however, the null hypothesis

can be rejected on a 99% confidence level, which tells us that both of these models convey a significant difference in cluster-distance comparing Exsitec's customers with unlabeled companies.

4.4 Classification

The classification results from both the clustering based classifier (CBC) and the PU-learning approach are presented below. Both of the models are evaluated using the F1 estimator criteria described in section 2.3.2.

4.4.1 Evaluation Measures

Table 5: *Recall*

	Model 1	**Model 2**	**Model 3**
Clustering based classifier	0.915	0.953	0.946
$PU - learning$	0.694	0.724	0.776

Recall is here interpreted as the share of Exsitec's customers that were correctly classified as positive. The clustering approach was significantly better on predicting whether an existing customer should be classified as a potential customer than the PU-approach.

Table 6: $P(f(company) = 1)$

	Model 1	**Model 2**	**Model 3**
Clustering based classifier	0.922	0.835	0.807
$PU - learning$	0.156	0.164	0.171

$P(f(company) = 1)$ is calculated as the share of companies that were classified positive, i.e. potential customer for Exsitec. The result above thus shows that a large share of the companies were classified positive when using the clustering approach - a significant difference to the PU-learning approach. One possible explanation for this is that Exsitec has a large customer database with a wide range of companies, meaning that using the 95th percentile as the distance-to-centroid threshold captures

companies similar to a random set of companies, in terms of the features used. When including more features however, the amount of positive classified companies decreases. This might mean that the model is getting more precise by capturing more information about what type of company it is. For the PU-learning approach, the opposite is happening. Assuming that the true parameter lies somewhere in between the clustering and PU-learning data, this would also mean that the PU-learning approach gets better when including more variables.

Table 7: *F1 estimator criteria*

	Model 1	**Model 2**	**Model 3**
Clustering based classifier	0.908	1.089	1.108
$PU - learning$	3.102	3.192	3.52

The F1 estimator criteria is interpreted as the accuracy of the model, combining both recall and precision. Table 7 shows that the PU-learning method was more accurate in its predictions than the clustering based classifier. This can be traced due to the high probability that the CBC predicts any company to be a positive. In other words, compared to the PU-learning method, the clustering based classifier is a somewhat blunt tool for prediction. Also, like the two earlier tables, having more variables gives a better performance as the F1 estimator criteria increases in model 2 and model 3.

5 Discussion

5.1 Model Analysis

Consider the clustering approach. For model 1, an optimum for the number of clusters could be identified in the elbow graph (figure 9). However, from the elbow graphs of model 2 and 3 (figure 10 and 11), it is not clear what the optimal numbers of clusters are. This indicates that when more features are included, the clustering is not satisfying. When examining the visualizations of the clusters in figure 12, this conclusion seems to be confirmed. There are no clear grouping by eye and outliers affect the clustering a lot. Even though an optimum of 6 clusters are identified using the elbow method in model 1, two clusters contain the majority of the instances, as seen in figure 12. This further indicates that the data of Exsitec's customers is not naturally dividable into clusters.

There are two main possible explanations that the clustering did not succeed as intended:

- There are no underlying market segments of Exsitec's customers
- The data is insufficient enough to represent the segments of Exsitec's customers

From a business perspective it is natural to assume that Exsitec's customers can be divided into customer segments. However, since they also have a large variation in their customer base, there might be many companies that falls in between these segments, making the clusters more diffuse. The second explanation would also make sense; aggregated financial data as used in this report might not be enough to explain the different customer segments. There could potentially be many other variables that creates a better basis for dividing the companies into segments that publicly available information does not account for. One example is transactional data. However, using such data is out of the scope of this paper, that focuses solely on publicly available information.

It is important to reflect on these findings when evaluating the clustering based classifier. However, these results does not rule out the possibility of a successful classifier developed from the clustering estimator. This is due to the visualizations in section 4.2 that indicate a difference between customers and unlabeled companies, which supports the assumption that the data can proxy the underlying factors that identifies a customer. This is further supported by the differences between customers and unlabeled examples as shown in table 3, and proven significant in table 4.

Studying the histograms displayed in figure 13-15, and the statistics in table 3, it is worth noticing that the distance to the nearest centroid for the unlabeled instances increase as the number of features increases, relative to the same distance for the current customers. This leads us to conclude that the clustering models' ability to separate customers from unlabeled instances increases with the number of features.

From the evaluation measures of the models (section 4.4.1), a large difference in the f1 estimator criteria between the clustering based classifier and the PU-learning classifier is noticed, where the PU-learning classifier perform better in general. In particular, the best performing classifier is the PU-learning model using the model 3 feature set. This model will be further analyzed to determine how well it performs compared to a baseline, and is therefore denoted f_1 in the following paragraph.

For comparison, a baseline model can be set up as a model f_0, which randomly classifies input feature vectors as positive with a given probability x_0, such that $P(f_0(X) = 1) = x_0$. Since the baseline model classifies randomly, the recall r would be equal to x_0, $(r = x_0)$. The best classifying model developed, f_1, has $P(f_1(X) = 1) = x_1 = 0.171$ and the recall is 0.766. To compare these, the baseline model is set such that the probability of classifying an input as positive is the same as the best performing, that is $x_0 = x_1$. This shows that the best classifier performs a recall 4.47 times over the baseline.

5.2 Business Implications

To be able to fully answer the problem definition, a discussion on how a desirable model should perform is needed. This is derived from a business analysis on how a classifier could be implemented in the daily operations and how the predictions would be used.

In the interview with business developer Toivonen (2017-03-23) at Exsitec, the current sales process was discussed along with how a classification model would best be used. Comparing the process with the sales funnel, most of the daily operations of sales personnel at Existec are operating in the last steps of the funnel (closer to becoming a customer). This includes talking to leads to get an understanding of their problems and thereby identifying how Exsitec's services can help them. Less time is spent on direct marketing, which in theory could generate additional leads.

Currently, most of the leads are generated by recommendations from Exsitec's software partners, such as Visma. (Toivonen, 2017-03-23)

The current direct marketing operations is mainly marketing via email to potential customers. The companies that Exsitec are targeting are chosen by simple heuristics, such as companies with a turnover in a specific interval or in a specific geographic location. These heuristics are based on assumptions from Exsitec's personnel. Today, an external service provider is used to find companies that match these heuristics. The consequence of this it that it is costly, and limits their control of how these companies are selected. The content of the marketing emails is typically invitations to events, such as seminars. (Toivonen, 2017-03-23)

A machine learning model similar to one of those developed in this paper would be a possible replacement to the heuristics, and could be implemented to improve the targeting of the marketing. The predictions could be used for email based marketing as well as for cold calling and other ways of reaching out to new customers. For such implementations, it is more desirable that a model generates fewer, but more accurate suggestions, rather than a lot of potential matches. This is due to the nature of B2B-sales where every sale costs more compared to B2C-sales, and because the sales personnel would not have the resources to process a large set of potential customers.

With these business implications in mind, our results show that the best model has a high recall despite classifying a relatively small amount of companies as potential customers. This indicates that the model can work for the type of implementation that is desired. It would generate relatively few suggestions of potential customers, and the suggestions would be similar to the current customers. For the intention of identifying new potential customers for marketing, the performance requirements are relatively low. It is not expected that the model generate perfect suggestions, but rather performs better than heuristics. We therefore conclude that the best model is good enough for this purpose.

6 Conclusion

In this paper, we have investigated the subject of binary classification with positive and unlabeled data. It was conducted in a company specific setting with the purpose of predicting new potential customers for B2B-sales. The classifiers were based on a dataset with publicly available data of existing customers (positives), and a set of randomly selected companies fulfilling certain conditions (unlabeled data). Two methods were used to solve this problem. The first method was a naive clustering based classifier based on the K-means algorithm for clustering, with a simple classifier built on top by calculating and comparing distances to each cluster centroid for the companies. In the second method, a more advanced technique known as PU-learning was used in accordance to a paper by Elkan and Noto (2008) which allowed use of the random forest algorithm. For both of these models, three different sets of features were used to see how that would impact the results.

Comparing the performance of the two classifiers, we found that the best clustering based classifier achieved an F1 estimator criteria of 1.1 while the best PU-learning classifier achieved an F1 estimator criteria of 3.51. This means that the PU-learning approach was substantially better at the given task of predicting new customers. When evaluating the best performing PU-learning model in comparison to a random baseline classifier, the recall was 4.72 times better than the baseline, while the clustering based classifier was merely marginally better than the random baseline case. The clustering based classifier's low result can be partly be explained by a failure in clustering the customers. Furthermore, we conclude that more features leads to a better classifier in both approaches, and there was no problem with the curse of dimensionality even with > 60 features.

For it to be beneficial to implement a model in the marketing process, it is important to consider the nature of B2B operations. Since each customer requires substantial time from the sales force, the model must have a high accuracy and not generate too many bad suggestions. We can therefore conclude that our clustering based classifier would not be useful to implement in a B2B marketing process. However, the significantly better results of the PU-learning method means that it is likely that it performs good enough for a marketing implementation. This is our main finding and contribution with this paper.

Further research in this subject is recommended to conduct a similar type of study for different types of industries and discuss if the difference in value proposition

affects the result. Intuitively, industries with more distinct market segments would produce more accurate results. We would also encourage the use of other machine learning algorithms that were left out of the scope of this paper. From a business perspective, a case study where this type of model is actually implemented would also be of great interest to see if it actually generates more sales.

Lastly, our findings show that machine learning is a potentially valuable tool for companies when trying identify new customers in their marketing processing in B2B-sales. This thesis establish a glimpse of the possibilities and limitations of machine learning in a B2B-sales context, and lays a foundation of how the sales process can be reimagined in the future.

References

P. Bholowalia and A. Kumar. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *IJCA Journal*, 105(9), 2014.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565.

S. Davies. Building a business-to-business sales process. *Open Source Business Resource*, 10/2010 2010. ISSN 1913-6102. URL `http://timreview.ca/article/386`.

A. Drouin. Positive and unlabled learning (pu-learning). `https://github.com/aldro61/pu-learning`, 2013.

T. Economist. Million-dollar babies. `http://www.economist.com/news/business/21695908-silicon-valley-fights-talent-universities-struggle-hold-their`, 2016. Accessed: 2017-05-25.

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 213–220, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-193-4.

G. Hamerly and C. Elkan. Learning the k in k-means. In *Advances in Neural Information Processing Systems*, volume 17, 2003.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Ironpaper. 4 tips to improve your b2b sales funnel. URL http://www.ironpaper.com/webintel/articles/4-tips-to-improve-your-b2b-sales-funnel/. Accessed: 2017-05-08.

M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015. ISSN 0036-8075. doi: 10.1126/science.aaa8415. URL http://science.sciencemag.org/content/349/6245/255.

E. Keogh and A. Mueen. *Curse of Dimensionality*, pages 257–258. Springer US, Boston, MA, 2010.

P. Kotler and G. Armstrong. *Principles of Marketing*. The Prentice-Hall series in marketing. Pearson, 2010.

P. Kotler, I. Michi, and W. Pfoertsch. *B2B Brand Management*. Springer Berlin Heidelberg, 2006.

M. Kubat, R. Holte, and S. Matwin. *Learning when negative examples abound*, pages 146–153. Springer Berlin Heidelberg, Berlin, Heidelberg, 1997.

S. learn. Scikit-learn - machine learning in python. URL http://scikit-learn.org/. Accessed: 2017-05-08.

W. S. Lee and B. Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003)*, pages 448–455, 2003.

X. Li and B. Liu. Learning to classify texts using positive and unlabeled data. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, pages 587–592, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

C. X. Ling. Introduction to special issue on machine learning for business applications. *ACM Trans. Intell. Syst. Technol.*, 2(3):18:1–18:2, May 2011.

D. J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002.

D. Pelleg and A. Moore. X-means: Extending K-means with efficient estimation of the number of clusters. In *Proc. 17th International Conf. on Machine Learning*, pages 727–734. Morgan Kaufmann, San Francisco, CA, 2000.

M. C. d. Plessis, G. Niu, and M. Sugiyama. Analysis of learning from positive and unlabeled data. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 703–711, Cambridge, MA, USA, 2014. MIT Press.

S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2 edition, 2003. ISBN 0137903952.

J. H. Stock and M. W. Watson. *Introduction to Econometrics*. Pearson, third edition, 2015.

Swedish House of Finance. Swedish house of finance, annual report 2015, 2015. URL `http://www.houseoffinance.se`.

P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

R. Toivonen. Personal Interview, 2017-03-23.

D. Törnvall. Personal Interview, 2017-03-23.

H. Wang, S. Li, S. Oyama, X. Hu, and T. Qian. Web-age information management: 12th international conference, waim 2011, wuhan, china, september 14-16, 2011, proceedings, 2011.

J. Wu and Z. Lin. Research on customer segmentation model by clustering. In *Proceedings of the 7th International Conference on Electronic Commerce*, ICEC '05, pages 316–318, New York, NY, USA, 2005. ACM.